Base genética del cambio de fase en la langosta del desierto *Schistocerca gregaria*

Rubén Martín Blázquez

Tesis Doctoral

27 de marzo de 2017

Editor: Universidad de Granada. Tesis Doctorales
Autor: Rubén Martín Blázquez
ISBN: 978-84-9163-246-7
URI: http://hdl.handle.net/10481/47052

Departamento de Genética

Facultad de Ciencias

Universidad de Granada

Base genética del cambio de fase en la langosta del desierto *Schistocerca gregaria*

Memoria de Tesis Doctoral presentada por el Licenciado Rubén Martín Blázquez para optar al grado de "Doctor por la Universidad de Granada".

Dirigida por el Doctor:

Dr. Mohammed Bakkali

El doctorado/ *The doctoral candidate* Rubén Martín Blázquez y el director de tesis / *and the thesis supervisor*, Dr. Mohammed Bakkali

Garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

/

Guarantee, by signing this doctoral thesis, that the work has been done by the doctoral candidate under the direction of the thesis supervisor/s and, as far as our knowledge reaches, in the performance of the work, the rights of other authors to be cited (when their results or publications have been used) have been respected.

Granada, a 27 de marzo de 2017

Director de la tesis / *Thesis supervisor*	Doctorando / *Doctoral candidate*
	R Martin
Dr. Mohammed Bakkali	Rubén Martín Blázquez

El presente trabajo se ha desarrollado en los laboratorios del Departamento de Genética en la Facultad de Ciencias de la Universidad de Granada. La investigación realizada ha sido financiada por el Ministerio de Economía y Competitividad del Gobierno de España a través del proyecto con referencia BFU2010-16438. Durante la duración de esta tesis he disfrutado de una beca del programa de Formación de Personal Investigador con referecia BES-2011-043627, asociado al proyecto mencionado anteriormente.

Dedicado a mi familia
y a mis amigos

Índice general

Índice de figuras

Índice de cuadros

Resumen

Las plagas de langosta son un mal que afecta a varias decenas de países, invirtiéndose mucho dinero y recursos tanto en su control como en compensar las pérdidas causadas por éstas. El fenómeno por el cual las langostas desarrollan una tendencia a agregarse y formar enjambres es un caso de polifenismo denominado cambio de fase. Pese a que se tiene un amplio conocimiento sobre aspectos moleculares y no moleculares relacionados con este fenómeno en varias especies de langosta, la regulación genética del mismo aún no está claramente dilucidada. Esta tesis trata de abordar las consecuencias genéticas del cambio de fase en la langosta del desierto *Schistocerca gregaria* Forskal comparando los patrones de expresión genética global que acontecen entre la fase gregaria y la fase solitaria de esta especie. Para ello hemos desarrollado métodos para cuantificar el grado de gregariedad que un individuo de langosta puede presentar, hemos secuenciado, ensamblado, analizado y comparado los transcritos expresados en el sistema nervioso central y el tubo digestivo de adultos gregarios y solitarios. Los resultados se han validado tanto con qPCR como mediante comparación bibliográfica de nuestros datos con los publicados en otros trabajos tanto sobre *S. gregaria* como sobre la langosta migratoria *Locusta migratoria*. Como caso de estudio, hemos analizado la familia de proteínas quimiosensoras (*chemosensory proteins*, CSPs) entre cuyos integrantes hay proteínas implicadas en la activación del cambio de fase.

En el **capítulo 1** de esta tesis desarrollamos una herramienta que permite estimar el estado de gregariedad de un individuo de langosta basándonos en modelos matemáticos con datos extraídos del análisis del comportamiento y morfología de las langostas. Para ello, realizamos grabaciones de cómo respondían individuos criados a distintas densidades de población (aisladas y a 150 individuos por jaula) ante un grupo de langostas gregarias, además

de tomar medidas morfológicas de los individuos experimentales. Con los resultados de estos estudios, se calcularon diez variables de comportamiento y tres de morfometría. Estudiamos el efecto del sexo, el tamaño y la especie sobre las variables y tras atenuar el efecto del tamaño, que se resolvió normalizando las variables de comportamiento con la longitud del fémur de cada individuo, confeccionamos varios modelos para cada especie, basados en formulas de regresión múltiple logística que fueron validadas con grupos externos de langostas criadas en densidades intermedias. Como resultado obtuvimos cuatro ecuaciones de modelos validados con y sin variables morfométricas para *S. gregaria* y para *L. migratoria*. Los modelos sin datos morfométricos los confeccionamos para el estudio comparativo de las mismas langostas que no hayan mudado antes y después de un tratamiento, y los modelos con datos morfométricos fueron concebidos para la comparación de distintas langostas de la misma especie y/o de langostas que hayan mudado entre análisis. Ambos modelos de *S. gregaria* lograron caracterizar a las ninfas con precisión pero sólo el modelo sin morfometría logró caracterizar a los adultos con precisión. Los modelos de *L. migratoria* funcionaron para caracterizar ninfas pero no lograron clasificar correctamente a los adultos. Para facilitar el uso de estas herramientas, confeccionamos en un programa informático escrito en R para extracción de datos de comportamiento a partir de las grabaciones en video y otro para aplicar las formulas a los datos extraídos tanto de los videos como de las medidas morfométricas.

El **capítulo 2** de esta tesis se centra en el estudio del transcriptoma del sistema nervioso central de las fases gregaria y solitaria de *S. gregaria.* Mediante un ensamblaje sin referencia de dos librerías de secuencias de ARN mensajero provenientes de tejido enriquecido en sistema nervioso central de ambas fases gregaria y solitaria. Analizamos funcional y cuantitativamente los genes encontrados en el mismo. La anotación de las secuencias permitió identificar 17.620 transcritos con resultado BLAST único, dejando 34.696 transcritos no redundantes sin anotación conocida. Los análisis de términos funcionales de las bases de datos *Gene Ontology*, *Kyoto Encyclopedia of Genes and Genomes* e *InterProScan* nos muestran enriquecimiento de términos relacionados con la neurotransmisión y la sinapsis, así como otras funciones relacionadas con el transporte de energía, la respuesta a estímulos y la respuesta al estrés. El análisis de expresión diferencial nos mostró que el sistema nervioso central de la fase gregaria presenta una mayor actividad transcripcional presentando casi 10 veces más transcritos sobre-expresados con respecto a la fase solitaria. Los resultados de este capítulo no solamente ofrecen una lista de candidatos a estudiar como secuencias que afecten al cambio de fase, sino que también nos permite estudiar el proceso a nivel de rutas metabólicos y procesos de regulación completos. También nos da una idea de la magnitud y complejidad del cambio de fase a nivel de regulación neural, afectando a la expresión de casi el 50 % de los transcritos (anotados y no anotados) estudiados.

En el **capítulo 3** elaboramos un estudio del transcriptoma del tubo digestivo con metodología análoga al capítulo anterior pero con un enfoque ligeramente distinto. En este caso obtuvimos un transcirptoma de referencia con secuencias de las fases gregaria y solitaria de todo el tubo digestivo, con 57.637 transcritos con único resultado BLAST y 16.491 transcritos sin anotar. Este alto número de resultados BLAST fue debido a la secuenciación de microorganismos simbiontes y patógenos de la flora intestinal. Encontramos que 5.247 transcritos con resultado BLAST provenientes de organismos no pertenecientes al reino animal, de los cuales 1.073 pueden derivar de un apicomplejo del género *Gregarina* spp. Observamos un enriquecimiento de transcritos de regulación por ubiquitina y apoptosis, siendo la profundidad de lecturas alineadas a estos transcritos significativamente mayor en la fase gregaria. Este estudio también permitió descubrir la presencia de transcritos bacterianos (pertenecientes a los géneros *Escherichia*, *Enterobacter* y *Bacteroides*, entre otros) y fúngicos (sobre todo del género *Fomitopsis*). En cuanto a secuencias específicas de la langosta, encontramos que los procesos relacionados con la lisis de proteínas (enzimas digestivos sobre todo), la detoxificación (como las glutatión S-transferasas o los citocromos P450) o el estrés (como las proteínas de choque térmico) son abundantes en general entre los transcritos anotados. Entre los genes diferencialmente expresados encontramos transcritos del sistema inmune, regulación del calcio y apoptosis en la fase gregaria comparado con transcritos de membrana peritrófica y estructuras musculares de la fase solitaria, dando a entender un contraste entre un tubo digestivo gregaria bajo infección y otro tubo digestivo solitario que invierte más en mantenimiento de estructuras como el músculo o la matriz epitelial.

El **capítulo 4** es un estudio comparativo entre los dos transcriptomas provenientes del tejido nervioso y digestivo, así como entre varios trabajos ómicos que abordan el cambio de fase tanto en *S. gregaria* como en *L. migratoria*. Comparando los resultados de la anotación, encontramos que casi 3.000 transcritos únicos presentaban idéntico resultado BLAST entre los dos transcriptomas, y que solamente 14 presentaron idénticos perfiles de expresión. Entre los que presentaron un patrón común de sobre-expresión en gregarios, cabe destacar la pacifastina 4, el inhibidor de serina proteasa 3 y el gen *black*; los dos primeros relacionados con el sistema inmune y el tercero relacionado con la melanización y las catecolaminas. Con respecto a las secuencias no anotadas, al agrupar mediante alineamientos las secuencias de nervioso y digestivo, descubrimos que un buen porcentaje de ellas están presentes en ambos transcriptomas, pudiendo ser posibles secuencias funcionales que no hayan sido caracterizadas aun. La validación mediante qPCR del transcriptoma del sistema nervioso fue un éxito, con 10 de 12 transcritos presentando un patrón de expresión significativo idéntico en los dos estudios, aunque la tendencia se atenúa, sin eliminarse, cuando se compara con muestras de distintos tejidos, estadíos de desarrollo o con otra especie. Para concluir este capítulo, se llevó a cabo un estudio comparativo entre los resultados de expresión de seis trabajos realizados por otros grupos de investigación con distintos objetivos, tecnologías, número de genes y especies, obteniéndose diferentes niveles de concordancia y apuntando a secuencias que posiblemente sean marcadoras o determinantes en el cambio de fase en langostas. Los transcritos con congruencia absoluta resultaron ser dos: el de la anexina IX y una proteína *unkempt* con un motivo RING finger, ambos sobre-expresadas en fase gregaria y con un papel por determinar en el cambio de fase.

El **capítulo 5** abarca el estudio, como caso, de una familia de proteínas, las CSP (*Chemosensory proteins*). Realizamos un estudio de secuencias genómicas y transcriptómicas de *L. migratoria* de acceso público y de nuestros datos de 5 transcriptomas de *S. gregaria* (los dos estudiados en esta tesis y tres más secuenciados por el grupo de investigación: músculo torácico, ovarios y testículo). Realizando un estudio inicial basado en búsquedas BLAST sobre el genoma de *L. migratoria* y alineamientos con fragmentos de secuencias expresadas (ESTs), logramos caracterizar 57 loci que contienen CSPs en esta especie, siendo confirmadas por su patrón de cuatro cisteínas y su estructura típica de dos exones. Análogamente, se realizó otra búsqueda BLAST sobre nuestros transcriptomas de *S. gregaria*, y tras un filtrado de secuencias redundantes basado en la identidad mostrada por las secuencias de *L. migratoria*, así como la búsqueda de posibles transcritos quiméricos, logramos identificar 42 loci de CSPs de *S. gregaria*. Mediante una filogenia con CSPs de otros insectos, logramos identificar parejas de CSPs homólogas entre las dos especies de langosta *S. gregaria* y *L. migratoria*. Además, caracterizamos los patrones de expresión de las CSPs de *S. gregaria* y *L. migratoria* mediante análisis de expresión diferencial de datos ómicos. En total, siete parejas de homólogos compartían un patrón de expresión significativo hacia la fase gregaria, una de las cuales (*LmigCSP3*) está relacionada con la detección de feromonas que propician la agregación, lo que puede indicar que su homólogo (*SgreCSP37*) pueda tener una implicación en el cambio de fase también. Todos los homólogos (tanto intra-específicos como inter-específicos) presentaron indicios de estar bajo selección purificadora. El número de parálogos de CSPs de ambas especies es el mayor conocido entre las CSPs estudiadas hasta el momento.

En conclusión, esta tesis presenta datos y herramientas para el estudio genético y funcional del cambio de fase en *S. gregaria*. Los modelos desarrollados en esta tesis servirán para realizar estudios funcionales sobre cómo afectan tratamientos experimentales el grado de agregación. Nuestros transcriptomas nos han permitido identificar una amplia lista de secuencias que pueden estar implicadas en el cambio de fase, además de otras que podrían servir como dianas para insecticidas. Además, nos indican que las rutas genéticas están más afectadas en el sistema nervioso central de la fase gregaria. Basados sobre los datos transcriptómicos, elaboramos así una interpretación integral de los cambios globales que acompañan el paso al estado gregario. También nos ha supuesto descubrir un potencial nuevo agente de control biológico: una gregarina. Toda esta información ha servido incluso para centrarnos en un caso de estudio, por el cual logramos identificar copias únicas de CSPs en dos especies de langosta, así como sus relaciones filogenéticas y sus perfiles de expresión, lo que es un indicio de la conservación de la función de algunas parejas de CSPs homólogas.

Summary

Locust outbreaks affect near two thirds of the Earth's dry surface, which requires a great investment in their control and repair of the damage they cause. The transformation of isolated and passive locusts into swarming and active locusts is regulated by a striking case of polyphenism called phase change. Despite the knowledge on some potential modulators of the locust phase change, its integrative genetic regulation is not yet understood. This dissertation studies the transcriptional consecuences of phase change in the desert locust *Schistocerca gregaria* Forskal by comparing the overall gene expression profile differences between the gregarious and solitarious phases. For this purpose, we first develope mathematical models in order to quantify the degree of locust gregariousness. Following that, we sequence, assemble, analyze and compare the transcripts and their expression profiles from two adult tissues (central nervous system and digestive tube) between the two phases, gregarious and solitarious. We validated the results using qPCR and by comparisons with the data from scientific publications on the phase change both in *S. gregaria* and in the migratory locust, *Locusta migratoria*. In addition, as case study, we characterize the copy number of chemosensory proteins (CSPs) in both species, and we found some of them to be linked to the phase change in one or both species.

In **chapter 1** we develope mathematical models based on behavioural and morphological data in order to calculate a locust's probability of being gregarious. To fulfill this task, we recorded videos of locusts reared in isolation and in crowding conditions. For that we used an arena with a stimulus group of gregarious locusts. We also measured various morphological traits of the locusts. After assessing the association between gregariousness and several colorimetric, morphometric and behavioural variables, and checking for the

effect of sex, size and species on these variables, we mitigated the effect of the animal's size on its movement-related variables by normalization by the hind femur length (sex had no effect). We then built multivariate logistic regression-based models both for *S. gregaria* and *L. migratoria* and validated them using independent samples of locusts reared in a gradient of intermediate population densities. In total we used thirteen variables for model building, ten behavioural and three morphological. For each of these species, we selected two validated models, one with and one without morphometrical variables. *S. gregaria* models succesfully quantified the gregariousness of nymphs, but only the one without morphometric variables managed to quantify gregariousness in adults. *L. migratoria* models however were noisy for nymphs and could not perform accurately for adults. The models were implemented in an R script that we make available to the locust research comunity.

In **chapter 2** we establish and compare the gregarious and solitarious *S. gregaria* central nervous system (CNS) transcriptome. We performed a de novo assembly from RNA high throughput sequencing reads obtained from adult gregarious and solitarious CNS-enriched samples, followed by a quantitative analysis. We retrieved 17,620 unique BLAST results from the assembled transcripts. 34,696 non redundant sequences had no annotation result. After mapping against Gene Ontology, Kyoto Encyclopedia of Genes and Genomes and InterProScan term databases we found neurotransmission and synapsis-related term enrichement, together with other functions such as energy transport, response to stimuli and stress response. Differential gene expression analysis shows that the number of gregarious CNS up-regulated transcripts is ten-fold higher than in solitarious phase. With these results, we do not only build a complete set of CNS locust sequences, but we also highlight sequences as well as whole genetic pathways as affected by the phase change. We also reveal the complex nature of the phase change at the gene expression level, with almost fifty percent of the transcript sequences being up-regulated in the gregarious CNS.

In **chapter 3** we perform the transcriptomic study of the digestive tube (DT) in an analogous way as we did for the CNS. The refference transcriptome for that tissue contained 57,637 single BLAST results, and 16,491 sequences without annotation. We can explain the high number of unique BLAST results due to the presence of sequences from commensalist and pathogenic microbes from the gut flora. We found near 5,247 transcripts with a BLAST result belonging to non-animal species, with 1,073 of them belonging to the apicomplexa *Gregarina* spp., with enriched representation of ubiquitylation and apoptosis processes, and a higher overall sequencing depth in the gregarious phase. This study also revealed the presence of bacterial (typical enteric bacteria such as *Escherichia*, *Enterobacter* and *Bacteroides*) and fungal (*Fomitopsis* spp.) transcripts. Regarding locust transcripts, we found those for protein lysis (mostly thanks to digestive proteases), detoxification (with glutathione S-transferase and cytochrome P450 variants) and stress (such as heat shock protein variants) to be abundant. Among the differentially expressed genes we found those related to immune response, calcium regulation and apoptosis to be over-expressed in the gregarious phase, while transcripts related to the peritrophic matrix and muscle structure were over-expressed in the solitarious phase. The gregarious gene expression pattern in the locusts' DT is therefore modulated by the phase sensu stricto and by the higher presence of pathogens. The solitarious locust DT, in contrast, shows a predominantly housekeeping and maintenance gene expression profile.

In **chapter 4** we compare *S. gregaria*'s CNS and DT transcriptomes between each other and to the data in several published works on locust phase change. Comparing the unique BLAST results from both transcriptomes we found 2,772 shared results, but only 14 of them show congruent significant differential expression. Among the gregariously up-regulated congruent transcripts were *black*, involved in melanization and catecholamine metabolism, and pacifastin 4 and serin protease inhibitor 3, both related to insect immune response regulation. We clustered by the sequences that had no BLAST result from both transcriptomes and discovered that a good percentage of them are homologous, indicating that they are transcribed in both transcriptomes and that they are genuine unknown transcripts. We succesfully validated the expression pattern of 10 out of 12 genes from the CNS transcriptome using qPCR. Further comparisons support our in silico data, although, the trends expectedly dissipated when comparing samples from different tissues, develpmental states and even with *L. migratoria.* We also compared our data to the data from six peer reviewed published works, obtaining different agreement rates and the confirmation of some phase-related genes. Only two transcripts consistently showed the same expression profile in all these works. These were Anexin IX and a RING finger containing protein unkempt, both up-regulated in gregarious phase on several tisues from both *S. gregaria* and *L. migratoria.*

Chapter 5 consisted on the case study of the chemosensory protein (CSP) family. For that we used all the genomic and transcriptomic sequences available for *L. migratoria* and *S. gregaria* (including the raw sequencing data that our research group has on other transcriptomes from toracic muscle, ovaries and testicles). An initial BLAST search for CSPs in the *L. migratoria* genome scaffolds and EST database followed by alignment with known CSPs let us to identify 57 CSP-containing loci. They all showed the CSP-specific cystein pattern and its two exon genomic structure. Similarly, another BLAST search using the several *S. gregaria* transcriptomics data of our research group identified 42 CSP transcripts probalby belonging to different CSP loci. These numbers of CSP paralogs in both locust species is the highest value reported in insects with known CSP number. It is to highlight that the identified sets of locust CSP transcripts were filtered for redundancy and potential chimeric sequences. Phylogenetical analysis using the aminoacid sequences of the identified locusts' CSPs and those other insects, downloaded from the NCBI database, allowed us to group several CSP homologous pairs from both locust species. All homologous pairs (both intra- and inter-specific) present evidence of being under purifying selection. After comparison of the expression patterns of all the locusts' CSPs between the solitarious and gregarious phase, we found that seven inter-specific homologous pairs shared the same expression pattern. These are therefore associated with the phase change and, one of them, *LmigCSP3*, is known to be involved in the detection of *L. migratoria*'s agregation pheromones. The homologous sequence that we report here for *S. gregaria*, *SgreCSP37*, might thus very likely be involved in detection of this species aggregating pheromones too.

In summary, this thesis updates the study of phase change in *S. gregaria* with a behavioural tool and high amounts of genetic data. The behavioural models developed will facilitate and standarize the functional study of the effect of an experimental treatment in phase change. Our transcriptomes contribute with a wide set of interesting sequences probably involved in phase change and other set of interesting sequences that might be explorable for targeting locusts. In addition, we find that the gregarious CNS presents the highest number of pathways affected or involved in the phase change. We also identified transcripts from a potential biological control agent: a gregarine. The case study presented here offers an example on how to identify putative genes, their copy number, phylogenetic relationships and expression profiles in the gregarious and solitarious phases, which ultimately leads to revealing the potential association of the concrete transcripts with the phase change.

Introducción

La relevancia de las plagas de langostas

El ser humano lleva estableciendo cultivos para su sustento desde hace más de 10.000 años. Inevitablemente, estas grandes extensiones de vegetales comestibles también atraen a un buen número de especies que, al igual que el ser humano, le sirven para alimentarse de manera eficiente. Por conveniencia llamamos plagas a estas especies, las cuales han obligado a la humanidad a desarrollar métodos para evitar el daño que infligen simultáneamente al desarrollo de la agricultura. De entre todas las plagas, una de las más sufridas por la especie humana es sin duda la plaga de langostas. Ya en la antigüedad aparecen nutridas referencias a este fenómeno: son mencionadas en informes de la antigua China, en el Antiguo Egipto y civilizaciones de la América precolombina las representa en murales y esculturas, incluso aparecen en textos religiosos tales como la Biblia o el Corán.

En la actualidad, las langostas siguen suponiendo un grave problema para los agricultores, que ven sus cultivos diezmados por los enjambres cuando las poblaciones de langostas experimentan explosiones demográficas. Además, la mayoría de zonas afectadas pertenecen a países en desarrollo, llegando a provocar hambrunas durante las plagas más graves. No es de extrañar que la Organización de Alimentos y Agricultura de las Naciones Unidas (*Food and Agriculture Organization*, FAO) dedique muchos esfuerzos al control y erradicación de plagas de langostas. En su página web, la sección Locust watch [FAO, 2009] se encarga de monitorizar el estado de los enjambres cada mes, además de recopilar información sobre brotes en el pasado y presentar modelos predictivos de formación, localización y desplazamiento de enjambres de langosta. Pese a los perjuicios que las plagas de langostas provocan, aparecen entre la lista de insectos que la misma FAO recomienda consumir como alternativa a otros alimentos [FAO, 2003a], y muchos trabajos actuales sobre langostas se centran en estudiar su valor nutricional para inculcar su consumo en otras culturas [Cerritos, 2009, Xiaoming et al., 2010, Oonincx and Van der Poel, 2011, Cheseto et al., 2015, Mohamed, 2016].

Podemos definir langosta como un grupo filogenéticamente heterogéneo de ortópteros pertenecientes a la familia Acrididae, cuya característica común es la capacidad de formar enjambres bajo condiciones de alta densidad de población. Con respecto a su ecología, las especies de langosta más comunes viven en llanuras aluviales de áreas desérticas o pobres en nutrientes para las plantas, donde las escasas plantas distribuidas de manera esparcida les permiten vivir aisladas, bajo la protección de su coloración críptica. Durante la época de lluvias, se forman parches de vegetación que atraen a las langostas circundantes, lo que hace que las langostas concentren sus efectivos. Además, el suelo húmedo y consolidado por las raíces de las plantas permite a las puestas de huevos tener mayor probabilidad de supervivencia, lo que produce un incremento considerable en el número de efectivos en un área relativamente pequeña. Algo similar ocurre con otras especies de langostas que, al vivir en praderas y debido a los cambios tanto de usos de suelo como de calidad nutricional de las plantas que crecen allí, se ven también concentradas en parches de vegetación preferida por las langostas [Latchininsky, 2013, Cease et al., 2015]. En cualquiera de estos dos escenarios aumenta la densidad de población de langostas, produciéndose cambios drásticos en su comportamiento y su morfología: tanto las ninfas (llamadas también saltones) como los adultos tienden a agregarse activamente en enjambres que pueden llegar hasta los veinte kilómetros de extensión [Simpson and Sword, 2008], su voracidad aumenta de tal modo que pueden llegar a comer su propio peso en un solo día [Davey, 1954], su capacidad de dispersión mejora llegando a cubrir en vuelo distancias de hasta 1.000 kilómetros al día [COPR, 1982] y sus colores cambian de cenicientos y crípticos a motivos mucho más llamativos como el negro, el amarillo o el rojo [Stower, 1959, Ellis, 1964]. Es en ese momento cuando se forma el enjambre, siendo el cambio de fase un proceso dependiente de la densidad de población mediante el cual se llega a pasar de una fase solitaria a una fase gregaria. Cuando el enjambre entra en declive por la escasez de recursos (proceso que ocurre tras unos meses o incluso durante varios años, dependiendo del tamaño del enjambre y de la especie), los individuos empiezan a morir y la población revierte poco a poco a su estado inicial de individuos aislados, lo que hace al cambio de fase un proceso reversible. Este proceso se suele repetir en ciclos de entre 7 y 15 años, dependiendo de la especie, coincidiendo con períodos de bonanza climática para los cultivos [Simpson and Sword, 2008, Latchininsky, 2013].

Sin duda, una de las especies de langosta que más perjuicios origina (presentando los enjambres más grandes documentados) es la langosta del desierto, *Schistocerca gregaria.* Su rango de distribución abarca la mitad norte del Continente Africano, Oriente Próximo, Oriente Medio y la India (figura 1). A pesar de no ser la especie de langosta con mayor rango geográfico de distribución (título ostentado por la langosta migratoria, *Locusta migratoria*), esta especie es altamente polífaga (el número de plantas de las que se alimenta supera las 500 especies [COPR, 1982], presenta los mayores enjambres documentados (llegando a cubrir unos 800 Km^2, con un número de individuos máximo estimado de 40 mil millones, [Simpson and Sword, 2008]) así como la mayor distancia recorrida por cualquier especie de langosta (ni más ni menos que un vuelo transatlántico de 5.000 Km, [Simpson and Sword, 2008]). Curiosamente es la única especie del género *Schistocerca* presente en el Viejo Mundo, estando el resto de sus representantes en el Continente Americano. Aunque se han documentado llegadas puntuales de enjambres de esta especie en el Algarve y las Islas Canarias, su rango de distribución no alcanza a Europa, quedándose entre el norte de África y el Sahel, Arabia Saudí y gran parte de Oriente medio (figura 1). Como se ha mencionado antes, hay casos documentados (aunque excepcionales) de migración transatlántica, en los que algunos enjambres de esta especie lograron alcanzar las costas norteamericanas desde África, posiblemente usando barcos y balsas de individuos muertos como plataforma de descanso ocasional [Simpson and Sword, 2008]. Este hecho explicaría la presencia de representantes del género tanto en el continente africano como en el americano.

La plaga de langosta del desierto más reciente ocurrió durante 2004, devastando las cosechas en el Noroeste Africano a lo largo del año. Actualmente el control de esta especie se basa en la fumigación de grandes superficies afectadas con una mezcla de insecticidas con fenilacetonitrilo (*phenylacetonitrile* en inglés, PAN), un compuesto químico que las propias langostas secretan como feromona de la agregación [Torto et al., 1994, Njagi et al., 1996] e inhibición del cortejo [Ferenz and Seidelmann, 2003], que en altas dosis es tóxico para los estadíos juveniles [Kane, 2012, Bal and Sidati, 2013]. También el uso de microorganismos se está utilizando para combatir la plaga: una cepa del microsporidio Metarhizium anisopliae var. acridum (comercializada como Green Muscle) ataca específicamente a los acrídidos y su uso en combinación con el PAN parece aumentar sensiblemente la mortalidad en la langosta del desierto [FAO, 2003b]. Ciertas cepas de otro microsporidio, Paranosema locustae, parecen prevenir la formación de enjambres en la langosta migratoria *L. migratoria* [Shi et al., 2014, Feng et al., 2015], pero su efecto en la langosta del desierto aún no ha sido estudiado y tampoco está comercializado. El problema de estos métodos de lucha anti-acridiana es su inespecificidad, lo que podría afectar a otras comunidades de insectos si se abusa de su uso, por no decir que podría llegar a afectar a los seres humanos (en este caso debido a los insecticidas utilizados en combinación con el PAN y Metarhizium). La tele-predicción de formación de enjambres también ayuda a combatir la plaga: con los métodos de análisis de imagen vía satélite, se puede inferir posibles focos de formación

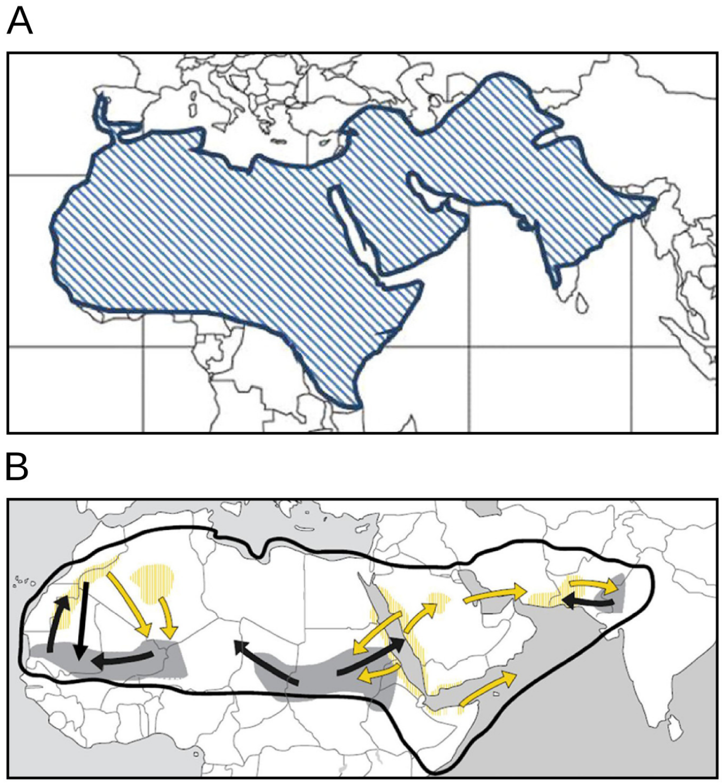

Figura 1: Distribución de la langosta del desierto, *Schistocerca gregaria*. A: La superficie rayada representa el área de invasión de *S. gregaria* (allí donde se ha documentado por lo menos alguna vez). B: Áreas de cría de *S. gregaria*. El área de distribución está delimitada por una línea continua negra, donde se muestran las áreas de cría de invierno (amarillo) y verano (gris oscuro), con sus respectivos flujos migratorios (flechas del color correspondiente a cada área de origen). Tomado de Latchininsky [2013].

de enjambres comparando valores de pluviometría e incrementos de verdor en áreas específicas, facilitando coordenadas donde brigadas anti-acridianas puedan utilizar preventivamente estos compuestos para prevenir la formación de plagas [Latchininsky, 2013].

Plasticidad fenotípica, polifenismo y cambio de fase

Al igual que las langostas, muchas especies que presentan un fenotipo específico para un carácter si se han desarrollado en unas condiciones ambientales concretas, pueden presentar otro fenotipo distinto para ese carácter de haberse desarrollado en condiciones ambientales distintas. Este fenómeno, llamado plasticidad fenotípica, implica que a partir de un único genotipo se puedan generar distintos fenotipos mediante mecanismos genéticos modulados por el ambiente durante el desarrollo del organismo, sin que esto necesite la aparición de mutaciones en su genoma. Los fenotipos resultantes para un carácter se pueden presentar como un gradiente (en cuyo caso hablaremos de norma de reacción) o bien como dos fenotipos extremos, cada uno adaptado a condiciones ambientales distintas (en cuyo caso hablaremos de polifenismo). Aunque el polifenismo se podría considerar una norma de reacción donde a pesar de la clara separación de los fenotipos extremos se pueden dar fenotipos intermedios, al encontrarse éstos a una frecuencia extremadamente baja en las poblaciones naturales se recurre a diferenciar entre norma de reacción y polifenismo.

Hay que puntualizar que la epigenética no es un fenómeno Lamarckista (el ambiente hace la función), aunque varios trabajos la enmarcan como un complemento a la teoría de la heredabilidad enmarcada en el neo-Darwinismo bajo el término neo-Lamarckismo, donde habría una componente “fuerte” (a nivel de genoma) y otra “débil” (a nivel de marca epigenética) en la herencia [Jablonka and Lamb, 1999, Jablonka and Raz, 2009]. Para comprender mejor esta visión podemos utilizar este ejemplo: imaginemos una población de un organismo que presenta plasticidad fenotípica en la que hay cinco individuos (genotipos) distintos. Esta población puede ser sometida a dos ambientes distintos, A y B, (figura 2), donde los mismos cinco individuos pueden variar su fenotipo dependiendo si se han desarrollado bajo las condiciones A o B. De esta manera, observaremos que hay una varianza intra-poblacional regida por la varianza genética, hay una varianza inter-poblacional regida por la varianza ambiental, y hay distintas capacidades de respuesta en cada individuo, regidas por la interacción genética-ambiente. Así pues, existe variabilidad en la capacidad de respuesta (epigenética), con la posibilidad de estar asociada a adaptaciones que aportarán más o menos a la eficacia biológica (fitness) de cada individuo, todo ello definido por unidades heredables (genes), lo que nos ofrece todos los requisitos para que haya selección natural. Por lo tanto, en cuanto a la evolución de la plasticidad fenotípica, podemos afirmar que se seleccionan caracteres con la capacidad de cambiar el fenotipo del individuo dependiendo del ambiente, no siendo el ambiente el que genere el fenotipo en sí.

En insectos, la plasticidad fenotípica está presente en muchas especies y

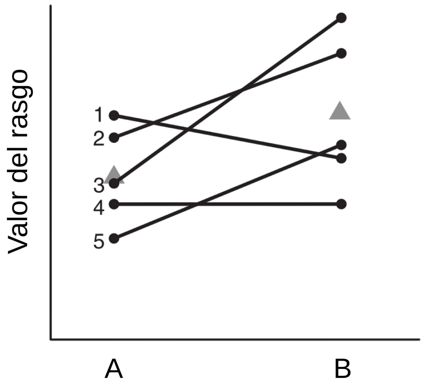

Figura 2: Gráfico explicativo de un caso hipotético de plasticidad fenotípica. En el eje X se muestran los dos ambientes a los que se ha sometido a las dos poblaciones, A y B. En el eje Y se presentan los valores fenotípicos que toma el rasgo para los individuos del 1 al 5. Los triángulos son los valores medios del rasgo en cada ambiente y las líneas indican la varianza del rasgo entre ambientes para cada individuo. Modificado de Whitman and Agrawal [2009].

además aparece modulada por distintos factores ambientales. Un ejemplo ilustrativo es el del escarabajo *Onthophagus taurus*, que dependiendo de la calidad del alimento que hayan tomado las larvas de machos, desarrollarán cuernos de mayor o menor longitud que determinarán su estrategia reproductiva [Moczek, 1998]. Los exitosos insectos eusociales, como las hormigas, las abejas y las termitas, también presentan plasticidad fenotípica, que determina su sistema de castas de manera más o menos estricta según la especie [Evans and Wheeler, 1999]. Muchas especies plaga también presentan plasticidad fenotípica, lo que en muchos casos les otorga ventajas con respecto a otras especies. Un claro ejemplo son los áfidos como *Aphis phabae*, que alternan varias generaciones de hembras partenogenéticas ápteras con otra de machos y hembras con alas en condiciones de sobrepoblación [Johnson, 1966, Shaw, 1970]. En la polilla africana *Spodoptera exempta*, también es la densidad de población la que modula su polifenismo: cuando se crían en condiciones de hacinamiento, las orugas se vuelven más voraces, arrasando grandes áreas de cultivo de cereales y desplazándose en busca de otras áreas por la noche [Faure, 1943, Simmonds and Blaney, 1986].

Las langostas nos ofrecen un claro ejemplo de polifenismo: presentan un morfotipo plaga llamado fase gregaria y un morfotipo aislado llamado fase solitaria. Los cambios producidos por el polifenismo afectan a muchos aspectos de la biología de estos organismos. Las diferencias entre fases pueden ser tan acentuadas que incluso individuos de la misma especie pero distinta fase han sido descritas en el pasado como dos especies totalmente distintas. De hecho, no fue hasta principios del siglo XX cuando el científico ruso Boris Uvarov identificó como una sola especie a las distintas fases de la langosta migratoria *L. migratoria* [Uvarov, 1921]. En sus trabajos iniciales sobre esta langosta, Uvarov describió los cambios más trascendentes entre la fase gregaria, la fase solitaria y las fases transitorias, pero mediante análisis morfométricos posteriores el número de fases reconocidas en la langosta migratoria se establecieron finalmente como dos: gregaria y solitaria [Stower et al., 1960]. Teniendo en cuenta toda la información mostrada en este epígrafe, podemos definir el cambio de fase en langostas como el proceso por el cual individuos pertenecientes a una fase sufren una serie de cambios regidos por el ambiente (normalmente mediante cambios dependientes de la densidad de población) que provocan que desarrollen los caracteres fenotípicos de la otra fase.

La diferencia más llamativa entre las dos fases es sin duda el color de las ninfas: mientras que las ninfas solitarias presentan tonos crípticos y de patrones claros (verde, marrón claro o gris dependiendo de la especie), las ninfas gregarias presentan un patrón negro, muchas veces acompañado de tonos rojos o amarillos (figura 3). El color de las ninfas solitarias depende de la dieta o incluso del sustrato donde la langosta se esté criando [Pener and Simpson, 2009], pero en el caso de la fase gregaria parece ser una señal aposemática que avisa a los depredadores sobre su toxicidad o mal sabor [Sword et al., 2000]. Se podría considerar también la posibilidad de que este cambio de color sea una consecuencia metabólica del cambio de fase (como se discutirá más adelante) que pese a su vistosidad no supone una desventaja para los individuos debido a que viven en forma de enjambre y así minimizan individualmente las posibilidades de ser depredados. Los adultos también pueden presentar diferencias de color: en *L. migratoria* los adultos solitarios suelen presentar vetas de color verde en cabeza y tórax, los adultos gregarios presentan un homogéneo color marrón; mientras que en *S. gregaria* los adultos solitarios presentan colores marrones y cenicientos y los adultos gregarios inmaduros presentan un color rosado más o menos intenso, cambiando a amarillo en machos y a marrón en hembras conforme van alcanzando la madurez sexual (figura 4).

Las dos fases también presentan diferencias en peso y tamaño, siendo los individuos gregarios en promedio más ligeros y más pequeños que los individuos solitarios. Además, existen diferencias en las proporciones morfométricas: los individuos solitarios presentan fémures traseros y alas relativamente más cortas que los individuos gregarios, lo que ha permitido la realización de índices morfométricos que pueden discriminar poblaciones solitarias de poblaciones gregarias basándose en los valores de las relaciones entre la longitud del ala (E), la anchura de la cabeza (C) y la longitud del fémur trasero (F). Concretamente, el índice E/F es mayor y el índice F/C es menor en individuos gregarios de varias especies de langostas [Stower et al., 1960, Ellis, 1963, Uvarov, 1977]. Se han documentado también diferencias tanto

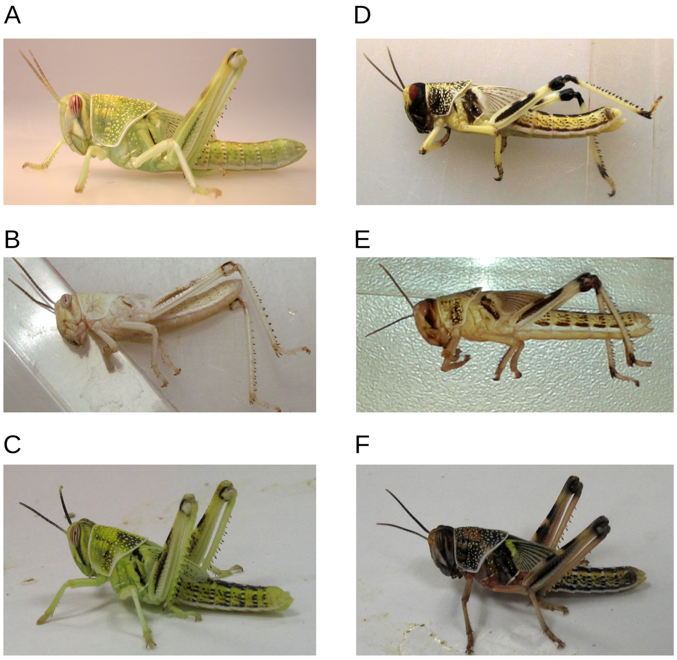

Figura 3: Ninfas de quinto estadío de la langosta del desierto *Schistocerca gregaria.* En estas fotos se observa que existe variabilidad incluso dentro de la misma fase. A: Ninfa solitaria verde. B: Ninfa solitaria pálida (cuarto estadío). C: Ninfa gregaria criada en aislamiento. D: Ninfa gregaria. E: Ninfa gregaria con ojos claros. F: Ninfa gregaria con alto grado de melanismo.

en el número como en los tipos de sensilas (órganos sensoriales que aparecen como prolongaciones cuticulares parecidas a pelos en insectos) entre individuos de fase solitaria y gregaria, siendo mayor el número de sensilas olfativas en antenas de individuos solitarios que en gregarios [Tawfik, 2012]. También está comprobado que los individuos solitarios tienen mejor percepción auditiva que los gregarios [Gordon et al., 2012], lo que apoyaría la hipótesis de que los individuos solitarios podrían ser más receptivos a estímulos del medio para poder detectar a posibles depredadores o incluso a otros individuos de su misma especie, al contrario que los individuos gregarios, con menor capacidad de detección pero protegidos por los números que ofrece el enjambre. Tanto los cambios morfométricos como los cambios en la densidad de sensilas sólo pueden ocurrir entre muda y muda, ya que estas estructuras se remodelan cuando se forma la nueva cutícula del individuo, y a veces se detectan gradualmente

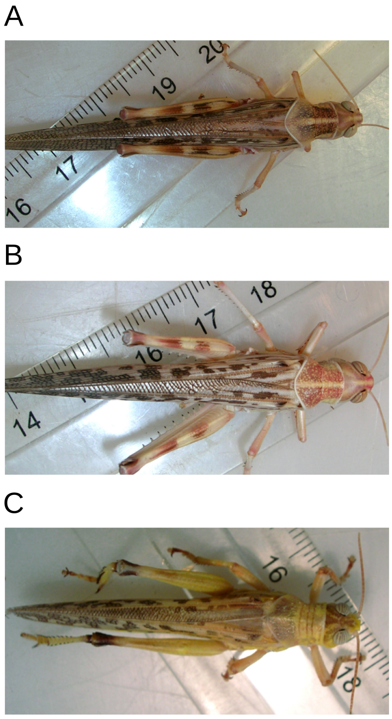

Figura 4: Adultos de la langosta del desierto *Schistocerca gregaria.* A: Adulto solitario. B: Adulto gregario inmaduro. C: Adulto gregario que ha alcanzado la madurez sexual.

(como el cambio de color) o a más largo plazo (en el caso de la variación de los índices morfométricos). Cabe destacar también las diferencias de longevidad entre fases, siendo los individuos solitarios más longevos que los individuos gregarios [Boerjan et al., 2011].

Otro bloque importante son los cambios en el comportamiento. Los individuos gregarios presentan una clara tendencia a la atracción hacia otros individuos y a llevar un modo de vida en agregación, cualidad que le da nombre a la fase. Por el contrario los individuos solitarios viven esparcidos, evitando activamente a otros individuos salvo durante su madurez sexual, cuando buscan más activamente miembros del sexo opuesto para aparearse. Hay también diferencias en la rapidez de la toma de decisiones: los individuos gregarios toman decisiones con mayor rapidez, pudiéndose decir que tienen más claro lo que quieren (acercarse a otros individuos o a fuentes de alimento), mientras que los individuos solitarios son más lentos a la hora de tomar una decisión, y cuando la toman suele implicar alejarse a un grupo de individuos o acercarse a una fuente de alimento [Behmer et al., 2005, Lancet and Dukas, 2012, Simões et al., 2013]. También las capacidades motoras cambian: los individuos gregarios muestran mayor grado de actividad que los solitarios, lo que permite que enjambres de millones de individuos se desplacen rápidamente a través de largas distancias. Hay trabajos que justifican la movilidad del enjambre como adaptación a la evitación del canibalismo, suceso bastante frecuente entre los acrídidos [Lockwood, 1988, Bazazi et al., 2008], y que la movilidad de los enjambres podría ser una adaptación para evitar estar cerca de individuos potencialmente caníbales, por lo menos en estadíos juveniles [Bazazi et al., 2008, Guttal et al., 2012]. El conjunto de estos cambios enumerados más arriba son utilizados para determinar el estado de fase de un individuo, pudiéndose utilizar en su conjunto para elaborar una estima de fase, como veremos en el primer capítulo de esta tesis.

Para que se dé el cambio de fase deben ocurrir principalmente dos acontecimientos: presencia (o ausencia) continuada de un estímulo "gregarizante" y una respuesta hormonal acoplada que desencadene el proceso de dicho cambio. Como se ha nombrado anteriormente, los enjambres de *S. gregaria* se forman cuando aparecen islas de vegetación muy localizadas que atraen a individuos de toda el área circundante. El cambio de fase se puede interpretar evolutivamente como una adaptación a la futura falta de alimento cuando el número de langostas aumenta desmesuradamente de manera local, activándose mediante la densidad de población como estímulo desencadenante, no con la abundancia de alimentos [Ellis, 1963]. Si las langostas utilizaran la falta de recursos en lugar que la densidad de población, los recursos ya esquilmados no podrían sustentar al enjambre. Lo más eficiente sería detectar un estímulo anticipado a la falta de recursos en lugar de activarse directamente ante la escasez. Por eso la densidad de población actúa como variable desencadenante del cambio de fase: cuando las langostas detectan una concentración de individuos mayor a la que están acostumbrada, empiezan a experimentar el cambio en anticipación a un aumento de la competencia, lo que le confiere ventaja a la hora de competir con otras especies. Cabe mencionar que el cambio de fase es reversible: un individuo gregario (o su descendencia) puede volverse solitario si se dan las condiciones, y viceversa [Nolte, 1963]. No obstante, la reversibilidad no es igual de rápida en los dos sentidos: mientras que un individuo solitario de *S. gregaria* sometido de una a cuatro horas a estímulos gregarizantes puede presentar signos de gregarización, conseguir el efecto contrario es mucho más costoso en tiempo y a veces es necesario que pase una o varias generaciones para conseguir un grado de solitarización completo [Simpson et al., 1999]. También es verdad que dependiendo de la especie, la tendencia de la velocidad del cambio puede invertirse: mientras que la langosta australiana *Chortoicetes terminifera* también presenta una rápida gregarización y una lenta solitarización [Gray et al., 2009], pero *L. migratoria* presenta una lenta gregarización y una rápida solitarización [Guo et al., 2011].

A pesar de que tanto los desencadenantes externos como el valor adaptativo del cambio de fase están claros, su origen evolutivo en los acrídidos no está claro del todo. Dentro de la filogenia de los acrídidos, podemos observar que no hay un clado que agrupe a las especies de langostas que forman plagas, sino que se sitúan dispersas entre distintas subfamilias. Esto conlleva que especies que forman plagas puedan estar más relacionadas evolutivamente con otras especies que no forman plagas, incluso dentro del mismo género o subfamilia [Song, 2005, Song and Wenzel, 2008]. Tampoco la localización geográfica de las langostas es restringida: varias especies del género Melanoplus o del género *Schistocerca* se ubican en el Continente Americano, mientras que otras especies como *S. gregaria*, *L. migratoria*, la langosta roja *Nomadacris septemfasciata*, la langosta italiana *Calliptamus italicus* o la langosta marroquí *Dociostaurus maroccanus* se distribuyen en regiones más o menos amplias del viejo mundo, sin contar la presencia de *C. terminifera* en Australia [Song, 2010]. Este hecho da lugar a dos posibles hipótesis evolutivas: o bien el cambio de fase (la adaptación para formar enjambres) puede haberse originado de manera independiente en varios taxones (homoplasia), o bien que este carácter estuviera presente en el antecesor común de todas las langostas y se haya perdido o mantenido dependiendo del taxón. En el caso de *Schistocerca*, los resultados de estudios recientes entre distintas especies de este género apoyan la segunda opción, ya que al someter a especies de este género que no forman enjambres a estímulos gregarizantes se observó la aparición de caracteres típicos de fase gregaria, pudiendo ser reminiscencias evolutivas del cambio de fase como carácter ancestral [Gotham and Song, 2013]. Curiosamente, a pesar de haber una mayor diversidad de especies del género en el Continente Americano, recientes estudios moleculares basados en marcadores mitocondriales ponen de manifiesto que la dirección de migración ancestral fue desde África (cuyo único representante del género es *S. gregaria*) hacia Sudamérica, prosiguiendo una diversificación del género en el Nuevo Mundo [Lovejoy et al., 2006], lo que explicaría por qué muchas especies del género *Schistocerca* que no forman plaga puedan presentar estas reminiscencias del cambio de fase.

Base fisiológica del cambio de fase

Se ha comprobado que varios estímulos están implicados en la inducción del cambio de fase, todos ellos obviamente relacionados con la densidad de población. Sin ir más lejos, tras ejercer fricción sobre las patas traseras de ninfas solitarias se manifiestan comportamientos similares a los de fase gregaria en un intervalo relativamente corto de tiempo, entre una y cuatro horas [Simpson et al., 2001], lo que demuestra que el mismo contacto físico entre individuos provoca el cambio de fase. No obstante, la parte del cuerpo que desencadena el estímulo al ser frotada depende de la especie, siendo el fémur trasero en *S. gregaria* [Rogers et al., 2003] o las antenas en *C. terminifera* [Gray et al., 2009] las zonas más sensibles a desencadenar el cambio de fase por fricción. El estar en contacto con individuos gregarios también hace que saltones solitarios se comporten como gregarios en relativamente poco tiempo [Simpson et al., 1999]. Para la langosta del desierto *S. gregaria*, se ha llegado a estimar un umbral de densidad de población crítica de 2,45 saltones por metro cuadrado, siendo cualquier valor superior susceptible de indicar el inicio de la formación de un enjambre [Cisse et al., 2015].

Como otros insectos, las langostas secretan una gran cantidad de compuestos químicos al medio, entre los cuales también se encuentran agentes iniciadores del cambio de fase. En exudados de machos adultos gregarios se ha encontrado principalmente PAN, el cual parece presentar un papel activo en la agregación [Torto et al., 1994, Deng et al., 1996] y en la inhibición del cortejo [Ferenz and Seidelmann, 2003], aunque no parece que actúe solo, sino acompañado de otras sustancias. Otra sustancia que encontramos en exudados es el guaiacol [Nolte, 1963, Torto et al., 1994], feromona de atracción cuya síntesis parece depender de la enterobacteria Pantoea agglomerans, alojada en el intestino medio o en el recto de las langostas [Dillon et al., 2000, Dillon and Charnley, 2002]. El estímulo visual parece tener un efecto gregarizante también, pero comparado con los estímulos mecánicos, apenas tiene relevancia en el cambio de fase [Simpson et al., 1999, Tanaka and Nishide, 2013, Rogers et al., 2014]. También se ha comprobado si otro disparador del cambio de fase podría encontrarse en la calidad nutritiva de la fuente de alimento, pero aunque sí es verdad que el alimento tomado influye en la coloración de la langosta, no parece afectar a su estado de fase [Simpson et al., 1999, 2002].

Una vez el estímulo gregarizante haya alcanzado al individuo, el sistema nervioso de la langosta lo interpreta elaborando una serie de respuestas que generarán los varios cambios descritos más arriba. Las hormonas y los neurotransmisores son los encargados de comunicar al resto de tejidos los cambios fisiológicos que determinan cambios biológicos como el comportamiento o el desarrollo. La secreción de ambas sustancias está regulada de primera mano por estímulos neurales. En el cambio de fase hay descritas varias de estas moléculas como activadoras del proceso. Como ocurre en otros casos de plasticidad fenotípica en insectos, parece haber una relación entre el cambio de fase y los ciclos de hormona juvenil y ecdisteroides. Estas dos hormonas específicas de artrópodos se expresan durante todo el ciclo de vida de forma antagónica: durante los períodos entre mudas la hormona juvenil mantiene un nivel de expresión alto y constante hasta que los ecdisteroides aumentan su concentración en la hemolinfa, momento en el que se produce la muda o ecdisis, tras la cual los niveles de ecdisteriodes en linfa bajan y aumentan los de hormona juvenil. Hay múltiples trabajos en los que se acoplan los cambios en polifenismo con picos de hormona juvenil, y no solamente en langostas [Roussel, 1993, Wiesel et al., 1996, Tawfik and Sehnal, 2003], también en insectos eusociales, lepidópteros gregarios y otros [Corona et al., 2007, 2013, Libbrecht et al., 2013, Sen et al., 2013]. No obstante, que se solape el cambio de fase con la muda puede ser debido a que en insectos sólo se pueden generar ciertos cambios (sobre todo morfológicos) tras mudar, por lo que esta relación con el polifenismo puede ser más bien una coincidencia inevitable más que una causa.

Como se ha mencionado antes, el cambio de fase se regula no sólo a través de hormonas, también a través de neurotransmisores. Concretamente dos de ellos desencadenan respuestas típicas del cambio de fase al ser inyectados en las langostas: la corazonina y la serotonina o 5-hidroxitriptamina (5HT). La corazonina, descrita por primera vez en la cucaracha Periplaneta americana, es un cardiopéptido específico de artrópodos cuya principal función está aumentar el ritmo cardíaco, como sugiere su propio nombre [Veenstra, 1989]. En la langosta del desierto se identificó como una sustancia relacionada con el cambio de fase que se producía en los cuerpos cardíacos (corpora cardiaca) de su sistema nervioso [Tawfik et al., 1999]. Al inyectar corazonina en ninfas solitarias de langosta del desierto y de langosta migratoria se consiguió que aparecieran patrones negros en la cutícula y que los índices morfométricos utilizados para asesorar el cambio de fase [Stower et al., 1960] presentasen valores más parecidos a los de individuos gregarios, proponiéndose así como desencadenante del cambio de fase [Tawfik et al., 1999, Maeno et al., 2004, Sugahara et al., 2015]. Sin embargo, no parecen haberse encontrado diferencias de comportamiento en la langosta del desierto tras las inyecciones, a pesar de que las ninfas inyectadas, al igual que en *L. migratoria*, emergieron como adultos con proporciones morfométricas más parecidas a las de la fase gregaria [Hoste et al., 2002]. En un estudio más reciente se ha logrado identificar que el albinismo en estas dos especies está ligado a mutaciones a componentes de la regulación de de la corazonina. Por ejemplo, un caso de albinismo en *L. migratoria* parecía producirse por una deleción de diez nucleótidos en la secuencia de la corazonina [Sugahara et al., 2017]. Otro caso, documentado en *S. gregaria*en el mismo trabajo, fue asociado a una mutación sin sentido (aparición de un codón de terminación prematuro) en la secuencia codificante de un receptor de corazonina.

Con respecto a la serotonina, al inyectar tanto 5HT como análogos de la misma en ninfas solitarias se consiguió hacer que el comportamiento de ninfas solitarias se tornara más parecido al de ninfas gregarias [Anstey et al., 2009, Rogers et al., 2014]. Además, inyectando antagonistas o inhibidores de síntesis de la 5HT en ninfas gregarias se revirtió el comportamiento gregario de la plaga a un estado más parecido al de fase solitaria [Anstey et al., 2009]. No obstante, no parece generar diferencias en el color tras el aumento de la concentración de serotonina en la hemolinfa, por lo que parece que active un cambio en el comportamiento a corto plazo. Es más, en el estudio de [Tanaka and Nishide, 2013] no se obtiene cambio de color ni cambio en la tendencia de agregación en *S. gregaria* tras la inyección de serotonina, mientras que [Guo et al., 2013] ocurre lo contrario en *L. migratoria*, promoviendo la aparición de caracteres solitarios. Podemos concluir que tanto los efectos de la corazonina como los de la serotonina, a pesar de los resultados tan contundentes obtenidos tras su inyección, parecen más bien una respuesta aguas abajo que un disparador del estímulo gregarizante, aunque está claro que ambos neuropéptidos están implicados en la modulación del cambio de fase.

Continuando con la serotonina, se sabe que este neurotransmisor interacciona con receptores acoplados a proteínas G o GPCRs [Wang et al., 2013, McCorvy and Roth, 2015], y una de las varias funciones de estos receptores es la síntesis de AMPc (adenosinmonofosfato cíclico) gracias a la activación de la adenilato ciclasa mediante proteínas G [Pierce et al., 2002]. Esta molécula activa proteínas kinasa dependientes de AMPc (PKA), que a su vez fosforilan otras proteínas. Entre las proteínas activadas por las PKA aparecen los factores de transcripción CREB (cAMP-response-element-binding), que, entre otras funciones, activan genes relacionados con la remodelación sináptica en neuronas sensoriales, lo que puede llevar a modular el comportamiento del animal [Hegde et al., 1993]. Siguiendo la lógica de que la serotonina está implicada en el cambio de fase y que las PKA se activan como respuesta a la recepción de serotonina, Ott y colaboradores realizaron un experimento que demuestra el papel de las PKA en el cambio de fase: bloqueando estos receptores con compuestos antagonistas de la PKA se consiguió inhibir el cambio de fase durante la gregarización de individuos solitarios, lo cual también se consiguió mediante ARN interferente (RNAi) de la PKA [Ott et al., 2012]. Sin embargo, al realizar el mismo experimento con el gen foraging (una proteina kinasa dependiente de GMPc implicada en comportamiento alimenticio social en *Drosophila melanogaster* y en *Apis melifera* [Pereira and Sokolowski, 1993, Ben-Shahar et al., 2003] no se obtuvo ningún resultado tras gregarizar individuos solitarios [Ott et al., 2012]. Sobre el gen foraging, se sabe que en la mosca del vinagre presenta un alelo que aumenta su movilidad mientras se alimenta [Pereira and Sokolowski, 1993], y que en la abeja de la miel está implicado en cambio de comportamiento dependiente de la edad en las obreras [Ben-Shahar et al., 2003]. De hecho, en *S. gregaria*, se ha visto que este gen se sobre-expresa en cerebros de ninfas gregarias comparado con ninfas solitarias, y que además presenta actividad diferencial siguiendo el mismo patrón [Lucas et al., 2010], pero el resultado del estudio de [Ott et al., 2012] pone de manifiesto que este gen no parece implicado en el cambio de fase después de todo.

También se han detectado marcadores moleculares de fase en distintos trabajos. En [Rahman et al., 2002] se menciona la detección de un péptido sin función conocida en adultos criados en alta densidad de población, además de un inhibidor de serina proteasas (SGPI-2). A nivel de transcritos, se encontraron dos secuencias marcadoras, SSG (para individuos solitarios) y GSG (para individuos gregarios), de las cuales solamente la segunda se identificó como una secuencia similar a la metaloproteasa SPARC [Rahman et al., 2003]. Las neuroparsinas, péptidos antagonistas de la hormona juvenil y de la hormona diurética, también presentan un patrón de expresión relacionado con el cambio de fase en *S. gregaria*, donde dos de ellas (SGNPP3 y SGNPP4) presentaban picos de expresión diferentes entre adultos gregarios y solitarios, y curiosamente también entre machos y hembras [Claeys et al., 2006]. Aunque hay desarrollados microsatélites para esta especie, no se ha encontrado ninguno relacionado con el cambio de fase, resultado esperable teniendo en cuenta la naturaleza del polifenismo [Ibrahim et al., 2000, Kaatz et al., 2007]. Otras proteínas analizadas son las pacifastinas, inhibidores de serina proteasas relacionados con la respuesta inmune en insectos [Liang et al., 1997], que presentan diferencias de expresión tanto a nivel de sexo como de fase en *S. gregaria* [Simonet et al., 2005]. En [Loof et al., 2006] se presenta una buena síntesis de todos los potenciales marcadores de fase, donde se detallan no solo los descritos, sino que otros marcadores llamativos, como es la proteína amarilla responsable del color amarillo de los machos gregarios sexualmente maduros de *S. gregaria* [Wybrandt and Andersen, 2001].

Estudios "ómicos" en langostas

Pese a todos los estudios presentes, y aún habiéndose detectado algunas de las moléculas que desencadenan la respuesta fisiológica y metabólica del cambio de fase, aún no se puede afirmar con claridad que conozcamos toda la cadena de regulación que disparan el cambio. Para desentrañarlo se necesitaría hacer un análisis de los patrones de expresión de genes que estén expresándose durante el cambio de fase, lo que implicaría el seguimiento de miles de genes. Esta tarea sería una gran inversión en tiempo y costes si no fuera porque a lo largo de la década del 2000 aparecen técnicas de secuenciación masiva de ácidos nucleicos (Next Generation Sequencing, NGS), siendo posible obtener tanto secuencias genómicas casi completas como miles de secuencias de RNA mensajero (si se secuencia ADN complementario) sin el coste en dinero y tiempo que requeriría hacerlo con métodos anteriores (secuenciación tipo Sanger). Desde entonces, gran cantidad de estudios de genomas y transcriptomas han aflorado a lo largo de la bibliografía científica a un ritmo casi exponencial [Schuster, 2007, Metzker, 2010, Mardis, 2013]. Entramos pues en una etapa donde los datos "ómicos" (término referido a estudios con ingentes cantidades de datos, como genómica, transcriptómica y proteómica) se están convirtiendo en algo usual y que cada vez se explota más y mejor. Como en muchos otros casos, estas técnicas se han aplicado al estudio del cambio de fase de langostas, pudiéndose obtener una vista de pájaro de todo (o por lo menos casi todo) lo que está ocurriendo durante el cambio de fase a nivel de genoma, transcritos o metabolitos [Bakkali, 2013]. De esta forma, se pueden rastrear mejor las rutas genéticas y metabólicas implicadas en el cambio de fase en langostas, y así conseguir más evidencias para poder construir un modelo genético completo de este fenómeno.

Un estudio pionero en abordar el cambio de fase con secuenciación Sanger fue llevado a cabo por el laboratorio del profesor Le Kang en

Pekín, consistiendo en la confección de una biblioteca genética de ESTs (Expressed Sequence Tags) a partir de ARN mensajero de las fases gregaria y solitaria de la langosta migratoria *L. migratoria* [Kang et al., 2004]. Uno de los resultados que se destacan en este trabajo, que aborda secuencias de cuerpo, cabeza, intestino medio y fémur trasero, es que varias proteínas pertenecientes a la familia JHPH (Juvenile hormone binding protein - Hexamerin - Prophenoloxydase - Hemocyanin) aparecen sobre-expresadas con patrones característicos, presentando transcritos específicos de fase y tejido en muchos casos. Las proteínas de unión a la hormona juvenil (Juvenile hormone binding protein, JHBPs) sirven para regular la concentración de hormona juvenil en la hemolinfa promoviendo su degradación, aunque también es posible que algunas de estas proteínas sirvan como transportadoras de hormona juvenil o intervengan en procesos regulados por esta molécula [Trowell, 1992, De Kort and Granger, 1996]. La hormona juvenil ya fue nombrada en el apartado de la regulación hormonal del cambio de fase como factor de mantenimiento no sólo de estadíos entre mudas sino de fase solitaria. Las hexamerinas transportan hormonas y otras proteínas [Burmester, 1999] y de hecho presentan un dominio de unión a la hormona juvenil que hace pensar que no sólo intervengan en su transporte sino también en su regulación [Zhou et al., 2007, Martins et al., 2010], por lo que también podrían tener un papel en el cambio de fase de estar implicada la hormona juvenil. La profenoloxidasa está implicada tanto en la respuesta inmune como en la melanización de la cutícula [Leonard et al., 1985, Söderhäll and Cerenius, 1998]. Además, parece que su producción se asocia también a una respuesta dependiente de fase en otros insectos como en el escarabajo *Tenebrio molitor* [Barnes and Siva-Jothy, 2000] y en polillas de distintas especies plaga [Wilson et al., 2001], siendo parte de su regulación también regida la hormona juvenil [Curtis et al., 1984, Hiruma and Riddiford, 2009]. La hemocianina es la molécula encargada del transporte e intercambio gaseoso en la hemolinfa de los insectos [Sánchez et al., 1998]. Que una familia de proteínas con funciones tan diversas presente patrones específicos de unión a la hormona juvenil y que además presenten expresión diferencial regida por la fase desde luego la pone en el punto de mira en el estudio del cambio de fase. En un estudio posterior, se identificaron varias JHBPs en la langosta del desierto, detectándose variabilidad en la concentración en hemolinfa y algunas propiedades enzimáticas más (unión a JH III, actividad lipoforina), pero sin profundizar en su caracterización funcional [Tawfik et al., 2006]. Los ESTs obtenidos en el trabajo de Kang y colaboradores de 2004 se muestran en LocustDB [Ma et al., 2006], una base de datos que incluye los resultados de anotación correspondientes a esas secuencias, como su función, dominios de proteínas, rutas metabólicas y homología con otras especies.

Varios años después, el mismo grupo secuenció transcriptomas de varios estadíos ninfales de las dos fases de *L. migratoria*, esta vez con la tecnología de NGS Illumina [Chen et al., 2010]. Aquí, además de obtener un mayor número de secuencias con niveles de expresión más precisos, se ve que los estadíos con mayor número de transcritos diferencialmente expresados entre fases son el cuarto (penúltimo) y el quinto (último) estadío ninfal, proponiendo el cuarto estadío como crítico en cuanto a oportunidad de modulación de la fase en esta especie. Esta asignación está basada también en las diferencias globales de expresión entre fases en los distintos estadíos de desarrollo, siendo el cuarto estadío ninfal el que mayores diferencias en el patrón de expresión presentó entre fases. Ese trabajo también puso de manifiesto que una extensa lista de neurotransmisores y sus receptores presentan sobre-expresión en la fase gregaria, entre los que encontramos enzimas como la tirosina-hidroxalasa o la glutamato descarboxilasa, receptores para neurotransmisores como la 5-HT, la DOPA y la octopamina y transportadores relacionados a vesículas sinápticas.

También en *L. migratoria* se realizó un experimento de microarrays donde se cuantificó ADNc de varios individuos solitarios gregarizados y de individuos gregarios aislados en distintos puntos de una línea temporal, para poder comprobar si existen genes cuyo patrón de expresión a lo largo del tiempo sea gradual [Guo et al., 2011]. En este caso se ponen de manifiesto genes relacionados con la recepción, como las proteínas quimiosensoras (*Chemosensory Proteins*, CSPs), demostrando mediante silenciamiento de expresión vía RNAi que una de las CSPs es responsable de recibir estímulos de atracción a otros individuos, por lo que puede estar relacionada con la respuesta a corto plazo del cambio de fase. Otro gen llamado Takeout parece tener el efecto contrario: al ser silenciado en individuos solitarios aumentó su tendencia a agregarse con otros individuos, lo que lo convierte en un gen que puede prevenir el cambio de fase.

En otro experimento de microarrays distinto, los genes relacionados con la ruta de las catecolaminas, que incluyen a genes clásicos relacionados con el color (como *ebony*, *tan* o *pale*) como genes relacionados con el metabolismo de la tirosina y sus subproductos (*henna*, DOPA-decarboxilasa o *Ddc*, fenoloxidasa o *PO*) aparecieron como el proceso metabólico con más diferencias de expresión entre ninfas de cuarto estadío gregarias y solitarias [Wu et al., 2012], lo que confirma parte de las conclusiones del trabajo sobre el transcriptoma de *L. migratoria* de Chen et al. [2010]. Los productos más interesantes que origina esta ruta metabólica incluyen la dopamina y sus derivados, la melanina (implicada en respuesta inmune y color de la cutícula) y la proteína *yellow* (implicada principalmente en el color de la cutícula). Lo más impactante en este estudio fue que los niveles de expresión de ebony y henna aumentaban tras ocho horas de gregarización de solitarios, y tras silenciar mediante RNAi estos genes en ninfas gregarias se produjo una reversión del comportamiento gregario. A nivel metabolómico también se aprecian diferencias entre las dos fases, sobre todo los relacionados con el metabolismo lipídico [Ma et al., 2011]. Moléculas como el diacilglicerol o las fosfatidiletanolaminas aparecen sobrerrepresentadas en la hemolinfa de individuos gregarios, así como la carnitina y sus derivados, relacionados con la movilidad y degradación de lípidos para obtener energía.

Uno de los hitos que abrirán muchas puertas a la hora de describir la base molecular del cambio de fase ha sido sin duda la secuenciación del genoma de *L. migratoria* [Wang et al., 2014b]. En principio, no es tarea fácil secuenciar genomas tan grandes con la tecnología actual: aumenta la proporción de errores de secuenciación, se necesitará un mayor número de secuenciaciones para obtener una cobertura aceptable, y al aumentar la proporción de secuencias repetidas en genomas más grandes existirán más problemas de ensamblaje. Los saltamontes, donde se incluyen las langostas, se encuentran entre las especies que mayor tamaño genómico presentan, como es el caso de las langostas migratoria (6,3 Gb) [Wang et al., 2014b] y del desierto (8,6 Gb) [Camacho et al., 2015]. En el trabajo de Wang et al. [2014b] se consiguieron ensamblar 11 grupos de ligamiento, con un grupo extra donde se quedan los contigs sin ensamblar, lo que no representa la dotación cromosómica de *L. migratoria* (n = 11 + X) y que seguramente ha sido causa de la esperada proporción de ADN repetido y la dificultad añadida al ensamblaje de este tipo de secuencias. En efecto, de la partición estudiada cabe destacar que alrededor del 60 % del genoma secuenciado consiste efectivamente en secuencias repetidas, hecho congruente con la hipótesis de que a mayor tamaño genómico mayor proporción de elementos repetidos [Flavell et al., 1974, Gregory, 2005]. Los intrones parecen también inusualmente grandes en este genoma, unas diez veces de mayor longitud que en otros insectos, con un tamaño medio cercano a las 11.000 pares de bases [Wang et al., 2014b].

Gracias a la obtención de una secuencia genómica de referencia se pudieron estimar diferencias en los patrones de metilación en el cerebro de *L. migratoria* de fase gregaria y solitaria mediante una secuenciación por bisulfito de representación reducida [Wang et al., 2014b]. Los resultados de ese estudio indicaron que de las cerca de 9,5 millones de islas CpG encontradas, incluyendo las secuencias de cerca de 12.000 genes, la mayor proporción de posiciones metiladas se observó en intrones, seguida por secuencias codificantes enteras (incluyendo regiones no traducidas e intrones) y secuencias repetidas, siendo estas posiciones mucho menos representadas en secuencias intergénicas. En cuanto a genes diferencialmente metilados entre fases, se encontraron 90 candidatos, la mayoría de ellos estrechamente relacionados con la plasticidad sináptica. Tras estudiar también los patrones de corte y empalme alternativo (splicing) del ARN mensajero y comparar los resultados con los transcriptomas ya existentes [Wang et al., 2014b], hay un alto grado de congruencia entre estas tres particiones de datos, acotando incluso más los mecanismos moleculares responsables en última instancia del cambio de fase. Parece ser que los genes implicados en plasticidad sináptica, como genes del citoesqueleto, de transporte de vesículas y de transducción de señales, pueden jugar un papel muy importante en la modulación del cambio de fase en la langosta migratoria.

Durante el desarrollo de las "ómicas", casi simultáneamente se han estado describiendo varios elementos de regulación genética conocidos como ARNs no codificante (ncRNAs). La estructura secundaria de estos ARNs interacciona con otras moléculas, y su potencial homología con otras secuencias de ARN les permite formar apareamientos tanto con ADN como con ARN [Mattick and Makunin, 2006]. Entre estas secuencias podemos encontrar micro-ARNs (miRNAs), ARNs de interferencia pequeños (siRNAs), ARNs que interaccionan con proteínas Piwi (piRNAs), ARNs largos no codificantes (lncRNAs) y ARNs circulares (circRNAs). Así pues, gracias a la progresiva caracterización de ncRNAs se abre un nuevo camino a la hora de estudiar el cambio de fase, ya que seguramente están implicados en la expresión diferencial mediante eliminación de mRNAs. Hay estudios que abarcan el estudio de miRNAs en *L. migratoria*, siendo el pionero el artículo de [Wei et al., 2009], donde se describe una lista de miRNAs y siRNAs, con perfiles de expresión diferencial entre fase gregaria y solitaria. Esta información ha sido actualizada con análisis más refinados que tienen en cuenta la información genómica disponible para *L. migratoria.* Como resultado, se ha logrado confeccionar una extensa base de datos de los miRNA expresados en este organismo [Wang et al., 2015], en la cual podemos ver que hay muchos de estos elementos específicos y otros conservados en otros órdenes de insectos. El estudio de patrones de expresión generales de miRNAs también sirvió para encontrar candidatos en cuanto a la modulación del cambio de fase. Un ejemplo interesante es el artículo de [Yang et al., 2014], donde se demuestra cómo miRNA-133 regula tanto la expresión de *henna* como de *pale*, dos genes implicados en la ruta de las catecolaminas. Este miRNA presenta homología con la región 3' no traducida de *henna* y con la secuencia codificante de *pale*, y cuando se inyectaron análogos de miRNA-133 a individuos gregarios de *L. migratoria* se logró revertirlos a la fase solitaria.

Paralelamente, en la langosta del desierto *S. gregaria* se han estado llevando a cabo estudios de expresión génica similares a los planteados con *L. migratoria.* El grupo de Jozef Vanden Broek confeccionó, mediante tecnología de secuenciación Sanger, una biblioteca de ESTs del sistema nervioso central de esta especie gracias a la cual se pudo hacer una lista extensiva de transcritos codificantes para multitud de receptores sinápticos y enzimas sintetizadoras de neuropéptidos, neurotransmisores y hormonas [Badisco et al., 2011a]. Entre los neuropéptidos mencionados en ese trabajo se encuentran tanto la *Ddc* como la triptófano-hidroxilasa (TH), enzimas que sintetizan la dopamina y la serotonina respectivamente. En cuanto a receptores cabe destacar la identificación de receptores de membrana acoplados a proteínas G (GPCRs) como Methuselah 2 (*mth2*, implicado en modulación de longevidad en *Drosophila* [Cvejic et al., 2004]), el receptor para el neuropéptido corto F (*short NeuroPeptide F Receptor*, sNPFR, interacciona con el neuropéptido corto F, implicado en la inhibición del apetito en *S. gregaria* [Dillen et al., 2014] a pesar de tener el efecto contrario en otras especies [Lee et al., 2004, Ament et al., 2011]), receptores de serotonina y receptores de opsina. Por último, presentan una lista de neuropéptidos putativos, que pueden abrir más puertas en el estudio de la fisiología de esta especie. El mismo grupo de investigación también realizó un estudio de microarray comparando adultos gregarios y solitarios [Badisco et al., 2011b]. En ese trabajo se encontraron 214 genes diferencialmente expresados entre las dos fases. Con respecto a la función de recepción, la opsina y varias CSPs aparecen sobre-expresadas en individuos gregarios, al igual que los genes relacionados con la respuesta al estrés, como la taumatina o proteínas de choque térmico (*Heat Shock Protein*, HSP). Aparecen también sobre-expresados en fase gregaria citocromos P450, relacionados con detoxificación. También se detectaron niveles más altos de ADN metil transferasa 2 (*Dnmt2*) en la fase solitaria, siendo otra evidencia más de cómo la epigenética está implicada en el cambio de fase. Toda la información referente a ESTs producida para la langosta del desierto está siendo recopilada en una página de la universidad de Illinois [Broeck et al., 2005], que pretende ser un referente como base de datos de secuencias de varias especies de langosta.

A nivel de epigenética, lo cierto es que en *S. gregaria* hay varios trabajos que tratan de cubrir los patrones de metilación de ADN, seguramente promovidos por el resultado obtenido en Badisco et al. [2011b] sobre la expresión diferencial entre fases del gen *Dnmt2*. En Boerjan et al. [2011] se realiza un estudio basándose en estos valores de expresión de *Dnmt2*, encontrando que hay expresión diferencial de este gen en diferentes tejidos, y también demostrando cómo la fase en la que los progenitores se encontraban influye en la longevidad de su progenie mediante impronta genética. En Falckenhayn et al. [2013] se demuestra que la proporción de metilación del genoma de *S. gregaria* es mayor en las regiones exónicas al igual que ocurre con otros invertebrados. No obstante, el grado de metilación sí resultó ser más acentuado que en otros genomas. Relacionado con la epigenética, mediante un análisis de polimorfismos de longitud de fragmentos amplificados sensibles a la metilación (MS-AFLP) en *S. gregaria*, Amarasinghe et al. [2015] encontraron que la fase gregaria presentaba mayor proporción de loci no metilados con respecto a la fase solitaria. Ese resultado concuerda con los resultados obtenidos para *L. migratoria* en Wang et al. [2014b]. No obstante, para otros aspectos de la regulación de la expresión génica, parece que apenas hay avances en lo que respecta a los mencionados ncRNAs, ni a nivel de identificación ni a nivel de caracterización, lo que deja un importante campo de trabajo aún por explotar.

¿Por qué realizar más trabajos en langostas?

A pesar de que en la langosta del desierto y la langosta migratoria presentan genes con patrones de expresión compartidos, también encontramos muchos genes que a pesar de ser descritos como posiblemente implicados en el cambio de fase presentan patrones de expresión distintos. El estudio conjunto de los patrones de expresión en estas dos especies podría acotar incluso más la base genética del cambio de fase si realmente éste estuviera evolutivamente conservado en las dos especies. No obstante, si comparamos la información basada en estudios de secuenciación masiva disponible para la langosta del desierto con la que disponemos para la langosta migratoria, da la impresión de que el estudio de esta última va un paso más adelante (cuadro 1). El genoma de *S. gregaria* contiene 8,6 Gb, lo que lo convierte en uno de los genomas animales más grandes conocidos. Normalmente secuenciar un genoma de este tamaño conlleva dificultades tanto económicas como técnicas. Las dificultades económicas pueden disiparse a lo largo del tiempo, cuando los precios de técnicas NGS bajen. En cuanto a las dificultades técnicas se incluye la dificultad de ensamblar genomas ricos en secuencias repetidas, como ocurrió con el genoma de *L. migratoria* (con un 60 % de secuencias repetidas) y puesto que a mayor tamaño de genoma mayor proporción de secuencias repetidas se espera, ensamblar el genoma de esta especie no será una tarea fácil. Como paso inicial, se han obtenido varias secuencias de ADN satélite de *S. gregaria* mediante NGS, además de obtener valores de expresión y localización cromosómica, datos que pueden contribuir en un futuro a mejorar la calidad de ensamblaje del genoma de esta especie [Camacho et al., 2015]. Una vez sorteado el problema del ensamblaje del genoma, el siguiente paso de estudio será indudablemente el estudio epigenético del cambio de fase, acompañado de la determinación de los ncRNAs. Entre todas las evidencias obtenidas tanto para esta especie y con toda la información disponible de *L. migratoria* se podrá entonces llegar a un modelo de regulación genética y epigenética mucho más detallado que el que tenemos presente actualmente [Bakkali, 2013].

Además de los estudios sobre la caracterización in vivo del grado de gregariedad, el establecimiento y comparaciones cuantitativas y cualitativas in silico de transcriptomas, y el análisis in vitro de la expresión génica, esta memoria de tesis también compila los datos obtenidos de varios trabajos que comprenden el estudio de genes y de la base genética del cambio de fase en la langosta del desierto *S. gregaria.* Como se ha mencionado antes, a nivel de secuencias genéticas, *L. migratoria* está mejor cubierta y *S. gregaria* todavía necesita ponerse al mismo nivel de información disponible. Es más, la posible distancia evolutiva que hay entre las dos especies y la posibilidad de una evolución independiente de cambios de fase entre linajes de acrídidos con polifenismo de fase puede significar que lo que funcione en una especie no funcione en la otra (o al menos no al 100 %). Además, esta tesis forma parte de un proyecto cuyo objetivo es asentar un nuevo grupo de investigación con la intención de establecer el estudio de la genética y base molecular de las adaptaciones de especies plaga como nueva línea de investigación en España. De hecho, entre los datos que hemos generado, revelamos hechos tanto esperables como sorprendentes o interesantes, y ofrecemos herramientas de trabajo que hagan que los datos predictivos sean comparables entre trabajos y entre grupos de investigación. Nuestros resultados permiten comprender mejor los cambios de expresión genética entre fases. De hecho, somos pioneros en ofrecer una síntesis de la dinámica global que acompaña al cambio de fase basándonos en una interpretación experta de los datos sobre los cambios en la expresión de los genes que secuenciamos.

	Locusta migratoria	*Schistocerca gregaria*
Hormona juvenil	Más expresada en solitarios [Roussel, 1993, Wiesel et al., 1996]	Más expresada en solitarios [Roussel, 1993, Wiesel et al., 1996]
Corazonina	Induce cambios de color y forma [Tawfik et al., 1999, Maeno et al., 2004]	Induce cambios de color y forma [Tawfik et al., 1999, Maeno et al., 2004]
Serotonina	Solitarización [Guo et al., 2013]	Gregarización [Anstey et al., 2009, Rogers et al., 2014] Sin cambios en comportamiento [Tanaka and Nishide, 2013]
Transcriptoma	Varios estadios ninfales y adultos [Kang et al., 2004, Chen et al., 2010] Sistema nervioso [Zhang et al., 2012, Wang et al., 2014b]	Sistema nervioso de ninfas y adultos [Badisco et al., 2011a]
Microarray	Cuerpo entero [Guo et al., 2011, Wu et al., 2012] Sistema digestivo [Spit et al., 2016]	Adultos [Badisco et al., 2011b]
Genoma	*Draft* [Wang et al., 2014b]	Secuenciación parcial [Camacho et al., 2015]
Epigenética	Menor metilación en gregarios [Wang et al., 2014b]	*Dnmt2* sobre-expresada en solitarios [Boerjan et al., 2011] Metilación diferencial en solitarios [Amarasinghe et al., 2015]
miRNAs	miRNA 133 regula catecolaminas [Yang et al., 2014] Base de datos [Wang et al., 2015]	

Cuadro 1: Resultados más destacables sobre la base fisiológica y molecular del cambio de fase en *S. gregaria* y en *L. migratoria*.

Objetivos

Estando encuadrada en el marco de una beca predoctoral para la Formación de Personal Investigador (referencia BES-2011-043627) asociada al proyecto "*The genetic basis of the gregarious phase associated with outbreaks of the swarming locust* Schistocerca gregaria" financiado por el Ministerio de Ciencia e Innovación español (código BFU2010-16438), los objetivos a cumplir en esta tesis son comunes a los de dicho proyecto. El objetivo general a resolver consiste en detectar cambios de expresión genética a gran y pequeña escala asociados al cambio de fase en esta especie de langosta. Para ello nos valdremos de técnicas moleculares, bioinformáticas y basadas en el estudio del comportamiento, tratando de resolver los siguientes objetivos específicos:

1. Construir un modelo basado en regresión múltiple logística, a partir de datos de comportamiento y morfológicos de individuos gregarious y solitarios de *S. gregaria*, que cuantifique el estado de agregación de las langostas estudiadas.

2. Comprobar la potencial asociación entre posibles diferencias en el cambio de fase y factores aparentemente intrínsecos al cambio de fase, como las intensidades colorimétricas del individuo o su tamaño, y extrínsecos, como el sexo o la especie.

3. Obtener transcriptomas de referencia para el sistema nervioso central y del tubo digestivo de adultos gregarios y solitarios de *S. gregaria.*

4. Detectar las diferencias de expresión entre las fases gregaria y solitaria en el sistema nervioso central y en el tubo digestivo de *S. gregaria* para cada transcrito ensamblado.

5. Validar los datos transcriptómicos mediante qPCR.

6. Comparar nuestros resultados con los presentados en los trabajos de otros grupos de investigación mediante comparación bibliográfica.

7. Comprobar si hay diferencias de diversidad y de expresión genética entre la flora intestinal de individuos solitarios y gregarios.

8. Comparar la información obtenida de los transcriptomas del sistema nervioso central y del tubo digestivo.

9. Dar el primer paso hacia el dibujo de la imagen global, a completar por estudios posteriores, de la asociación entre eventos moleculares y aspectos no moleculares que acompañan al cambio de fase.

10. Como paso previo al estudio comparativo de su expresión como caso particular, determinar el número de copias que la familia de proteínas quimiosensoras (CSP) presentan en *L. migratoria* y *S. gregaria* y establecer sus relaciones evolutivas mediante análisis de secuencias y filogenias.

11. Analizar los patrones de expresión de las CSPs para determinar su relación con el cambio de fase, tanto al nivel específico como general para langostas.

Metodología general

Cría de langostas

Tanto las ninfas como adultos de la langosta del desierto *S. gregaria* utilizados para los experimentos que forman parte de esta tesis doctoral se criaron en la Facultad de Ciencias de Granada. El cultivo fue establecido por el Profesor Mohammed Bakkali a partir de cuatro puestas gregarias que le fueron cedidas en junio 2009 por el Profesor Jozef Vanden Broek (Universidad de Leuven, Bélgica). La alta consanguineidad del cultivo, incrementada por las varias generaciones endogámicas, proporcionó langostas solitarias y gregarias genéticamente homogéneas. Tanto langostas solitarias como gregarias se mantuvieron en una habitación con la temperatura controlada para que estuviera a 30 °C, con ventilación, y con un ciclo de luz y oscuridad de 14 h : 10 h. La dieta era basada en col ecológica bien lavada y pienso de cereales.

Los individuos gregarios se criaron en cajas de madera de 60 cm x 60 cm x 60 cm con una ventana de cristal y una puerta de madera de 10 cm x 10 cm en uno de los costados para facilitar el acceso al interior de la caja cuando fuera necesario (figura 5). Estas cajas contuvieron alrededor de 300 individuos para mantener las condiciones de alta densidad poblacional. Cuando fue necesario realizar un experimento de gradiente de densidad de población, se mantuvieron cajas con 10, 20, 80, 150 y 300 individuos. Se utilizaron bombillas de 60 W para iluminar el interior y, al mismo tiempo, generar un gradiente de temperatura dentro de la caja. De esta manera las langostas se iban situando dentro de la jaula en función de sus necesidades de temperatura corporal. La parte frontal inferior de cada caja albergaba cuatro botes de plástico incrustados en la madera mediante un cierre de rosca, conectando así el suelo de la jaula con el interior de cada bote. Los botes estaban normalmente llenos de agua y tapados con una esponja para garantizar cierto grado de humedad dentro de la caja, pero cuando se presenciaban cópulas se reemplazaba del agua y la esponja por vermiculita húmeda para que las hembras grávidas pudieran poner sus huevos. Las aperturas para botes también servían para recogida de los deshechos generadas por las langostas, que se barrían dentro de la caja con una brocha y se dirigían hacia una de las aperturas con bote para recogida de deshechos. Además, las cajas contenían una o dos ramas secas de unos 70 cm de longitud y cartones de huevo para ofrecer puntos de sujeción durante la muda y refugio. Tanto la col cortada como el pienso se situaron en varias placas de Petri abiertas para evitar que se generase moho en el conglomerado a causa de la humedad de la col lavada así como para facilitar su retirada de la caja.

A diferencia de los individuos gregarios, los individuos solitarios fueron criados individualmente en cajas de madera de 30 cm x 15 cm x 15 cm. Estas cajas contaban con la presencia de sólo dos botes en su parte frontal inferior, pero se destinaban también para humidificar, proporcionar vermiculita para puestas y recogida de deshechos. Para su iluminación interna se usaron bombillas de bajo consumo. También se añadió un sistema de renovación de aire en cada caja para evitar que las feromonas desprendidas por la langosta o sus heces tuvieran un efecto gregarizante sobre los individuos. Cuando los individuos solitarios maduraban sexualmente, se colocaban por parejas en estas mismas cajas hasta que la cópula tenía lugar, para así perpetuar la línea solitaria (figura 5).

Las puestas de huevos se recogieron en los botes con vermiculita humedecida antes mencionados llenos de vermiculita humedecida. Una vez sustraídas, se llevaron a una placa de Petri limpia, donde se cubrían con vermiculita y se humedecían con agua, retirando el exceso para evitar infecciones producidas por hongos. Tras esto, las placas con puestas se colocaron en una incubadora a 28 °C, donde tras un período de 10-12 días los huevos eclosionaban. Los saltones recién nacidos se colocaron en cajas grandes con saltones de su misma edad para producir animales gregarios o bien se aislaron en cajas pequeñas para producir animales solitarios.

Cuando fue necesario también se criaron ninfas y adultos de la langosta migratoria *L. migratoria* (figura 6). Se mantuvo un cultivo inicial de 50 individuos de *L. migratoria* en ausencia de *S. gregaria*, pero tras criarlas durante dos generaciones en nuestro insectario su número se redujo a 14 especímenes a los que consideramos como criados a baja densidad de población. Cada vez que se fuera a realizar un experimento que necesitara de individuos gregarios de esta especie, se compraron tandas de 100 ninfas gregarias y se manipularon después de una semana de aclimatación a sus nuevas condiciones de vida. Las condiciones de cría fueron las mismas para ambas especies y tanto el diseño, construcción y montaje de cajas para cría,

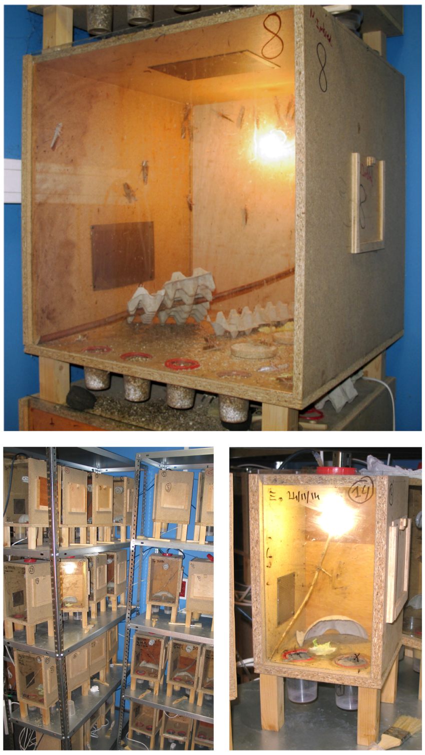

Figura 5: Terrarios de cría. Arriba se muestra una jaula para individuos gregarios. Abajo a la izquierda se muestra la estantería con las jaulas pequeñas conectadas a un sistema de ventilación. Abajo a la derecha se muestra en detalle una jaula pequeña para aislar langostas para producir individuos solitarios. El diseño, construcción y montaje de cajas para cría, estanterías y del sistema de ventilación, así como el establecimiento de protocolos para cría y mantenimiento de langostas, fueron obra del Profesor Mohammed Bakkali.

estanterías, sistema de ventilación y condiciones de la habitación de cría, así como el establecimiento de protocolos para cría y mantenimiento de langostas, fueron obra del Profesor Mohammed Bakkali. En los experimentos de qPCR de los capítulos 4 y 5 de esta tesis, los individuos de *L. migratoria* utilizados pertenecieron al cultivo del laboratorio del profesor Le Kang (Instituto de Zoología, Pekín, China), por lo que su cría se llevó a cabo siguiendo el procedimiento descrito en Kang et al. [2004].

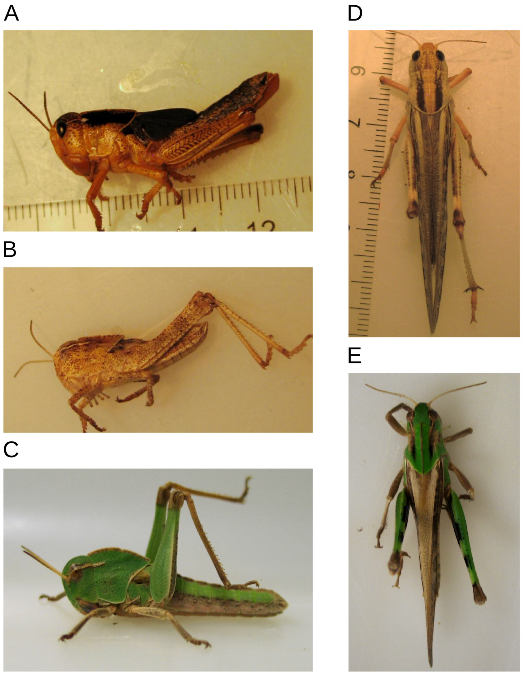

Figura 6: Muestra de individuos de *L. migratoria* utilizados en los experimentos. A: ninfa criada en alta densidad de población. B: ninfa criada en baja densidad de población. C: ninfa solitaria, capturada del campo. D: adulto gregario. E: adulto solitario capturado del campo.

Estudio de comportamiento y morfometría

Se realizaron grabaciones de 189 y 103 individuos de *S. gregaria* y *L. migratoria* criados a distintas densidades de población. Para ello utilizamos una arena de observación de comportamiento diseñada y construida por el Profesor Mohammed Bakkali. El grupo estímulo de langostas de la misma especie en fase gregaria situadas en uno de los lados de la arena de observación. Para cada langosta estudiada se extrajeron datos basados en el comportamiento para calcular 10 variables relacionadas con el cambio de fase. Además se tomaron medidas morfológicas y colorimétricas de toda langosta estudiada, calculándose tres índices morfométricos y otros tres colorimétricos asociados con el cambio de fase. Con los datos obtenidos a partir de las langostas criadas a densidades de población 1 y 150 (para *S. gregaria*) y de baja y alta densidad (para *L. migratoria*) se confeccionaron modelos con y sin datos morfológicos basados en regresión múltiple logística (Multiple Logistic Regression, MLR) para determinar un valor que cuantificase el estado de fase (nivel de gregariedad) de los individuos estudiados. Los detalles referentes a la confección, resultados y validación de dichos modelos de predicción de estado de fase están descritos de manera más detallada en el capítulo 1 de esta tesis.

Extracción de ARN

Para la extracción de ARN total se utilizó el reactivo RNAzol RT (Sigma-Aldrich) tomando las precauciones pertinentes para evitar la degradación del ARN (entre otras, mantener las muestras en hielo y trabajar con material libre de nucleasas). Primero se diseccionaron los tejidos pertinentes o se utilizaron individuos enteros, según requiriera el experimento, adormilando con frío a los animales. Por cada 100 mg de muestra se añadió 1 mL de RNAzol en un tubo de 50 mL o de 15 mL (según el volumen de RNAzol utilizado), e inmediatamente se procedió a la homogeneización mecánica de la muestra para asegurar que el ARN quedara en contacto con el RNAzol. Una vez que no quedaron restos visibles de tejido sin homogenizar, se trasvasó 1 mL de la mezcla a tubos Eppendorf de 1,5 mL de capacidad que contenían 0,4 mL de agua ultra-pura libre de nucleasas (se utilizaron tantos tubos como permitió el volumen de homogenizado), agitándola durante 15 segundos. Dada la capacidad de carga de tubos de la centrífuga utilizada, y dependiendo del volumen total del homogeneizado, se usaron hasta un máximo de 18 tubos por muestra de tejido y fase. Tras la agitación, se centrifugaron los tubos a 12.000 RCF durante 15 minutos a 4 °C. Una vez centrifugados, se trasvasó el 80 % del sobrenadante (donde la mayoría del ARN se encontraba) de cada tubo con la mezcla (cerca de 1 mL) a tubos Eppendorf nuevos, quedando en los tubos usados un precipitado de proteínas, carbohidratos, lípidos y parte de ADN genómico. A estos sobrenadantes se le añadieron 0,4 mL de etanol 75 % frío para iniciar la precipitación del ARN. Se dejaron 5 minutos a −20 °C y tras esto se centrifugaron a 12.000 RCF durante 10 minutos a 4 °C. Al acabar, se trasvasaron los sobrenadantes (con ARNs pequeños como miRNA y siRNA) en tubos nuevos y se almacenaron a −80 °C, mientras que a los precipitados (que contienen el ARN total) se les añadió 1 mL de etanol 70 % frío para limpiarlo de restos de reactivos. Se centrifugaron a 8.000 RCF durante 5 minutos, se descartó el sobrenadante, se volvió a añadir 1 mL de etanol 70 % frío a todos los tubos y de nuevo se centrifugaron a 8.000 RCF durante 5 minutos, descartando el sobrenadante de cada tubo otra vez y finalmente dejando secar durante unos minutos el precipitado para evitar que restos de alcohol puedan interferir en reacciones posteriores. Resuspendimos el ARN total en 200 µL de agua ultra-pura para su posterior manejo.

Para eliminar restos de ADN genómico que pudieran quedar en la muestra, se utilizó la nucleasa DNAsa I (Thermo Scientific). Se añadieron 25 µL de tampón de DNAsa I y 25 µL de DNAsa I a los 200 µL, se agitó suavemente la mezcla y se incubó a 37 °C durante 30 minutos. A continuación se añadieron otros 25 µL de EDTA del kit a la reacción, se agitó suavemente y se incubó a 65 °C durante 10 minutos para parar la reacción. Para limpiar restos de DNAsa I y sales, se añadieron 275 µL de fenol-cloroformo ácido (mezcla de fenol ácido, cloroformo e isoamil-alcohol en la proporción 25:24:1) a la reacción, se agitó suavemente hasta que la mezcla se volvió blanca y se centrifugó a 12.000 RCF durante 5 minutos. Se recuperó la fase acuosa superior con cuidado de que no se mezclara con la interfase y se trasvasó a un tubo Eppendorf limpio con dos volúmenes de cloroformo, tras lo cual se agitó y se centrifugó a 12.000 RCF durante 5 minutos. Se volvió a recuperar la fase acuosa superior y se trasvasó a un tubo Eppendorf limpio con 250 µL de isopropanol frío, se agitó y se enfrió la mezcla durante 5 minutos a −20 °C para seguidamente centrifugarla a 12.000 RCF durante 25 minutos a 4 °C. Se añadió 1 mL de etanol 70 % frío y se centrifugó a 8.000 RCF durante 5 minutos, se descartó el sobrenadante, se añadió otra vez 1 mL de etanol 70 % frío y de nuevo se centrifugó a 8.000 RCF durante 5 minutos, se descartó el sobrenadante otra vez y se dejó secar durante unos minutos el precipitado para que se evaporen los restos de alcohol. Se resuspendió el ARN total ya limpio en 250 µL de tampón citrato a pH 6,4. La calidad del ARN se valoró mediante electroforesis, con 2 µL de la muestra en un gel de agarosa al 2 % de peso por volumen, cerciorándonos de que la banda de ARN ribosómico se visualizaba y que el rastro de ARN fuese continuo a lo largo de todo el gel. La cantidad de ARN se midió con un nano-espectrofotómetro Quawell Q5000, comprobando que había cantidad suficiente para continuar con el siguiente paso (más de 500 ng/µL de ARN total por tubo), y que los valores de absorbancia fueran correctos (cocientes de absorbancia 260/280 y 260/230 próximos a 2).

Para obtener la mayor proporción de ARN mensajero posible en nuestras muestras, se utilizó el kit de *GenElute mRNA Purification Miniprep Kit* (Sigma-Aldrich). Tras asegurar que tanto la calidad como la cantidad de ARN total eran suficientes, se añadieron a los 250 µL de la extracción otros 250 µL del binding buffer solution del kit y se mezclaron. A continuación se añadió a la mezcla 30 µL de la solución de microesferas con oligómeros de desoxitimina, llamados oligo(dT), y se agitó muy suavemente hasta que toda la reacción quedase uniformemente de color blanco. Los oligo(dT) hibridan con las colas poliadeniladas de los ARN mensajeros y los anclan a las esferas, proceso que optimizamos en nuestro laboratorio mediante un primer paso de desnaturalización a 70 °C durante 3 minutos y luego un paso de incubación a 55 °C durante 20 minutos, donde ocurre la hibridación entre el ARN poliadenilado y los oligo(dT). Acto seguido se centrifugó la mezcla a 12.000 RCF durante 2 minutos y se descartó el sobrenadante. Se lavó el precipitado añadiendo 500 µL de washing solution a 55 °C para mantener la temperatura de hibridación y se trasvasó la mezcla a la columna del kit, colocada sobre un tubo de colección vacío. Se centrifugó la columna con la mezcla a 12.000

RCF durante 2 minutos, descartando la fase acuosa del tubo de colección. Se añadieron otros 500 µL de washing solution a 55 °C, se centrifugó a 12.000 RCF durante 2 minutos y se descartó la fase acuosa del tubo de colección otra vez. Se centrifugó la columna sobre el tubo de colección vacío a 12.000 RCF durante 1 minuto para eliminar los restos de alcohol de la columna. Tras esto, se trasladó la columna a un tubo Eppendorf limpio y se añadieron 50 µL de elution buffer solution a 70 °C, se incubó durante 5 minutos a 70 °C y se centrifugó a 12.000 RCF durante 1 minuto. Se añadieron otros 50 µL de elution buffer solution a 70 °C, se incubó 5 minutos a 70 °C y se centrifugó a 12.000 RCF durante 1 minuto otra vez. La cantidad de ARN se midió de nuevo en el espectrofotómetro Quawell Q5000. Para almacenar o enviar a secuenciar el ARN mensajero aislado, a los 100 µL resultantes de dilución se le añadieron 10 µL de acetato sódico 3 molar a pH 7 y 220 µL de etanol absoluto, almacenándolo a −80 °C.

Secuenciación y control de calidad

Para obtener las librerías con lecturas de cada tejido de las fases gregaria y solitaria se utilizó la tecnología de secuenciación masiva Illumina (Solexa). Esta tecnología, a pesar de presentar una menor longitud de las lecturas secuenciadas, ofrece la posibilidad de secuenciar el extremo 5' y el extremo 3' de un solo fragmento de ADN complementario originado del ARN mensajero (tecnología *pair-end*). La profundidad de secuenciación (número de veces estimado que se secuencia un mismo nucleótido) es del orden de uno o dos órdenes de magnitud mayor que la tecnología de Roche 454 (a pesar de la mayor longitud de lectura ofrecida, entre 400-500 nucleótidos en promedio), por lo que no solo ofrece un mejor respaldo del ensamblaje sino que además compensa la corta longitud de lectura y permite que las estimas de expresión diferencial de transcritos sea más precisa, pudiéndose ensamblar y detectar la expresión de secuencias con muy pocos transcritos presentes en la muestra. Así pues, el ARN mensajero obtenido de las muestras fue enviado a la sede coreana de Macrogen, donde tras superar un control de calidad tras ser comprobado mediante un bioanalizador, fue secuenciado con un secuenciador modelo 2000 Hiseq. En Metzker [2010] se pueden ver con detalle las principales diferencias entre ambas tecnologías de secuenciación masiva.

Cada muestra originó dos ficheros de secuencias con lecturas emparejadas (uno correspondiente a las del extremo 5' y otro al extremo 3' del fragmento de ADN complementario secuenciado a partir del ARN mensajero) gracias a la tecnología *pair-end*. Aunque en la mayoría de las veces se conocen de cada fragmento sólo las secuencias de sus extremos, gracias a la profundidad de secuenciación se puede reconstruir más fácilmente un transcrito con esta opción. Para asegurarnos de la calidad de la secuenciación, analizamos los ficheros de lecturas en formato ".fastq" con el software FastQC [Andrews,

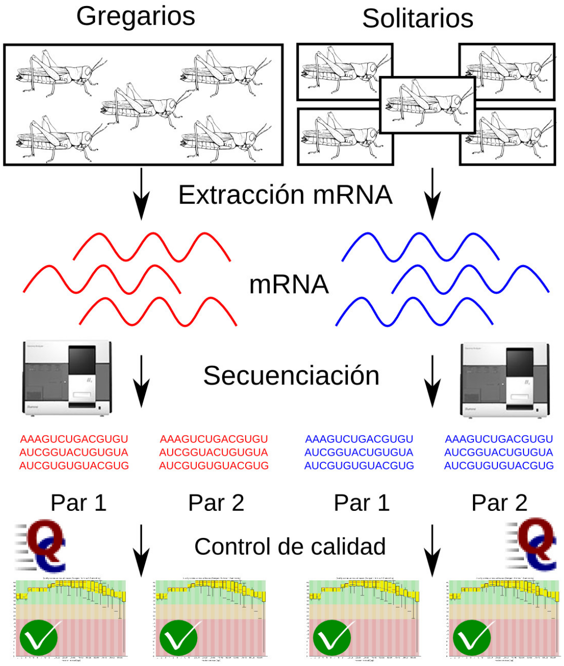

Figura 7: Esquema del diseño experimental de la secuenciación y del control de calidad de las librerías.

2010]. Este programa genera varios resultados, como los porcentajes de cada nucleótido presentes en la librería, número de lecturas con la misma secuencia repetida y probabilidad de que un nucleótido haya sido erróneamente secuenciado por posición dentro de la lectura. Utilizamos este último análisis para asesorar la calidad de la secuenciación, basándose en los valores medios de “Phred” por posición (logaritmo negativo en base 10 del porcentaje de error de secuenciación para un nucleótido) como estimadores de calidad de secuenciación. También se compararon los porcentajes de guanina y citosina (% GC) para confirmar la ausencia de diferencias significativas en la proporción de nucleótidos entre las librerías gregaria y solitaria de cada tejido. El diseño experimental de la secuenciación y el control de calidad están esquematizados en la figura 7.

Ensamblaje, anotación y expresión diferencial

Una vez comprobada la calidad de las lecturas, realizamos un ensamblaje *de novo* (sin referencia) de los transcriptomas secuenciados. En este caso nos valimos de dos algoritmos diferentes para realizar los ensamblajes de referencia: los grafos de De Bruijn y el consenso por patrón de solapamiento (*overlap layout consensus*). Los grafos de De Bruijn se construyen mediante la fragmentación *in silico* de la lectura en k-meros (subunidades de longitud k menor a la longitud total de la lectura) seguido de su solapamiento sistemático entre ellas para finalmente generar una secuencia consenso. El primer paso consiste en colocar los k-meros de forma consecutiva en función de los solapamientos de secuencia que presenten, formando nodos de longitud k − 1. Los nodos son relacionados mediante sus respectivos k-meros a modo de aristas del grafo. No obstante, puede haber varios caminos posibles para una misma secuencia, formando lo que se llaman burbujas en el grafo. El siguiente paso consiste en colapsar estas burbujas (es decir, se elige la posibilidad de ensamblaje más adecuada) para lograr una o varias opciones de secuencia contigua (contig), que se validarán gracias a la profundidad de secuenciación obtenida. Este algoritmo es particularmente útil para la secuenciación de ARN mensajero, ya que al ensamblar varias alternativas de un mismo transcrito se pueden estimar isoformas y además se puede utilizar la profundidad de lectura del grafo resultante como soporte, lo que le da ventaja con respecto al algoritmo de ensamblaje de consenso por patrón de solapamiento, como se ilustra en [Martin and Wang, 2011]. No obstante, este último algoritmo, basado en solapar sistemáticamente las lecturas mediante búsqueda de patrones de longitud k y ensamblarlas cuando superan un umbral de identidad, puede ser útil para rescatar fragmentos de transcritos que han sido parcialmente ensamblados por el otro algoritmo.

La estrategia de ensamblaje consistió en realizar un primer ensamblaje con el programa basado en grafos de De Bruijn ABySS v.3.3.1 [Birol et al., 2009, Simpson et al., 2009] con valores de parámetros que permitiesen el ensamblaje de la mayoría de secuencias contiguas posibles. Los parámetros utilizados en ABySS fueron: *abyss-pe OVERLAP_OPTIONS = 'no-scaffold' SIMPLEGRAPH_OPTIONS = 'no-scaffold' MERGEPATHS_OPTIONS = 'greedy' np = 6 k = X n = 1 c = 1 e = 0 v = -v in = INPUT_PATH -o OUTPUT_PATH name = OUTPUT* , donde “X” es el tamaño de k-mero utilizado, “INPUT_PATH” es la situación de los ficheros de las lecturas, “OUTPUT_PATH” la situación del fichero de salida y “OUTPUT” el nombre del fichero con los resultados del ensamblaje de ABySS. Este comando fue integrado en un pequeño programa para ser ejecutado recursivamente, tomando valores impares de longitud de k-mero (parámetro k) comprendidos entre 19 y 95. El resto de parámetros se ajustaron bajo las recomendaciones presentadas en el manual en línea de ABySS para ensamblar transcritos y no genomas. Tras un profundo estudio del modo de acción del algoritmo, modificamos las opciones del ensamblaje a n = 1 (sólo necesitamos 1 pareja de k-meros solapados para unirlos en una sola secuencia contigua o contig), c = 1 (eliminamos contigs con una covertura media menor a 1) y e = 0 (recortar nucleótidos del extremo de un contig cuando su cobertura es 0); de esta manera logramos aumentar la probabilidad de obtener todas las secuencias presentes en el cDNA. El siguiente paso consistió en refinar el ensamblaje con la extensión de ABySS especializada en ensamblar transcritos: Trans-ABySS v.3.3.1 [Robertson et al., 2010], que utilizará las llamadas burbujas de los grafos de De Bruijn para establecer isoformas. El comando de Trans-ABySS utilizado fue *trans-abyss.sh -i OUTPUT -0* , siendo "OUTPUT" la salida de ABySS y "-0" el modo de re-ensamblaje de transcritos.

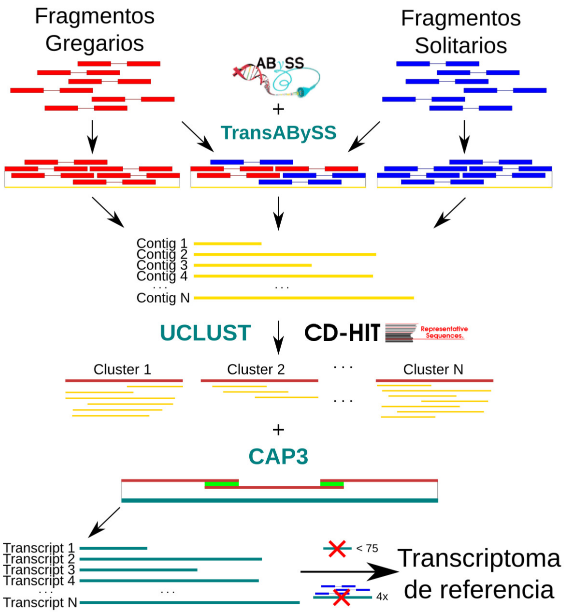

Figura 8: Esquema del proceso de ensamblaje del transcriptoma.

Para eliminar secuencias redundantes se utilizó primero UCLUST [Edgar, 2010], que agrupa las secuencias mediante un umbral de identidad considerando toda la secuencia, o CD-HIT [Li and Godzik, 2006], que agrupa los contigs mediante una ventana deslizante de un tamaño concreto en una agrupación de secuencias (cluster), usando como umbral de identidad el 95 % en ambos casos. El comando de UCLUST es el siguiente: *uclust –sort OUTPUT.fasta –output OUTPUT.sorted.fasta && uclust –input OUTPUT.sorted.fasta –uc OUTPUT.uc –id 0.95 –maxlen value –rev && uclust –uc2fasta OUTPUT.uc –input OUTPUT.fasta –output OUTPUT.uc.fasta –types S* , donde "value" debe representar la máxima longitud de secuencia. Para CD-HIT utilizamos el siguiente comando: *cd-hit-est -i OUTPUT.uc.fasta -c 0.95 -o OUTPUT.uc.c095.fasta* . Una vez obtenidos los clusters de contigs, se utilizó el programa de ensamblaje basado en consenso de solapamiento CAP3 [Huang and Madan, 1999] para rescatar contigs ensamblados parcialmente. El comando de CAP3 utilizado fue *cap3 -o 16 -p 95 -k 0* , usándose recursivamente hasta que no generase nuevos contigs. Como paso final, desechamos cualquier secuencia de longitud menor a 75 nucleótidos y alineamos las lecturas de ambas librerías (gregaria y solitaria) a los contigs mediante el protocolo detallado más adelante, desechando toda secuencia con menos de cuatro fragmentos alineados, considerando los contigs que han pasado el filtro como transcritos. Todo el proceso del ensamblaje *de novo* y el refinamiento de secuencias para obtener un transcriptoma de referencia está esquematizado en la figura 8. Para el transcriptoma del tubo digestivo (en el capítulo 3), optamos por eliminar secuencias menores de 100 nucleótidos.

Una vez obtenidas las secuencias de referencia, procedimos a la anotación de las mismas. Para ello se utilizó BLASTx [Altschul et al., 1990, 1997] de manera local para alinear los transcritos ensamblados contra varias bases de datos de secuencias de proteínas o nucleótidos de manera secuencial. Cada transcriptoma se lanzó contra una secuencia de bases de datos BLAST, seleccionándose siempre el mejor resultado BLAST, todos ellos con una resolución de valor E máximo de 10^{-6}, y lanzando el resto de secuencias sin resultado BLAST aceptable contra otra base de datos BLAST, repitiendo el procedimiento hasta agotar todas las posibles bases de datos. Obtenidos los resultados de anotación por BLASTx, se utilizó el programa BLAST2GO v.2.8 [Conesa et al., 2005, Conesa and Götz, 2008] para realizar una búsqueda de términos asociados a identificadores BLAST en las bases de datos de *Gene Ontology* (GO), de *Inter Pro Scan* (IPS) y de la *Kyoto Encyclopaedia of Genes and Genomes* (KEGG). Una vez etiquetadas las secuencias con estos términos se elaboraron gráficos de barras o sectores con la abundancia de términos GO para las tres categorías principales de GO: proceso biológico, función molecular y componente celular. También se utilizaron los resultados de BLASTx para categorizar qué especies presentaban homología con las secuencias obtenidas en el transcriptoma, así como para calcular la abundancia de transcritos pertenecientes a cada especie.

Completada la anotación de la referencia, utilizamos las lecturas de cada librería para ser "mapeadas" (alineadas en las secuencias contiguas) y así obtener una estima cuantitativa de los niveles de expresión genética en cada muestra. Para ello utilizamos el programa *Burrows-Wheeler Transform Aligner* (BWA) [Li and Durbin, 2009] para alinear cada conjunto de lecturas a sus secuencias de referencia correspondientes. La línea de comandos utilizada fue: *bwa index -a is REF.fa && bwa aln REF.fa reads1.fastq >lib1.sai && bwa aln REF.fa reads2.fastq >libB1.sai && bwa sampe -s REF.fa lib1.sai lib2.sai reads1.fastq reads2.fastq >lib.sam* . Una vez alineadas, utilizamos el programa “xa2multi.pl” (incluído como programa adicional en BWA) para desdoblar registros de lecturas alineadas en posiciones múltiples en registros independientes de lecturas alineadas en una sola posición y el porgrama “SAMtools” [Li and Durbin, 2009] para comprimir los ficheros de salida de BWA a formato “.bam”. La línea de comandos utilizada fue: *cat lib.sam |./xa2multi.pl - >lib.xa.sam && awk ’!/\t*\t0\t0\t\*\t*\t0\t0\t/’ <lib.xa.sam >lib.clean.sam && samtools view -bSh lib.clean.sam >lib.bam && samtools sort -n lib.bam lib.sorted* . Tras esto, extrajimos los conteos de lecturas alineadas a cada secuencia con el programa “HTSeq-count” [Anders, 2010, Anders et al., 2014], obteniéndose un fichero con los identificadores de los contigs y sus lecturas alineadas en cada muestra. La línea de comandos utilizada fue: *samtools view lib.sorted.bam |htseq-count -q -s no -m intersection*-nonempty - lib.gff >lib.htseq* . La ejecución secuencial de todos estos pasos ha sido automatizada en un programa escrito en *bash*. Una vez obtenidos los conteos, se realizó una prueba χ^2 con corrección de Yates [1934] seguida de una prueba de descubrimiento de falsos positivos (*False Discovery Rate*, FDR [Benjamini and Hochberg, 1995]) para detectar diferencias de expresión significativas entre muestras a la par que se corrige el valor P resultante para reducir la presencia de falsos positivos dado el gran número de comprobaciones a realizar. Al obtener la lista de transcritos diferencialmente expresados, extrajimos los términos GO, los términos KEGG y los de la especie con la que más homología presentaban estos transcritos, y mediante pruebas exactas de Fisher (*Fisher’s Exact Test*, FET [Fisher, 1935]) comprobamos si había enriquecimiento en la proporción de algunos términos GO, términos KEGG y de especies con homología entre los transcritos sobre-expresados en las fases gregaria y solitaria.

Para realizar las gráficas de expresión diferencial se calcularon cuatro variables derivadas de los datos obtenidos: los fragmentos alineados por kilobase del gen de referencia y por millón de fragmentos totales alineados a la referencia (*Fragments mapped Per Kilobase of referente sequence and Million of total mapped fragments*, FPKM), el logaritmo en base dos de la razón entre el valor FPKM solitario y el valor FPKM gregario (Log_2FC o M), el logaritmo en base dos de la media aritmética de los valores FPKM solitario y gregario (A) y el logaritmo en base 10 del valor P corregido por FDR ($\text{Log}_{10}\text{Pval}$). El FPKM es una medida extendida derivada del RPKM (calculado de igual forma pero con lecturas, *reads*, en vez de con fragmentos) para normalizar los conteos a la hora de representarlos o compararlos con el resto del transcriptoma [Mortazavi et al., 2008]. Para tener un valor práctico de patrón de expresión se utiliza el llamado fold change (FC), consistente en la división de FPKMs de dos tratamientos experimentales distintos para un mismo gen, y que según sea mayor o menor a 1, indicará si el gen está sobre-expresado más en la librería del numerador (si es mayor a 1) o en la librería del denominador (si es menor a 1). Al tomar logaritmos, esta razón puede ser mayor o menor a cero respectivamente para cada caso y se comporta de manera simétrica, pudiéndose comparar más fácilmente los valores tanto positivos como negativos para distintos genes, aunque es una medida meramente orientativa de la intensidad de sobre-expresión. Con la intensidad media por gen lo que hacemos es calcular la media de FPKM por gen en todos los tratamientos estudiados, lo que facilita la detección de posibles sesgos de cobertura entre tratamientos. Para poder comparar la significación con estas variables en gráficas, usamos el logaritmo del P-valor corregido por FDR. El proceso de anotación y análisis de expresión se muestra esquematizado en la figura 9.

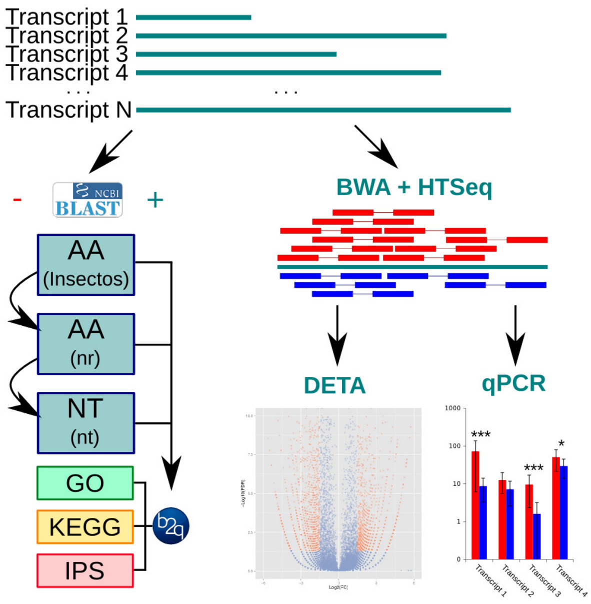

Figura 9: Esquema del proceso de anotación (izquierda) y del análisis de expresión-diferencial de transcritos (derecha) del transcriptoma.

Síntesis de ADN complementario

Para sintetizar ADN complementario se utilizamos el kit *First Strand cDNA Synthesis* (Thermo Scientific). Se utilizaron hasta 5 µg de ARN total por cada 20 µL de reacción. El protocolo de este kit consta de dos reacciones. En la primera reacción se hibrida el cebador oligo(dT) con el ARN, para lo cual se mezclaron 1 µL de oligo(dT), 1 µL de desoxirribonucleótidos (dNTPs) y ARN diluido en agua ultra-pura hasta los 15 µL; tras mezclar la reacción se incubó a 65 °C durante 5 minutos y luego se colocó en hielo durante 1 minuto. En la segunda reacción es donde ocurre la retrotranscripción del ARN sintetizándose ADN complementario, para lo cual se añadieron 4 µL de tampón para la retrotranscriptasa y 1 µL de retrotranscriptasa del kit (obteniéndose un volumen final de 20 µL); tras mezclar la reacción se incubó en cuatro ciclos de 48 °C durante 5 minutos y 60 °C durante 15 minutos (para alineamiento y extensión de la retrotranscriptasa), seguidos de una incubación final de 85 °C durante 5 minutos para parar la reacción. Comprobamos la integridad del ADN resultante mediante electroforesis en gel de agarosa.

Reacción en cadena de la polimerasa (Polymerase Chain Reaction, PCR)

Para realizar reacciones en cadena de la polimerasa (PCR) utilizamos el kit de PCR de Biotools. La reacción se preparó conforme al protocolo, normalmente en un volumen final de 50 µL, usándose 5 µL de tampón del kit, 5 µL de cebador forward y 5 µL de cebador reverse (concentración a 0,01 nmol μL^{-1}), 5 µL de dNTPs (concentrados a 0,002 nmol μL^{-1} cada dNTP por separado, 0,008 nmol μL^{-1} los cuatro en total), entre 1 y 5 µL de ADN diluido en agua (dependiendo de la concentración), 2:5 µL de la polimerasa termoestable de Thermus aquaticus (Taq-polimerasa) y enrasando con agua hasta los 50 µL. Una vez mezclada y distribuida en un tubo de 200 µL, la reacción se introdujo en un termociclador modelo Techne (Techgene) bajo un programa de ciclos de temperatura basado en lo siguiente: tras precalentar la placa superior a 104 °C, se procedió a iniciar el programa con 1 ciclo de 2 minutos a 94 °C, entre 35 y 45 ciclos de 30 segundos a 94 °C, 30 segundos a la temperatura óptima de los primers y 1 minuto a 72 °C, seguidos de un último ciclo de 72 °C durante 5 minutos, tras lo cual se dejó la reacción durante tiempo indefinido a 4 °C. Utilizamos electroforesis en gel de agarosa para comprobar la integridad del producto resultante.

Electroforesis mediante gel de agarosa

Preparamos los geles de agarosa en función del material a visualizar. Para eso usamos cantidades de agarosa comprendidas entre porcentajes en peso de 1,5 % (para ADN complementario total y productos de PCR mayores a 400 pares de bases) y 2 % (para ARN total y productos de PCR de menos de 400 pares de bases de longitud) en 50 o 100 mL de tampón Tris-borato-EDTA (TBE), dependiendo del tamaño deseado del gel. La agarosa y el TBE se mezclaron bien en un matraz de 330 mL y se calentaron durante 1 o 2 minutos en el microondas hasta que la agarosa se disolvió en el TBE, tras lo cual se dejó enfriar unos minutos antes de añadirle entre 1 y 2 µL de SYBR, se agitó y se vertió en un molde previamente sellado y con un peine colocado para dejar espacio a los pocillos de carga, dejándolo reposar unos minutos hasta la solidificación. Las muestras (normalmente 1 o 2 µL de ADN o ARN) se mezclaron con 2 µL de tampón de carga y agua hasta los 20 µL, y tras recoger cada muestra en la micropipeta se cargó en el pocillo del gel previamente colocado en la cubeta de electroforesis con tampón TBE. Incubamos las muestras de ARN total antes de ser cargadas en el gel durante 2 minutos a 65 °C para linearizar las bandas de ARN ribosómico, y tras cargarlas en el gel se sometió a un campo electroforético de 100 V durante 15 o 20 minutos para minimizar la degradación del ARN. Las muestras de ADN simplemente se cargaron en el gel y se sometieron a un campo electroforético de 100 V durante 30 o 45 minutos, dependiendo de la velocidad de migración de la banda o el rastro en el caso del ADN complementario total. En cada gel cargamos también 10 µL de marcador molecular (ladder) con distintas longitudes y cantidades de bandas de ADN para tener una estima de la cantidad y longitud de las bandas cargadas en ese gel.

PCR cuantitativa (quantitative PCR, qPCR)

Para cuantificar los niveles de expresión de algunos genes in vitro se utilizó la técnica de la PCR cuantitativa (qPCR). Los ADNs complementarios se diluyeron 1:50 en agua destilada libre de ARNasas y a continuación se utilizaron para las reacciones mezclando el kit *SensiMix SYBR* (Bioline), realizándose tres réplicas técnicas por gen y por muestra de cDNA comprobada. En cada reacción se mezclaron 5 µL del kit de *SensiMix SYBR*, 5 µL de ADN complementario diluido (correspondiendo a 5 ng de ARN usado como molde de la síntesis de este ADN), 1 µL de cada cebador (10 pmol) y 3 µL de agua destilada libre de ARNasas. Los ciclos de la qPCR consistieron en un primer paso de 95 °C durante 10 minutos, seguido de 40 ciclos con el programa 94 °C durante 15 segundos, 60 °C durante 15 segundos y 72 °C durante 15 segundos cada uno, con una lectura de fluorescencia después de cada fase de extensión (en este caso el paso de 72 °C). Para comprobar que solo un producto de PCR se amplificó en cada reacción sometimos a las reacciones a una rampa de temperaturas desde 72 hasta 95 °C y realizamos una lectura de fluorescencia por cada incremento de 1 °C para obtener las curvas de fusión. Las reacciones tuvieron lugar en un termociclador modelo DNA Engine Peltier Thermal Cycler acoplado a un detector continuo de fluorescencia modelo Chromo 4 System CFB-3240 (BIO-RAD).

Antes de realizar los experimentos con genes de interés necesitamos comprobar genes de referencia que se expresen de manera estable entre los ADNs complementarios que vayamos a estudiar. Hicimos diluciones seriadas para los ADNs complementarios con concentraciones de 25, 5, 1, 0,2 y 0,04 ng/µL. A continuación realizamos varias qPCRs para cada cDNA y para cada

Gen (orientación)	Secuencia	Longitud	Tm	GC %
CSP12 (+)	CGCGCTACGACAACATCAAC	20	60,5	55
CSP12 (-)	TCCGCTTGTAGATGCCAGTG	20	60,11	55
Mth2 (+)	GGCCGAAGAAGTAGGACCAG	20	59,82	60
Mth2 (-)	AGCTCACCAGCCACGTTATC	20	60,11	55
TAT (+)	CCCTACGACATTCGGCAACC	20	62,5	60
TAT (-)	TCGACGCCAATAGACTCAGC	20	59,9	55
PEPCK (+)	ACACAAGAGGGACACCCTGAG	21	61,38	57,14
PEPCK (-)	GTTGCTGCCACTAATGATGGAG	22	62,1	50
Yellow H (+)	ACCGGAAGTTGAACGAGTCC	20	59,97	55
Yellow H (-)	CACCGCACTACGGAGGATAC	20	59,97	60
NA202 (+)	TTGAAGACTACGACGCCGAC	20	59,42	55
NA202 (-)	CAGCAGTCACTCGTTGTGTG	20	60,11	55
PPI (+)	ATGTTTCTTCTGCAACTAGTGTTTC	25	60,9	36
PPI (-)	TTTGTCATGGTCCTCTAGCATG	22	60,1	45
Peroxiredoxin (+)	CTTGACTCCTCGCTTCTGG	19	59,5	58
Peroxiredoxin (-)	ATCGTACCAAACTTCACGGC	20	58,4	50
ASNS (+)	TCACTGCTTTGCTTGCAACTTTG	23	60,9	43
ASNS (-)	ACATAGTCTGCCACACATTTGC	22	60,1	45
HSP20 (+)	ACGTAGTCGACTGAGAACACC	21	61,2	52
HSP20 (-)	GTGTTGCACGAATAGCGCAG	20	60,5	55
GMF (+)	ACAAACCTTGGCTGATGGCT	20	60,2	50
GMF (-)	GGCGCTGAAAGCATTTCGTT	20	60,3	50
LCP9 (+)	ACTGGCGTAGGATTGTTCCG	20	59,8	55
LCP9 (-)	ACCATCAGCGTCAGGTACAC	20	60,3	55
RNA helicase (+)	TCATCGTCTCTTCCTGGTGC	20	59,5	55
RNA helicase (-)	GGAAAGACATATTCGCGGCA	20	58,7	50
Troponin-C (+)	CCTGTTTTCAAGGCAGTTGGG	21	59,9	52,4
Troponin-C (-)	ACCACAGTCCTCCCAAGAGA	20	59,8	55
Tubulin A1 (+)	TGACAATGAGGCCATCTATG	20	56,4	45
Tubulin A1 (-)	CGCAAAGATGCTGTGATTGA	20	56,4	45

Cuadro 2: Cebadores utilizados para validar mediante qPCR los resultados del transcriptoma del sistema nervioso central de *S. gregaria*

dilución paralos seis genes de referencia utilizados por Van Hiel et al. [2009], siendo estos genes la actina, *Armadillo*, la GADPH, el factor de elongación 1, la proteína ribosomal 49 y la tubulina 1. Mediante el programa geNorm [Vandesompele et al., 2002] se seleccionaron los genes tubulina y actina como genes con niveles de expresión más estables entre los cDNAs comprobados, usándolos como genes de referencia en las posteriores qPCRs. Para estimar los niveles de expresión relativa de transcritos utilizamos la fórmula del ΔC_t descrita en Livak and Schmittgen [2001] y Pfaffl [2001]. Los cebadores utilizados para las reacciones y sus características están detallados en las tablas 2 y 3.

Gen (orientación)	Secuencia	Longitud	Tm	GC %
LmigCSPI-1 (+)	CCTGCCTACTCTCATCTCTG	20	60,5	55
LmigCSPI-1 (-)	GACAATACAGGTAGTTACAGGAG	23	60,9	43
LmigCSPI-2 (+)	CTGCAGGCTCCGCTCTAC	18	60,8	67
LmigCSPI-2 (-)	CACCAGAGTGGCCTTCATGT	20	60,5	55
LmigCSP3 (+)	GAGGTCCACTGTGATAGCTC	20	60,5	55
LmigCSP3 (-)	GCTTGCTCGCTTAACAATAATTTAC	25	60,9	36
LmigCSP4 (+)	ACTCCTGACGTCTACCAGAC	20	60,5	55
LmigCSP4 (-)	GATATCCCGTGCAAACAACGG	21	61,2	52
LmigCSP24 (+)	CGTCTGGGCCAGTCGCAG	18	62,9	72
LmigCSP24 (-)	AAGGTGCTTATCACAGTCGGGT	22	62,1	50
LmigCSP33 (+)	AGGCGGGCGCTGACAGCT	18	62,9	72
LmigCSP33 (-)	TGGCAACTCAACAATGGTCACTG	23	62,9	48
RP49 (+)	CGCTACAAGAAGCTTAAGAGGTCAT	25	64,1	44
RP49 (-)	CCTACGGCGCACTCTGTTG	19	61,6	63

Cuadro 3: Cebadores utilizados para confirmar mediante qPCR el patrón de expresión de seis CSPs de *L. migratoria*

Chapter 1: Standardized multivariate regression models for estimation of gregariousness in the main pest locusts

1.1 Introduction

With the availability of transcriptomics [Badisco et al., 2011a,b, Chen et al., 2010, Guo et al., 2011, Zhang et al., 2012] and even genomics [Wang et al., 2014b] data, research on the molecular basis of locust gregariousness is entering a function-testing phase. In addition, RNAi has been succesfully applied on different Orthoptera species (e.g., Boerjan et al. [2011], Cabrero et al. [2013], Ruiz-Estévez et al. [2014], Dong and Friedrich [2005], He et al. [2006]), which is resulting a key technique for checking how genes detected by transcriptomic analysis affects phase change. Such works need tools for quantifying the degree of locust gregariousness in order to quantitatively compare its levels before and after the experimental alteration. There is no qualitative, let alone quantitative, molecular marker for the locust phase and, thus far, a custom behavioural model has to be built for each experiment. Yet, the interest in locust gregariousness is as high as ever, with a clear increase and foreseen intensification of the research on the topic (see Bakkali [2013]). It would therefore be very useful and time saving to build an accessible, ready to use and standardized tool for evaluating locust gregariousness.

Several methods have been developed in order to characterize the phase state of the locusts. The initial ones, based on morphometry [Stower et al., 1960, Symmons, 1969], are useful for discriminating between solitarious and gregarious adults. But their indices are extracted from adults and not applicable to nymphs —given the significant morphometric differences between nymphs and adults and between nymphs at different developmental stages. In addition, the morphometric indices are not of much use for testing the same animals before and after a treatment. A solution to that is to use different traits. One way is based on the analysis of the spatial distribution of groups of locusts [Cisse et al., 2015]. Resembling cluster analysis approaches such as Taylor's power law [Taylor, 1961] and Iwao's patchiness regression [Iwao, 1968], this recent actualization and improvement of an untested method initially used by Ellis [1963] seems a valid and even promising method, as it uses the tendency to aggregate as criterion and relays on the statistical significance for testing various individuals at once. However, locust gregariousness is not only about aggregating. It also involves changes in a plethora of aspects including morphology and degree of activity. Hence, the most used method is based on multivariate logistic regressions (MLR, Roessingh et al. [1993]). It relies on extracting several behavioural variables from individuals of the two extremes of the polymorphism and establishing MLR formulae that allow estimating the probability of being gregarious.

Every work that uses the MLR approach has to include video tracking of the experimental locusts before and after the experimental treatment, extraction of the behavioural parameters from the video tracks, and application of the MLR formula in order to assess the levels of gregariousness and detect any potential effect of the treatment. Prior to that, the researchers also have to build the MLR formula (e.g., Guo et al. [2011], Ma et al. [2011], Roessingh et al. [1993], Simpson et al. [1999]), thus adding steps to the ones needed for the research sensu stricto. Researchers have to undergo the hassles of video tracking several solitarious and gregarious locusts, extracting several behavioural parameters from the video tracks, choosing which parameters to use, and performing regression analyses in order to get the MLR formula that will then be applied to test the effect of the molecule or gene being investigated. They risk using sample-specific models if they build them using locusts that are not the experimental ones and use them with no prior testing on independent samples. There is also the risk of habituation of the experimental locusts if these were the ones used for model building as well as for the experiment itself.

Surprisingly, the formulae used for the different works were always different from each other and never tested on different locust populations (see references above), thus increasing the risks of heterogeneity in work standards and criteria. We thus aim at establishing accessible formulae that work for different locusts' populations (samples) so that they can be reliably used by researchers to save time and effort. We foresee that the use of these formulae would standardize criteria between experiments and research groups to the benefit of the homogeneity and validity of the results. We chose the desert locust *Schistocerca gregaria* and the migratory locust *L. migratoria* as study organisms, since they are the most widely used species for research on locusts and gregariousness. We evaluated the usefulness of several morphometric and behavioural parameters and we facilitate a script that makes their extraction from the video tracks easy and accessible. We also test the effect of sex, size and developmental stage on the behavioural variables and we establish formulae based on the MLR approach. We offer two formulae for the researcher to use on the main pest locust, *S. gregaria*. The formula that includes morphometric variables works on locust nymphs, and is intended for comparative works between different locust samples or between different developmental stages of the same sample (after molts), whereas the formula that does not include morphometric variables is intended for comparative works on the same locust sample before and after a treatment, be it nymphs or adults. Both models (formulae) were tested and found to work both for nymphs and, at a lesser level (at least the one without morphometric variables), for adults. They also work for different populations and seem to be quantitative. They could therefore be used for works that need quantifying locust gregariousness.

1.2 Material and Methods

1.2.1 Locust rearing

Gregarious *S. gregaria* locusts were collectively reared in the large wooden cages described in the methodology section, then they were redistributed in other large wooden cages containing 10, 10, 16, 20, 22, 80, 150, or 300 individuals each. Solitarious locusts were individually reared, also for over three generations, in small wooden cages. We bought 50 *L. migratoria* specimens from a pet store and, after two generations of rearing at low densities in our insectarium, only 14 individuals remained. As gregarious *L. migratoria*, we bought two more sets of 100 nymphs, some of which we used for behavioural model building. The rearing conditions for *S. gregaria* and *L. migratoria* are discribed in detail in the general methodology section.

1.2.2 Video tracking and photography

Observation of the behaviour was carried out in a 60 cm x 60 cm x 60 cm arena (same size as the large rearing cages). Two of the arena's opposite sides were open so that at one of them the glass side of an empty large rearing cage was placed whereas, at the opposite side, was the glass side of a large rearing cage containing the stimulus (100 gregarious locusts). The needle end of a 30 ml plastic syringe was cut and the syringe was placed at a bottom whole at the middle of one of the observation arena's wooden sides. The experimental animals were introduced into the observation arena through that syringe and a PowerShot G6 digital camera (Canon) was placed at the exact middle of the observation arena's top side for behaviour recording. Figure 1.10 shows a schematic representation of this experimental setting.

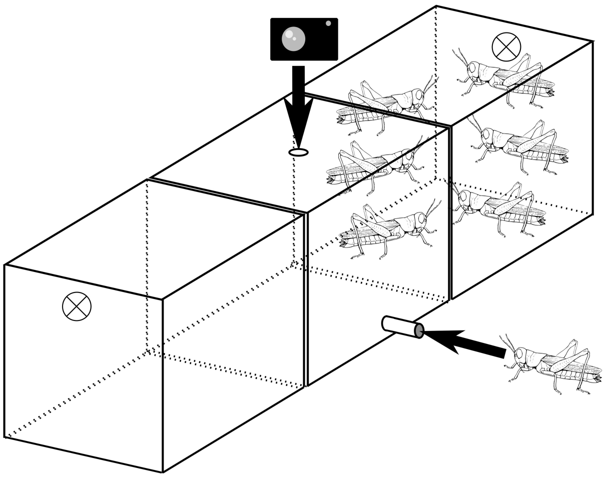

Figure 1.10: Schematic representation of the observation arena where the behavioural observation and video recording of the individual experimental locusts were conducted. The crossed circle in the central top region of the rear side of each lateral box represents a light bulb. The various locusts kept at one side of the arena were the stimulus group.

Videos were recorded for 180 seconds (three minutes) or until the experimental animal reached the stimulus or blank side of the arena —if it did in less than three minutes. The recording was stopped if the experimental animal climbed the arena's wall and went out of the camera's range for 9 seconds (5 % of the 180 seconds recording cut-off). Videos were taken for adults and late nymphs of both sexes and phases. Nymphs smaller than 1.5 cm in length were not used to avoid possible interference from excessive handling and mechanical damage to such fragile specimens. Biotrack software (`https://github.com/biotracking/biotrack`) was used to get the X and Y coordinates of the experimental animal in the observation arena every 67 milliseconds time frame of the total duration of the recording. Behavioural parameters were extracted from the Biotrack outputs using the custom built R script *btf_analyzer.R* available at `http://www.ugr.es/~mbakkali/mlr_sup.zip`.

For collection of morphometric and colourimetric data, photographs of the dorsal and lateral sides of each experimental animal were taken using the same digital camera as above and in constant conditions of space and light (same white surface, place, lamp and light incidence angle). A scale was placed near the specimen in order to obtain standardized leg, thorax and head measures from the pictures using the TPSDig v1.40 software [Rohlf, 2004]. The RGB colour channel values were extracted from the specimen's pronotum area of the picture using GIMP software —the colour values of the specimen-less background of the picture were used for normalization.

Each experimental animal was photographed and video recorded. We therefore recorded video trackings and took dorsal and lateral photographs for 189 *S. gregaria* individuals distributed as follows: 28 solitarious (15 nymphs and 13 adults) and 161 gregarious individuals corresponding to 8 individuals from two 10 individuals density cages, 32 individuals from one 16, two 20, and one 22 individuals density cages (the animals of these four cages were treated as if coming from a single ∼20 density cage), 40 individuals from a 80 individuals density cage, 51 individuals from a 150 individuals density cage and 30 individuals from a 300 individuals density cage. In addition, we similarly photographed and took videos of 14 solitarious and 30 gregarious *L. migratoria* nymphs, as well as 40 more gregarious nymphs (30 crowd-reared and 10 reared at low density for 1 week), 4 solitarious adults and 15 gregarious adults of the same species.

1.2.3 Behavioural, morphometric and colourimetric data collection

Ten behavioural variables were calculated from the video trackings:

1. Elapsed time (ET, total time of the recording)
2. Total distance (TD, total distance traveled by the experimental animal during the recording time)
3. Average speed (AS, average of dividing the distance increments by the time increments at each time frame)
4. Average acceleration (AA, average of dividing the speed increments by the time increments at each time frame)
5. Stop ratio (SR, number of 67 milliseconds time frames with no distance increment divided by the total number of time frames of the recording)
6. Choice (CH, a binary variable describing the side of the arena where the experimental animal was positioned at the end of the recording, with 0 being the blank side and 1 being the stimulus side)
7. Last coordinate (LC, value of the observation arena's X-axis coordinate where the animal was positioned at the end of the recording, this ranged between -150, at the blank side and +150, at the stimulus side)
8. Choice/time (CT, last coordinate divided by the elapsed time)

9. Turn ratio (TR, number of time frames when the animal turned divided by total number of time frames of the recording. A deviation from the path is considered as turn if the angle increment between two consecutive time frames exceeds 9 degrees —5% of a full, 180 degrees, turn)

10. Erratic movement (EM, the sum of the product of the turn angle by the distance at each time frame).

Six morphometric (meassured as in figure 1.11) and three colourimetric variables were obtained from the pictures of each experimental animal:

1. Hind femur length (the length of the left hind femur)

2. Pronotum dorsal length (the length of the largest sagital section of the antero-dorsal pronotum)

3. Head width (the length of the widest transverse head section)

4. Pronotum-femur index (PF, pronotum dorsal length divided by hind femur length)

5. Pronotum-head index (PH, pronotum dorsal length divided by head width)

6. Femur-head index (FH, hind femur length divided by head width)

7. Corrected R value (red value from RGB channel normalized by blank —blank being the red value of the entire picture excepting the animal itself)

8. Corrected G value (green value from RGB channel normalized by blank —blank being the green value of the entire picture excepting the animal itself)

9. Corrected B value (blue value from RGB channel normalized by blank —blank being the blue value of the entire picture excepting the animal itself)

The sex, phase, density of the rearing cage and stage (nymph or adult) were also recorded for each experimental animal.

1.2.4 Confirmative testing of the phase state of the locusts

To further ensure that the animals that we are using for model building are the right ones, morphometric, colourimetric and molecular comparisons

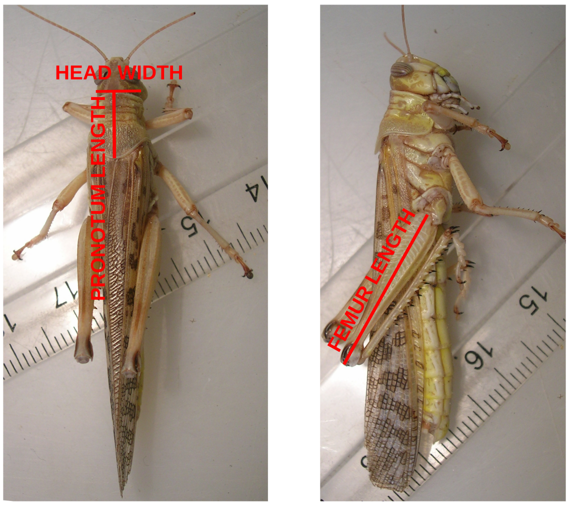

Figure 1.11: Detail of the taken morphometric meassures. Both pictures show a gregarious adult male of *S. gregaria.* A: Dorsal view picture with the Pronotum length and Head width meassures. B: Sagital view picture with the Femur length meassure.

were carried out between some solitarious and gregarious *S. gregaria* locusts. Kruskal-Wallis test with the animals' phases as factor and their respective femur, pronotum, hind leg lengths, and the values given by the red, green, and blue colour channels as variables was carried out in order to confirm the effect of our lab animals' phase on their morphology. In addition, qPCR estimation of the relative expression levels of the three genes chemosensory protein 12, phosphoenolpyruvate carboxykinase and yellow-h was carried out on cDNAs from 4 solitarious and 4 gregarious *S. gregaria* locusts with the tubulin gene as housekeeping, as described in the general methodology section of this thesis.

1.2.5 Exploration and normalization of the variables

Linear regressions were carried out and Pearson's correlation coefficients were obtained in order to check for effect of the morphometric variables on the behavioural variables of animals belonging to the same phase, cage, stage and sex. The behavioural variables elapsed time, distance, average speed, average acceleration, choice/time and erratic movement were then normalized by the length of the hind leg femur in order to reduce the effect of the animals' size on the speed-related behavioural variables within the same phase group. Barttlet's test was then carried out between non-normalized and normalized variables to confirm that the variance of the normalized values of the behavioural variables of the animals belonging to the same phase was significantly lower than the variance of the non-normalized variables. All statistical analyses were performed in R environment version 2.15.1 [Gentleman et al., 1997].

Only extreme solitarious and gregarious individuals (density 1 and 150 or more animals per cage) were used for exploring the variables that might be useful for building the behavioural model. Principal Components Analysis (PCA) was thus carried out to explore the relationships between all the morphometric, colourimetric, behavioural, sex, stage and phase variables. Pearson's correlation coefficient was calculated for all the possible pairs of variables in order to detect the highly correlated ones that might bias the behavioural models. To decide whether separate multivariate models have to be built for each sex, the effect of the animal's sex on the variables was assessed using Kruskal-Wallis tests.

1.2.6 Building of the multivariate model for inferring the locusts phase

Three steps were taken in order to build the most inclusive and accurate multivariate model using the lowest possible number of morphometric and behavioural variables. We made sure that we are using animals that show extreme behaviours. For that, we built our first models using all the variables (except the colourimetric ones) and all the extreme solitarious and gregarious locusts and, when applicable, we removed the gregarious locusts that show less than 0.8 probability of being gregarious. A first approach was building an extended model using extreme solitarious and fully gregarious locusts and all the 13 variables. A second approach was leaving only one representative of each set of highly correlated variables (the one that has the highest sum of the absolute values of its 13 PCs loadings), then building a low redundancy model and comparing its performance to that of the extended one (using pairwise t-test). As third and last approach, a series of successive models were built by successive elimination of the weakest variable (the one with the lowest sum of the absolute values of its 13 PCs loadings) until obtaining a minimal working model.

The best models were selected and applied to independent animal samples to test their reproducibility. To improve the sensitivity of the models and quantitative nature the formulae for calculating their P-values were corrected using adequate correction factors c (see below). After further testing of the best model on adults, we generated a simplified version of it that does not include morphometric variables. We also built, corrected and tested models for

L. migratoria, for which the *S. gregaria* models didn't work. In supplementary figure 1.12 we show a flow-diagram with the specified steps of model selection criteria.

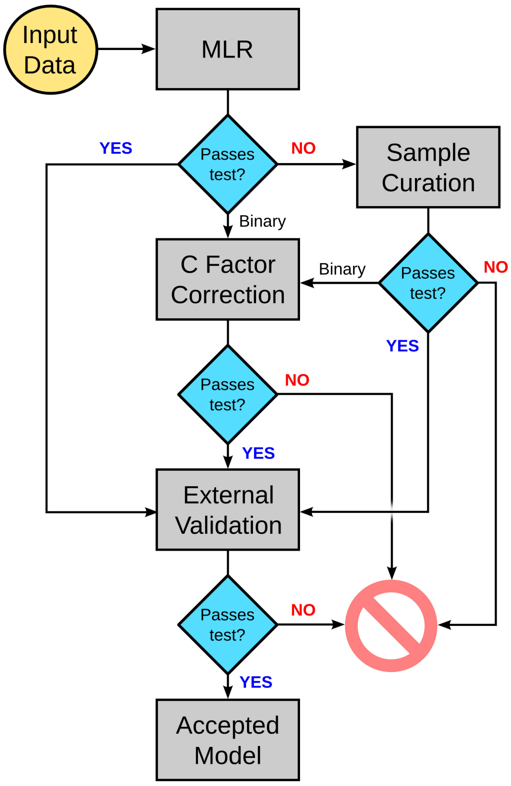

Figure 1.12: Workflow diagram of model validation.

The models were based on MLR formulae. We first calculated $\eta = \beta_0 + \beta_1 V_1 + \beta_2 V_2 + \ldots + \beta_k V_k$, where β_0 is the intercept constant, β_k the regression coefficients of the variable k and V_k is the value of the variable k of the same individual. A locust's probability of being gregarious was then calculated as $P_{greg} = e^{c\eta}/(1 + e^{c\eta})$, with e being Euler's number (2.718), η the result of the regression, and c a correction factor (c = 1 / xk, where x is a real number and k is the number of variables in the MLR). Akaike's information criterion, AIC [Akaike, 1974], was calculated for each MLR to help us decide which model to suggest when the results of different models were not significantly different from each other .

1.2.7 Model performance testing

To test our models on other locusts (independent cages) and assess their performance on adults as well as their abilities to discriminate between locusts colonies of different densities (gregariousness), we applied them to locusts from six additional and independent nymph and adult locust cages of different densities. We used animals from seven cages that contained one solitarious adult each, two cages that contained 10 nymphs each, four cages that contained 16, 20, 20 and 22 nymphs (considered as ∼20), a cage of 80 nymphs and a cage of 300 adults. We also tested the usefulness of our *S. gregaria* model for other pest Orthoptera by applying it to solitarious and gregarious *L. migratoria*. Species-specific models were then built for that species, using 18 solitarious and 30 gregarious nymphs, as we did for *S. gregaria*. The *L. migratoria* models were also improved and tested on two independent samples of 30 and 10 gregarious nymphs, as well as on a 14 gregarious adults sample.

1.3 Results

1.3.1 Adequacy of the material

After several generations of rearing in individual isolation or at high densities the solitarious and gregarious *S. gregaria* locusts of our lab colony showed clear differences both at the colourimetric (table 1.4, figure 1.13A), morphometric (table 1.4) and gene expression (figure 1.13B) levels. In terms of colours, and in agreement with the known dissemblance between solitarious and gregarious *S. gregaria*, the differences were especially pronounced in nymphal stages with the solitarious nymphs being greenish whilst the gregarious ones having a mixed pattern of yellow and black. Solitarious and gregarious nymphs also showed significant differences for all the measured morphometric traits whereas only the head width showed significant differences between adults of the two phases.

The clear non-behavioural differences between solitarious and gregarious *S. gregaria* in our lab colony extend to the molecular level. Based on transcriptomics data, we earlier found both the chemosensory protein 12 (*csp12*), phosphoenolpyruvate carboxykinase (*pck*) and yellow h (*y-h*) genes to have consistently higher expression levels in gregarious *S. gregaria* locusts compared to solitarious ones. A similar result was reported for a *L. migratoria* CSP [Guo et al., 2011]. qPCR testing of our *S. gregaria* locusts using those three genes, and tubulin A1 as reference gene, confirmed the RNAseq data and showed higher expression levels in our gregarious *S. gregaria* than in solitarious ones (figure 1.13B) —0.8 fold increase for *csp12* and *pck* and 0.5 for *y-h*.

	Nymphs		Adults	
	Chi-square	p-value	Chi-square	p-value
Pronotum	Pronotum	13.288	0.0003	0.556
Femur	Femur	12.6	0.0004	2.83
Head width	Head width	14.718	0.0001	7.315
R channel	R channel	7.536	0.0061	0.96
G channel	G channel	6.664	0.0098	0.879
B channel	B channel	7.616	0.0058	0.0005

Table 1.4: Comparison of the morphometric and colourimetric variables between our laboratory-reared solitarious and gregarious *S. gregaria* nymphs and adults. The table shows the values of Kruskal-Wallis test with 1 degree of freedom and its p-value.

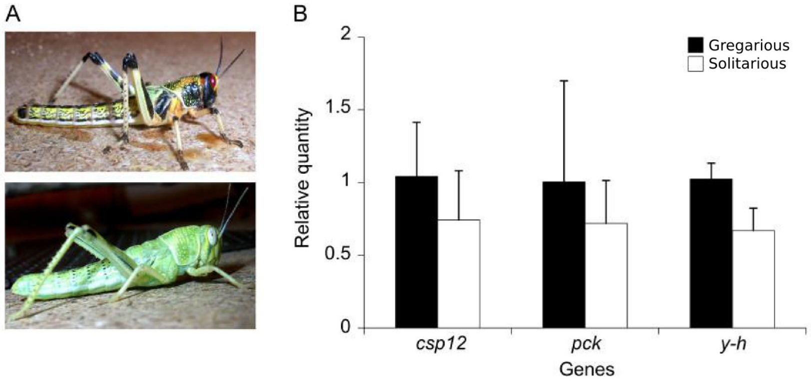

Figure 1.13: (A) Colour differences between the gregarious (up) and solitarious (down) locust nymphs from our laboratory colonies. (B) Comparative qPCR estimation of the expression levels of three gregariousness potential marker genes in individuals from our gregarious (black bars) and solitarious (white bars) *S. gregaria* colonies. Both the chemosensory protein 12 (*csp12*), the phosphoenolpyruvate carboxykinase (*pck*), and the yellow-h (*y-h*) genes were reported elsewhere and confirmed in our RNA-Seq based work (in preparation) to be over-expressed in gregarious locusts. Accordingly individuals from our gregarious *S. gregaria* colony showed less expression levels than individuals from our solitarious colony of the same species. On average gregarious locusts show about 0.8 fold increases in *csp12* and *pck* expression and about 0.5 increase for *y-h* expression.

1.3.2 Data exploration and processing

Noticeably, regression analyses using *S. gregaria* animals of the same cage (phase), sex and stage showed that the morphometric variables significantly correlate only with the elapsed time variable (table 1.5). Division by the hind femur length for normalization of the speed-related behavioural variables attenuated such effect (table 1.5). Accordingly, the normalized behavioural variables showed significantly smaller variances between animals of the same cage, sex, and stage than did the non normalized ones (table 1.6). It should be noted that, when the animals are of similar sizes (the solitarious sample used for this analysis), the normalization has a significant but clearly weaker effect (table 1.6).

	Femur		Pronotum		Head width	
	Raw	Normalized	Raw	Normalized	Raw	Normalized
Elapsed time	**0.363**	0.116	**0.332**	0.149	**0.387**	0.178
Total distance	0.091	0.083	0.141	0.135	0.169	0.164
Average speed	0.076	0.047	0.122	0.094	0.147	0.121
Average Acceleration	-0.088	**-0.360**	-0.129	**-0.399**	-0.169	**-0.415**
Choice by time	0.027	-0.087	-0.052	-0.144	-0.032	-0.127
Erratic Movement	0.215	-0.262	0.187	-0.228	0.228	-0.103

Table 1.5: Pearson's correlation coefficients between the morphometric and behavioural variables of the *S. gregaria* locusts before and after normalization of the behavioural variables of each animal by its hind femur length. We used 106 animals and all the significant correlations, in bold, had p values lower than 0.001.

	All	Nymphs	Adults	Gregarious	Solitarious	Males	Females
Elapsed time	39.487	7.999	21.444	33.323	8.416	23.943	14.086
Total distance	49.844	5.227	18.328	37.792	13.371	27.807	18.031
Average speed	49.930	17.103	18.317	37.965	9.938	26.029	18.080
Average Acceleration	39.573	12.545	16.813	40.688	7.182	25.076	8.670
Choice by time	28.314	17.044	16.386	27.299	4.594	19.433	7.958
Erratic Movement	51.857	7.116	19.399	36.549	16.533	28.589	19.464

Table 1.6: Homogenization of the behavioural variables after their normalization by the length of the animal's hind femur. The table shows the Bartlett's K-squared statistic values, all of them with p values ranging between 2.224 10^{-2} and 1.158 10^{-12}.

Initial exploration of the data using the Kruskal-Wallis test showed significant differences between locusts' phases for the femur/head, pronotum/head, elapsed time and total distance variables (table 1.7). However, a principal components analysis did not reveal any clear clustering of the variables, nor did it detect any variable as allowing unambiguous separation between groups of individuals (figure 1.14, table 1.9). None of the absolute values from principal components (PCs) loadings per variable was equal or over 0.7 (table 1.9) so the first three PCs were needed in order to explain more than 70% of the cumulative variance (table 1.9). Accordingly, and in the absence of a single marker, multivariate models need to be built in order to estimate locusts' probability of being gregarious (P_{greg}).

	Chi-squared	p-value
Pronotum-Femur index	2.047	0.153
Pronotum-Head index	9.412	0.002
Femur-Head index	11.178	0.001
Elapsed time	7.048	0.008
Total distance	4.173	0.041
Average speed	0.774	0.379
Average Acceleration	0.774	0.379
Stop Ratio	0.886	0.347
Last Coordinate	0.991	0.319
Choice by time	0.312	0.576
Turn Ratio	8.319	0.004
Eratic Movement	2.606	0.107
Choice	2.764	0.096

Table 1.7: Comparison of the values of 13 morphometric and behavioural variables between solitarious and gregarious locusts. The table shows the values of Kruskal-Wallis test with 1 degree of freedom and its p-value.

To detect highly correlated variables that might reinforce or bias the model towards a particular trait, we carried out pairwise correlations between all the morphometric, colourimetric and normalized behavioural variables. Several of these correlations showed Pearson's correlation coefficient whose absolute value was of at least 0.7 (table 1.10). These were average speed vs. average acceleration (r = 1), vs. stop ratio (r = -0.705) and vs. choice by time (r = 0.704), choice vs. last coordinate (r = 0.933), pronotum-femur index vs. femur-head index (r = -0.849), stop ratio vs. turn ratio (r = -0.814) and vs. average acceleration (r = -0.703), total distance vs. erratic movement (r = 0.829), and vs. elapsed time (r = 0.828), and choice by time vs. average acceleration (r = 0.704). Of each of these pairs, we eliminated the variable that had lower sum of the absolute values of the 13 PC loadings (table 1.9). The femur-head index, choice, total distance and turn ratio variables were therefore retained. These four variables, together with the remaining less correlated variable (pronotum-head index) were selected for building a *Sg_low-redundancy* multivariate behavioural model for estimating *S. gregaria* locusts probability of being gregarious (see bellow). None of the 13 variables were significantly affected by the sex of the animal, neither before nor after adulthood, meaning that no separate models need to be built for males and females (table 1.11).

1.3.3 Model building

A first *Sg_extended* model (AIC = 28) was built using all the 13 variables extracted from 15 solitarious and 51 gregarious *S. gregaria* nymphs. The

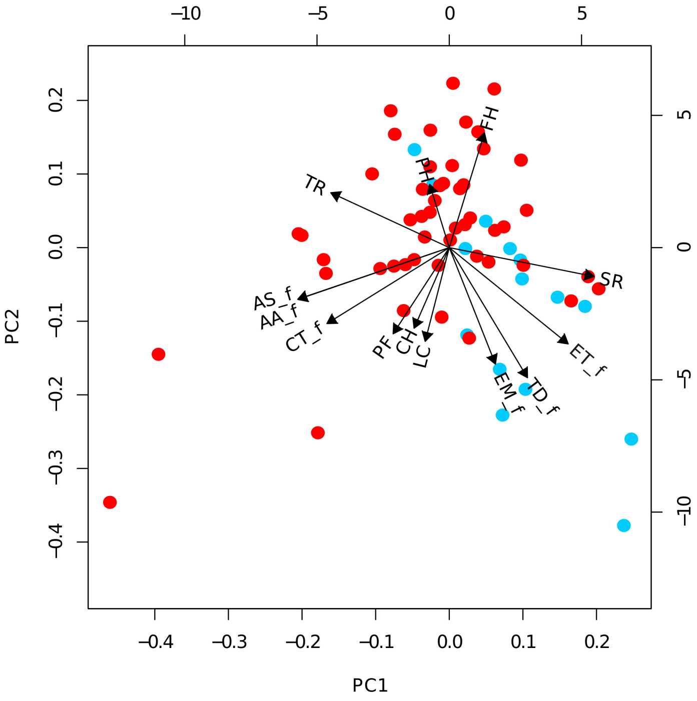

Figure 1.14: Biplot of the principal components (PCs) 1 and 2. It shows how the variables are related to both PCs and to the extreme *S. gregaria* nymphs used in the study (gregarious in red and solitarious in blue). PF = Pronotum-Femur index, PH = Pronotum-Head index, FH = Femur-Head index, ET = Elapsed time, TD = Total distance, AS = Average speed, AA = Average acceleration, SR = Stop ratio, CH = Choice, LC = Last coordinate, CT = Choice by time, TR = Turn ratio, EM = Erratic movement. An "*_f*" indicates that the variable has benn normalized by femur length.

resulting formula is described in table 1.8 and the P_{greg} detected all the 51 gregarious nymphs as gregarious with 100% probabilities and attributed 0% gregariousness probability to all our 15 solitarious nymphs (figure 1.15A). Obtaining the three morphometric indexes used as variables here is as easy as measuring the length of the hind leg femur, the dorsal length of the pronotum and the dorsal width of the head. In addition, we provide a script that automates the extraction from video tracks of the ten behavioural variables used for this work (script at `http://www.ugr.es/~mbakkali/mlr_sup.zip`). Consequently, applying the *Sg_extended* model should be accessible and straightforward. Still, we also tried less complex models after elimination

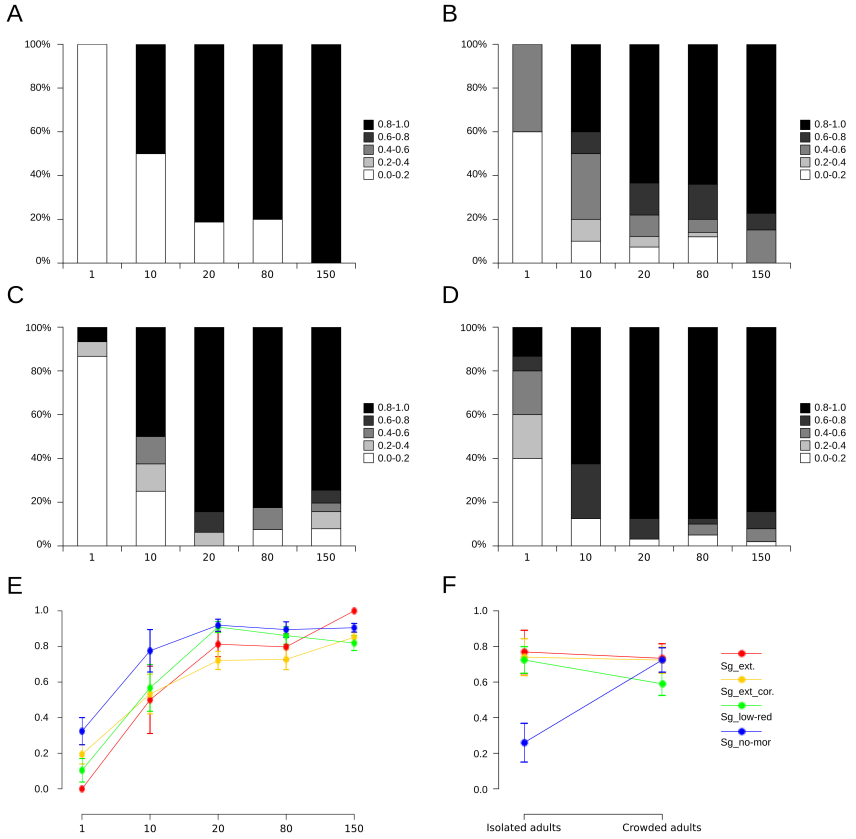

Figure 1.15: Performance of the tested models for *S. gregaria*. X-axes represent the tested population (density 1 for solitarious locusts; 10, 20, 80 and 150 for crowd-reared gregarious locusts). Y-axes are the percentage of individuals of a sample (A-D) or the average P_{greg} value of the entire sample (E, F). We divided the P_{greg} values into five intervals represented from white to black as follows: 0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, 0.8–1. (A) *Sg_extended* model (B) *Sg_extended_corrected* model (C) *Sg_low-redundancy* model (D) *Sg_non-morphometric* model. (A) *Sg_extended* model (B) *Sg_extended_corrected* model (C) *Sg_low-redundancy* model (D) *Sg_non-morphometric* model. (E, F) Average P_{greg} values with the associated standard error for each model in *S. gregaria* nymph (E) and adult samples (F). Sg_ext. = Sg_extended; Sg_ext_cor. = *Sg_extended_corrected*; Sg_low-red. = *Sg_low-redundancy*; Sg_no-mor. = *Sg_non-morphometric*.

of some of the variables based on two logics.

The first logic relies on reducing redundancy based on elimination of one of each two highly correlated variables. In this case, the resulting *Sg_low-redundancy* model was built using the five less-correlated variables (see above). Its formula (AIC = 39.413) is as described in table 1.12 and, as figure 1.15C shows, the resulting distribution of the P_{greg} values of our solitarious and gregarious *S. gregaria* locusts were not as clear-cut and different compared to the results of the *Sg_extended* model — 0.104 ± 0.066 and 0.819 ± 0.041 being the average P_{greg} that this *Sg_low-redundancy* model gave to 15 solitarious and 51 gregarious locusts, respectively (figure 1.15C). As mentioned above, the *Sg_low-redundancy* model seems not as accurate in predicting gregarious locusts as the *Sg_extended* model, with both models giving similar results for the solitarious sample ($t = -1.581$, $p = 0.136$) but significantly different results for the gregarious sample ($t = 4.347$, $p < 0.001$). The second logic relies on reducing the number of variables used for model building by stepwise elimination of the weakest variables until the step previous to the first significant alteration of the model's performance. Variables' weakness was decided based on the sum of the absolute values of their PC loadings (see table 1.9). This attempt at building a *Sg_minimal* model led to no way as the model's results became significantly different after removal of the first variable (the pronotum-head width index).

1.3.4 Model testing and improvement

The two models (*Sg_extended* and *Sg_low-redundancy*) were applied to additional independent samples of 30, 32 and 8 locusts grown at densities of 80, ∼20 and 10 individuals per cage, respectively. The *Sg_extended* model showed almost no differences in average P_{greg} between locusts rared at 80 and 20 densities and lower values for locusts grown at 10 individuals density (0.798 ± 0.073, 0.813 ± 0.070 and 0.500 ± 0.189, respectively). A similar tendency was given by the *Sg_low-redundancy* model, with 0.861 ± 0.050, 0.909 ± 0.031 and 0.567 ± 0.132 as average P_{greg} values for the locusts of the 80, ∼20 and 10 individuals per cage, respectively (figure 1.15E). Overall, our two *S. gregaria* behavioural models not only seem to work for independent samples, but also appear to be quantitative. However, when looking beyond the means and into individual values, *Sg_extended* model seems to behave in an almost binary way (figure 1.15A,E). The very high positive and very low negative regression results (η) inevitably make the P_{greg} values tend to either 1 or 0 for gregarious and solitarious locusts, respectively. The range of values of an MLR result might depend on the number of variables of its regression formula. In fact, the *Sg_low-redundancy* model contains less variables than the *Sg_extended* one and gives less extreme results. We thus conducted a series of homogeneous corrections in order to increase the sensitivity of our models in the near 0 and near 1 P_{greg} areas (i.e., in case of very high and very low regression results). We tested dividing the results of the model's MLR formula by multiples of the number of its variables. We carried out simulations and considered a corrected model to be altered beyond our thresholds if it gave average P_{greg} above 0.2 (20%) for solitarious locusts, or below 0.8 (80%) for gregarious ones. As figure 1.16 shows the *Sg_extended* model loses power, both for the solitarious and gregarious samples, after division of the regression results by more than 8 times (x) the number of its variables (k). Division of the result of the *Sg_extended* model regression by 104 therefore results in a significantly more sensitive *Sg_extended_corrected* model (figure 1.15B) both for the solitarious sample (t = -3.561, p < 0.001) and for the gregarious one (t = 5.616, p < 0.001) —the average P_{greg} values were 0.194 ± 0.055 and 0.852 ± 0.026, respectively. The simulations suggest that the *Sg_low-redundancy* model tolerates no correction as it gave average P_{greg} values that surpass the thresholds after the first correction step (i.e., after dividing by the number of variables in the MLR formula). The *Sg_extended_corrected* model, however, performed in a non-binary and significantly different way from the *Sg_extended* model for the extreme phase samples (t = -3.561, p = 0.003 for the solitarious sample and t = 5.616, p < 0.001 for the gregarious sample). It produced more adjusted values for the intermediate density samples, with the average P_{greg} values of locusts reared at 80, 20 and 10 individuals per cage being 0.728 ± 0.058, 0.721 ± 0.051 and 0.533 ± 0.112, respectively. Earlier we explained that we did not use colourimetric data for simplicity and because nymphs and adults of the same phase have significantly different colours. The latter reason aimed at building a single model for both stages. However, in addition to colour, nymphs and adults are different in other aspects that even the normalization by the hind leg femur length that we adopted for our model building might not attenuate (table 1.6). To test whether our *Sg_extended_corrected* model also works for adults, we applied it to 30 individuals randomly taken from a cage of 300 gregarious adults and 8 solitarious adults. The model classified the tested gregarious adult locusts as gregarious with about 72 % probability (0.723 ± 0.073), but the average P_{greg} values of the 8 solitarious adults was of 0.740 ± 0.103 (figure 1.15F). Worse was the *Sg_low-redundancy* model, as it failed to properly predict both the gregarious (0.590 ± 0.066) and solitarious adults (0.724 ± 0.075).

The *Sg_extended_corrected* model thus differentiates between solitarious and gregarious *S. gregaria* nymphs and seems even to be quantitative (figure 1.15E). However, its morphometric variables are not needed and might even introduce unnecessary noise when testing the same animals unmolted (for instance before and after a treatment). They might be the reason why the *Sg_extended_corrected* model did not perform well on adults. For that, we built a new *Sg_non-morphometric* model (AIC = 54.306) in a similar way as before using the 10 non-morphometric variables and the solitarious and extreme gregarious nymphs. The resulting formula is in table 1.8 and, the

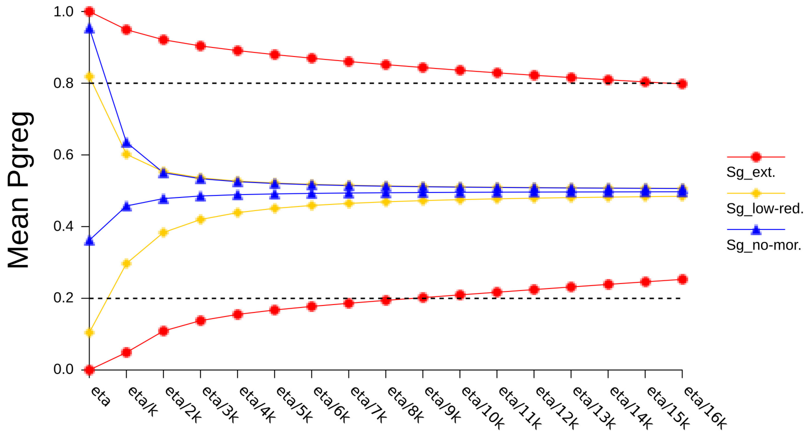

Figure 1.16: Identification of the models' correction factors by assessment of the changesof the average P_{greg} of extreme solitarious (bottom charts) and gregarious (upper charts) *S. gregaria* and *L. migratoria* nymphs that the models gave after application of different correction factors. X-axis shows the correction factor applied to the multivariateregression results before calculating the P_{greg}, where k is the number of variables of the MLR of the corresponding behavioural model. The Y-axis shows the P_{greg}. Ouracceptance thresholds for the corrected models' on solitarious (P_{greg}= 0.2) and gregarioussamples (P_{greg}= 0.8) are marked by a horizontal dashed line. Sg_ext_cor. = Sg_extended_corrected; Sg_low-red. = *Sg_low-redundancy*; Sg_no-mor. = *Sg_non-morphometric*.

model classified the solitarious and gregarious locusts as such, with average P_{greg} of 0.324 ± 0.076 and 0.905 ± 0.024, respectively. This model does not tolerate correction as its results were not binary and the average P_{greg} values it gave to locusts reared at intermediate densities were intermediate: 0.776 ± 0.119 for the density 10, 0.919 ± 0.033 for the density 20 and 0.895 ± 0.044 for the density 80 (figure 1.15D,E). Interestingly, the *Sg_non-morphometric* model properly classified the adult samples attributing average P_{greg} values of 0.724 ± 0.067 to the sample of 30 gregarious *S. gregaria* adults and 0.259 ± 0.109 to the 8 solitarious adults (figure 1.15F).

1.3.5 Modeling for *Locusta migratoria*

We tested our working *S. gregaria* models (*Sg_extended_corrected* and *Sg_non-morphometric*) using solitarious and gregarious *L. migratoria*. None of them seems to work for that species as they gave average solitarious

	*Sg_extended_corrected***	*Sg_non-morphometric**
C	1/104	1
Intercept (β_0)	-2.83 10^7	1.12 10^4
Pronotum-Femur index	5.11 10^7	—
Pronotum-Head index	-1.69 10^7	—
Femur-Head index	9.39 10^6	—
Choice	1.03 10^6	8.45 10^3
Elapsed Time	-3.53	5.30 10^{-3}
Total Distance	-6.68 10^2	-2.13 10^1
Average Speed	7.91 10^8	-3.41 10^5
Average Acceleration	-2.53 10^5	-1.47 10^4
Last Coordinate	-3.68 10^3	-1.83 10^1
Choice by time	1.90 10^7	-6.46 10^4
Stop Ratio	-9.04 10^5	-7.53 10^3
Turn Ratio	-5.33 10^{10}	2.04 10^7
Erratic Movement	2.91 10^1	4.35 10^{-1}

Table 1.8: Intercepts (β_0) and variable regression coefficients ($\beta_{variable}$) for each of the two valid models suggested for estimating locusts probability of being gregarious. Each model's formula is: $\eta = \beta_0 + \beta_{P/F}$ x P/F + $\beta_{P/H}$ x P/H ... $\beta_{\text{Erratic move}}$ x Erratic move, and the probability of being gregarious is P_{greg}= $e^{\eta c}/(1+e^{\eta c})$, e being Euler's number 2.718 and c the correction factor (c = 1 / xk, where x is a real number and k is the number of variables in the MLR formula). The corrected models (e.g., *Sg_extended_corrected*) had the same coefficients as their non corrected counterparts (e.g., *Sg_extended*). *: recommended for testing the same nymphs if they do not molt in between and **: recommended for testing adults and different locust samples (including different nymphs, the same nymphs after molting and adults).

and gregarious P_{greg} values of 0.530 ± 0.118 and 0.314 ± 0.073 for the *Sg_extended_corrected* model, and 0.651 ± 0.101 and 0.627 ± 0.071 for the *Sg_non-morphometric* model, respectively. We therefore decided to check for potential differences and species-specific trends of each of the 13 variables used for the model. As figure 1.17 shows, *S. gregaria* has consistently higher average values for the variables pronotum-head index, femur-head index, average speed and average acceleration, whereas *L. migratoria* shows consistently higher values for pronotum-femur index and elapsed time. Total distance, stop ratio, last coordinate, choice, choice by time, turn ratio and erratic movement, however, do not seem to have species-specific tendencies between *L. migratoria* and *S. gregaria.*

Given the apparent species-specific requirements of the behavioural models, and in an attempt to further enrich this work, we decided to build and provide useful behavioural models for *L. migratoria* too. We used 14 solitarious *L. migratoria* nymphs and 30 gregarious ones to build an initial *L. migratoria* behavioural model based on MLR using the 13 variables used ear-

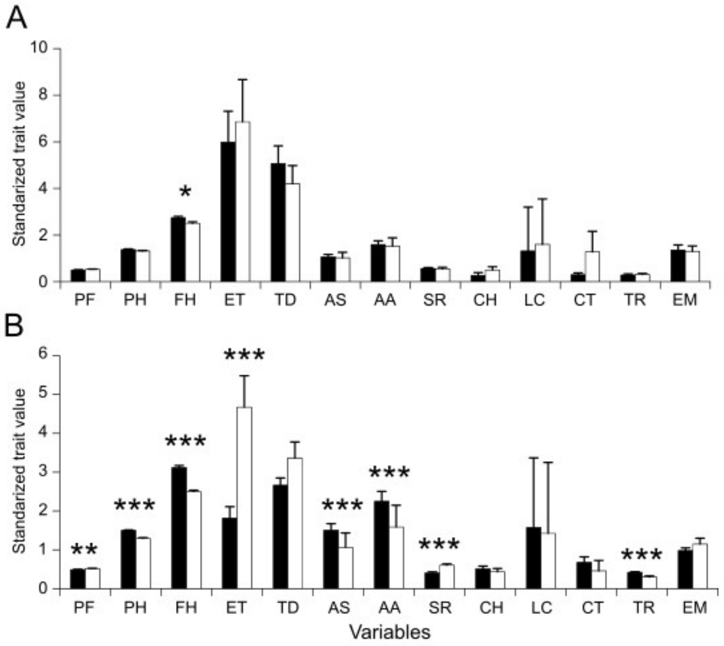

Figure 1.17: Standarized average values of different morphometric and behavioural traits of solitarious (A) and gregarious (B) *S. gregaria* (black bars) and *L. migratoria* (white bars) nymphs. Asterisks indicate the degree of significance of the Kruskal-Wallis test. The values of the different variables were scaled so to fit in a single graph. PF = Pronotum-Femur index, PH = Pronotum-Head index, FH = Femur-Head index, ET = Elapsed time, TD = Total distance, AS = Average speed, AA = Average acceleration, SR = Stop ratio, CH = Choice, LC = Last coordinate, CT = Choice by time, TR = Turn ratio, EM = Erratic movement.

lier for *S. gregaria.* The resulting *Lm_extended model* (formula in table 1.12) separated the two phases but mis-categorized some of the solitarious as well as some of the gregarious locusts. To improve this first model we eliminated the gregarious *L. migratoria* specimens that, for being probably sick/old/altered, showed gregariousness probabilities below 80 % (no solitarious locust was eliminated) and recalculated the regression coefficients for each of the 13 variables to get an *Lm_extended_filtered_sample* model (formula in table 1.12, AIC = 28). The new model successfully categorized all the 19 gregarious *L. migratoria* locusts as gregarious at a 100% probability and all the 14 solitarious ones as such at an almost 0% probability. Following the logic used for *S. gregaria*, a series of divisions of the regression result by multiples of 13 (the number of variables in the regression formula) allowed us to identify c =

1/39 as the model correction factor (figure 1.18C). However, in contrast to *S. gregaria* models, the *Lm_extended_filtered_sample_corrected* model, which kept identifying the solitarious and gregarious *L. migratoria* as such (0.168 ± 0.040 and 0.809 ± 0.033 average P_{greg}, respectively), did not perform well when tested on two independent samples. While it correctly identified one of these samples (it gave an average P_{greg} of 0.956 ± 0.027 for a sample of 10 gregarious *L. migratoria* nymphs), it seemed to underestimate the gregariousness level of a second sample of 30 gregarious *L. migratoria* nymphs, for which it gave an average P_{greg} of 0.558 ± 0.073 (figure 1.18A,D). Removing the morphometric variables and building an *Lm_non-morphometric* model (AIC = 22), as we did for *S. gregaria*, gave average P_{greg} values of 0 and 1 for the solitarious and gregarious *L. migratoria* samples used to build the model, respectively. Application of a correction factor c = 1/40 enhanced the model's sensibility (with average P_{greg} values of 0.126 ± 0.046 and 0.804 ± 0.048 for the solitarious and gregarious samples, respectively) but the new, Lm_non-morphometric_corrected, model failed to correctly classify the two independent *L. migratoria* samples of gregarious nymphs used earlier (average P_{greg} of 0.244 ± 0.068 and 0.412 ± 0.155) (figure 1.18B,D).

Furthermore, none of the *L. migratoria* models succeeded in correctly classifying two samples of 4 solitarious and 14 gregarious *L. migratoria* adults from the same solitarious breed and gregarious cage whose individuals were used to build the model. Both models, and in both cases, gave average P_{greg} values of rather solitarious ranges (the *Lm_extended_filtrered_sample_corrected* model gave average P_{greg} values of 0.250 ± 0.250 and 0.292 ± 0.099 for the solitarious and the gregarious samples, respectively, and the *Lm_non_morphometric_corrected* model gave average P_{greg} values of 0.204 ± 0.204 and 0.310 ± 0.094 for the same samples).

1.4 Discussion

Pest locust outbreaks, as mentioned in the introduction, are one of the most serious threats not only to human agriculture but also to livestock resources and to the equilibrium of the ecosystems of the affected areas. Their infamous importance is further enhanced by the large geographical areas that they affect (see Waloff [1966], Pedgley and David [1981]). The vertiginous developments in genetics and molecular biology mean that the locust problem can currently be approached in a more systematic way and at greater details [Bakkali, 2013]. The research on this topic has therefore shifted from a predominantly entomological [Ellis, 1963, Uvarov, 1921, Gillett, 1973], physiological [Anstey et al., 2009, Tawfik et al., 1999, Wiesel et al., 1996, Gillett, 1975] and cytogenetic level [Fox, 1973, Nolte, 1969] to the transcriptome [Badisco et al., 2011a, Chen et al., 2010, Kang et al., 2004, Zhang et al., 2012], metabolome [Ma et al., 2011] and even genome [Wang et al., 2014a] levels. The wealth of

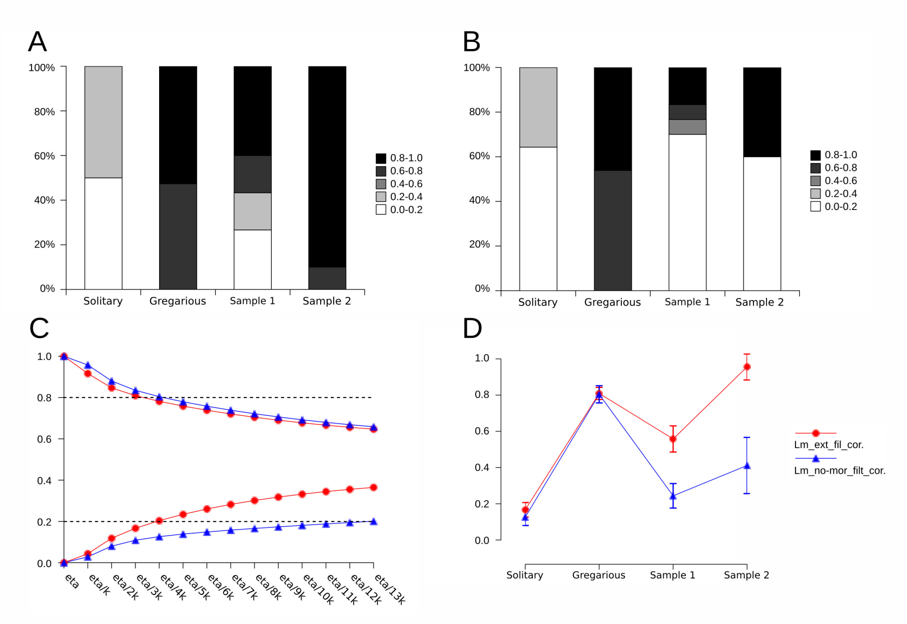

Figure 1.18: Performance of the *L. migratoria* behavioural models on different locust samples of this species. (A) Lm extended filtered sample corrected model. (B) Lm non-morphometric filtered sample corrected model. (C) Identification of the correction factor for each *L. migratoria* model. (D) Average P_{greg} values of each sample as given by the corrected models. Lm_ext_fil_cor. = *Lm extended filtered sample corrected*; *Lm_no-mor_fil_cor.* = *Lm non-morphometric filtered sample corrected.*

data that the current research is generating set the ground for the upcoming era of functional testing. Experimental testing of the effect of several genetic or molecular alterations on the locusts' propensity to outbreak and swarm means testing locusts' outbreak/swarming-related phenotypes both before and after the treatment. The question here is what could we consider as outbreak/swarming-related phenotype?

Locust population outbreaks and swarming are density dependent [Ellis, 1963, Gillett, 1975, Uvarov, 1921], meaning that the first outbreak/swarming-related phenotype that one can possibly think of is the population density parameter. The population density is of course a very valid indicator of the possible state of the population and might even be useful for field-based cause-consequence studies (a field-oriented researcher wouldn't need more than a glance at the locust population density to tell its state). However, population density is of no good use for low scale testing of the effects of experimental molecular manipulation on the locusts' propensity to outbreak and swarm. Indeed, any change in locust population density is more related to reproductive success and survival than to any other locust condition. Furthermore, most works on the topic are lab-based and test the locust sample prior to and after a treatment, without changes in population density. For such studies, the phenotype to consider as indicator of the increase or decrease of the outbreak and swarming propensity should not be inter-generational (must be able to show differences within a generation). Furthermore, being itself just an indicator and not the research question, and in order for it to be quantifier, the outbreak/swarming-related phenotype to use should also be easily measurable. Being a phenomenon that affects almost every aspect of the locusts' biology, locust population outbreaks and swarming offer a plethora of potential indicators of the state of the population. Apart from their developmental, survival, reproductive, immunological and physiological differences, the locusts of the non-outbreak (solitarious) and outbreak (gregarious) phases also differ in their morphology (both colour and shape) and behaviour. The behavioural differences between the two phases have to do with the tendency to aggregate and with the degree of activity (both of which pronounced in the gregarious phase). Both traits lead to a higher propensity to swarm and migrate and allow a more efficient search for new resources as well as reducing the chaos and 'negative' locust-locust interactions in the enormously dense gregarious population (e.g., Bazazi et al. [2012, 2008], Buhl et al. [2006]). The morphological and colourimetric differences between phases, however, have to respectively do with differences in the need for camouflage/warning and the effect of stress and low food resources on the animal's development and resulting size.

We therefore chose to look for potential outbreak/swarming-related phenotype indicators among the traits related to morphology and behaviour. In fact, some of these phenotypes were used as comparative testers of the state of our locust colonies. It is well known that solitarious *S. gregaria* nymphs are greenish whereas the gregarious ones are black and yellow. So were our solitarious and gregarious *S. gregaria* nymphs. The morphometric differences between our solitarious and gregarious locusts further confirmed their states. Still, we went down to the molecular level and found differences in the expression levels of three genes between our solitarious and gregarious *S. gregaria* colonies. The data thus definitely confirm the solitarious and gregarious states of our locusts, which we can therefore use for our work as such with complete confidence —working with the right material is essential to the reliability of the results.

It should be noted that our results also mean that the expression levels of the genes chemosensory protein 12, phosphoenolpyruvate carboxykinase and yellow-h could be used as comparative indicators, not markers, of the phase state; since they consistently appear more expressed in the gregarious locusts both in our transcriptome data (Chapters 2 and 3) and in the qPCRs of the present work. However, we based our work neither on colour nor on the expression levels of the three genes tested here since none of these indicators is proven to be quantitative (a necessary quality to expect from any trait that should differentiate between the states of a locust before and after a treatment) and are not easy to measure (a quality to expect from a trait to use as research tool). Furthermore, the colourimetric differences are specific to *S. gregaria* and, more precisely, to its nymphal states (it can't be used as a general indicator —not even for nymphs and adults of the same species). That being said, our work shows that of two sets of *S. gregaria* locusts, the one that shows greater levels of expression of the genes chemosensory protein 12, phosphoenolpyruvate carboxykinase and yellow-h is the one that is more likely to be more gregarious. A fact that makes sense as these genes respectively have chemosensory, high metabolism and behaviour-related functions —all of which expected to be enhanced in the gregarious locusts.

Once the phase state of our *S. gregaria* locusts colonies has been confirmed, the search for valid outbreak/swarming-related phenotypes to use as indicators of the degree of gregariousness started by measuring 13 traits relating to shape, degree of activity and tendency to aggregate. These and other traits are usually used for building models for testing locust gregariousness [Guo et al., 2011, Ma et al., 2011, Roessingh et al., 1993, Rogers et al., 2014, Simpson et al., 1999]. However, separate models with different combinations of traits are usually built and no single ready-to-use model is hitherto available to the research community. That increases the risks of heterogeneity in the work criteria and even possible lab- and experiment-specificity of the results. A possible case of this could be the controversy around the involvement of serotonin; which was reported as *S. gregaria* gregarizer [Anstey et al., 2009], as *L. migratoria* solitarizer [Guo et al., 2013], with no gregarizing effect on *S. gregaria* [Tanaka and Nishide, 2013] and, again, re-defended as *S. gregaria* gregarizer [Rogers et al., 2014, Rogers and Ott, 2015]. Furthermore, building a model and applying it to the same locusts increases the risks of alteration of the result due to habituation of the animals to the observation arena. On the other hand, building a model using a sample of locusts and applying it to another sample without thoroughly testing it on several independent samples increases the risks of error, due to possible sample-specificity of the model. Our goal is therefore to build, test and provide valid models for the research to come —which should homogenize methods, tools and criteria, with the consequent increase of the validity of the research results and interpretations across the research groups.

The significant correlation between the morphometric variables and the speed-related behavioural ones means that some kind of normalization is needed in order to build models that might work for locusts of different stages and sizes. To our knowledge, that kind of normalization has thus far never been applied in any of the published research that used locust behavioural modeling and testing —meaning that the behavioural models used might have in occasions been sample-specific and could possibly benefit from normalization if used for larger or smaller locusts. The obvious morphological trait that relates to speed is the length of a leg (for instance the abdomen of the same animal has varying lengths, depending on its relaxation and food intake states, and the thorax or head are not directly related to movement). To minimize uncertainties we choose the length of the hind leg femur as normalizer, since the femur is the usually less deformed and damaged part of the longest locust leg (the tibia often gets damaged at the molting time and the uncertainties from length measurements of small legs would be higher). Normalization of the speed-related variables by the hind leg femur length seems thus to work as it had a homogenizing effect on the data from animals of the same size, stage and cage, and since the magnitude of the homogenization depended on the degree of the initial heterogeneity of the data (locust sizes). There are well known size, physiological and behavioural differences between male and female locusts, so one should not simply discard the possibility of significant sex-related differences in the gregarious behaviours between locusts of the same species, stages and rearing densities. However, our data suggest that, after normalization by size, the sex of the locusts did not significantly affect any of the 13 variables. Hence, no normalization by sex, inclusion of the sex as variable, or building of separate sex-specific models are needed.

The fact that four of the 13 traits that we used showed significant differences between locust phases would in principle mean that we might be able to use these single or combined traits as quantifiers of the degree of gregariousness. However, PCA analysis suggested that at least three PCs are needed in order to reach the conventionally used 70% of the cumulative variance (e.g., Viscosi and Cardini [2011]). So, just as we cannot completely trust any single molecular trait as marker, a combination of indicators is needed for morphology and behaviour-based quantification of the degree of gregariousness. With the multivariate approach confirmed as the way to quantify locusts gregariousness, the issues become which of the variables and what animals to use. As to the second question, it is clear that using clearly solitarious and clearly gregarious animals is the response. Yet, while the slow (lazy or shy) solitarious animals are inherently unable to surpass a threshold of activity (due to their constitutive muscle excitation and activity capacities), the gregarious animals can show misleading poor activity —due to sickness, for instance. In other words, a solitarious locust cannot suddenly become as active as a gregarious one and, if sick, it will only look as ‘too solitarious’, whereas a sick or old gregarious animal might show similar degree of activity to solitarious locusts. For that, exploration of the animals was carried out by building an initial model then applying its formula to the data from the same animals in order to decide whether to discard some gregarious locusts or not. The results of the *Sg_extended* model showed that all our *S. gregaria* locusts were properly classified so there was no need for eliminating any gregarious animal. Our gregarious *S. gregaria* nymphs were not only yellow and black and with increased expression of the three tested genes, they also showed

P_{greg} values of 1 —meaning that they were vigorous, fully gregarious and valid for model building. The solitarious nymphs were also adequate as they were greenish, with less expression of the three genes and showed P_{greg} values of 0.

As to the variables themselves, the more variables we use for model building the closer we are to the reality of the animal. However, our aim is to provide a tool for other researchers to use, the model should therefore be as simple as possible. In addition, we cannot use all the possible variables and we need to avoid bias towards a group of highly correlated traits. We thus removed the weakest among each pair of highly correlated variables and built an *Sg_low-redundancy* model. It performed as good as the *Sg_extended* model, probably because reinforcement was not an issue for our 13 variables, but its AIC value was higher than that of the *Sg_extended* model —meaning that the latter is better. Although we provide a script for data extraction from the video trackings, we tried to reduce complexity by stepwise removal of the less important variable up to a minimal model. This strategy led to no good outcome as the prediction power of the potential *Sg_minimal* model worsened significantly, compared to the *Sg_extended* model, after removal of the first variable.

The difficulty of maintaining solitarious locusts —they need to be kept and cared for in isolation and in individual cages —is a limiting factor that forced us to initially use the same animals both for model building and initial testing of the model's performance. Consequently, we incurred the risks of being circular and of having models that only work for the specific sample that we used. Furthermore, using extremes (fully solitarious and fully gregarious) meant that we could not be sure of whether our models are quantitative. To deal with that, we tested our models using independent samples of different developmental stages and rearing densities. On average, both the *Sg_extended* and the *Sg_low-redundancy* models performed well on the different sets of independent nymph samples. We attribute this to the accuracy of the MLR formulae and correcting effect of the normalization of the speed-related variables by the size of the animal. Our models seem quantitative and gave results in accordance with the locusts states —with low values for the solitarious, high values for the fully gregarious, somewhat low values for the somewhat solitarious (density 10 individuals in a big cage), and intermediate values for the locusts reared at intermediate densities (20 and 80 individuals per 60 x 60 x 60 cm cage, although distinguishing between locusts at these intermediate densities falls beyond the models' sensitivity).

The *Sg_extended* and *Sg_low-redundancy* models therefore discriminate between solitarious, gregarious and low density-reared nymphs, and between these and the locusts reared at intermediate densities. However, the results seemed only applicable at the population (whole sample) level. In fact, when looking at the data for each individual locust from the intermediate samples, the results contained a noticeably high mixture of close-to-zero and close-to-one values rather than being all intermediate values. In principle this is no issue, since building models based on logistic regressions and limited-size samples means that their individual results might have tendency to be binary-like (close to 0 or close to 1). nvolve statistically testable samples (populations) of various individuals Furthermore, experiments on locusts involve statistically testable samples (populations) of various individuals. Still, a homogeneous scale-tuning could be applied to adjust the models' sensitivity. If we want P_{greg} to serve as quantifier, it has to discriminate between locusts of the same phase —this way we can detect changes in the degree of gregariousness before and after an experimental testing and not only the shifts from the solitarious to the gregarious states and vice versa. One can never ensure that all the individuals from a population of gregarious locusts are absolutely gregarious, neither can we state that individuals from a solitarious locust population are absolutely solitarious. The tendency to 0 or 1 of our models' results is due to the way the P_{greg} value is calculated from the MLR result, which makes the too low regression results give near-zeros and the too high regression results give near-ones —a -∞ regression result will give 0 as P_{greg} and a +∞ will give 1. We thus needed to homogeneously soften the extreme results of the MLR of our models (bring the MLR values down in a parallel way). We opted for a correction strategy based on the equation that calculates P_{greg} from the value given by the MLR formula. In that equation, eta (η) is assumed to be multiplied by 1 in both terms of the fraction. We therefore changed the value of this constant by the inverse of a series of whole-number (x) multiples of the number of variables in the equation (k). Based on that series of simulations we choose the constant, correction factor ($c = 1/xk$), that moved the solitarious average P_{greg} values from 0 and the gregarious ones from 1 without crossing the thresholds of 20% and 80%, respectively. Our models were hence corrected and their quantitative nature improved. To keep just one model we discard the *Sg_extended* model (for being binary) as well as the *Sg_low-redundancy* model (for having a higher AIC), and we suggest using the *Sg_extended_corrected* model.

The *Sg_extended_corrected* is a good tool that works quantitatively for estimating the gregariousness level of *S. gregaria* locusts —nymphs and adults. However, that model includes morphometric data and might be better used for testing different samples or the same locusts if they molt between testings. The morphometric data would be unnecessary and might even introduce noise if one tests the same, morphologically unchanged, locusts at different time points (i.e., if they do not molt between tests). To deal with that and provide a model that better fits testing the same locusts before and after a treatment, we built a new non-morphometric model. The resulting *Sg_non-morphometric* model needed no correction, worked both for solitarious and gregarious nymphs, and gave quantitative results. We therefore recommend it for testing the same nymphs between different time points —if they do not molt in between. We hoped that the normalization of the morphometric data by size (see above) would be enough to attenuate the clear size differences between *S. gregaria* adults and nymphs and in such a way that our models would work both for nymphs and adults. However, only the *Sg_non-morphometric* model was able to accurately estimate the phase state of solitarious and gregarious adult samples. We attribute that to differences between nymphs and adults that even the normalization by femur length cannot remove. These might include potential attenuation of the morphometric differences between solitarious and gregarious adults compared to nymphs, which might make the morphometric data unnecessary for adults, as well as physiological differences between adults and nymphs which might make the normalization by size insufficient for correcting speed-related variables.

The locusts outbreak problem is not exclusive to *S. gregaria*, neither is the research on the phenomenon. Given that gregariousness have similar characteristics between locusts (morphological changes, increased activity and tendency to aggregate), that our models use morphology, speed-related and tendency to aggregate parameters, and that the variables used for the MLRs are normalized by the size of the animal, we dared to test whether our *S. gregaria* models can also be useful for other pest locusts. The answer was a clear 'no'. The reason seems to lay on the fact that both species have such different relative shape proportions, velocity and walking patterns (stops, turns) so that the MLR coefficients had to be recalculated. Accordingly, we built, corrected and tested behavioural models for *L. migratoria* in a similar way as we did for *S. gregaria*. We built two models for that species, one with the morphometric variables and the other without. Both models worked well on the animals that we used for building them. However, they did not perform satisfactorily when tested on two independent samples. One might think that this could be due to the use of "wrong" (sick, old or altered) individuals for building the models. For that we filtered the sample by removing all the gregarious locusts whose P_{greg} was bellow 80% and rebuilt the models. Still, this was ineffective. We do not usually rear *L. migratoria* and the samples that we used were only for the purpose of this work. That means that we do not know the history nor do we have the same confidence in the degree of gregariousness of any of the *L. migratoria* samples as we do in the case of *S. gregaria* (a species on which we do work). We know that the *L. migratoria* solitarious sample that we used for model building is for sure less gregarious than the gregarious one but we cannot confirm, as we did with *S. gregaria*, that it is extreme solitarious, neither can we ensure that the gregarious sample was extreme gregarious. This is why we consider the models that we built for *L. migratoria* as invalid. The reason why we are discussing them here is because we want to highlight and alert about the fact that using fully characterized solitarious and gregarious samples for model building is a must if the model is to be valid. Equally important is testing the model on known independent samples, that are different from the ones that were used for building it, before assuming its validity —in fact the models that we built for *L. migratoria* do work for the samples on which they were built but not for the independent ones. In the case of *S. gregaria*, we knew the history of the samples, we had morphological and colorimetric indicators of their state and we confirmed our data by comparing the relative expression of three genes. The models that we built thus performed well when we tested them on several independent samples and are thus to be trusted. The case of our attempt at model building with *L. migratoria*, if anything, provides a strong support to the reason of being of the current work. We unrecommend building a model and using it straightaway, as one might have misleading results as we had for the independent (test) *L. migratoria* samples, a species for which valid models are still to be provided by the experts.

To conclude, here we provide two working and tested quantitative behavioural models to use in experimentation with both adults and nymphs of *S. gregaria* —one of the two main pest locusts and paradigm species for most the locust-related research. We suggest using the *Sg_extended_corrected* model (that includes morphometric variables) for comparing different *S. gregaria* nymph samples. Whereas, for testing adults or the same nymphs at different time points (if they do no molt), we suggest using the *Sg_non-morphometric* model (that does not include morphometric variables). We also encourage the experts to build similar models for *L. migratoria*, a species on which we do not normally work and, consequently, our attempt at building models for it failed. Tested on a total of 189 *S. gregaria* individuals, the models that we provide here are thus useful for comparing different populations as well as for testing the effects of drugs and gene and other manipulations on the tendency of *S. gregaria* locusts to be gregarious. We offer them with the hope that they would save time and effort, and that they will provide a much needed opportunity for standardizing and homogenizing methodologies and tools for the *S. gregaria* locust research community, to the benefit of less lab- or experiment-specificity and higher universality and reliability of the research results.

1.5 Supplementary material

Loadings	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10	PC 11	PC 12	PC 13	Abs. sum
Pronotum-Femur index	-0,159	-0,280	-0,281	0,591	-0,041	0,107	-0,017	0,064	-0,052	-0,073	0,032	0,665	0,007	2,368
Pronotum-Head index	-0,055	0,206	0,077	0,539	-0,636	-0,342	0,046	0,031	-0,032	-0,028	-0,004	-0,364	-0,003	2,361
Femur-Head index	0,100	0,373	0,331	-0,324	-0,280	-0,356	0,026	0,128	0,054	-0,109	0,071	0,627	0,004	2,786
Choice	0,337	-0,313	-0,103	-0,067	-0,088	-0,140	0,680	0,274	0,096	0,224	-0,386	0,024	0,001	2,733
Elapsed Time	0,221	-0,423	-0,164	-0,205	-0,267	-0,051	0,163	-0,188	0,063	-0,440	0,604	-0,070	0,003	2,863
Total Distance	-0,429	-0,168	-0,094	-0,144	0,019	-0,371	0,053	-0,306	0,061	0,081	-0,116	0,018	0,707	2,567
Average Speed	-0,429	-0,169	-0,095	-0,140	0,018	-0,369	0,057	-0,307	0,077	0,088	-0,104	0,031	-0,707	2,592
Average Acceleration	0,410	-0,094	0,136	0,242	0,169	-0,214	-0,258	-0,254	0,666	0,293	0,099	0,042	0,010	2,886
Last Coordinate	-0,100	-0,262	0,567	0,044	-0,123	0,202	0,158	-0,166	-0,303	0,539	0,322	0,066	0,003	2,854
Choice by time	-0,068	-0,305	0,545	0,044	-0,112	0,239	-0,052	-0,157	0,187	-0,498	-0,472	-0,003	-0,004	2,687
Stop Ratio	-0,346	-0,247	0,234	0,028	0,212	-0,209	-0,072	0,721	0,244	-0,063	0,259	-0,137	0,002	2,776
Turn Ratio	-0,335	0,178	-0,172	-0,169	-0,391	0,511	0,139	0,043	0,568	0,186	0,070	0,026	0,009	2,798
Erratic Movement	0,131	-0,379	-0,169	-0,278	-0,427	-0,016	-0,619	0,205	-0,135	0,247	-0,216	0,014	-0,004	2,840
Cumulative variance	0.325	0,571	0,729	0,832	0,906	0,943	0,966	0,982	0,992	0,996	0,999	1,000	1,000	–

Table 1.9: Principal Components Analysis result. Cumulative variances per principal component are shown in the last row and variable loadings for each principal component are shown in rows below. Note that the cumulative variance reaches 0.7 at the 3rd principal component. Abs. sum is the sum of the absolute values of the 13 principal components' loadings for each variable.

	Pronotum Femur index	Pronotum Head index	Femur Head index	Choice	Elapsed Time	Total Distance	Average Speed	Average Acceleration	Last Coordinate	Choice by time	Stop Ratio	Turn Ratio	Erratic Movement
Pronotum-Femur index		0.054	0.000	0.849	0.657	0.174	0.004	0.003	0.662	0.007	0.411	0.587	0.223
Pronotum-Head index	0.239		0.027	0.896	0.033	0.007	0.519	0.529	0.758	0.347	0.751	0.100	0.045
Femur-Head index	**-0.849**	0.271		0.987	0.085	0.004	0.006	0.006	0.666	0.016	0.790	0.672	0.025
Choice	0.024	0.016	0.002		0.907	0.459	0.157	0.158	0.000	0.000	0.896	0.285	0.566
Elapsed Time	0.056	-0.262	-0.214	0.015		0.000	0.001	0.001	0.594	0.025	0.000	0.000	0.000
Total Distance	0.169	-0.329	-0.353	0.093	**0.828**		0.483	0.488	0.143	0.237	0.003	0.003	0.000
Average Speed	0.354	-0.081	-0.332	0.176	-0.388	-0.088		0.000	0.264	0.000	0.000	0.000	0.838
Average Acceleration	0.358	-0.079	-0.336	0.176	-0.387	-0.087	**1.000**		0.265	0.000	0.000	0.000	0.843
Last Coordinate	0.055	-0.039	-0.054	**0.933**	0.067	0.182	0.139	0.139		0.000	0.353	0.163	0.178
Choice by time	0.330	-0.118	-0.295	0.544	-0.276	-0.148	**0.704**	**0.704**	0.538		0.001	0.227	0.797
Stop Ratio	-0.103	-0.040	0.033	0.016	0.570	0.359	**-0.705**	**-0.703**	0.116	-0.406		0.000	0.191
Turn Ratio	0.068	0.204	0.053	-0.134	-0.566	-0.355	0.483	0.484	-0.174	0.151	-0.814		0.218
Erratic Movement	0.152	-0.248	-0.276	0.072	0.554	**0.829**	0.026	0.025	0.168	-0.032	0.163	-0.154	

Table 1.10: Pearson's correlation coefficients (lower half of the matrix) and p-values (upper half of the matrix) of the pair-wise correlations between the 13 variables used for building the behavioural models for estimating *S. gregaria* probability of being gregarious. In bold are the strong correlations coefficients of at least 0.7.

	Nymphs				Adults			
	Gregarious		Solitarious		Gregarious		Solitarious	
	Chi-squared	p-value	Chi-squared	p-value	Chi-squared	p-value	Chi-squared	p-value
Pronotum-Femur index	0.999	0.318	0.000	1.000	1.119	0.290	0.098	0.754
Pronotum-Head index	1.189	0.276	0.121	0.729	0.035	0.852	3.153	0.076
Femur-Head index	1.239	0.266	0.214	0.643	0.413	0.520	1.844	0.175
Elapsed time	0.052	0.820	0.482	0.488	0.527	0.468	0.011	0.917
Total distance	0.025	0.874	0.054	0.817	0.124	0.724	0.098	0.754
Average speed	0.019	0.892	0.054	0.817	0.155	0.694	1.320	0.251
Average Acceleration	0.013	0.910	0.054	0.817	0.035	0.852	0.011	0.917
Stop Ratio	0.785	0.376	0.121	0.729	0.073	0.788	0.273	0.602
Last Coordinate	0.766	0.381	0.165	0.685	0.007	0.934	1.098	0.295
Choice by time	1.859	0.173	0.013	0.908	0.028	0.863	1.844	0.175
Turn Ratio	0.706	0.401	0.086	0.770	0.052	0.820	1.484	0.223
Eratic Movement	1.092	0.296	1.736	0.188	0.527	0.468	0.798	0.372
Choice*	1.201	0.273	0.184	0.668	0.041	0.830	0.024	0.876

Table 1.11: Comparison of 13 morphometric and behavioural variables between male and female solitarious and gregarious *S. gregaria* adults and nymphs. The table shows the results of Kruskal-Wallis test with 1 degree of freedom and its p-value. No significant sex-dependent differences were found for any of these variables.

	Sg_low-red	Sg_min	Lm_ext_fil_cor	Lm_no-mor_fil_cor
Intercept (β_0)	$-2.448\ 10^{1}$	$-3.446\ 10^{4}$	$1.666\ 10^{7}$	$-2.210\ 10^{6}$
Pronotum-Femur index	—	$3.763\ 10^{4}$	$-3.069\ 10^{7}$	—
Pronotum-Head index	$1.494\ 10^{1}$	—	$1.082\ 10^{7}$	—
Femur-Head index	2.765	$9.527\ 10^{3}$	$-5.993\ 10^{6}$	—
Choice	2.465	$4.833\ 10^{3}$	$-7.541\ 10^{4}$	$-7.931\ 10^{5}$
Elapsed time	—	$-1.919\ 10^{-2}$	-2.369	-4.261
Total distance	$-6.287\ 10^{-3}$	$-1.222\ 10^{1}$	$1.018\ 10^{3}$	$1.527\ 10^{3}$
Average speed	—	$1.407\ 10^{6}$	$8.380\ 10^{7}$	$-1.014\ 10^{8}$
Stop ratio	-5.307	$-1.434\ 10^{4}$	$5.054\ 10^{5}$	$2.874\ 10^{6}$
Last coordinate	—	-9.450	$2.140\ 10^{2}$	$3.209\ 10^{3}$
Choice/time	—	$4.192\ 10^{4}$	$-2.213\ 10^{6}$	$-6.448\ 10^{6}$
Turn ratio	—	$-9.096\ 10^{3}$	$2.127\ 10^{5}$	$2.655\ 10^{6}$
Average acceleration	—	$-9.966\ 10^{7}$	$-5.358\ 10^{9}$	$7.193\ 10^{9}$
Erratic movement	—	$3.179\ 10^{-1}$	$-2.600\ 10^{1}$	$-3.353\ 10^{1}$

Table 1.12: Formulae from the discarded models. The first column shows the variable names whereas the following columns show wether the variable was not included in the current pointed in as column title (in which case is marked with three dashes) or wether it was included (in which case the variable's regression coefficient value is shown). Sg_low-red = *Sg_low-redundancy*; Sg_min = *Sg_minimum*; Lm_ext_fil_cor = *Lm_extended_filtered_sample_corrected*; Lm_no-mor_fil_cor = *Lm_non-morphometric_filtered_sample_corrected*.

Chapter 2: *S. gregaria*'s central nervous shows an association between gregarious phase and gene expression

2.1 Introduction

While the search for key molecules for the development of the gregarious phase is on its way—and the current work pretends to be part of it, another aspect relating to the molecular basis of the phase change is to establish the dynamics of gene interactions and ultimately the cascade of molecular events that accompany the shift from solitarious to gregarious. Obtaining the large amount of necessary data for inferring such a big picture is now possible thanks to the available high-troughput technologies (e.g., Schuster [2007], Metzker [2010], Mardis [2013]). Since locust phase polyphenism occurs due to changes in gene expression rather than mutations, RNA-seq is one way to analyze it [Bakkali, 2013]. As by the date of the present work, Illumina seems the best choice due to its price-quality-quantity. In fact, Illumina-based RNA-Seq analysis of locust transcriptomes has already been carried out for *L. migratoria* [Kang et al., 2004, Chen et al., 2010, Zhang et al., 2012, Wang et al., 2014b]. Surprisingly, for the main pest locust species, *S. gregaria*, the only available transcriptomics work to date [Badisco et al., 2011a] is Sanger-sequencing based. No doubt, that was a gigantesque work (see the comparative section of the current manuscript's results) and, as such, provided a good wealth of sequencing data. Still, it hardly provided quantitative information on differential gene expression between phases. One might think that such information could be inferred from the works on *L. migratoria*, but phase change might present species specific differneces. However, apart from a non-quantitative study [Zhang et al., 2012], there is only one RNAseq report on *L.*

migratoria's CNS [Wang et al., 2014b]. In addition, the molecular aspects of the phase change seem to show differences between species (see the examples above). We give such importance to the analysis of gene expression in the CNS due to its obvious importance for a phenomenon so tightly linked to perception of the environment and changes in behaviour.

The works carried out on the molecular basis of locusts' phase change have no doubt highlighted an extensive list of genes involved in functions such as neurotransmition, hormone reception, ion channels, neuronal structure and development and circadian rythm. Here we present the first RNA-Seq work on the main pest locust. We assemble, annotate and quantitatively compare the transcriptomes of CNS-enriched libraries from solitarious and gregarious locusts. After validation, via qPCRs, we compare our data to those obtained in previous works on locusts in general. We thus not only produce sequences, both annotated and with no known annotation, but also lists of genes whose expression is significantly different between phases and, based on our comparisons, we highlight those that seem important for the phenomenon. Based on our data and knowledge on both gene functions and the locust phase change phenomenon, we infer a "big picture" of the association between the cascade of events that accompany the phase change and the observed changes in gene expression.

2.2 Material and methods

The locust rearing, RNA isolation, RNA sequencing and RNA assembly protocols were performed as described in the general methodology section of this thesis. We downloaded, assembled using CAP3 and its defaults options, and annotated the sequences obtained by Badisco et al. [2011a] (Accession numbers: JG662739.1 to JG697409.1). In order to comparatively assess the efficiency of our transcriptome assembly, local BLAST databases were separately built using our reference transcriptome and the sequences assembled from Badisco et al. [2011a]'s ESTs. Reciprocal local BLASTn searches were then carried out in order to compare the completeness of each of these sequence sets. A local database was built using the available amino acids sequences of *Acyrthosiphon pisum*, *Anopheles gambiae*, *Apis mellifera*, *Bombyx mori*, *Drosophila melanogaster*, *Nasonia vitripennis*, *Pediulus humanus* and *Tribolium castaneum*. Local BLASTx searches against that amino-acids database allowed an initial annotation of our reference transcriptome contigs. BLAST2GO [Conesa et al., 2005, Conesa and Götz, 2008] was then used to further BLAST the sequences that gave no BLAST results against the NCBI nr database. After that, we carried out BLASTn searches against the NCBI nt database, using BLAST2GO and the sequences that had no BLAST result or were homologous to prokaryotic, fungi or plant proteins. Sequences homologous to non-animal proteins or non-arthropodan nucleotides were considered as contaminants. 10^{-6} was the E-value cut-off in all BLAST searches. Similar BLAST searches were carried out to annotate the assembled sequences from Badisco et al. [2011a]'s ESTs —which are not annotated in the NCBI database. BLAST2GO was also used for functional annotation against the KEGG database and establishment of the GO terms and enzymatic maps. We compared our results on individual genes to those reported in Kang et al. [2004], Badisco et al. [2011b], Guo et al. [2011] and, as further tests, we used qPCRs to validate the RNAseq results. 14 genes were chosen for qPCR testing, eight of them (*troponin C*, *larval cuticle protein 9* and *asparagine synthetase* from Kang et al. [2004], *heat shock protein 20* and *peptidyl-prolyl cis-trans isomerase* from Guo et al. [2011] and *RNA helicase*, *glia maturation factor* and *peroxiredoxin* from Badisco et al. [2011b]) were selected because our RNAseq results on them did not agree with the published results. The remaining set of six genes were selected from our reference transcriptome based on the differential expression they showed in our RNAseq data and their potential relevance, but regardless of their level of expression or magnitude of the differential expression. These were *chemosensory protein 12*, *G-protein coupled receptor Mth2-like*, *tyrosine aminotransferase-like*, *phosphoenolpyruvate carboxykinase*, *yellow-h* and Sg_CNS_NA202—the latter being a sequence with no known annotation. Primer sequences of these genes are detailed in table 2 from the general methodology section of this thesis. qPCRs were carried out both on the same RNA samples as well as on samples from other locusts that also were at different developmental stages (four gregarious and four solitarious 4th instar *S. gregaria* nymphs). The primers of the sequences troponin C and RNA helicase from our transcriptome didn't work. We used the actin and tubulin sequences from Van Hiel et al. [2009] as reference for calculating relative quantities in the qPCR experiment. cDNA preparation and qPCR testing were as described in Cabrero et al. [2013]. This way, our RNAseq data were compared to the literature and tested using qPCR on the same and on different materials.

Literature-check of the data on the function of the differentially expressed genes allowed us to classify them and interpret their reason of being over-expressed in the gregarious locusts. Merging all the interpretation results allowed us to suggest a big picture on the cascade of events that accompany gregariousness and their interrelations with the differentially expressed genes. We also compared our results to those reported on differences in gene expression between locust phases in six peer-reviewed scientific articles [Kang et al., 2004, Chen et al., 2010, Badisco et al., 2011a,b, Guo et al., 2011, Wang et al., 2014b]. We thus gathered a list of genes that we grouped based on the consistency of their differential expression between works.

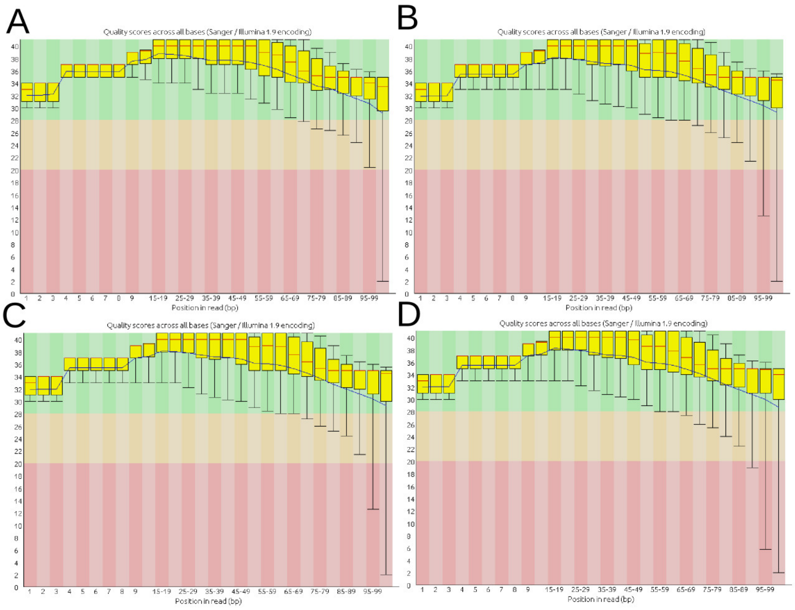

Figure 2.19: Distribution of the mean sequencing quality values along the 101 positions of the Illumina Hiseq2000 Paired End forward (a and c) and reverse (b and d) sequencing reads of the CNS-enriched tissue from solitarious (a and b) and gregarious (c and d) *S. gregaria* locusts. The positions of the sequencing reads are in the x-axis and the y-axis shows the quality values. The Q30 value is defined as less than 1 in 1,000 probability of error.

2.3 Results

2.3.1 The sequencing

From table 2.13 we can see how the sequencing library of the CNS-enriched tissues from solitarious locusts (henceforth solitarious library) yielded around 25% more sequencing reads than those obtained for the library of the CNS-enriched tissues from the gregarious locusts (henceforth gregarious library). It is to note that both libraries showed similar GC contents and that about 90% of their nucleotides do satisfy the Q30 quality threshold. Furthermore, the positional distribution of the nucleotide quality scores looks homogeneous along the 101 positions of the sequencing reads (figure 2.19), especially the first 75 positions. In accordance with these good sequencing results, the percentage of unidentified nucleotides (N) was less than 0.005%.

Library	Total bases	Total reads	%GC	%Q30	%N
Solitarious	9,621,251,112	95,259,912	42.98	89.52	< 0.005
Gregarious	7,701,076,684	76,248,284	41.92	90.06	< 0.005

Table 2.13: Depth and quality of the Illumina Hiseq 2000 Paired Ends sequencing results obtained for our CNS-enriched sequencing libraries from solitarious and gregarious *S. gregaria* locusts.

2.3.2 The assembly and annotation

After assembly, we prudentially considered all the contigs that were shorter than 75 bp and/or were assembled from less than 4 sequencing reads as potential sequencing or assembly artifacts. Their removal, as first filter, left 117,309 contigs (table 2.14). Around 51% of these had significant BLAST hits to protein sequences in our local insect protein database, around 5% had significant BLAST hits to sequences in the NCBI nr database, around 9% to sequences of the NCBI nt database, and around 30% corresponded to contigs that had no significant BLAST hit to any sequence in any of these databases. The remaining 5% of the assembled contigs were prudentially considered as potential contaminants as they corresponded to sequences that had significant BLAST hits to sequences of either non-animal proteins (168 contigs) or non-arthropodan nucleotides (6,049 contigs). The 76,396 contigs of our assembly that had significant and acceptable BLAST hit to sequences in any of the three databases used here corresponded to a total of 17,620 unigenes. The remaining 34,696 contigs had no significant BLAST hit to sequences in any of the databases used. The assembled-, size-, coverage- and BLAST-filtered contigs that we retained as reference CNS transcriptome for our work on *S. gregaria* CNS comparative transcriptomics can be found in the supplementary file Sg_CNS_NGS.fasta. Of the *Acyrthosiphon pisum*, *Anopheles gambiae*, *Apis mellifera*, *Bombyx mori*, *Drosophila melanogaster*, *Nasonia vitripennis*, *Pediculus humanus* and *Tribolium castaneum* proteins that formed our local insect BLAST protein database, the contigs of the reference *S. gregaria* CNS transcriptome showed higher similarity to the beetle *Tribolium castaneum* proteins (figure 2.20A). The remaining contigs that had acceptable BLAST results against the NCBI nr database highlighted similarity to the ant Harpegnathos saltator (figure 2.20B) and, naturally, the sequences of the locusts *S. gregaria* (2,215) and *L. migratoria* (590) were notorious among the results of the BLAST searches against the NCBI nt database. As figure 2.21A, B and more extensively table S4 show, categorization of the annotated contigs by GO process showed that we covered a deep portion of the processes (1,043 processes distributed among 16 GO levels), with the usual predominance of ribosome biogenesis, homeostatic, synthesis, metabolic (including nucleotide, DNA, RNA and protein synthesis and modification), energy, structure-related, cell differentiation and cell organization processes.

Local database of Insect proteins		NCBI nr		NCBI nt		No BLAST result	Total (with BLAST result)	
Contigs	Unigens	Contigs	Unigens	Contigs	Unigens	N/A	Contigs	Unigens
59,513	14,700	6,243	1,794	10,313	1,126	34,696	76,396	17,620

Table 2.14: Number of contigs and unigens that had significant BLAST hits to sequences in the three databases used in the current work.

Relevant to this work, however, we had a notorious presence of processes related to:

- biological regulation (877 sequences at level 2, including 526 sequences for gene expression at level 5)
- signaling and cell communication (777 sequences at levels 2 and 4, including 749 sequences for signal transduction at level 5, 14 of them classified as belonging to the G-protein coupled receptor signaling pathway at level 7)
- response to stress (218 sequences at level 3)
- response to stimuli (943 sequences at level 2, including response to cellular, biotic, aboitic, endogenous and external stimuli at higher GO levels)
- 99 vesicle-mediated transport sequences at level 5
- 372 trans-membrane transport sequences at level 6, and
- 36 neurological system process sequences at level.

2.3.3 Comparative analysis

To test the quality and depth of our NGS sequencing and assembly, we compared our results to the *S. gregaria* 34,672 Sanger sequenced ESTs obtained by Badisco et al. [2011a] from *S. gregaria* CNS (accession numbers JG662739.1-JG697409.1). Our assembly of these raw ESTs gave 13,079 sequences of between 200 and 4,954 bases corresponding to 4,714 contigs and 8,365 singlets (a conservative assembly result but similar to the 4,785 unigenes and 7,924 singlets reported in Badisco et al. [2011a]). Both assemblies of our NGS-obtained reference transcriptome (see supplementary file Sg_CNS_NGS.fasta) and the sequences we assembled from the ESTs by Badisco et al. [2011a] (see supplementary file Sg_CNS_EST.fasta) show similar GC content (39.53% and 40.31% respectively). With N50 values of 1,296 vs. 827, our assembly of the NGS transcriptome (75 and 35,064 as minimum and maximum contig sizes, respectively) seems to have given

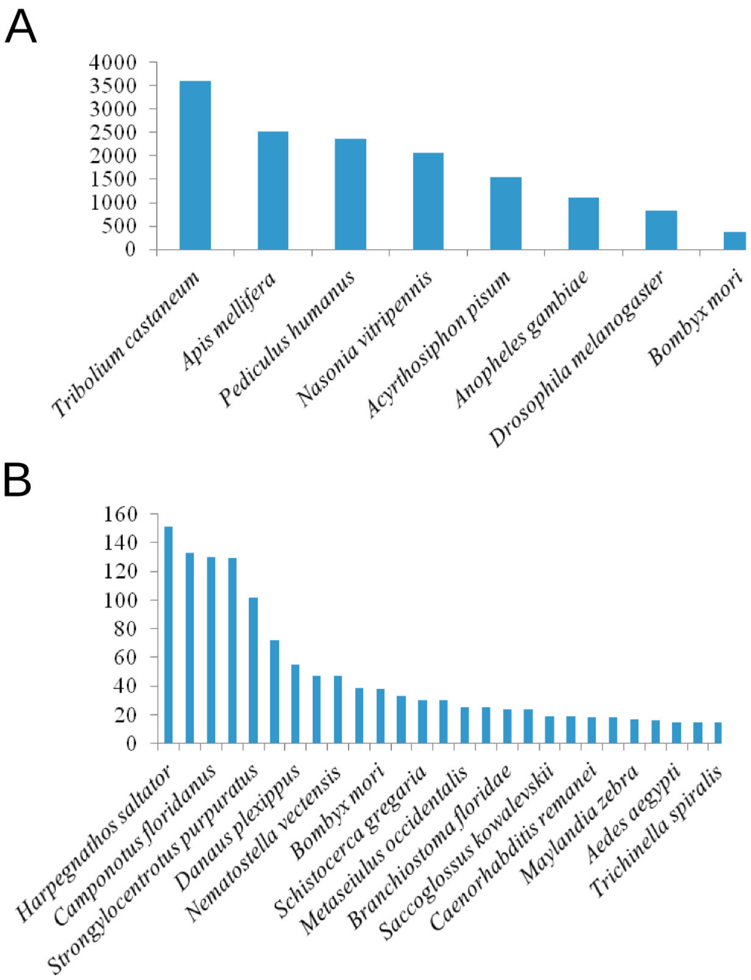

Figure 2.20: Distribution of the species against which the sequences of pooled non-redundant reference transcriptome of the solitarious and gregarious *S. gregaria* CNS-enriched tissues gave a significant top BLASTx result when annotated using our local insect proteins database (a) then using the NCBI nr database (b).

results at least as good in contig size as the Sanger-based sequences—with the differences in homogeneity of the sequence lengths that one would expect from the differences between the raw results of the two sequencing technologies in question.

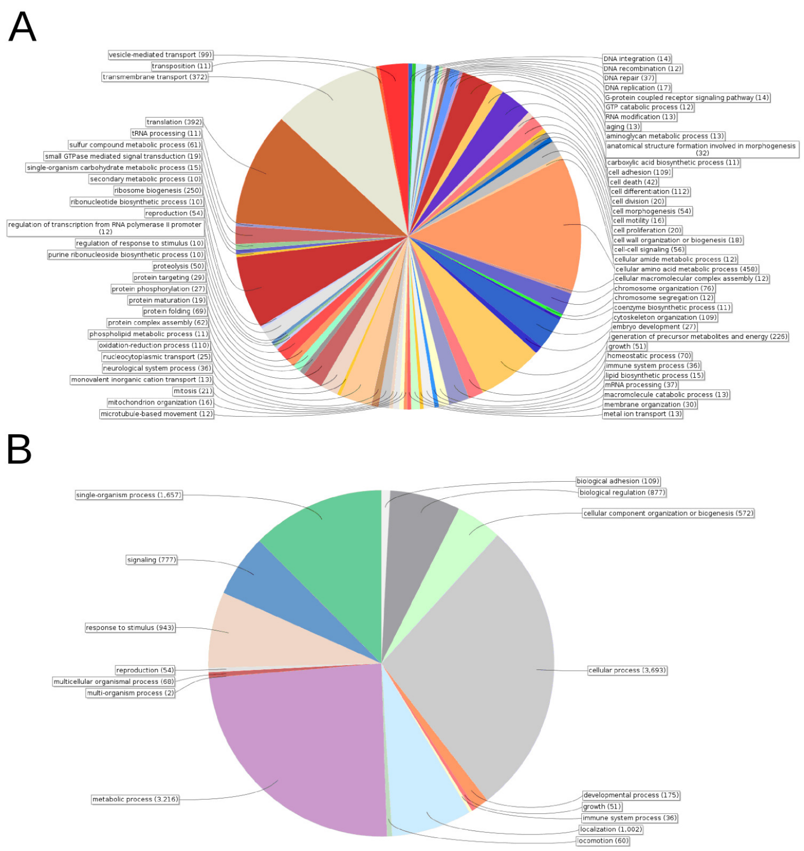

Figure 2.21: Distribution of the number of sequences in each of the GO terms in the functionally annotated part of the reference transcriptome of the CNS-enriched tissues of the gregarious (a) and solitarious (b) *S. gregaria* adults (see also table 2.16 and supplementary table S4). Online link in `https://drive.google.com/open?id=0B5K51Iajv0_jX180WXAzc1JNaDg`.

7,977 of the assembled Sanger ESTs from Badisco et al. [2011a] had significant BLAST results, while 5,102 had no significant hit neither to sequences of the NCBI nr nor to sequences of the NCBI nt databases (see supplementary file Sg_CNS_EST.fasta and table S5). 29,666 contigs of the NGS assembled transcriptome (see supplementary file Sg_CNS_NGS.fasta) matched 7,719 sequences assembled from the Sanger-obtained ESTs from Badisco et al. [2011a], 5,780 of which had significant BLAST hits while the remaining 1,939 had not (table S6). As a whole, 23,497 of the 29,666 NGS assembled contigs that matched assembled Sanger ESTs had significant

BLAST hits to sequences either in the NCBI nr or in the NCBI nt databases while the remaining 6,169 contigs had no significant BLAST hit (table S7). The remaining 81,098 contigs of our NGS assembled transcriptome were not present in the assembled Sanger ESTs from Badisco et al. [2011a]. 52,572 of these had significant BLAST hits to 13,057 unigenes either in the NCBI nr or in the NCBI nt databases while the remaining 28,526 contigs had no acceptable BLAST hit to sequences of any of those two databases (table S8). Of the 5,360 assembled Sanger ESTs from Badisco et al. [2011a] that our NGS sequencing project missed, about half (2,556) had significant BLAST hits to sequences either in the NCBI nr or in the NCBI nt databases and the remaining 2,804 had no BLAST hit (table S9). Among the 51,448 entries (unigenes and non-annotated contigs) that have at least 4 reads in our NGS-based reference transcriptome of the *S. gregaria* CNS, some 48% (24,770) were significantly over-expressed in one locust phase or the other. About 10% of these (2,537) were significantly over-expressed in the solitarious sample (table S10) while about 90% (22,233) of the unigenes and non-annotated contigs were significantly over-expressed in the gregarious sample (table S11)—the remaining 26,678 entries had no significant differences in expression level between solitarious and gregarious locusts (table S12).

As a first overall assessment of these results, we looked for genes from our NGS assembled transcriptome that were specifically tested or mentioned in the literature as significantly over-expressed either in solitarious or gregarious locusts, we used either qPCR, microarrays, or NGS-based works that were published up to the year 2014 both on *S. gregaria* or *L. migratoria.* Of the 86 genes analyzed, 28 had significant over-expression in the opposite phase in our results compared to the literature, 22 had no significant differential expression in our data, compared to the significant difference reported elsewhere (9 of them showing over-expression in the same phase and 13 in the opposite one), and 36 genes were confirmed by our results as significantly over-expressed either in the solitarious (4 genes) or gregarious (32 genes) phase. Our data therefore support the published results in 42% of the cases, could not do so in 26% of the cases and oppose them in 32% of the cases (being 56% of the cases in agreement with the data reported in the examined literature, if we do not consider the significance of the over-expression) (table S13). Interestingly, our data show a similar agreement rate with the data published on *L. migratoria* (29 agreements vs. 22 opposed, 56 % agreement) and the data published on *S. gregaria* (7 agreements vs. 6 opposed, 53 % agreement). As table S13 shows we have noticeably higher agreement with Chen et al. [2010], who used Illumina-based RNAseq technology for detecting differences in gene expression between solitarious and gregarious whole *L. migratoria* nymph and adult bodies. Unfortunately the lack of overlap between the genes highlighted in the different works impeded us from cross-testing the agreement between these works (details in table 2.17). The genes whose significantly different expression levels between solitarious and gregarious locusts we confirm here are in table 2.17. A further test of our results consisted on qPCR testing of 14 genes. Six of them were chosen from our reference transcriptome based on their potential (they had to have significant differential expression but did not have necessarily to be among the most differentially expressed—we avoided genes with the highest differences). The remaining eight genes were chosen among those genes whose published data are either unsupported or contradicted by our RNAseq data, and for which working qPCR primers were available. We first used the same mRNA samples that we previously used for the RNAseq. The primers for two genes, *troponin C* and *RNA helicase* (see table 2.15), did not work. As figure 2.22 and table 2.15 show, the tendencies of the qPCR results were in agreement with those of our RNAseq results in 11 out of the 12 cases (the *peroxiredoxin* case being the only exception, while *HSP20* shows clearly lower expression differences by qPCR than by RNAseq). As additional test we also used RNAs from animals that were different from the ones that we used for RNAseq (nymphs this time). In this case the qPCRs of eight of the 12 tested genes showed over-expression in the same phase as revealed by the RNAseq data (figure 2.22 and table 2.15) and the remaining four genes (*peroxiredoxin*, *G-protein coupled receptor Mth2-like*, *tyrosine aminotransferase-like* and *peptidyl-prolyl cis-trans isomerase*) showed opposing direction of over-expression between these qPCRs and the RNAseq. The qPCRs on the sequenced cDNA show that the results for five out of the six genes that we chose from the literature were, as it can be seen in figure 2.22, were in agreement with our RNAseq data and not with the expression differences reported elsewhere. Only one of these five genes shows opposite result to our RNAseq data when the qPCRs were carried out on other cDNAs. Overall, eight of the 12 tested genes showed consistent direction of over-expression both in our RNAseq data as well as in our two sets of qPCR data on the same cDNAs and on cDNAs from other locusts. These were: *chemosensory protein 12*, *yellow-h*, *larval cuticle protein 9*, *asparagine synthetase*, *heat shock protein 20*, *phosphoenolpyruvate carboxykinase*, *glia maturation factor* and Sg_CNS_NA202. Table 2.17 lists the genes that we so far confirm are associated with the locusts phase change. It is interesting to note that the in silico gene expression results were concordant between the solitarious and gregarious libraries and show similar sets of most expressed genes in each of these libraries. In numbers, near 90% (4,618) of the 10% most expressed genes in each of the two libraries (5,145) were the same. The remaining 10% (527 genes) of the 10% most expressed genes in each library are either absent from or do not belong to the 10% most expressed genes in the other library (table S15). Even more, as a whole, the expression levels of these 90% shared genes among the 10% most expressed genes in both the solitarious and the gregarious libraries were statistically similar (t = 1.882, p = 0.06). However, when it comes to the least expressed genes, only less than 7% (349) of the 10% least expressed genes (5145) were shared between both libraries (table S15). Even more, the expression levels of these 349 shared

Conting in our assembly	BLAST Description	Accession	Reported in	ID in qPCR
88095	chemosensory protein 12	ABH88185	Our work	CSP12
102208	PREDICTED: G-protein coupled receptor Mth2-like	XP_003250433	Our work	Mth2
47160	tyrosine aminotransferase-like	XP_001603572	Our work	TAT
34202	phosphoenolpyruvate carboxykinase, isoform B	NP_725802	Our work	PEPCK
49891	yellow-h	ABB81847	Our work	Yellow H
Sg_CNS_NA4Plus202	—	—	Our work	NA202
29324	lcp9_drome ame: full=larval cuticle protein 9 ame: full=larval cuticle protein ix flags: precursor	P82384	Kang et al. [2004]	LCP9
85647	asparagine synthetase, putative	EEB18196	Kang et al. [2004]	ASNS
Sg_CNS_NA4Plus277	locusta migratoria heat shock protein 20.6 mRNA, complete cds	DQ355964	Guo et al. [2011]	HSP20
33921	peptidyl-prolyl cis-trans isomerase-like	XP_001602615	Guo et al. [2011]	PPI
103230	glia maturation factor	EEB15105	Badisco et al. [2011b]	GMF
81383	PREDICTED: similar to peroxiredoxin	XP_968419	Badisco et al. [2011b]	Peroxiredoxin
Sg_CNS_NA4Plus4221	PREDICTED: *Nasonia vitripennis* troponin C (TpnC), mRNA *Nasonia vitripennis* troponin c mrna	XM_003424121	Kang et al. [2004]	Troponin C*
86113	PREDICTED: ATP-dependent RNA helicase p62-like	XP_001604593	Guo et al. [2011]	RNA helicase*

Table 2.15: List of genes from *S. gregaria* CNS transcriptome for qPCR validation. Primers from genes marked with an asterisk did not work.

genes among the 10% least expressed genes in each library were significantly different (t = 8.971, p < 0.0001). Even when we consider only those sequences that have at least 10 reads in one or other library (40,573 cases), the number of shared genes among the least expressed 10% (4,057) of the total was barely above 15% (618) and the expression levels of these genes were significantly different between libraries too (t = -8.184, p < 0.0001). Accordingly, figure 2.23 shows how the correspondence between the rankings of the genes based on their expression levels in each library weakens as the expression level goes down.

When looking at the distribution of the GO processes where the differentially expressed sequences fall, figure 2.21 shows how the set of annotated contigs that were over-expressed in the solitarious sample is dominated by translation, catabolism and metabolic processes. Whereas, in addition to these processes, the set of annotated contigs which expression was significantly higher in the gregarious sample shows a notorious presence of processes relating to protein modification, ion transport, energy as well as behaviour and the response to the different stimuli and to stress (a full list of genes and their relative expression values is in table S10 and S11). A further analysis shows how 26 of the 114 GO terms have significantly different number of sequences in the set of sequences over-expressed either in the solitarious or in the gregarious locusts. Of these 26 GO terms, 14 were significantly over-represented in the set of genes that were significantly over-expressed in

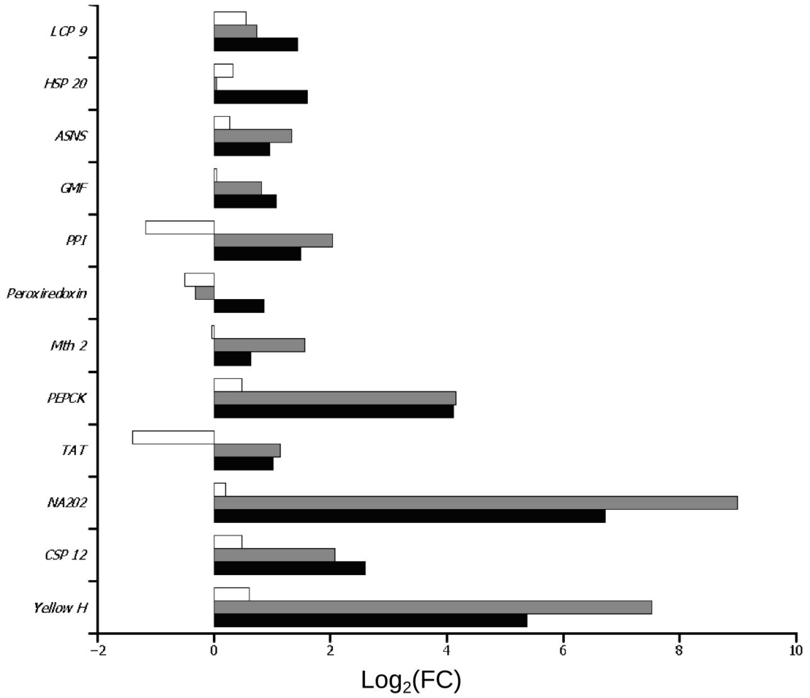

Figure 2.22: Direction and magnitude of the differential expression for the 12 genes tested by RNAseq (black bars), qPCRs on the cDNA used for RNAseq (grey bars) and qPCRs on cDNAs from other *S. gregaria* locusts of different nature (heads only) and developmental stage (nymphs). The y-axis shows the tested genes (*LCP9*: larval cuticular protein 9, *HSP20*: heat shock protein 20, *ASNS*: asparagine synthase, *GMF*: glia maturation factor, *PPI*: cis-trans peptidyl-prolyl isomerase, *Mth2*: methuselah 2-like, *PEPCK*: phosphoenolpyruvate carboxykinase, *TAT*: tyrosine aminotransferase-like, *NA202*: a transcript with no known annotation, *CSP12*: Chemosensory protein 12). The x-axis shows the 2-based logarithm of the ratio between the expression level in gregarious and the expression level in solitarious (positive values thus reflect over-expression in the gregarious cDNA and negative values are due to over-expression in the solitarious one). Primers are described in the general methodology section (table 2)

the solitarious locusts. They mainly corresponded to growth, metabolism and biosynthesis processes (table 2.16). From their part, the remaining 12 GO terms that were significantly over-represented in the set of genes that were significantly over-expressed in the gregarious locusts corresponded mainly to molecule modification, response to stimuli and cell communication and signaling (table 2.16). The shared genes among the 10% most expressed genes in each library corresponded mainly to ribosome biogenesis processes (figure 2.24A). Whereas, the small set of genes that belong to the set of the 10% most expressed genes in each library but were not shared to both libraries, showed noticeable presence of transmembrane transport, translation and carbohydrate metabolic processes in the gregarious phase (figure 2.24C) and ribosome biogenesis and cellular aminoacid metabolic processes in the solitarious library (figure 2.24E). On the other hand, the small set of shared genes among the 10% least expressed genes in each library that had at least 10 reads in one or other library, showed noticeable presence of biosynthetic processes (figure 2.24B). The ones that, being among the 10% least expressed genes in each library, that had at least 10 reads in a library and are absent from the other library included noticeable presence of proteolysis, cell differentiation cell adhesion and signal transduction in the gregarious sample (figure 2.24D) and transmembrane transport, translation and nucleobase-containing compound catabolic processes in the solitarious sample (figure 2.24F). Highlighting all the differentially expressed genes is not possible given the so many genes that seem differentially expressed as by a 5% statistical significance cutoff. Furthermore, the comparatively reduced set of sequences with significant over-expression in the solitarious phase contained sequences with little information, as inferred from their annotations. Accordingly, in addition to the sequences with no known annotation, among the sequences over-expressed in the solitarious phase we can highlight the several protease genes as well as those encoding for vitellogenin, vitellogenin receptor, yellow-y, yellow-12, hexamerin, and hemocyanin. As to the set of sequences that show the most pronounced significant increase of expression in the gregarious sample, and in addition to the sequences that have no known annotation, the huge and complex amount of annotated sequences means that the general picture is rather complex. To have a glimpse of it, we opted for a meticulous search for the functions and implications reported in the literature for each of the thousands of annotated genes that are significantly over-expressed in the gregarious phase. We then selected a resumed sample of relevant genes for highlighting based on a classification by functional implication. Table S17 thus classifies some of the most significant genes among the ones significantly over-expressed in the gregarious phase into functional categories (each sheet of the file) and subcategories (different columns in each sheet), and, when necessary, clarifies the reason why we consider such gene important for the phenomenon that we are studying here.

Based on that, we can see how among the annotated genes that are significantly over-expressed in the gregarious phase we had:

I. Genes whose differential expression is very likely proper to other tissues and not to the CNS tissue per se, these include:

   a. Gene related to muscle, such as those encoding for different troponins,

GO level	GO tag	GO term	Total counts	Gregarious	Solitarious	Log(FC)	P-value
Biological process							
Gregarious							
2	GO:0050896	response to stimulus	943	581	9	1,462	0,002
6	GO:0006464	cellular protein modification process	560	373	4	1,993	0,004
5	GO:0036211	protein modification process	560	373	4	1,993	0,004
5	GO:0043412	macromolecule modification	573	373	4	1,993	0,004
4	GO:0007154	cell communication	777	477	7	1,540	0,004
2	GO:0023052	signaling	777	464	7	1,501	0,006
3	GO:0044700	single organism signaling	777	464	7	1,501	0,006
2	GO:0065007	biological regulation	877	774	17	0,959	0,007
5	GO:0007165	signal transduction	749	455	7	1,472	0,007
4	GO:0050794	regulation of cellular process	808	455	7	1,472	0,007
3	GO:0051716	cellular response to stimulus	787	455	7	1,472	0,007
3	GO:0050789	regulation of biological process	812	762	17	0,936	0,008
Solitarious							
3	GO:0016049	cell growth	51	5	4	-4,228	0,000
6	GO:0006412	translation	392	204	24	-1,463	0,000
5	GO:0009059	macromolecule biosynthetic process	473	204	24	-1,463	0,000
5	GO:0034645	cellular macromolecule biosynthetic process	473	204	24	-1,463	0,000
4	GO:0044249	cellular biosynthetic process	523	204	24	-1,463	0,000
4	GO:1901576	organic substance biosynthetic process	525	204	24	-1,463	0,000
5	GO:0010467	gene expression	526	210	24	-1,421	0,000
3	GO:0009058	biosynthetic process	1287	596	49	-0,946	0,000
2	GO:0008152	metabolic process	3216	2143	116	-0,343	0,012
2	GO:0040007	growth	51	22	4	-2,091	0,016
4	GO:0006091	generation of precursor metabolites and energy	226	93	9	-1,181	0,030
4	GO:0019538	protein metabolic process	1071	801	47	-0,459	0,039
4	GO:0043170	macromolecule metabolic process	1470	986	56	-0,412	0,042

Table 2.16: GO terms over-represented in gregarious and solitarious phase CNS.

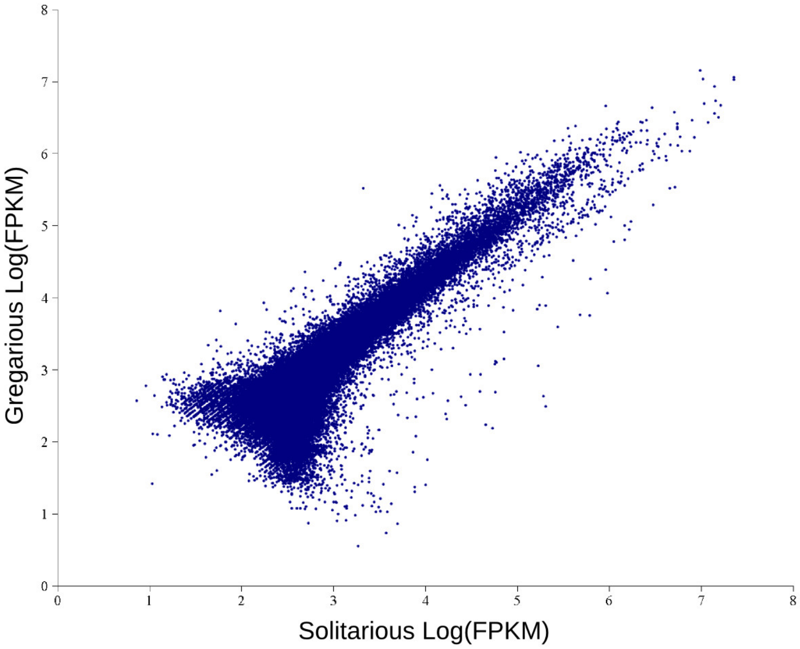

Figure 2.23: Correlation of the rankings by expression of each sequence in the solitarious and gregarious transcriptomes of the CNS-enriched *S. gregaria* tissues. The rankings were calculated as the positions of the sequence in each library after sorting the sequences by expression level in each library (the least expressed being first and the most expressed last). The two rankings of each sequence (one for each library) were log-transformed for better representation in the figure. The narrower is the cloud of dots around the 45 degree axis the more similar are the rankings of each of a set of sequences are in the two libraries.

tropomyosin, myophilin, myosins, zintin, twitching, wings-up, zeelin and colmedin—the latter being involved in maintaining a structural microenvironment that allows efficient neuromuscular signaling at the synapses.

b. Genes related to the cuticle, such as those coding for different cuticle proteins, chitinase and laccase—the latter being involved in cuticle tanning.

c. Genes related to the haemolymph, such as the one coding for hemolymph lipopolysaccharide-binding protein.

d. Genes related to the Malpighi tubes, such as the one coding for inverted formin.

e. Genes related to the fatty tissue, such as those coding for acyl-CoA oxidase, sterol carrier protein x, lipases, apolipoproteins, apolipophorins and adipokinetic hormone—the latter being reported as associated with gregariousness [Ayali and Pener, 1992].

II. Genes related to general and metabolic processes of active cells including neural ones:

a. Genes related to sugar metabolism, including glucose transporters, glucolysis (such as triosephosphate isomerase and enolase), gluconeogenesis (phosphoenolpyruvate carboxykinase and serine dehydratase), glucose homeostasis (GSK-3-binding protein) and metabolism (such as D-arabinose 1-dehydrogenase).

b. Genes related to mitochondrial function, such as mitochondrial ribosomal proteins, NADH dehydrogenase, Thioredoxin-dependent peroxidase, cytochrome c oxidase, succinyl-coa synthetase, peroxiredoxin-6, lipoyltransferase 1 and adenine nucleotide translocase.

c. Genes related to different metabolic processes, such as ATP synthase and adenylate cyclase (general), and to lipid (e.g., very long-chain-fatty-acid–CoA ligase bubblegum and serine palmitoyltransferase), sugar (e.g., N-acetylglucosamine-6-phosphate deacetylase), transport (e.g., aquaporin, Abc transporter), CNS-related processes (e.g., glutamate synthase and acetylcholinesterase).

d. Genes related to detoxification, including a large list of cytochrom P450 isoforms as well as other genes implicated in detoxification processes such as: carboxylesterase-6, glutathione s-transferase M2, multidrug resistance-associated protein 4 and microsomal glutathione S-transferase-like.

e. Genes related to cell death, including autophagy proteins as well as genes such as croquemort, apoptosis-inducing factor 3, caspase-8, Bax inhibitor-1 and alsin—the last two being involved in protection against ER stress-induced apoptosis and in motor neuron degeneration, respectively [Otomo et al., 2003, Bultynck et al., 2012].

f. Genes related to defense and immunity, including different pacifastins, serine protease inhibitors, toll-related and melanization-related proteins, proteins related to pathogen recognition (e.g., peptidoglycan recognition protein) and genes that link the immune response to other processes such as endotoxin tolerance, response to stress, cytokins and free radicals (e.g., immune responsive gene 1 (IRG1) and nuclear factor related to kappa-B-binding protein).

g. Genes related to other processes known to be affected by the change of phase, such as those related to genome instability (including recombination and transposition).

III. Genes related to processes at the interface between the environment and the intra-cellular molecular responses, such as those involved in:

   a. Sensorial functions, such as Odorant binding and chemosensory proteins, proteins related to mechanosensation (e.g., no-mechanoreceptor potential A, hemicentin, serine proteinase stubble and groucho) and genes related to the visual system (e.g., class A rhodopsin-like G-protein coupled receptor GPRnna14, opsin blue sensitive, eye-specific protein kinase C, unc119 and rhodopsin-specific isozyme).
   b. Signaling functions, such as G-protein coupled receptor Mth2, geranylgeranyl transferase, GTPase-Activating Proteins, G protein and G-protein signaling modulator.
   c. Hormone-related functions, such as the prothoracicostatic peptide, juvenile hormone-inducible protein and steroid receptor seven-up.
   d. Response to stress, such as heat shock proteins 60, 70 and 75, apolipoprotein D and Methuselah.

IV. Genes that are more directly related to the neural cells and their functioning including:

   a. Genes which alteration relates to neural-related conditions in humans, such as Down syndrome cell adhesion molecule-like protein CG42256-like, AMME syndrome candidate gene 1 protein, fragile X mental retardation syndrome-related protein 1 and huntingtin-interacting protein 1.
   b. Genes related to ventralization and neural development and/or maintenance, such as hunchback, dorsal, enhancer of split, spaetzle, vrille, saxophone protein, survival motor neuron protein, spindle F and short gastrulation SOG.
   c. Genes related to neuronal function, such as synaptic vesicular amine transporter, nicotinic acetylcholine receptor, sodium-potassium-chloride cotransporter, scribble and roadblock.
   d. Genes related to the regulation of the neuronal function, such as macoilin-1, septin 4, nose resistant to fluoxetine and neuropeptide receptor A27,
   e. Genes related to neuronal plasticity, such as slit, fasciclins 1, 2 and 3, small bristles, longitudinal lacking, semaphorin, as well as microtubule and actin proteins.

V. Genes directly related to behaviour, including ebony, yellow, fruitless, genes of the circadian rhythm (e.g., timeless, period and pigment dispersing factor), and genes related to memory, activity and social behaviour (e.g., visgun, protein takeout, HERC2 and major royal jelly protein).

VI. A large list of genes related to all the steps of gene expression including genes involved in nucleotide synthesis, tRNA synthesis, transcription, cytoplasm-nucleus traffic, RNA editing, aminoacids synthesis, ribosome synthesis, translation, protein editing and vesicle traffic.

To establish the potential link between all these genes that we found significantly over-expressed in the gregarious locusts and between them and what we already know about the biology of gregariousness in general, we drew figure 2.25. In that figure we interpret why and how a subset of genes is associated with gregariousness and we offer an integrative view on the cascade of events that lead to the gregarious phase and how these events affect or are affected by the genes that belong to different biological processes and functions whose levels of transcription we found significantly altered in the library of CNS-enriched tissues from gregarious *S. gregaria* adults. As to the sequences that have no known annotation, all what we can say is that at least some of them must be genuine (due to the large amount of sequencing reads that align to them) and that some must be very significantly linked either to the solitarious or to the gregarious phase. To cite just one example for each phase, a sequence of 608 bases aligned to 133 reads in the solitarious library and to 14,079 reads in the gregarious library. Furthermore, our qPCRs confirmed its presence and differential expression. From the other side, a sequence of 1,849 bases, with no known annotation, aligned to 175,987 and 3,181 reads in the solitarious and gregarious libraries, respectively. In total, of the 34,696 sequences with no known annotation that assembled in our reference transcriptome, 2,059 were significantly over-expressed in the solitarious phase (380 of them have over 10 reads and are absent from the gregarious library, 1,380 had over 10 reads and over 1 fold increase of expression in that phase, and the remaining 299 have either less than 10 reads in the solitarious phase or show less than 1 fold increase of expression in that phase). From the gregarious side, 11,061 of the 34,696 sequences with no known annotation were significantly over-expressed in that phase (only 28 of them have over 10 reads and are absent from the solitarious library, 5,940 have over 10 reads in the gregarious library and over 1 fold increase of expression in that phase, and the remaining 5,093 have either less than 10 reads in the gregarious phase or less than 1 fold increase of expression in the same phase).

2.4 Discussion

In the present work we quantitatively determine, for the first time, the global changes in gene expression levels that differentiate between solitarious and gregarious *S. gregaria* locusts. For that, both the use of adequate material and its adequate handling are crucial. To reach our objective it was suitable to use individuals of both extremes of such polyphenism (i.e., fully solitarious

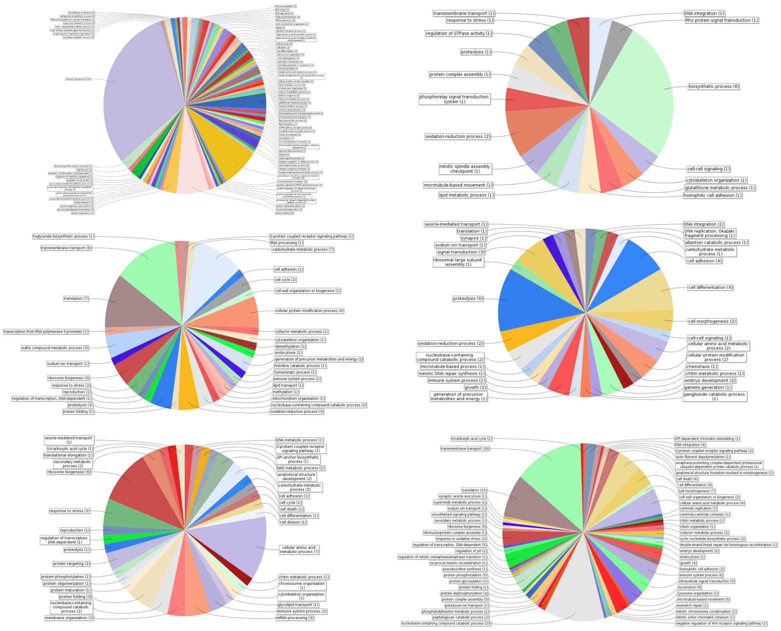

Figure 2.24: Distribution of the number of sequences in each of the GO terms in the functionally annotated part of the several fractions of the reference transcriptome of the CNS-enriched tissues (sequencing libraries) of the gregarious and solitarious *S. gregaria* adults. A: Fraction of the transcriptome containing the shared part of the 10% most expressed genes in each library, B: Fraction of the transcriptome containing the shared part of the 10% least expressed genes in each library, C: Fraction of the transcriptome containing the part of genes present only in the 10% most expressed genes in the gregarious library, D: Fraction of the transcriptome containing the part genes present only in the 10% least expressed genes in the gregarious library, E: Fraction of the transcriptome containing the part of genes present only in the 10% most expressed genes in the solitarious library, F: Fraction of the transcriptome containing the part genes present only in the 10% least expressed genes in the solitarious library. Online link in `https://drive.google.com/open?id=0B5K51IajvO_jQ2RkejdOR1NUbXc`.

and fully gregarious). So were the locusts that we used here, as they comply with the following criteria: (i) the solitarious locusts were reared in individual cages in complete mechanical, visual and olfactive isolation from other locusts, and the gregarious locusts were reared in crowded conditions. (ii) The solitarious and gregarious locusts were kept separated, at their respective rearing conditions, for three generations in order to erase any homogenizing maternal effect. (iii) The solitarious and gregarious locusts used showed the typical overall differences including nymph color differences (see figure 3 in the Introduction from this thesis), adult size and activity differences—as evidenced by the multivariate logistic regression modeling results in Chapter 1. Furthermore, since we are not interested in the inter-population differences in the genetic control of the locust polyphenism, both our fully solitarious and fully gregarious colonies came from the same locust stock. Neither are we interested in the inter-individual (intra-population) differences in the genetic control of the locust polyphenism and, for that reason, we used a pool of five individuals for each locust phase. There should be sex-dependent differences in the genetic control of the locust polyphenism [Wybrandt and Andersen, 2001, Rahman et al., 2008, De Loof et al., 2010], which are not the subject of the current work, for that the pooled sample contained 3 males and 2 females (the size of adult females are about double that of the adult males—see Chapter 1). Using nymphs implies having to make sure that all the animals are of the same developmental stage and deciding whether to use one of the six solitarious instars or five gregarious instars stages or a carefully proportioned pool of shared instars. To avoid uncertainties from such possible comparisons we used sexually mature adults (i.e., no developmental differences between samples). With these precautions taken, the differences that we should detect would be mainly due to general differences between solitarious and gregarious locusts.

However, the results might still be altered due to inadequate handling of the material. Solitarious locusts should, especially, be handled with the utmost gentleness and speed in order to minimize stressing them. Of course the gregarious animals should be treated the same exact way as their solitarious counterparts. The anesthesia method is relevant here, so we opted for cold anesthesiation in order to avoid the effect on the CNS of synthetic chemical volatiles. Even more, when it comes to the tissue itself, we opted for quick slashing of the central part of the ventral side and the head of each anesthetized locust followed by immediate immersion in liquid nitrogen in order to avoid altering the tissues and the gene expression of their cells by direct and lengthy mechanical handling and extraction of the CNS ganglia and their immersion in buffered solutions—as it was the standard way in previous works. We therefore recon that we used calculated procedures to get one of the most unaltered materials from appropriately proportioned samples of the right locusts. We also performed a standard RNA extraction in ten replicates for each sample (to minimize errors from the RNA extraction step) before pooling the resulting

RNAs for sequencing at a high-depth at one of the world-leading sequencing companies (at Macrogen Inc. Korea). The results should therefore represent the reality. Accordingly, the over 75 million Illumina Hiseq2000 Paired End 100-base reads that we obtained for each library not only showed good quality indicators (about 90% of their nucleotides have less than 10^{-3} probability of being erroneous and a fraction of them lesser than 5 10^{-5} being undetermined), but also contained similar G+C content in the solitarious and gregarious libraries—implying little contamination too (the 1% difference in the G+C content between the solitarious and gregarious libraries could be expected from libraries of the same species that contain different proportions of the different transcripts). Wang et al. [2011] estimated that 30 million 75-base sequencing reads (2.25 10^9 bases) are enough for detecting all the annotated genes in a transcriptome. It is thus worth highlighting that our sequencing produced over 3 times that number of bases for each transcriptome (over 7 times in total). Thus, the sequencing step also went well both in terms of coverage, quality and lack of evidence of differential contaminants between libraries. Another issue, this time relating to the handling of the RNAseq reads, is how to assemble the reads into contigs. We chose ABySS and its complementing software TransABySS as they are both well established [Birol et al., 2009, Simpson et al., 2009, Martin and Wang, 2011] and require affordable computer processing and memory power. As to the k-mer lengths, instead of arbitrarily choosing one or using the median length of the assembled sequences as a less subjective criterion, we opted for a complete multi-K-mer strategy. We thus assembled, merged and filtered the 33 assemblies of the odd number k-mers from the lax 19 to the very strict 95 bases. Being this a transcriptome assembly, we used only odd numbered k-mers in order to avoid reverse complement matches and the possibility of inverting a de Bruijn graph in case of palindromic k-mers [Zerbino and Birney, 2008, Miller et al., 2010]. As to the k-mer-range itself it had as a maximum the largest odd number k-mer that a 64-bit-based computer processor can handle (95) and the smallest odd number k-mer that would not produce excessive mis-assembly (the probability of having a particular sequence of 19 bases in a set of equi-base sequences is of 3.638 10^{-12} so, given one particular 19 base sequence, the probability of it happening again by chance in the set of 8.2 10^9 sliding 19-base windows and their reverse complements from a 100 million reads library is of less than 6%).

Illumina sequencing reads typically start losing quality by the 75% of their length [Dohm et al., 2008], so any contig that is shorter than 75 bases is most likely produced by less than one full length read. We therefore eliminated those contigs from the assembly. As to the remaining contigs, they were kept if they had significant BLAST hits (E-value 10^{-6}) in the NCBI nr or nt databases or were produced by the assembly of at least 4 sequencing reads (i.e., they come from at least 2 sequenced molecules). In terms of sizes, the N50 indicator of our transcriptome assembly was even higher than that of the Sanger-sequenced transcriptome in Badisco et al. [2011a] even thought we included in our reference transcriptome sequences as short as 75 bases long (which lower the N50 and increase the deviation from the mean).

We could confirm the expression of most of the sequences contained in Badisco et al. [2011a]'s Sanger-sequenced transcriptome and we provide a good wealth of additional sequences. Furthermore, the fact that only few sequences of our transcriptome had significant BLAST results against non-animal proteins or non-arthropodan nucleotides means that our libraries were not significantly contaminated (as also suggested by the similar G+C content of the sequencing libraries). The distribution of the BLAST results further confirmed that our libraries mainly contained arthropodan RNAs (as evidenced by the higher proportion of insect BLAST-positive sequences). Furthermore, the fact that the annotated sequences showed a noticeable presence of CNS-linked molecules implies that indeed our sequencing libraries were enriched in CNS tissue. We therefore have a well assembled, large and valid transcriptome from good quality Illumina Hiseq 2000 Paired End reads obtained by sequencing uncontaminated RNA libraries from CNS-enriched tissues that were extracted in an as less altering as possible way from confirmed fully solitarious and fully gregarious non-sex-biased and low-inter-population and inter-individual noise pooled samples of locusts. Accordingly, 12 of the 14 qPCR attempts were successful, meaning that at most only two of the 14 tested sequences (less than 15%) might be misassembled, the remaining 12 (over 85%) were indeed genuine and correctly assembled and annotated (at least in the part between the primers).

In addition to the fact that the assembled contigs included sequences as short as 75 bases and as little expressed as those having only four mapped reads, the fact that our filtered reference transcriptome contained over 111 thousands contigs implies that, as it is always the case in RNAseq assemblies, it contains partial sequences and isoforms of the same genes. Further proof of that are the same BLAST hits that different contigs have (the 76 thousand BLAST positive sequences correspond to 17,620 unigenes) and the BLAST similarities of different sequences to different isoforms of the same gene. It is interesting to note that about a third of the assembled sequences had no known BLAST-similar sequences in the NCBI's nr or nt databases as by 2013. We cannot discard that some of these sequences might be sequencing artifacts (especially those with few mapped reads) neither can we discard that others might be genuine (especially those with high number of mapped reads with and without differential expression between locust phases). Some of these sequences probably belong to the realm of lncRNAs that confer diversity and dynamism to the transcriptomes (e.g., see Brown et al. [2014]).

Another interesting result is the large number of unigenes and contigs with no known annotation that appear differentially expressed between locust phases (58% of the total). The fact that 11 out of 12 qPCR testings showed the same expression tendency as the one shown by the RNAseq when tested on the same cDNAs means that, not only the assembly and annotation were correctly done, but also that the sequencing did not alter the molecular proportions of the cDNA and that the in silico determination of the expression levels was correct. Further confirmation of that is the fact that the qPCRs gave similar tendencies for eight out of the 12 tested genes even when the cDNAs were from other individuals that moreover belonged to different developmental stages (nymphs) and tissue composition (heads only). Even more interesting is the fact that the qPCRs both on the same cDNA and on different cDNAs confirmed the expression tendencies suggested by our RNAseq data for four out of six genes whose expression tendencies were not congruent between our RNAseq data and the data reported elsewhere. Certainly, we did not test our RNAseq data on clear-cut cases but we rather included the "risky" cases of genes whose data do not support the data published elsewhere. With RNAseq and two qPCR experiments "from our part", the fact that our data do not support the published data for four genes does not necessarily mean error in other peoples' work. The source of variation might be at any level from species-specificity (some of these genes were tested in other species) to inter-individual variation.

The datum on the number of differentially expressed unigenes and contigs with no known annotation is therefore trustable. It is indeed surprising but not unexpectable; given the fact that the differences between solitarious and gregarious locusts affect almost every aspect of the locust biology: external morphometry [Stower et al., 1960, Symmons, 1969, Maeno et al., 2004, Hamouda et al., 2011], size of internal organs [Ott and Rogers, 2010], color [Stower, 1959, Ellis, 1964, Sword and Simpson, 2000, Lester et al., 2005], behaviour [Gillett, 1973, Roessingh et al., 1993, Bouaichi et al., 1995, Heifetz et al., 1996, Roessingh et al., 1998, Simpson et al., 1999], diet [Sword and Simpson, 2000, Simpson et al., 2002, Lester et al., 2005], metabolism [Ma et al., 2011, Wu et al., 2012], physiology [Injeyan and Tobe, 1981, Ayali and Pener, 1992, Wiesel et al., 1996, Tanaka and Yagi, 1997, Anstey et al., 2009], reproduction [Islam et al., 1994, Simpson et al., 1999, Ferenz and Seidelmann, 2003] or development [Islam et al., 1994, McCaffery et al., 1998, Hägele et al., 2000]. Of course, being the phase change of the locusts not due to changes in the genome per se, all these differences must be due to differences in gene expression. Not surprising neither is the fact that about 90% of the differentially expressed unigenes and sequences with no known annotation have increased expression in the gregarious phase. Gregarious locusts are more active and exposed to a more stimulating and challenging environment than the lethargic solitarious locusts that live in a less difficult environment.

As a whole, the reference transcriptome that we assembled is enriched in processes that are tightly linked to CNS functions and that should be of interest to the current work. In fact, the over 1,950 sequences that belong to the signaling and cell communication, response to stress, response to stimuli and neurological system processes should contain important sequences to the phenomenon analyzed here. As to the differentially expressed unigenes that have functional annotation, we had a clear bias towards metabolic, catabolic and translation processes among the ones over-expressed in the solitarious locusts and a noticeable presence of processes relating to the response to different stimuli, stress and behavior among the sequences over-expressed in the gregarious locusts. What's more, the processes that are significantly over-represented in the set of sequences that are over-expressed in the solitarious phase mainly correspond to growth, metabolism and biosynthesis processes—processes expected to be overrepresented in organisms that live in not so challenging conditions (availability of food, little competition, little infections...). On the other hand, the processes that are significantly over-represented among the set of sequences that are over-expressed in the gregarious locusts mainly correspond to molecule modification, response to stimuli and cell communication and signaling—processes expected to be over-represented in organisms facing infections, high levels of interactions and competition, and crucial need for processing the increased and amplified inputs that originate from such conditions. The data therefore reflect the fact that gregarious locusts are more exposed to stimuli, live in stressful conditions and have altered behaviours compared the solitarious locusts.

Another interesting yet not surprising datum is that most of the most expressed genes are shared between libraries and their expression levels are largely similar between the solitarious and gregarious locusts, whereas most of the least expressed genes are not. It is indeed expected that the set of most expressed genes should be enriched with constitutive and housekeeping functions and, as such, would be mainly common and similarly expressed between phases, whereas the set of least expressed genes could be enriched by state-specific functions. Accordingly, the set of the 10% most expressed genes is enriched with ribosome biogenesis processes whereas the disperse set of least expressed genes had a noticeable presence of the vaguely described biosynthetic processes.

The main aim of the current work was not only to identify and list the set of genes that are differentially expressed between solitarious and gregarious locusts, but to reach a more a global understanding of the overall differences at the molecular level between these two phases of the locusts. The large number of sequences that our analyses revealed as differentially expressed between locust phases (42%) appears at first surprising if not potentially exaggerated. Yet, the phenomenon analyzed here is not a simple explosion in the population size of a living being. We are dealing with deeply rooted adaptations to recurrent drastic changes in the living conditions. Indeed the locust has to adapt to changes that potentially affect to every aspect of its biology. Be it for warning or camouflage the color of the locust changes between phases. Crowdedness does not only result in higher chemical, mechanical and visual stimuli, it also leads to more stress, infections, competition and movement. The differences in the availability of resources lead to differences at the metabolic, physiological, fat content, and body size levels, among other aspects. Facing the unavoidable exhaustion of the food sources lays behind the changes at the developmental and structural levels in order to develop wings and muscles earlier and, thus, be able to move and migrate in search of new resources when needed. Even the increased perception of stimuli requires restructuring of the nervous ramifications and termini as part of larger changes in the CNS (such as the relatively larger brain size of gregarious locusts [Ott and Rogers, 2010]). All these changes require adaptations and changes in form of differences in the expression of the genes involved in a plethora of functions—including those involved in the regulation of the metabolism and in gene expression itself. The result in terms of number of genes that we found as differentially expressed is therefore logical. In fact, there are even more pronounced cases; such as the over 85% of the 5,500 genes found by Whitfield et al. to be differentially expressed between nurse and forager honey bees [Whitfield et al., 2006]. Another striking aspect is the biased balance between the number of genes that are differentially expressed in one state or the other. Is it logical to expect more genes to be differentially expressed in the gregarious phase than in the solitarious one? Yes, the gregarious locusts are expected to have higher levels of sensorial functions, metabolism, detoxification, DNA and cell damage and apoptosis, neuronal activity and remodeling, active immune system, even gene expression. The remaining striking datum is the large number of sequences that have no known annotations. Given the state of the art and the number of assembled sequences, some of these are no doubt sequencing or assembly artifacts. However, the number of sequences with no known annotation in our assembled transcriptome appears reasonable when we take into account that the transcriptome of an organism with a genome as small as *D. melanogaster*'s (less than 2% the size of *S. gregaria*'s) has recently been discovered to produce a diverse number of transcript splice variants [Brown et al., 2014] and to contain thousands of unannotated transcripts [Graveley et al., 2011, Hoskins et al., 2011]—one of these works reported 111,195 new elements of the *D. melanogaster*'s developmental transcriptome.

The question now is how to fit all these gene expression differences within the framework of what we know about the events and changes that differentially affect the environment and the locust itself between phases. Figure 2.25 summarizes how we fit the differentially expressed genes, that we already categorized in table S17, into the different steps that form the cascade of events initially triggered by the increase in the population density due to the environmental changes that ultimately lead to outbreaks. Unusual survival of the locust eggs and hoppers, due to favourable environmental conditions, unavoidably leads to increased population size. The resulting crowdedness leads to increased visual, olfactive and mechanical stimulation of the locusts which, in turn, leads not only to increased activity of the genes related to those functions, but also affects the genes involved in remodeling the neural structures (genes for pathfinding, fasciclines, actins. . .) in order to support and enhance pereception of the stimuli. The increased population density also results in increased contagion rate and, thus, increased expression of immune response genes. The depletion of resources and the inter- and intra-specific competition induce stress and even more changes in the neuronal structures and activities (remodeling, signalling, hormone production...), which affect the expression of behavior-related genes. The latter would mean more activity and, thus, changes in the muscles, the fat-bodies and the metabolism which, in turn, increases the toxic metabolites inside the cells and, thus, activity of the detoxification, repair of DNA damage and cell death genes. Both the hormonal, immune and detoxification functions promote the production of melanin, which changes the animal's color. All these processes would not only need changes in every step of the gene expression regulation but also increased expression of the genes involved in these steps too—so genes related to gene expression are affected too. The genes whose expression is increased in the gregarious locusts could thus easily be categorized and fitted into a coherent and logical cascade of events that is highly concordant with the changes that we know locusts undergo when they shift from solitarious to gregarious. While the increase in the expression of some of these genes is already known (e.g., JH [Injeyan and Tobe, 1981, Wiesel et al., 1996, Tawfik et al., 2006], circadian genes [Gaten et al., 2012], chemosensory genes (Guo et al. [2011] and Chapter 5 of this thesis), the increase in the expression of some categories of genes might seem surprising. Still, even the over-expression of genes related to the cytoskeleton and cell structure (actins for instance) is well understood within the context of neural remodeling in response to increased stimulation. Remodeling of the CNS would also explain the increase in the expression of the developmental genes (most of them involved in ventralization). We also think that the increase in the expression of the DNA repair and recombination genes is well understood in a high intracellular toxicity context (due to the increased metabolism) and is in accordance with the observation that gregarious locusts show more recombination than the solitarious ones [Fox, 1973]. All in all, our work reveals for the first time the parallelism between increased behavioural activity and the increment of gene expression. It shows how the gregarious phase affects every aspect of the biology of *S. gregaria* through significant changes of the expression levels of a large number of genes involved in every aspect of the normal functioning and reactions of the cells of a living being. Whether such magnitude of changes is an adaptation that is specific to locusts, due to their recurrent exposure to the episodic selective pressures that crowdedness cause, or it is rather a generalized characteristic, at least in part, that differentiates crowded from non-crowded populations of living beings is an interesting question for future research. The fact that stress, for instance, affects the expression of many genes in humans (e.g., see [Dusek et al., 2008, Nayak et al., 2014]) would suggest that the psychological effects of the level of crowdedness, stimuli and other social interactions might generally affect the expression of as many genes as the ones we show affected in gregarious locusts.

2.5 Supplementary material

Given their large extension, Sg_CNS_NGS.fasta file and supplementary tables S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12 and S13 are uploaded in this link: `https://drive.google.com/open?id=0B5K51Iajv0_jNVR5dGtwM2FaSzg`.

Table 2.17: Genes with confirmed association with locusts phase change. Accession: a sample accession number of a homologous sequence; Phase: the locust phase where the gene is over-expressed; Confirmation: the methods and laboratories that report similar results on the gene; Contig: code of the contig in our transcriptome that correspond to the respective gene.

Confidence	Gene description	Accession	Phase	Confirmation	Tested species	Contig
Agreement between our data and data published elsewhere	hexamerin 5 precursor	NP_001164204	Solitarious	Microarrays in Guo et al. 2011 and our RNAseq	*L. migratoria* and *S. gregaria*	104268
	PREDICTED: protein takeout-like	XP_001950683	Solitarious			37604
	PREDICTED: similar to glutathione-s-transferase theta, gst	XP_975048	Solitarious	Microarrays in Badisco et al. 2011b and our RNAseq	*S. gregaria*	103559
	PREDICTED: arylphorin subunit alpha	XP_001600430	Solitarious			85614
	PREDICTED: similar to adenylate cyclase	XP_975639	Gregarious	qPCR in Kang et al. 2004 and our RNAseq	*L. migratoria* and *S. gregaria*	83877
	PREDICTED: similar to Annexin IX CG5730-PC	XP_967931	Gregarious			85392
	adipokinetic hormone receptor	NP_001161243	Gregarious	RNAseq in Chen et al 2010 and our RNAseq	*L. migratoria* and *S. gregaria*	83922
	PREDICTED: similar to calcium/calmodulin-dependent serine protein kinase membrane-associated guanylate kinase (cask)	XP_972920	Gregarious			86801
	PREDICTED: diuretic hormone receptor-like isoform 2	XP_003427176	Gregarious			92428
	gamma-aminobutyric-acid receptor alpha-2 subunit precursor, putative	EEB18262	Gregarious			102894
	glutamate receptor Gr1	ABD36124	Gregarious			103430

Confidence	Gene description	Accession	Phase	Confirmation	Tested species	Contig
	PREDICTED: protein Malvolio-like isoform 2	XP_003424930	Gregarious			37276
	tyramine/octopamine receptor	NP_001164311	Gregarious			47154
	synaptic vesicle protein, putative	EEB16248	Gregarious			43963
	PREDICTED: ATP-citrate synthase-like isoform 2	XP_003425261	Gregarious	Microarrays in Guo et al. 2011 and our RNAseq		85903
	CSP precursor	AAO16783	Gregarious			108631
	pacifastin-related serine protease inhibitor precursor	CAD11970	Gregarious			77531
	chemosensory protein 2	AAC25400	Gregarious			108608
	Locusta migratoria clone LmigCSP3 hypothetical protein mRNA, complete cds	GU722578	Gregarious			Sg_CNS_NA4Plus8496
	chemosensory protein 4	ABH88177	Gregarious			88200
	Fasciclin-1 precursor, putative	EEB13818	Gregarious	qPCR in Badisco et al. 2011a and our RNAseq	*S. gregaria*	52795
	PREDICTED: slit homolog 2 protein	XP_001603014	Gregarious			42904
	cytochrome P450, putative	EEB11101	Gregarious			91157
	PREDICTED: 10 kDa heat shock protein, mitochondrial-like	XP_001599992	Gregarious	Microarrays in Badisco et al. 2011b and our RNAseq		77390
	PREDICTED: RNA-binding protein Musashi homolog Rbp6-like	XP_001606007	Gregarious			41415
	PREDICTED: unc-112-related protein-like isoform 1	XP_392367	Gregarious	RNAseq in Wang et al. 2014 and our RNAseq	*L. migratoria* and *S. gregaria*	47840
	PREDICTED: basement membrane-specific heparan sulfate proteoglycan core protein-like	XP_393220	Gregarious			86324
	PREDICTED: similar to Switch-associated protein 70 (SWAP-70)	XP_974449	Gregarious			43957
	PREDICTED: E3 ubiquitin-protein ligase hyd-like isoform 1	XP_001605335	Gregarious			93157

Confidence	Gene description	Accession	Phase	Confirmation	Tested species	Contig
	PREDICTED: similar to mitogen-activated protein-binding protein-interacting protein	XP_967919	Gregarious			31687
	PREDICTED: glycyl-tRNA synthetase-like	XP_001606827	Gregarious			103685
	PREDICTED: polyphosphoinositide phosphatase	XP_394455	Gregarious			35564
	PREDICTED: dynein heavy chain, cytoplasmic-like	XP_001951535	Gregarious			93074
	PREDICTED: serine/threonine-protein kinase mTOR isoform 1	XP_625130	Gregarious			42592
	PREDICTED: niemann-Pick C1 protein-like isoform 2	XP_624752	Gregarious			32806
	PREDICTED: vacuolar protein sorting-associated protein 16 homolog	XP_392642	Gregarious			48556
Agreement between two techniques used in the same and different samples in our data and no disagreement with previously published works	CSP12	ABH88185	Gregarious	our RNAseq and qPCRs	*S. gregaria*	88095
	Yellow h	ABB81847	Gregarious			49891
	—	—	Gregarious			Sg_CNS_NA4Plus202
Agreement between two techniques used in the same and different samples in our data but disagreement with previously published works	asparagine synthetase, putative	EEB18196	Gregarious		*S. gregaria*	85647
	Larval cuticle protein 9	P82384	Gregarious			29324
	Locusta migratoria heat shock protein 20.6 mRNA, complete cds	DQ355964	Gregarious			Sg_CNS_NA4Plus277
	Glia maturation factor beta, putative	EEB15105	Gregarious			103230

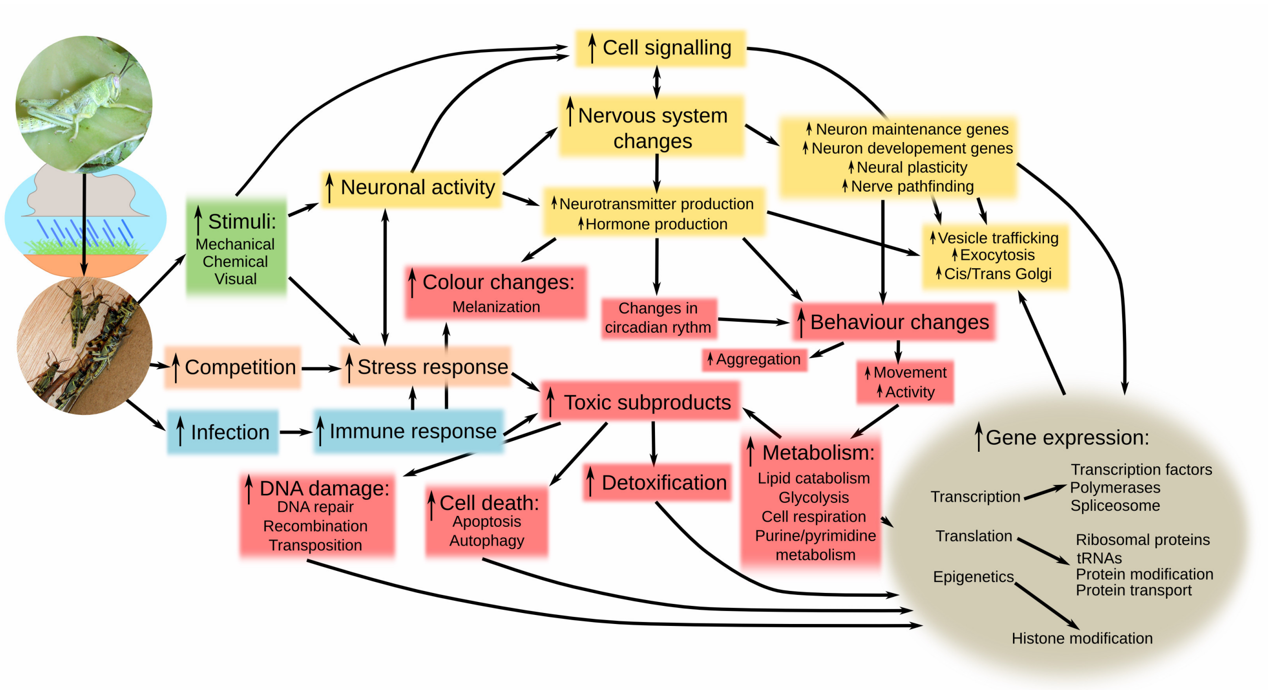

Figure 2.25: Schematic representation of how do genes and events fit and interrelate between each other in the cascade of happenings that leads to gregariousness in *S. gregaria*.

Chapter 3: Pathogens, detoxification and immune response predominate in *S. gregaria*'s gregarius digestive tube

3.1 Introduction

Phase change is tightly linked to locust outbreaks: when the conditions are good for mating, feeding and reproduction, the locust population grows and the solitarious individuals become gregarious members of a plague that consumes all the vegetation in the affected area before swarming to another. This aggregation state has disadvantages not only for the affected area but for the gregarious locusts as well. Gregarious locusts have indeed increased stress responses (see Chapter 2). The high levels of catecholamines in the hemolymph of the migratory locust *Locusta migratoria* [Ma et al., 2011], the involvement of the cardiopeptide corazonin in both *L. migratoria* and the desert locust *Schistocerca gregaria* phase change [Tanaka and Yagi, 1997] and the link with the neurotransmitter serotonin in *S. gregaria* [Anstey et al., 2009], are more evidences that further support the coupling between stress and the gregarious phase of the locusts. One of the reasons behind the increased stress in the gregarious locust population is the increased intra-specific competition. In addition, it is not only difficult to access resources in an extremely dense population, it is also very hard to maintain one's self healthy. Contagion rate rises due to high population density in locust swarms, making the crowd-reared locusts more susceptible to contract diseases, and in consequence they have enhanced immune system activity [Wang et al., 2013, Adamo, 2016]. Moreover, when the resources decline, gregarious locusts not only turn on each other and recur to cannibalism [Bazazi et al., 2008, Guttal et al., 2012], which further enhances contagion, they also have to consume toxic plants that they otherwise wouldn't consume and, thus, have to deal with increased toxicity.

The digestive tube (DT), besides being the gateway for nutrient assimilation, is also subjected and reacts to the general stress that the individual is subjected to and acts as the first zone of interaction with food-bourne pathogens and toxins. It must thus be a barrier against free-entry of pathogens and toxins into the spaces and tissues between the endoderm and ectoderm-borne limits. For this purpose, the DT provides proper defense mechanisms to the individual, such as physical isolation of the food bolus [Barbehenn, 2001, Terra, 2001], enzymatic digestion and pattern recognition proteins inside the digestive tube [Kim et al., 2000, Dziarski, 2004]. In addition, pathogens induce a constitutive immune response in the DT [Engel and Moran, 2013, Adamo, 2016]. Among the actions on which this response relays is the release of reactive oxygen species (ROS) to the intestinal lumen, which helps fight the pathogen, but can damage the host tissues if the exposition to them is prolonged and not countered [Engel and Moran, 2013].

In addition to the immune function, the DT has a detoxification one; as some plants, as well as some microorganisms, produce/contain toxins that have to be processed by the insect in order to avoid poisoning [Bernays, 1982, Bernays and Chapman, 2000, Glendinning, 2002]. Reports on the gregarious *S. gregaria* feeding on a different, more toxic, diet include [COPR, 1982, Bernays and Lewis, 1986]. As a counter measure, *S. gregaria* has developed tolerance to a series of substances used by plants as chemical defenses, such as thioglycosides [Mainguet et al., 2000], glucosinolates [Falk and Gershenzon, 2007] or tannins [Bernays and Chamberlain, 1980]. Even compared to the other important locust species, *Locusta migratoria*, the detoxification ability of *S. gregaria* has a wider range so that the locusts suffer less harmful effects when toxic substances derived from plants were injected in its haemolymph [Cottee et al., 1988]. Furthermore, Sword et al. [2000] show that the ingestion of toxic plants might make locusts more unpallatable to predators, probably due to the very same toxic substances in the DT.

Of course, the presence of bacteria, fungi and protozoa in the DT represents an advantage or a disadvantage depending on the proportions and taxa. Besides pathogens, insects, as the rest of metazoans, also have a symbiotic DT microbiota. Some of these DT microorganisms were studied in the desert locust *S. gregaria*, and an idea on the diversity of the bacterial community that settles in the locust DT was formed [Dillon and Charnley, 2002]. Among the several advantages that a healthy DT microbiota confers are the several anti-fungal substances that prevent the establishment of fungi such as the microsporidium *Metarhizium anisopliae*, an important entomopathogen used in anti-acridian pest control [Dillon and Charnley, 1986, Wilson et al., 2002].

A healthy DT microbiota might also facilitate digestion or even detoxification by degrading molecules that otherwise would not be processed by the locust's digestive machinery alone.

Even more, there are evidences that some DT molecules and bacteria are also involved in the locust phase change. For instance, *Klebsiella pneumoniae*, *Enterobacter casseliflavus* and *Pantoea agglomerans* produce guaiacol from the breakdown of plants material, that the locust uses as precursor to metabolize its aggregation pheromones [Dillon et al., 2000]. By contrast, other microorganisms, such as the microsporidium Paranosema locustae, are reported to prevent gregarization in the migratory locust *L. migratoria* [Shi et al., 2014].

All these DT-level changes in response to stress, pathogens and toxins are predicted to be diverse in *S. gregaria*, because it is a polyphagous grasshopper that faces a wider range of toxic agents in the wide variety of plants that this species consumes, compared to other oligophagous species [COPR, 1982, Bernays and Lewis, 1986]. Furthermore, there must be clear and deep gene expression differences between the DT of the solitarious and gregarious locusts. These changes are expected to be due to the differences in content and also to the differences in the proper DT tissue due to the differences in the actions, responses and interactions to which the solitarious and gregarious locusts' DT are subjected.

In the present NGS-based comparative work, we provide a large set of *de novo* assembled transcript sequences from solitarious and gregarious desert locust's DT. The obtained transcripts were characterized, including functional annotation, and their expression levels were estimated and compared between the gregarious and solitarious phases. Thanks to these analyses, we do not only contribute with a large amount of locust transcripts, many of them putative and newly described, but also with lists of several differentially represented transcripts and biological processes between gregarious and solitarious locusts. The work thus produces and highlights gene sequences that can be used for functional studies or explored as target sequences for fighting locust outbreaks. Interpretation of the data, taking into account the annotations of the locust transcripts and the differences in their expression levels allows us to draw a larger picture on the changes that differentiate the gregarious locusts from the solitarious ones. In addition to the locust data, we provide many bacterial, fungi and protozoan annotated sequences as well as qualitative and quantitative comparative data on the microbial diversity of the locust DT (both solitarious and gregarious).

3.2 Materials and Methods

The locust rearing, RNA isolation, RNA sequencing and RNA assembly protocols were performed as described in the general methodology section of this thesis. The assembled transcriptome was analyzed using BLASTx against the *Drosophila melanogaster* amino acid sequence database. The sequences that did not present a BLAST result were further analyzed using BLASTx against the non redundant (nr) amino acid sequence database, and the sequences that still did not present a BLAST result were analyzed using BLASTn against the non redundant nucleotide sequence (nt) database. The BLAST results were considered as valid at an E-value threeshold of 10^{-6}. Unlike the procedure in Chapter 2, we used a *D. melanogaster* sequence database first in order to avoid the loss of BLAST homologies between contigs, since they could be masked by the same BLAST results from different species. Once we annotated the insect-specific transcripts, we had to use databases of sequences from all kind of species, in order to detect sequences from the DT flora. Term annotation and differential expression analysis were also performed as described in the general methodology section of this thesis.

As mentioned above, we did not empty the contents of the DT in order to avoid excessive manipulation of the tissue, which would alter the gene expression and produce noise in the RNA-seq downstream analyses. As a consequence, we expected the presence of transcripts from symbiotic microbial species in the DT, and we thus could infer their diversity, annotations and transcriptional activity. For that, we counted all the unigenes of the assembled DT transcriptome that had a BLAST result belonging to protist, fungi, bacteria and plants. We did the same with both lists of gregarious and solitarious differentially expressed transcripts, and we elaborated contingency tables for each taxon with the abundance values in the whole transcriptome, the gregarious set of diffferentially expressed transcripts (DETs), and the solitarious set of DETs. Using a χ^2 contingecy test corrected by FDR, we checked for signifficant differences in the representation of each taxonomical group between the gregarious and solitarious phases. We also retrieved all the transcripts and their expression profiles of every non-animal genus that presented more than three DETs. We then compared the mean FPKM values for these transcripts between gregarious and solitarious phases using Student's t test in order to check for significant differences between phases at the transcriptional level. Despite this study being not as accurate for determining the microbial diversity as a metagenomic-focused experiment (i.e. 16S ribosomal amplicon sequencing), it allowed us estimating the transcriptional activity and its potential differences between phases for the main groups of bacterial, protist and fungal species in the locust's DT.

3.3 Results

3.3.1 Sequencing and assembly statistics

RNA sequencing and quality check results are detailed in table 3.18 and figure 3.26. We obtained 32,404,835 101-base reads in the gregarious library and

41,225,365 101-base reads in the solitarious library. The quality of these sequencing reads proved to be high, with the average Phred score per position being higher than 30 in most of the read, with the exception of the 25 last positions, which presented a gradual decline in the phred value. Such decline is expected in the case of Illumina GA output sequences [Dohm et al., 2008]. The percentage of guanine and cytosine was almost identical between libraries, and the proportion of undetermined nucleotidic bases was lower than 10^{-3} %. The combined *de novo* assembly of the two libraries resulted in 695,585 contigs. After filtering redundant contigs we obtained 16,491 clusters of non-annotated transcript and 57,637 clusters of sequences with a single BLASTx result, hereafter referred to as unigenes. Table 3.19 details the assembly statistics that show N50 values surpassing 1,000 base pairs and a similar guanine-cytosine percentage in the solitarious as well as the gregarious libraries—as was the case with the CNS sequencing (see Chapter 2). It is also remarkable that, in the whole transcriptome, 5,247 BLAST results belong to non animal species, probably because they correspond to sequences from microorganisms of the intestinal flora.

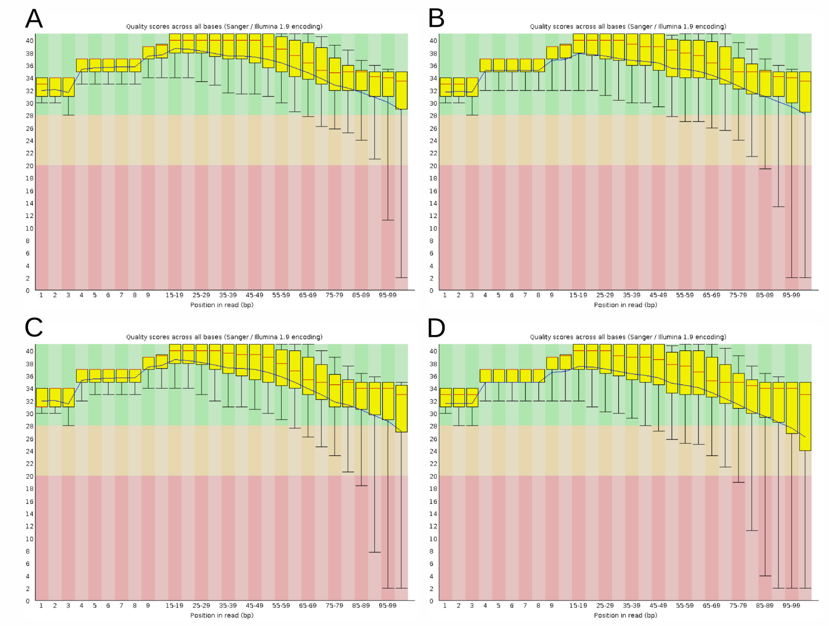

Figure 3.26: Distribution of the sequencing quality values across the 101 positions of the reads in each RNA-seq library. A, B: gregarious library pairs one and two, respectively. C, D: solitarious library pairs one and two, respectively.

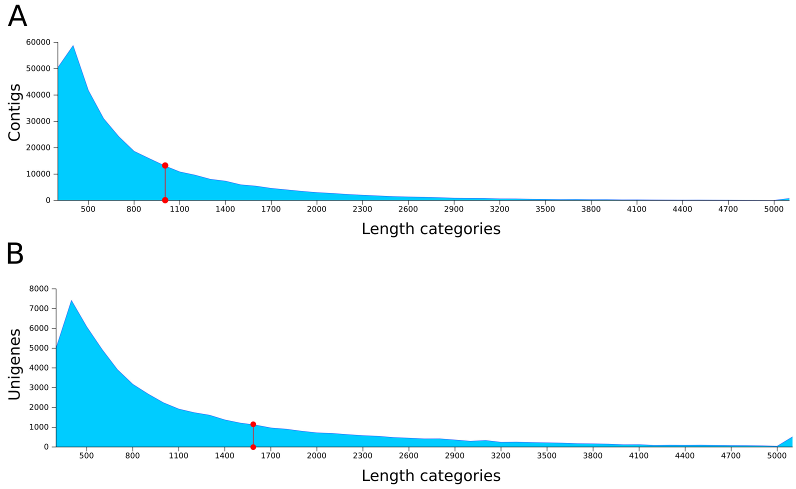

Figure 3.27: Sequence length distribution. X axis shows the length categories in intervals of 100 bases length. Y axis shows the number of contigs (A) or unigenes (B) belonging to each length interval, the first interval being 100 to 200 bases. Red bar in each distribution shows the N50 value.

Library	Total bases	Total reads	% GC	% Q30	% N
Solitarious DT	4,163,761,865	41,225,365	46.617	84.447	> 0.001
Gregarious DT	3,272,888,335	32,404,835	44.397	87.649	> 0.001

Table 3.18: Statistics of the read-pairs obtained for the RNA-seq libraries of the solitarious and gregarious *S. gregaria* DT.

3.3.2 Functional annotation and estimated diversity

Gene Ontology (GO) term counts from the biological process, molecular function and cellular component categories to which the annotated unigenes of the transcriptome belong are represented in figure 3.28A, and their total counts are shown in supplementary table 3.20. Metabolism of proteins, nitrogen and aromatic compounds (GO:0019538, GO:0006807 and GO:0006725, respectively) represent the highest amount of term counts in the transcriptome (2,433, 2,944 and 2,378, respectively), possibly due to the digestive functions and the primary metabolization of recently endocyted molecules in intestinal epithelial cells. Transport also gets a high term count (GO:0006810; 2,251 counts), probably because it contains down-stream terms that are involved in both lumen-to-cell and cell-to-lumen vesicle transport, indicating that not only import to the cells but also export to the intestinal lumen is highly active in the digestive tube. Cell component terms themselves show that the vast

Feature	Value
Assembled contigs	695,585
BLAST positive contigs	472,163
Unigene sequence clusters	74,128
Unique BLAST results	57,637
Unique sequences without BLAST result	16,491
N 50	1,028
Mean sequence length	601.44
Median sequence length	367
Maximum sequence length	18,753
Minimum sequence length	100
Guanine-Cytosine percentage	43.22

Table 3.19: Statistics of the *de novo* assembled transcriptome for *S. gregaria* digestive tube.

majority of the counts belong to the intracellular localization (GO:0005622, 5,845 counts), and that a high portion belong to membrane-bound organelle (GO:0043227, 3,259 counts). As to the molecular function terms, we found that the vast majority are included in the catalytic activity (GO:0003824; 5,875 counts) and binding terms (GO:0005488; 5,418 counts), with the hydrolase activity also presenting a high term count (GO:0016787, 2,213 counts). The GO term enrichement analysis casts 533 over-represented terms in one or the other phase, (288 in gregarious phase, 238 in solitarious phase and 7 in both phases). They were distributed as follows: 371 in biological process category (215 in gregarious phase, 156 in solitarious phase and none in both), 68 in cellular component category (36 in gregarious phase and 61 in solitarious phase and 1 in both) and 94 in molecular function category (37 in gregarious phase, 51 in solitarious phase and 7 in both). Suplementary table S2 lists the strongly over-represented terms (Fisher's exact test P-value < 0.001), of which 50 (Fisher's exact test P-value < 0.0001) are represented in figure 3.28B.

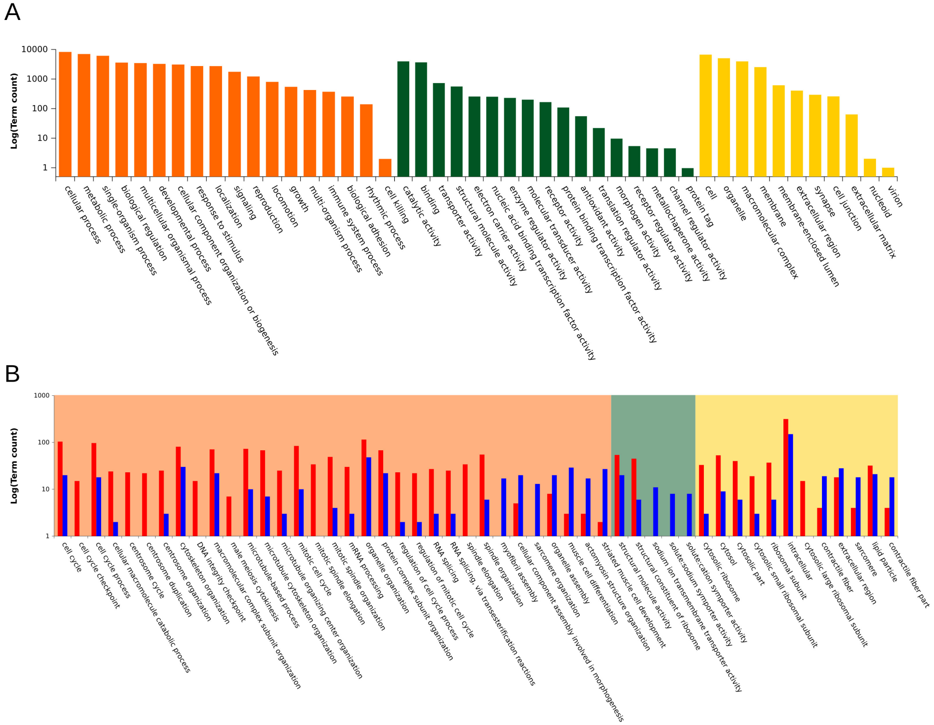

Figure 3.28: GO term counts and enrichment analysis results. A: GO term counts for all the annotated transcripts (excluding non-animal annotations). B: Enriched GO terms and their counts in each phase. All the results shown in this figure are statistically significant (Fisher's exact test < 0.0001). Color key: biological process in orange, molecular function in green and cellular component in yellow, gregarious in red, solitarious in blue. Online link: `https://drive.google.com/open?id=0B5K51IajvO_jV3d4MHlhSWV1NlE`

	Whole DT transcriptome	Gregarious DET set	Solitarious DET set	Present in both DET sets	Total from both DET sets
Biological process	45772	215	156	0	371
Cellular component	18935	36	31	1	68
Molecular function	14769	37	51	6	94
All	79476	288	238	7	533

Table 3.20: Statistics from the functional annotation of the differentially expressed transcripts. GO term counts for the three highest level categories are represented.

Regarding KEGG pathways, 1,893 unigenes were tagged by a total of 472 different enzyme codes that belong to 123 KEGG pathway categories containing a total of 4,336 KEGG category counts. KEGG term enrichment analysis threw a total of 25 unigens tagged by 26 enzyme codes from five pathways. 14 of these unigens were tagged by 18 enzyme codes belonging to four pathways exclusive to the solitarious phase and the remaining 11 unigens were tagged by eight enzyme codes from a single pathway exclusive to the gregarious phase (details in supplementary table 3.25). InterProScan terms were associated to 7,442 unigenes, 398 of which over-expressed, 223 in the gregarious phase and 175 in the solitarious one.

The distribution of the E-values (figure 3.29A) ranged from 10^{-6} (the maximum we set) to 10^{-60} for almost all BLAST results (near 85 % of the sequences). The species distribution of the BLAST hits shows mainly species belonging to the phylum arthropoda, with the exception of the protist *Gregarina niphandrodes*, whose 1,073 BLAST hits ranked it as the 10th most represented species (figure 3.29B). Among the most represented species, besides *G. niphandrodes*, there were two chelicerates (the subsocial spider *Stegodiphus mimosarum* and the mite *Ixodes scapularis*) and one crustacean (the water flea *Daphnia pulex*), the remaining were all part of the hexapoda. Among hexapoda, and ignoring *Drosophila melanogaster* (whose sequences were used as first BLAST database for annotation) the most represented species was the termite *Zootermopsis nevadensis.* Neither *S. gregaria* nor *L. migratoria* were among the top ten species in rank because of the scarcity of their aminoacids sequences in the nr database of the NCBI. Comparaison of the species distribution among the BLAST results of the over-expressed unigens revealed that *G. niphandrodes* infections are almost proper to the gregarious locusts (figures 3.29C and 3.29D) and that these show more microbial sequences than solitarious locusts do (see figures 3.31A and 3.31B).

3.3.3 Analysis of the transcript's differential expression analysis

The analysis of expression revealed a total of 4,412 differentially expressed transcripts (DETs), 2,019 over-expressed in solitarious phase and 2,393 over-

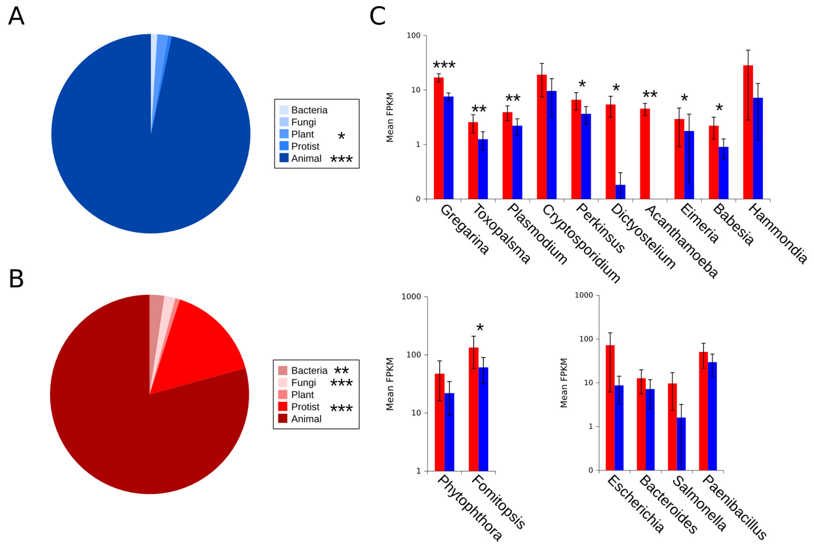

Figure 3.29: Distribution of the annotation metadata. A: E-value distributions, shown as the negative base 10 logarithm (-$\log_{10}$). B: Species distribution for all the transcriptome. C: Species distribution for the over-expressed transcripts in the gregarious phase. D: Species distribution for the over-expressed transcripts in the solitarious phase.

expressed in gregarious one. Accordingly, a volcano plot comparing the fold-change values and the FDR-corrected p-values for all the tested transcripts (Supplementary figure 3.30A) shows that the tendency of the expression is practically symmetrical. The MA plot shown in supplementary figure 3.30B indicates that a low number of transcripts have a low coverage despite presenting rather high fold change values, but the number of this low-coverage transcripts is clearly surpassed by the numbers of transcripts that had high absolute fold change values and significantly different coverage between phases. Of the differentially expressed transcripts, about two thirds (2,924) had significant BLAST hits (1,321 in the solitarious phase and 1,603 in gregarious one). The remaining third (1,488) were sequences with unknown annotation. 297 of the gregarious DETs and 35 of the solitarious ones had BLAST hits against sequences belonging to non-animal species.

We also estimated and compared the diversity of microorganisms in the solitarious and gregarious locust DT, at the kingdom level, based on the number of species to which the BLAST results of the DETs belonged (ilustrated in figures 3.31A and 3.31B). In our study, transcripts from species of the bacteria, fungi and protozoa kingdoms were significantly over-represented in the gregarious phase, whilst transcripts from the plants kingdom were

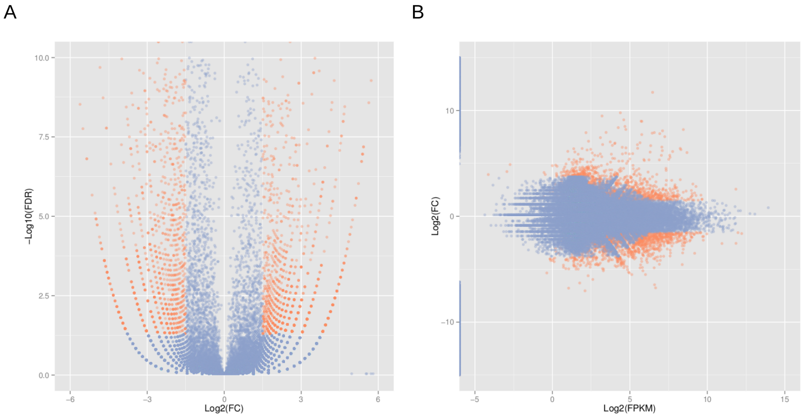

Figure 3.30: Overall results of the differential expression analysis between the solitarious and gregarious RNA-seq libraries. A: Volcano plot. X axis shows the base 2 logarithm of the fold change for each transcript. Y axis shows the negative base 10 logarithm of the FDR-corrected p-value for each transcript. B: MA plot. X axis shows the average coverage for each transcript (calculated from the base 2 logarithm of its solitarious and gregarious FPKMs). Y axis shows the base 2 logarithm of the fold change for each transcript. Transcripts with an absolute logarithm of fold change value higher than 1.5 and with an FDR value higher than 0.05 are colored in orange, the rest are in blue.

significantly over-represented in solitarious phase (χ^2 test p-values = 0.004 for bacteria, < 0.001 for fungi, < 0.001 for protists, and 0.015 for plants). Among the DETs with protist species assignation, 192 were homologous to sequences of the apicomplexa *G. niphandrodes*, 189 of them over-expressed in the gregarious phase. The most expressed transcripts from that species were transmembrane channels, structural proteins, ubiquitylation related proteins and several chitinase—all of them over-expressed in the gregarious phase. The coverage shows that the mean FPKM for all the 1,073 *G. niphandrodes* transcripts in gregarious phase doubles that mean in the solitarious phase (mean FPKM in the solitarious phase = 8.063 ± 1.311; mean FPKM in the gregarious phase = 19.199 ± 3.326; Student's t = 5.389, d.f. = 1,072, p-value < 0.001). We also detected differential coverage between phases for nine protist, four bacteria and two fungi species that were represented by at least three DETs, shown in figure 3.31C.

Fungi transcript count was significantly enriched in the gregarious phase (figure 3.31B), with two main representative species (*Fomitopsis* and *Phytophthora*). In fact, just one solitarious DET belongs to a fungus. However, although *Phytophthora* transcripts showed a higher number of gregarious DETs, there was no significant difference between their gregarious and solitarious

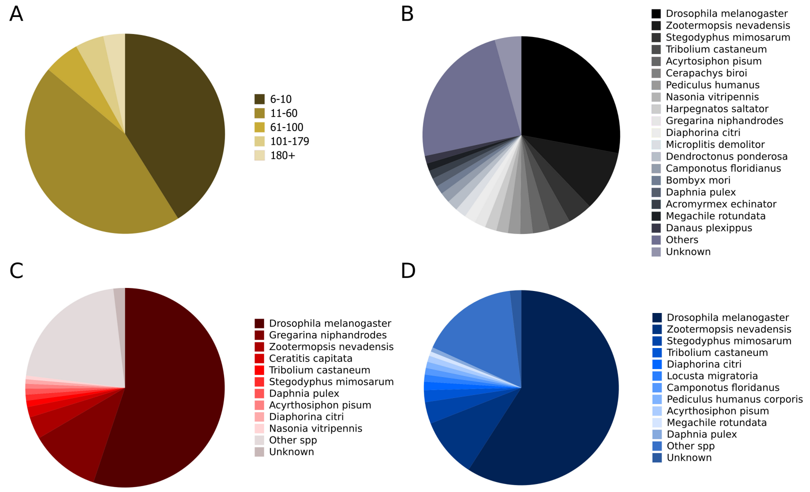

Figure 3.31: A: Taxonomical composition of the BLAST results of the DETs over-expressed in the solitarious phase. B: Taxonomical composition of the BLAST results of the DETs over-expressed in the gregarious phase. The signification of the enrichment is represented by asterisks in the legend. C: Overall expression levels of non-animal taxa at genus level, that show more than three DETs. X axis shows the genus and Y axis shows the mean FPKM of all the transcripts from that genus in the gregarious (red) and solitarious (blue) phase. The statistical significance for the paired Student's t tests is represented by asterisks (* = 0.05 > P > 0.01; ** = 0.01 > P > 0.001; *** = P < 0.001).

FPKM means. In contrast, the FPKM mean for the transcripts from the genus *Fomitopsis* did show significantly higher value for the gregarious DETs compared to the solitarious ones (figure 3.31C). As to the nature of these fungal DETs, we find putative chitinases, a ribosomal protein, several ABC transporters, a calcineurin and ATP synthases, all of them over-expressed in gregarious phase, although having relatively high coverage in the solitarious phase.

Bacterial species were represented by 380 assembled transcripts, with the genus *Escherichia* showing the highest number of transcripts (56). *Escherichia* also presented the highest number of DETs: four in gregarious phase and two in the solitarious one. The mean FPKM value for all the *Escherichia* transcripts was higher in the gregarious phase, compared to the solitarious one (figure 3.29C), but the difference was not significant. Other three genera with at least three DETs (*Bacteroides*, *Salmonella* and *Paenibacillus*) were represented by twelve, ten and seven transcripts, respectively. The six DETs annotated as from *Escherichia* were an IS1 transposase InsAB, two beta-lactamases and three hypothetical proteins with accession numbers EEH89633.1, EHN95545.1, ELC06673.1. The three DETs annotated as from *Bacteroides* encode ATPases type AAA. The three DETs annotated as from *Salmonella* were a hypothetical protein (accession number ESE59707.1) and two beta-lactamases. The three DETs annotated as from *Paenibacillus* were a subtilisin/peptidase S8, a thiol reductant ABC exporter subunit CydD and a tryptophan-rich sensory protein. Significant BLAST hits against other described enteric bacterial genera, such as *Enterobacter*, *Enterococcus* or *Streptococcus*, were also present but without enough differentially expressed transcripts as to take them into account. In fact, most of the minority bacterial transcripts belong to typical enteric bacterial genera, with *Pseudomonas*, *Enterobacter* and *Klebsiella* being among the most represented ones, but with no differences in expression between phases. We also had significant BLAST hits against pathogenic genera, such as *Streptomyces* (35 transcripts) and *Bacillus* (20 transcripts), again with low sequencing coverage.

Since the number of DETs is quite extensive (Supplementary table 3.26), we will focus on only those that show clear differential expression and who are involved either in the most represented or relevant biological processes. The following ten processes are to highlight:

1. Juvenile hormone regulation: we found four juvenile hormone metabolism and signalling-related DETs all of them over-expressed in the solitarious locusts, and 30 transcripts for the juvenile hormone-binding proteins (JHBPs), six of them DETs (four over-expressed in the solitarious phase and two in the gregarious one).

2. Regulation of gene expression: we found two DETs for DNA helicases, another four for genes related to transcription and seven related to translation. All these transcripts were over-expressed in the gregarious phase.

3. Peritrophic matrix: we found 139 transcripts for peritrophin (one of them significantly over-expressed in the gregarious phase and five in solitarious one) and several cuticular protein transcripts (all of them over-expressed in the solitarious phase).

4. Immune response: among the 43 transcripts for putative peptidoglycan recognition proteins (PGRP), which are involved in defence against Gram-positive bacteria and fungi, only four were DETs (two over-expressed in the gregarious phase and the remaining two in the solitarious one). From a total of 14 transcripts annotated as members of the Gram-negative bacteria binding proteins family, only one (GNPB1) was over-expressed (in the gregarious phase). Two out of nine hemolymph lipopolysaccharide-binding proteins (LBP) were over-expressed in the gregarious phase. Different defensins transcripts were over-expressed in

both phases, with a locustin being over-expressed in the gregarious phase and two defensins over-expressed in the solitarious one.

5. Oxidative stress and detoxification: among DETs belonging to these processes, we found three transcripts related to thioredoxin and a peroxiredoxin to be over-expressed in the gregarious phase. We also detected a good amount of cytochrome P450 and glutathione S-transferase transcripts to be over-expressed either in the gregarious or the solitarious phase (discussed latter on).

6. Apoptosis: as gregarious DETs belonging to this category, we detected at least three from the p53 apoptosis pathway, one apoptosis inhibitor (BAX inhibitor protein), one apoptosis promoter (FAM32b protein) and two cytochrome C related proteins. The only obvious apoptosis related DET detected in the solitarious phase is a protein that cleaves the p53 protein, inhibiting apoptosis.

7. Ubiquitylation: there were at least 24 DETs that relate to this function, almost all of them over-expressed in the gregarious phase. As gregarious DETs that belong to this category, and had the highest FPKM, we detected three ubiquitin ligase components (two E1 and one E3), one ubiquitylation protein and, also, two deubiquitylation proteins. Only one DET in this category, the Fbox only protein, which mediates protein ubiquitylation, was over-expressed in the solitarious phase.

8. Vesicle formation: the top DETs from this category were over-expressed in the gregarious phase. Of these we can highlight four adenoribosylation factors (ARF) and related transcripts and three SNARE related proteins.

9. Nervous system: we detected several transcripts belonging to this category, probably originating from the nerves that surround the digestive tube. Among the DETs belonging to this category we found three axon guidance transcripts (ninjurin, slit and spondin 1) to be over-expressed in the solitarious phase, whereas three receptors (GABA receptor, NMRP and nischarin) were over-expressed in the gregarious phase. Other nervous system-related DETs, such as those relating to vesicle formation and neural structure, had mixed expression pattern; with some being over-expressed in the gregarious locusts and others being over-expressed in the solitarious ones.

10. Calcium regulation: a good part of the Wnt/calcium regulation transcripts appear as over-expressed in the gregarious phase, including phospholipase C, casein kinase II and other proteins related to the regulation of intermediates (such as inositol 3 phosphate kinase) or to calcium transport (such as calmodulin).

To highlight is the extensive list of 468 proteases that we found in *S. gregaria*'s DT transcriptome. Serine-proteases are the most prevalent among these proteases, followed by carboxypeptidases (C-peptidases) and aminopeptidases (N-peptidases). Other kinds of proteases, such as metalloproteases or cysteine-proteases, had only a relatively minor representation (Supplementary figure 3.32A). Among the serine-proteases, trypsin-like proteases were the most abundant and closely followed by chymotrypsin-like proteases (Supplementary figure 3.32B). The number of DETs for serine-proteases (detailed in supplementary figure 3.32C) was higher in the solitarious phase than in the gregarious one (18 versus 12 DETs, respectively), and this tendency is maintained for the trypsin and trypsin-like peptidases (including trypsin itself, with seven solitarious DETs, chymotrypsin with one gregarious and two solitarious DETs, and neurotrypsin with one gregarious DET). Aminopeptidases, however, showed a gregarious tendency, with four DETs in the gregarious phase and two in the solitarious one. Several other peptidases had only one DET (such as a gregarious DET metalloprotease and a solitarious DET carboxypeptidase, supplementary figure 3.32C). We also detected transcripts for several putative digestive enzymes, including glycosyl hydrolases (28), α-amylase (109) and lipases (152), although only four α -amylases and four lipases were differentially expressed (two in each phase in both cases) and no glycoside hydrolase transcript was differentially expressed.

Other transcripts worth mentioning are those related to cytochrome P450 (CYP), heat shock proteins (HSP) and glutathione S-transferases (GST) protein families. A total of 289 CYP transcripts belonging to 25 different subfamilies were assembled, being the family CYP6 the most represented followed by the family CYP4 (Supplementary figure 3.33A). Twenty CYP transcripts are differentially expressed: one in the gregarious phase and the remaining 19 in the solitarious one (Supplementary figure 3.33B). We also assembled a total of 181 HSP transcripts, with HSP70 being the most represented family (figure 3.34A). We detected twelve HSP DETs, five of them belonging to HSP70, the expression of three of them was higher in the gregarious phase while the expression of the remaining two was higher in the solitarious phase, although the differences were not significant in any of these five cases (Supplementary figure 3.34B). Regarding GSTs, we assembled a total of 879 sequences grouped in 135 unique BLAST results. Six classes of GSTs were represented, being class sigma the most frequent one (figure 3.35A). Among these GST transcripts, six were over-expressed in the gregarious phase and seven in the solitarious one (figure 3.35B). The gregarious GSTs belong to microsomal and sigma classes, whereas the solitarious ones belong to delta and omega classes.

We compare our data on the expression of some transcripts with those published for these transcripts elsewhere (table 3.21). A good portion (87) of 107 transcripts reported elsewhere as up or down regulated in the gregarious phase show non-significant differences of expression in our transcriptome (20

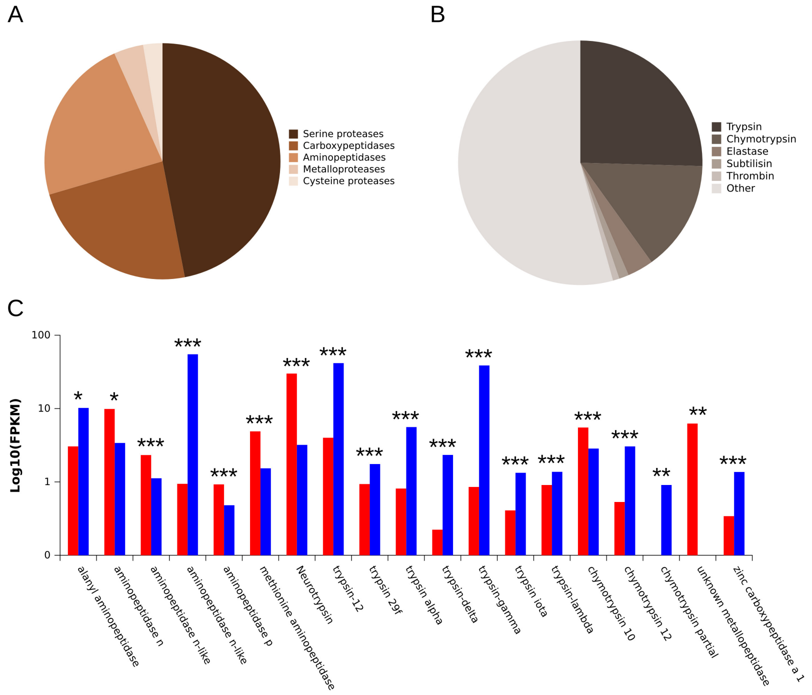

Figure 3.32: Proteases expressed in the digestive tube of *S. gregaria*. A: Relative abundance of all the proteases. B: Relative abundance of the serine proteases. C: Differentially expressed transcripts with homology to putative digestive proteases. X axis shows the transcript and Y axis shows the FPKM in the gregarious (red) and solitarious (green) libraries. FDR-corrected P-values for Yate's corrected chi-square test are represented with asterisks as follows: * = $0.05 > P > 0.01$; ** = $0.01 > P > 0.001$; *** = $P < 0.001$.

of them with zero as fold change value). Still, the proportion of transcripts that show the same expression tendency both in our work and others (12 significant, 36 non significant) was higher than that of the transcripts with higher expression in one phase in our work and in the other phase elsewhere (8 significant, 31 non significant).

Of the transcripts whose data we compared (Supplementary table 3.27), we found 12 DETs to show congruent differential expression both in our transcriptome and in four scientific publications. Four of these transcripts show over-expression towards the gregarious phase and eight towards the solitarious phase (details in table 3.22). We also compared our results to those of a microarray study, which compared brain against DT gene expression in

	Transcripts	Significant		Non significant		
Work		Same direcction	Opposite direction	Same direcction	Opposite direction	Zero FC
Kang et al. 2004	9	3	1	1	2	2
Chen et al. 2010	15	1	1	2	6	5
Guo et al. 2011	20	6	0	9	2	3
Badisco et al. 2011a	4	0	2	0	1	1
Badisco et al. 2011b	17	2	3	4	6	2
Wang et al. 2014	42	0	1	20	14	7
All	107	12	8	36	31	20

Table 3.21: Statistics of the comparative study between our digestive tube RNA-seq results and the results reported in the bibliography. We compared transcripts data from six scientific publications and we show the breakdown of the comparison results into those significant in our data and elsewhere, with over-expression both of the same or opposite direction, and those significant elsewhere but not in our data, with higher expression level both in the same or opposite direction (phase). The last column shows transcripts the number of transcripts reported as differentially expressed elsewhere but with a logarithm of FC equals to zero in our work. More details on this comparative study are in supplementary table 3.27.

L. migratoria [Spit et al., 2016]. We compared the presence, absence and abundance of transcripts studied in both works. The results are summarized in supplementary table 3.23.

3.4 Discussion

3.4.1 An extensive record of new transcripts expressed in the desert locust digestive tube

To date, transcriptome studies of *S. gregaria* were focused on the central nervous system (including Badisco et al. [2011a] and our analysis from Chapter 2). Aimed at explaining the differences in hormonal and neuronal gene expression differences between locust phases, these studies reported circa 20,000 annotated transcripts plus several thousands of transcripts that still have no known annotation. With the present *de novo* assembly of the transcriptome of the gregarious and solitarious *S. gregaria* digestive tube we obtained 57,637 unigenes, a number that triplicates the number of unique entries in other locust transcriptomes (see Chen et al. [2010], Badisco et al.

Transcript name	Over-expression phase	Accession number	Sequence length	Solitarious FPKM	Gregarious FPKM
		Kang et al. [2004]			
Annexin IX	Gregarious	XP_008471197	681	3.231	15.159
Larval cuticle protein precursor	Solitarious	XP_001861377	1092	28.789	0.270
Troponin C	Solitarious	P47949	5297	4.451	1.615
		Chen et al. [2010]			
Tyrosine hydroxalase	Gregarious	AAA62877	7096	1.595	4.988
		Guo et al. [2011]			
Protein TU-36B	Solitarious	P19967	13242	0.641	0.178
Lethal 2 essential for life protein	Solitarious	P82147	23547	1.789	0.927
Cytochrome p450 4g15	Solitarious	ACA04895	495	25.404	1.788
Cytochrome p450 6a2	Solitarious	P33270	145270	1.026	0.422
Heat shock protein Hsp20	Solitarious	AEV89751	3166	6.454	0.373
Peptidyl-prolyl cis-trans isomerase	Solitarious	AAX33414	29678	1.737	0.865
		Badisco et al. [2011b]			
Pacifastin-like peptide precursor 4	Gregarious	CAC82510	13210	5.164	12.995
Chromodomain helicase DNA binding protein	Gregarious	XP_005234509	749	2.518	17.721

Table 3.22: Transcripts whose statistically significant differential expression is congruent between our work and the data reported elsewhere. We also attach the accession number of the best significant BLAST hit, the sequence length and the solitarious and gregarious FPKMs for each of these transcripts as assembled in our *S. gregaria* DT transcriptome.

[2011a], Zhang et al. [2012], Wang et al. [2014b]). This higher number of assembled transcripts could be explained by the fact that RNA isolation and sequencing from intestinal contents unavoidably results in higher number of transcripts due to presence of genetic material from different species—the digestive contains bacteria, fungi, protists and plants.

The study of GO terms abundance (shown in figure 3.28A) revealed processes, functions and components expected to be present in the digestive tube of any insect. This way, the most represented biological process terms were those related to metabolism, development, transport and regulation, as the digestive tube in a place for digestive enzymes export, import of digestion products, regeneration of the intestinal epithelium and all the signalling needed for regulating all these and the other processes that take place in the digestive tube. In terms of molecular functions: the high number of hydrolase, binding and catalytic activity terms counts might be derived from the fact that the main activities that take place in the digestive tube are related to digestive

	L. migratoria DT	*S. gregaria* DT
	Proteases	
Proteases	132	481
Serin proteases	102	232
Metalloproteases	23	20
Cystein proteases	4	13
	Cytochrome P450	
Cytochrome P450	17	289
CYP2	Reported	4
CYP3	Reported	3
CYP4	Not reported	59
CYP6	Not reported	134
	Glutathions S-transferases	
Glutathione S- transferases	5	135
Delta	Reported	17
Epsilon	Not reported	11
Microsomal	Not reported	5
Mu	Reported	0
Omega	Not reported	8
Sigma	Reported	31
Theta	Not reported	6
	Transporters	
Major Facilitator Transporter	18	116
ABC transporter	7	32

Table 3.23: Sequence counts for several protein families from Spit et al. [2016]'s *L. migratoria* study and our *S. gregaria* DT transcriptome.

enzymes and defense molecules. As to the cellular components, the categories related to the membrane and cytoskeleton presented abundant term counts since a many channels, transporters and vesicle related proteins are recruited either for the secretion of enzymes or for transporting the products of the digested food.

The GO enrichment analysis detected various different processes that confirm the results from the central nervous system transcriptomics study. For instance, RNA processing (GO:0006396) and its subordinate categories are enriched in the gregarious phase, as a consequence of the enhanced transcriptional activity present in the gregarious phase, as occurs in the central nervous system of this species (see Chapter 2). One of these processes, RNA splicing (GO:0008380), is directly involved in the genesis of transcript diversity, with genes such as splicing factor 3a subunit 3 (required for U2 snRNP spliceosome assembly [Legrain et al., 1993]), highlights the link between mRNA splicing and the regulation of the phase change. In fact, several ribosomal protein related cell component tags (GO:0030529, GO:0044391 and GO:0015935) are also over-represented in gregarious phase, which reinforces the evidence of an enhanced gene expression in this phase. Term tags related to the proteasomal protein catabolic process (GO:0010498) and to the response to DNA damage stimulus (GO:0006974) were also over-represented in the gregarious phase, which might be reflection of the stressed status of the tissue generated by the phase stimuli and by the pathogen activity, which leads to apoptosis and oxidative stress, among other consequences.

Cell cycle (GO:0007049) and mitosis (GO:0000278) process terms are also over-represented in gregarious phase, indicating that cell proliferation is taking place in the epithelium in order to regenerate damaged cells. Of the transcripts belonging to this process we can cite scribbled (promotes epithelial cell proliferation [Qin et al., 2005]), cyclin-dependent kinase 1 (key control of cell cycle by phosphorilation of at least 75 target proteins [Enserink and Kolodner, 2010]) and fumble isoform F (panthotenate kinase required for cell division [Afshar et al., 2001], among other functions).We also found that the microtubule-based process (GO:0007017) and cell cycle regulators related to spindle formation, such as the serine/theronine proteases microtubule star (involved in the attachment of microtubules to the kinetocore during cell division [Snaith et al., 1996], PAR-1 isoform S (PAR-1 regulates the cell polarity through the microtubules of the cytoskeleton [Guo and Kemphues, 1995] among other functions), and the acidic leucine-rich nuclear phosphoprotein 32 family member (seems to act as a cell proliferation factor [Chen et al., 1996]). Other microtubule-related proteins, such as actin-like protein 87c (which seems to be part of the dynactin complex according to its GO tag, GO:0005869), are related to intracellular transport.

From their part, GO terms that were over-represented in the solitarious phase were related to muscle cell differentiation (GO:0042692), muscle cell homeostasis (GO:0046716), calcium regulation (GO:0055074) and sodium regulation (GO:0015081), indicating an apparently higher inversion in the muscular tissue rather than in other functions—probably due to the differences in size between solitarious and gregarious phases (solitarious individuals are usually bigger than gregarious ones, see Chapter 1) or a higher transcriptional investment towards other functions in the gregarious phase. Curiously, carboxylic acid biosynthesis (GO:0046394) is over-represented in the solitarious phase, with sequences such as the beta-alanine synthase (related to pyrimidine metabolism [Andersen et al., 2008] and involved in several processes including cuticle colouration and linking pyrimidine metabolism to beta-alanine metabolism [Rawls, 2006]), the argininosuccinate synthase (involved in arginine and urea metabolism [Haines et al., 2011]), and alkylglycerol monooxygenase (involved in alkylglycerol ether lipids hidroxylation [Tietz et al., 1964]). The over-expression of these transcripts might indicate an extra inversion in amino acid and lipid metabolism (more precisely, in anabolism) in the solitarious phase, opposed to the proteasomal processes that we found enriched in the gregarious phase—probably due to higher need for defense.

The KEGG pathway enrichment analysis indicates that pyrimidine metabolism (KEGG:00240) is over-represented in the gregarious phase, with sequences such as the ectonucleoside triphosphate diphosphohydrolase 5 (EC:3.6.1.6, involved in protein glycosilation and consumption of excessive ATP via dephosphorilation of NDPs [Fang et al., 2010]), a thioredoxin reductase-1 (EC:1.8.1.9, although it is related to oxidative stress regulation by regenerating thioredoxin, which deoxidyzes ribose in pyrimidinic bases (UTP and CTP) [Arnér and Holmgren, 2000], some isoforms can regulate differentiation and adhesion of diverse cell types [Nalvarte et al., 2015]), an UMP-CMP kinase (EC:2.7.4.14, with a key role in pyrimidine biosynthesis [Sugino et al., 1966]), and an aspartate carbamoyltransferase (EC:2.1.3.2, directly related to the first steps of pyrimidine synthesis [Simmer et al., 1990]). Several RNA polymerase subunits (EC:2.7.7.6) were also over-expressed in the gregarious phase. Gene expression takes an important portion of transcript abundance in the gregarious phase, so the over-represented terms shown both by the GO and KEGG analyses are in concordance with that.

All the four KEGG pathway terms that we found over-represented in the solitarious phase were related to amino acid and fatty acid metabolism. In accordance with the higher anabolism in the solitarious phase, we found several fatty acid biosynthesis enzymes. These include: acetyl-CoA carboxylase, isoform A (EC:6.4.1.2, biogenesis of long-chain fatty acids [Colbert et al., 2010]), fatty acid synthase 1, isoform A (EC:3.1.2.14, synthesis of long-chain fatty acids [Jayakumar et al., 1995]) and 4-Coumarate-CoA ligase 1 (EC:2.3.1.86 and EC:6.2.1.3, it is a fatty-acyl-CoA synthase homologous enzyme [de Azevedo Souza et al., 2009]). Regarding amino acid metabolism, mainly three pathways were enriched. They included enzymes such as: fumarylacetoacetase (EC:3.7.1.2, from tyrosine metabolism, hydrolyses 4-fumarylacetoacetate into acetoacetate and fumarate [Phaneuf et al., 1991]), mitochondrial trifunctional protein alpha subunit, isoform A (EC:4.2.1.17, tangentially related with phenylalanine metabolism, with enoyl-CoA hydratase activity used in the oxidation of long chained fatty acids [Uchida et al., 1992]), peptidoglycan-recognition protein-lb isoform 2 (EC:3.5.1.4, also from phenylalanine metabolism, with amidase activity and related to a protein family involved in immune response), and a set of alanine, aspartate and glutamate metabolism pathway enzymes, such as serine pyruvate aminotransferase (EC:2.6.1.44), argininosuccinate synthase (EC:6.3.4.5), protein N-terminal asparagine amidohydrolase isoform X1 (EC:3.5.1.1) and glutamate synthase 1 (EC:1.4.1.14, EC:1.4.1.13). Interestingly, among the different amino acid metabolism tags, some enzymes have overlapping functions, such as tyrosine aminotransferase (EC:2.6.1.5), succinate-semialdehyde dehydrogenase [NADP(+)] GabD (EC:1.2.1.24), tyrosine decarboxylase 2 (EC:4.1.1.28) and glutamate oxaloacetate transaminase 1, isoform B (EC:2.6.1.1).

3.4.2 The microbial enteric communities present differences between phases

The DT of insects are populated with a plethora of microorganisms, ranging from mutualistic bacteria to pathogenic protists [Dillon and Charnley, 2002]. As the cDNA libraries for the transcriptome analyzed here were synthesized from an eukaryotic mRNA-enriched template and not exclusively from polyA+ RNAs, we took advantage of the presence of non-eukaryotic transcripts to estimate how different the microbial communities of both phases were, based on transcript BLAST homology. This approach does not allow us to elaborate an accurate diversity study but, thanks to the presence of transcripts that have significant BLAST homology with sequences from bacteria, protist and fungi, we could perform a low resolution functional-based diversity analysis. This way we found that the diversity and genetic activity from the DT's microbes is higher in the gregarious phase than in the solitarious phase.

Given that gregarious locusts suffer higher contagion rates, and that in this study we did not stablished sterile colonies, we expected to find higher representation of pathogen transcripts in this phase. Our data support that expectation as the FPKMs of the microbial transcripts were higher in the gregarious phase than in the solitarious one (figure 3.29). The apicomplexa parasite *Gregarina niphandrodes* was the most represented non-arthropod organism, and the fact that the expression level of its transcripts in the gregarious phase almost doubles that of the solitarious phase proves that there the infection is more prevalent in our gregarious colony. However it is the parasite *Gregarina garnhami*, the one that is commonly found in locust and grasshopper laboratory colonies [Canning, 1956, Dillon and Charnley, 2002], while *G. niphandrodes* seems to parasitize other groups of insects, such as the mealworm beetle *Tenebrio molitor* [Clopton et al., 1992]. The transcripts that we found as homologous to *G. niphandrodes* (and even those homologous to other protist species) might therefore belong to *G. garnhami* whose BLAST data are not as complete and prevalent as those of other protist species. In fact, the GenBank database shows that for *G. niphandrodes* are 9,302 nucleotide, 1,919 EST and 12,751 protein entries, whereas for *G. garnhami* these entries are empty. Besides some transport, structural and signalling related transcripts, chitinases are to highlight among the annotated DETs from *G. niphandrodes.* They might be secreted by the pathogen in order to perforate the peritrophic matrix (which contains cuticular proteins and chitin) and gain access to the epithelial cells.

It is also remarkable that transcripts with BLAST annotations against *Fomitopsis* and *Phytophthora* sequences show the highest FPKM means among transcripts from non animal genera, meaning that fungi are at least transcriptionally active in the locusts' DT. The species of both fungi genera are plant parasites, which means that the fungal transcripts that we assembled from the locusts DT surely belong to other unidentified species with small, if any, representation in sequence databases. Although our study cannot determine whether these locust DT fungi are beneficial or harmful, among the fungal gregarious DETs there are several putative chitinases, that probably have the same function as the ones secreted by the protozoan parasites. There are also several fungi ABC transporters, which might import or export metabolites with coupled ATP consumption [Goffeau et al., 2004]. It would be interesting to identify the fungi species present in the locust DT, which might clarify whether they are parasitic or not, and probably allow for exploration of potential new pest control tools.

The *S. gregaria* DT microbial community is predominantly composed of Enterobacteriaceae, from the phylum Proteobacteria [Dillon and Charnley, 2002]. Among these, sequences from *Escherichia coli* were the most abundant. *E. coli* is broadly found in the intestinal lumen of a many insect species, both as a commensalist, mutualist or pathogen, depending of the strain [Bentley and Meganathan, 1982, Hudault et al., 2001]. Since only six out of 56 *E. coli* sequences were differentially expressed towards gregarious phase, and two of these were genes involved in antibiotic resistance (beta-lactamase homologs, [Abraham and Chain, 1940]), we can deduce that these bacteria might be defending themselves against the enhanced immune response that we saw taking place in the gregarious phase (see bellow), or against antibiotic substances emited by other micro-organisms. The differential presence between locust phases of sequences from other bacterial genera, such as two beta-lactamases from *Salmonella*, seem to support our interpretation. Although *Pantoea agglomerans* (formerly *Enterobacter agglomerans*) is reported to produce aggregation pheromone precursors inside the locust DT [Dillon et al., 2000, Dillon and Charnley, 2002], we did not detect any sequence from this species. The reasons might be the low prevalence of sequences from species other than eukaryotic in the poly A+ enriched RNA sample that we used for sequencing. Another plausible cause for the absence of *P. agglomerans* sequences might be the annotation of its sequences under the old genus name, *Enterobacter*. In fact we had 12 sequences annotated as from *Enterobacter*: a set of hypothetical proteins, a fimbrium component and four beta-lactamases, although none of them shows differential expression towards the gregarious phase.

In addition to *E. coli* and *Salmonella* (for which 10 sequences were assembled), both enterobacteria, we found differential presence of sequences from bacteria belonging to the phylum Firmicutes, such as *Paenibacillus*. The *Paenibacillus* sequences that we found as more prevalent in the gregarious phase also have defensive and stress-related functions. Indeed, we found a subtilisin/peptidase S8 (belongs to a family of endopeptidases with several functions [Siezen and Leunissen, 1997], including the cleavage of harmful proteins), a thiol reductant ABC exporter subunit CydD (transports glutathione across the bacterial cytoplasmic membrane [Pittman et al., 2005], which contributes to redox homeostasis) and a tryptophan-rich sensory protein (involved in signal transduction coupled to stress response in some bacteria [Davey and de Bruijn, 2000]). Bacteria from this genus appear in various ecosystems, from soil to marine sediments and rizhosphere [McSpadden Gardener, 2004, Lal and Tabacchioni, 2009], but not in intestinal lumen. The sequences that we assembled might thus belong to another phylogenetically related, spore-forming, bacteria. The other Firmicutes genus for which we found several sequences as differentially present in the gregarious phase is *Bacteroides*. Its sequences encode ATPases type AAA, which might be involved in multiple biological functions, such as molecule transport and protein remodelling [Snider et al., 2008]). Species of *Bacteroides* have been found in the DT of the termite *Reticulitermes flavipes*; they play a crucial role in recycling nitrogenous residues via uricolysis [Potrikus and Breznak, 1981]. But, apart from our work, there is no reported evidence of the presence of species from this genus in the locust DT.

The fact that we found several sequences belonging to plant species is due to not emptying the DT's contents before RNA extraction and sequencing. In fact, almost all of the BLAST results for these sequences are from the *Brassica* genus—our locusts were fed on cabbage. The relatively more prevalent plant transcripts in the solitarious phase, if not a false positive, might be due to higher amounts of cabbage in the DT of the solitarious individuals compared to that in the DT of the gregarious ones (probably because gregarious individuals have to share the food with their cage mates).

3.4.3 Phase-specific transcripts of *S. gregaria*'s digestive tube

The results of the transcripts' differential expression analysis revealed a good number of genes that are already known to be over-expressed either in the gregarious or the solitarious phase, which speaks to the overall quality of our work and its results. A good expample is the juvenile hormone (JH) synthase and the up- and down-stream regulation genes of its well known hormonal pathway. Mostly represented in solitarious phase, the JH is involved in maintaining the inter-molt state of the insect. When the insect is about to molt, JH levels decrease and ecdysteroid levels rise in the hemolymph, promoting the ecdysis [Schneiderman and Gilbert, 1964]. It is well known that JH concentration is higher in the solitarious phase hemolymph titre [Wiesel et al., 1996]. Accordingly, we found the JH-related genes to be over-expressed in the solitarious locusts. These include JH acid methyltransferase (an enzyme involved in JH activation), a group of juvenile hormone binding proteins (JHBP) with no specific function described, a 50KD midgut protein described as up-regulated by JH in termites [Sen et al., 2013], a methoprene tolerant receptor activated by a JHBP [Jindra et al., 2013], an ecdysteroid kinase homologue sequence probably related to inactivation of ecdysteroids (the JH function-complementary hormones involved in molt regulation [Ito and Sonobe, 2009]), and the lipid transport and storage protein vitellogenin (regulated by JH in insects [Chen et al., 1979]). Other works on JH in *S. gregaria* also support its over-expression in the solitarious phase, suggesting that solitarious females have higher JH titers, probably to facilitate the mobilization of vitellogenin or to induce solitarious specific responses [Injeyan and Tobe, 1981, Wiesel et al., 1996]. The fact that our study is based on adult tissues discards the option that the differential expression of these genes is due to their involvement in maintaining the inter-molt state. The higher expression of JH-related genes might be linked to an enhanced lipid transport or the extended life span in the larger solitarious locusts that also live longer than the gregarious ones. In agreement with the differential expression of the JH, the number of reads mapped to JHBPs was higher in the solitarious phase, although the proportion of JHBPs that were over-expressed in the solitarious phase is not significantly different from that of the over-expressed JHBPs in the gregarious phase (four out of 20 in the solitarious phase compared to two out of 20 in gregarious phase, see supplementary table 3.26). A more extensive study of the differentially expressed JHBPs might reveal its implication in JH down-stream response in each phase.

We found that RNA synthesis is a prominent biological process in the gregarious phase. For instance, among the top DETs were the 60S ribosomal protein and nucleolin (ribosomal metabolism), elongation factor Elf1 (regulator of transcription) and tRNA synthases and tRNA maturation enzymes (translation). It is also to note that several helicases, involved in several processes related to gene expression regulation, are over-expressed in the gregarious phase. For example, CHD-1 and DDX17 act as RNA polymerase II-dependent transcription regulators, although there are several alternative functions described for these helicases (such as chromatin structure maintenance for CHD-1 [Delmas et al., 1993] and RNA processing for DDX17 [Fuller-Pace and Ali, 2008]). Coupled to the expression of these helicases, we find that components of the holoenzyme RNA polymerase II are also over-expressed in the gregarious phase. At the translational level, we can also count the eukaryotic translation initiation factor 4 proteins (needed for initiating almost every translation event in the cell). A transcript for vigilin was over-expressed in the gregarious phase; it has functions as diverse as tRNA transport from the nucleus to the cytoplasm [Kruse et al., 1998], RNA metabolism or vitellogenin regulation [Dodson and Shapiro, 1997]. The higher expression of transcripts related to the gene expression machinery in the gregarious phase is concordant with the results that we report for this category of genes in the solitarious and gregarious transcriptomes of *S. gregaria*'s central nervous system (in Chapter 2).

In brief, the transcriptional machinery, from chromatin structure regulation to RNA editing, seems highly active in the gregarious digestive tube, as happens in the gregarious central nervous system, where we find these processes to be over-represented. Both the phase change and the immune challenges that the gregarious individuals face might contribute to the higher activity of the genes involved in gene expression in order to produce a higher yield of proteins related to the regulation of both processes. However, while in central nervous system the amount of DETs is asymmetrically biased towards the gregarious phase, the proportion of DETs from the two phases is more balanced in the DT, which indicates that the differences in gene expression affects the expression levels rather than the transcript diversity (number of genes expressed).

3.4.4 Digestion and detoxification-related transcripts of the locust DT

One of the specific characteristics of the insect digestive tube is the peritrophic matrix. It is composed by cuticular proteins assembled by peritrophin, a mucin derivated protein [Terra, 2001]. During digestion, the peritrophic matrix separates from the epithelium and surrounds the food bolus, thus optimizing the action of the digestive enzymes. In the locusts *S. gregaria* and *L. migratoria*, the peritrophic membrane covers the midgut epithelium and caeca [Bernays, 1981]. The fact that the transcription of genes related to the peritrophic membrane is increased in the solitarious phase might be understood in two ways: I) the body size of the solitarious individuals is larger than that of the gregarious ones, so the expression of peritrophic matrix component transcripts is higher in the former phase—the same logic as for the case of the genes involved in structural components, such as muscular fibers; or II) the higher stress level to which the gregarious locusts are exposed makes them invest more in other functions, such as innate immune response and phase change related genes. Both hypotheses seem to be supported by our data, since we report higher abundance of transcripts related to the maintennance of tissue and cell structures in the solitarious phase (which supports the first hypothesis), but we also report over-expression of several stress response-related transcripts in the gregarious phase (supporting the second hypothesis).

The principal process in digestion is the breaking of big molecules into smaller ones in order to facilitate their absorption by the intestinal epithelium cells. After scrutinising all the probable digestive enzymes from our transcriptome, we found that the most abundant lytic enzymes were proteases. Serine proteases, which are involved in diverse functions [Krem and Di Cera, 2001], were the most abundant type. Belonging to this family are trypsin and chymotrypsin, digestive enzymes that are conserved among animals. Other members of the same protease family include the gregarious phase DET neurotrypsin (which cleaves agrin, a proteoglycan involved in neuromuscular junction formation [Stephan et al., 2008]) and several trypsin-like proteins that are potentially involved in many insect processes (see Lazarević and Janković-Tomanić [2015] for a list of functions). A good portion of the proteases that we assembled in *S. gregaria*'s DT transcriptome (almost 40 % of the total) were N- and C-peptidases, which contain digestive enzymes. Most of the proteasess were not differentially expressed, probably because of their phase-independent digestive functions. The same might be said about lipases and glycosidases: none of them was prevalent, compared to proteases, and only four lipases were differentially expressed between locust phases, which might be due to the constitutive/housekeeping functions of these enzymes.

Ingestion of plants involves dealing with toxic substances such as glucosynolates or tannins. This requires a detoxification response to get rid of these substances along with their metabolites. One of the most represented families of such detoxification-related proteins that we assembled is the cytochrome P450 (CYP) family. Its monooxygenase function is involved in a wide range of processes [Feyereisen, 1999]. CYP transcripts are involved in detoxification of insecticides and xenobiotics, and take part of a good number of metabolic pathways in insects. The only gregarious over-expressed CYP is homologous to CYP6K1, a member of CYP6 family, linked to resistance to insecticides in several insects [Daborn et al., 2007, Ding et al., 2013, Yang et al., 2016]. The nineteen solitarious over-expressed CYP transcripts mostly belong to subfamilies 4, 6 and 9, and on average are expressed four folds higher that in the gregarious phase. As explained for the CNS transcriptome (Chapter 2), the higher metabolism of the gregarious locusts might be what drives the increase in the expression of these detoxifying molecules. The over-expression of some CYPs in the solitarious phase can be explained by the high number of possible functions in which these molecules are involved.

3.4.5 Innate defense: from the intestinal epithelium to the hemolymph.

Due to their life style, gregarious locusts compete for resources and space with the rest of the swarm, fly very long distances to find new food resources, once they have devastated the area where they were, have to even actively escape from the reach of their own kind in order to avoid cannibalism and, logically, recure to cannibalism if they can. These life conditions lead to stress response activation, modifying a high portion of the locust physiology. They also increase the probability of getting infected by parasites and pathogens, thus enhancing the activity of the constitutive immune response in order to control the pathogenic microorganisms to which the gregarious locust is exposed. This way, the gregarious locusts have to face a double-cost immune response: one

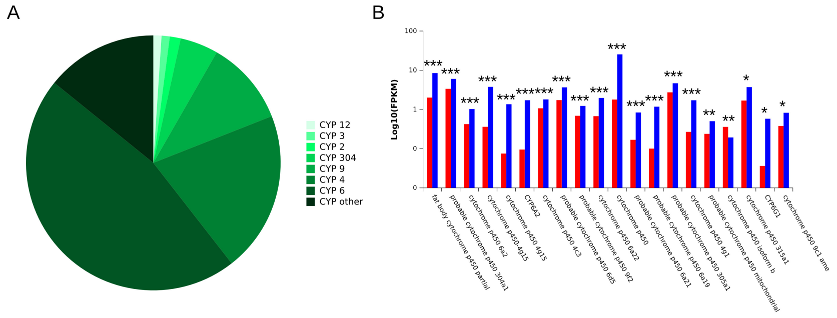

Figure 3.33: Cytocrome P450 (CYP) transcripts expressed in the digestive tube of *S. gregaria*. A: Relative abundance. B: Expression levels, in FPKM, of the transcripts (X axis) between solitarious (blue) and gregarious (red) libraries. FDR-corrected P-values for Yate's corrected chi-square test are represented with asterisks as follows: * = $0.05 > P > 0.01$; ** = $0.01 > P > 0.001$; *** = $P < 0.001$.

induced by the stress produced by the crowd conditions and other induced by the higher pathogenic load (see Adamo [2016]). We therefore expect to find the key genes that activate both immune responses (stress driven and constitutive) among the over-expressed contigs in the gregarious phase.

As part of the insect innate immune response, peptidoglycan recognition proteins (PGRPs, Dziarski [2004]) and Gram-negative bacteria binding proteins (GNBPs, [Kim et al., 2000]) are constitutively expressed in the intestinal epithelial cells, and exported to the intestinal lumen [Engel and Moran, 2013]. Once there, they bind to peptidoglycan and other bacterial cell wall components. If the enteric bacteria population exceeds certain threshold, these proteins trigger a primary immune response that leads to the excretion of antimicrobial peptides, such as defensins [Zasloff, 2002], that inhibit pathogens, and of dual oxidases that induce an oxidant environment inside the intestinal lumen. In addition to PGRPs and GNBPs, there are haemolymph lipopolysaccharide-binding proteins which recognize Gram-negative bacteria such as *E. coli* [Schumann et al., 1990].

The balance of antibacterial expressed genes was inclined towards the gregarious phase, which clearly indicates the active status of the antimicrobial response during that phase. The defense response in the solitarious phase, although present, was not as strong as that of the gregarious phase. More evidences reinforcing our interpretation are the gregarious over-expression of up-stream immune response genes, such as the Spaetzle-processing enzyme, that cleaves the Toll-signalling antimicrobial response precursor of the protein Spaetzle into an active form [Levashina et al., 1999]. Spaetzle regulates the immune response against Gram-positive bacteria and fungi by activating a series of proteins that eventually promote the release of antimicrobial peptides [Ligoxygakis et al., 2002, Hoffmann, 2003].

Another group of proteins with important roles in the immune response are the serin protease inhibitors (also known as serpins). We found serpin 3 to be over-expressed in the gregarious digestive transcriptome. Serpin 3 negatively regulates the phenoloxidase activation proteinase PAP-3 in the tobacco worm *Manduca sexta* [Zhu et al., 2003], thus inhibiting the immune response mediated by melanin. Two transcripts belonging to another serpin family, pacifastin, were also differentially expressed between locust phases. One (a partial sequence homologous to pacifastin-related precursor 4t) was over-expressed in the gregarious phase, which is concordant both with our study of the *S. gregaria* CNS (Chapter 2) and with the results in Simonet et al. [2005], Badisco et al. [2011b]. Some pacifastins are involved in the inhibition of digestive serin proteases, such as trypsin and trypsin-like proteases, but others also inhibit enzymes involved in other biological functions. Interestingly, it seems that a pacifastin also inhibits PAP-3 in the crayfish *Pacifastacus leniusculus* [Aspán et al., 1990, Liang et al., 1997], thus having the same role as the insects' serpin 3. PAP-3 activates prophenoloxidase and converts it into functional phenoloxidase, an enzyme that produces melanin and can inhibit pathogens via melanization. The fact that both pacifastin 4 and serpin 3 are over-expressed in the gregarious DT might indicate control of the immune response, probably via a negative feedback, in order to avoid excess of melanine production, which might be harmful to the insect itself.

The involvement in the immune response of some of transcripts that we found over-expressed in the gregarious phase rather part of their pleiotropic effects. A good example of these is apolipophorin III. It is a lipoprotein that can unbind its lipidic part, normally used for lipid transport, to then recognize bacterial cell wall elements and induce the immune response [Ogoyi et al., 1995, Malik and Amir, 2011]. When the locust faces an immune challenge, apolipophorin III works as a pathogen recognition protein [Chung and Ourth, 2002, Zdybicka-Barabas et al., 2013]. However, under other stressfull conditions (such as intense flight) its function shifts towards lipid transport in order to optimize resource usage. This results in a trade-off between metabolic and stress response and immunological investment in a similar way as it was described for the cricket *Gryllus texenis* [Adamo et al., 2008]. The ganglioside GM2 activator also accomplishes both a lipid transport and immune response functions. It does not only transports ganglioside glycerolipids, but it was also found to activate beta-hexosaminidase (a facultative peptidoglycan hydrolase) in mammals [Koo et al., 2008]. However, there is no evidence of beta-hexosaminidase differential expression in our transcriptome (despite its presence among the assembled transcripts) and no hexosamidase activation function of the ganglioside GM2 activator has been reported in insects yet. Other proteins that were over-expressed in the solitarious phase, such as elongation of very long chain fatty acids 7-

like protein, are specific to lipid transport and metabolism. Their differential expression could be a consequence of the tendency to accumulate fat by the well fed and less active solitarious locusts.

Several of the assembled transcripts contain domains that might indicate their involvement in the immune response. For example, several transcripts contain an immunoglobulin-like domain but, due to the multiple functions of this domain, we cannot confirm the involvement of the transcripts that contain it in the immune functions. The same could be said of lectins, whose ability to bind to saccharides is sometimes used for structural purposes, but also in antigen recognition (C-type lectins, for instance, are involved in hemaglutination of hemocytes [Weis et al., 1998] and in innate immunity as pattern recognition proteins [Robinson et al., 2006]). Another protein group also abundant in the transcriptome are the leucine-rich repeat-containing proteins, some of them being ribonuclease inhibitors (related to RNA lifespan regulation, [Kobe and Deisenhofer, 1993]), Toll-like receptors (related to Toll signalling, [Bell et al., 2003]) or more specific proteins as the tropomyosin regulator tropomodulin (regulates actin filament length in muscle, [Matsuzaka et al., 2004]). Despite the presence of the leucine-rich repeat-containing protein 40 among the top over-expressed transcripts in the gregarious phase (which might indicate its importance in phase change), the lack of specific information about this sequence makes difficult its assignation to some of the above cited functions.

Among the most prominent protein families that are associated with stress response is the heat shock protein (HSP) family. The main function of HSPs is to maintain the folding pattern of proteins under several stress conditions [Lindquist and Craig, 1988]. However, some HSPs are also involved in antigen recognition and binding, thus working as facultative immune response agents [Thériault et al., 2005, Nishikawa et al., 2008]. Besides, HSPs do not only come in different functional and expression (constitutive versus inducible) variants, they also vary in terms of cellular localization (cytosolic versus mitochondrial). Several transcripts of this protein family seem to be differentially expressed towards gregarious phase in *L. migratoria*, which might relate them to phase change [Wang et al., 2007, Chen et al., 2015]. Regarding *S. gregaria*, HSP70 is the most abundant family of HSPs in our transcriptome (figure 3.34A), presenting several members with a plethora of functions, including protection against protein aggregation and acting as an endoplasmatic reticulum transporter protein [Tavaria et al., 1996, Daugaard et al., 2007]. This presence of multiple copies and the multiple functions are probably the reason why we found HSP70 DETs both in the gregarious (three transcripts) and solitarious (two transcripts) phases (see figure 3.34B). One HSP20 transcript, related with heat tolerance [Groenen et al., 1994], is over-expressed in the solitarious phase (figure 3.34B). It is also worth highlighting that two mitochondrial variants (HSP10 and HSP60) were over-expressed in the gregarious phase, while other HSPs are differentially expressed in one phase or the other. Given the data, this protein family seems to contain different transcripts that are associated with one phase or the other but, unlike the case of *L. migratoria* [Wang et al., 2007], the HSP family shows no clear patterns of association with any of the two *S. gregaria* phases. The reason for this difference might be the fact that only six different HSP transcripts were studied in *L. migratoria*, whereas we had a high number of HSP in the *S. gregaria* transcriptome that we assembled—which might indicate that a good number of *L. migratoria* HSP variants might present different expresion patterns. Whatever the case, more studies are needed in order to determine the exact functions of these HSP transcripts and their relation to the locusts' phase change.

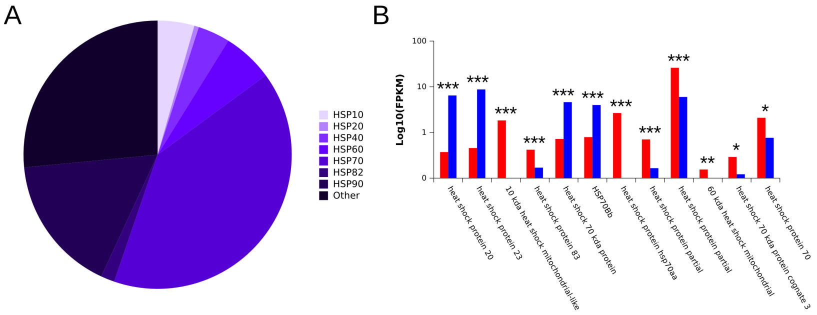

Figure 3.34: Heat shock protein (HSP) transcripts expressed in the digestive tube of *S. gregaria*. A: Relative abundance. B: Expression levels, in FPKM, of the transcripts (X axis) between solitarious (blue) and gregarious (red) libraries. FDR-corrected P-values for Yate's corrected chi-square test are represented with asterisks as follows: * = 0.05 > P > 0.01; ** = 0.01 > P > 0.001; *** = P < 0.001.

3.4.6 Oxidative stress and the differential expression between phases

Recognition of pathogens in the digestive tube drives a response involving redox reactions that increase the concentration of reactive oxygen species (ROS) inside the intestinal lumen and set a hostile environment for pathogens (see Engel and Moran [2013]). The increase in pathogen-driven and oxidative stress levels is harmful for the organism and requires counter-measures. Because of the higher microbial activity in the gregarious phase, we expect an increment of ROS levels in that phase. The resulting increase in molecular and organelle damage would eventually lead to cell apoptosis.

Accordingly, several redox homeostasis transcripts are over-expressed in the gregarious phase. Examples include prostaglandin synthase, thioredoxin and peroxiredoxin. These three transcripts are involved in the redox reactions, where they reduce hydrogen peroxide to water by peroxidation of the active cysteine of the peroxiredoxin, then coupling the reduction of the inactive peroxidized peroxiredoxin to the oxidation of thioredoxin [Rhee et al., 2005]. Thioredoxin reductase 1, the enzyme that reduces peroxidized thioredoxin with NADPH, is also over-expressed in the gregarious phase, closing so the regeneration pathway of peroxiredoxin. Sharing this expression pattern is also the ATP-dependent NAD(P)HX dehydratase, an enzyme that repairs the hydrated form of NADPH (NADPHX), which is a dehydrogenase inhibitor [Regueiro et al., 1970]. A thioredoxin-like prostaglandin synthase, prostamide/prostaglandin F2 alpha synthase, is also over-expressed in the gregarious phase. This enzyme reduces prostaglandin H2 (PGH2) to prostaglandin F2 alpha (PGF2a), which contributes to the coupled oxidation of other thioredoxin, thus potentially increasing the amount of ROS [Moriuchi et al., 2008]. In addition to their hormonal functions, prostaglandins are also eicosanoids involved in the immune response against bacteria [Tunaz et al., 2001, Stanley et al., 2009], so their production might be linked to the enhanced immune response of the locust.

On the other hand, other transcripts involved in redox processes are clearly over-expressed in the solitarious individuals. An example is the enzyme gamma-glutamyl cyclotransferase which competes with glutathione synthase for gamma-glutamil-cysteine [Richman and Meister, 1975], thus potentially lowering the regeneration of glutathione. It is also related to amino acid transport [Thompson and Meister, 1975], which might be the main reason for its increase in the solitarious locusts. Linked to detoxification, we also find that vanin is over-expressed in the solitarious phase. It is an enzyme that breaks pantetheine into vitamin B5 and cysteamine [Maras et al., 1999, Boersma et al., 2014]—the latter being related to redox regulation because it raises the levels of glutathione and thioredoxin inside the cells [Khomenko et al., 2003, Wilmer et al., 2011]. Transcripts for redox regulation are over-expressed in solitarious phase while oxidative stress is increased in the gregarious phase, meaning that the expression of redox regulators in the solitarious phase might be more related to redox-based signalling rather than to oxidative stress palliation.

One of the most representative families of proteins involved in detoxification by redox reactions are the glutathione S-transferases (GSTs, EC: 2.5.1.18). They utilize glutathione as substrate for detoxification of xenobiotic molecules by addition of a glutathione radical to the target molecule [Sheehan et al., 2001]. Higher glutathione synthase levels indicate a more efficient, if not forced, oxidative status of the organism. Besides, GSTs are involved in several other steps of the detoxification of xenobiotic substances (KEGG pathway 00980)—where CYPs are also involved. This way, GSTs are involved in detoxification of several insecticides and peroxidized products. For instance, delta class GSTs have been reported to process substrates such as the insecticide

DDT [Udomsinprasert et al., 2005], omega class GSTs are involved in detoxification of heavy metals [Laliberte et al., 2003, De Chaudhuri et al., 2008], and microsomal and sigma classes interacts with peroxidized lipids in the endoplasmatic reticulum or cytoplasm [Agianian et al., 2003]. GSTs are probably related more to signalling and xenobiotic detoxification in the solitarious phase (delta and omega classes) and to oxidative stress palliation in the gregarious one (microsomal and sigma classes), as we suggest in the former paragraph. Overall, all the GST classes are involved in avoiding cell damage.

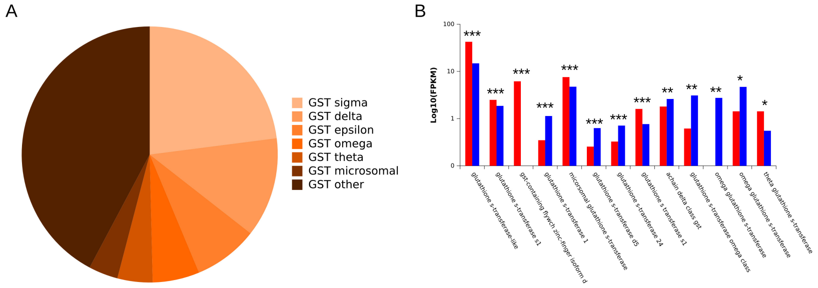

Figure 3.35: Glutathione S-transferase (GST) transcripts expressed in the digestive tube of *S. gregaria*. A: Relative abundance. B: Expression levels, in FPKM, of the transcripts (X axis) between solitarious (blue) and gregarious (red) libraries. FDR-corrected P-values for Yate's corrected chi-square test are represented with asterisks as follows: * = $0.05 > P > 0.01$; ** = $0.01 > P > 0.001$; *** = $P < 0.001$.

As stated above, excess of ROS can damage molecules, organelles and cells and activates molecule-modification pathways. One of these pathways is the polyubiquitylation signalling, which mainly regulates the addition of various ubiquitin peptides to the proteins that are destined to degradation by the proteasome. Congruently, most of the transcripts that relate to polyubiquitylation are over-expressed in the gregarious phase, as an expected consequence of the cummulation of pathogen-induced damage and ROS. Overall, E3 ubiquitin ligases are notorious in the gregarious phase: E3 ubiquitin-protein ligase RNF181 is directly involved in the ubiquitylation of target poteins [Lee et al., 2014], while E1 ubiquitin-activating protein AOS1 is related to the the activation of that ubiquitin [Johnson et al., 1997]. Some of the genes involved in the polyubiquitylation pathway are also involved in other regulation pathways (such as the microbial DNA-induced immune response thanks to E3 ubiquitin-protein ligase AMFR-like, and mitosis regultaion by Apc11 anaphase-promoting complex subunit). In addition, we found several genes that show homology with protein domains related to ubiquitin ligases (such as WWE domains [Gmachl et al., 2000, Aravind, 2001, Wang et al., 2014a]) to be over-expressed in gregarious phase. Not only strict ubiquitylation is over-represented in the gregarious phase, several proteins involved in ubiquitin-like signalling (such as the ubiquitin-like protein 5) share this expression pattern. It is also remarkable that ubiquityl-protein hydrolases and other deubiquitylation enzymes (such as UBX domain-containing protein 6-like [Madsen et al., 2011] or protein FAM152b) also are over-expressed in the gregarious phase. Not surprisingly, an F-box containing transcript (involved in ubiquitilation mediating [Bai et al., 1996]) is the only transcript related to ubiquitylation among those over-expressed in the solitarious phase.

Polyubiquitylation signalling, together with the presence of excessive ROS and pathogen agents, also activates apoptosis pathways, such as p53 and caspase. Some ubiquitin-related modifiers are also related to the p53 apoptosis pathway and are over-expressed in the gregarious phase. For example, ubiquitin-like modifier-activating enzyme ATG7 regulates vacuole formation in autophagy and p53/caspase-8 apoptosis processes [Lee et al., 2012]. The CDC14A protein regulates p53 apoptosis [Li et al., 2000], among other functions (such as stopping mitosis to begin DNA replication and DNA repair [Mocciaro and Schiebel, 2010]). Since leak of Cytochrome C from mitochondria to cytoplasm is part of the apoptosis signal [Kadenbach et al., 2004], the over-expression of components of the cytochrome C oxidase complex in the gregarious phase could also be caused by higher levels of apoptosis. The Fam32A protein, a promoter of apoptosis in humans, is also over-expressed in the gregarious phase, but little is known about this protein in insects. In contrast, the solitarious locusts show over-expression of the Tp53rk-binding protein, a protein kinase that phosphorylates the tumour suppressor p53 protein, thus activating it and leading to apoptosis inhibition [Huh et al., 2007]. Unexpectedly, some apoptosis inhibitors and antagonists are over-expressed in the gregarious phase. This might be due to possible pleiotropic effects. For instance, BAX inhibitor 1 is an apoptosis inhibitor [Krajewska et al., 2011] but also as a regulator of calcium inside the endoplasmatic reticulum, leaking Ca^{2+} to the cytoplasm [Bultynck et al., 2012]. Death-associated inhibitor of apoptosis 2 (DIHA) is an apoptosis inhibitor dependent of PGRPs that polyubiquitylates the caspase DREDD [Meinander et al., 2012], but it has been reported that its expression is needed for the immune response against Gram negative bacteria [Huh et al., 2007].

The presence of residues in cytoplasm, as well as the presence of apoptosomes and autophagosomes, promote the formation and traffic of vesicles, whose related DETs belong to several protein families of which ADP ribosylation factors (ARFs) and SNARE proteins were notorious. ARFs belong to the RAS protein family, consisting of G proteins that induce vesicle formation, actin modulation, and some of their members are also involved in apoptosis. While there are over-expressed transcripts of the RAS family in both phases, ARFs appear specifically over-expressed in the gregarious phase. Overall, the lack of phase specificity of the RAS DETs might be explained because, besides formation and trafficking of apoptosomes and autophagosomes, the intestinal epitellium cells are constantly secreting (exporting) enzymes and immune response peptides, as well as absorbing (importing) nutrients to the cell cytoplasm—none of these normally occurring processes is affected by the phase of the locust. SNARE proteins, which are involved in the fusion of vesicles (such as fusion of lysosomes and peroxisomes, to promote cellular digestion) or other biological membranes (such as the cis and trans-Golgi networks and the plasmatic membrane, to promote secretion), are also specifically over-expressed in the gregarious phase. More transcriptional evidences in support of an enhanced secretory activity in the gregarious phase include, among others: the signal peptide receptor (SSR) beta subunit, involved in the inclusion of proteins from the ribosomes into endoplasmatic reticulum [Görlich et al., 1990], and tetraspanin CD63, a molecular facilitator that is present in late endosomes and interacts with the plasmatic membrane to form an exosome [Pols and Klumperman, 2009]—in addition its potential indirectly involvement in the immune response [Levy and Shoham, 2005].

3.4.7 Transcriptional evidence of the digestive tube stimulation by the nervous system in the gregarious phase.

The digestive tube is covered by part of the peripheral nervous system, refered as enteric nervous system, which function is to establish DT-CNS and CNS-DT crosstalk and regulate peristaltism and intestinal secretion. We therefore expect our RNA-seq study of the solitarious and gregarious digestive tubes of *S. gregaria* to show differences in the expression levels of genes related to the nervous system. Accordingly, we found that several of these transcripts were differentially expressed towards the same phase both in the CNS and DT transcriptome. For example, the GABA-receptor associated protein (GRAP) is over-expressed in the gregarious phase both in the digestive tube and CNS. This protein is involved in the cytoplasmic transport of the GABA receptor to the plasmatic membrane [Kittler et al., 2001], meaning that the reception of GABA is promoted in the digestive tube of the gregarious locusts. In insects, GABA is mainly an inhibitory neuropeptide, although it can be excitatory, depending on the type of GABA-ergic neuron that receives the stimulus, affecting to physiology and behaviour [Pitman, 1971, Ffrench-Constant et al., 1993]. In addition, GRAP is also related to apoptosis signalling and autophagy promotion, so a pleiotropic effect might occur in the digestive system, both promoting muscle stimulation and autophagy in the gregarious locusts [Kabeya et al., 2004]. Also over-expressed in gregarious phase is the receptor white, which is an ABC transporter for tryptohpan. This receptor is more known for being involved in transport of pigment precursors, such as tryptophan or guanine, to the compound eyes during insect developement [Mackenzie et al., 1999]. However, tryptophan is also the precursor for the neurotransmitter serotonin [Borycz et al., 2008], which has been reported to be involved in phase change [Anstey et al., 2009, Guo et al., 2013, Tanaka and Nishide, 2013, Rogers et al., 2014], among other functions. So white might be involved in tryptophan import to the CNS for production of serotonin.

The receptor no mechanoreceptor potential C protein (NMRP-C), involved in positive peristaltism regulation, is also over-expressed in the gregarious phase and, interestingly, works as a mechanorreceptor involved in aggregation behaviour in *Drosophila* [Cheng et al., 2010]. A calmodulin dependent nitric oxide (NO) synthase is over-expressed in the gregarious phase, producing NO, which is involved in many biological aspects, including intestinal mucosa integrity [Whittle et al., 1990] and immune response in mammalian DT [McCafferty et al., 1997] by interacting with neuropeptides. However, NO synthases can contribute to peroxide formation in some conditions, so they might contribute to the ROS formation [Xia et al., 1998]. It is known that stressful conditions can activate the peristaltism in other organisms, so if we consider the fact that the stress response is stronger in gregarious phase, we should expect over-expression of the transcripts that relate to increased peristaltism in the gregarious locusts. Curiously, we found Ly6/neurotoxin 2 (Lynx2), a protein homologous to Lynx1 [Dessaud et al., 2006], to be over-expressed in the solitarious locusts with high FPKM values (Supplementary table 3.26). The Lynx proteins interact with nitrosamine and acetylcholine receptors, maintaining the cholinergic pathway [Tsetlin, 2015] and, depending on the variant, they modulate the expression of specific nitrosamine and acetylcholine receptors [Wang et al., 2015]. In the same sense, we also found a nicotinic acetylcholine receptor to be over-expressed in the solitarious phase.

Over-expression of the potassium voltage-gated channel protein Shal in the gregarious phase might be an indicator of the need for rapid repolarization of neurons in the digestive tube. This channel is involved in the recovery of the electric potential difference of the membrane by transporting K+ ions to the needed compartment. A transcriptional increment of this protein might thus be used for the recuperation of the nervous impulse [Ping and Tsunoda, 2012]. Complexin, a nervous system-specific protein that interacts with SNARE proteins during vesicle fusion [Trimbuch and Rosenmund, 2016], is also over-expressed in the gregarious phase. As stated in the oxidative stress section, vesicle formation and transport seem to be enhanced in the gregarious locusts. That increase is therefore not only for importing defensive products and exosomes to the intestinal lumen, but also for releasing neurotransmitters in the synapsis and neural plates. Following the same tendency, clavesin-1, a molecule expressed in neurons, where it associates with phosphatidilinosytol-3,5-biphosphate (PIP2), interacts with clatrin and is present in the vesicles whose destiny are endosomes and lysosomes [Katoh et al., 2009]. The actin-like protein 87c (commented in the GO analysis section), synonymous of the actin-related protein 1, is also over-expressed in the gregarious phase. It is part of the dynactin complex and it is involved in synapse retraction [Eaton et al., 2002]. Overall, the transport of ions, organelles and vesicles (all of them related to neurons) seems to be enhanced in our gregarious RNA-seq libraries, both of the CNS and the digestive tube.

3.4.8 Calcium regulation: roles in immune response, neuronal signalling and muscle contraction in the digestive tube

The expression of several enzymes from the Wnt/calcium signalling pathway is increased in the gregarious phase. On case of these is phospholipase C, that metabolizes the membrane lipid phosphatidylinositol 2-phosphate (PIP2) to inosytol trisphosphate (I3P) and diacylglycerol (DAG) [Kühl et al., 2000]. While DAG remains in the plasmatic membrane, I3P is released to the cytoplasm and interacts with endoplasmatic reticulum (ER) calcium-associated channels that release calcium. Among the proteins whose genes are over-expressed in the gregarious locusts is a protein that contains a domain similar to an I3P receptor (activates ER calcium release), the preprotachykinin (precursor of a vasodilator and peristaltism regulation neuropeptide whose receptor is a GPCR that activates phospholipase C [Torrens et al., 1989]), and the I3P kinase (enzyme that sequentially phosphorilates I3P to I6P for gene expression regulation [Odom et al., 2000]). Cytosolic calcium transporters, such as calmodulin and calmodulin-related proteins, are also over-expressed in the gregarious locusts. These transporters interact with other proteins contributing to calcium signalling. DAG works also as a precursor that can be processed by prostaglandin synthase, enzyme commented in the oxidative stress section among those over-expressed in the gregarious phase.

Wnt regulation presents several pathways with complementary activation. For example, Wnt/calcium pathway elements can down-regulate the Wnt/beta-catenin pathway [Sugimura and Li, 2010]. This might explain why we do find inhibitors or antagonists of the Wnt/beta-catenin pathway (such as trophoblast glycoprotein [Zhao et al., 2014] and nucleoredoxin [Funato et al., 2006]) over-expressed in the gregarious phase. Calcium is a very important and ubiquitous secondary messenger involved in the regulation of multiple biological processes, including neuronal transmission, cytoskelleton regulation, enzyme activation and immune system response, among others. Calcium is also involved in muscle contraction, both in synaptic regulation and myosin contraction. Contraction of the intestinal muscle is regulated at a higher level by hormones such as prostaglandin and NO, as mentioned above. However, the transcripts coding for structural components of the muscles, such as myosin, tropomyosin and actin filaments related to muscles are clearly over-expressed in solitarious phase. This can be explained by de differences in size between solitarious and gregarious individuals, being the former smaller and thus might invest less in expression of genes related to muscle structure and maintenance. Another explanation might be the differences in investment in other processes: whilst gregarious locusts face stressful conditions that require higher expression of stress and immune response genes, solitarious locusts invest in housekeeping functions such as muscular fibre maintenance.

A study on commensalist bacterial communities in the mammalian DT revealed that *Bacteroides thetaiotaomicron* packs an inositol hexaphosphate phosphatase in vesicles that are then endocyted by the epithelial cells of the DT [Stentz et al., 2014]. By doing so, these bacteria can induce calcium signalling in the host cells, by cleaving inositol hexaphosphate into inositol polyphosphates and rising levels of I3P. It therefore looks as if the symbiotic bacteria were "warning" the host about the presence of pathogens in the intestinal lumen, thus, being a sort of accessory pathway to the stimulation of immune response. Interestingly, one of the bacteria with both more transcript counts and more than three DETs towards the gregarious phase belongs to the genus *Bacteroides*. The fact that the expression of I3P signalling transcripts is up-regulated in the gregarious phase, as do several immune response transcripts, might indicate that this inter-kingdom communication could be happening also in the desert locusts' DT.

3.4.9 Comparison with the published data:

The transcriptome that we assembled for the digestive tube of *S. gregaria* contains only a few transcripts whose significant differential expression was previously reported elsewhere (Supplementary table 3.27). The number is expectedly higher when we do not consider the statistical significance and take into account only the tendency of the difference in expression level. Since the analysis of differential expression applies the statistics test to thousands of cases, and despite correcting for type I errors (false positives) with post-hoc statistics (like the FDR), some cases will be false positives and others false negatives—the correction for type I error is not absolute. By taking into account the non-significant cases where the tendency of the difference in expression is the same as the significant over-expression reported in the bibliography, we do not intend to claim the confirmation of these transcripts. We rather pinpoint these results as to keep in mind. It is to highlight also that the published scientific works that we used for the comparaisons are on differences tissues and even species, and use different methodologies. All of these are reasons for the low number of genes that share the same expression pattern between works (see supplementary table 3.27 in chapter 4 for the specific information on each of scientific works used for the comparison).

Only two transcripts were comparable between our results and those published for *S. gregaria*. Both, pacifastin-like peptide precursor 4 and a chromodomain helicase DNA binding protein (CHD) homologous sequence, were congruently over-expressed in the gregarious locusts in both sources. Being a protease inhibitor linked to immune response [Liang et al., 1997] and with several studies marking it as typically up-regulated in the gregarious phase [Simonet et al., 2005, Badisco et al., 2011b], pacifastin-like peptide precursor 4 may thus be considered as a solid marker sequence for the gregarious phase. Chromodomain helicase DNA binding protein (CHD) is usually involved in chromatin structure modification, and our data contain several CHDs, such as CHD-1 and CHD-3 (related to histone deacetylation [Tong et al., 1998]), all of them congruently over-expressed in the gregarious phase.

The most relevant coincidence when comparing to the data on *L. migratoria* is the tyrosine hydroxalase, the enzyme that synthetises L-DOPA. It is reported both in our transcriptome and in Chen et al. [2010] as up-regulated in the gregarious phase, thus confirming the role of catecholamines in the regulation of phase change [Ma et al., 2011]. Other gene families also present the same over-expression pattern in both sources. Examples are the CYPs and the HSP20, both of them over-expressed in the solitarious phase in our transcriptome and in [Guo et al., 2011]. Also up-regulated in our solitarious *S. gregaria* and *L. migratoria* from Guo et al. [2011] are cuticle proteins (present in insect DT because their role as part of the peritrofic matrix) and a peptidyl-prolyl cis-trans isomerase (involved in protein folding [Takahashi et al., 1989, Liu et al., 1990] and sometimes in the promotion of viral propagation [Anderson et al., 2011]). The latter two protein groups were also validated in *S. gregaria* CNS transcriptome both via qPCR and RNA-seq (Chapter 2). The remaining transcripts that we found up-regulated in the solitarious phase both in the bibliography and our DT transcriptome are components of muscular tissues. These include lethal 2 essential for life (HSP homologous sequence involved in lifespan increase [Vos et al., 2015] and muscle development [Wójtowicz et al., 2015]), protein TU-36B (a muscle specific cytochrome B5 with unknown function, [Levin et al., 1989]) and troponin C (component of cytoskelletal filaments from muscle cells, it binds to calcium and regulates muscle contraction [Herzberg et al., 1986]). These last results add to the evidences gathered in our GO enrichment and differential expression analysis, where muscle-related transcripts appear up-regulated in the solitarious phase.

No transcriptomics or gene expression data are available for *L. migratoria*'s digestive tube— work from Spit et al. [2016] compared the expression of crowd-reared *L. migratoria* digestive tube against the brain. Furthermore, Spit et al. [2016]'s work is based on microarray technology, thus sequence information is determined a priori from *L. migratoria* and *S. gregaria* known sequences, which limits the number of sequences studied compared to an NGS *de novo* assembled transcriptome. Still, we could mine that work and retrieve information about the presence, absence and diversity of transcripts (Supplementary table 3.23). We coincide in several aspects. For instance, the diversity of proteases is similar, with the most abundant ones being serine proteases, followed by metalloproteases then cysteine proteases. Two of the most abundant GSTs in our transcriptome (types delta and sigma) are reported in the work from Spit et al. [2016]. We both report a sequence homologous to the major allergen protein Bla G1, probably involved in digestion in hemimetabolous insects [Gore and Schal, 2004], thus highlighting the presence of allergenic proteins and their association with the phase in both species of locusts. As to CYPs, however, the most represented classes were CYP4 and CYP6 in our work, and CYP2 and CYP3 in Spit et al. [2016]'s work.

3.5 Conclusions

In summary, this work offers a complete record of transcripts from the DT of Schistocerca gregaria, with thousands of transcripts identified by BLAST search and other thousands with no BLAST identity. We also infer how the phase change affects gene transcription in the locust DT. Immune response, RNA metabolism and vesicle formation related transcripts are more represented in the gregarious phase due to the action of the two drivers of stress (population density and pathogens). On the other hand, more constitutive functions are more represented in the solitarious phase. We also see how GABA and catecholamine signalling over-expression is concordant between the digestive tube and the CNS, and that calcium signalling and ubiquitilation are also abundantly represented in the gregarious phase. We report significant taxonomical differences in transcript abundance and coverage between both phases, with bacteria, fungi and protozoa having higher presence in the DT of the gregarious locusts. In fact, this last result may also be understood as differences both at gene expression level and the taxon diversity. Among these microbes, the most differentially represented genus was the apicomplexa pathogen *Gregarina* spp, which might be considered for locust control (although we have to determine if the species is *G. niphandrodes* or *G. garnhami*). The higher density of pathogens in the gregarious phase may lead innate immune responses driven, some calcium-regulated and some lead to increased ROS levels in intestinal lumen, which in turn triggers ROS buffering and detoxification processes as well as cell damage, apoptosis and autophagy. It is to highlight that most of the DETs might have pleiotropic functions, not necessarily linked to phase change. What is clear is that the differences between phases in the digestive tube are modulated at two levels: host's phase and microbial abundance, which rises the question of whether the microbial community is what affect the host's phase or viceversa.

3.6 Supplementary material

Table 3.24: Over-represented GO terms in the gregarious and solitarious phase. We show the GO terms belonging to the specified principal category and hierarchical level, showing both the P-value for the Fisher's exact test derived from the enrichment analysis and the logarithm of fold change between phases.

GO	GO tag	GO term	Total	Gregarious	Solitarious	FET	Log(FC)
level			counts			pvalue	
		Gregarious phase					
		Biological process					
3	GO:0016043	cellular component organization	2519	158	85	1,783E-04	-0,894
4	GO:0007049	cell cycle	898	104	20	1,656E-15	-2,379
4	GO:0007017	microtubule-based process	570	73	10	3,450E-13	-2,868
4	GO:0043933	macromolecular complex subunit organization	737	71	22	5,077E-08	-1,690
4	GO:0006996	organelle organization	1554	115	48	2,670E-06	-1,261
5	GO:0000278	mitotic cell cycle	558	84	10	1,210E-18	-3,070
5	GO:0022402	cell cycle process	745	97	18	1,704E-17	-2,430
5	GO:0071822	protein complex subunit organization	673	68	22	1,726E-08	-1,628
5	GO:0007010	cytoskeleton organization	915	81	30	1,546E-07	-1,433
5	GO:0009057	macromolecule catabolic process	255	26	4	3,535E-04	-2,700
6	GO:0000226	microtubule cytoskeleton organization	472	68	7	2,258E-15	-3,280
6	GO:0010564	regulation of cell cycle process	175	23	2	1,407E-05	-3,524
6	GO:0044265	cellular macromolecule catabolic process	201	24	2	3,810E-05	-3,585
6	GO:0007346	regulation of mitotic cell cycle	185	22	2	8,283E-05	-3,459
6	GO:0000075	cell cycle checkpoint	98	15	1	9,346E-05	-3,907
6	GO:0006974	response to DNA damage stimulus	272	27	4	2,242E-04	-2,755
6	GO:0044770	cell cycle phase transition	149	18	4	2,960E-04	-2,170
7	GO:0007051	spindle organization	254	55	6	1,185E-18	-3,196
7	GO:0031023	microtubule organizing center organization	140	25	3	7,927E-08	-3,059
7	GO:0031570	DNA integrity checkpoint	85	15	0	3,380E-05	-20,517
7	GO:0044772	mitotic cell cycle phase transition	141	18	2	2,529E-04	-3,170
7	GO:1901987	regulation of cell cycle phase transition	128	16	2	6,684E-04	-3,000
7	GO:0048568	embryonic organ development	60	10	3	0,000988331	-1,737
8	GO:0007052	mitotic spindle organization	197	49	4	5,989E-19	-3,615
8	GO:0051231	spindle elongation	98	34	1	3,730E-17	-5,087
8	GO:0051297	centrosome organization	140	25	3	7,602E-08	-3,059
8	GO:0007112	male meiosis cytokinesis	18	7	1	7,757E-05	-2,807
8	GO:0000077	DNA damage checkpoint	80	13	0	2,258E-04	-20,310
8	GO:0006396	RNA processing	406	36	7	3,551E-04	-2,363
8	GO:0016071	mRNA metabolic process	336	31	4	4,809E-04	-2,954
8	GO:1901990	regulation of mitotic cell cycle phase transition	127	16	2	6,100E-04	-3,000
9	GO:0000022	mitotic spindle elongation	97	34	1	1,562E-18	-5,087
9	GO:0007098	centrosome cycle	109	23	1	2,959E-09	-4,524
9	GO:0008380	RNA splicing	264	27	3	5,396E-05	-3,170
9	GO:0006397	mRNA processing	314	30	3	6,745E-05	-3,322
9	GO:0010498	proteasomal protein catabolic process	61	11	1	1,363E-04	-3,459
9	GO:0031572	G2 DNA damage checkpoint	76	12	0	2,088E-04	-20,195
9	GO:0031123	RNA 3'-end processing	56	10	2	2,917E-04	-2,322

GO level	GO tag	GO term	Total counts	Gregarious	Solitarious	FET pvalue	Log(FC)
10	GO:0051298	centrosome duplication	86	22	1	1,651E-10	-4,459
10	GO:0000375	RNA splicing, via transesterification reactions	252	25	3	9,744E-05	-3,059
10	GO:0007093	mitotic cell cycle checkpoint	87	12	1	4,867E-04	-3,585
10	GO:0031124	mRNA 3'-end processing	44	8	0	8,732E-04	-19,610
11	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	252	25	3	9,065E-05	-3,059
11	GO:0044774	mitotic DNA integrity checkpoint	78	12	0	1,949E-04	-20,195
11	GO:0043161	proteasomal ubiquitin-dependent protein catabolic process	58	10	1	2,944E-04	-3,322
12	GO:0000398	mRNA splicing, via spliceosome	252	25	3	1,435E-04	-3,059
12	GO:0044773	mitotic DNA damage checkpoint	77	12	0	2,413E-04	-20,195
13	GO:0007095	mitotic G2 DNA damage checkpoint	76	12	0	1,062E-04	-20,195
Molecular function							
2	GO:0005198	structural molecule activity	765	54	20	4,439E-08	-1,433
3	GO:0003735	structural constituent of ribosome	575	45	6	3,254E-07	-2,907
5	GO:0016779	nucleotidyltransferase activity	101	12	1	2,738E-04	-3,585
6	GO:0034062	RNA polymerase activity	25	6	0	5,310E-04	-19,195
7	GO:0003899	DNA-directed RNA polymerase activity	25	6	0	4,991E-04	-19,195
Cellular component							
3	GO:0044422	organelle part	2658	183	79	1,45E-04	-1,212
4	GO:0005622	intracellular	5845	315	150	1,35E-06	-1,070
5	GO:0044424	intracellular part	5724	313	150	4,77E-06	-1,061
6	GO:0030529	ribonucleoprotein complex	1204	89	13	0,001	-2,775
8	GO:0005829	cytosol	293	53	9	1,03E-10	-2,558
8	GO:0005811	lipid particle	260	32	21	0,000532396	-0,608
9	GO:0044445	cytosolic part	174	40	6	4,54E-10	-2,737
9	GO:0044391	ribosomal subunit	202	37	6	3,70E-07	-2,624
10	GO:0022626	cytosolic ribosome	99	33	3	2,85E-11	-3,459
10	GO:0015935	small ribosomal subunit	98	20	4	0,000133454	-2,322
11	GO:0022627	cytosolic small ribosomal subunit	56	19	3	3,63E-07	-2,663
11	GO:0022625	cytosolic large ribosomal subunit	44	15	0	6,25E-06	-20,517
11	GO:0005730	nucleolus	101	19	3	4,41E-04	-2,663
11	GO:0071013	catalytic step 2 spliceosome	140	23	2	0,00062322	-3,524
Solitarious phase							
Biological process							
5	GO:0070925	organelle assembly	229	8	20	4,803E-05	1,322
5	GO:0048878	chemical homeostasis	170	5	15	3,644E-04	1,585
5	GO:0048646	anatomical structure formation involved in morphogenesis	599	24	34	5,513E-04	0,503
6	GO:0010927	cellular component assembly involved in morphogenesis	218	5	20	3,254E-05	2,000
6	GO:0042692	muscle cell differentiation	405	3	29	5,033E-05	3,273
6	GO:0046716	muscle cell cellular homeostasis	62	2	9	2,447E-04	2,170
6	GO:0055082	cellular chemical homeostasis	99	3	11	4,345E-04	1,874
6	GO:0050801	ion homeostasis	118	4	12	5,066E-04	1,585
6	GO:0016053	organic acid biosynthetic process	256	8	19	6,221E-04	1,248
7	GO:0031032	actomyosin structure organization	171	3	17	7,273E-05	2,502
7	GO:0051146	striated muscle cell differentiation	394	2	28	1,267E-04	3,807
7	GO:0055001	muscle cell development	374	2	27	1,306E-04	3,755
7	GO:0006873	cellular ion homeostasis	98	3	11	5,219E-04	1,874
8	GO:0055002	striated muscle cell development	374	2	27	8,182E-05	3,755
8	GO:0032787	monocarboxylic acid metabolic process	211	6	17	5,207E-04	1,502
8	GO:0055065	metal ion homeostasis	87	0	10	6,332E-04	19,932
8	GO:0046394	carboxylic acid biosynthetic process	256	8	19	6,455E-04	1,248
8	GO:0030003	cellular cation homeostasis	89	3	10	7,438E-04	1,737

GO level	GO tag	GO term	Total counts	Gregarious	Solitarious	FET pvalue	Log(FC)
8	GO:0072507	divalent inorganic cation homeostasis	59	0	8	8,266E-04	19,610
9	GO:0030239	myofibril assembly	138	1	17	4,611E-06	4,087
9	GO:0006875	cellular metal ion homeostasis	79	0	10	3,416E-04	19,932
9	GO:0055074	calcium ion homeostasis	53	0	8	4,654E-04	19,610
9	GO:0072503	cellular divalent inorganic cation homeostasis	54	0	8	5,205E-04	19,610
10	GO:0045214	sarcomere organization	103	0	13	4,235E-05	20,310
10	GO:0006874	cellular calcium ion homeostasis	53	0	8	4,342E-04	19,610
		Molecular function					
4	GO:0022804	active transmembrane transporter activity	277	15	19	1,368E-04	0,341
4	GO:0005214	structural constituent of chitin-based cuticle	26	1	5	8,686E-04	2,322
5	GO:0005523	tropomyosin binding	27	0	6	1,573E-04	19,195
5	GO:0015291	secondary active transmembrane transporter activity	94	3	10	3,042E-04	1,737
6	GO:0015293	symporter activity	52	1	8	1,788E-04	3,000
7	GO:0015294	solute:cation symporter activity	50	1	8	7,455E-05	3,000
8	GO:0046873	metal ion transmembrane transporter activity	165	3	14	4,377E-04	2,222
9	GO:0015081	sodium ion transmembrane transporter activity	72	1	11	2,885E-06	3,459
10	GO:0015370	solute:sodium symporter activity	34	0	8	3,146E-05	19,610
		Cellular component					
2	GO:0005576	extracellular region	392	18	28	1,001E-05	0,637
8	GO:0043292	contractile fiber	185	4	19	4,689E-06	2,248
8	GO:0005811	lipid particle	260	32	21	4,416E-05	-0,608
9	GO:0030016	myofibril	182	4	18	7,290E-05	2,170
9	GO:0044449	contractile fiber part	184	4	18	8,301E-05	2,170
10	GO:0030017	sarcomere	181	4	18	1,383E-05	2,170
10	GO:0036379	myofilament	86	1	10	4,107E-04	3,322
11	GO:0005865	striated muscle thin filament	55	1	9	1,674E-04	3,170

Pathways	Sequence	BLAST result	Enzyme codes	Function
		Gregarious		
Pyrimidine metabolism	Dm138808	UMP-CMP kinase	EC:2.7.4.14	kinase
	Dm15751	DNA polymerase delta catalytic subunit	EC:2.7.7.7	DNA polymerase
	Dm1760	Thioredoxin reductase-1	EC:1.8.1.9	reductase
	Dm99307	Ectonucleoside triphosphate diphosphohydrolase 5	EC:3.6.1.6	diphosphate phosphatase
	Dm109353	DNA-directed RNA polymerases I, II and III subunit RPABC2	EC:2.7.7.6	RNA polymerase
	Dm123977	DNA-directed RNA polymerase III subunit RPC1	EC:2.7.7.6	RNA polymerase
	Dm2879	DNA-directed RNA polymerases I, II and III subunit RPABC5	EC:2.7.7.6	RNA polymerase
	Dm42941	DNA-directed RNA polymerases I, II, and III subunit RPABC3	EC:2.7.7.6	RNA polymerase
	Dm160182	RNA polimerase hdc06513	EC:2.7.7.6	RNA polymerase
	Dm171332	Ubiquitin-like protein 5	EC:2.7.7.6	RNA polymerase
	Dm16114	Carbamoyl-phosphate:L-aspartate carbamoyltransferase	EC:2.1.3.2	carbamoyltransferase
			EC:3.5.2.3	carbamoylaspartic dehydrase
			EC:6.3.5.5	synthase (glutamine-hydrolysing)
		Solitarious		
Alanine, aspartate and glutamate metabolism	Dm150142	Serine pyruvate aminotransferase	EC:2.6.1.44	transaminase
	Dm184823	Argininosuccinate synthase	EC:6.3.4.5	synthase
	Dm197294	Protein N-terminal asparagine amidohydrolase isoform X1	EC:3.5.1.1	asparaginase II
	Dm36642	Succinate-semialdehyde dehydrogenase [NADP(+)] GabD	EC:1.2.1.24	dehydrogenase (NAD+)
			EC:1.2.1.16	dehydrogenase [NAD(P)+]
	Dm44811	Glutamate oxaloacetate transaminase 1, isoform B	EC:2.6.1.1	transaminase
	Dm155754	Glutamate synthase 1 [NADH]	EC:1.4.1.14	synthase (NADH)
			EC:1.4.1.13	synthase (NADPH)
Fatty acid biosynthesis	Dm116127	Acetyl-CoA carboxylase, isoform A	EC:6.4.1.2	carboxylase
	Dm69106	Fatty acid synthase 1, isoform A	EC:3.1.2.14	hydrolase
	Dm85720	4-coumarate–CoA ligase 1	EC:2.3.1.86	synthase
			EC:6.2.1.3	ligase
Phenylalanine metabolism	Dm117575	Mitochondrial trifunctional protein alpha subunit, isoform A	EC:4.2.1.17	hydratase
	Dm953	Peptidoglycan-recognition protein-lb isoform 2	EC:3.5.1.4	acylamidase
	Dm96945	tyrosine decarboxylase 2	EC:4.1.1.28	decarboxylase
	Dm44811	Glutamate oxaloacetate transaminase 1, isoform B	EC:2.6.1.1	transaminase
	Dm116199	Tyrosine aminotransferase	EC:2.6.1.5	transaminase
Tyrosine metabolism	Dm49965	Fumarylacetoacetase	EC:3.7.1.2	beta-diketonase
	Dm96945	tyrosine decarboxylase 2	EC:4.1.1.28	decarboxylase
			EC:4.1.1.25	decarboxylase
	Dm36642	Succinate-semialdehyde dehydrogenase [NADP(+)] GabD	EC:1.2.1.16	dehydrogenase [NAD(P)+]
	Dm44811	Glutamate oxaloacetate transaminase 1, isoform B	EC:2.6.1.1	transaminase
	Dm116199	Tyrosine aminotransferase	EC:2.6.1.5	transaminase

Table 3.25: KEGG pathways where the differentially expressed transcripts are included. The fields of each column represent the sequence name, simplified BLAST result, enzyme code, KEGG pathways where the enzyme is involved, main function of the enzyme and the phase during which it is over-expressed.

Table 3.26: Sets of differentially expressed transcripts with common functions in *S. gregaria* DT transcriptome. Differential expression was considered at FDR < 0.05.

Name	Accesssion number	Sequence length	Solitarious FPKM	Gregarious FPKM
Juvenile hormone regulation and metabolism				
JHBP EFN75367	EFN75367	994	0,298	9,460
50MGP	AGT15837	9785	9,799	0,482
JH acid methyltransferase	BAC98836	11841	14,098	11,508
JHBP AAL48609	AAL48609	29714	2,719	0,715
JHBP AAM29553	AAM29553	6528	8,669	0,090
JHBP AAY55827	AAY55827	4010	3,058	0,074
JHBP ABL75751	ABL75751	8671	0,375	1,627
JHBP XP_004928501	XP_004928501	1555	75,410	37,936
Methoprene-tolerant receptor	AHA44478	11207	9,930	7,053
Vitellogenin	AFW97644	324469	0,069	0,003
Gene expression regulation				
Chromodomain-helicase-dna-binding protein 1	Q7KU24	5723	1,034	5,586
ATP-dependent rna helicase ddx17	KDR17555	12282	67,503	90,469
Eukaryotic translation initiation factor 4	XP_002411156	1902	0,000	2,802
Eukaryotic initiation factor 4e-2	BAG30778	3675	3,380	10,832
Gly-tRNA ligase	XP_003700775	1678	1,586	8,219
Lsm12	XP_003702748	664	0,000	15,105
Nucleolin	XP_003395590	748	0,000	11,314
RNA polimerase hdc06513	DAA02521	4426	1,337	6,940
tRNA methyltransferase (h)	EFX87167	6824	2,861	8,497
tRNA-pseudouridin synthase	ABV82369	13875	0,128	1,130

Name	Accesssion number	Sequence length	Solitarious FPKM	Gregarious FPKM
Ubiquitin-like protein 5	Q9V998	13544	16,073	22,795
Vigilin	AAD34767	39570	5,158	9,141
Peritrophic matrix components				
Cuticular protein 49aa	NP_001097285	11381	17,872	11,610
Cuticular protein 49ag	NP_610776	6789	18,106	4,040
Cuticular protein Dm172325	AAV36951	913	11,019	0,323
Cuticular protein Dm21177	AAK77312	3344	9,965	1,940
Larval cuticle protein 352	XP_001861377	1092	28,789	0,270
Peritrophin 2936	XP_003702718	17059	23,091	6,069
Peritrophin 90825	ETN65702	3423	30,033	2,844
Peritrophin Dm106757	NP_728732	9841	3,929	1,618
Peritrophin Dm135168	CBA35308	993	1,489	21,148
Peritrophin Dm26312	AAK93496	1650	2,667	0,179
Peritrophin-1	KDR14645	715	7,035	0,000
Immune response				
Apolipophorin III	AAM48468	2519	2,466	11,572
C-type lectin	AAS93748	3394	7,410	0,087
C-type lectin 8	EHJ72392	1721	0,000	3,278
Defensin	AIL24687	753	115,648	57,188
Defensin 1	AGE89781	5728	191,472	103,422
Ganglioside gm2 activator	KDR17702	689	0,000	23,201
GNBP1	CAJ18915	7379	0,000	1,657
IbpA	EFX75568	4523	75,067	62,146
Ig-like 143507	KDR22486	12254	14,341	9,941
Ig-like Dm22995	AAK92995	26275	10,706	16,331
Lectin 33A	NP_001097145	1848	16,672	3,192

Name	Accesssion number	Sequence length	Solitarious FPKM	Gregarious FPKM
Lectin-related protein	BAA18916	1643	13,394	0,180
Locustin	P83428	6778	81,778	199,587
LPS-binding protein KDR16864	KDR16864	4376	1,149	4,298
LPS-binding protein XP_003708137	XP_003708137	1273	0,000	15,512
Leucine-rich repeat-containing protein 70	KDR13390	7499	25,699	20,689
Pacifastin-related	CAD24808	8742	4,783	0,101
Pacifastin-related 4t	CAC82510	13210	4,859	13,809
PGRP sb2	CAD89142	5974	2,315	5,233
PGRP sc1a	Q9V3B7	3575	6,068	9,900
PGRPsc1b	CAD89167	3419	3,770	1,208
PGRP-lb	AAG23732	11151	5,977	2,830
Serpin 3	O46163	6907	9,892	19,559
Spaetzel-processing enzyme	AAK93062	8176	0,036	6,709
Vesicular mannose-binding lectin	DAA34055	1905	8,229	18,592
Oxidative stress				
ATP-dependent -NAD(H)-hydrate dehydratase	Q9VVW8	11510	1,902	3,867
Gamma-glutamyl cyclotransferase	AAM50196	9053	4,063	1,824
Peroxiredoxin-2-like	XP_008482365	394	45,043	116,943
Thioredoxin	EFA07249	1393	1,486	9,675
Thioredoxin like protein/Prostaglandin synthase	ETE68771	634	0,000	15,326
Thioredoxin reductase-1	AAG25639	29291	1,151	2,814
Vanin-like	XP_002428166	10746	15,827	10,595
Ubiquitylation				
Apc11 anaphase-promoting complex subunit	CAB63945	13943	7,446	21,019
E1 ubiquitin-activating protein AOS1	CAL26275	2218	2,934	11,305
E3 ubiquitin-protein ligase AMFR-like	XP_003707525	484	121,612	251,917

Name	Accesssion number	Sequence length	Solitarious FPKM	Gregarious FPKM
E3 ubiquitin-protein ligase RNF181 homolog	Q9VE61	45020	3,298	4,644
FAM152b	EFN65714	7845	4,901	11,067
F-box only protein	AAM27519	5170	17,330	8,672
Ubx domain-containing protein 6-like	XP_003397465	1099	8,881	58,752
WWE domain containing protein	AAZ41787	3355	0,000	2,149
Apoptosis				
Ubiquitin-like modifier-activating enzyme ATG7	EZA62304	8212	0,072	0,992
BAX inhibitor 1	Q9VSH3	38430	12,561	17,568
dual specificity protein phosphatase CDC14A	EFN73005	4269	2,702	6,608
cytochrome c oxidase subunit partial	AFC37464	16022	0,646	2,387
cytochrome c oxidase assembly factor 6	NP_001138136	4855	11,880	28,729
FAM32A	AGM32593	1643	2,160	12,591
TP53RK-binding protein	KDR16589	3652	5,682	1,858
Vesicle formation and mobility				
ADP-ribosylation factor-like protein 2	XP_394559	2855	4,455	15,590
Bet1	XP_004535540	1665	6,395	18,825
Complexin	XP_001628076	1911	1,857	8,037
Coatomer subunit protein beta	P45437	59156	0,420	1,208
GTPase activating protein for ARF	AAM11391	12605	3,754	9,797
Sec1	AAM51046	25616	2,921	6,253
SSR-beta	AAD46863	9115	0,195	1,891
Tetraspanin CD63	XP_008180318	3358	0,000	3,454
Vacuolar protein sorting isoform a	NP_611651	30015	0,118	0,449
Vacuolar protein sorting-associated protein 35	AAL28782	19384	0,687	1,682
Nervous system related transcripts				
Actin-like protein 87c	P45889	33168	2,622	4,838

Name	Accesssion number	Sequence length	Solitarious FPKM	Gregarious FPKM
Calmodulin dependent NO synthase	2122379A	14386	3,290	5,687
Clavesin-1	AAL29003	2869	0,309	10,707
Early growth response 2	AHB17950	6529	40,543	30,042
Fasciclin	NP_001138065	21264	3,430	1,900
FRP domain containing protein	ABC86248	39172	3,483	2,417
GABA receptor associated protein GRAP	AAM52664	27282	19,244	27,838
Ly-6/neurotoxin 2	XP_002430283	7108	41,442	33,362
Ninjurin	NP_001138042	8428	6,229	3,745
Nischarin	XP_001945341	2644	1,007	4,386
NMRP	ACO72861	4533	0,522	3,181
Protocadherin beta-15-like	XP_008216384	5255	6,754	12,406
Potassium voltage-gated channel protein SHAL	P17971	14174	1,607	5,307
Slit	AAD26567	2289	14,970	8,247
Spondin-1	KDR20370	6853	4,266	1,722
Tetraspanin-1	XP_967238	3120	25,594	3,498
Calcium regulation				
Calmodulin	ACO12252	3312	7,234	20,158
Calmodulin-related protein	2021248D	303	0,000	24,827
Casein kinase II subunit alpha	P08181	27703	1,110	1,788
Casein kinase II subunit beta	P08182	14912	3,213	7,483
I3P kinase	XP_001952324	6671	13,523	30,775
MIR domain containing	AAM50667	15740	2,631	6,352
Nucleoredoxin	XP_002425676	6504	38,428	86,022
Phospholipase C	XP_308977	28796	0,555	1,557
Phospholipase C domain containing	KFM70036	12907	12,558	34,411
Preprotachykinin	AAX11212	5091	2,033	5,849

Name	Accesssion number	Sequence length	Solitarious FPKM	Gregarious FPKM
Trophoblast glycoprotein	KDR23388	3651	0,000	2,747

Table 3.27: List of transcripts from the comparative study between our RNA-seq data on the *S. gregaria* digestive tube and other works on locust phase change. Transcripts with congruent expression pattern between the cited work and ours are highlighted in bold.

	Expression pattern		DT transcriptome data				
Transcript name	Scientific publication	S. g. DT RNA-seq	S. g. DT RNA-seq sequence ID	Accession number	Sequence length	Solitarious FPKM	Gregarious FPKM
			Kang et al. [2004]				
similar to brain adenylate cyclase 1	Gregarious	N. S.	159478	XP_008560532	361	0,001	0,001
annexin IX	**Gregarious**	**Gregarious**	150444	XP_008471197	681	3,231	15,159
similar to asparagine synthetase	Solitarious	N. S.	149268	CDO39383	870	0,001	0,001
cytochrome c oxidase chain 3	Gregarious	N. S.	Dm156703	ACI28586	8405	1,571	1,158
larval cuticle protein precursor	**Solitarious**	**Solitarious**	352	XP_001861377	1092	28,789	0,270
NADP-dependent malic enzyme	Gregarious	N. S.	77893	AGC84405	2674	0,705	0,110
troponin C	**Solitarious**	**Solitarious**	Dm19668	P47949	5297	4,451	1,615
beta-N-acetyl-hexosaminidase activity	Gregarious	Solitarious	Dm80034	AAM48390	8138	1,893	1,051
putative fatty acid elongase	Gregarious	N. S.	Dm175109	CAI40769	6193	0,406	0,524
			Chen et al. [2010]				
serotonin receptor	Gregarious	N. S.	Dm3176	P20905	821	0,001	0,001
allatostatin receptor	Gregarious	N. S.	Dm179570	AAF05299	146	0,001	0,001
blot	Gregarious	N. S.	92395	AHH29252	351	0,001	0,001
cask	Gregarious	N. S.	Dm35315	NP_001097863	10588	0,356	0,306
fmr1	Gregarious	N. S.	Dm191726	Q9VF87	18488	0,085	0,064
diuretic hormone receptor	Gregarious	N. S.	58789	EGI58221	901	0,000	0,655
GABA receptor	Gregarious	N. S.	Dm136186	Q08832	601	7,323	2,454
glutamate receptor	Gregarious	N. S.	15004	XP_001655465	336	0,001	0,001
malvolio	Gregarious	N. S.	Dm153651	NP_996251	2862	8,568	9,584
neurotransmitter transporter	Gregarious	N. S.	Dm78496	CAA69649	3952	0,318	0,299
octopamine receptor	Gregarious	N. S.	16956	XP_001869400	545	2,307	0,001
synaptic vesicle 2-related protein	Gregarious	N. S.	134872	KDR17464	276	0,001	0,001
synaptic vesicle protein	Gregarious	Solitarious	226	XP_001658023	1742	50,531	24,890
synaptobrevin	Solitarious	N. S.	68583	EHJ77205	285	0,000	1,035
tyrosine hydroxylase	**Gregarious**	**Gregarious**	Dm176256	AAA62877	7096	1,595	4,988
			Guo et al. [2011]				
alcohol dehydrogenase	Gregarious	N. S.	81868	BAN21280	307	1,024	10,568
Alpha crystallin	Solitarious	N. S.	151152	XP_008475973	372	0,845	2,379
Peptidase M2, peptidyl-dipeptidase A	Gregarious	N. S.	169762____k43_f_13402665	EAW94320	1260	0,001	0,001
arylphorin hexamerin-like protein 2	Solitarious	N. S.	4419	AAX14951	5765	0,055	0,000
ATP-citrate synthase	Gregarious	N. S.	65031	XP_003425261	326	0,001	5,429
ATP-dependent RNA helicase p62	Solitarious	N. S.	135204	KDR17554	1511	0,208	0,000
cellular retinaldehyde-binding protein	Solitarious	N. S.	Dm60670	NP_523939	623	4,037	0,000

	Expression pattern		DT transcriptome data				
Transcript name	Scientific publication	S. g. DT RNA-seq	S. g. DT RNA-seq sequence ID	Accession number	Sequence length	Solitarious FPKM	Gregarious FPKM
protein TU-36B	**Solitarious**	**Solitarious**	Dm186680	P19967	13242	0,641	0,178
cystathionine gamma-lyase	Gregarious	N. S.	67754	EHJ71352	585	0,000	0,504
Fatty acid desaturase	Solitarious	N. S.	Dm11263	BAB21537	2699	1,165	0,984
lethal 2 essential for life protein	**Solitarious**	**Solitarious**	Dm15247	P82147	23547	1,789	0,927
lambda-crystallin homolog	Gregarious	N. S.	199	XP_001601340	966	0,001	0,001
takeout 1	Solitarious	N. S.	6593	KDR17338	932	13,155	8,228
purine nucleoside phosphorylase	Solitarious	N. S.	73655	BAM18213	401	0,001	0,001
GTP-binding protein Rheb homolog	Solitarious	N. S.	Dm204957	Q9VND8	18065	0,157	0,278
cytochrome p450 4g15	**Solitarious**	**Solitarious**	15798	ACA04895	495	25,404	1,788
cytochrome p450 6a2	**Solitarious**	**Solitarious**	Dm12701	P33270	145270	1,026	0,422
probable cytochrome p450 4ac1	Solitarious	N. S.	Dm4355	Q9VMS9	3197	0,197	0,000
Heat shock protein Hsp20	**Solitarious**	**Solitarious**	2738	AEV89751	3166	6,454	0,373
peptidyl-prolyl cis-trans isomerase	**Solitarious**	**Solitarious**	Dm177758	AAX33414	29678	1,737	0,865
Badisco et al. [2011a]							
slit homologue	Gregarious	Solitarious	Dm164661	AAD26567	2289	14,970	8,247
RNA helicase ddx1	Solitarious	N. S.	Dm112949	Q9VNV3	26703	0,094	0,133
fasciclin-like precursor	Gregarious	Solitarious	Dm91568	NP_001138065	21264	3,430	1,900
sparc	Gregarious	N. S.	90139	AAV88596	2188	0,001	0,001
Badisco et al. [2011b]							
Pasilla	Gregarious	N. S.	Dm5593	AAG36788	11802	1,678	1,849
arrestin	Gregarious	N. S.	Dm187172	AAF32365	13268	4,715	3,624
Cullin-3	Gregarious	N. S.	151333	XP_008477011	385	0,001	0,001
Glia Maturation Factor	Solitarious	N. S.	24945	XP_002427843	302	0,001	0,001
Slowpoke homologue	Solitarious	N. S.	Dm179850	NP_001014656	3074	1,227	0,864
mitochondrial ATP-synthase coupling factor 6	Solitarious	N. S.	90614	ETN60559	796	0,000	0,371
Pacifastin-like peptide precursor 4	**Gregarious**	**Gregarious**	90431	CAC82510	13210	5,164	12,995
5-oxoprolinase	Solitarious	N. S.	70450	XP_003699466	125	0,001	0,001
ribophorin	Solitarious	N. S.	275	XP_001647637	147	0,001	0,001
transaldolase homologue	Solitarious	N. S.	Dm173384	Q9W1G0	34011	0,092	0,217
thaumatin-like protein	Gregarious	Solitarious	3321	AAS83110	6035	0,365	2,346
transient receptor potential-like	Gregarious	N. S.	139128	KDR19828	706	3,562	1,253
Osa	Gregarious	N. S.	160487	XP_008551463	582	2,161	4,561
Musashi	Gregarious	N. S.	Dm196699	Q9VVE5	21855	1,280	0,850
chromodomain helicase DNA binding protein	**Gregarious**	**Gregarious**	488622_____313405	XP_005234509	749	2,518	17,721
microsomal glutathione-S-transferase	Solitarious	Gregarious	90269	AHC08054	10276	4,742	7,578

Transcript name	Expression pattern		DT transcriptome data				
	Scientific publication	S. g. DT RNA-seq	S. g. DT RNA-seq sequence ID	Accession number	Sequence length	Solitarious FPKM	Gregarious FPKM
peroxiredoxin	Solitarious	Gregarious	152968	XP_008482365	394	47,875	110,045
Wang et al. [2014b]							
Aladin	Solitarious	N. S.	137865	KDR19153	1136	1,384	0,779
Ankyrin-2	Solitarious	N. S.	122457	KDR10348	10182	8,059	6,141
Basement membrane-specific heparan sulfate proteoglycan core protein	Gregarious	N. S.	114214	XP_008193278	984	0,000	0,300
Ribosome biogenesis protein BOP1 homolog	Solitarious	N. S.	Dm9807	Q7K0Y1	35142	0,555	0,369
Brefeldin A-inhibited guanine nucleotide-exchange protein 3	Solitarious	N. S.	144811	KDR23060	912	0,345	0,000
Dynein heavy chain, cytoplasmic	Gregarious	N. S.	75468	XP_003742362	419	0,000	4,928
Early endosome antigen 1	Solitarious	N. S.	120298	KDR08560	1018	0,000	0,579
Endophilin-B1	Solitarious	N. S.	253368	KFM81156	1037	0,303	1,138
Eukaryotic translation initiation factor 4B	Solitarious	N. S.	145099	KDR23229	5329	14,512	14,944
GRIP and coiled-coil domain-containing protein 1	Solitarious	N. S.	81689	BAN20798	372	0,845	3,171
HEAT repeat-containing protein 5B	Gregarious	N. S.	150384	XP_008470740	1047	1,501	1,409
E3 ubiquitin-protein ligase hyd	Gregarious	N. S.	Dm186987	P51592	45892	1,103	1,208
Kinesin-associated protein 3	Solitarious	N. S.	160159	XP_008547973	888	0,354	0,000
Glycyl-tRNA synthetase	Gregarious	N. S.	2499	XP_003399968	687	0,001	0,001
Sterile alpha and TIR motif-containing protein 1-like	Solitarious	N. S.	70083	XP_003698758	113	0,001	0,001
mitogen-activated protein-binding protein-interacting protein	Gregarious	N. S.	137398	KDR19052	401	0,001	0,001
Niemann-Pick C1 protein	Gregarious	N. S.	148171	KDR24042	850	0,740	0,000
NUAK family SNF1-like kinase 1	Solitarious	N. S.	132349	KDR16310	18368	0,205	0,209
Phosphatidylinositol-4, 5-bisphosphate 3-kinase catalytic subunit delta isoform	Solitarious	N. S.	140899	KDR21011	692	0,909	1,279
Phosphatidylinositol glycan anchor biosynthesis class U protein	Gregarious	N. S.	6441	KDR16382	961	0,000	0,307
Phosphatidylinositol-4-phosphate 5-kinase type-1 alpha	Gregarious	N. S.	48449	EFN84793	622	0,000	1,897
Polyhomeotic-like protein 1	Gregarious	N. S.	60868	EGI63416	195	0,001	0,001
Polyphosphoinositide phosphatase	Gregarious	N. S.	128825	KDR14022	1368	0,230	0,000

	Expression pattern		DT transcriptome data				
Transcript name	Scientific publication	S. g. DT RNA-seq	S. g. DT RNA-seq sequence ID	Accession number	Sequence length	Solitarious FPKM	Gregarious FPKM
Probable protein-cystein N-palmitoyl-transferase porcupine	Gregarious	N. S.	Dm195333	Q9VWV9	647	0,972	0,000
PR domain zinc finger protein 4	Solitarious	N. S.	135233	KDR17559	2676	1,645	1,433
Uncharacterized protein C10orf118 homolog	Solitarious	N. S.	5048	XP_008201271	1390	0,679	0,424
Probable phospholipid-transporting ATPase IA	Solitarious	N. S.	75839	XP_003744036	1394	1,353	1,693
Protein Daple	Solitarious	N. S.	134610	KDR17251	2049	0,460	0,576
protein FAN	Gregarious	N. S.	26467	XP_002432451	299	0,000	4,932
Rho GTPase-activating protein 190	Solitarious	N. S.	Dm193307	Q9VX32	11302	0,334	0,183
RING finger protein unkempt	Solitarious	Gregarious	21075	ACO12796	61194	2,959	5,659
RNA exonuclease 1 homolog	Solitarious	N. S.	120262	KDR08483	2805	6,613	6,414
SAP domain-containing ribonucleoprotein	Solitarious	N. S.	149444	XP_008475226	2410	1,044	0,612
Serine/threonine-protein kinase mTOR	Gregarious	N. S.	118542	KDR06409	588	0,001	0,001
Serine–pyruvate aminotransferase, mitochondrial	Solitarious	N. S.	81320	XP_004533730	4151	0,227	0,071
Protein son of sevenless	Solitarious	N. S.	Dm197965	AAB04680	8619	0,292	0,034
Telomerase-binding protein EST1A	Solitarious	N. S.	5608	KDR09736	1423	0,000	0,207
Tubulin glycylase 3B	Gregarious	N. S.	123735	KDR10777	163	0,001	0,001
Ubiquitin-protein ligase E3B	Gregarious	N. S.	116753	XP_001602271	1133	0,000	0,260
Probable uridine-cytidine kinase	Solitarious	N. S.	Dm74416	Q9VC99	7189	0,175	0,205
Unc-112-related protein	Gregarious	N. S.	Dm159460	Q9VZI3	13308	0,001	0,001
Zinc finger protein 91	Solitarious	N. S.	62350	EGI68339	433	6,534	0,681

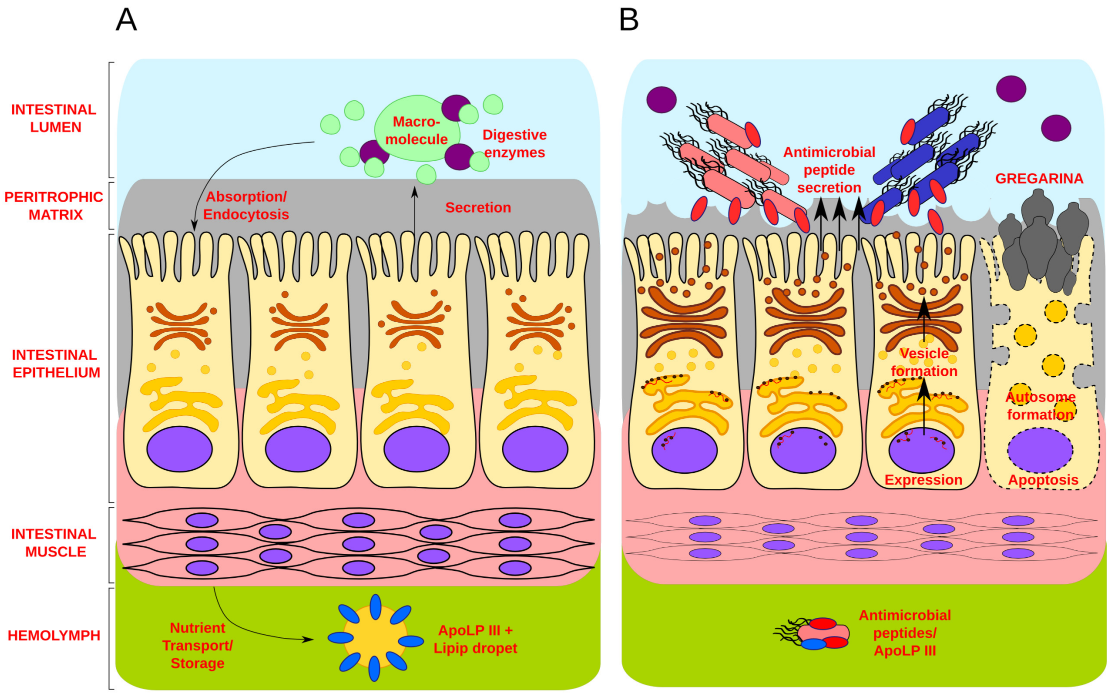

Figure 3.36: Simplified representation of our inferred summary of the differences highlighted by our RNA-seq analysis of the transcriptome from solitarious and gregarious *S. gregaria*'s digestive tube. A. Digestive tube of the locust during the solitarious phase, with regular activities, such as secretion of digestive enzymes (e.g., trypsin) and peritrophic matrix components (e.g., peritrophin and cuticular proteins) taking place. B. Digestive tube of the locust during the solitarious phase.

Chapter 4: Comparatiev study between CNS and DT transcriptomes from *S. gregaria*

4.1 Introduction

In order to search for candidate genes and validate our transcriptomical results, in this chapter we compare the central nervous system (CNS) and digestive tube (DT) expression profiles in gregarious and solitarious phase from *S. gregaria*. We first focus on the comparison of unique BLAST results between tissues and their associated expression profiles to find transcripts that share their expression pattern between transcriptomes. We also clustered the assembled transcripts without BLAST result from both transcriptomes in order to obtain a set of homologous unknown transcripts. To compare the results of our study with other related studies, we compile expression data from other publications regarding locust transcriptomics from *S. gregaria* and *L. migratoria*, and stablish a list of transcripts with akin expression patterns, as well as transcripts with different results.

4.2 Material and Methods

We extracted all the accession numbers from the BLAST annotations of the sequences of our *de novo* assembled CNS and DT reference transcriptomes and gathered a non-redundant list of unique entries for each of them. We compared the two lists using the 'fgrep' unix command ('*fgrep –w –f CNS_list DT_list*') in order to extract the list of shared accession numbers between both transcriptomes. We also used the commands '*fgrep –w –v –f CNS_list DT_list*' and '*fgrep –w –v –f DT_list CNS_list*' in order to extract the accession numbers that only appear in one of the two reference transcriptomes. We then calculated the solitarious and gregarious CNS and DT **F**ragments **P**er contig **K**ilobase and **M**illion mapped fragments in the sequencing library (FPKM) values for the annotated sequences in the shared list of BLAST accession numbers. After calculating the fold change (FC) as the 2-based logarithm of the ratio between the FPKM of the solitarious and gregarious libraries of each tissue, we could check whether the differential expression of the shared genes between solitarious and gregarious states was congruent in both transcriptomes. Four groups of accession numbers with significant over-expression pattern were thus established: congruent (significant over-expression in the same sense in both transcriptomes), incongruent (significant over-expression in opposite senses in the two transcriptomes), inconsistent (significant over-expression in one transcriptome but not in the other), and non-significant (no sign of statistically significant over-expression in any of the two transcriptomes).

Both transcriptomes contain a high number of sequences with no known annotation (henceforth unknown sequences). To deal with the challenge of comparing them, we started by building local BLAST databases of the annotated sequences in each transcriptome and carried out a pairwise reciprocal BLASTn analysis of the sequences that have the same BLAST annotation in both transcriptomes. We then extracted the BLAST identity values between query and database sequences and calculated the mean and the minimum identity values between sequences that previously were found to have the same annotation in both transcriptomes. We then used the mean identity value as threshold for the algorithm used to cluster groups of unknown sequences—here we assume that the annotated sequences have an overall similar degree of conservation as the unknown sequences. For such clustering, we used CD-HIT [Li and Godzik, 2006] on a fasta-formatted file containing the unknown sequences of both CNS and DT reference transcriptomes and applied the mean identity between sequences of the same annotation as minimum identity threshold for grouping unknown sequences in a cluster (the command was '*cd-hit-est –i input.fasta –c 0.98 –d 50 –o output.fasta*'— -c 0.98 indicates the identity threshold and -d 50 indicates the number of characters to place in the output from fasta headers (to avoid incomplete fasta header names that would affect downstream sequence analyses). This way we obtained clusters and singlets (sequences which do not belong to any cluster under the identity threshold used). The clusters of CNS and DT unknown sequences (designated as shared clusters) identified the unknown sequences that were shared between both transcriptomes, while the singlets were specific to one or the other transcriptome. To obtain data on the differential expression of the sequences of each cluster between solitarious and gregarious states in each transcriptome we wrote several Unix-based command lines (using combinations of ‘split’, ‘cut’, ‘sed’, ‘grep’ and ‘awk’ commands) in order to retrieve the counts of NGS reads that mapped to each sequence. The fragment counts were summed for each group of sequences that belong to the same cluster, state and tissue, which resulted in four groups of fragment counts per cluster: solitarious CNS counts, gregarious CNS counts, solitarious DT counts and gregarious DT counts. We also calculated the FC between the solitarious and gregarious libraries of each tissue as before. We then grouped the clusters of unknown sequences based on the congruency of their over-expression as we did with the annotated sequences. Since sequence grouping based on identity might be sensitive to sequence length, we repeated the same analysis excluding all the unknown sequences that are shorter than 500 bases in order to detect any possible bias that might emanate from such effect.

In order to compare our data with other studies on locusts’ phase change, we gathered information and elaborated lists of genes that were reported in eight published works [Kang et al., 2004, Chen et al., 2010, Badisco et al., 2011a,b, Guo et al., 2011, Zhang et al., 2012, Wang et al., 2014b, Spit et al., 2016]. The published data on the presence, absence, abundance and/or expression pattern of these genes were compared against our CNS and DT transcriptomic data, depending on the scope of the published study. In this chapter, we will focus on the comparison between the three components (bibliography and CNS and DT transcriptomes), since the comparioson between CNS and DT transcriptomes against bibliography is already included in former chapters. The methodology is analogous to that used in Chapters 2 and 3, but comparing three data sets instead of two. To compare the expression of validated genes between species, qPCR validated genes in *S. gregaria* adults and nymphs were also checked via qPCR in *L. migratoria* adults and nymphs following the protocol described in the general methodology chapter of this memory.

4.3 Results

We extracted 16,749 and 57,637 unique accession numbers from the BLAST results of the *de novo* assembled sequences of the *S. gregaria* CNS and DT transcriptomes, respectively. 2,772 BLAST accession numbers where shared between the two transcriptomes whereas 13,977 (CNS) and 54,865 (DT) were not. The sequences with shared BLAST accession numbers show very different expression profiles between transcriptomes (table 4.28). In the CNS transcriptome we detected 1,789 sequences with significant differential expression between solitarious and gregarious locusts and 983 sequences with no significant expression difference between the two locust states. In the DT transcriptome, however, the results seemed the other way around as we detected only 240 sequences with significant differential expression while the sequences with no significant differential expression were 2,532.

About a third (987) of the sequences with shared BLAST result also shared the same pattern of expression levels between transcriptomes, although 922 of them are not significantly over-expressed in any locust phase. Only 43 and 22 accession numbers show congruent gregarious and solitarious over-expression, respectively (tables 4.29, 4.36 and 4.37). 110 shared BLAST accession numbers show significant over-expression of opposite solitarious vs. gregarious direction in both tissues (tables 4.38 and 4.39). It is therefore worth highlighting that the overall expression pattern of the shared annotated genes differs between tissues, with nearly 90 % of the shared annotated genes that show significant differential expression between locust phases in the CNS transcriptome being over-expressed in the gregarious phase, whereas over-expression in the solitarious phase is the predominating tendency in the DT transcriptome (table 4.28 and figure 4.37). Accordingly, when we use congruency of the expression pattern between tissues in order to classify the annotated transcripts whose BLAST accession numbers appear both in the CNS and DT transcriptomes (shared) and whose differential expression between solitarious and gregarious locusts is significant in both transcriptomes, we observe that most of these genes —about two thirds— are over-expressed in the gregarious phase in the CNS and in the solitarious phase in the DT.

	Number of sequences	Solitarious DET	Gregarious DET	Not Significant
CNS DETs	1,789	128	1,661	983
DT DETs	240	168	72	2,532

Table 4.28: Distribution of the shared genes between CNS and DT reference transcriptomes based on their differential expression between solitarious and gregarious locusts.

	CNS Solitarious	CNS Gregarious	CNS Non-Significant
DT Solitarious	22	101	45
DT Gregarious	9	43	20
DT Non-Significant	97	1,517	918

Table 4.29: Congruency of the shared differentially expressed transcripts between locust phases and tissues.

The pairwise BLASTn-based comparative study between sequences that had the same BLAST annotation in both transcriptomes gave a mean sequence identity value of 97.557 % (maximum value of 100 %, minimum value of 72.7 %). We merged the fasta files of the 34,696 and 16,491 sequences that had no known BLAST annotation in the CNS and DT transcriptomes, respectively (see chapters 2 and 3), and used that identity value as cut-off sequence identity

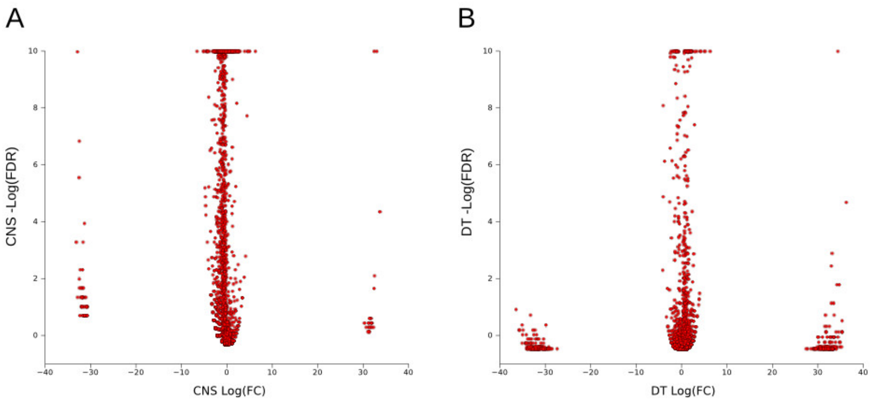

Figure 4.37: Volcano plots of the gregarious and solitarious expression levels of the transcripts that share BLAST accession number between the *S. gregaria* CNS and DT transcriptomes. The X axis represents the 2-based logarithm of the fold change in expression level between gregarious and solitarious transcriptomes and the Y axis represents the negative 10-based logarithm of the False Discovery Rates (FDR) associated with the fold change in gene expression level. A: *S. gregaria* adult CNS transcriptome. B: *S. gregaria* adult DT transcriptome.

threshold for building clusters of unknown sequences in both transcriptomes (since CD-HIT allows only two decimals we used 0.98 as threshold). The resulting 44,440 clusters were composed of 38,974 'single sequence clusters' (singlets) and 5,466 multi-sequence clusters (table 4.30). Almost all the clusters (5,435) contained both CNS and DT transcripts. Barely 31 clusters were composed of only CNS sequences and none had only DT sequences. 2,374 multi-tissue clusters showed congruent direction of differential expression between locust phases in the two transcriptomes, 1,202 over-expressed in the solitarious phase and 1,172 in the gregarious (table 4.31). When the clustering analysis was restricted to sequences that are larger than 500 bases it casted a total of 21,015 clusters, 18,114 of them containing singlets and 2,901 multi-sequence. 2,872 multi-sequence clusters were also multi-tissue and 29 clusters had CNS sequences only. The number of multi-tissue clusters that show congruent direction of significant over-expression between tissues was 1,411, 630 over-expressed in the solitarious phase and 781 in the gregarious one (table 4.32). The mean number of sequences in a cluster of anonymous sequences was 2.234 for sequences equal or larger than 100 bases and 2.105 for sequences equal or larger 500 bases, respectively. The ratio between the number of clustered and singlet anonymous sequences was 31.34 % for sequences $\geq$ 100 bases and 33.71 % for sequences $\geq$ 500 bases. For its part, the ratio between the number of clusters of anonymous sequences and the total number of anonymous sequences was 10.68 % for sequences $\geq$ 100 bases and 11.98 % for sequences $\geq$ 500 bases. 23.86 % and 25.21 % were the percentages of anonymous sequences that grouped into clusters for $\geq$ 100 and $\geq$ 500 bases, respectively.

Larger than 100 bases	Total	Singlets	Clusters
Number of clusters	44,440	–	5,466
Sequences involved	51,187	38,974	12,213

Larger than 500 bases	Total	Singlets	Clusters
Number of clusters	21,015	–	2,901
Sequences involved	24,220	18,114	6,106

Table 4.30: Clusters generated from the unknown sequences in the CNS and DT transcriptomes.

		Solitarious	Gregarious	DETs	Total
CNS DETs	Multi-tissue	2,449	2,909	5,358	5,435
	Specific	13	18	31	31
	All	2,462	2,927	5,389	5,466
DT DETs	Multi-tissue	708	1,987	2,695	5,435
	Specific	0	0	0	0
	All	708	1,987	2,695	5,466
CNS vs. DT	Congruent	1,202	1,172	2,374	5,435

Table 4.31: Distribution of the clusters of unknown sequences in the CNS and DT reference transcriptomes that are larger than 100 bases. DEGs: Significant differential expression between locust phases.

		Solitarious	Gregarious	DETs	Total
CNS DETs	Multi-tissue	1,126	1,718	2,844	2,872
	Specific	12	17	29	29
	All	1,138	1,735	2,873	2,901
DT DETs	Multi-tissue	1,438	1,184	2,622	2,872
	Specific	0	0	0	0
	All	1,438	1,184	2,622	2,901
CNS vs. DT	Congruent	630	781	1,411	2,872

Table 4.32: Distribution of the clusters of unknown sequences in the CNS and DT reference transcriptomes that are larger than 500 bases. DEGs: Significant differential expression between locust phases.

MA plots using the clusters of anonymous sequences confirmed that the clustering of sequences did not bias the tendency of the expression profiles compared to the overall CNS and DT transcriptomes. In all cases (CNS and DT clusters of sequences $\geq$ 100 or $\geq$ 500 bases, as well as the unclustered anonymous sequences), the comparison between the mean intensity and the fold change shows a near to 0 slope (figure 4.38). The correlations between the logarithm of the CNS fold change and that of the DT fold change show all the possible expression pattern states for the shared annotated sequences (the nine possible combinations between gregarious, solitarious and non significant differential expression categories between the levels in the CNS and DT), whereas the same correlations for the anonymous sequences show a biased tendency towards solitarious CNS (figure 4.39).

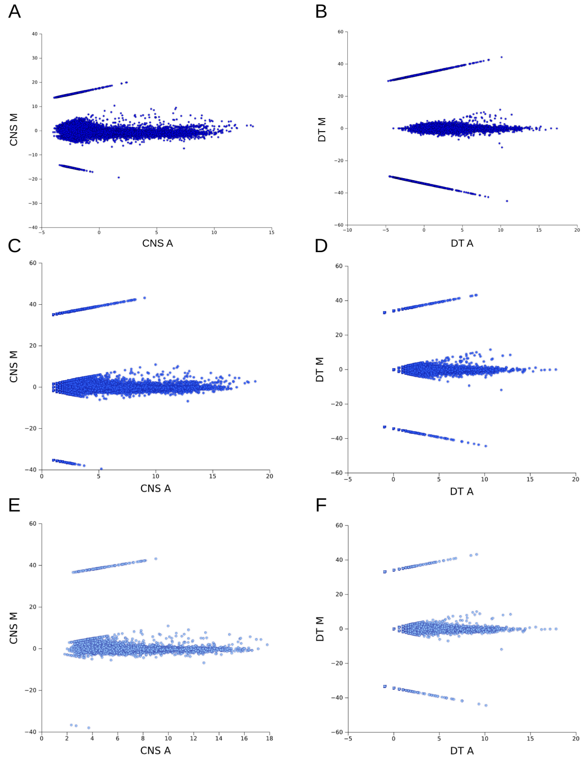

Figure 4.38: MA plots from the expression data. The X axis represents the logarithm of mean intensity (A) and the Y axis represents the logarithm of fold change (M). A, C, E: *S. gregaria* adult CNS transcriptome. B, D, F: *S. gregaria* adult DT transcriptome. A-B: Anonymous sequences before clustering. C-D: Clusters of anonymous sequences longer than 100 nucleotides. E-F: Clusters of anonymous sequences longer than 500 nucleotides.

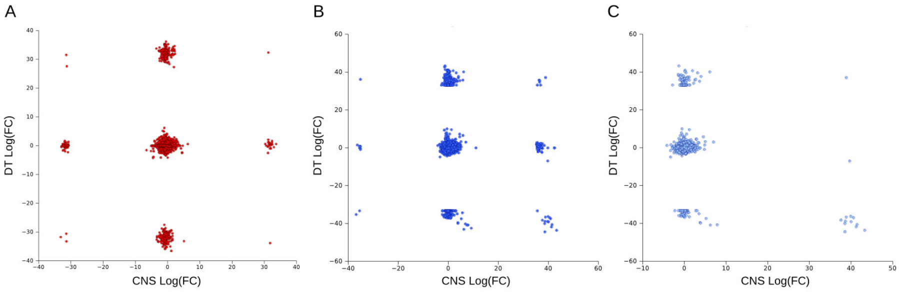

Figure 4.39: Correlation plots between the phase-related changes in the levels of gene expression in the CNS and DT. X axis represents the 2-base logarithm of the fold change in the gene expression levels between the gregarious and solitarious *S. gregaria* adult CNS transcriptomes and Y axis represents the 2-base logarithm of the fold change in the gene expression levels between the gregarious and solitarious *S. gregaria* adult DT transcriptomes. A: Transcripts sharing BLAST accession number between the two transcriptomes. B: Clusters of shared sequences with no known annotation that are larger than 100 nucleotides. C: Values from the clusters of shared sequences with no known annotation that are larger than 500 nucleotides.

Earlier we recompiled data on the expression tendencies reported for 124 genes in the literature on the phase change in *S. gregaria* and *L. migratoria*. 87 and 109 of these were present in our *S. gregaria* CNS and DT transcriptomes, respectively (table 4.40). We found that only 72 out of these 124 genes were present in the two transcriptomes, while 15 were present only in CNS and 37 transcripts were present only in DT. Only three transcripts (annexin IX and two pacifastin transcripts) were differentially expressed towards the gregarious phase in both our transcriptomes and in the bibliography, whereas none was differentially expressed in the solitarious phase in all the data sources. We also detected 10 transcripts which expression pattern was congruent. 7 showed higher expression in the gregarious phase and 3 in the solitarious one—although none of the differences was statistically significant in our transcripts (table 4.41). When comparing the list of shared accession numbers between transcriptomes with the accession numbers from the bibliographical data, we find that pacifastin (CAC82510.3) is the only gene present in both lists—it is over-expressed in the gregarious phase no matter the source of the data. This comparison of the expression tendencies, with regards to the locust phases, between our transcriptomics data and the data reported elsewhere for those genes (see table 2.17 in Chapter 2 and table 3.27 in Chapter 3) allowed us to further confirm the genes in table 9 as either (*i*) consistently over-expressed in the gregarious phase regardless of the tissue, species and work, (*ii*) consistently over-expressed in the solitarious phase regardless of the tissue, species and work, (*iii*) tissue-specific and consistently over-expressed in the gregarious phase regardless of the species and work, (*iv*) tissue-specific and consistently over-expressed in the solitarious phase regardless of the species and work and (*v*) tissue-specific with changing expression tendency depending on the work and/or species.

Reference		Kang et al. 2004	Chen et al. 2010	Guo et al. 2011	Badisco et al. 2011a	Badisco et al. 2011b	Wang et al. 2014	Total
Transcripts analyzed		9	19	27	5	19	45	124
Present in both transcriptomes		6	12	13	4	10	27	72
Congruently gregarious	Significant	1	0	1	0	1	0	3
	Non-significant	0	2	1	0	0	4	7
Congruently solitarious	Significant	0	0	0	0	0	0	0
	Non-significant	0	0	3	0	0	0	3
Only CNS and bibliography	Gregarious	2	10	4	3	3	8	30
	Solitarious	1	0	1	0	2	0	4
Only DT and bibliography	Gregarious	1	2	2	0	3	3	11
	Solitarious	2	1	9	0	1	13	26
Incongruent in bibliography	Solitarious in bibliography	0	0	0	1	3	6	10
	Gregarious in bibliography	0	0	0	0	0	1	1
Absent	In CNS	3	4	8	0	7	15	37
	In DT	0	3	6	1	2	3	15
Other inconsistent results		2	4	6	1	6	10	29

Table 4.33: Comparative analysis of the numbers of shared transcripts between our CNS and DT transcriptomes and those reported in the literature.

Of the 12 genes whose RNAseq-inferred expression tendency we confirmed earlier by qPCR, 2 (3 without taking into account *S. gregaria* nymph qPCRs) show similar results no matter the tissue whereas, and 8 (without including contig NA202 since it was a CNS specific sequence; without a DT homolog) seem to have tissue-specific tendencies in their differential expression between solitarious and gregarious locusts (figure 4.40A). The congruent genes were the phosphoenolpyruvate carboxykinase (PEPCK) and a yellow-h homolog protein. However, qPCRs using *L. migratoria* adults and nymphs did not show a completely congruent expression in any case. Still, some similarities could be discerned in the expression patterns of some of those genes. In fact, tyrosin aminotransferase (TAT) shows a gregarious expression pattern both in *S. gregaria* and *L. migratoria* adults, while both TAT, peroxiredoxin and peptidyl-prolyl cis-trans isomerase (PPI) show solitarious expression pattern in nymphs of both species (figure 4.40B).

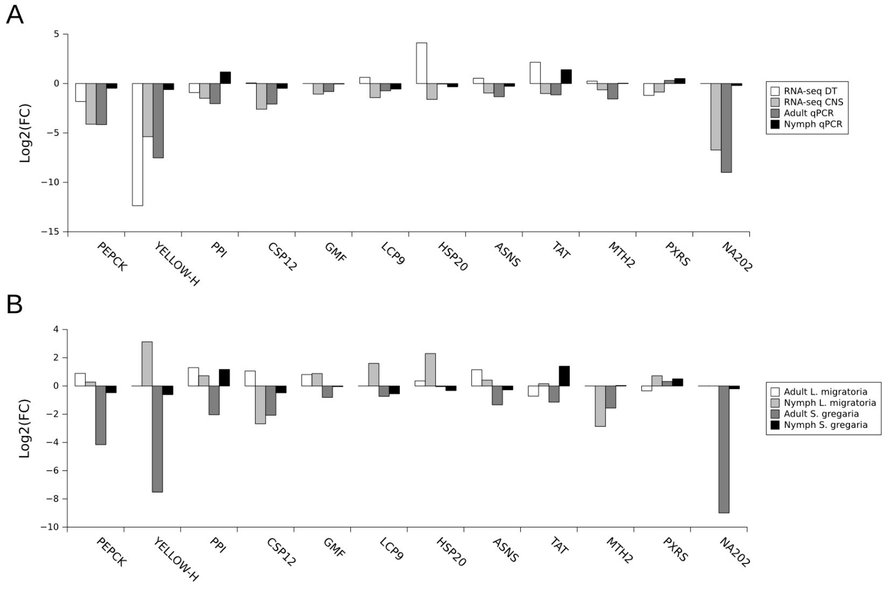

Figure 4.40: Comparison of the expression pattern of qPCR-tested genes between *S. gregaria* tissues and methodologies used (A) and species and developmental stages (B). The fold change values were calculated from the ratio between solitarious by gregarious FPKMs (for RNA-seq data) or relative quantity (for qPCR data).

Another set of genes from *L. migratoria* CNS [Zhang et al., 2012] and DT [Spit et al., 2016] transcriptomes were also used, although the first work contained no differential expression analysis and the comparison in the second is between gut and brain. The *L. migratoria* CNS transcriptome in Zhang et al. [2012] contained a total of 41,179 unigenes (compared to the 115,262 in our *S. gregaria* CNS transcriptome) that include 15,895 sequences with unique BLAST results (a similar number to the one in our *S. gregaria* CNS transcriptome, 16,749, but far from the 57,637 sequences of our *S. gregaria* DT transcriptome). 56 sequences were highlighted in that work, of which 25 and 16 had homologs in our *S. gregaria* CNS and DT transcriptomes, respectively (table 4.34 and table 4.42). However, since this work has no gene expression profiles, we only checked the presence of the described transcripts in both transcriptomes. It is to mention that several sequences described in that work belonged to protein families or genes represented by several variantsin our *S. gregaria* transcriptomes. For its part, the *L. migratoria* transcriptome analysis in Spit et al. [2016] is based on a microarray analysis that used 48,802 sequences (42,861 from *L. migratoria* and 5,941 from *S. gregaria*), focusing on a set of 2,756 transcripts that were over-expressed in the gut compared to the brain (table 4.35).

	Genes in Zhang et al. 2012	Homologs in our *S. gregaria* transcriptomes	Over-expression of the homologs in our *S. gregaria* transcriptomes		
			Gregarious	Solitarious	Non-significant
CNS	10	51*	19	0	32
	15	15**	9	0	6
DT	8	46*	2	1	43
	9	9**	0	2	7

Table 4.34: Presence in our *S. gregaria* CNS transcriptome of the sequences highlighted in Zhang et al. [2012] work on *L. migratoria*'s CNS transcriptome. *: Genes from Zhang et al. [2012] that are represented with more than one sequence in our transcriptome. **: Genes from Zhang et al. [2012] that are represented with one sequence in our transcriptome.

4.4 Discussion

4.4.1 Transcript diversity between tissues

The first striking difference between the CNS and DT transcriptomes is the amount of single BLAST accession numbers they present. The annotated sequences from the DT transcriptome correspond to almost three times the number of unique accession numbers from the CNS transcriptome annotation against the same BLAST databases and using the same cutoffs. The reason for such difference is not the activity of more genes in the DT than in the CNS but it is the different nature of both tissues. The CNS is formed by locust-only material whereas the DT includes a mixture of locust material and materials from the high number of species corresponding to the guts content and microflora (bacteria, protozoa and fungi, both symbiotic and

	L. migratoria DT	*S. gregaria* DT	*S. gregaria* CNS
	Proteases		
Proteases	132	481	180
Serin proteases	102	232	32
Metalloproteases	23	20	43
Cystein proteases	4	13	1
	Glutathione S-transferase		
Glutathione S- transferases	5	135	16
Delta	Reported	17	4
Epsilon	Not reported	11	0
Microsomal	Not reported	5	1
Mu	Reported	0	0
Omega	Not reported	8	1
Sigma	Reported	31	1
Theta	Not reported	6	1
	Cytocrhome P450		
Cytochrome P450	17	289	65
CYP2	Reported	4	0
CYP3	Reported	3	1
CYP4	Not reported	59	4
CYP6	Not reported	134	11
	Transporters		
Major Facilitator Transporter	18	116	9
ABC transporter	7	32	30

Table 4.35: Comparison of the sequence counts for several protein families between Spit et al. [2016] *L. migratoria* study and our *S. gregaria* DT and CNS transcriptomes.

parasites). In fact, the heterogeneity of the DT material, compared to the CNS one, is what explains the differences in G+C content (40.3 % and 43.22 % for CNS and DT, respectively) and N50 value (1,261 and 1,028) between the transcriptomes that we assembled *de novo* based on a similar amount of Illumina Hiseq 2000 Paired End reads obtained by sequencing cDNAs from the CNS and DT tissues of the same *S. gregaria* individuals. In contrast, the CNS transcriptome contains more sequences with unknown annotation (34,696) than the DT does (16,491), which might be due to the different proportion of RNAs from other species present in both transcriptomes—which would in the DT increase the number of annotated sequences from fully sequenced microbial model organisms while reducing the chances for the *S. gregaria* sequences that have no known annotation (explained further in this chapter). Still, the CNS and DT transcriptomes that we compare here came, as mentioned above, from the same species (even from the same individuals), so one would expect most of the BLAST accession numbers from the annotation of the CNS transcripts to be present among the BLAST accession numbers from the annotation of the DT transcripts. However, only 17 % of the unique CNS BLAST accession numbers were present among the accession numbers from the BLAST annotation of the DT transcriptome, corresponding to just 5 % of the DT unique BLAST accession numbers that belong to insect species. The first obvious reason for that small number is the heterogeneity of the DT transcriptome that contains sequences from other species at the expense of *S. gregaria* sequences—many of the least expressed genes in the DT were therefore missed. Besides, bacterial and other microorganisms genomes usually present a more complete annotation and are abundant in the databases, which makes sequences from these group of species have a higher chance to present a BLAST result, hence elevating the number of annotated genes in the DT. Furthermore, these sequences competed with *S. gregaria*'s RNAs during the sequencing, thus shifting the proportion of sequences towards the annotated sequences in the DT compared to the CNS. Another reason would be the expression of different, tissue-specific, variants of the same gene in each transcriptome (which leads to different BLAST accession numbers for the different transcripts of the same gene in the different transcriptomes). The difference between the CNS and DT transcriptomes in the number of sequences that show significant differential expression between locust phases is also striking. In fact, the number of differentially expressed sequences in the DT transcriptome (1,036 unigenes and 3,399 sequences without known annotation) is just about 20.12 % the number of sequences with significant differential expression between locust phases in the CNS transcriptone (10,238 unigenes and 11,804 sequences without known annotation)—see Chapters 2 and 3. Such result does not only confirm that locust phase change is principally CNS-linked, but is also yet another indicator of the genuine and accurate nature of our RNAseq assembly, annotation and expression analyses of both the CNS and DT transcriptomes.

Congruent with the higher number of differentially expressed genes between locust phases in the CNS than in the DT transcriptomes, the number of shared annotated transcripts that show significant differential expression between phases is also asymmetric, with almost two thirds of the total (1,579) showing significant over-expression between phases in the CNS transcriptome, and only 240 transcripts showing over-expression between phases in the DT transcriptome (tables 1 and 2). 175 of these genes show significant differential expression between locust phases in both tissues, although not always in a congruent way (65 congruent vs. 110 incongruent transcripts). The relatively lower number of annotated genes that show the same direction of over-expression in both tissues compared to those that show an opposite (tissue-specific) one—about half—seems to suggest that gregariousness affects gene expression in a largely tissue-specific manner. This is also in agreement with the fact that most of the differentially expressed transcripts of the CNS are over-expressed during the gregarious phase (about 90%), whereas the differentially expressed transcripts of the DT are more evenly distributed between phases (45 % towards solitarious and 55 % towards gregarious phase)—see Chapters 2 and 3.

The anonymous sequences, however, present a different pattern in CNS: while the mean intensity values are in a similar range, the expression pattern seems biased towards solitarious phase in the CNS (figure 4.38A), although this effect is due to the presence of extreme positive (solitarious) FC values (figure 4.38B). Curiously, this bias becomes more conspicuous when analyzing sequences equal or longer than 500 nucleotides, where the most extreme FC negative (gregarious) values dissipate (figure 4.38C), with the anonymous clustered DT sequences showing evenly distributed FC values between the solitarious and gregarious phases (figure 4.38D). Hence, the anonymous sequences show divergent expression patterns in both transcriptomes, which can be understood as profound differences in gene expression between both tissues. We have to take into account that the DT also shows a high number of sequences belonging to different microorganism, thus contributing to the differences in sequence expression levels. Consistent with the above-exposed interpretations is the fact that the number of differentially expressed transcripts in the CNS, especially those with over-expression in the gregarious phase, is about ten times the number of genes that are differentially expressed in the DT (although, in the latter case, most of the over-expressed genes appear in the solitarious phase). Gregariousness implies living in an over-charged environment of increased stimuli, movement, competition, infections, little resources and challenging conditions, whereas the solitarious phase is associated with attenuated locust-locust interactions and abundance of resources. It is therefore reasonable to expect, at least in a speculative way, a convulsed gregarious CNS that regulates all the responses and pathways related to the response to the increased locust-locust interactions and challenging conditions, and an active solitarious DT with increased constitutive processes that deal with digestion of the easily available resources.

4.4.2 Congruency of the over-expression between phases and tissues

Thanks to the current comparative analysis we can confirm the genuine nature of the association between the locust phase change and the transcripts whose presence is shared and whose differential expression between locust phases is congruent between two transcriptomes. Among the annotated transcripts that are congruently over-expressed in the gregarious phase (table 4.36) we highlight the presence of several transcripts related to the immune response, metabolism and energy production, transcription and translation. For instance, among the transcripts associated with immunity, and in addition to the genes that we already discussed in former chapters (such as *black* arginine carboxylase, pacifastin precursor 4 and serine protease inhibitor 3), we can highlight two transcripts that are homologous to lysozymes (accession numbers AAY85086.1 and NP_648151.2). The over-expression of genes related to the immune system reflects, as discussed in previous chapters, the fact that gregarious individuals do not only face stress from intra-specific competition but also have to deal with the increased infections that a gregarious lifestyle propitiates. To highlight are also two transcripts encoding the enzyme phosphoenolpyruvate carboxykinase (PEPCK, AAM11205.1 and P20007.2), related to the Krebs cycle and gluconeogenesis, that were also confirmed in our CNS by qPCR, as shown in figure 2.17 from Chapter 2 and figure 4.40A of this chapter. The increase in the activity of genes related to energy and metabolism directly relates to the increased activity of the locusts during the gregarious phase. In fact, in mice, PEPCK over-expression leads to increase of the animals' activity and aggressiveness [Hakimi et al., 2007, Hanson and Hakimi, 2008], a strikingly similar outcome to the locusts' behaviour during the gregarious phase. The phase change is about a change of gene expression and not the genes themselves [Bakkali, 2013]. Accordingly, it is reasonable for transcripts related to transcription and translation to be up-regulated in the gregarious phase. Examples include ribosomal proteins (P15357.2, P18101.2, Q9VTP4.2), tRNA transferases (Q9VSZ6.1, Q9VK89.1), elongation factor (AFH89818.1), and RNA polymerases (Q9VC49.1, NP_001153714.1). Interestingly, and in relation with gene transcription and translation, epigenetics and protein processing are also among the processes with over-expressed transcripts. We have, for instance, transcription adapter 2A (AAK92962.1), which is part of an histone acetylation complex that has chromatin remodelling functions [Guelman et al., 2009], and E3 ubiquitin ligase RNF 126-B (AAK93431.1), which might be a sign of enhanced ubiquitin signalling regulation and protein degradation activity.

Despite the fact that we are comparing CNS and DT transcriptomes, we find a noticeable number of clearly CNS-related transcripts to be shared and consistently over-expressed in the gregarious phase in both tissues. Their shared presence is very likely due to the unavoidable presence of neuronal termini in the DT tissue, and their congruent expression trend confirms their genuine association with the gregarious phase. Examples include sniffer (AAM49926.1), a protein which at low levels promote neurodegeneration [Botella et al., 2004]. Zinc/iron regulated transporter-related protein 1 (CAC14873.1) also seems to be linked with neurodegeneration promotion, lifespan shortening and cognitive failures in *D. melanogaster* [Lang et al., 2012]. NimA-like kinase (NP_651293.1) regulates mitosis in CNS [Sepp et al., 2008]). Myospheroid (EFA10689.1) has a number of CNS-related functions, including axon guidance, central nervous system development, and sensory perception of smell [De La Pompa et al., 1989, Goddeeris et al., 2003, Sepp et al., 2008, Delon and Brown, 2009]. Genes related to neuronal plasticity have been described to be up-regulated in the gregarious phase in *S. gregaria* and *L. migratoria* [Badisco et al., 2011a,b, Wang et al., 2014b], and our data confirm that a number of CNS development genes do have a role in the triggering or maintenance of the gregarious phase very likely via their involvement in neural plasticity.

Among the genes that have shared over-expression during the gregarious phase both in the CNS and DT, and whose annotation is directly related to neurotransmition, we can highlight annexin IX (AAN71504.1) and CutA (AAM50725.1). The first, despite being part of a phospholipid binding protein family sometimes related to vesicle formation [Benz and Hofmann, 1996, Goebeler et al., 2003], presents a low reception affinity for acetylcholine and is also involved in autoimmunity [Nguyen et al., 2000]. The expression of the second, which anchors acetylcholinesterase (an enzyme that degrades acetylcholine) to biological membranes, may be induced in order to export this enzyme from vesicles to synaptic membranes for inactivating acetylcholine and inhibiting its response.

Regarding the annotated genes that were shared between the CNS and DT transcriptomes and that show consistent over-expression during the solitarious phase, we found 22 transcripts with general functions such as energy and metabolism, and functions that seem proper to either the CNS or the DT (shown in table 4.37). We detected energy transport and storage related genes, such as vitellogenin (XP_971398.1), which stores and transports lipids, a fatty acid binding protein (AAN71654.1), which transport lipids, and a methionin-rich hexamerin (AAX14950.1), which stores and transports amino acids [Wheeler et al., 2000, Pan and Telfer, 2001]. *bent/twitchin* homologue (ABX00765.1) is involved in muscle development [Schnorrer et al., 2010] and contraction [Greene et al., 2008]; its congruent expression might be due to the function of the smooth muscle of the DT and the presence of muscle tissue in the CNS-enriched library (see the Chapter 2). Among the neurotransmission-related genes we found an alpha-l1 nicotinic acetylcholine receptor homologous sequence (EFN71375.1), which encodes a protein that modulates the opening of transmembrane ion channels when bound to acetylcholine, and among the immune response function genes we found the A3 subunit of the phenoloxidase (Q9V521.1), related to immune response and melanization via cathecholamines pathway. The presence of genes congruently over-expressed during the solitarious phase although the functions that they belong to also contain genes that are over-expressed during the gregarious phase is not, as it might appear, incongruent. The fact that solitarious locusts show increased activity of certain housekeeping processes (e.g. digestion) and that they also face certain of the challenges that the gregarious locusts fase, although at lower intensity (e.g., exposition to pathogens), means that some transcripts related to those functions should be expected to be over-expressed during the solitarious phase too. While the functional categories that a living organism uses are the same, the type of genes of each of these categories changes according to the demands. Surprisingly, six out of the 22 sequences that were congruently over-expressed between tissues during the solitarious phase were homologous to probable transposable elements (AAM11674.1, BAC82628.1, EFA11647.1, EFA11995.1, EFA13106.1 and XP_003741356.1), compared to just a single transposon sequence (AAL89932.1) that was congruently over-expressed between tissues during the gregarious phase. Genome instability is proper to the gregarious phase [Nolte, 1969, Fox, 1973]. Accordingly, our separate tissue RNAseq data show over-expression of transposons and DNA recombination and repair genes during the gregarious phase (see the Chapters 2 and 3). We would therefore expect more intersection between tissues in the gregarious phase rather than in the solitarious one and, short of valid explanation, our best guess is that the difference in the number of congruently over-expressed transposons between tissues might be just a stochastic result (a sort of type I error).

As to the sequences that show opposite expression pattern between tissues, they surely are not phase markers but, despite showing opposite patterns of over-expression between phases and tissues, the significant alteration of their expression levels between locust phases highlights the tissue-specific nature of some of the alterations in biological processes between phases. From table 4.38, we can see how the shared annotated genes that are over-expressed in the gregarious CNS and in the solitarious DT relate to processes like glycoside, amino acid and lipid transport and degradation, protease inhibition, ion transport and detoxification. Whereas, from table 4.39, we can see how the shared annotated genes that are over-expressed in the solitarious CNS and gregarious DT transcripts relate to protease activity, translation and cell cycle regulation. Such result is expectable, given the nature of the tissues compared here and the nature and degrees of challenges and activities of both phases, and is congruent with the data on each tissue.

Another issue are some of the annotations that show incongruent pattern of over-expression between tissues and phases and that, at the same time, are variants, have similar (or like) annotations, or belong to the same gene family of functional category as genes that are congruently over-expressed between tissues and phases. Examples of incongruently expressed transcripts that are similar to congruently expressed ones are a histone-lysine N-methyltransferase SETMAR-like protein (EFN89173.1), a *bent* homologous sequence (AAK77295.1), a zinc/iron regulated transporter-related protein 71B (ABO52850.1), and an alpha-l1 nicotinic acetylcholine receptor (EFN65035.1). In the absence of deeper individual sequence analyses and/or functional testing of these sequences we cannot confirm them as genuine variants that have tissue-specific differential expression between locust phases, nor can we discard them as sequences whose differential expression might potentially be type I errors that the application of the false discovery rate could not correct. Pleiotropy could also be a reason for such apparent incongruence; for instance, a hexamerin like protein 1 homologous sequence (ACU78068.1) is over-expressed the in gregarious CNS and solitarious DT, despite other hexamerin transcripts being congruently over-expressed during the solitarious phase in both tissues. Hexamerins can be used as amino acid storage proteins, in favourable conditions, but also as anti-bacterial proteins, in immune challenging conditions [Burmester, 1999, Wang et al., 2007], which might explain their over-expression both during the solitarious and gregarious phases—the former phase allows storing amino acids whereas the latter shows enhanced immune response and energy mobilization.

Worth highlighting among the genes that belong to a protein family that shows incongruent over-expression pattern between phases and tissues are the odorant binding proteins (OBPs), ACR39391.1 and AEX33167.1, and several cytochrome P450 (CYP) variants (a CYP 4g15 (Q9VYY4.1), a CYP 4c3 (Q9VA27.1), a CYP 6d5 (Q9VFP1.1) and a CYP 304a1 (Q9VG17.2)). Like the chemosensory proteins (see Chapter 5 of this thesis), OBPs have a plethora of target ligands and tissue-dependent expression patterns [Pelosi and Maida, 1995, Pelosi et al., 2005]. The direction of their over-expression pattern thus depends on the tissue, the conditions and the chemicals that the animal has to deal with. When facing the same chemicals at similar intensities, the same tissues and physiological conditions are expected to show congruent expression of the OBPs that deal with those chemicals, otherwise the expression pattern would be tissue and/or condition-specific and, thus, apparently incongruent. Similarly, CYP is an essential and multi-faceted protein family with functions ranging from signalling to detoxification [Feyereisen, 1999]. The signalling cascades are expected to have both common and phase and tissue-specific components, and the same applies for detoxification that has to specifically deal both with common as well as tissue and phase specific conditions of toxicity.

Similar interpretation as the one that applies to the incongruently over-expressed genes that belong to the same family is valid for incongruently over-expressed genes that belong to the same functional category. For instance, within genes that have immune system-related function, serine protease inhibitor 3 (SERPIN 3) and pacifastin SgPI-4 are consistently over-expressed during the gregarious phase both in the CNS and DT, while alpha trypsin inhibitor heavy chain 3 (XP_001943110.2), SERPIN SgPI-1 (O46162.1) and SERPIN B6 (XP_968005.1) are over-expressed in gregarious CNS and in solitarious DT.

What applies to the shared genes whose over-expression is in the gregarious CNS and solitarious DT also applies to the ones whose incongruent over-expression is between the solitarious CNS and the gregarious DT. These include two Jonah 65Aiv trypsin like protease homologs (AAL49280.1 and AAL29107.1), two ribosomal proteins (ACS78170.1 and AAM50819.1), a checkpoint-like protein (ABK29471.1, involved in DNA repair and cell cycle regulation [McNeely et al., 2014]), a transitional endoplasmic reticulum ATPase TER94 (XP_001605497.2, involved in maintenance of endoplasmatic reticulum [Thomson et al., 2008]) and a transposable element (EFX62991.1).

4.4.3 At least not all the sequences that have no BLAST annotation are potential sequencing or assembly artifacts

The sequences that have no known BLAST annotation represent 30.102 % and 22.247 % of the assembled *S. gregaria* CNS and DT transcriptomes, respectively (see Chapters 2 and 3). Both proportions are not too different from each other and, although they seem high, the fact is that the number of non annotated transcripts in other insect CNS and DT transcriptomes is also high. It ranges from 50 % to 66 % in other locust CNS transcriptomes [Chen et al., 2010, Badisco et al., 2011a, Zhang et al., 2012] and from 40 % to 50 % in DT transcriptomes from other insects [Pedra et al., 2003, Hughes and Vogler, 2006, Coates et al., 2008]. These sequences may be either sequencing and/or assembly artifacts, non-conserved parts of known mRNAs, mRNAs from unknown coding sequences, or non-coding RNAs (ncRNAs).

Sequencing and assembly artifacts are unavoidable in every NGS and *de novo* assembly (in the absence of reference genome or transcriptome, the assembly algorithms join the sequencing reads based on small windows, k-mers, and in all the possible manners, with no guarantee that all the resulting contigs represent real transcripts). Besides, we cannot control the RNA edition status, thus adding complexity to the assembly; since different isoforms and a variable degree of mRNA maturation are present in the sample. Similarly, contig assembly from short sequencing reads unavoidably produces partial sequences of genuine known transcripts, some of which might be variable enough as to not produce significant BLAST match to the corresponding aminoacids and nucleotide homologs. We cannot even discard that some transcripts might correspond to highly diverging known genes that would not pass the BLAST homology thresholds (i.e., higher than threshold E-value or lower than threshold HSP). The NCBI database is huge and contains ever increasing scores of aminoacids and nucleotide sequences of all the known genes. Yet, not all the existing genes are sequenced, known and deposited in that database. Thus, a number of the sequences that gave no significant BLAST result might in fact correspond to coding sequences that are not described yet. The last possibility includes ncRNAs. These are sequences that are currently at the cutting edge of the RNA and gene expression regulation research. Even the compact genome of *D. melanogaster* has recently been found to contain scores of such sequences [Brown et al., 2014]. The problem with these sequences is their low degree of sequence conservation, which does not allow their identification via BLAST. Nevertheless, ncRNA specific databases and ncRNA prediction software have now been developed [Griffiths-Jones et al., 2005, Kong et al., 2007, Xie et al., 2014], so one of the potential future research on our data would be identifying locust ncRNAs. With the sequences that have no known annotation being most likely a mixture of the above mentioned types, the question is how can we trust them as genuine transcripts, partial or not, known or unknown, coding or not? Already in the CNS and DT transcriptome chapters we highlighted that being artifacts is not likely at least for the sequences that are formed from a high number of sequencing reads (we even amplified and analyzed by qPCR one of these sequences—see Chapter 2). Furthermore, the fact that 12,213 (23.86 %) of these sequences appear both in the CNS (corresponding to 35.20 % its unknown sequences) and DT transcriptomes (corresponding to 74.06 % of its unknown sequences) is a tangible proof of their mostly genuine nature.

Another problem that one faces when studying *de novo* assembled transcripts that lack a BLAST result is how to discriminate between variants of the same gene and different genes. In fact, the high number of sequences with no known annotation is no doubt partly due to the presence of several variants of the same genes. We opted for an objective, threshold-based, strategy. We calculated the mean nucleotide sequence identity from pairwise comparisons between all the sequences that have the same BLAST annotation in both tissues, and we applied that mean value as sequence identity threshold for grouping the sequences that have no known annotations into cluster of sequences that potentially belong to the same gene. The prior to this logic is the assumption that the degree of conservation between sequences from the different transcriptomes of the same species is the same for every sequence. A potential shortcoming could be the possibly higher conservation of the annotated sequences compared to the ones that have no known annotation. We therefore risk having false positives (sequences that do not cluster together although they belong to the same gene, because their variability may be higher). Still, we have to remember that our transcriptomes come from the same animals, so even for highly variable sequences the chances of generating false positives are just as high as the differences between ten haplotypes can get (we used pooled cDNAs from five locusts—see the general methodology section).

We are also aware that the minimum sequence length for our transcriptomes is 100 nucleotides, a value that might induce to think that some of the generated clusters of sequences without annotation are produced in a spurious way. However, performing the same analyses after excluding sequences shorter than 500 nucleotides indicated that the tendencies were conserved: both the mean number of sequences per cluster, the cluster sequence number by singlet sequence number ratio, and the number of clusters by total number of sequences were very similar. Thus, the abundance of sequences shorter than 500 nucleotides is not mainly due to assembly artifacts or a fragmented assembly (since they would contribute to clusters) and thus, eliminating these might lead to a potential loss of information.

After clustering, the number of sequences with no known annotation in the CNS and DT transcriptomes is reduced to 27,949 (19.45 % reduction) and 9,744 (42.51 % reduction), respectively. As to their patterns of over-expression, we have the same expected mixture of situation as with the annotated sequences (congruent, incongruent and tissue-specific over-expression). The lack of annotation makes discussing the case of these sequences pointless, until we determine their most likely interesting functional implications.

4.4.4 Comparison with the data obtained elsewhere

The importance of data validation in the 'omics' field of research is crucial. We previously validated some of our RNAseq data by the use of other methodology (qPCR), we then compared with the data published elsewhere and, here, we compared and looked for consistency between different materials (tissues). In a further step, in this part we compare the results of our comparison between tissues to those published in the literature on locust phase change. It is however worth highlighting that we gathered a set of 124 genes about which other works specifically provided differential expression data between solitarious and gregarious locusts. Still, this is obviously no complete validation of the hundreds of thousands of sequences reported by us and elsewhere—individual validation would still be needed for most of the individual genes that might be of interest to the researchers and that were not fully validated here. As expected, our data sometimes agree with the published ones and sometimes do not. Of course, if a gene is truly associated with a biological phenomenon one has naively to expect 100% agreement between the data no matter the laboratory from which they come. However, several types of differences between the experimental settings (different species, populations, individuals, developmental stages, tissues, and techniques used) end up constituting several sources of variation that might result in differences in the results of some genes between experiments and laboratories. Of course one can no doubt apply the expert knowledge and assess the possibility of inclining the balance towards one of the discordant results based on consideration of the type of material (for instance, same species versus different species) and the reliability and accuracy of the experimental design and technology (RNAseq versus microarrays, for instance). We can therefore in no way discard the association of a gene with the phase change based on the disagreement between experiments that have different settings. In case of coincidence of the results, however, we can definitely confirm the association of the gene in question with the biological phenomenon.

A striking case of apparently discordant results is the one of the tyrosine hydroxylase (TH, the enzyme that synthesizes L-DOPA). The expression of the gene for this enzyme was up-regulated towards the solitarious phase in our CNS transcriptome (11.154 FPKM in solitarious phase vs 10.372 FPKM in gregarious phase, accession number ABQ95974), while being up-regulated towards the gregarious phase in our DT transcriptome (1.595 FPKM in solitarious phase vs 4.988 FPKM in gregarious phase, accession number AAA62877), although with less coverage than in CNS. As it happened, Chen et al. [2010] show TH over-expression in gregarious *L. migratoria* whole nymphs and Ma et al. [2011] found that downstream catecholamine pathway enzymes, such as DOPA decarboxylase (*Ddc*), *pale*, *henna* or *ebony* are also over-expressed in gregarious *L. migratoria* nymphs. In fact, dopamine injection induces gregarious behaviour on isolated reared nymphs, as well as *pale* gene silencing induces a colour change towards black [Ma et al., 2011]. Regarding *S. gregaria*, Miller et al. [2008] report a probable L-DOPA analogous molecule in gregarious egg pod foam but, apart from that work, there is no other study on the effect of DOPA on phase change in this species. Given the species- and/or stage-specific differences between the cited works and ours, the differences in the expression of DOPA-related genes that we report here might be expected. Accordingly, one would infer that either our CNS datum for that gene is erroneous or that the gene has a tissue-specific pattern of solitarious versus gregarious expression (we cannot include potential inter-individual, inter-population or inter-species variation here as we used the same animals both for CNS and DT sequencing). In fact, apparent incongruence between different works and materials is no stranger to the literature, for instance: Anstey et al. [2009] report serotonin as a solid trigger of the gregarious phase in *S. gregaria*. However, Guo et al. [2013] reported that serotonin enhances solitariousness in *L. migratoria*, and its role in *S. gregaria* was even criticized in Tanaka and Nishide [2013] —Rogers et al. [2014] latter published a work in support of the previous work by their research laboratory that suggested serotonin as the trigger of gregariousness in locusts [Anstey et al., 2009]. The saga of serotonin might thus be just due to the possibility that this peptide might have varying association types with the phase depending on the locust, tissue or condition —not to mention the possible errors from the behavioural modelling that we discuss in Chapter 1.

Exceptions apart, comparison of our data with those of other published works that use high throughput data not only confirms the quantitative aspects of the assembly (number of annotated sequences and number of sequences with no known annotation), but also shows an agreement in the general tendencies of the alterations in biological functions that are associated with locust phase change, and allow even pinpointing concrete genes as no doubt associated with the phase change both in *S. gregaria* and in locusts as a whole. For instance, the work on the CNS of the migratory locust by Zhang et al. [2012] provides a list of highlighted transcripts, all of them present in our CNS transcriptome, with an elevated proportion of gregarious over-expressed transcripts and no solitarious over-expressed transcripts (table 8). While most ion channels or peptide transporters reported in that work are not significantly over-expressed in *S. gregaria* (table 4.42), we observe that enzymes that relate to structures from synapsis, catecholamines and circadian clock are over-expressed mainly in *S. gregaria*'s CNS during the gregarious phase. Similarly, the work from Spit et al. [2016], where they compare the transcript expression between the gut and the brain of the migratory locust, is also concordant with ours. They, for instance, report similar protease diversity, being the serine-proteases (more precisely the trypsin-like type) the most abundant proteases in the gut, and the most abundant type of glutathione S-transferases happens to be the sigma subfamily (table 4.35). The results from our CNS transcriptome also point to the same abundance trends as stated in both Spit et al. [2016] and our DT transcriptome works, with some minor differences (such as the abundance of cystein proteases being higher than serin proteases or the absence of the major allergen protein in CNS). However, since work the work from Spit et al. [2016] is based on a microarray study, the novelties are limited by the sequences used as probes.

As to the specific genes, of the 12 that show congruent over-expression pattern between our CNS and DT data, only one (pacifastin 4) was reported in the literature on *S. gregaria* as also showing concordant tendency of over-expression. These are therefore genes with confirmed association with the phase change in *S. gregaria*. One on them, pacifastin 4 is, as already mentioned in Chapter 3, a serine protease inhibitor related to the immune response in arthropods [Aspán et al., 1990, Liang et al., 1997] and has been already reported as over-expressed in gregarious desert locusts [Simonet et al., 2005]. Even more, a good number of genes that were reported in Kang et al. [2004], Chen et al. [2010], Guo et al. [2011], Wang et al. [2014b] to be significantly associated with the phase change in *L. migratoria*, also show significant association with that phase change in our data on *S. gregaria* CNS and DT. In fact, 12 of the genes that show congruent over-expression in both CNS and DT transcriptomes of *S. gregaria* were reported to have the same expression tendency in *L. migratoria*. These are genes that we can confirm as associated with the phase change in locusts in general, or at least the two main locust pest species, thus pointing towards the presence of conserved locust phase-associated pathways. The most remarkable among these genes were the ones related to neurotransmition and its reception and signalling. Almost all of them show over-expression during the gregarious phase, and their list includes the adipokinetic hormone receptor, CASK protein, diuretic hormone receptor, GABA receptor, glutamate receptor, Malvolio protein, octopamine receptor and a synaptic vesicle protein. Other functions and genes worth highlighting among the ones that show similar over-expression patterns between species are ubiquitylation (with E3 ubiquitin-protein ligase hyd), chemoreception (with several chemosensory proteins) and translation (Glycyl-tRNA synthase), all of which up-regulated during the gregarious phase. Annexin IX and pacifastin inhibitor are also reported as up-regulated during the gregarious phase in all the bibliographical sources, thus highlighting the increased exocytosis and immune activities during that phase in locusts in general. Interestingly, hexamerin was also found to show similar expression pattern between our CNS transciptome and the bibliography. Some variants of this protein family seem to modulate juvenile hormone titers in termites and modulate caste-related phenotypic plasticity [Zhou et al., 2007], thus highlighting the association of both molecules with crowd (social) behaviours in insects in general.

4.5 Supplementary material

Table 4.36: Transcripts that show over-expression during the gregarious phase both in the CNS and DT transcriptomes. The results are limited to the sequences that share BLAST accession number between CNS and DT transcriptomes.

Accession number	BLAST Description	CNS Refference	CNS FC	CNS FDR	DT Refference	DT FC	DT FDR
AAD55415.1	AF181629_1BcDNA.GH04637 [Drosophila melanogaster]	78365	-1,045	5,913E-11	Dm175703	-1,618	9,571E-05
AAK92962.1	GH19029p [Drosophila melanogaster]	52979	-1,051	1,649E-17	Dm22685	-1,575	2,214E-03
AAK93014.1	GH23453p [Drosophila melanogaster]	53013	-1,504	1,156E-07	Dm23086	-0,718	7,503E-05
AAK93431.1	LD47007p [Drosophila melanogaster]	53845	-1,504	4,885E-107	Dm51678	-0,518	1,187E-04
AAL13472.1	GH01077p [Drosophila melanogaster]	52833	-0,562	1,284E-20	Dm35839	-0,762	6,345E-05
AAL39897.1	LP11089p [Drosophila melanogaster]	53917	-1,380	3,088E-02	Dm54017	-2,586	2,514E-07
AAL89932.1	RH07106p [Drosophila melanogaster]	82606	-0,668	6,994E-30	Dm65874	-0,994	3,889E-16
AAM11132.1	LD10758p [Drosophila melanogaster]	53599	-0,605	8,290E-06	Dm72278	-0,658	2,167E-02
AAM11205.1	RE12569p [Drosophila melanogaster]	82324	-4,567	2,757E-40	Dm72584	-4,230	5,068E-03
AAM49926.1	LD36273p [Drosophila melanogaster]	53728	-0,579	4,833E-65	Dm81257	-0,449	1,179E-02
AAM50253.1	LD20420p [Drosophila melanogaster]	53639	-0,947	6,521E-31	Dm82296	-0,861	7,438E-04
AAM50725.1	GM24986p [Drosophila melanogaster]	53072	-0,962	3,966E-32	Dm83373	-0,801	9,771E-05
AAM75077.1	RE61939p [Drosophila melanogaster]	82558	-0,621	1,056E-05	Dm88253	-2,300	3,668E-04
AAN71504.1	RH01338p [Drosophila melanogaster]	82594	-1,174	5,133E-179	Dm111945	-1,868	1,616E-32
AAX33543.1	LD14383p [Drosophila melanogaster]	53613	-1,307	6,147E-114	Dm178564	-0,590	2,210E-05
AAY85086.1	IP04203p [Drosophila melanogaster]	53158	-1,658	0,000E+00	Dm186094	-0,804	1,426E-05
AFH89818.1	FI20117p1 [Drosophila melanogaster]	52791	-1,081	3,282E-65	Dm151520	-0,963	1,199E-06
CAC14873.1	zinc/iron regulated transporter-related protein 1, DZIP1 protein [Drosophila melanogaster]	51120	-0,697	1,079E-26	Dm2798	-0,766	4,541E-08
CAC82510.3	pacifastin-related precursor 4t [Schistocerca gregaria]	Sg_CNS_NA4Plus3143	-2,897	0,000E+00	90431	-1,331	1,003E-30
CAD11970.1	pacifastin-related serine protease inhibitor precursor [Locusta migratoria migratorioides]	77531	-4,457	2,408E-61	17113	-3,815	7,457E-07
EFA06655.1	hypothetical protein TcasGA2_TC009580 [Tribolium castaneum]	17020	-1,822	1,878E-36	29531	-0,188	6,094E-07
EFA10689.1	myospheroid [Tribolium castaneum]	32284	-0,901	6,120E-05	30747	-0,394	2,612E-08
EFN71403.1	hypothetical protein EAG_00286 [Camponotus floridanus]	72207	-1,348	7,410E-04	42871	-1,001	4,266E-09
NP_001153714.1	poly-(ADP-ribose) polymerase [Tribolium castaneum]	35503	-0,596	2,977E-30	21363	-0,733	9,895E-05
NP_648151.2	CG8492 [Drosophila melanogaster]	52482	-2,611	3,981E-10	Dm155448	-1,511	6,212E-05
NP_651293.1	nimA-like kinase [Drosophila melanogaster]	32808	-1,846	5,160E-03	Dm104577	-0,403	2,640E-02
O46163.1	Serine protease inhibitor 3 [Schistocerca gregaria]	Sg_CNS_NA4Plus23248	-1,280	0,000E+00	3883	-0,808	3,879E-10
P15357.2	Ubiquitin-40S ribosomal protein S27a [Drosophila melanogaster]	41681	-1,324	3,114E-48	Dm135382	-1,618	8,631E-15
P18101.2	Ubiquitin-60S ribosomal protein L40 [Drosophila melanogaster]	41312	-0,881	4,031E-60	Dm135266	-1,057	4,745E-18

Accession number	BLAST Description	CNS Refference	CNS FC	CNS FDR	DT Refference	DT FC	DT FDR
P20007.2	Phosphoenolpyruvate carboxykinase [GTP] [Drosophila melanogaster]	33809	-5,218	2,116E-272	Dm12575	-1,817	5,782E-11
Q05547.1	Full=Serine/threonine-protein phosphatase alpha-3 isoform [Drosophila melanogaster]	35768	-31,862	4,478E-02	Dm17175	-1,951	1,159E-14
Q9VC49.1	DNA-directed RNA polymerases I, II, and III subunit RPABC5 [Drosophila melanogaster]	41583	-0,616	3,230E-22	Dm2879	-1,364	7,224E-21
Q9VK89.1	Probable tRNA (guanine(26)-N(2))-dimethyltransferase [Drosophila melanogaster]	46825	-0,797	2,627E-29	Dm10221	-1,590	3,385E-10
Q9VQX4.2	Nicotinate phosphoribosyltransferase [Drosophila melanogaster]	34371	-1,297	5,436E-55	Dm192707	-0,466	2,478E-02
Q9VSZ6.1	Queuine tRNA-ribosyltransferase subunit QTRTD1 homolog [Drosophila melanogaster]	38151	-1,395	8,364E-12	Dm195174	-0,609	9,268E-04
Q9VTP4.2	60S ribosomal protein L10a-2 [Drosophila melanogaster]	38158	-0,692	2,140E-04	Dm170707	-1,345	9,895E-07
XP_001607107.1	PREDICTED: pyridoxal kinase isoform X2 [Nasonia vitripennis]	38125	-1,664	3,302E-06	197	-1,041	1,352E-07
XP_001942617.2	PREDICTED: N-(5-amino-5-carboxypentanoyl)-L-cysteinyl-D-valine synthase-like [Acyrthosiphon pisum]	29512	-1,066	4,193E-32	57806	-1,418	3,395E-05
XP_001944507.2	PREDICTED: leucine-rich PPR motif-containing protein, mitochondrial-like [Acyrthosiphon pisum]	29374	-0,590	5,832E-03	2280	-0,442	6,807E-03
XP_002588454.1	hypothetical protein BRAFLDRAFT_197365 [Branchiostoma floridae]	55946	-2,151	0,000E+00	1164591___k93_f_741953	-4,078	1,314E-05
XP_003248175.1	PREDICTED: uncharacterized protein LOC100573341 [Acyrthosiphon pisum]	72320	-32,529	1,447E-07	2299	-1,594	4,326E-27
XP_003248559.1	PREDICTED: mannitol dehydrogenase-like, partial [Acyrthosiphon pisum]	48042	-6,590	1,325E-40	2310	-1,493	1,094E-03
XP_975528.1	PREDICTED: transmembrane protein 110 [Tribolium castaneum]	46124	-0,887	1,502E-78	256747	-1,369	1,288E-09

Table 4.37: Transcripts that show over-expression during the solitarious phase both in the CNS and DT transcriptomes. The results are limited to the sequences that share BLAST accession number between CNS and DT transcriptomes.

Accession number	BLAST Description	CNS Refference	CNS FC	CNS FDR	DT Refference	DT FC	DT FDR
AAG33250.1	trypsin-lambda [Drosophila melanogaster]	46984	6,211	6,132E-16	Dm2974	0,597	2,527E-04
AAL48992.1	RE40129p [Drosophila melanogaster]	82484	2,661	1,092E-121	Dm59061	2,294	4,313E-05
AAM11674.1	AF492764_2pol protein [Drosophila melanogaster]	84339	3,350	4,721E-02	Dm71266	0,471	2,941E-02
AAN71654.1	SD12036p, partial [Drosophila melanogaster]	82995	0,771	9,130E-43	Dm112549	1,639	4,547E-182
AAR82792.1	LD07466p [Drosophila melanogaster]	53579	0,471	2,090E-66	Dm157319	1,396	2,914E-07
AAX14950.1	methionine-rich hexamerin-like protein 1 [Romalea microptera]	Sg_CNS_NA4Plus995	1,664	0,000E+00	4410	34,328	7,979E-14
ABK57073.1	IP02555p [Drosophila melanogaster]	53134	33,624	4,346E-05	Dm7802	0,601	9,100E-08
ABX00765.1	LP08376p [Drosophila melanogaster]	53897	0,341	7,222E-13	Dm33489	1,414	2,973E-03
ACH48223.1	tumor differentially expressed protein [Haplopelma schmidti]	Sg_CNS_NA4Plus6238	1,007	5,230E-163	2870	0,675	4,867E-24
BAC82628.1	pol-like protein [Anopheles gambiae]	35438	0,230	1,785E-05	2531	0,577	2,096E-04
BAM20251.1	unknown protein, partial [Papilio xuthus]	Sg_CNS_NA4Plus9696	1,949	0,000E+00	3056	2,048	0,000E+00
EFA11647.1	hypothetical protein TcasGA2_TC010626 [Tribolium castaneum]	19625	1,142	0,000E+00	2848	1,281	0,000E+00
EFA11995.1	hypothetical protein TcasGA2_TC005145 [Tribolium castaneum]	11452	0,231	2,190E-61	31297	0,469	4,679E-04
EFA13106.1	hypothetical protein TcasGA2_TC016329 [Tribolium castaneum]	27228	2,137	6,556E-09	33445	34,652	1,618E-02
EFA13558.1	hypothetical protein TcasGA2_TC002334 [Tribolium castaneum]	6436	1,400	0,000E+00	35128	1,160	6,848E-05
EFN71375.1	hypothetical protein EAG_03557 [Camponotus floridanus]	19496	0,346	2,418E-170	42776	1,723	0,000E+00
ELU14262.1	hypothetical protein CAPTEDRAFT_151631, partial [Capitella teleta]	Sg_CNS_NA4Plus8183	1,823	0,000E+00	2892340____370090	1,694	4,793E-78
Q9V521.1	Phenoloxidase subunit A3 [Drosophila melanogaster]	37829	2,374	0,000E+00	Dm176353	1,230	9,350E-21
XP_001653948.1	hypothetical protein AaeL_AAEL001772 [Aedes aegypti]	Sg_CNS_NA4Plus2407_2	0,808	2,207E-210	241	1,230	4,254E-16
XP_001852902.1	GLP_748_1200_211 [Culex quinquefasciatus]	Sg_CNS_NA4Plus27093	1,393	0,000E+00	3854	0,538	1,623E-11
XP_003741356.1	PREDICTED: uncharacterized protein LOC100901629 [Metaseiulus occidentalis]	Sg_CNS_NA4Plus29302	0,743	9,484E-03	75042	0,479	9,461E-03
XP_971398.1	PREDICTED: vitellogenin [Tribolium castaneum]	81680	1,524	0,000E+00	256543	32,867	3,559E-03

Table 4.38: Transcripts that show over-expression during the gregarious phase in the CNS transcriptome and during the solitarious phase in the DT transcriptome. The results are limited to the sequences that share BLAST accession number between CNS and DT transcriptomes.

Accession number	BLAST Description	CNS Refference	CNS FC	CNS FDR	DT Refference	DT FC	DT FDR
1101404B	1101404BORF 2	78446	-0,714	8,393E-03	Dm93244	0,606	2,786E-03
A1ZAI5.1	FACR1_DROMERecName: Full=Putative fatty acyl-CoA reductase CG5065	101782	-3,779	3,023E-07	Dm97965	0,742	1,088E-06
AAA50836.1	BTB-IV protein domain, partial [Drosophila melanogaster]	86655	-0,932	7,356E-12	Dm173421	1,054	8,211E-03
AAA70221.1	putative ORF1 [Drosophila melanogaster]	37993	-0,726	1,719E-107	Dm203761	0,529	4,658E-07
AAA70222.2	putative ORF2 [Drosophila melanogaster]	38109	-0,806	1,690E-20	Dm170955	0,674	1,712E-14
AAA92147.1	reverse transcriptase [Bombyx mori]	39370	-0,551	1,085E-63	78274	2,025	5,657E-45
AAB26515.1	glutathione S-transferase D24, DmGST24 {EC 2,5,1,18} [Drosophila melanogaster, Peptide, 215 aa]	103519	-0,685	4,432E-74	Dm12007	1,124	8,040E-04
AAD00903.1	sorbitol dehydrogenase [Drosophila melanogaster]	43241	-0,678	1,675E-08	Dm159299	1,521	6,411E-19
AAK77295.1	GH07636p [Drosophila melanogaster]	52900	-1,461	6,122E-24	Dm21057	0,866	5,099E-34
AAK93413.1	LD45641p [Drosophila melanogaster]	53825	-1,071	5,478E-259	Dm25717	1,399	1,017E-02
AAK93560.1	SD09370p [Drosophila melanogaster]	82982	-1,272	2,436E-99	Dm26758	0,348	2,359E-02
AAL13528.1	GH06241p [Drosophila melanogaster]	52882	-0,667	1,099E-03	Dm37567	1,172	9,813E-08
AAL13704.1	GH27944p [Drosophila melanogaster]	53034	-0,654	4,188E-10	Dm36742	1,350	2,973E-06
AAL25321.1	GH12380p [Drosophila melanogaster]	52935	-1,396	7,050E-06	Dm40562	0,803	4,882E-05
AAL25334.1	GH13775p [Drosophila melanogaster]	52945	-1,064	5,306E-18	Dm40675	1,114	8,965E-03
AAL29063.1	LD46766p [Drosophila melanogaster]	53844	-1,596	1,063E-153	Dm45977	2,314	4,910E-13
AAL39892.1	LP08646p [Drosophila melanogaster]	53898	-1,155	1,080E-19	Dm53962	2,628	1,342E-03
AAL40415.1	AF369891_4endonuclease/reverse transcriptase [Branchiostoma floridae]	Sg_CNS_NA4Plus6841	-0,929	0,000E+00	187594____k45_f_12718349	0,428	6,458E-04
AAL48714.1	RE15779p [Drosophila melanogaster]	82341	-0,477	1,200E-08	Dm57571	0,729	1,914E-28
AAM11327.1	GH01626p [Drosophila melanogaster]	52840	-0,551	4,758E-18	Dm73191	1,282	1,218E-04
AAM27504.1	LD12605p [Drosophila melanogaster]	53603	-1,187	3,533E-28	Dm75953	4,950	1,842E-17
AAM48347.1	HL01080p [Drosophila melanogaster]	53093	-1,557	1,755E-02	Dm79871	5,179	1,344E-49
AAM49882.1	LD14179p [Drosophila melanogaster]	53608	-0,578	3,976E-16	Dm81029	0,915	4,374E-02
AAM50797.1	LD24679p [Drosophila melanogaster]	53665	-2,879	2,773E-02	Dm83624	0,451	4,132E-04
AAM51136.1	SD26211p [Drosophila melanogaster]	83030	-1,385	7,682E-89	Dm85241	0,512	3,895E-02
AAN71572.1	RH39779p [Drosophila melanogaster]	82655	-1,996	2,960E-246	Dm112242	1,577	7,133E-03
AAO39460.1	RH34107p [Drosophila melanogaster]	82649	-0,746	1,015E-90	Dm125964	0,935	7,163E-04
AAQ22426.1	RH23644p [Drosophila melanogaster]	82629	-0,941	9,426E-166	Dm144063	2,144	7,698E-53
AAR86939.1	reverse transcriptase [Drosophila melanogaster]	39140	-0,479	2,007E-32	Dm157946	0,387	1,412E-08
AAS15657.1	RH44796p [Drosophila melanogaster]	82663	-2,557	9,849E-19	Dm161107	1,547	1,162E-42
AAT27274.1	RE22403p [Drosophila melanogaster]	82375	-0,651	1,604E-263	Dm166339	0,854	2,253E-05
AAX28844.1	reverse transcriptase [Drosophila melanogaster]	39667	-0,518	3,692E-02	Dm177022	0,359	9,647E-03
AAX33500.1	LP18549p [Drosophila melanogaster]	53931	-1,249	4,341E-13	Dm178075	1,057	2,408E-06
ABB36428.1	RE74917p [Drosophila melanogaster]	82592	-1,107	6,150E-12	Dm199001	1,956	2,143E-146
ABO52850.1	IP18018p [Drosophila melanogaster]	53455	-1,126	1,845E-213	Dm16837	1,498	2,103E-03

Accession number	BLAST Description	CNS Refference	CNS FC	CNS FDR	DT Refference	DT FC	DT FDR
ABY20489.1	IP22166p [Drosophila melanogaster]	53464	-1,601	0,000E+00	Dm39544	0,626	1,909E-05
ACH92236.1	FI03751p [Drosophila melanogaster]	52751	-2,016	2,541E-03	Dm69754	1,185	5,570E-06
ACH95265.1	FI03238p [Drosophila melanogaster]	52747	-0,590	1,856E-258	Dm74144	2,135	6,822E-11
ACI62137.1	polyprotein [Drosophila melanogaster]	35711	-0,551	4,981E-05	Dm75566	0,760	1,355E-08
ACR39391.1	odorant-binding protein 3b, partial [Locusta migratoria manilensis]	Sg_CNS_NA4Plus297	-0,547	1,549E-75	21415	38,012	2,069E-05
ACT67257.1	MIP12805p [Drosophila melanogaster]	54024	-1,109	3,488E-72	Dm113667	1,627	1,868E-17
ACU78068.1	hexamerin-like protein 1 [Locusta migratoria]	Sg_CNS_NA4Plus5757	-2,762	3,074E-16	27039	37,481	1,281E-03
ACX61601.1	RE23506p [Drosophila melanogaster]	82384	-1,051	1,452E-34	Dm118604	0,591	1,565E-35
ADE06724.1	FI14109p [Drosophila melanogaster]	52776	-0,466	5,861E-218	Dm130780	1,309	1,273E-04
ADI44149.1	MIP21321p [Drosophila melanogaster]	54047	-1,919	6,217E-14	Dm133887	3,274	4,695E-19
ADU79242.1	AT10981p [Drosophila melanogaster]	52112	-0,871	1,877E-39	Dm138450	0,620	6,252E-03
ADY39495.1	putative RNA-binding protein [Hottentotta judaicus]	Sg_CNS_NA4Plus37603	-1,354	0,000E+00	2241	0,420	2,157E-02
AEX33167.1	OBP11 protein [Locusta migratoria]	Sg_CNS_NA4Plus22563	-0,763	1,508E-53	69261	37,012	1,618E-02
BAA18916.1	lectin-related protein [Periplaneta americana]	Sg_CNS_NA4Plus29828	-1,075	0,000E+00	17699	6,221	8,558E-15
BAM35674.1	putative reverse transcriptase [Marsupenaeus japonicus]	Sg_CNS_NA4Plus22878	-0,525	1,464E-02	76397	0,368	5,075E-07
CAA16814.1	EG:95B7,6 [Drosophila melanogaster]	52712	-1,149	4,644E-06	Dm127410	1,464	1,033E-02
CAA70289.1	NADH-ubiquinone oxidoreductase acyl-carrier subunit [Drosophila melanogaster]	32553	-0,674	3,412E-29	Dm40249	1,761	2,380E-10
CAA73031.1	putative organic cation transporter [Drosophila melanogaster]	33523	-1,017	1,923E-26	Dm64711	1,197	2,692E-03
CAP09075.1	minos transposase [Drosophila hydei]	31435	-1,008	0,000E+00	Dm31065	0,244	2,739E-03
EEZ99133.1	hypothetical protein TcasGA2_TC012914 [Tribolium castaneum]	23244	-0,836	2,690E-140	1026	0,255	3,770E-02
EEZ99812.1	hypothetical protein TcasGA2_TC002592 [Tribolium castaneum]	6861	-0,850	1,264E-171	28745	0,725	3,835E-03
EFA00566.1	hypothetical protein TcasGA2_TC003436 [Tribolium castaneum]	7549	-1,494	0,000E+00	28910	1,002	3,633E-02
EFA09378.1	hypothetical protein TcasGA2_TC001939 [Tribolium castaneum]	3269	-0,819	3,075E-35	30228	0,665	3,191E-02
EFA11632.1	hypothetical protein TcasGA2_TC000010 [Tribolium castaneum]	368	-0,439	8,085E-95	30819	0,429	4,321E-03
EFA11829.1	hypothetical protein TcasGA2_TC005278 [Tribolium castaneum]	13224	-2,214	1,534E-06	1155	1,569	1,012E-35
EFA12688.1	hypothetical protein TcasGA2_TC001995 [Tribolium castaneum]	3832	-0,826	0,000E+00	32768	1,938	6,518E-37
EFA13138.1	hypothetical protein TcasGA2_TC002010 [Tribolium castaneum]	3904	-0,470	4,406E-15	1210	0,688	9,447E-04

Accession number	BLAST Description	CNS Refference	CNS FC	CNS FDR	DT Refference	DT FC	DT FDR
EFA13284.1	hypothetical protein TcasGA2_TC010304 [Tribolium castaneum]	19261	-0,847	0,000E+00	1220	0,145	1,077E-03
EFN65035.1	hypothetical protein EAG_01178 [Camponotus floridanus]	Sg_CNS_NA4Plus968	-0,276	1,847E-127	1562	1,479	5,831E-129
EFN72056.1	hypothetical protein EAG_14155 [Camponotus floridanus]	73034	-0,434	6,055E-05	1668	0,797	1,500E-08
EFN89173.1	hypothetical protein EAI_12527 [Harpegnathos saltator]	Sg_CNS_NA4Plus35420	-0,791	7,209E-49	50049	0,600	6,781E-05
NP_001138042.1	CG14394, isoform F [Drosophila melanogaster]	28271	-1,315	1,339E-20	Dm91538	0,734	1,566E-03
NP_001246122.1	CG43346, isoform B [Drosophila melanogaster]	28216	-1,000	3,380E-67	Dm155174	0,398	2,853E-02
NP_048267.1	ORF MSV196 ALI motif gene family protein [Melanoplus sanguinipes entomopoxvirus]	Sg_CNS_NA4Plus483	-0,647	5,159E-75	176110___330022	0,677	2,619E-02
NP_609244.1	CG9287 [Drosophila melanogaster]	52490	-0,455	7,625E-21	Dm99721	1,262	1,243E-03
NP_610270.1	CG1707 [Drosophila melanogaster]	52413	-0,293	1,594E-13	Dm70934	2,092	6,960E-13
NP_649023.1	CG7402 [Drosophila melanogaster]	52477	-0,533	5,672E-10	Dm108296	2,805	3,777E-08
NP_650519.1	CG10407 [Drosophila melanogaster]	52377	-1,602	6,622E-10	Dm103586	1,310	7,289E-12
O46162.1	Serine protease inhibitor I/II [Schistocerca gregaria]	Sg_CNS_NA4Plus16535	-1,784	0,000E+00	86493	0,482	2,382E-05
P16423.1	Retrovirus-related Pol polyprotein from type-2 retrotransposable element R2DM [Drosophila melanogaster]	35472	-1,473	2,911E-33	Dm14801	0,267	5,194E-04
P16425.1	Putative 115 kDa protein in type-1 retrotransposable element R1DM [Drosophila melanogaster]	49772	-0,715	7,968E-05	Dm18560	0,862	3,461E-06
P52905.1	Trypsin iota [Drosophila melanogaster]	46974	-1,758	5,764E-03	Dm47872	1,703	2,534E-18
P82384.2	Larval cuticle protein 9 [Drosophila melanogaster]	29324	-1,428	1,517E-07	Dm167130	1,433	2,939E-02
Q24048.2	Sodium/potassium-transporting ATPase subunit beta-2 [Drosophila melanogaster]	86216	-0,841	2,575E-51	Dm11621	2,916	2,400E-12
Q25313.1	Putative defense protein [Locusta migratoria]	Sg_CNS_NA4Plus893	-0,538	0,000E+00	11614	0,680	3,655E-09
Q26365.4	ADP,ATP carrier protein [Drosophila melanogaster]	84025	-1,255	2,433E-64	Dm204585	1,153	3,118E-55
Q7JWG9.1	39S ribosomal protein L52, mitochondrial [Drosophila melanogaster]	41329	-0,663	2,900E-51	Dm9008	1,183	3,294E-02
Q8MKK4.1	Facilitated trehalose transporter Tret1-2 homolog [Drosophila melanogaster]	46761	-1,228	1,536E-121	Dm195811	0,696	1,067E-13
Q9VA27.1	Cytochrome P450 4c3 [Drosophila melanogaster]	90582	-0,320	9,572E-04	Dm11780	0,753	1,445E-20
Q9VFJ3.1	Serine protease HTRA2, mitochondrial [Drosophila melanogaster]	108506	-1,200	2,003E-54	Dm191752	1,314	3,299E-02
Q9VFP1.1	Probable cytochrome P450 6d5 [Drosophila melanogaster]	90586	-2,201	7,682E-08	Dm5011	1,081	1,397E-16
Q9VG17.2	Probable cytochrome P450 304a1 [Drosophila melanogaster]	90580	-0,800	0,000E+00	Dm4306	0,842	5,572E-31
Q9VTZ6.1	Probable phosphomannomutase [Drosophila melanogaster]	34363	-0,339	1,630E-05	Dm11046	0,598	3,153E-06

Accession number	BLAST Description	CNS Refference	CNS FC	CNS FDR	DT Refference	DT FC	DT FDR
Q9VYF8.1	2-methoxy-6-polyprenyl-1,4-benzoquinol methylase, mitochondrial [Drosophila melanogaster]	90499	-0,270	9,372E-14	Dm86614	1,926	1,656E-15
Q9VYY4.1	Cytochrome P450 4g15 [Drosophila melanogaster]	86710	-0,252	2,949E-03	Dm4401	4,179	1,937E-22
Q9W1C9.2	Ejaculatory bulb-specific protein 3 [Drosophila melanogaster]	33842	-0,954	0,000E+00	Dm131614	1,463	1,131E-179
S60466	S60466transposase - fruit fly (Drosophila melanogaster) transposon element S	41987	-0,546	6,644E-16	Dm79396	0,719	5,988E-04
XP_001943110.2	PREDICTED: inter-alpha-trypsin inhibitor heavy chain H3 isoform X2 [Acyrthosiphon pisum]	28047	-0,794	0,000E+00	57737	0,724	2,941E-02
XP_003707219.1	PREDICTED: uncharacterized protein LOC100883491 [Megachile rotundata]	54215	-0,430	1,351E-02	72936	1,294	1,310E-03
XP_003729088.1	PREDICTED: RNA-directed DNA polymerase from mobile element jockey-like [Strongylocentrotus purpuratus]	Sg_CNS_NA4Plus937	-0,573	7,551E-47	250850____k25_f_25558396	0,709	1,109E-02
XP_003729561.1	PREDICTED: uncharacterized protein LOC100891120 [Strongylocentrotus purpuratus]	Sg_CNS_NA4Plus7648	-0,996	7,108E-50	73880____k37_f_16221168	0,414	1,359E-02
XP_003730218.1	PREDICTED: uncharacterized protein LOC100889850 [Strongylocentrotus purpuratus]	Sg_CNS_NA4Plus7876	-0,806	5,313E-220	52710____87945	0,277	5,730E-03
XP_003737968.1	PREDICTED: uncharacterized protein LOC100904251 [Metaseiulus occidentalis]	77893	-1,942	8,667E-173	74117	1,899	2,424E-04
XP_003742976.1	PREDICTED: protein GTLF3B-like [Metaseiulus occidentalis]	Sg_CNS_NA4Plus1341	-0,287	6,291E-06	75703	3,899	3,253E-02
XP_968005.1	PREDICTED: serpin B6 [Tribolium castaneum]	42452	-1,683	9,302E-244	114921	0,865	1,335E-02
XP_969001.1	PREDICTED: alkylglycerol monooxygenase [Tribolium castaneum]	46164	-1,550	5,387E-20	11757	1,355	3,586E-06

Table 4.39: Transcripts that show over-expression during the solitarious phase in the CNS transcriptome and during the gregarious phase in the DT transcriptome. The results are limited to the sequences that share BLAST accession number between CNS and DT transcriptomes.

Accession number	BLAST Description	CNS Refference	CNS FC	CNS FDR	DT Refference	DT FC	DT FDR
AAL29107.1	LP10918p [Drosophila melanogaster]	53916	4,441	4,279E-96	Dm46370	-2,311	4,661E-31
AAL49280.1	RE74144p [Drosophila melanogaster]	82586	4,397	1,828E-08	Dm60320	-2,508	5,921E-11
AAM50819.1	LD37859p [Drosophila melanogaster]	53753	0,044	2,651E-03	Dm84012	-1,296	3,758E-23
AAS15661.1	RH08789p [Drosophila melanogaster]	82609	0,738	6,368E-106	Dm161150	-0,850	1,211E-03
ABK29471.1	CHK1 checkpoint-like protein [Helicoverpa armigera]	Sg_CNS_NA4Plus4828	0,731	0,000E+00	39	-1,838	2,850E-06
ACS68170.1	TA01007p [Drosophila melanogaster]	83083	1,128	0,000E+00	Dm98624	-1,429	1,234E-03
CAM36311.1	hypothetical protein [Thermobia domestica]	Sg_CNS_NA4Plus15493	1,003	0,000E+00	13109	-1,272	0,000E+00
EFX62991.1	hypothetical protein DAPPUDRAFT_67491 [Daphnia pulex]	Sg_CNS_NA4Plus7868	0,528	0,000E+00	2022	-1,525	0,000E+00
XP_001605497.2	PREDICTED: transitional endoplasmic reticulum ATPase TER94 [Nasonia vitripennis]	45922	1,469	0,000E+00	65050	-1,036	8,013E-08

Table 4.40: List of transcripts cited in the bibliography and comparison of their expression pattern between bibliography and CNS and DT transcriptomes. The transcripts are sorted considering the bibliographical reference that cites them. NS indicates the lack of statistical significance.

Transcript name	Bibliography	CNS RNA-seq	DT RNA-seq
	Kang et al. [2004]		
annexin IX	Gregarious	Gregarious	Gregarious
beta-N-acetylhexosaminidase activity	Gregarious	Absent	Solitarious
cytochrome c oxidase chain 3	Gregarious	Absent	Solitarious NS
larval cuticle protein precursor	Solitarious	Gregarious	Solitarious
NADP-dependent malic enzyme	Gregarious	Gregarious NS	Solitarious NS
putative fatty acid elongase	Gregarious	Absent	Gregarious NS
similar to asparagine synthetase	Solitarious	Solitarious NS	No differences
similar to brain adenylate cyclase 1	Gregarious	Gregarious	No differences
troponin C	Solitarious	Gregarious	Solitarious
	Chen et al. [2010]		
adipokinetic hormone receptor	Gregarious	Gregarious	Absent
allatostatin receptor	Gregarious	Absent	No differences
blot	Gregarious	Absent	No differences
cask	Gregarious	Gregarious	Solitarious NS
corazonin receptor	Gregarious	Gregarious NS	Absent
diuretic hormone receptor	Gregarious	Gregarious	Gregarious NS
fmr1	Gregarious	Absent	Solitarious NS
GABA receptor	Gregarious	Gregarious	Solitarious NS
glutamate receptor	Gregarious	Gregarious	No differences
malvolio	Gregarious	Gregarious	Gregarious NS
neurotransmitter transporter	Gregarious	Gregarious NS	Solitarious NS

Transcript name	Bibliography	CNS RNA-seq	DT RNA-seq
octopamine receptor	Gregarious	Gregarious	Solitarious NS
serotonin receptor	Gregarious	Gregarious NS	No differences
synapsin	Gregarious	Gregarious NS	Absent
synaptic vesicle 2-related protein	Gregarious	Absent	No differences
synaptic vesicle protein	Gregarious	Gregarious	No differences
synaptobrevin	Solitarious	Gregarious	Solitarious
tyrosine hydroxylase	Gregarious	Solitarious	Gregarious NS
vesamicol binding protein	Gregarious	Solitarious NS	Gregarious
Guo et al. [2011]			
alcohol dehydrogenase	Gregarious	Solitarious	Gregarious NS
Alpha crystallin	Solitarious	Absent	Gregarious NS
arylphorin hexamerin-like protein 2	Solitarious	Solitarious NS	Solitarious NS
ATP-citrate synthase	Gregarious	Gregarious	Gregarious NS
ATP-dependent RNA helicase p62	Solitarious	Gregarious	Solitarious NS
cellular retinaldehyde-binding protein	Solitarious	Solitarious NS	Solitarious NS
chemosensory protein 1	Gregarious	Solitarious NS	Absent
chemosensory protein 2	Gregarious	Gregarious	Absent
chemosensory protein 3	Gregarious	Gregarious	Absent
chemosensory protein 4	Gregarious	Gregarious	Absent
csp precursor (various)	Gregarious	Gregarious	Absent
cystathionine gamma-lyase	Gregarious	Absent	Gregarious NS
cytochrome p450 4g15	Solitarious	Gregarious	Solitarious
cytochrome p450 6a2	Solitarious	Gregarious NS	Solitarious
Fatty acid desaturase	Solitarious	Absent	Solitarious NS
GTP-binding protein Rheb homolog	Solitarious	Absent	Gregarious NS
Heat shock protein Hsp20	Solitarious	Gregarious NS	Solitarious
hexamerin precursor	Solitarious	Solitarious	Absent

Transcript name	Bibliography	CNS RNA-seq	DT RNA-seq
lambda-crystallin homolog	Gregarious	Solitarious	No differences
lethal 2 essential for life protein	Solitarious	Absent	Solitarious
Pacifastin inhibitor	Gregarious	Gregarious	Gregarious
Peptidase M2, peptidyl-dipeptidase A	Gregarious	Absent	No differences
peptidyl-prolyl cis-trans isomerase	Solitarious	Gregarious	Solitarious
probable cytochrome p450 4ac1	Solitarious	Absent	Solitarious NS
protein TU-36B	Solitarious	Absent	Solitarious
purine nucleoside phosphorylase	Solitarious	Gregarious NS	No differences
takeout 1	Solitarious	Solitarious	Solitarious NS
Badisco et al. [2011a]			
fasciclin-like precursor	Gregarious	Gregarious	Solitarious
probable cytochrome p450	Gregarious	Gregarious	Absent
RNA helicase ddx1	Solitarious	Gregarious	Gregarious NS
slit homologue	Gregarious	Gregarious	Solitarious
sparc	Gregarious	Solitarious	No differences
Badisco et al. [2011b]			
5-oxoprolinase	Solitarious	Gregarious	No differences
arrestin	Gregarious	Gregarious NS	Solitarious NS
arylphorin-like	Solitarious	Solitarious	Absent
chromodomain helicase DNA binding protein	Gregarious	Absent	Gregarious
Cullin-3	Gregarious	Absent	No differences
Glia Maturation Factor	Solitarious	Gregarious	No differences
heat-shock proteins	Gregarious	Gregarious	Absent
microsomal glutathione-S-transferase	Solitarious	Solitarious	Gregarious
mitochondrial ATP-synthase coupling factor 6	Solitarious	Gregarious	Gregarious NS
Musashi	Gregarious	Gregarious	Solitarious NS
Osa	Gregarious	Absent	Gregarious NS

Transcript name	Bibliography	CNS RNA-seq	DT RNA-seq
Pacifastin-like peptide precursor 4	Gregarious	Gregarious	Gregarious
Pasilla	Gregarious	Absent	Gregarious NS
peroxiredoxin	Solitarious	Gregarious	Gregarious
ribophorin	Solitarious	Absent	No differences
Slowpoke homologue	Solitarious	Gregarious NS	Solitarious NS
thaumatin-like protein	Gregarious	Absent	Solitarious
transaldolase homologue	Solitarious	Gregarious NS	Gregarious NS
transient receptor potential-like	Gregarious	Absent	Solitarious NS
Wang et al. [2014b]			
Aladin	Solitarious	Gregarious	Solitarious NS
Ankyrin-2	Solitarious	Gregarious	Solitarious NS
ATP-binding cassette sub-family D member 3	Gregarious	Gregarious NS	Absent
Basement membrane-specific heparan sulfate proteoglycan core protein	Gregarious	Gregarious	Gregarious NS
Brefeldin A-inhibited guanine nucleotide-exchange protein 3	Solitarious	Gregarious	Solitarious NS
Dynein heavy chain, cytoplasmic	Gregarious	Gregarious	Gregarious NS
E3 ubiquitin-protein ligase hyd	Gregarious	Gregarious	Gregarious NS
Early endosome antigen 1	Solitarious	Absent	Gregarious NS
Endophilin-B1	Solitarious	Gregarious	Gregarious NS
Eukaryotic translation initiation factor 4B	Solitarious	Absent	Gregarious NS
Glycyl-tRNA synthetase	Gregarious	Gregarious	No differences
GRIP and coiled-coil domain-containing protein 1	Solitarious	Absent	Gregarious NS

Transcript name	Bibliography	CNS RNA-seq	DT RNA-seq
HEAT repeat-containing protein 5B	Gregarious	Solitarious NS	Solitarious NS
Kinesin-associated protein 3	Solitarious	Gregarious	Solitarious NS
mitogen-activated protein-binding protein-interacting protein	Gregarious	Gregarious	No differences
Niemann-Pick C1 protein	Gregarious	Gregarious	Solitarious NS
NUAK family SNF1-like kinase 1	Solitarious	Absent	Gregarious NS
Phosphatidylinositol glycan anchor biosynthesis class U protein	Gregarious	Absent	Gregarious NS
Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit delta isoform	Solitarious	Gregarious	Gregarious NS
Phosphatidylinositol-4-phosphate 5-kinase type-1 alpha	Gregarious	Absent	Gregarious NS
Polyhomeotic-like protein 1	Gregarious	Absent	No differences
Polyphosphoinositide phosphatase	Gregarious	Gregarious	Solitarious NS
PR domain zinc finger protein 4	Solitarious	Absent	Solitarious NS
Probable phospholipid-transporting ATPase IA	Solitarious	Gregarious	Gregarious NS
Probable protein-cystein N-palmitoyltransferase porcupine	Gregarious	Absent	Solitarious NS
Probable uridine-cytidine kinase	Solitarious	Gregarious NS	Gregarious NS
Protein Daple	Solitarious	Absent	Gregarious NS
protein FAN	Gregarious	Gregarious NS	Gregarious NS
Protein son of sevenless	Solitarious	Absent	Solitarious NS
Rho GTPase-activating protein 190	Solitarious	Gregarious	Solitarious NS
Ribosome biogenesis protein BOP1 homolog	Solitarious	Absent	Solitarious NS

Transcript name	Bibliography	CNS RNA-seq	DT RNA-seq
RING finger protein unkempt	Solitarious	Gregarious	Gregarious
RNA exonuclease 1 homolog	Solitarious	Gregarious	Solitarious NS
SAP domain-containing ribonucleoprotein	Solitarious	Absent	Solitarious NS
Serine/threonine-protein kinase mTOR	Gregarious	Gregarious	No differences
Serine–pyruvate aminotransferase, mitochondrial	Solitarious	Gregarious NS	Solitarious NS
Sterile alpha and TIR motif-containing protein 1-like	Solitarious	Gregarious	No differences
Switch-associated protein 70	Gregarious	Gregarious	Absent
Telomerase-binding protein EST1A	Solitarious	Gregarious	Gregarious NS
Tubulin glycylase 3B	Gregarious	Absent	No differences
Ubiquitin-protein ligase E3B	Gregarious	Absent	Gregarious NS
Unc-112-related protein	Gregarious	Gregarious	No differences
Uncharacterized protein KIAA0467	Solitarious	Gregarious NS	Solitarious NS
Vacuolar protein sorting-associated protein 16 homolog	Gregarious	Gregarious	Absent
Zinc finger protein 91	Solitarious	Gregarious	Solitarious NS

Transcript name	Expresion pattern				CNS transcriptome					DT transcriptome				
	Cited in	Bibliography	CNS RNA-seq	DT RNA-seq	Sequence ID	Accession number	Solitarious FPKM	Gregarious FPKM	Log(FC)	Sequence ID	Accession number	Solitarious FPKM	Gregarious FPKM	Log(FC)
Annexin IX	Kang et al. 2004	Gregarious	Gregarious	Gregarious	85392	XP_967931	50,479	96,925	-0,941	150444	XP_008471197	3,231	15,159	-2,230
Pacifastin-like peptide precursor 4	Guo et al. 2011, Badisco et al. 2011b	Gregarious	Gregarious	Gregarious	77531	CAC82510.3	0,325	7,136	-4,457	90431	CAC82510.3	5,164	12,995	-1,331
Arylphorin hexamerin-like protein 2	Guo et al. 2011	Solitaroius	Solitaroius	Not significant	85614	XP_001600430	324,993	148,461	1,130	4419	AAX14951	0,055	0,001	5,769
ATP-citrate synthase	Guo et al. 2011	Gregarious	Gregarious	Non-significant	85903	XP_003425261	17,466	29,555	-0,759	65031	XP_003425261	0,001	5,429	-12,406
Basement membrane-specific heparan sulfate proteoglycan core protein	Wang et al. 2014	Gregarious	Gregarious	Non-significant	86324	XP_393220	0,481	0,930	-0,949	114214	XP_008193278	0,001	0,300	-8,228
Cellular retinaldehyde-binding protein	Guo et al. 2011	Solitaroius	Non-significant	Non-significant	87267	XP_392138	0,435	0,046	3,251	Dm60670	NP_523939	4,037	0,001	11,979
Diuretic hormone receptor	Chen et al. 2010	Gregarious	Gregarious	Non-significant	92428	XP_003427176	0,147	0,647	-2,142	58789	EGI58221	0,001	0,655	-9,355
Dynein heavy chain, cytoplasmic	Wang et al. 2014	Gregarious	Gregarious	Non-significant	93074	XP_001951535	22,682	48,572	-1,099	75468	XP_003742362	0,001	4,928	-12,267
E3 ubiquitin-protein ligase hyd	Wang et al. 2014	Gregarious	Gregarious	Non-significant	93157	XP_001605335	0,964	1,964	-1,027	Dm186987	P51592	1,103	1,208	-0,132
Malvolio	Chen et al. 2010	Gregarious	Gregarious	Non-significant	37276	XP_003424930	2,140	5,091	-1,250	Dm153651	NP_996251	8,568	9,584	-0,162
Protein FAN	Wang et al. 2014	Gregarious	Non-significant	Non-significant	37135	XP_001120591	0,371	0,614	-0,727	26467	XP_002432451	0,001	4,932	-12,268
Takeout 1	Guo et al. 2011	Solitarious	Solitarious	Non-significant	37604	XP_001950683	49,137	38,401	0,356	6593	KDR17338	13,155	8,228	0,677

Table 4.41: Details about the transcripts that show the same expression pattern in the bibliography, our CNS transcriptome and our DT transcriptome. In addition to expression pattern in these three sources, the tables shows the reference number, the accession number, the FPKM values, and the fold-change value in each transcriptome.

Table 4.42: Details of the sequences highlighted in Zhang et al. [2012] study of *L. migratoria*'s CNS that were present in our *S. gregaria* CNS transcriptome. The search was based on best BLAST result's name.

Described in Zhang et al. [2012]	Expression in our CNS transcriptome	*S. gregaria* CNS Refference	Log(FC)	FDR	Expression in our DT transcriptome	*S. gregaria* DT Refference	Log(FC)	FDR
Acetylcholine receptor protein subunit delta	Gregarious	83493	-0,562	6,659E-05	—	—	—	—
Acetylcholinesterase	No significant	83498	-1,972	2,152E-01	No significant	100116	-16,610	2,855E+00
Calcineurin subunit B	No significant	86773	-0,557	1,004E-01	No significant	Dm153915	-0,138	2,995E+00
Chloride channel protein ClC-Ka	Gregarious	88275	-1,410	3,211E-10	—	—	—	—
Circadian clock protein PERIOD	Gregarious	Sg_CNS_NA4Plus29518	-2,437	7,725E-70	Solitarious	58	0,858	1,062E-07
Circadian clock protein TIMELESS	Gregarious	45485	-1,220	7,900E-08	No significant	2240	-0,918	1,686E+00
	No significant	45487	0,443	1,958E+00	—	—	—	—
	Gregarious	88404	-0,972	4,335E-02	—	—	—	—
Cysteine rich motor neuron	Gregarious	91101	-1,257	5,039E-223	No significant	6240	-16,610	2,855E+00
Dopamine beta-hydroxylase precursor	Gregarious	92790	-0,784	3,304E-02	—	—	—	—
Dopamine transporter	No significant	92794	-1,557	9,848E-01	No significant	Dm46943	-18,610	1,326E+00
	No significant	92795	0,028	1,562E+00	No significant	Dm201431	0,000	2,394E+00
Foraging	No significant	102055	-0,034	1,126E+00	Solitarious	Dm95032	1,663	5,462E-03
GABA neuro-transmitter transporter-1A	Gregarious	102262	-0,219	2,598E-02	—	—	—	—
Glutamate receptor 1	No significant	103417	0,166	1,776E+00	—	—	—	—
	No significant	103424	-0,345	1,342E+00	—	—	—	—
Neuroglian precursor	Gregarious	79769	-1,763	9,104E-04	No significant	25596	0,000	2,394E+00
Nicotinic acetylcholine receptor alpha 9	Gregarious	32783	-0,500	1,246E-28	No significant	92562	-0,263	3,013E+00
Nitric oxide synthase	Gregarious	32828	-1,021	2,032E-02	No significant	60180	0,000	2,394E+00
	Gregarious	32838	-0,313	3,662E-19	No significant	141610	0,000	2,394E+00
	Gregarious	86983	-3,142	7,167E-03	No significant	Dm64407	-0,755	2,403E-01
NMDA-type glutamate receptor 1	No significant	32855	-0,631	6,543E-01	—	—	—	—
Preprotachykinin	No significant	44075	-0,195	5,029E-02	Gregarious	255993	-1,044	1,529E-13
	—	—	—	—	Gregarious	4380	-1,441	7,035E-05
Slowpoke protein	No significant	86816	-1,142	1,562E+00	No significant	Dm179849	0,000	2,394E+00
	—	—	—	—	No significant	Dm179850	0,415	2,445E+00
	—	—	—	—	No significant	Dm179855	18,932	7,412E-01
	—	—	—	—	No significant	Dm179856	0,415	2,999E+00
	—	—	—	—	No significant	Dm179860	1,000	2,933E+00
	—	—	—	—	No significant	Dm179861	-16,610	2,855E+00
	—	—	—	—	No significant	Dm104561	17,610	2,454E+00
	—	—	—	—	No significant	Dm163860	-16,610	2,855E+00
Sodium channel	No significant	23321	21,375	7,265E-01	No significant	102792	0,000	2,394E+00
	Gregarious	33717	-1,189	2,187E-02	No significant	121386	0,585	2,975E+00
	No significant	33719	0,443	1,958E+00	No significant	139717	0,000	2,394E+00
	No significant	43008	1,559	4,109E-01	—	—	—	—
	No significant	43020	1,443	1,514E+00	—	—	—	—
	No significant	43023	-2,142	9,852E-01	—	—	—	—

Described in Zhang et al. [2012]	Expression in our CNS transcriptome	*S. gregaria* CNS Refference	Log(FC)	FDR	Expression in our DT transcriptome	*S. gregaria* DT Refference	Log(FC)	FDR
	No significant	49468	-0,779	1,217E+00	—	—	—	—
	No significant	49526	-0,557	1,939E+00	—	—	—	—
	No significant	103726	1,443	1,514E+00	—	—	—	—
Synaptic glycoprotein SC2	Gregarious	43962	-0,489	8,116E-78	No significant	44755	0,000	2,394E+00
Synaptic vesicle protein	Gregarious	43963	-0,989	2,253E-17	No significant	Dm189749	0,000	2,394E+00
	No significant	43978	-0,557	7,243E-01	No significant	4124	0,000	2,394E+00
	Gregarious	43982	-0,535	1,581E-14	No significant	225	0,000	2,394E+00
	—	—	—	—	No significant	14753	0,000	2,394E+00
	—	—	—	—	Solitarious	226	0,930	1,573E-10
	—	—	—	—	No significant	16053	0,000	2,394E+00
	—	—	—	—	No significant	16054	0,000	2,394E+00
	—	—	—	—	No significant	16055	0,000	2,394E+00
	—	—	—	—	No significant	115965	1,000	2,933E+00
	—	—	—	—	No significant	16698	0,000	2,394E+00
	—	—	—	—	No significant	23772	0,000	2,394E+00
	—	—	—	—	No significant	25234	0,000	2,394E+00
Voltage dependent L-type calcium channel	Gregarious	49487	-1,654	2,308E-11	—	—	—	—
	Gregarious	49489	-1,096	2,067E-06	—	—	—	—
	No significant	49491	-1,727	2,348E-01	—	—	—	—
Voltage-activated ion channel	Gregarious	49473	-2,142	4,430E-02	—	—	—	—
	No significant	49474	-1,142	1,562E+00	—	—	—	—
Potassium channel	No significant	86816	-1,024	1,562E+00	No significant	121585	0,000	2,394E+00
	No significant	49518	-19,048	1,981E-01	No significant	134604	-1,000	2,933E+00
	No significant	35737	-2,761	2,862E-01	No significant	Dm167333	0,000	2,394E+00
	No significant	42923	-0,397	5,220E-01	No significant	123890	0,000	2,394E+00
	No significant	47135	-0,024	1,788E+00	No significant	127118	-0,241	2,839E+00
	No significant	42926	-1,761	5,918E-01	No significant	143838	0,000	2,394E+00
	No significant	35736	-2,761	2,862E-01	No significant	Dm61598	-0,302	2,387E+00
	No significant	86820	-0,542	3,611E-01	No significant	Dm15441	0,000	2,527E+00
	No significant	86132	-1,277	7,834E-10	No significant	91955	1,379	8,299E-01
	Gregarious	44040	-1,237	0,000E+00	No significant	112887	0,000	2,394E+00
	Gregarious	35738	-0,599	7,276E-55	No significant	113619	0,000	2,394E+00
	No significant	86815	-0,303	5,624E-02	No significant	152177	0,000	2,394E+00
	Gregarious	86811	-1,302	3,914E-20	—	—	—	—
Glutamate receptor	No significant	103391	0,146	1,562E+00	No significant	Dm184938	0,000	2,394E+00
	No significant	103413	-15,684	1,981E-01	No significant	15004	0,000	2,394E+00
	No significant	103392	-0,439	1,867E+00	No significant	25784	0,000	2,394E+00
	No significant	103415	-1,608	2,348E-01	No significant	86098	0,000	2,394E+00
	No significant	32855	-0,513	6,543E-01	—	—	—	—
	No significant	103410	-0,801	1,402E-01	—	—	—	—
	Gregarious	103427	-2,164	9,208E-14	—	—	—	—
	Gregarious	103399	-1,547	1,478E-07	—	—	—	—
	Gregarious	28094	-0,656	6,166E-13	—	—	—	—
	Gregarious	103396	-0,869	4,417E-248	—	—	—	—
	Gregarious	103430	-0,518	0,000E+00	—	—	—	—

Chapter 5: Diversity, evolution and expression of chemosensory proteins (*CSP* family) of the main pest locust

5.1 Introduction

Olfactory protein families such as the chemosensory proteins (CSPs) are among the molecules in charge of the perception of environmental stimuli that have a high probability to be involved in the locust phase change. In fact, over-expression of a CSP gene, namely *LmigCSP3*, has already been reported in gregarious *L. migratoria* [Guo et al., 2011]. The same work further showed that knocking down *LmigCSP3* expression in gregarious *L. migratoria* nymphs led to a decreased ability to detect volatiles and to agregate with other conspecifics. CSPs seem therefore to be part of the initial set of switches for turning the phase change on.

CSPs are a conserved family of soluble proteins involved, as their name suggests, in chemoreception. They have been found in different numbers (variants) in all Arthropoda orders. They are related to another family of chemoreceptors called odorant binding proteins (OBPs) [Pikielny et al., 1994]. Contrary to OBPs, CSPs seem specific to the phylum Arthropoda and have no known homology to any known mammalian protein sequence [McKenna et al., 1994, Vieira and Rozas, 2011]. These proteins share a conserved amino acids pattern consisting of a predicted N-terminal signal peptide region, a conserved pattern composed of a cysteine followed by 6-8 residuals, another cysteine, 18 residuals, another cysteine, 2 residuals then a fourth cysteine (Cys-X_{6-8}-Cys-X_{18}-Cys-X_2-Cys) [Angeli et al., 1999]. CSPs are also characterized by the presence of six α-helices. The signal peptide might act as a transmembrane protein interactor, and the CSP can be included in the endoplasmic reticulum for its later secretion via vesicles. Although the four cysteines pattern allows the formation of disulfide bridges between cysteines 1 and 2 and between cysteines 3 and 4, it is confirmed that the secondary structure of a CSP, in form of a globular protein with a ligand interacting cavity, depends on the position of its α-helices [Campanacci et al., 2003]. CSPs are eminently present in the haemolymph of the insect sensilla [Pelosi, 1996]. They bind to olfaction-related ligands [Ban et al., 2002] and interact with GPCRs of the olfactory neurons in order to activate a phosphorilation cascade [Mombaerts, 1999]. Some CSP homologous sequences have been found to be expressed in the ejaculatory bulb of *Drosophila melanogaster* [Bohbot et al., 1998], in the legs of the cockroach *Periplaneta americana* [Kitabayashi et al., 1998] and in embryonic states of the honey bee *Apis mellifera* [Maleszka et al., 2007]. These have likely conserved the capacity of ligand-binding and interaction with GPCRs but seem to have acquired novel biological functions, including in reproduction [Bohbot et al., 1998], regeneration [Kitabayashi et al., 1998] and embryonic development [Maleszka et al., 2007], respectively.

Although their phylogenetic origin remains unsolved, CSPs come in several and different copies and numbers in different arthropods. They very likely have followed a birth and death evolutionary dynamic [Vieira and Rozas, 2011]. The known number of CSPs in insects ranges from three to twenty-two [Vieira and Rozas, 2011]. Some groups, like Drosophila genus, have a reduced number [Sanchez-Gracia et al., 2009], whereas others (such ants or butterflies) seem to have undergone an increase in the copy number of their genomic CSPs [Gong et al., 2007, Kulmuni and Havukainen, 2013, Li et al., 2015]. In addition to the five CSPs reported for *S. gregaria* in [Angeli et al., 1999] and the other five reported for *L. migratoria* in [Picimbon et al., 2000], a search into the GenBank databases reported another 15 putative CSPs in the latter species [Ban et al., 2003]. Both sets of sequences were obtained by sequencing cloned cDNAs, meaning that they come from genuine expressed genes. Still, intra-specific comparison of the five *S. gregaria* and five *L. migratoria* CSP sequences casts high identity values [Angeli et al., 1999, Picimbon et al., 2000], meaning that some might be alleles of the same gene or gene copies with a very recent phylogenetic origin. Estimating the number of all or almost all the CSP sequences of *L. migratoria* is now feasible thanks to the availability of transcriptomics data [Kang et al., 2004, Chen et al., 2010] and a draft genome [Wang et al., 2014a]. For *S. gregaria*, however, the definitive determination of the total number of genomic CSP copies will have to wait until a genome assembly is available and, for now, CSP detection in that species has to be based on the available trancriptomics data from [Badisco et al., 2011a] and from our own laboratory (as seen in Chapters 2 and 3).

Being of such potential importance for the development of the locust gregarious phase, the determination of which of the detected locust CSPs might be involved in locust gregariousness is of obvious relevance and can be achieved in at least two ways. One is a functional genomics approach, direct but laborious if not preceded by preliminary data on what CSP to use and how, where, and when to observe. That approach is based on knockdown of each CSP gene in different individuals, stages and phases followed by in-depth examination of the knockdown phenotypes. A second approach, indirect and preliminary but less laborious and its data might guide the subsequent functional genomics works, could be based on a comparative sequence and expression phylogenetic study of the CSPs of different locusts. Microarrays revealed changes in the expression of some CSPs between solitarious and gregarious states in *L. migratoria*, and a functional study (RNAi) of one of them, *LmigCSP3*, showed that it is involved in the detection of conspecifics and gregarization. Still, *L. migratoria* genome might provide more CSP sequences than the reported ones and more data on the number and classification of *S. gregaria* CSPs are needed in order to identify and characterize all of these locusts CSPs and infer more CSPs with potential involvement in locust phase change. Here we carry out a phylogenetic comparative analysis of the locust CSPs sequence and expression patterns. We compare the CSPs obtained from the published works and databases and by high throughput sequencing on the two most important locust species: *S. gregaria* (transcriptomic data) and *L. migratoria* (transcriptomic and genome sequencing data). Our first objective is to present an accurate number of CSP paralogs in *L. migratoria* based on homology searches of the contigs of its draft genome as available in 2016. We also identified all the transcribed CSPs from solitarious and gregarious transcriptomes of different *S. gregaria* tissues and compared the *L. migratoria* data with those of *S. gregaria* in order to characterize homology between the CSPs of both species. A more global phylogenetic reconstruction using the *S. gregaria* and *L. migratoria* as well as several confirmed CSPs from other arthropod species allowed us to identify and differentiate the different CSP lineages and detect possible genomic duplication events in the studied locusts. Once the two locust CSP sequences were characterized, we checked wether there is a relationship between their phylogenetic and solitarious versus gregarious expression patterns in the two species. CSP phylogenies were thus coupled with differential expression analysis between the two phases of both *L. migratoria* and *S. gregaria.* The work thus allowed us to characterise the CSPs of the two most destructive locust species and to identify homologs as well as CSPs with similar differential expression patterns between solitarious and gregarious locusts, thus inferring ancestry, specificity and importance for the development of the gregarious phase in locusts. The work also highlights specific CSPs for posterior functional testing.

5.2 Materials and Methods

5.2.1 Locust rearing

S. gregaria specimens for qPCR were reared in an insectarium at the Faculty of Sciences of the University of Granada, as described in the general methodology section. *L. migratoria* specimens for qPCR were reared at the Institute of Zoology of Chinese Academy of Sciences facilities as described in Kang et al. [2004].

5.2.2 Putative CSP retrieval, characterization, filtering and assignment

Locust CSP sequences were identified following three approaches: (i) based on published works on locust CSPs, (ii) by scanning the confirmed genomic sequences from *L. migratoria* and *S. gregaria* for the conserved CSP aminoacids pattern as well as other sequence attributes (presence of two exons and their relative orientation) and (iii) by BLAST searches [Altschul et al., 1997] on the Sanger sequenced EST databases of solitarious and gregarious (a) *L. migratoria* head, midgut and hind leg tissues [Kang et al., 2004] (EST accession numbers: CO819675 to CO832059 and CO832067 to CO865130) and (b) *S. gregaria* central nervous system [Badisco et al., 2011a], as well as on five Illumina sequenced *de novo* transcriptome assemblies from the central nervous system, toracic muscle, digestive tube, ovaries and testicles from both phases of *S. gregaria*.

We retrieved the known CSP sequences, in both nucleotide and protein formats, from the works of Picimbon et al. [2000] and Ban et al. [2003] for *L. migratoria* and Angeli et al. [1999] for *S. gregaria*. For locating CSPs on the *L. migratoria* genome we began by a BLASTx exploration of the available gene set from the *L. migratoria* genomic assembly version 2.4.1 as query [Wang et al., 2014a] (contigs accession number: AVCP000000000, and `http://159.226.67.243/` for downloading the assembly) and a local BLAST database of all the available arthropod CSP protein sequences from the NCBI. The translated protein sequences of the genes that had positive BLASTx hits against arthropod CSP proteins (E-value threshold of 10^{-10}) were used for pattern-based searches. These searches consisted of an initial extraction of the sequences that contain part of the conserved CSP pattern (the three cysteines C-X_{18}-C-X_2-C) followed by confirmation of the presence of the first cysteine 6 or 8 residues upstream the second cysteine. Only those sequences that presented a CSP conserved pattern with exactly four cysteines and the configuration C-X_6-C-X_{18}-C-X_2-C or C-X_8-C-X_{18}-C-X_2-C were kept for further analyses, the remaining sequences were discarded as non-CSPs even though they showed positive BLAST hits against arthropod CSP proteins.

tBLASTn searches using *L. migratoria* genomic assembly version 2.4.1 as query and the above mentioned local arthropod CSP protein database allowed us to identify the available *L. migratoria* genomic scaffolds that potentially contain CSPs. This allowed us to further filter our results based on the presence and orientation of the two CSP exons. We thus checked the highlighted genomic scaffolds for the presence of the typical CSP exons 1 and 2 structure considering all the six possible reading frames. We checked for the presence of the first pair of cysteines separated by six or eight residues in exon 1, the presence of the remaining pair of cysteines separated by two residues in exon 2, and we verified whether both exons of the potential CSP in the same genomic scaffold had coherent locations and the same orientation (i.e., exon 1 located upstream of exon 2 and congruent orientation of both exons, taking into account the strand of the genomic scaffold). For that, we took as reference the structure of the genomic CSP sequences reported in species such as the honey bee *A. mellifera* [Forêt et al., 2007] and the silkworm *B. mori* [Gong et al., 2007]. This way we confirmed a first set of putative CSP genes in the available *L. migratoria* genomic scaffolds. We also retained orphan exons 1 and 2 (exons 1 and 2 in loci where there is no exon 2 and 1, respectively) for downstream analysis in order to cheque whether they might be part of partially sequenced genes. tBLASTn also allowed us to determine the exonic coordinates of the putative CSP genes in each *L. migratoria* genomic scaffold.

BLASTx searches with the *L. migratoria* ESTs from Kang et al. [2004] as queries against our local CSP protein database allowed us to detect CSP transcripts from that species. The *L. migratoria* ESTs that gave positive BLAST results (E-value lower than 10^{-10}) were further analysed using TransDecoder [Haas et al., 2013] in order to obtain their translated amino acid sequences and check for the conserved four cysteines CSP pattern, as we did for the genomic loci. We then used the CD-HIT-EST command from CD-HIT [Li and Godzik, 2006] in order to remove redundancies by eliminating identical ESTs. We assigned ESTs to genomic loci by reciprocal BLASTn searches. Searching a database of the ESTs that we characterized from Kang et al. [2004], together with the transcripts reported as *L. migratoria* CSPs in the literature, for significant hits against sequence queries from the *L. migratoria* genomic loci with full CSP gene sequence or orphan exons that we characterized earlier allowed us to determine the loci that have prove of transcription. The assignation of genomic loci to ESTs was straightforward when a locus gives a best significant BLAST hit against an EST that does not appear as best hit against any other genomic locus. These cases allowed us to establish a minimum BLAST identity threshold that an EST and a locus had to reach in order to be assigned one to the other. We determined that value for the whole sequence as well as for exon 1 and for exon 2 separately. This way, when various loci give best BLAST hit against the same EST, we assigned the EST to the loci that gave above threshold identity with that EST, both as a full sequence as well as in their exon 1 and exon 2 parts (these loci are thus considered as potential gene duplicates). When the "conflict" is between two full CSP sequence loci, the EST is assigned to the locus whose BLAST results are equal or above the thresholds both for the whole sequence as well as its exon 1 and exon 2 parts. When the conflict is between a full CSP sequence locus and an orphan exon, the orphan exon had to reach the threshold established for that exon or the full length CSP sequence locus had to reach the threshold for both its full length, its exon 1 and its exon 2 in order to be assigned. The remaining loci and ESTs, the ones that had no BLAST hit or did not reach the three BLAST identity thresholds, were considered as with no evidence of transcription. Reciprocally, searching the database of genomic loci (both complete sequences and orphan exons) using the ESTs as queries allowed us to determine the ESTs that seem alleles of the same gene. These were all the ESTs that give above thresholds identity against the same locus or against a group of loci previously identified as gene duplicates. When the best BLAST hit of an orphan exon was against an unasigned EST and the reciprocal best BLAST hit of that EST was against the same orphan exon, we assigned the whole EST sequence to that locus. The ESTs that gave no acceptable hit against any genomic locus were considered as transcripts of CSP genes whose genomic loci are still unsequenced. This reciprocal BLAST step also allowed us to identify the exon junctions in *L. migratoria* CSP genes and to detect the presence of new *L. migratoria* CSPs —the ESTs, homologous to an arthropod CSP protein, that did not align against any *L. migratoria* genomic locus.

For *S. gregaria*, and in the absence of a draft genome, we initially performed a tBLASTn search using our local arthropod CSP protein database and the assembled contigs from our partial *S. gregaria* genomic DNA library [Camacho et al., 2015] as query. The downstream analyses were as described for *L. migratoria*. As to the ESTs, we used the wealth of data from the Sanger sequenced solitarious and gregarious CNS ESTs from [Badisco et al., 2011a] as well as our ten NGS (Illumina HiSeq2000 paired end) *de novo* assembled solitarious and gregarious EST libraries from the CNS, digestive tube, toracic and hind leg muscles, ovaries and testicles (over 500 Gb of data). The *S. gregaria* ESTs that had significant BLASTx hits against our local CSP protein database (at an E-value threshold of 10^{-10}) were scanned using Transdecoder [Haas et al., 2013] to detect complete or partial CSP open reading frames (ORFs) then translated into amino acids sequences. Only nucleotide sequences whose translated amino acid sequence contained the complete conserved four cysteines CSP pattern were retained for further analyses. We then removed redundancies using the CD-HIT-EST command from CD-HIT [Li and Godzik, 2006], as we did for *L. migratoria.*

Before we went further in the analysis of the locust CSPs (phylogeny and expression), we analysed the relationships between the inferred putative *L. migratoria* and *S. gregaria* CSPs in order to detect and remove any remaining redundancy in the *S. gregaria* CSP transcripts that we identified. We built a *L. migratoria* phylogeny using the full genomic loci and EST nucleotide sequences, and identified the different clades of putative *L. migratoria* CSP loci and their respective alleles based on the tree as well as the BLAST results. We then calculated the minimum within-clade sequence identity value and used it as threshold above which nucleotide sequences of the same species could be considered as alleles of the same CSP gene. We only took into account the clades that did not contain gene duplicates when calculating the identity values, since we could not assign propperly the alleles to each gene duplicate inside the clades that contain gene duplicates and their transcripts. We then calculated the pairwise sequence identities for the *S. gregaria* ESTs and removed redundancy (we removed *S. gregaria* ESTs expressed by potential alleles of the same CSP gene) based on the above mentioned threshold.

For the former analysis, we obtained the sequence alignments using the MAFFT-LINSI option of MAFFT v7 [Katoh and Standley, 2013]. We used MAFFT-LINSI because it focuses on aligning a conserved core region and gives less importance to the unconserved flanking regions —so we considered it suitable for aligning CSPs given that their sequences have a conserved core. We built maximum likelyhood trees using PhyML v3.1 [Guindon et al., 2010], with 1000 bootstrap iterations, and edited tree graphics from the PhyML Newick output format using the online version of the interactive Tree of Life tool iTOL [Letunic and Bork, 2011]. We used the CD-HIT-EST command from CD-HIT with the lowest identity possible (80%) in order to analyse the *L. migratoria* CSP clades for obtaining sequence identity matrices from the genomic and EST sequences of the different *L. migratoria* CSPs. This way we identified the minimum identity threshold between different CSPs of the same species and clade. We then used the CD-HIT-EST command from CD-HIT to calculate pairwise sequence identities and remove all but one of each putative set of *S. gregaria* ESTs that show higher identity than the threshold identified from the *L. migratoria* sequences (putative alleles of the same gene).

We had to deal with two additional issues in the case of *S. gregaria*, where no genomic scaffolds are available. Detecting the exon junctions is not as straightforward as in the case of *L. migratoria*, and the assembled ESTs from the NGS libraries may have assembly artifacts in form of chimeric sequences. For the first issue we used the *L. migratoria* exonic sequences from the genomic and transcriptomic sequences to trace the exon juntions on the *S. gregaria de novo* assembled transcripts. Still, this method did not allow us to detect such junctions for most of the *S. gregaria* ESTs. Consequently, we built BLAST databases using exon 1 and exon 2 of the *S. gregaria* ESTs whose junctions were identified by comparison with the *L. migratoria* putative CSP sequences and then carried out BLASTn analyses using the *S. gregaria* ESTs whose exon junctions were not previously located. We updated both exon BLAST databases by adding the newly identified exon junctions of the *S. gregaria* ESTs and we performed another BLAST search for exon junctions of the *S. gregaria* ESTs whose exon junctions were not located yet. We repeated this process

—database update and BLAST search for exon junctions in ESTs— until we cessed to obtain new significant BLAST results. We then built consensus sequences from the well identified exon 1 and exon 2 alignments and aligned them to each *S. gregaria* EST that still had no located exon junction. This way we successfully characterised the exon junctions for all the *S. gregaria* transcripts.

As to the potential chimeric *S. gregaria* sequences that might have resulted as *de novo* NGS assembly artifact, we used the MAFFT-LINSI command of MAFFT v7 [Katoh and Standley, 2013] to separately align all the exon 1 and all exon 2 nucleotide sequences from all the different putative *L. migratoria* CSPs (including the genomic loci, orphan exons, and the ESTs with no available genomic sequences) and all the putative CSP transcripts from *S. gregaria.* This way we generated an alignment for all exon 1 sequences and another for all exon 2 sequences. We then generated trees from these exon 1 and exon 2 alignments using PhyML v3.1 [Guindon et al., 2010] with 1000 bootstrap iterations. We proceeded by calculating the identity values between the *S. gregaria* exon sequences that were incongruently placed in exon 1 and exon 2 trees and their nearest neighbour sequences in the corresponding trees. To assign an incongruent *S. gregaria* EST to a clade or discard it as potentially chimeric we had to deal with three possible options: 1) when the identity values between the exons of the incongruent *S. gregaria* EST and their nearest sequence in the tree both were below-threshold (see below), 2) when the identity value between one of the two exons of the incongruent *S. gregaria* EST and its nearest sequence in the tree was below or within-threshold, whereas the identity value between the other exon of the incongruent *S. gregaria* EST and its nearest sequence in the tree was above or within-threshold, and 3) when the two exons of the incongruent *S. gregaria* EST showed above or within-threshold identities to their respective nearest neighbour sequences in the trees. We considered the putative *S. gregaria* ESTs that fit the first two cases as not chimeric and assigned them to a clade according to the location of their full length sequences (exons 1 and 2) in the tree of the full length locust CSP sequences (see below). We considered a potential putative *S. gregaria* EST as chimeric and discarded it from further analyses if it fit the third case. For this purpose, we firstly extracted the identity matrices between exon 1 sequences of the *L. migratoria* CSP loci, ESTs with no available genomic scaffold and orphan exon 1 and, in a similar way, between the exon 2 sequences of the putative CSPs of the same species. We then respectively identified the highest identity values between *L. migratoria* CSP sequences for both exon 1 and exon 2 matrices (excluding the identity values of the potentially duplicated loci, located in the same scaffold) and used them as respective exon sequence identity thresholds for attributing exons to CSP variants and, thus, identifying potentially chimeric *S. gregaria* ESTs that might have resulted as *de novo* assembly artefact. To avoid false negatives, we used the threshold values from the identities between the *L. migratoria* exons that belonged to different phylogenetic clades (i.e., the low distance values between recently duplicated CSP copies might mask the detection of *S. gregaria* exons with marginally higher distances). By including the *L. migratoria* orphan exons in the analysis we could characterise their similarity to other CSPs.

5.2.3 Locust CSP evolution

For the evolutionary relationships between the different *L. migratoria* and *S. gregaria* CSPs that we identified earlier we built maximum likelyhood phylogenies based on full length sequences (exons 1 and 2) of each CSP. The software and bootstrap iterations were as described earlier. In addition, we translated the nucleotide sequences to amino acid sequences and aligned them using the MAFFT-LINSI command before running the online version of ProtTest 2.4 [Abascal et al., 2005] to obtain the fittest amino acid substitution model which we used for maximum likelyhood phylogenetic analysis using PhyML and 1000 bootstrap iterations. The reason of building both nucleotides and amino acids trees was to help us confirm the place of the CSPs whose individual exons earlier gave incongruent placements in the nucleotides trees, and to check whether the functional products (amino acid sequences) followed a similar evolutionary path as their nucleotide counterparts. To have a wider idea on the evolution of locust CSPs we built an amino acids phylogeny including locust and non-locust CSPs. We retrieved from the NCBI protein database all the available CSP amino acid sequences of insect species with confirmed genomic CSP copy number. We thus had CSPs from the fruit fly *Drosophila melanogaster*, the mosquito *Anopheles gambiae*, the red flour beetle *Tribolium castaneum*, the silkworm *Bombyx mori*, the honey bee *Apis mellifera*, the pea aphid *Acyrthosiphon pisum* and the head louse *Pediculus humanus* (the sequences were selected based on the work of Vieira and Rozas [2011], excluding the CSP copies marked as pseudogenes or incomplete sequences).

The workflow described above for locust CSPs was also followed for the multispecies phylogenetic analysis of the CSPs. The very nature of the multiple copies sequences and sequences from gene families makes rooting the trees with a single external sequence from a related species ineffective (i.e., a single outgroup sequence for a multispecies phylogeny of such sequences does not guarantee shared ancestry between all the analyzed sequences and the outgroup sequence in question). One way to deal with that and clarify the phylogenetic relationships of the different lineages of the multiple copies or gene family sequences is to use as outgroups all the sequences of the same multiple copy or gene family from a related species. We used all the CSPs from the water flea *Daphnia pulex* (crustacean) as outgroup sequences in the multispecies phylogeny. This way we could also locate the locusts' last CSP ancestry point by outgrouping at the Arthropoda phylum level.

Once we had the multispecies phylogeny, we proceeded to revising (without renaming) the names that were already attributed to the locust CSPs reported elsewhere and to naming the CSPs that the present work reports as new. We based this nomenclature standardization step on the phylogenetic proximity of the subject locust CSP to the well established insect CSP that should give it a name. We named all the new *L. migratoria* CSPs first and then we followed with the new *S. gregaria* ones. We did not change the name of any CSP that is already reported elsewhere, even when we considered it pertinent, in order to avoid introducing more noise and/or confusion.

We analyzed the nucleotide diversity and non-synonymous to synonymous substitution rates for the CSP sequences of all the phylogenetic clades (both those with sequences of a single species and the ones with sequences of both species). For this, we separately aligned the CSP sequences' coding region belonging to each clade using MAFFT, as described above, and we used DNAsp v.5 [Librado and Rozas, 2009] to calculate the π and θ nucleotide diversity estimators, with their respective standard errors, and the number of synonymous and non-synonymous mutations (in order to calculate the non-synonymous to synonymous mutation rate; K_a/K_s) for each sequence pair or sequence group. The mean value of the pairwise K_a/K_s ratios were calculated for each of the clades that contained more than two CSP sequences.

5.2.4 Differential gene expression analysis

For testing differential expression of the CSPs between gregarious and solitarious locusts, we mapped the *L. migratoria* and *S. gregaria* RNA-seq reads to their respective transcriptomes and obtained the data for the CSP sequences. We used the reads obtained by Chen et al. [2010] from gregarious and solitarious whole adult *L. migratoria* bodies (SRA accessions SRR058455 and SRR058449, respectively) and the solitarious and gregarious Illumina Hiseq2000 Paired End reads that are currently being analyzed in our laboratory for comparative transcriptomics works on adult *S. gregaria* tissues (central nervous system, digestive tube, muscles, ovaries and testicles). As *L. migratoria* reference transcriptome, we used the published gene set derived from predicted transcripts in the *L. migratoria* genome home page (`http://159.226.67.243/download.htm`) and the new putative CSP sequences identified in the present work. For *S. gregaria*, we separately used *de novo* assemblies form our five NGS libraries. Each *S. gregaria* assembly was complemented with the CSPs that we identified only in the other *S. gregaria* assemblies. We mapped the reads to their respective refference transcriptome and summarized the read counts as described in the General Methodology chapter of this thesis. Read counts were normalized by the total number of mapped reads for the corresponding library and 2-based logarithm of the fold change of the normalized read counts (comparing the solitarious against gregarious phase for each species) was used to generate heatmaps and their corresponding dendrograms with the default command using R v2.15.0 environment [Gentleman et al., 1997].

To obtain an overall expression profile for *S. gregaria* gregarious and solitarious adults, we summed al the read counts from all the *S. gregaria* gregarious tissues and all the read counts from all the *S. gregaria* solitarious tissues to obtain total read counts from the whole solitarious and gregarious adults. The total counts were then normalized by the combined numbers of mapped reads in each tissue libraries in a similar way as described before. We did not include the solitarious and gregarious *S. gregaria* digestive libraries in this overall expression analysis because the *L. migratoria* adult whole body sequencing reads were obtained by Chen et al. [2010] from cDNAs of adults after dissection of the digestive tube —this way the results on both species would be comparable. Since NGS reads from different developmental stages were also available from Chen's work [Chen et al., 2010], we followed the same steps as above in order to obtain normalized read counts from the NGS libraries of gregarious and solitarious *L. migratoria* eggs (NCBI accession numbers SRR058432 and SRR058451 for gregarious and solitarious phases respectively), 1st and 2nd nymphal instars combined (SRR058446 and SRR058452 for gregarious and solitarious phases respectively), 3rd nymphal instar (SRR058447 and SRR058453 for gregarious and solitarious phases respectively), 4th nymphal instar (SRR058492 and SRR058457 for gregarious and solitarious phases respectively) and 5th nymphal instar (SRR058448 and SRR058454 for gregarious and solitarious phases respectively).

One of the CSPs was notoriously expressed in the testis library (see Results). To check whether it is a locust homolog of the ejaculatory bulb specific protein III (EBP3, GenBank Accession No U08281) —a protein that seems to be homologous to CSPs [Bohbot et al., 1998, Angeli et al., 1999, Picimbon et al., 2000]— we carried out an amino acid sequence phylogenetic analysis adding the EBP3 sequences from the pea aphid (*ApisEBP3*, accession number NP_001156287.1), the red floor beetle (*TcasEBP3*, accession number XP_008196341.1) and the fruit fly (*DmelEBP3*, accession number NP_524966.1), using the same methodology and software as detailed earlier.

Quantitative polymerase chain reactions (qPCRs) were carried out to complement and double-check the results obtained based on RNA-seq data. Because CSP nucleotide sequences sometimes present high degrees of conservation, we designed primers from the non-conserved regions of these genes in order to avoid non-specific amplification and the consequent quantification error. We initially focused the primer design on the signal peptide region and, when this region allowed no good quality primers, we mapped the *L. migratoria* CSP sequences from the EST database to their genomic contigs in order to identify several 5' and 3' untranslated regions (UTRs) that allowed us to design CSP-specific primer pairs in confirmed UTR regions. Still, we could obtain primers for only six putative *L. migratoria* CSPs based on their asigned ESTs. We dissected the heads of eight *L. migratoria* adults (4 gregarious and 4 solitarious), eight *L. migratoria* 4th instar nymphs (4 gregarious and 4 solitarious) and eight *S. gregaria* 4th instar nymphs (4 gregarious and 4 solitarious) and used central nervous system tissues from *S. gregaria* adults (five gregarious and five solitarious) and followed the RNA isolation cDNA synthesis and qPCR protocols described in the general methodology chapter. Primer sequences are listed in table 3 from the general methodology section of this thesis.

5.3 Results

We identified 42 possible CSP loci (located in 30 different genomic scaffolds) from the extensive list of sequences with detected CDSs in the *L. migratoria* genome sequences [Wang et al., 2014a]. Subsequent tBLASTn analysis allowed us to detect seven new complete putative loci in that genome (one locus in scaffold 101, four in scaffold 71401 and two in scaffold 33302). Table 5.43 lists the 49 identified loci and figure 5.41A shows the proportions of each configuration of the CSP loci (single gene, gene cluster, tandem repeats...). We also detected four orphan exon 1 (figure 5.41B) and 14 orphan exon 2 sequences (figure 5.41C) in a total of 18 different genomic scaffolds. In addition, two partial loci with exon 2 upstream of exon 1 were found in the scaffold 15074 and the contig C189039548 (figure 5.41D). In summary, we identified a total of 49 loci containing putative full length CSPs in 33 different *L. migratoria* genomic scaffolds, and the number of putative CSPs in this locust genome could rise to at least 55 if we consider the orphan exons (i.e., at least four complete CSPs among the four orphan exon 1 and 14 orphan exon 2, and up to two complete CSPs among the two exons 2 that we found upstream of two exons 1 in two separate genomic scaffolds). The mean length of the whole CDS (meassured as the distance in base pairs from the start of exon 1 to the end of exon 2) was 9050 ± 1459 bp, whereas the mean intron length was 8705 ± 1459 and the mean exon sequence length was 345 ± 7. Surprisingly, four CSP loci contain no introns (less than seven bases separated exon 2 from exon 1, figure 5.41E). Nine of the genomic scaffolds presented more than one CSP gene, being eight the maximum number of detected CSP genes found in a single locus (scaffold 71401). Notoriously, these eight putative CSPs are in tandem, with no bases separating the exon 1 of a putative CSP from the exon 2 of the upstream one (figure 5.41F). The prevalence of genomic CSP structures is pictured in figure 5.41G and table 5.44 shows the genomic location and BLAST results for each locus.

Table 5.43: Location of the putative CSPs detected in the available *L. migratoria* genomic scaffolds and their association to the available ESTs from the same species. This table shows all the retrieved sequences that had a positive BLASTx result against an arthropodan CSP protein. A sequence was considered to be a putative locust CSP if it had the CSP conserved pattern of four cysteines (see the main text), was composed of two exons (1 and 2) belonging to the same scaffold and in the correct relative orientation. Additional BLAST data and on the attribution of ESTs and genomic loci to each other and to known CSPs can be found in table 5.44 and table 5.45.

L. migratoria Genome loci	Scaffold	Sense	Exon1 start	Exon 2 end	Expression	Assigned EST
101	Scaffold 101	+	2366516	2400577	No	—
103059	Scaffold 103059	+	2171	10788	No	—
12585	Scaffold 12585	-	179306	167247	Yes	LM_SH5_003244*
13671	Scaffold 13671	-	595509	590973	No	—
15810	Scaffold 15810	-	78980	63220	Yes	*LmigCSPII-6*
18858cds1	Scaffold 18858	+	97840	108678	Yes	LM_GH5_000761*
18858cds2	Scaffold 18858	+	141703	149698	Yes	LM_GH5_000761*
18858cds3	Scaffold 18858	+	168506	170451	No	—
21551	Scaffold 21551	-	122154	97824	Yes	LM_SH5_001382
22826cds1	Scaffold 22826	-	159147	153386	Yes	LM_SH5_003413*
22826cds2	Scaffold 22826	+	127283	129558	Yes	*LmigCSPII-1*
235750	Scaffold 235750	+	7652	9847	Yes	*LmigCSPII-14*
24400	Scaffold 24400	+	13865	18627	Yes	LM_GH5_000758
25611	Scaffold 25611	+	14519	63192	No	—
2564	Scaffold 2564	+	89690	97425	No	—
30358	Scaffold 30358	-	35537	22421	No	—
31810	Scaffold 31810	-	78016	67182	No	—
320887	Scaffold 320887	+	589	33184	Yes	LM_GH5_003053
3212cds1	Scaffold 3212	-	1325340	1315951	No	—
3212cds2	Scaffold 3212	-	1382008	1363494	No	—
325580	Scaffold 325580	+	830	24951	Yes	LM_GB5_001536
33302cds1	Scaffold 33302	+	4672	5022	No	—
33302cds2	Scaffold 33302	+	10024	10374	No	—
37289	Scaffold 37289	+	4533	10346	No	—
374630	Scaffold 374630	+	5695	6078	No	—
392768	Scaffold 392768	+	61	21224	Yes	LM_GH5_003725*
41553	Scaffold 41553	+	66870	72899	Yes	LM_GH5_003053*
46375	Scaffold 46375	+	44829	62108	No	—
5214cds1	Scaffold 5214	-	122926	116531	No	—
5214cds2	Scaffold 5214	+	143789	145156	No	—
57579	Scaffold 57579	+	3592	15413	No	—
647	Scaffold 647	-	198201	176012	Yes	LM_GL5_000034
699cds1	Scaffold 699	+	80447	89313	No	—
699cds2	Scaffold 699	-	152813	144007	No	—
71401cds1	Scaffold 71401	-	45335	44165	Yes	LM_GH5_002985*
71401cds2	Scaffold 71401	-	44176	41897	Yes	LM_GH5_002985
71401cds3	Scaffold 71401	-	41890	40695	Yes	LM_GH5_002985*
71401cds4	Scaffold 71401	-	40637	38659	Yes	LM_GH5_002985*
71401cds5	Scaffold 71401	-	38626	35392	Yes	LM_GH5_002985
71401cds6	Scaffold 71401	-	35385	33956	Yes	LM_GH5_002985*
71401cds7	Scaffold 71401	-	33949	31212	Yes	LM_GH5_002985
71401cds8	Scaffold 71401	-	30988	28604	Yes	LM_GH5_002985*
757cds1	Scaffold 757	-	3626	2064	Yes	LM_SH5_003270
757cds2	Scaffold 757	-	60814	58709	Yes	Lmig_CSPII-10
757cds3	Scaffold 757	-	95948	91849	Yes	LM_SH5_003782
75957	Scaffold 75957	-	4048	3689	No	—
78016	Scaffold 78016	-	21591	16775	No	—
9174cds1	Scaffold 9174	+	10000	11523	Yes	LM_SH5_003244
9174cds2	Scaffold 9174	+	36759	38397	Yes	*LmigCSPII-13*
LM_GH5_003055	C178632750	+	20	175	Yes	LM_GH5_003055
LM_GB5_004555	C187757636	+	479	655	Yes	LM_GB5_004555
LM_GH5_003725	scaffold50720	-	1643	1521	Yes	LM_GH5_003725
	scaffold401450	+	218	412	Yes	LM_GH5_003725
LM_GH5_003400	scaffold53850	-	4232	4044	Yes	LM_GH5_003400
LM_GH5_000760	scaffold68729	-	7415	7227	Yes	LM_GH5_000760
LM_GH5_000761	scaffold281155	+	5573	5761	Yes	LM_GH5_000761
LM_SH5_003268	C157799226	-	98	6	Yes	LM_SH5_003268
LM_GM5_003208	—	—	—	—	Yes	LM_GM5_003208

Regarding *L. migratoria* ESTs, out of 96 sequences with positive BLAST result against our local arthropod CSP protein sequence database, 36 showed the conserved CSP cysteine pattern, 34 of which were unique sequences (not

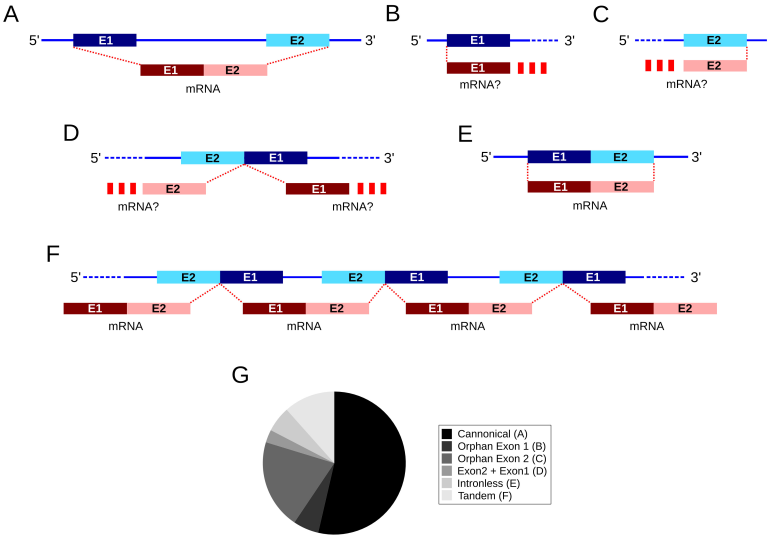

Figure 5.41: Configurations and frequencies of the CSP sequences detected in the available draft genome od the migratory locust (*L. migratoria*). A-F: Schematic representation of each configuration. G: Sector graph with the number of times a genomic CSP configuration is detected.

alleles of the same CSP). Since the EST sequence names from Kang et al. [2004] contain a code referring to the tissue from which the cDNA came, we could infer that 22 of these 34 transcripts were expressed in the head, five in the hind legs, and two in the mindgut of *L. migratoria* 5th instar nymphs. The remaining five transcripts were expressed in 5th instar female nymphs (for more information, see table 5.45). To these 34 CSP transcripts we summed the 20 clonned CSP mRNAs from the works of Picimbon et al. [2000] and [Ban et al., 2002], thus totalling 54 *L. migratoria* CSP transcripts.

BLASTn of the 54 *L. migratoria* ESTs against the 49 complete sequences plus four orphan exon 1 and 14 orphan exon 2 CSP sequences that we identified earlier in different *L. migratoria* genomic loci showed that 27 complete genomic sequences have best significant hits against 14 different ESTs, due to 18 genes corresponding to the same five ESTs whereas nine genes have best BLAST hits against unique ESTs, and four exon 1 against three different ESTs (all but one of these ESTs with significant hit against complete genomic CSP loci), and the 14 exon 2 against seven different ESTs (one of them sharing BLAST result against two orphan exon 1, four with no BLAST results agains any orphan exon 1 or complete CSP and the rest having already significant results against complete CSP loci). The remaining 22 complete gemonic sequences and seven orphan exon 2 had no significant BLAST hit to any *L. migratoria* EST (table 5.44). Reciprocally, of the 54 putative ESTs 46 show best positive BLAST result against 13 different complete *L. migratoria* putative CSP loci (table 5.45), whereas seven have their best positive BLAST results against orphan exons (two transcripts against orphan exon 1, four against orphan exon 2, and one against an orphan exon 1 and an orphan exon 2). Although one CSP transcript (LM_GM5_003208) shows no significant BLAST result against any CSP locus, its amino acids sequence cumplies with the requisites that a sequence needs to have for it to be considered another putative CSP (table 5.45). Thanks to this analysis we also found that the two exons of the EST LM_GH5_003725 are located in two genomic scaffolds (exon 1 in scaffold 50720 and exon 2 in scaffold 401450). In addition, two ESTs (LM_SH5_003268 and LM_GH5_003400) showed high similarity to two of the genomic orphan exon 1 sequences (contig C157799226 and scaffold 53850, respectively) and another four ESTs (LM_GH5_000760, LM_GH5_000761, LM_GH5_003055 and LM_GB5_004555) showed high similarity to four of the genomic orphan exon 2 sequences (scaffold 68729, scaffold 281155, contig C178632750 and contig C187757636, respectively). Thus, the total number of *L. migratoria* putative CSPs detected in the present work appears to be 57, 35 of which expressed, and distributed as follows: 49 loci with completely sequenced genes —27 of which expressed—, two loci whose exon 2 is still missing from the draft genome, four whose exon 1 is still missing from the draft genome, an exon 1 and an exon 2 that belong to different scaffolds but align to the same EST, and one EST whith no sequenced genomic locus.

A similar analysis of the genomic contigs of *S. gregaria* that we assembled from a previous partial (low coverage) genome sequencing work in our lab [Camacho et al., 2015] allowed us to retrieve only 11 genomic contigs that might potentially contain putative CSP loci. However, they were discarded for being too fragmented and with unreliable sequence patterns (table 5.46). We thus had to rely on BLAST searches and homology to *L. migratoria* putative CSP sequences in order to determine the minimum number of different CSP transcripts from the available *S. gregaria* transcriptomics data (ours and from Badisco et al. [2011a]). After BLASTx searches on our local arthropod CSP protein database using the Sanger sequenced ESTs [Badisco et al., 2011a] and our NGS assembled transcriptomes from the five *S. gregaria* tissues (see general methodology), and after selecting only those sequences that contain the conserved CSP cysteine pattern and removing identical redundant sequences, we obtained a total of 179 *S. gregaria* putative CSP transcript sequences. Five of these were already reported in Angeli et al. [1999]. Of the remaining unreported transcripts, we detected nine in the EST sequences from Badisco et al. [2011a] whereas 18 belonged to our NGS *de novo* assembled central nervous system transcriptome, 92 to our NGS *de novo* assembled digestive tube transcriptome, 48 to our NGS *de novo* assembled muscle transcriptome and seven to our NGS *de novo* assembled testicle transcriptome —we found no valid putative CSP transcript in our NGS *de novo* assembled ovary transcriptome (table 5.47). BLASTn searches against the *L. migratoria* genomic sequences of putative CSPs showed that all these 179 *S. gregaria* transcripts have at least one significant BLAST hit against 25 putative *L. migratoria* CSP genes.

A preliminary phylogeny with the nucleotide sequences of the previously identified *L. migratoria* CSP genes and transcripts is shown in figure 5.47. That phylogeny shows seven clear cases of a single CSP genomic sequence and few transcript sequences sharing a single clade (with two to six alleles per gene), and five genomic sequences with a single assigned transcript. We calculated the pairwise sequence identity values between the sequences of each of these clades (table 5.48), since the rest of the genes did not present associated transcripts or were too close in the phylogeny (i.e. recently diverged genomic copies) to take them as sequences divergent enough as to discriminate between their alleles and unequivocally attribute them to the respective gene copy. The lowest sequence identity value was between sequences 22826cds2 and LM_SL5_002527 (93.8 %). We applied that sequence identity value as threshold to the 179 *S. gregaria* ESTs in order to discriminate between the sequences transcribed from different "good" genes and those transcribed from different alleles of the same gene. This way the number of *S. gregaria* CSP transcripts went from 179 to 42 (marked in bold in table 5.47).

Exon characterisation showed a more variable in length exon 1 (80 to 242 bp; mean length = 167.755 ± 5.047 bp) and a sensibly larger exon 2 (113 to 290 bp; mean length = 179.694 ± 4.247 bp). We built maximum likelyhood phylogenies for each of these two CSP exons using the identified *L. migratoria* genomic CSP sequences and *S. gregaria* transcripts. This way we could group the majority of these sequences, including the orphan exons detected in *L. migratoria* genome, within congruent clades in both exons' trees. However, the exon 1 and 2 sequences belonging to 13 *S. gregaria* CSP transcripts (about 30 % of the total) were at incongruent positions of the two single-exon-based trees (figures 5.48 and 5.49). Nucleotide sequence distance matrices were separately established for exons 1 and 2 of the distinct CSP genes (defined as *L. migratoria* CSP sequences at different genomic loci, see general methodology). The highest identity value were 97 % between the exon 1 part of the CSP genes in scaffolds 78016 and 103059, and 95 % between the exons 2 part of the CSP genes in scaffolds 30358 and 46375. No exons 1 and 2 of the same incongruent *S. gregaria* CSP simultaneously showed more than the threshold values to their respective nearest *L. migratoria* or *S. gregaria* neighbour sequence in the tree. Hence, none of the sequences whose exons were incongruently located in their respective phylogenetic tree could simultaneously be considered as belonging to different clades and, consequently, none of the apparently incongruent full CSP sequences from the *S. gregaria de novo* assembled transcriptome seem to be chimeric —at least there seems to be no artifactual exon shuffling (details in table 5.49).

A maximum likelihood tree built using the whole nucleotide sequences of the non-redundant set of identified genomic and transcriptomic locust CSPs is shown in figure 5.50. In contrast with the trees that we previously built with single-exon sequences (figures 5.48 and 5.49), that showed a polarized topology between specific *L. migratoria* and specific *S. gregaria* sequences, the whole CSP sequence tree in figure 5.50 indicates that most of these CSPs have orthologous sequences in the two locust species analyzed here. In fact, 26 out of the 57 full length *L. migratoria* CSPs and 27 out of the 42 *S. gregaria* CSPs are grouped in interspecific independent clades. 10 of these clades contained an orthologous pair of sequences whereas 11 contained both orthologs and paralogs (6 clades contained a *L. migratoria* ortholog and two *S. gregaria* paralogs, and 5 contained a *S. gregaria* ortholog and two *L. migratoria* paralogs). The remaining 31 *L. migratoria* and 15 *S. gregaria* full length CSP sequences are grouped in various species-specific clades. Alignment of the full amino acid sequences of both *L. migratoria* (figure 5.52) and *S. gregaria* (figure 5.53) CSPs clearly display the conserved CSP cysteine pattern as well as several more conserved regions, with the N-terminal region being the most variable. Before building a maximum likelihood tree for these amino acid sequences, we analyzed the most likely amino acid substitution model using ProtTest, which suggested LG + I + G [Le and Gascuel, 2008] as optimal model. Under this model, the topology of the tree of the identified locust CSP amino acid sequences (figure 5.51) is almost identical to that of the tree built using the full nucleotide sequences of the same CSPs. The amino acids tree shows overall shorter branch lengths, compared to the nucleotide-based locust CSP tree, probably due to a high amount of synonymous mutations. Comparison of the nucleotide and amino acid trees shows that none of the sequences occupy incongruent positions within its respective congruent clade.

A maximum likelihood tree of the amino acid sequences of the CSPs from multiple arthropod species further defines orthology between groups of locust CSPs and between these and other arthropod CSPs. Besides the multiple species clades, the tree shows two locust specific CSP expansions that include the vast majority of the putative locust CSP sequences described in this work (86 out of 99, figures 5.42, 5.43 and 5.44). The ancestral locust specific CSP expansion, which we name locust CSP expansion 1, is supported by *A. pisum* CSP sequences. It includes 31 locust CSPs (16 from *L. migratoria* and 15 from *S. gregaria*) of which at least nine pairs are clearly orthologous. The recent locust specific CSP expansion, or locust CSP expansion 2, is supported by the probable locust CSP4 clade and is sister to a multiple species clade. It includes 54 locust CSPs (35 from *L. migratoria* and 19 from *S. gregaria*) of which seven pairs are orthologous. Of the 14 locust CSP sequences (six from *L. migratoria* and eight from *S. gregaria*) that lay outside of the two locust specific clades (expansions), at least five pairs are orthologous. The tree shows no clear orthology between similarly named CSPs that were reported for different species in different works (see Discussion). Nevertheless, we maintained the names of these sequences as reported in the original works. As to the newly discovered putative locust CSPs that we report here, we named them depending on their position in the topology of the tree. We gave lower numbers to the most ancestral CSPs unless the newly discovered putative locust CSP happens to be orthologous to a CSP that has already been reported and named in the other locust, in which case we attributed the same number to the new CSP as the one given to the CSP already reported in the other locust (details in table 5.50). With only 1.3 % of the amino acids alignment sites being completely conserved and pairwise aminoacids identities between 24 and 40% , arthropod CSPs are not especially conserved (figure 5.52I) and locust CSPs are no exeption (figures 5.53 and 5.54)— 3.5 % and 2.3 % complete site conservation and 40.3 and 33.9 % pairwaise identities in the *L. migratoria* and *S. gregaria* aminoacids alignment, respectively. Still, and in addition to the four cysteines region, CSPs of the terrestrial arthropods show three islands with somewhat conserved aminoacids signatures (figure 5.54).

The differential expression analysis did not reveal general conserved expression patterns between the gregarious and solitarious adult individuals from both species. However, it showed several orthologous sequences with a shared expression pattern. In figures 5.42, 5.43 and 5.44 we can observe that 21 out of 57 *L. migratoria* CSPs and 38 out of 42 *S. gregaria* CSPs show significant differential expression between phases. Among these, 14 orthologous CSPs share the same direction of the significant differential gene expression pattern: *LmigCSP8* and *SgreCSP8*, *LmigCSP9* and *SgreCSP9*, *LmigCSP10* and *SgreCSP10*, *LmigCSP12* and *SgreCSP12*, *LmigCSP14* and *SgreCSP14*, *LmigCSP19* and *SgreCSP19* and *LmigCSP3* and *SgreCSP37*. Interestingly, all of these orthologs show higher gene expression levels in the gregarious phase compared to the solitarious one (figure 5.45). The overall expression of CSPs throughout the *L. migratoria* developmental stages analyzed here shows a minimum level at the 3rd instar and a maximum level at the 4th instar, both in the gregarious phase, with generally higher expression levels in the gregarious phase (figure 5.55A). Regarding *S. gregaria*, the overall CSP expression levels were clearly higher in the gregarious phase for all the analyzed tissues, with the maximum expression level being in the central nervous system, followed by the muscles (figure 5.55B). While *S. gregaria* ovaries showed overall low coverage of CSPs (with a maximum of 29 mapped reads), the testicles showed a CSP (*SgreCSP14*) with a considerably high number of mapped reads (5249 mapped reads) compared to the rest of CSP sequences (with a maximum of 46 mapped reads), although it shows no differential expression between phases (details of the studied CSP expression profiles can be found in figure 5.46 and table 5.53). Phylogenetic positioning of the *SgreCSP14* amino acids does not support its potential orthology to the Ejaculatory Bulb Protein 3 (EBP3) in spite of its notorious expression in the testicle library. In fact the phylogeny (figure 5.56) grouped each of the EBP3 from the three species with a respective homologous CSP sequence from the same species (*TcasEBP3* with *TcasCSP3*, *DmelEBP3* with *DmelCSP2*, and *ApisEBP3* with *ApisCSP6*), and the EBP3-containing clades contained no locust CSP sequences. Strikingly, the position of the EBPs in the phylogeny is so different between species and so close to specific CSPs that their distinction from CSPs and inference on the EBP as opposed to CSP nature of a sequence cannot be made based on sequence homologies.

qPCR analyses supported the RNA-Seq data for five out of six tested CSP sequences in *L. migratoria* adults (figure 5.57A). The remaining CSP, *LmigCSP4*, showed an opposite differential expression pattern than the one revealed by RNA-seq (figure 5.46A). The qPCR results on *L. migratoria* 4th instar nymphs were similar: four out of six tested CSPs supported the RNA-seq data (figure 5.57C). *LmigCSP3* and *LmigCSP24* were the genes that showed opposite qPCR differential expression pattern between phases compared to the one suggested by RNAseq. We also used qPCR to test the expression of *SgreCSP18*, a CSP that our RNAseq data suggest to be overexpressed in the central nervous system of gregarious *S. gregaria* adults. We confirmed the biased expression pattern of that CSP towards the gregarious phase in central nervous system tissues of both *S. gregaria* adults and 4th instar nymphs (figures 5.57B and 5.57D). Overall, 10 out of the 14 replicated qPCR testings (nine in *L. migratoria* and one in *S. gregaria*) showed the same direction of differential, not necessarily significant, gene expression as the RNAseq data. Four of these qPCR testings showed significant differential expression between phases and only three out of the total 14 replicated qPCR testings showed opposite direction of the differential expression as the one obtained by RNAseq—expression of *LmigCSP3* in the 4th instar nymphs of *L. migratoria* being the only case of significant qPCR differences between gregarious and solitarious locusts that was incongruent with the RNAseq data (figures 5.57 and 5.58).

5.4 Discussion

The interest of the research on locust phase change being obvious and the involvement of CSPs in such natural phenomenon being pivotal (see Introduction), the present work was planned in order to determine the full set of locust CSPs from the currently available genomic and transcriptomic data. With the CSP sequences being identified for the two main pest locusts, we procured to understand the evolution of these genes, infer how their functional constraints might have shaped their numbers and sequences, and highlight the ones that seem important for gregariousness based on conservation of the differential expression between phases of the two locusts.

5.4.1 Locust genomes contain a strikingly huge number of CSP genes

Prior to the current work, the molecular cloning and Sanger sequencing works by Angeli et al. [1999], Picimbon et al. [2000] and Ban et al. [2002] reported a total of 25 locust CSP transcripts (20 in *L. migratoria* and five in *S. gregaria*). Surprisingly, that number did not increase hitherto the current work, in spite of the wealth of transcriptome and even genome next generation sequencing works in *L. migratoria* —probably because none of those works focused especially on the CSP gene family. Still, one of the open questions on locust CSPs remained being whether the full set of such important genes has been identified correctly or not. Just by looking at the numbers, it is clear that unless the number of CSPs identified in *L. migratoria* is artifactually higher than the real one, at least the full set of *S. gregaria* CSPs was not completely identified —especially since this species' genome is larger than that of *L. migratoria* [Wang et al., 2014a, Camacho et al., 2015]. As to *L. migratoria*, 20 might initially appear a reasonable number of CSPs if it wasn't for the fact that identified sequences were transcripts whose genomic loci were not identified and might, as we here show is the case, not represent the full set of CSP genes of that species genome and/or contain alleles of the same gene.

5.4.2 The described set of locust CSP genes is representative

The accuracy of this statement obviously depends on the logic and search method used. In silico detection of the full set of genes in a genome is conditioned by the completion state of such genome and by the presence of features that allow distinguishing the genes in question from the rest of the genome. The definitive number of genes would be obtained straightforward if the genome is completely sequenced and the genes have unambiguous conserved features. Fortunately for us, CSP genes have distinctive features that include the presence of two exons and, more importantly, a conserved pattern of four cysteines [Bohbot et al., 1998, Angeli et al., 1999, Jacquin-Joly et al., 2001, Ban et al., 2003, Campanacci et al., 2003, Forêt et al., 2007]. However, our genomic sequences of *S. gregaria* [Camacho et al., 2015] presently cover only a small fraction of this species huge genome and the available draft genome of *L. migratoria* is only nearly completed —the results of our genome-based search thus wouldn't detect the full set of CSPs. To solve this, we complemented the public databases and genome searches using ESTs and transcriptome assembled sequences. Here the issue is multi-fold. On the one hand, detection of the full set of CSPs from a single transcriptomics project is impossible and the number of detectable sequences depends on the animal, tissue and RNA handling methodology as well as the sequencing depth. It also depends on the gene expression level, as well as on the timing (developmental stage, moment of the day, circadian rhythm...) and conditions of the experiment (environment, exposure to chemicals, exposure to other specimens, stress...), and on the material (sex and organ). On the other hand, the detected number of CSPs via transcript-based searches; in vivo (cloning and sequencing) or in silico (transcriptomics), might be inflated due to false positives (alleles of the same gene being mistakenly identified as different genes). Our solution to that had two general aspects: (i) to search as much data as possible in order to detect the most complete set of genes and (ii) to use a logical combination of rules and distinctive CSP sequence features in order to avoid false positives.

The number of locust CSPs hitherto reported in other works is undoubtedly incomplete, especially for *S. gregaria*, in spite of our exhaustive analysis of all the EST, genome, protein and NGS sequences of the public databases, as well as our over 500 million Illumina Hiseq 2000 Paired End sequencing reads. However, the share amount of data that we analyzed and the large number of CSPs that we report here suggest that the number of the potentially undetected locust CSPs is very likely low (a single digit), so that the set of locust CSPs is now very likely almost completely identified. As to whether the detected genes are genuine or not, and whether they contain false positives or not; the gene-search rules were clear, logical and based on objective cut-offs. They included significant BLAST hit to known arthropod CSP proteins, presence of the conserved CSP cysteine pattern, presence of the two CSP gene exons, concordant best BLAST result of each of the two exons against the same arthropod CSP protein, concordant relative position of the two exons, concordant relative orientation of the exons and adecuate sequence similarity thresholds. Our choice for the latter criterion was no doubt the most objective way to distinguish between sequences of different genes (those that differ more than the least similar sequences in a clade of *L. migratoria* ESTs and their respective and unique genomic CSP sequence) and the alleles of the same gene (those that differ at most as the least similar sequences in a clade of *L. migratoria* ESTs and their respective, only one, genomic CSP sequence). This way, here we show that CSPs that other works reported as different genes (for example *LmigCSP5* and *LmigCSPII-10*) seem in fact alleles of the same gene, and we also show and alert of the incongruencies in CSP sequence numbering between different species (for instance *SgreCSP2*, of *S. gregaria*, is an ortholog of *LmigCSP24* rather than *LmigCSP2* in *L. migratoria*, which is instead orthologous to *SgreCSP35*). We tried to use a phylogeny-guided (sequence similarity) numbering of the sequences that we newly report here and we suggest this as a rule for the CSPs that future works might detect in these or other species. However, we didn't want to suggest changes in the naming of the CSPs that are already reported elsewhere in order to avoid confusion and "water mudding".

The *S. gregaria* sequences reported in this work came from RNA-seq libraries and their *de novo* assembly might have produced chimeric sequences that would misleadingly increase the number of CSPs in that species. Nevertheless, our filtering method, based on sequence similarities and congruency between the results of the separate trees of each of the two exons, suggested that we had no missasemblies. Neither does our assembly seem to be too incomplete given the fact that the whole set of described CSP sequences from Angeli et al. [1999] showed several orthologous sequences in our *S. gregaria de novo* assembled transcripts.

The combination of EST and genomic data highlighted partially sequenced genomic CSP loci (the orphan exons detected in this work) and allowed confirmation of the transcription of some of them (the two orphan genomic exon 1 sequences of the ESTs *LmigCSP5* and *LmigCSP47* and the four orphan genomic exon 2 sequences of the ESTs *LmigCSPI-3*, *LmigCSPI-4*, *LmigCSP8* and *LmigCSP18*). Furthermore, works of similar nature as ours might be of some help to the complete genome assembly goal as they allow relating scaffolds based on their exon content. In the case of the current work, among the several scaffolds of the *L. migratoria* draft genome, scaffold 50720 should be the one located immediately downstream of scaffold 401450 (since *LmigCSP1* exon 1 is in scaffold 50720 and its exon 2 in scaffold 401450). In addition, the fact that we detected a *L. migratoria* EST corresponding to a putative CSP (*LmigCSP9*) with no assigned locus indeed suggests that some more CSPs might be in the still not sequenced parts of that species genome. We also found 17 *L. migratoria* genomic CSP loci that have no assigned *L. migratoria* ESTs but who show potential orthology to 14 putative *S. gregaria* CSP transcripts (supplementary table 5.44 and figure 5.47), meaning that those loci are genuine and might even be functional. Only five putative CSP loci from the *L. migratoria* draft genome lacked transcriptional evidence in any of the two species, they are thus the only ones whose existence relays solely on in silico mining of a draft genome.

5.4.3 CSP gene sequences seem to have proliferated in the locust genomes through independent duplication

With essential implication in the so important phenomenon that defines locusts as organisms and distinguishes them from grasshoppers (see Introduction), the study of locust CSPs is of clear importance at least for understanding the evolution of the chemoreception genetic toolkit under the functional and selective constrains imposed by episodic exposure to high populations densities and its resulting conditions. Furthermore, from a genomic side of view, it is intriguing to know what the consequence of increased genome sizes on such key gene family might have been.

The first main result unveiled by the present work is the detection of an unexpectedly much higher number of CSPs in the genomes of the two main pest locust species than the numbers hitherto reported for any living organism (see above). Staggering expansion of this gene family must have happened in order to reach the at least 57 and 42 complete CSP gene copies that we respectively report here for *L. migratoria* and *S. gregaria.* Locusts therefore have more CSPs than those found in the sequenced genomes of insect species such as the red flour beetle, the honey bee, the mosquito and several *Drosophila* species (summarized in Vieira and Rozas [2011]). In fact, locusts seem to even have more CSPs than species with confirmed duplication and diversification of CSPs such as the silkworm *Bombyx mori*, that has 21 described CSP paralogs [Gong et al., 2007], and the red flour beetle *Tribolium castaneum*, for whom 19 CSP paralogs have been described [Engsontia et al., 2008]. Even the champions of chemoreception, ants (with 11 to 21 CSP paralogs depending on the species [Kulmuni and Havukainen, 2013]), remain far behind of the locusts studied here. Although it is true that the estimated less than 400 Mb mean ant genome size [Tsutsui et al., 2008] is less than a 5% of the estimated about 10 Gb size of the *S. gregaria* genome, meaning that locusts have less CSPs per Mb of their genomes than ants (see table 5.54). However, the latter datum cannot allow for function-oriented interpretation and it would be misleading to conclude that "as ants have more CSPs per genome size than locusts do, then CSP functions are more important or needed more in ants than in locusts". In fact, in spite of grasshoppers having some of the biggest metazoans genome sizes [Wilmore and Brown, 1975], the two locust species studied here are not poliploid —they have a diploid set of 22 autosomes plus one X chromosome in males and two in females— and CSP gene expansion in their genomes must thus have happened due to gene duplication events not to whole genome ploidy. Not all locust genes have multiple copies and only those that either their DNA content or function allow or require expansion have experienced it (CSPs among them). The DNA content of the CSP genes does not show relevant repetitions that might easily explain gene duplication events, still we found tandem repetitions of CSP genes with up to eight copies, meaning that unequal crossing-over and recombination between homologous chromosomes, followed by selection or drift, could have been a cause for part of the mechanisms that allowed CSP expansion in the locust genome. There are no clear footprints of transposition and CSPs are no transposable elements, still the presence of highly similar CSP sequences in different genome parts (scaffolds), such as *LmigCSP26* and *LmigCSP27* that share over 98 % nucleotide sequence similarity, suggests that some movement of the CSPs between different genomic loci might have happened—with such large genome sizes, genome wide reorganizations may have taken place more frequently in locusts than in other species [Flavell et al., 1974].

Assuming that the original CSP copy in each multi-CSP locus is the one at the 5' end of the locus, due to proximity to the promoter, we can observe that the further we are from the original copy the more dissimilar the CSP copies are (see scaffolds 18858 and 757, details in tables 5.55 and 5.56

respectively). The most likely explanation for this is a stepwise downward expansion either by repeated duplication of the original copy or by intial duplication of that copy and subsequent duplication of the resulting copy—the latter scenario is more likely given the higher similarity between the first and second duplicated copies compared to the similarity between the presumed original and second duplicated copies. Such declining pattern of sequence similarity within the multi-CSP loci is not observed in the case of the eight CSP copies at the same locus in scaffold 71401. These therefore seem to have originated by independent duplications of different CSP paralogs in the same locus. The higher nucleotide sequence similarity between these paralogs suggests that their duplication is more recent than that of the paralogs in scaffold 18858—some duplications seem indeed very recent (paralogs 4, 5, 6 and 7)—but earlier than duplication of the paralogs in scaffold 757.

The inferred number of ancestral CSPs at the arthropod phylum level is only seven, with four conserved paralogs, a fifth presenting an extra α-helix and the remaining two showing a higher mutation rate [Kulmuni and Havukainen, 2013]. They seem to evolve following a birth-and-death dynamic [Sanchez-Gracia et al., 2009, Vieira and Rozas, 2011, Kulmuni and Havukainen, 2013]. The phylogeny in figure 5.42 suggests that the ancestral number of CSPs for both locusts might be also around the seven clades shown by the tree. Two of these are major CSP expansions in locusts, one of them sub-dividable into four clades and the second into six. Twelve species-specific CSP groups and 21 interspecific groups could be formed for within and between species sequence similarity comparisons (table 5.57).

As to the locust CSP expansions, the ancestral one shares ancestry with three *A. pisum* CSPs indicating that it probably derived from a hemimetabolous lineage. However, the weak branch support between that locust expansion, its *A. pisum* neighbours, and a clade composed by CSPs from several species does not support this option. Despite this, it is clear that the CSPs inside this locust specific expansion derive from a locust lineage, meaning that the expansion took place before *L. migratoria* and *S. gregaria* were separated. From its part, the more recent locust expansion is supported by a clade rich in CSP4 ortho- and paralogs, and includes nearly half of the locust CSPs described in this work, including some of the CSP sequences described in Angeli et al. [1999], Picimbon et al. [2000], Ban et al. [2002] and the majority of the *L. migratoria* paralogs—thus indicating a genuine locust CSP expansion.

Moreover, as if the number of complete locust CSPs identified here were not large enough, our search criteria were too restrictive as to retain pseudogenes, which presence has been reported in other species (summarized in [Sanchez-Gracia et al., 2009]). We have also detected 20 orphan CSP exons in the still unfinished genome sequence of *L. migratoria*, 12 of which are not assigned to a complete CSP sequence yet, and some show high identity to ESTs from the same species —meaning that they might be genuine and might even be expressed. The numbers of locust CSPs reported here might therefore still be incomplete, just like the available locust genome, and few more might still be for discovering at least in *L. migratoria* (the same logic should apply to *S. gregaria* where the absence of known genome sequence limited us to a transcriptomic study).

Despite the fact that the genomic information on *S. gregaria* is far more limited than on *L. migratoria*, whose draft genome is available, 16 out of 42 *S. gregaria* CSP sequences identified in the transcriptome from Badisco et al. [2011a] and in our RNA-seq assemblies did not associate directly with any *L. migratoria* genomic locus. Similarly, 31 out of the 57 *L. migratoria* CSPs reported here do not resemble any of the *S. gregaria* CSP transcripts (table 5.47). Species-specific CSPs expansions must thus have occurred with or without functional specialization.

5.4.4 Most of the large number of transcribed CSPs in locusts are linked to the phase change

As to whether the reported CSPs are functional or not, the case is clear for *S. gregaria*, whose CSPs reported here are transcripts, but not for *L. migratoria*, for which we report both transcripts and genomic sequences (22 of them with no detected *L. migratoria* transcript). The nucleotide diversity values of the comparisons between *L. migratoria* and *S. gregaria* CSP orthologs, between all the CSPs of the same species, and even between the CSP paralogs (those of the same species and same phylogenetic clade and those of the same genomic cluster (locus)) are relatively high (table 5.61). They are, for instance, between ten to a hundred times higher than those reported for the cis-regulatory sequences of *Drosophila*'s fushi tarazu gene [Bakkali, 2011] [0.045-0.505 versus 0.001-0.008, respectively]. This might be due to divergence of the different CSPs and functional relaxation or even loss of function of the phylogenetically related and duplicated sequences due to potential redundancy. Loss of function might seem in agreement with the fact that only one *L. migratoria* CSP transcript (LM_GH5_002985) was associated with the genes in the multi-CSP locus from the scaffold 71401, where eight conserved paralogs were detected (*LmigCSP38* to *LmigCSP45*). However this is no tangible argument as possible transcripts of seven of these paralogs might have been filtered out as alleles of the same gene during the assembly or sequence editing processes employed in Kang et al. [2004]—the work that reported the single LM_GH5_002985 CSP transcript. Despite the high nucleotide diversity, K_a/K_s values tell another story as their lower than one mean values indicate more synonymous than non-synonymous substitutions per site between these sequences—which explains the branch length differences between the congruent amino acids and nucleotide trees.

The mean data on nucleotide variability and K_a/K_s ratios hence suggest that locust CSPs seem in general under purifying selection, although the high standard deviations imply that this might variate considerably from one homologue pair to another. The K_a/K_s value is marginally lower for *S. gregaria* CSPs than for *L. migratoria*'s, probably due to the fact that the former are transcripts (all functional) whereas the latter are genomic sequences (not necessarily all of them functional). More importantly, the K_a/K_s values of the *L. migratoria* paralogs at the same genomic locus are not higher neither than those of the paralogous or orthologous CSPs that share the same phylogenetic clade but not the same physical location, nor than those that share neither phylogenetic clade nor physical location (tables 5.55 and 5.56). Signs of purifying selection are therefore evident for the different (duplicated) CSPs of the same multi-CSP loci—indication of them conserving function. The K_a/K_s values of inter-specific orthologs are lower than one, suggesting conservation of amino acids sequences, hence function. Only the different CSPs that do not share neither locus nor phylogenetic clade (paralogy or orthology) show K_a/K_s values above one due to their ancestral divergence and functional differences within the same genome.

The differences in CSP function usually associate with specialization in ligands. For example, in the cotton bollworm *Helicoverpa armigera*, a CSP specifically binds hormones whereas the rest of the CSPs bind plant compounds [Li et al., 2015]. A similar situation was described for *A. mellifera AmelCSP3* [Briand et al., 2002]. Furthermore, several studies report that some CSPs may also be involved in processes not related to chemoreception, such as reproduction [Bohbot et al., 1998], regeneration [Kitabayashi et al., 1998] and development [Maleszka et al., 2007]. The case of CSP19 is however striking due to the low proportion of non-synonymous substitutions between its *S. gregaria* and *L. migratoria* sequences —probably due to a conserved important function of that CSP for locusts (table 5.59).

The first and necessary step for a CSP gene function obviously is its transcription. We did not look into the cis-regulatory sequences upstream of each locus that contains a putative CSP in the *L. migratoria* genome. Still, it is worth highlighting that in the case of the multi-CSP locus in scaffold 71401, the genes are repeated in tandem with no promoter sequences between them. They are therefore possibly under the control of a single promoter. In fact, the *L. migratoria* EST LM_GH5_002985 contains an exon 1, an exon 2, and another exon 1 —proof of it being the product of the simultaneous (carry on) transcription of two CSP loci under the control of the same cis-regulatory region. This also raises questions about the post-transcriptional editing of the pre-mRNAs from these tandemly repeated genes and whether it results in a variety of CSP forms based on the number and combination of the repeated genes' exons.

We expected CSPs to either trigger, maintain, and/or be affected by the locust phase change. Accordingly, 90 % of the adult *S. gregaria* CSPs (38 out of 42) show significant differential expression between the solitarious and gregarious states. In adult *L. migratoria* however, 'only' 38 % (22 out of 57) of the CSPs show such differential expression. Yet, such huge difference does not mean that there are more CSPs involved in the phase change in *S. gregaria* than in *L. migratoria* as the sequencing libraries of both species were different. Our *S. gregaria* sequencing libraries contained CNS-enriched tissues whereas the available data on *L. migratoria* [Chen et al., 2010] did not. The higher amount of CSP read counts and the more pronounced differential expression of the CSPs in the CNS library, compared to the libraries from other tissues (figure 5.55B), thus explain the inter-species differences in the numbers of CSPs with significant differential expression between phases. Moreover, given the sensorial functions of the CSPs and the essential involvement of the CNS in the locust phase change (whose behavioural component is major), our data seem reasonable as they show more expression and more differences in expression between phases of the CSPs in the CNS, then in the muscles, then the guts, and hardly in the gonads. The latter showed almost no expression of the CSPs except for one, *SgreCSP14*, in the testicles. This CSP, despite not being significantly differentially expressed, shows high coverage (in number of sequencing reads) both in the gregarious and the solitarious testicle RNA-seq libraries. It is known that the gene ejaculatory bulb protein 3 (EBP3) is a CSP homolog [Bohbot et al., 1998], so we tested whether *SgreCSP14*, together with its orthologous *LmigCSP14*, could be locust EBP3 homologs that might have a reproduction-related function. Both locust CSPs appear in the first locust-specific CSP expansion and it seems that their origin is more recent than that of other CSPs. When EBP3 sequences from other species are included in a phylogenetic analysis of CSPs from various species, neither *SgreCSP14*, nor *LmigCSP14* appear in any EBP3-containing clade. Given that the possition of the EBP3 sequences, in general, is not conserved in the phylogeny and no locust CSP appears neighbouring an EBP3 in the CSP + EBP3 phylogeny, we cannot conclude on the potential EBP nature of our locust *SgreCSP14* sequence. However, the dispersed close relatedness between each EBP3 and a CSP also means that we cannot use sequence homology as a criterium for discarding the potential EBP nature of *SgreCSP14*, given its unexpectedly high expression level in *S. gregaria* testicle.

The CNS library is enriched with transcripts from antennae, palps and other locust sensory organs. It is therefore the most adequate material for studying CSPs. It is also the most obvious material for studying locusts' phase change. Our data show that 40 out of the 42 CSPs that we detected in *S. gregaria* CNS have significant differential expression between the solitarious and the gregarious phases. Most of the locust CSPs are therefore involved and/or affected by the phase change. Even more, the fact that 36 out of these 40 differentially expressed CSPs are over-expressed in the gregarious phase is concordant with the fact that that phase is sensorially more charged than the solitarious one —due to the increased and diversified amount of stimuli in the crowded populations. Such tendency seems a general characteristic to locusts as, overall, 37 % of the differentially expressed CSPs in *S. gregaria* and 40 % in *L. migratoria* show higher expression in the gregarious phase.

Still, CSPs are involved in a plethora of biological processes some of them could be expected to be general to different species (detection of food, chemicals...) while others could be species-specific (detection of mates, competitors...). In fact, Lester et al. [2005] reported that *S. gregaria* and *L. migratoria* can mutually trigger gregariousness in each other, although with much lower efficiency compared to the conspecific stimuli. In addition we reported species-dependent differences in the characteristics of the phase change between *L. migratoria* and *S. gregaria*, described in Chapter 1. One therefore would expect some of the phase-related CSPs to either be specific (i.e., not present in other species) or with species-dependent differential expression (i.e., linked to the phase change in one species but not in others), while other CSP orthologs would be linked to phase change in all locusts —these would be of great potential value to the fight against locusts' outbreaks. The latter group of CSPs (the ones linked to the phase change in all locusts) is expected to be ancestral to all swarming locusts. Interestingly, the most modern locusts' CSPs expansion clade shows a tendency to present higher expression levels in the gregarious phase, indistinctly of the instar or tissue (figure 5.46A and 5.46B). That expansion therefore predates gregariousness in locusts.

Despite the large amount of phase-linked CSPs in both *L. migratoria* and *S. gregaria* and the fact that most of these CSPs are over-expressed in the gregarious phase, only a few orthologs share similar expression tendency between both species —in all these cases the expression was increased in the gregarious phase.

As stated in the introduction, a CSP is already shown as involved in the phase change in *L. migratoria* [Guo et al., 2011]. We found its nearest *S. gregaria* ortholog (*SgreCSP37*) to share its expression pattern of higher expression in the gregarious phase in both nymphs and adults. With nucleotide sequence similarity of 90 % , amino acid sequence similarity of 92 % and conserved differential expression towards gregarious *L. migratoria* and *S. gregaria* locusts, *LmigCSP3* and its homolog *SgreCSP37* are expected to detect the same molecules and seem to not only be a general locust phase-related CSP but are very likely among the molecules that allow detection of stimuli from non-conspecifics. That CSP is therefore confirmed as general locust phase-related molecule.

Six more clades, containing the orthologous sequences *LmigCSP8* and *SgreCSP8*, *LmigCSP9* and *SgreCSP9*, *LmigCSP10* and *SgreCSP10*, *LmigCSP12* and *SgreCSP12*, *LmigCSP14* and *SgreCSP14* and *LmigCSP19* and *SgreCSP19* also show a conserved gregarious over-expression pattern. Applying the same logic as for *LmigCSP3* and *SgreCSP37*, the CSPs of these four additional clades provide more candidate molecules with potentially common implication in the phase change at least in the two main pest locust species. On the other hand, while the elevated expression levels of a locust CSP in the testicules might suggest that its is a locust EBP3, the phylogenetic relationships between the EBPs of different species and between these and the CSPs of the same species do not allow for a tangible confirmation of such possibility. Functional testing is therefore needed in order to determine the sequence of the locust EBP.

Out of the remaining 14 clades of CSPs with orthologs in both species (21 CSPs in total), 13 contain CSPs whose orthologs show opposite expression pattern in the other species. Furthermore, 15 *S. gregaria* and 31 *L. migratoria* CSPs seem species-specific, as they have no similar sequences. Most of these sequences which presence or expression patterns are different between species show over-expression in the gregarious phase —a congruent fact with the presumed greater need for stimuli detection in crowded conditions. These, therefore, offer a set of molecules for further genetic workflow testing (see Bakkali [2013]), some of which might be of potential interest to species-specific actions on locusts. No CSP shows conserved over-expression in solitarious *S. gregaria* and *L. migratoria* —a datum in accordance with the expected little need for stimuli detection in very low population densities.

In conclusion, we identified the nearly complete set of CSPs in two locust species. The fact that these organisms have the highest number and more diversified set of CSPs is mainly due to gene duplications and speaks to the essential nature of these molecules for the locust biology —locust phase change in particular. Accordingly, most of these CSPs show significant differential expression between locust phases and, in accordance with the greater need for stimuli detection in crowded conditions, most of the differentially expressed locusts' CSPs show higher expression in the gregarious phase. CSPs therefore offer potentially interesting molecules for dealing with locusts' outbreaks. Indeed, some of the CSPs share similar sequences and expression patterns between species and, hence, might be of general usefulness against all locusts, whereas others have either sequences or expression patterns that are species-specific and might be of use for species-specific works. Our work also allows a certain degree of speculation as to the functionality of such huge amount of locust CSPs but our interpretations on CSP functions need further functional genomics testing which we are planning for possible future works.

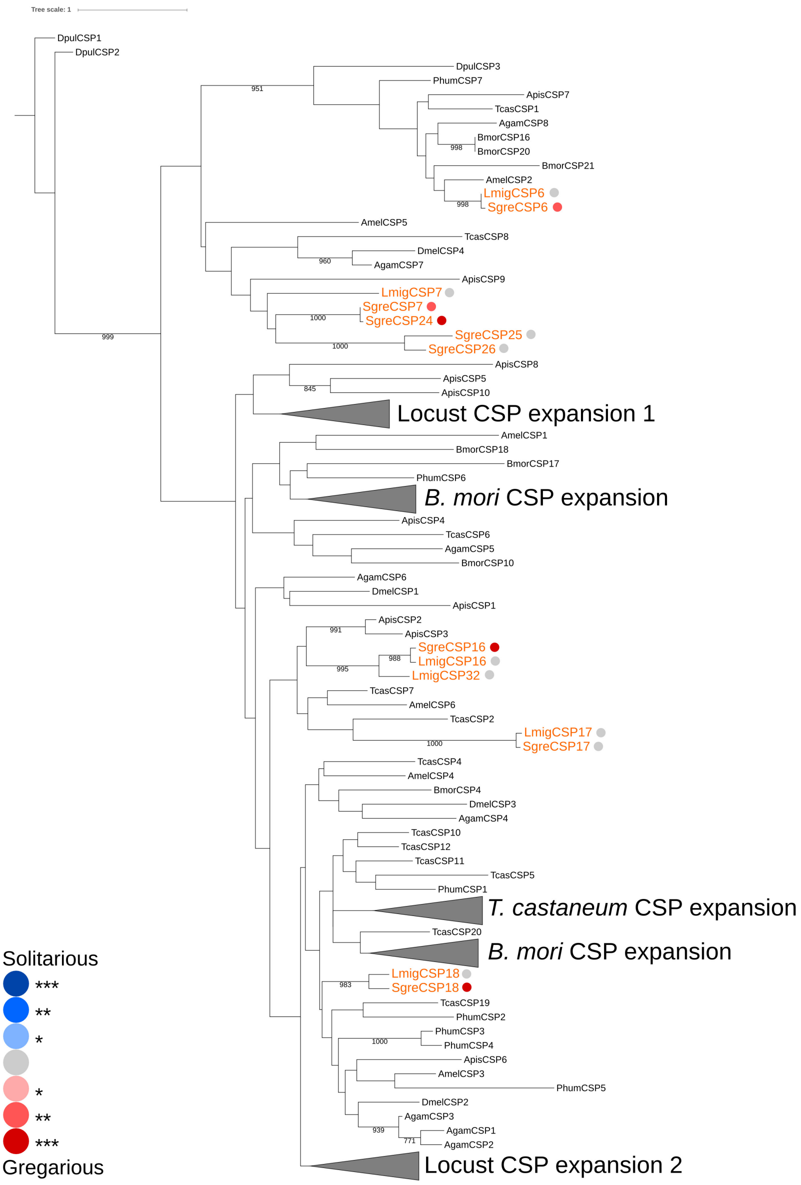

Figure 5.42: Maximum likelyhood phylogenetic tree of the amino acids sequences of CSPs from multiple arthropod species. Locust CSPs are marked in orange, and the large clades are compacted in triangles and developed when necessary. Branch lengths follow the scale shown in the figure, showing only branch supports higher than 75 %.

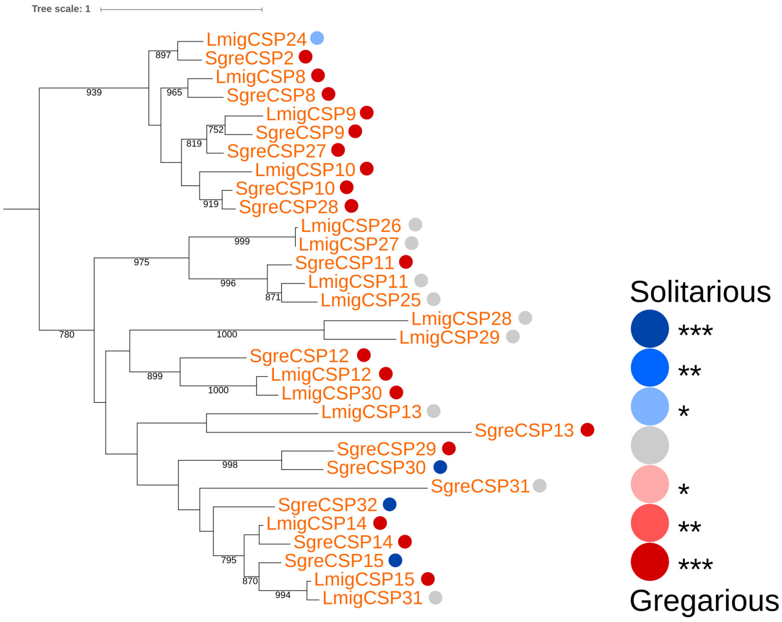

Figure 5.43: Maximum likelyhood phylogenetic tree of the amino acids sequences of CSPs for first locust expansion.

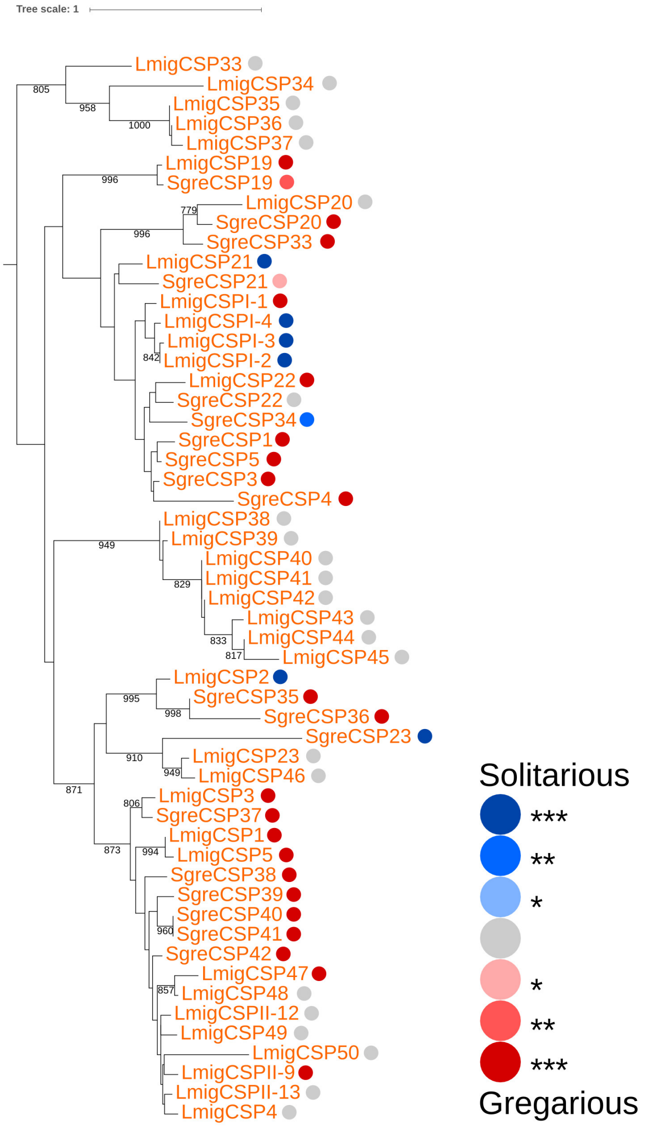

Figure 5.44: Maximum likelyhood phylogenetic tree of the amino acids sequences of CSPs for the second locust expansion.

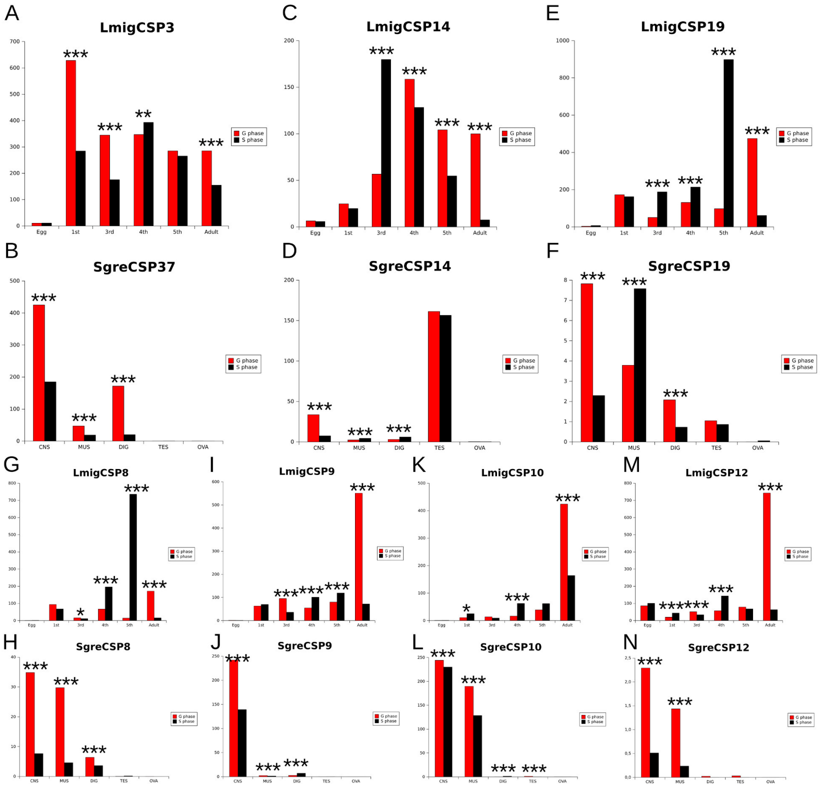

Figure 5.45: Expression profiles of the seven homologous *S. gregaria* and *L. migratoria* CSP pairs. The Y axis represents the NGS reads mapped to the CSP per million of total mapped reads (RPM). The X axis shows the instars for *L. migratoria* and the tissues for *S. gregaria* (CNS = central nervous system, MUS = toracic muscle, DIG = whole digestive tube, OVA = ovaries, TES = testicles). The asterisks indicate the significance level of the test after FDR correction (* = 0.05 - 0.01; ** = 0.01 - 0.001 and *** = 0.001 - 0).

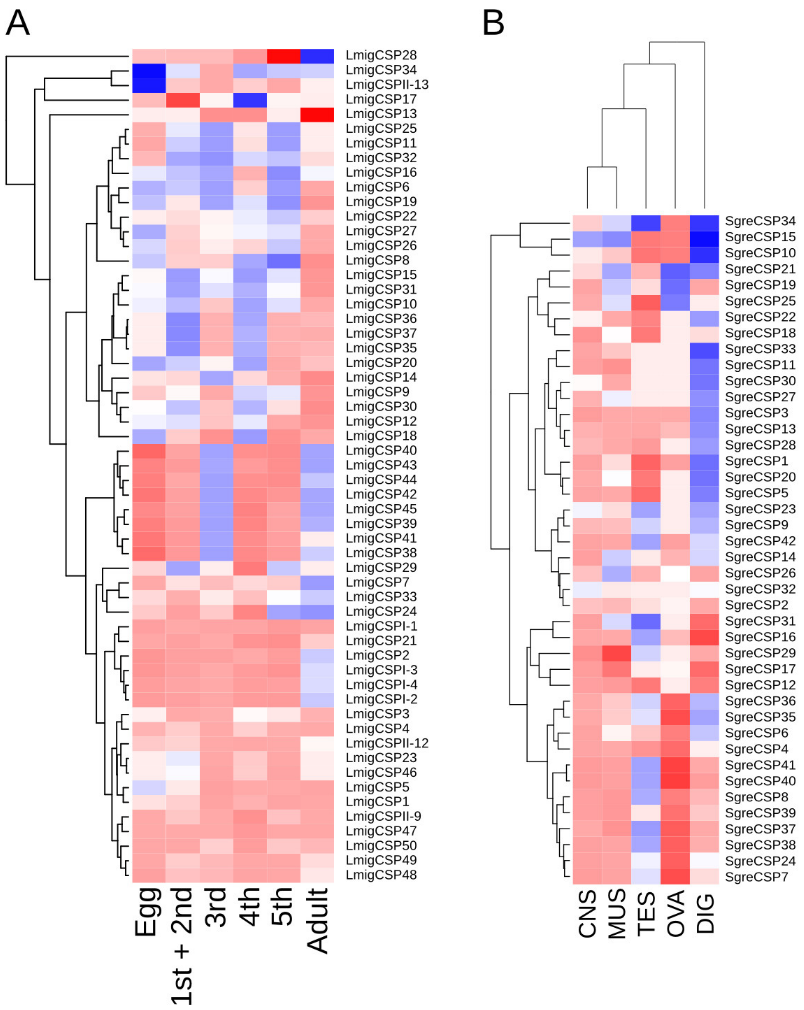

Figure 5.46: Solitarious to gregarious expression profiles of the locust CSPs identified in the present work. (A) Comparison between nymphal instars and adults of *L. migratoria*. (B) Comparison between tissues of *S. gregaria*. The expression levels are shown as color hues proportional to the fold change (see general methodology). Blue hues represent solitarious over-expression whereas red hues represent gregarious over-expression. The hues become lighter as the differences on expression become weaker, and white hues represent non-differential expression. Dendrograms agrupate the samples based on similarity of the expression profiles from CSPs (A and B) and tissues (only B). CNS = central nervous system, MUS = toracic muscle, DIG = whole digestive tube, OVA = ovaries, TES = testicles.

5.5 Supplementary material

Table 5.44: Detailed BLAST data on the *L. migratoria* genomic loci that putatively contain CSP genes and their corresponding transcripts. The table also includes data on single CSP exon sequences detected in scaffolds of the *L. migratoria* genome (orphan exons 1 and 2), exons 2 and 1 detected in this same unconventional order (orphan exon 2 + 1).

L.migratoria CSP loci	Exon 1 start	Exon 1 end	Exon 2 start	Exon 2 end	Best tBLASTn result	Identity (%)	E - value	Assigned EST	Identity (%)	E - value	ID exon1	ID exon2
101	2366516	2366626	2400458	2400577	BAS29776.1	73,17	4.00E-43	—	—	—	—	—
103059	2171	2379	10600	10788	AGB14643.1	40	5.00E-19	—	—	—	—	—
12585	179306	179145	167417	167247	2GVS	66	2.00E-46	LM_SH5_003244	83.64	3.00E-54	86.05	—
13671	595509	595372	591098	590973	AJP61952.1	45,05	5.00E-27	—	—	—	—	—
15810	78980	78801	63405	63220	CAB65178.1	99,17	1.00E-72	*LmigCSP2*	98.62	0.0	98.17	99.19
18858cds1	97840	97977	108499	108678	AAC25403.1	84,26	7.00E-64	LM_GH5_000761	82.53	5.00E-92	84.26	—
18858cds2	141703	141840	149531	149698	2GVS	81,65	1.00E-62	LM_GH5_000761	86.13	1.00E-117	85.03	90.48
18858cds3	168506	168673	170293	170451	AAC25403.1	62,89	1.00E-35	—	—	—	—	—
21551	122154	122026	97991	97824	JAA74384.1	55,17	3.00E-33	LM_SH5_001382	98.68	5.00E-156	100.00	98.13
22826cds1	159147	158959	153574	153386	CAB65181.1	89,6	3.00E-68	LM_SH5_003413	93.93	6.00E-166	88.83	98.95
22826cds2	127283	127504	129370	129558	CAB65181.1	88,8	6.00E-69	LM_GH5_003400	98.31	0.0	96.99	99.47
235750	7652	7834	9659	9847	CAB65181.1	88	7.00E-72	LM_SH5_003413	94.93	3.00E-169	91.94	97.87
24400	13865	14005	18439	18627	2GVS	87,16	2.00E-67	LM_GH5_000758	98.44	0.0	98.98	98.38
25611	14519	14635	63010	63192	AFQ07771.1	38,61	2.00E-28	—	—	—	—	—
2564	89690	89894	97291	97425	AGO81736.1	40,4	5.00E-24	—	—	—	—	—
30358	35537	35333	22564	22421	2GVS	40,4	1.00E-24	—	—	—	—	—
31810	78016	77812	67358	67182	AGZ04930.1	45,05	1.00E-24	—	—	—	—	—
320887	589	782	32999	33184	AAC25400.1	48,48	8.00E-33	LM_GH5_003053	99.22	0.0	100.00	98.45
3212cds1	1325340	1325136	1316189	1315951	AAP57461.1	41,89	1.00E-16	—	—	—	—	—
3212cds2	1382008	1381804	1363732	1363494	AAP57461.1	34,25	1.00E-11	—	—	—	—	—
325580	830	1072	24763	24951	2GVS	89,11	2.00E-63	LM_GB5_001536	96.11	6.00E-171	98.98	94.27
33302cds1	4672	4843	4844	5022	AAV68929.1	51,72	6.00E-39	—	—	—	—	—
33302cds2	10024	10195	10196	10374	AAV68929.1	51,72	2.00E-39	—	—	—	—	—
37289	4533	4737	10158	10346	AJP61955.1	39,42	3.00E-23	—	—	—	—	—
374630	5695	5866	5900	6078	AAV68929.1	50,43	8.00E-39	—	—	—	—	—
392768	61	249	21057	21224	AAC25401.1	69,16	6.00E-52	LM_GH5_003725	81.36	8.00E-60	—	—
41553	66870	67074	72744	72899	AIT38547.1	47,93	3.00E-33	LM_GH5_003053	79.91	2.00E-41	—	81.77
46375	44829	45033	61950	62108	2GVS	36	9.00E-21	—	—	—	—	—
5214cds1	122926	122795	116743	116531	CAB65180.1	66,04	4.00E-52	—	—	—	—	—
5214cds2	143789	143920	144992	145156	CAB65178.1	63,39	5.00E-53	—	—	—	—	—
57579	3592	3796	15234	15413	AGO81736.1	40,4	5.00E-24	—	—	—	—	—
647	198201	198088	176248	176012	XP_008193777.1	58,24	1.00E-37	LM_GL5_000033	95.94	0.0	96.10	95.78
699cds1	80447	80651	89125	89313	AIU68827.1	34,55	3.00E-20	—	—	—	—	—
699cds2	152813	152649	144189	144007	AIU68827.1	37,27	6.00E-23	—	—	—	—	—
71401cds1	45335	45210	44368	44165	CAB65179.1	60,78	2.00E-43	LM_GH5_002985	88.65	7.00E-46	87.96	—
71401cds2	44176	44000	42067	41897	2GVS	61,76	2.00E-47	LM_GH5_002985	97.49	1.00E-137	100.00	97.06
71401cds3	41890	41732	40886	40695	2GVS	60,4	3.00E-44	LM_GH5_002985	88.69	1.00E-97	100.00	—
71401cds4	40637	40509	38808	38659	CAB65179.1	61,36	2.00E-35	LM_GH5_002985	88.62	1.00E-82	—	97.71
71401cds5	38626	38468	35562	35392	2GVS	62,38	1.00E-46	LM_GH5_002985	98.94	5.00E-146	99.07	98.82
71401cds6	35385	35227	34126	33956	CAB65179.1	60,38	1.00E-45	LM_GH5_002985	92.26	5.00E-121	99.07	97.65

L.migratoria *CSP loci*	Exon 1 start	Exon 1 end	Exon 2 start	Exon 2 end	Best tBLASTn result	Identity (%)	E - value	Assigned EST	Identity (%)	E - value	ID exon1	ID exon2
71401cds7	33949	33791	31325	31212	2GVS	61,76	5.00E-47	LM_GH5_002985	98.92	2.00E-144	99.07	98.82
71401cds8	30988	30830	28783	28604	2GVS	61	8.00E-45	LM_GH5_002985	96.97	1.00E-77	100.00	—
757cds1	3626	3495	2231	2064	AAP57461.1	93	2.00E-67	LM_SH5_003413	97.50	9.00E-159	96.82	98.15
757cds2	60814	60692	58897	58709	CAB65181.1	97,6	4.00E-72	*LmigCSP4*	99.21	0.0	100.00	98.95
757cds3	95948	95760	92043	91849	CAB65179.1	100	9.00E-76	LM_SH5_003782	100.00	0.0	100.00	100.00
75957	4048	3907	3979	3689	AMA98187.1	59,79	2.00E-36	—	—	—	—	—
78016	21591	21387	16963	16775	AGB14643.1	40,51	3.00E-19	—	—	—	—	—
9174cds1	10000	10080	11335	11523	CAB65181.1	88,76	6.00E-56	LM_SH5_003244	96.90	3.00E-124	—	99.48
9174cds2	36759	36968	38209	38397	CAB65181.1	89,6	5.00E-71	LM_SH5_003413	94.44	3.00E-168	97.69	95.24
ORPHAN EXON 1												
100785	961	1110	—	—	AAC25403.1	68.00	2.00E-17	LM_GH5_000761	83.25	8.00E-50	83.17	—
157799226	98	6	—	—	CAB65177.1	96.77	5.00E-17	LM_SH5_003268	100.00	2.00E-46	100.00	—
173797773	5	160	—	—	CAB65178.1	65.38	2.00E-20	LM_GH5_000761	84.62	2.00E-37	84.51	—
50720	1643	1521	—	—	CAB65180.1	97.56	4.00E-19	LM_SH5_003413	98.68	3.00E-73	98.67	—
ORPHAN EXON 2												
123385	—	—	6441	6280	AAP57461.1	70.37	3.00E-20	*LmigCSP1*	82.73	2.00E-29	—	83.33
165997	—	—	1263	1442	AAC25400.1	83.33	2.00E-28	—	—	—	—	—
178632750	—	—	20	175	ACZ58021.1	48.08	1.00E-15	LM_GH5_003055	100.00	3.00E-107	—	100.00
187757636	—	—	479	655	AMA98187.1	64.41	1.00E-21	LM_GB5_004555	98.29	3.00E-86	—	98.80
221995	—	—	518	706	AAC25401.1	69.84	3.00E-25	—	—	—	—	—
281155	—	—	5573	5761	2GVS	90.48	1.00E-32	LM_GH5_000761	98.19	1.00E-138	—	98.19
286923	—	—	1496	1695	ALG36156.1	41.67	6.00E-14	—	—	—	—	—
316421	—	—	5589	5470	AAC25403.1	72.50	1.00E-13	—	—	—	—	—
392833	—	—	1068	940	AAC25403.1	65.12	3.00E-14	—	—	—	—	—
401450	—	—	218	412	CAB65177.1	95.38	2.00E-37	LM_SH5_003651	100.00	2.00E-100	—	100.00
53850	—	—	4232	4044	CAB65179.1	88.89	1.00E-31	LM_GH5_003400	99.48	2.00E-97	—	99.48
68729	—	—	7415	7227	2GVS	88.89	1.00E-31	LM_GH5_000760	97.19	4.00E-119	—	97.18
71274	—	—	23123	23311	AAC25401.1	77.78	1.00E-27	—	—	—	—	—
75485	—	—	3816	3628	EFA07417.1	49.21	3.00E-12	—	—	—	—	—
ORPHAN EXON 2 + 1												
189039548	—	438	250	—	AIT38541.1	57.14	5.00E-23	LM_GH5_002985	83.50	3.00E-78	—	—
15074	—	193525	193355	—	AIT38540.1	66.67	2.00E-19	LM_GH5_002985	99.11	2.00E-173	—	—

Table 5.45: Attribution of the available *L. migratoria* CSP transcripts to CSP proteins and *L. migratoria* genomic loci. The codes of the loci and their location in *L. migratoria* genomic scaffolds are as in table 5.44. The table also includes data on a CSP transcript to which no genomic locus could be assigned (no assigned locus), transcripts whose best BLAST result were orphan exon sequences detected in *L. migratoria* draft genome (orphan exon BLAST hit), partial ESTs whose sequences are identical to a larger sequences (100 % redundant), and sequences that do not present the CSP conserved cysteine pattern (no cysteine pattern).

L. migratoria *CSP EST*	Best BLASTx result	Identity (%)	E - value	Assigned locus	Identity (%)	Loci vs EST best hit	Identity (%)	Exon1 best hit	Identity (%)	Exon2 best hit	Identity (%)
LmigCSP1	CAB65177.1	100	0.0	757cds2	93.29	123385*	83.33	50720*	93.33	401450*	99.48
LmigCSP2	CAB65178.1	100	0.0	15810	98.62	15810	98.62	15810	98.17	15810	99.19
LmigCSP3	CAB65179.1	100	0.0	757cds3	99.74	757cds3	99.74	757cds3	100.00	53850*	92.63
LmigCSP4	CAB65180.1	100	0.0	757cds2	99.21	757cds2	99.21	50720*	98.66	757cds2	98.95
LmigCSP5	CAB65181.1	100	0.0	757cds2	97.88	757cds2	97.88	757cds1	96.77	757cds2	98.43
LmigCSPI-1	AY149648.1	100	0.0	325580	95.49	325580	95.49	325580	100.00	281155*	94.57
LmigCSPI-2	AY149649.1	100	0.0	24400	97.53	24400	97.53	24400	96.35	68729*	98.42
LmigCSPI-3	AY149650.1	100	0.0	24400	96.30	24400	96.30	24400	97.08	281155*	99.47
LmigCSPI-4	AY149651.1	100	0.0	24400	95.51	24400	95.51	24400	95.20	281155*	98.42
LmigCSPI-5	AY149652.1	100	0.0	24400	94.80	24400	94.80	24400	94.29	281155*	97.89
LmigCSPI-6	AY149653.1	100	0.0	24400	97.22	24400	97.22	24400	97.81	281155*	100.00
LmigCSPII-10	AY149658.1	100	0.0	757cds2	99.69	757cds2	99.69	757cds2	100.00	757cds2	100.00
LmigCSPII-11	AY149649.1	100	0.0	22826cds2	99.38	22826cds2	99.38	22826cds2	99.26	53850*	99.45
LmigCSPII-12	AY149650.1	100	0.0	235750	96.88	235750	96.88	757cds2	98.50	9174cds1	97.88
LmigCSPII-13	AY149651.1	100	0.0	9174cds2	95.71	9174cds2	95.71	9174cds2	94.82	9174cds2	96.81
LmigCSPII-14	AY149652.1	100	0.0	235750	98.12	235750	98.12	235750	99.24	235750	98.40
LmigCSPII-6	AY149654.1	100	0.0	15810	99.37	15810	99.37	15810	100.00	15810	99.17
LmigCSPII-7	AY149655.1	100	0.0	15810	98.13	15810	98.13	15810	97.22	15810	98.37
LmigCSPII-8	AY149656.1	100	0.0	15810	99.07	15810	99.07	15810	100.00	15810	98.37
LmigCSPII-9	AY149657.1	100	0.0	22826cds2	98.75	22826cds2	98.75	22826cds2	98.53	53850*	98.91
LM_GB5_001536	AAC25403.1	91	3.00E-69	325580	96.11	325580	96.11	325580	98.98	325580	94.27
LM_GB5_004555	AFQ07769.1	65	7.00E-45	C187757636*	98.80	C187757636*	98.80	22826cds2	82.50	C187757636*	98.80
LM_GB5_007405	CAB65177.1	99	1.00E-57	757cds2	90.31	757cds2	90.31	50720*	91.72	401450*	99.48
LM_GB5_007551	CAB65177.1	99	7.00E-58	757cds2	93.85	757cds2	93.85	50720*	93.94	401450*	100.00
LM_GB5_007735	CAB65179.1	100	4.00E-75	757cds3	99.47	757cds3	99.47	757cds3	100.00	53850*	91.05
LM_GH5_000758	2GVS	87	3.00E-67	24400	98.44	24400	98.44	24400	98.98	68729*	97.29
LM_GH5_000759	2GVS	87	3.00E-67	24400	98.18	24400	98.18	24400	98.48	68729*	96.86
LM_GH5_000760	2GVS	87	3.00E-67	24400	97.92	68729*	97.18	24400	97.97	68729*	97.17
LM_GH5_000761	2GVS	90	2.00E-68	24400	96.61	281155*	98.19	24400	97.97	281155*	98.19
LM_GH5_002985	2GVS	61	1.00E-40	71401cds5	98.94	71401cds5	98.94	71401cds2	100.00	71401cds7	98.82
LM_GH5_003053	AAC25400.1	48	1.00E-32	320887	99.22	320887	99.22	320887	100.00	320887	98.45
LM_GH5_003055	ACZ58021.1	48	5.00E-31	C178632750*	100.00	C178632750*	100.00	—	—	C178632750*	100.00
LM_GH5_003400	CAB65179.1	89	1.00E-65	22826cds2	98.31	53850*	99.48	22826cds2	96.99	53850*	99.47
LM_GH5_003478	CAB65181.1	96	6.00E-71	757cds2	96.83	757cds2	96.83	757cds1	98.71	22826cds1	96.83
LM_GH5_003489	CAB65181.1	86	5.00E-66	22826cds2	97.46	22826cds2	97.46	22826cds2	96.97	53850*	97.89
LM_GH5_003725	CAB65177.1	99	2.00E-57	757cds2	90.55	392768	81.36	50720*	100.00	401450*	100.00
LM_GH5_003820	CAB65179.1	100	8.00E-75	757cds3	100.00	757cds3	100.00	757cds3	100.00	53850*	92.11
LM_GH5_003822	CAB65179.1	98	5.00E-71	757cds3	99.68	757cds3	99.68	757cds3	100.00	53850*	91.58
LM_GL5_000033	CAI64033.1	53	5.00E-34	647	95.94	647	95.94	647	96.10	647	95.78
LM_GL5_000034	XP_008193777.1	58	1.00E-33	647	98.23	647	98.23	647	97.07	647	100.00

L. migratoria *CSP EST*	Best BLASTx result	Identity (%)	E - value	Assigned locus	Identity (%)	Loci vs EST best hit	Identity (%)	Exon1 best hit	Identity (%)	Exon2 best hit	Identity (%)
LM__GL5__000035	CAI64033.1	55	7.00E-38	647	96.17	647	96.17	647	95.61	647	96.88
LM__GM5__003208	AMA98182.1	50	3.00E-30	—	—	—	—	—	—	—	—
LM__GM5__003915	CAB65181.1	95	3.00E-70	757cds2	95.24	757cds2	95.24	50720*	98.67	22826cds1	98.95
LM__SH5__001382	JAA74384.1	56	1.00E-37	21551	98.68	21551	98.68	21551	100.00	21551	98.13
LM__SH5__003244	CAB65181.1	92	1.00E-71	22826cds2	95.69	9174cds1	96.90	22826cds2	96.93	9174cds1	99.48
LM__SH5__003268	CAB65177.1	99	6.00E-58	757cds2	91.39	C157799226*	100.00	50720*	91.72	401450*	98.95
LM__SH5__003270	CAB65181.1	94	4.00E-62	757cds1	100.00	757cds1	100.00	757cds1	100.00	757cds1	100.00
LM__SH5__003326	CAB65177.1	97	2.00E-70	757cds2	90.05	757cds2	90.05	50720*	92.65	401450*	99.48
LM__SH5__003413	CAB65181.1	95	2.00E-70	757cds2	95.24	50720*	98.67	50720*	98.67	22826cds1	98.95
LM__SH5__003512	CAB65179.1	100	2.00E-74	757cds3	100.00	757cds3	100.00	757cds3	100.00	53850*	92.11
LM__SH5__003651	CAB65177.1	98	5.00E-71	757cds2	93.54	401450*	100.00	50720*	93.18	401450*	100.00
LM__SH5__003782	CAB65179.1	100	2.00E-74	757cds3	100.00	757cds3	100.00	757cds3	100.00	53850*	92.11
LM__SL5__002526	CAB65181.1	88	3.00E-65	22826cds2	97.45	22826cds2	97.45	22826cds2	96.32	53850*	98.43
LM__SL5__002527	CAB65181.1	84	2.00E-57	22826cds2	96.08	22826cds2	96.08	22826cds2	96.36	53850*	95.81

Genome contig	Sense	Start	End	Best tBLASTn result	Identity (%)	E-value
k21_64501851 138 189	+1	1	120	gi|6688652|emb|CAB65179.1|	90	2.00E-22
k23_14559871 148 305	-3	146	3	gi|982743619|gb|AMA98187.1|	65	2.00E-16
k23_58371876 170 169	-1	158	3	gi|3283940|gb|AAC25403.1|	63	3.00E-20
k25_71126441 441 482 64156211+,...,43200394-	-1	252	4	gi|270010975|gb|EFA07423.1|	58	5.00E-29
k27_22627662 204 261	+2	2	202	gi|3283934|gb|AAC25400.1|	60	9.00E-25
k27_49827897 167 144	-1	167	87	gi|3283934|gb|AAC25400.1|	100	3.00E-15
k29_39894968 94 67	-3	92	3	gi|3283934|gb|AAC25400.1|	100	8.00E-18
k33_38727381 146 185	-1	146	3	gi|281426845|emb|CBA11329.1|	60	3.00E-17
k41_24247034 192 270	+1	1	192	gi|281426845|emb|CBA11329.1|	59	1.00E-23
k45_5189504 111 88	-2	110	3	gi|31747330|gb|AAP57461.1|	92	4.00E-20
k49_37213911 160 107 4851896+, 29241020-	-3	146	3	gi|6688652|emb|CAB65179.1|	79	1.00E-23

Table 5.46: Putative CSP sequences detected in our partial, low coverage, assembly of the *S. gregaria* genome. Due to low sequencing depth, only partial information on CSP genomic loci of *S. gregaria* could be obtained.

Table 5.47: List of the putative CSP transcripts from our NGS assemblies of *S. gregaria* transcriptomes as well as from the Sanger-sequenced ESTs by Badisco et al. [2011a]. The largest sequences among a group of putative alleles, at 93 % identity threshold, are marked in bold.

S. gregaria CSP transcript	Accession number	Identity (%)	E-value	*L. migratoria* loci homology	Identity (%)	E-value
SgEST001	gi|459277241|gb|JAA74384.1|	57.45	5,00E-38	21551	98.84	3,00E-57
SgEST011	gi|982743603|gb|AMA98179.1|	84.42	1.00E-44	101	89.61	2.00E-46
SgCNS002	gi|761545970|gb|AJP61955.1|	35.51	1.00E-18	37289	43.81	7.00E-32
SgCNS003	gi|401786704|gb|AFQ07771.1|	40.62	2.00E-30	25611	94.95	1.00E-65
SgCNS004	gi|982743617|gb|AMA98186.1|	43.00	2.00E-26	699cds2	57.80	2.00E-43
SgCNS005	gi|3283934|gb|AAC25400.1|	49.49	5.00E-33	320887	82.41	1.00E-66
SgCNS006	gi|3283940|gb|AAC25403.1|	68.09	3.00E-42	18858cds3	82.26	4.00E-50
SgCNS007	gi|3283940|gb|AAC25403.1|	65.96	8.00E-43	18858cds3	79.03	4.00E-48
SgCNS009	gi|761545970|gb|AJP61955.1|	38.46	4.00E-21	37289	88.79	5.00E-66
SgCNS010	gi|761545970|gb|AJP61955.1|	36.54	5.00E-22	37289	79.23	6.00E-60
SgCNS012	gi|3283932|gb|AAC25399.1|	83.33	2.00E-63	18858cds2	74.07	5.00E-65
SgCNS016	gi|443908519|gb|AGD80083.1|	40.18	2.00E-25	30358	73.98	2.00E-61
SgCNS023	gi|159164602|pdb|2GVS|A	91.94	7.00E-36	24400	85.48	7.00E-36
SgCNS024	gi|6688652|emb|CAB65179.1|	36.63	3.00E-23	37289	44.44	3.00E-32
SgDIG003	gi|48994220|emb|CAG26926.1|	58.73	3.00E-23	13671	53.85	1.00E-23
SgDIG004	gi|982743609|gb|AMA98182.1|	51.58	2.00E-30	41553	71.29	5.00E-54
SgDIG006	gi|281426845|emb|CBA11329.1|	46.60	2.00E-32	LM_GM5_003208	63.96	3.00E-52
SgDIG012	gi|6688656|emb|CAB65181.1|	91.20	3.00E-70	235750	93.40	5.00E-71
SgDIG017	gi|6688652|emb|CAB65179.1|	92.00	4.00E-70	757cds3	92.06	1.00E-70
SgDIG020	gi|6688656|emb|CAB65181.1|	88.80	9.00E-65	22826cds2	86.41	3.00E-73
SgDIG021	gi|159164602|pdb|2GVS|A	51.00	5.00E-33	41553	71.30	2.00E-58
SgDIG023	gi|6688656|emb|CAB65181.1|	92.00	3.00E-59	235750	95.28	4.00E-59
SgDIG025	gi|3283940|gb|AAC25403.1|	100.00	2.00E-75	24400	87.16	1.00E-67

S. gregaria CSP transcript	Accession number	Identity (%)	E-value	*L. migratoria* loci homology	Identity (%)	E-value
SgDIG027	gi|3283932|gb|AAC25399.1|	98.17	3.00E-74	24400	81.65	5.00E-65
SgDIG028	gi|6688656|emb|CAB65181.1|	93.02	6.00E-56	757cds2	90.80	9.00E-57
SgDIG030	gi|3283936|gb|AAC25401.1|	99.08	1.00E-75	24400	87.16	3.00E-68
SgDIG033	gi|159164602|pdb|2GVS|A	85.32	2.00E-66	24400	82.57	5.00E-65
SgDIG036	gi|4836779|gb|AAD30551.1|AF139197_1	37.23	3.00E-18	37289	64.00	1.00E-46
SgDIG038	gi|6688650|emb|CAB65178.1|	88.12	4.00E-65	15810	87.13	7.00E-63
SgDIG052	gi|6688656|emb|CAB65181.1|	94.74	1.00E-49	235750	96.05	4.00E-50
SgDIG078	gi|443908519|gb|AGD80083.1|	50.00	1.00E-29	LM_GM5_003208	79.00	9.00E-59
SgDIG094	gi|387158157|gb|AFJ54032.1|	31.33	3.00E-13	13671	31.08	1.00E-11
SgDIG096	gi|6688656|emb|CAB65181.1|	54.90	2.00E-39	5214cds2	59.66	1.00E-48
SgTES006	gi|48994224|emb|CAG26928.1|	51.95	6.00E-26	13671	53.16	5.00E-28
SgMUS004	gi|347943436|gb|AEP27186.1|	42.86	8.00E-30	LM_GM5_003208	80.53	1.00E-64
SgMUS005	gi|823091233|gb|AKI28975.1|	43.21	1.00E-19	13671	36.47	9.00E-19
SgMUS011	gi|291088334|dbj|BAI82449.1|	60.95	6.00E-46	757cds3	63	2.00E-40
SgMUS014	gi|3283936|gb|AAC25401.1|	68.22	1.00E-39	392768	96.26	3.00E-61
SgMUS025	gi|3283934|gb|AAC25400.1|	31.46	2.00E-14	37289	45.00	3.00E-31
SgMUS027	gi|761545966|gb|AJP61953.1|	27.42	7.00E-16	78016	42.22	5.00E-14
SgMUS029	gi|3283940|gb|AAC25403.1|	84.62	1.00E-35	325580	86.15	3.00E-37
SgMUS035	gi|6688650|emb|CAB65178.1|	86.05	7.00E-54	15810	84.88	2.00E-52
SgreCSP1	gi|3283932|gb|AAC25399.1|	100.00	3.00E-75	24400	81.65	5.00E-65
SgreCSP2	gi|3283934|gb|AAC25400.1|	100.00	7.00E-72	24400	80.39	2.00E-60
SgreCSP3	gi|3283936|gb|AAC25401.1|	100.00	2.00E-76	24400	86.24	1.00E-67
SgreCSP4	gi|159164602|pdb|2GVS|A	100.00	2.00E-76	24400	87.16	3.00E-68
SgreCSP5	gi|3283940|gb|AAC25403.1|	100.00	6.00E-76	24400	87.16	7.00E-68
SgCNS001	gi|761545970|gb|AJP61955.1|	35.58	1.00E-18	37289	43.81	2.00E-32
SgCNS015	gi|6688650|emb|CAB65178.1|	86.14	1.00E-62	15810	85.15	1.00E-60
SgCNS017	gi|6688656|emb|CAB65181.1|	53.33	9.00E-40	5214cds1	61.26	3.00E-47

S. gregaria CSP transcript	Accession number	Identity (%)	E-value	*L. migratoria* loci homology	Identity (%)	E-value
SgCNS018	gi|159164602|pdb|2GVS|A	50.00	2.00E-32	41553	70.37	2.00E-57
SgCNS019	gi|982743609|gb|AMA98182.1|	51.58	3.00E-30	41553	71.29	6.00E-54
SgCNS022	gi|3283936|gb|AAC25401.1|	97.25	9.00E-74	24400	87.16	8.00E-69
SgDIG001	gi|485220384|gb|JAA76614.1|	36.36	1.00E-13	37289	59.74	3.00E-32
SgDIG002	gi|433288640|gb|AGB14643.1|	39.06	1.00E-13	37289	74.49	1.00E-38
SgDIG007	gi|347943436|gb|AEP27186.1|	43.81	2.00E-30	LM_GM5_003208	79.65	8.00E-64
SgDIG008	gi|3283934|gb|AAC25400.1|	49.49	3.00E-33	320887	81.48	3.00E-66
SgDIG009	gi|3283934|gb|AAC25400.1|	49.49	6.00E-33	320887	82.41	1.00E-66
SgDIG010	gi|3283934|gb|AAC25400.1|	50.51	2.00E-33	320887	82.41	9.00E-67
SgDIG013	gi|6688656|emb|CAB65181.1|	92.00	5.00E-71	235750	94.34	1.00E-71
SgDIG014	gi|6688652|emb|CAB65179.1|	92.00	3.00E-70	757cds3	92.06	1.00E-70
SgDIG015	gi|3283936|gb|AAC25401.1|	68.22	9.00E-41	392768	96.26	1.00E-61
SgDIG018	gi|761545970|gb|AJP61955.1|	38.46	3.00E-21	37289	88.79	5.00E-66
SgDIG019	gi|761545970|gb|AJP61955.1|	38.46	2.00E-21	37289	88.79	5.00E-66
SgDIG022	gi|6688656|emb|CAB65181.1|	93.02	3.00E-56	757cds2	90.80	4.00E-57
SgDIG024	gi|6688656|emb|CAB65181.1|	91.86	1.00E-55	757cds2	89.66	5.00E-62
SgDIG026	gi|6688656|emb|CAB65181.1|	93.02	4.00E-56	757cds2	90.80	8.00E-57
SgDIG029	gi|159164602|pdb|2GVS|A	100.00	3.00E-76	24400	87.16	6.00E-68
SgDIG034	gi|291088334|dbj|BAI82449.1|	61.17	1.00E-45	757cds3	63	1.00E-40
SgDIG037	gi|281426845|emb|CBA11329.1|	46.60	2.00E-32	LM_GM5_003208	63.96	3.00E-52
SgDIG039	gi|3283932|gb|AAC25399.1|	82.41	9.00E-62	18858cds1	77.27	2.00E-59
SgDIG040	gi|982743609|gb|AMA98182.1|	51.58	2.00E-30	41553	72.28	1.00E-54
SgDIG041	gi|6688650|emb|CAB65178.1|	88.12	4.00E-65	15810	87.13	7.00E-63
SgDIG043	gi|3283932|gb|AAC25399.1|	88.99	2.00E-67	18858cds2	87.27	4.00E-66
SgDIG044	gi|4836779|gb|AAD30551.1|AF139197_1	37.23	7.00E-18	37289	64.00	2.00E-46
SgDIG045	gi|31747330|gb|AAP57461.1|	100.00	3.00E-48	757cds2	92.96	9.00E-48
SgDIG046	gi|6688650|emb|CAB65178.1|	88.12	4.00E-65	15810	87.13	7.00E-63

S. gregaria CSP transcript	Accession number	Identity (%)	E-value	*L. migratoria* loci homology	Identity (%)	E-value
SgDIG047	gi|982743609|gb|AMA98182.1|	51.58	2.00E-29	41553	71.29	5.00E-53
SgDIG048	gi|982743609|gb|AMA98182.1|	48.78	1.00E-24	320887	61.90	4.00E-38
SgDIG049	gi|6688656|emb|CAB65181.1|	92.00	3.00E-59	235750	95.28	4.00E-59
SgDIG050	gi|3283932|gb|AAC25399.1|	96.97	7.00E-66	325580	84.85	7.00E-59
SgDIG051	gi|6688656|emb|CAB65181.1|	53.92	7.00E-39	5214cds2	58.82	7.00E-48
SgDIG053	gi|6688650|emb|CAB65178.1|	88.68	1.00E-68	15810	87.74	2.00E-65
SgDIG055	gi|982743609|gb|AMA98182.1|	51.58	2.00E-30	41553	71.29	5.00E-54
SgDIG056	gi|347943436|gb|AEP27186.1|	42.86	7.00E-30	LM_GM5_003208	80.53	1.00E-64
SgDIG057	gi|347943436|gb|AEP27186.1|	42.86	5.00E-30	LM_GM5_003208	80.53	1.00E-64
SgDIG058	gi|761545970|gb|AJP61955.1|	38.46	3.00E-21	37289	88.79	4.00E-66
SgDIG059	gi|761545970|gb|AJP61955.1|	38.46	3.00E-21	37289	88.79	4.00E-66
SgDIG060	gi|6688656|emb|CAB65181.1|	91.20	3.00E-58	235750	94.34	2.00E-58
SgDIG061	gi|761545970|gb|AJP61955.1|	38.46	2.00E-21	37289	88.79	4.00E-66
SgDIG062	gi|761545970|gb|AJP61955.1|	38.46	2.00E-21	37289	88.79	4.00E-66
SgDIG063	gi|6688656|emb|CAB65181.1|	92.45	2.00E-70	235750	94.34	3.00E-71
SgDIG064	gi|3283932|gb|AAC25399.1|	98.17	4.00E-74	24400	81.65	5.00E-65
SgDIG065	gi|761545970|gb|AJP61955.1|	38.46	2.00E-21	37289	88.79	1.00E-65
SgDIG067	gi|3283932|gb|AAC25399.1|	98.17	5.00E-75	24400	81.58	7.00E-68
SgDIG068	gi|761545970|gb|AJP61955.1|	39.02	6.00E-17	37289	83.33	1.00E-47
SgDIG069	gi|3283932|gb|AAC25399.1|	82.41	7.00E-62	18858cds2	73.15	7.00E-60
SgDIG070	gi|3283936|gb|AAC25401.1|	99.02	7.00E-70	24400	86.27	2.00E-62
SgDIG071	gi|3283940|gb|AAC25403.1|	98.98	4.00E-66	325580	86.73	2.00E-57
SgDIG072	gi|6688656|emb|CAB65181.1|	93.40	2.00E-58	235750	78.70	1.00E-47
SgDIG073	gi|433288640|gb|AGB14643.1|	40.28	2.00E-15	37289	87.65	1.00E-48
SgDIG074	gi|31747330|gb|AAP57461.1|	90.54	2.00E-44	757cds2	87.84	9.00E-54
SgDIG075	gi|3283940|gb|AAC25403.1|	100.00	8.00E-59	18858cds1	85.06	3.00E-51
SgDIG076	gi|31747330|gb|AAP57461.1|	98.25	8.00E-36	757cds2	92.98	1.00E-37

S. gregaria CSP transcript	Accession number	Identity (%)	E-value	*L. migratoria* loci homology	Identity (%)	E-value
SgDIG077	gi|982743609|gb|AMA98182.1|	53.73	2.00E-20	41553	66.67	2.00E-32
SgDIG079	gi|433288640|gb|AGB14643.1|	35.38	2.00E-11	37289	59.15	1.00E-28
SgDIG080	gi|959478398|gb|ALR72515.1|	66.13	1.00E-28	18858cds2	70	4.00E-25
SgDIG081	gi|959478398|gb|ALR72515.1|	65.62	4.00E-29	18858cds2	68	3.00E-25
SgDIG082	gi|761545970|gb|AJP61955.1|	34.69	6.00E-19	37289	81.19	2.00E-55
SgDIG083	gi|31747330|gb|AAP57461.1|	100.00	6.00E-33	757cds2	94.23	6.00E-35
SgDIG084	gi|31747330|gb|AAP57461.1|	90.54	1.00E-44	757cds2	87.84	7.00E-54
SgDIG085	gi|3283940|gb|AAC25403.1|	98.90	3.00E-60	325580	87.91	2.00E-53
SgDIG086	gi|3283940|gb|AAC25403.1|	97.80	2.00E-59	325580	86.81	1.00E-52
SgDIG087	gi|6688652|emb|CAB65179.1|	89.66	6.00E-41	757cds3	89.66	7.00E-43
SgDIG088	gi|6688650|emb|CAB65178.1|	89.47	3.00E-47	15810	88.16	8.00E-46
SgDIG089	gi|3283940|gb|AAC25403.1|	69.33	1.00E-33	18858cds1	65.33	3.00E-37
SgDIG090	gi|6688652|emb|CAB65179.1|	90.00	5.00E-37	757cds3	91.14	5.00E-39
SgDIG091	gi|6688652|emb|CAB65179.1|	91.49	4.00E-47	757cds3	91.49	2.00E-48
SgDIG092	gi|3283940|gb|AAC25403.1|	98.73	6.00E-51	18858cds1	86.08	6.00E-46
SgDIG093	gi|6688652|emb|CAB65179.1|	89.36	3.00E-46	757cds3	89.36	9.00E-47
SgDIG095	gi|48994220|emb|CAG26926.1|	58.73	3.00E-23	13671	53.85	9.00E-24
SgDIG097	gi|761545968|gb|AJP61954.1|	49.35	3.00E-20	LM_GM5_003208	77.11	4.00E-43
SgDIG098	gi|239790047|dbj|BAH71609.1|	50.70	7.00E-20	LM_GM5_003208	80.77	6.00E-42
SgDIG099	gi|6688656|emb|CAB65181.1|	54.90	2.00E-39	5214cds2	58.82	7.00E-48
SgDIG100	gi|6688656|emb|CAB65181.1|	54.90	2.00E-39	5214cds1	60.87	2.00E-48
SgDIG101	gi|642928414|ref|XP_008193777.1|	44.87	4.00E-21	LM_GM5_003208	76.19	9.00E-44
SgEST002	gi|982743603|gb|AMA98179.1|	84.42	6.00E-45	101	89.61	2.00E-46
SgEST005	gi|373212596|gb|AEY60886.1|	80.46	4.00E-50	101	90.24	8.00E-50
SgEST006	gi|982743603|gb|AMA98179.1|	84.42	7.00E-45	101	89.61	2.00E-46
SgEST007	gi|982743603|gb|AMA98179.1|	84.42	3.00E-45	101	89.61	2.00E-46
SgEST008	gi|982743603|gb|AMA98179.1|	84.42	5.00E-45	101	89.61	2.00E-46

S. gregaria CSP transcript	Accession number	Identity (%)	E-value	*L. migratoria* loci homology	Identity (%)	E-value
SgEST009	gi|982743603|gb|AMA98179.1|	84.42	7.00E-45	101	89.61	2.00E-46
SgEST012	gi|459277241|gb|JAA74384.1|	57.45	5.00E-38	21551	98.84	3.00E-57
SgMUS001	gi|48994224|emb|CAG26928.1|	51.56	2.00E-19	13671	54.84	2.00E-21
SgMUS002	gi|761545966|gb|AJP61953.1|	26.61	3.00E-15	78016	42.22	2.00E-13
SgMUS006	gi|357611229|gb|EHJ67380.1|	40.00	7.00E-14	13671	31.08	6.00E-12
SgMUS007	gi|823091233|gb|AKI28975.1|	43.21	1.00E-19	13671	36.47	9.00E-19
SgMUS008	gi|3283940|gb|AAC25403.1|	64.89	1.00E-42	18858cds3	79.03	1.00E-46
SgMUS009	gi|347943436|gb|AEP27186.1|	41.90	2.00E-28	LM_GM5_003208	79.65	3.00E-63
SgMUS010	gi|347943436|gb|AEP27186.1|	42.86	6.00E-30	LM_GM5_003208	80.53	1.00E-64
SgMUS012	gi|823091233|gb|AKI28975.1|	41.98	4.00E-19	13671	36.47	5.00E-19
SgMUS013	gi|761545970|gb|AJP61955.1|	39.42	7.00E-22	37289	87.85	1.00E-65
SgMUS015	gi|3283936|gb|AAC25401.1|	68.22	2.00E-40	392768	96.26	3.00E-61
SgMUS016	gi|3283936|gb|AAC25401.1|	69.16	1.00E-40	392768	97.20	4.00E-62
SgMUS017	gi|3283936|gb|AAC25401.1|	69.16	2.00E-40	392768	96.26	2.00E-61
SgMUS018	gi|3283936|gb|AAC25401.1|	68.22	3.00E-39	392768	96.26	5.00E-61
SgMUS019	gi|6688652|emb|CAB65179.1|	36.63	4.00E-23	37289	44.44	9.00E-33
SgMUS021	gi|761545966|gb|AJP61953.1|	27.42	6.00E-16	78016	42.22	4.00E-14
SgMUS022	gi|48994224|emb|CAG26928.1|	51.95	6.00E-26	13671	53.16	5.00E-28
SgMUS023	gi|3283934|gb|AAC25400.1|	31.46	2.00E-14	37289	45.54	2.00E-31
SgMUS024	gi|48994224|emb|CAG26928.1|	51.95	6.00E-26	13671	53.16	5.00E-28
SgMUS030	gi|823091233|gb|AKI28975.1|	43.21	1.00E-19	13671	36.47	9.00E-19
SgMUS031	gi|347943436|gb|AEP27186.1|	42.86	8.00E-30	LM_GM5_003208	80.53	1.00E-64
SgMUS032	gi|6688656|emb|CAB65181.1|	55.91	1.00E-35	5214cds1	59.18	7.00E-41
SgMUS034	gi|4836779|gb|AAD30551.1|AF139197_1	37.23	5.00E-18	37289	65.00	8.00E-48
SgMUS036	gi|3283936|gb|AAC25401.1|	73.33	6.00E-26	392768	94.81	5.00E-38
SgMUS037	gi|3283940|gb|AAC25403.1|	68.09	3.00E-42	18858cds3	82.26	2.00E-49
SgMUS038	gi|6688656|emb|CAB65181.1|	91.86	1.00E-55	757cds2	89.66	1.00E-56

S. gregaria CSP transcript	Accession number	Identity (%)	E-value	*L. migratoria* loci homology	Identity (%)	E-value
SgMUS039	gi|6688656|emb|CAB65181.1|	91.20	6.00E-59	235750	94.34	7.00E-59
SgMUS040	gi|3283936|gb|AAC25401.1|	96.33	1.00E-72	24400	86.36	8.00E-68
SgMUS041	gi|6688656|emb|CAB65181.1|	89.60	3.00E-65	22826cds2	86.41	6.00E-73
SgMUS043	gi|31747330|gb|AAP57461.1|	96.92	1.00E-42	757cds1	98.46	3.00E-50
SgMUS046	gi|6688656|emb|CAB65181.1|	54.90	1.00E-39	5214cds2	59.66	1.00E-48
SgMUS047	gi|3283936|gb|AAC25401.1|	96.67	6.00E-59	24400	85.56	4.00E-54
SgMUS048	gi|3283940|gb|AAC25403.1|	68.09	2.00E-42	18858cds3	82.26	3.00E-49
SgMUS049	gi|6688656|emb|CAB65181.1|	91.51	7.00E-56	757cds2	89.72	2.00E-58
SgMUS050	gi|4836779|gb|AAD30551.1|AF139197_1	37.23	5.00E-18	37289	65.00	8.00E-48
SgMUS051	gi|31747330|gb|AAP57461.1|	98.72	1.00E-53	757cds1	91.14	4.00E-58
SgMUS052	gi|6688656|emb|CAB65181.1|	90.36	5.00E-53	235750	93.98	1.00E-59
SgMUS055	gi|31747330|gb|AAP57461.1|	97.96	4.00E-55	235750	93.00	1.00E-53
SgMUS056	gi|6688656|emb|CAB65181.1|	57.89	4.00E-30	5214cds2	53.26	7.00E-35
SgMUS057	gi|433288640|gb|AGB14643.1|	35.90	6.00E-15	37289	61.45	1.00E-35
SgMUS058	gi|6688656|emb|CAB65181.1|	54.90	1.00E-39	5214cds2	58.82	3.00E-48
SgTES001	gi|817050604|gb|AKF17719.1|	59.18	4.00E-33	24400	44	5.00E-23
SgTES003	gi|817050604|gb|AKF17719.1|	58.76	2.00E-32	24400	44	8.00E-23
SgTES004	gi|817050604|gb|AKF17719.1|	58.76	3.00E-32	24400	44	8.00E-23
SgTES007	gi|642928414|ref|XP__008193777.1|	47.71	2.00E-32	41553	71.30	2.00E-58
SgTES008	gi|642928414|ref|XP__008193777.1|	47.71	4.00E-32	41553	71.30	1.00E-57
SgTES009	gi|6688656|emb|CAB65181.1|	92.00	1.00E-65	235750	93.00	9.00E-66

Locus	Alleles	Identity	Locus	Alleles	Identity
21551	LM_SH5_001382	98.0	757cds2	*LmigCSPII-10*	98.8
				LM_GH5_003478	96.8
320887	LM_GH5_003053	99.2		*LmigCSP4*	99.2
				LmigCSP5	97.9
71401cds5	LM_GH5_002985	97.6			
			22826cds2	*LmigCSPII-9*	97.9
757cds1	LM_SH5_003270	97.1		*LmigCSPII-11*	98.5
				LM_GH5_003489	95.2
325580	LM_GB5_001536	96.1		LM_SL5_002526	95.2
				LM_SL5_002527	93.8
235750	*LmigCSPII-12*	96.0			
	LmigCSPII-14	97.2	24400	*LmigCSPI-2*	97.2
				LmigCSPI-3	96.0
647	LM_GL5_000033	95.7		*LmigCSPI-6*	96.9
	LM_GL5_000034	98.2		LM_GH5_000758	98.4
	LM_GL5_000035	96.2		LM_GH5_000759	98.2
15810	*LmigCSP2*	98.6	757cds3	*LmigCSP3*	99.7
	LmigCSPII-6	98.8		LM_GB5_007735	99.5
	LmigCSPII-7	98.1		LM_GH5_003820	100.0
	LmigCSPII-8	99.1		LM_GH5_003822	98.7
				LM_SH5_003512	100.0
				LM_SH5_003782	100.0

Table 5.48: Sequence identity between the different *L. migratoria* loci and their assignated alleles. Only unduplicated (single copy) loci are included in the analysis. Identity in bold indicates the identity threshold for clustering the *S. gregaria de novo* assembled transcripts.

	Exon 1		Exon 2	
Incongruent sequences	Associated sequences	Identity values	Incongruent sequences	Associated sequences
SgMUS011	SgMUS035	65	SgMUS011	SgMUS035
SgMUS035	21551	68	SgMUS035	21551
SgMUS029	18858cds1	77	SgMUS029	18858cds1
SgMUS005	57579	53	SgMUS005	57579
SgMUS027	46375	56	SgMUS027	46375
SgTES006	78016	50	SgTES006	78016
SgMUS025	9174cds1	50	SgMUS025	9174cds1
SgDIG003	SgCNS003	35	SgDIG003	SgCNS003
SgDIG038	LM3400	74	SgDIG038	LM3400
SgDIG096	LM3400	59	SgDIG096	LM3400
SgEST001	SgEST011	57	SgEST001	SgEST011
SgCNS003	SgEST001	59	SgCNS003	SgEST001
SgDIG094	SgCNS004	48	SgDIG094	SgCNS004
Reference loci pair	78016 vs 103059	97	30358 vs 46375	95

Table 5.49: *S. gregaria* putative CSP transcripts that show incongruent phylogenetic positions of their exons 1 and 2. The table indicates the phylogenetically nearest CSP to each incongruent exon, as well as the distance between both sequences. In the lower part of the table are the *L. migratoria* CSPs whose exons' distance were used as threshold for determining the potential homology of the incongruent *S. gregaria* CSP exon sequences.

CSP name	Code	Published homolog sequences	Accession
LmigCSP1	LM_GH5_003725	*LmigCSP1*	gi|6688648|emb|CAB65177.1|
LmigCSP2	15810	*LmigCSP2*	gi|6688650|emb|CAB65178.1|
		LmigCSPII-6	gi|27543485|gb|AY149654.1|
		LmigCSPII-7	gi|27543487|gb|AY149655.1|
		LmigCSPII-8	gi|27543489|gb|AY149656.1|
LmigCSP3	757cds3	*LmigCSP3*	gi|6688652|emb|CAB65179.1|
LmigCSP4	757cds2	*LmigCSP4*	gi|6688654|emb|CAB65180.1|
LmigCSP5	LM_SH5_003268	*LmigCSP5*	gi|6688654|emb|CAB65181.1|
		LmigCSPII-10	gi|27543493|gb|AY149658.1|
LmigCSPI-1	325580	*LmigCSPI-1*	gi|27543473|gb|AY149648.1|
LmigCSPI-2	24400	*LmigCSPI-2*	gi|27543475|gb|AY149649.1|
LmigCSPI-3	LM_GH5_000760	*LmigCSPI-3*	gi|27543477|gb|AY149650.1|
LmigCSPI-4	LM_GH5_000761	*LmigCSPI-4*	gi|27543479|gb|AY149651.1|
		LmigCSPI-5	gi|27543481|gb|AY149652.1|
		LmigCSPI-6	gi|27543483|gb|AY149653.1|
LmigCSPII-9	22826cds2	*LmigCSPII-9*	gi|27543491|gb|AY149657.1|
		LmigCSPII-11	gi|27543495|gb|AY149649.1|
LmigCSPII-12	235750	*LmigCSPII-12*	gi|27543497|gb|AY149650.1|
		LmigCSPII-14	gi|27543501|gb|AY149652.1|
LmigCSPII-13	9174cds2	*LmigCSPII-13*	gi|27543499|gb|AY149651.1|
SgreCSP1	SgDIG027	*SgreCSP1*	gi|3283932|gb|AAC25399.1|
SgreCSP2	SgCNS005	*SgreCSP2*	gi|3283934|gb|AAC25400.1|
SgreCSP3	SgDIG030	*SgreCSP3*	gi|3283936|gb|AAC25401.1|
SgreCSP4	SgCNS023	*SgreCSP4*	gi|159164602|pdb|2GVS|A
SgreCSP5	SgDIG025	*SgreCSP5*	gi|3283940|gb|AAC25403.1|

Table 5.50: Naming and homologies of the *L. migratoria* and *S. gregaria* CSPs already identified in other works.

L. migratoria		*S. gregaria*	
CSP name	Code	CSP name	Code
LmigCSP6	101	*SgreCSP6*	SgEST011
LmigCSP7	13671	*SgreCSP7*	SgTES006
LmigCSP8	LM_GH5_003055	*SgreCSP8*	SgDIG006
LmigCSP9	LM_GM5_003208	*SgreCSP9*	SgMUS004
LmigCSP10	41553	*SgreCSP10*	SgDIG004
LmigCSP11	46375	*SgreCSP11*	SgCNS016
LmigCSP12	699cds2	*SgreCSP12*	SgCNS004
LmigCSP13	31810	*SgreCSP13*	SgMUS027
LmigCSP14	37289	*SgreCSP14*	SgCNS009
LmigCSP15	78016	*SgreCSP15*	SgCNS010
LmigCSP16	21551	*SgreCSP16*	SgEST001
LmigCSP17	25611	*SgreCSP17*	SgCNS003
LmigCSP18	LM_GB5_004555	*SgreCSP18*	SgMUS011
LmigCSP19	392768	*SgreCSP19*	SgMUS014
LmigCSP20	18858cds1	*SgreCSP20*	SgCNS006
LmigCSP21	18858cds2	*SgreCSP21*	SgCNS012
LmigCSP22	18858cds3	*SgreCSP22*	SgDIG033
LmigCSP23	5214cds2	*SgreCSP23*	SgDIG096
LmigCSP24	320887	*SgreCSP24*	SgDIG003
LmigCSP25	30358	*SgreCSP25*	SgDIG094
LmigCSP26	57579	*SgreCSP26*	SgMUS005
LmigCSP27	2564	*SgreCSP27*	SgDIG078
LmigCSP28	3212cds1	*SgreCSP28*	SgDIG021
LmigCSP29	3212cds2	*SgreCSP29*	SgCNS002
LmigCSP30	699cds1	*SgreCSP30*	SgCNS024
LmigCSP31	103059	*SgreCSP31*	SgMUS025
LmigCSP32	647	*SgreCSP32*	SgDIG036
LmigCSP33	12585	*SgreCSP33*	SgCNS007
LmigCSP34	75957	*SgreCSP34*	SgMUS029
LmigCSP35	33302cds1	*SgreCSP35*	SgDIG038
LmigCSP36	33302cds2	*SgreCSP36*	SgMUS035
LmigCSP37	374630	*SgreCSP37*	SgDIG017
LmigCSP38	71401cds4	*SgreCSP38*	SgDIG028
LmigCSP39	71401cds6	*SgreCSP39*	SgDIG020
LmigCSP40	71401cds7	*SgreCSP40*	SgDIG012
LmigCSP41	71401cds5	*SgreCSP41*	SgDIG052
LmigCSP42	71401cds2	*SgreCSP42*	SgDIG023
LmigCSP43	71401cds8		
LmigCSP44	71401cds3		
LmigCSP45	71401cds1		
LmigCSP46	5214cds1		
LmigCSP47	LM_GH5_003400		
LmigCSP48	22826cds1		
LmigCSP49	757cds1		
LmigCSP50	9174cds1		

Table 5.51: Naming and homologies of the *L. migratoria* and *S. gregaria* CSPs identified in the current work.

LmigCSP	Egg	1st + 2nd	3rd	4th	5th	Adult
LmigCSP1	-0.185	-0.462***	-2.000***	-0.923***	-1.771***	-1.504***
LmigCSP2	-2.849*	-2.084**	-2.119***	-1.083***	-2.303***	1.018***
LmigCSP3	0.029	-1.143***	-0.974***	0.180**	-0.102	-0.881***
LmigCSP4	-0.849	-0.518***	-1.204***	-0.459***	-0.946***	-1.043
LmigCSP5	0.844**	-0.088	-2.066***	-1.091***	-1.438***	-1.774***
LmigCSPI-1	-2.180**	-1.133***	-1.792***	-1.836***	-2.618***	-1.763***
LmigCSPI-2	-2.466***	-2.163***	-1.511***	-2.488***	-2.532***	1.035***
LmigCSPI-3	-2.849***	-2.028***	-1.420***	-2.480***	-3.038***	0.769***
LmigCSPI-4	-2.264***	-1.768***	-1.442***	-2.295***	-2.213***	0.780***
LmigCSPII-12	-0.527	-0.394	-1.334***	-1.716***	-1.118*	0.124
LmigCSPII-13	12.124	-0.556	-1.363***	-0.475**	-1.708*	0.033
LmigCSPII-9	-1.013	-0.575**	-1.358***	-3.211***	-0.670**	-1.364***
LmigCSP6	1.387	1.026	2.145***	-0.449	2.606***	-1.648
LmigCSP7	-1.024***	-0.178	-0.744	-0.310	-0.515	2.522
LmigCSP8	1.057	-0.467	-0.491*	1.547***	5.663***	-3.448***
LmigCSP9	-0.527	0.143	-1.393***	0.891***	0.574***	-2.934***
LmigCSP10	0.473	1.199*	-0.583	1.937***	0.667*	-1.371***
LmigCSP11	-1.527	0.974	2.265***	-0.223	2.265	0.000
LmigCSP12	0.455	0.668	-0.632***	0.587***	-1.687***	-3.010***
LmigCSP13	0.000	0.000	-3.322	-3.322*	0.000	-13.446
LmigCSP14	-0.149	-0.331	1.665***	-0.306***	-0.930***	-3.745***
LmigCSP15	0.151	2.578***	0.337	2.598***	0.655	-2.773***
LmigCSP16	0.582	1.138**	1.597***	-0.896***	3.555***	0.366
LmigCSP17	-0.761***	-8.925	0.104	10.020	0.102	0.000
LmigCSP18	1.473	-0.5	-3.2***	2.56**	-16.612	17.338
LmigCSP19	1.151	-0.091	1.889***	0.705***	3.201***	-2.953***
LmigCSP20	1.644	0.950***	0.104*	1.977***	-1.401***	-0.666
LmigCSP21	-1.527	-2.027***	-1.466***	-3.083***	-3.353***	-0.518*
LmigCSP22	0.000	-0.248	0.076	0.516	1.070	-0.478
LmigCSP23	0.000	0.559	-1.581***	-0.529	-1.581	0.000
LmigCSP24	-0.527	-2.179***	-0.529*	-4.047***	1.823***	3.024*
LmigCSP25	-0.949	0.585	2.262***	-0.129	2.262	0.000
LmigCSP26	0.837	-0.500	-0.005	-0.375	1.070	-1.060
LmigCSP27	1.473	-0.441	0.127	0.434	0.655	-1.385
LmigCSP28	-0.691***	-0.763	-0.763	-2.710	-13.343	10.648
LmigCSP29	-0.383***	1.822	0.000	-4.857***	1.070	0.000
LmigCSP30	0.227	1.131***	-0.599***	1.324***	-0.212	-3.542***
LmigCSP31	0.321	2.010***	0.689	1.403***	0.292	-2.852***
LmigCSP32	-0.808***	1.509***	3.095***	0.821***	1.144***	-0.248
LmigCSP33	-0.305	-0.949***	-0.060	-0.678**	0.271	1.004
LmigCSP34	13.124	0.697	-1.234***	1.511***	1.070	0.937
LmigCSP35	0.000	3.407	-0.903***	1.425***	-0.903	-0.801
LmigCSP36	0.000	3.822*	-0.910***	1.618***	-0.910	-0.818
LmigCSP37	0.000	3.822*	-0.889***	1.390***	-0.889	-0.961
LmigCSP38	-6.051***	-2.667***	2.009***	-3.869***	-2.421***	0.980
LmigCSP39	-4.615***	-2.521***	2.067***	-4.145***	-1.950***	1.411
LmigCSP40	-6.360***	-1.881***	1.914***	-3.377***	-3.591***	1.813
LmigCSP41	-5.142***	-2.666***	1.876***	-3.531***	-2.190***	0.000
LmigCSP42	-5.190***	-2.055***	2.029***	-3.842***	-2.853***	1.714
LmigCSP43	-4.214***	-2.547***	1.601***	-2.411***	-3.394***	1.937
LmigCSP44	-4.577***	-1.674***	2.036***	-3.450***	-3.541***	1.082
LmigCSP45	-3.791***	-2.079***	1.926***	-4.290***	-1.816***	1.526
LmigCSP46	0.000	0.309	-1.842***	-0.518*	-1.842	0.000
LmigCSP47	-1.421	-1.314***	-1.237***	-2.759***	-1.714***	-1.671***
LmigCSP48	-2.112	-0.659	-0.803***	-1.756***	-1.549***	-0.099
LmigCSP49	-1.527	-0.507	-0.795***	-1.610***	-1.555**	-0.281
LmigCSP50	-1.527	-1.522**	-0.449*	-2.307***	-0.558	-0.818

Table 5.52: Differences in the expression level of the CSPs between solitarious and gregarious *L. migratoria*. The differences are expressed in fold change (see general methodology). The significance of the test after FDR correction is shown by adding an asterisk to the fold change values as follows: * = 0.05 - 0.01; ** = 0.01 - 0.001; *** = 0.001 - 0.

SgreCSP	CNS	MUS	DIG	TES	OVA	ALL
SgreCSP1	-2.010***	-0.254***	5.685***	-6.589	-2.077	-1.214***
SgreCSP2	-0.659***	-0.773***	-1.137***	-0.224	0.075	-0.410***
SgreCSP3	-2.356***	-1.067***	3.979***	-1.293	-1.340	-1.686***
SgreCSP4	-1.874***	-1.446***	0.018	-3.091	-6.153	-1.550***
SgreCSP5	-1.861***	-1.389***	4.586***	-6.011	0.000	-1.500***
SgreCSP6	-1.960***	0.096	1.116	-0.606	-4.608	-1.256**
SgreCSP7	-2.277***	-2.001***	-0.262	0.687	-8.253	-1.648***
SgreCSP8	-2.184***	-2.693***	-0.813***	1.494	-4.608	-2.200***
SgreCSP9	-0.797***	-0.720**	1.326***	0.909	0.000	-0.284***
SgreCSP10	-0.086**	-0.560***	10.368***	-5.339***	-4.608	-0.075***
SgreCSP11	-2.045***	-3.170***	5.336	0.000	0.000	-1.827***
SgreCSP12	-2.165***	-2.604***	-4.623	-5.034	0.000	-2.081***
SgreCSP13	-0.918***	-1.134***	3.438*	-1.615*	-0.755	-1.031***
SgreCSP14	-2.168***	0.973***	0.844***	-0.044	-0.870	-0.224***
SgreCSP15	2.051***	3.681***	13.042***	-5.034	-4.608	2.669***
SgreCSP16	-1.749***	-1.112	-8.567***	1.996	-0.755	-1.112***
SgreCSP17	-2.303*	-5.241	-6.168	0.000	0.130	-0.775
SgreCSP18	-3.138***	0.217	-0.259	-5.034	0.000	-2.549***
SgreCSP19	-1.771***	0.999***	-1.512***	-0.277	5.819	-0.285**
SgreCSP20	-1.365***	0.268***	5.438***	-5.034	0.000	-0.581***
SgreCSP21	-0.297***	1.509***	4.223***	-0.869	6.806	0.617***
SgreCSP22	0.046	-1.192***	2.485***	-4.179**	0.000	-0.894***
SgreCSP23	0.391***	-0.261***	1.860***	1.996*	0.000	0.349***
SgreCSP24	-2.096***	-1.854***	0.323	0.449	-6.882	-1.496***
SgreCSP25	-0.981	0.738	0.001	-7.000	4.844	-0.251
SgreCSP26	-0.651*	1.447**	-1.765*	-0.898	0.245	0.241
SgreCSP27	-0.907***	0.478***	3.381**	0.000	0.000	-0.314***
SgreCSP28	-0.774***	-1.429***	2.579***	-2.615***	0.000	-1.013***
SgreCSP29	-3.371***	-8.719***	-1.262	0.909	0.000	-3.481***
SgreCSP30	0.205***	-1.033***	5.336	0.000	0.000	0.697***
SgreCSP31	-2.122***	0.818	-6.168	5.922	0.000	-0.730
SgreCSP32	0.528***	-0.072	0.291	0.000	0.000	0.899***
SgreCSP33	-1.830***	-0.574***	8.305**	0.000	0.000	-1.130***
SgreCSP34	-0.471***	0.884***	10.057***	9.148**	-4.608	0.182**
SgreCSP35	-1.491***	-0.512***	1.825	0.687	-8.253	-0.942***
SgreCSP36	-1.234***	-0.535***	1.323	0.909	-6.563	-0.840***
SgreCSP37	-1.200***	-3.101***	-1.330***	2.494	-7.143	-1.146***
SgreCSP38	-2.084***	-1.775***	-0.930***	1.572	-6.882	-1.860***
SgreCSP39	-2.133***	-2.702***	-0.568*	-0.091	-5.578	-2.469***
SgreCSP40	-2.533***	-2.024***	-2.294***	1.842**	-9.408***	-2.212***
SgreCSP41	-2.497***	-2.171***	-1.414***	1.909	-9.135*	-2.247***
SgreCSP42	-1.943***	-1.596***	0.820	1.909	-2.077	-1.684***

Table 5.53: Differences in the expression level of the CSPs between solitarious and gregarious *S. gregaria*. The differences are expressed in fold change (see general methodology). The significance of the test after FDR correction is shown by adding an asterisk to the fold change values as follows: * = 0.05 - 0.01; ** = 0.01 - 0.001; *** = 0.001 - 0.

Order	Species	Number of CSPs	C-value
Blattaria	*Periplaneta americana*	7	2.72
Coleoptera	*Leptinotarsa decemlineata*	4	0.46
	Tribolium castaneum	20	0.21
Diptera	*Anopheles gambiae*	8	0.27
	Chironomus tentans	2	0.22
	Drosophila melanogaster	4	0.18
Hemiptera	*Acyrthosiphon pisum*	10	0.31
	Aphis gossypii	8	0.67
	Myzus persicae	8	0.32
	Rhodnius prolixus	4	0.59
Hymenoptera	*Nasonia vitripennis*	2	0.34
	Solenopsis invicta	29	0.77
	Apis meliphera	6	0.24
	Vespula squamosa	8	0.17
Lepidoptera	*Antheraea mylitta*	6	1
	Bicyclus anynana	10	0.49
	Bombyx mori	21	0.52
	Danaus plexippus	4	0.29
	Heliconius melpomene	9	0.3
Orthoptera	*Gryllus bimaculatus*	20	2.68
	Laupala kohalensis	8	1.93
	Locusta migratoria	57	6.35
	Schistocerca gregaria	42	8.96
Phthiraptera	*Pediculus humanus*	7	0.11

Table 5.54: Relation between number of CSPs and genome size in insects. The number of CSPs and C-values were retrieved from the works summarized in Xu et al. [2009] and the the animal genome size database Gregory [2001], respectively. The Pearson correlation coefficient between both variants is $R^2 = 0.623$.

Scaffold ID	CSP	*LmigCSP45*	*LmigCSP42*	*LmigCSP44*	*LmigCSP38*	*LmigCSP41*	*LmigCSP39*	*LmigCSP40*	*LmigCSP43*
71401cds1	*LmigCSP45*	100.00	78.20	89.50	79.60	83.00	79.00	79.40	81.90
71401cds2	*LmigCSP42*	78.20	100.00	86.60	87.90	98.00	89.40	96.30	84.40
71401cds3	*LmigCSP44*	89.50	86.60	100.00	80.50	88.50	79.60	89.20	88.00
71401cds4	*LmigCSP38*	79.60	87.90	80.50	100.00	90.20	96.60	90.20	77.50
71401cds5	*LmigCSP41*	83.00	98.00	88.50	90.20	100.00	92.20	100.00	84.50
71401cds6	*LmigCSP39*	79.00	89.40	79.60	96.60	92.20	100.00	91.30	76.50
71401cds7	*LmigCSP40*	79.40	96.30	89.20	90.20	100.00	91.30	100.00	84.70
71401cds8	*LmigCSP43*	81.90	84.40	88.00	77.50	84.50	76.50	84.70	100.00
Scaffold ID	CSP	*LmigCSP20*	*LmigCSP21*	*LmigCSP22*	Scaffold ID	CSP	*LmigCSP3*	*LmigCSP4*	*LmigCSP49*
18858cds1	*LmigCSP20*	100.00	79.20	70.00	757cds3	*LmigCSP3*	100.00	85.10	84.40
18858cds2	*LmigCSP21*	79.20	100.00	71.00	757cds2	*LmigCSP4*	85.10	100.00	90.20
18858cds3	*LmigCSP22*	70.00	71.00	100.00	757cds1	*LmigCSP49*	84.40	90.20	100.00

Table 5.55: Pairwise nucleotide identity matrices of the CSP sequences that belong to the scaffold 71401, 18858 and 757 of the *L. migratoria* draft genome.

Scaffold ID	CSP	*LmigCSP45*	*LmigCSP42*	*LmigCSP44*	*LmigCSP38*	*LmigCSP41*	*LmigCSP39*	*LmigCSP40*	*LmigCSP43*
71401cds1	*LmigCSP45*		0.221	0.603	0.230	0.225	0.233	0.225	0.104
71401cds2	*LmigCSP42*	0.221		0.181	0.339	0.181	0.301	0.181	0.092
71401cds3	*LmigCSP44*	0.603	0.181		0.211	0.202	0.214	0.202	0.059
71401cds4	*LmigCSP38*	0.230	0.339	0.211		0.456	0.520	0.456	0.094
71401cds5	*LmigCSP41*	0.225	0.181	0.202	0.456		0.384	0.000	0.086
71401cds6	*LmigCSP39*	0.233	0.301	0.214	0.520	0.384		0.384	0.104
71401cds7	*LmigCSP40*	0.225	0.181	0.202	0.456	0.000	0.384		0.086
71401cds8	*LmigCSP43*	0.104	0.092	0.059	0.094	0.086	0.104	0.086	
Scaffold ID	CSP	*LmigCSP20*	*LmigCSP21*	*LmigCSP22*	Scaffold ID	CSP	*LmigCSP3*	*LmigCSP4*	*LmigCSP49*
18858cds1	*LmigCSP20*		0.143	0.152	757cds3	*LmigCSP3*		0.205	0.195
18858cds2	*LmigCSP21*	0.143		0.368	757cds2	*LmigCSP4*	0.205		0.283
18858cds3	*LmigCSP22*	0.152	0.368		757cds1	*LmigCSP49*	0.195	0.283	

Table 5.56: Pairwise ratio of the mean non-synonymous to synonymous substitutions between the sequences that belong to the scaffold 71401, 18858 and 757 of the *L. migratoria* draft genome.

CSP sequences included	π	SD π	θ	SD θ	(Mean) $\mathrm{K}_a/\mathrm{K}_s$	SD $\mathrm{K}_a/\mathrm{K}_s$
LmigCSP3,LmigCSP4,LmigCSP49	0.086	0.027	0.085	0.052	0.110	0.074
LmigCSP20,LmigCSP21,LmCSP22	0.270	0.079	0.255	0.154	0.267	0.152
LmigCSP38,LmigCSP39,LmigCSP40, LmigCSP41 LmigCSP42,LmigCSP43, LmigCSP44,LmigCSP45	0.122	0.021	0.111	0.049	0.304	0.327

Table 5.57: Intra- and inter-specific sequence variability of the CSPs that belong to the same genomic locus and/or phylogenetic clade.

CSP sequences included	π	SD π	θ	SD θ	(Mean) $\mathrm{K}_a/\mathrm{K}_s$	SD $\mathrm{K}_a/\mathrm{K}_s$
LmigCSP33,LmigCSP34,LmigCSP35, LmigCSP36, LmigCSP37	0.205	0.060	0.200	0.101	0.173	0.134
LmigCSPII-12, LmigCSPII-13, LmigCSP48, LmigCSP49, LmigCSP50	0.140	0.038	0.147	0.074	0.548	0.223
LmigCSP26,LmigCSP32	0.500	0.250	0.500	0.355	0.718	N. A.
LmigCSP28,LmigCSP29	0.260	0.130	0.260	0.185	0.819	N. A.
LmigCSPI-3, LmigCSPI-4, LmigCSPI-1, LmigCSPI-2	0.055	0.015	0.015	0.031	0.241	0.091
SgreCSP1,SgreCSP3,SgreCSP4, SgreCSP5	0.126	0.033	0.126	0.069	0.249	0.244
SgreCSP39,SgreCSP40,SgreCSP41	0.070	0.028	0.070	0.043	0.027	0.024
SgreCSP25,SgreCSP26	0.206	0.103	0.206	0.146	0.801	N. A.
SgreCSP29,SgreCSP30	0.201	0.101	0.201	0.143	0.926	N. A.
SgreCSP31,SgreCSP32	0.467	0.234	0.467	0.331	0.520	N. A.
LmigCSP1, LmigCSP4	0.021	0.010	0.021	0.016	0.094	N. A.
LmigCSP47, LmigCSPII-9	0.014	0.0073	0.014	0.011	0.449	N. A.

Table 5.58: Intra-specific sequence variability of the CSPs that belong to sequences sharing a intra-specific phylogenetic clade or sequences in a clade that contains a single inter-specific pair of orthologs.

CSP sequences included	π	SD π	θ	SD θ	(Mean) K$_a$/K$_s$	SD K$_a$/K$_s$
LmigCSP3,SgreCSP37	0.095	0.048	0.095	0.068	0.125	N. A.
LmigCSP6,SgreCSP6	0.057	0.028	0.057	0.042	0.051	N. A.
LmigCSP7,SgreCSP7	0.389	0.194	0.389	0.276	0.511	N. A.
LmigCSP8,SgreCSP8	0.137	0.069	0.137	0.098	0.260	N. A.
LmigCSP13,SgreCSP13	0.505	0.253	0.505	0.358	0.756	N. A.
LmigCSP14,SgreCSP14	0.138	0.069	0.138	0.099	0.203	N. A.
LmigCSP17,SgreCSP17	0.111	0.056	0.111	0.080	0.109	N. A.
LmigCSP18,SgreCSP18	0.155	0.077	0.155	0.110	0.235	N. A.
LmigCSP19,SgreCSP19	0.046	0.023	0.046	0.034	0.091	N. A.
LmigCSP21,SgreCSP21	0.212	0.106	0.212	0.151	0.195	N. A.
LmigCSP24,SgreCSP2	0.132	0.132	0.132	0.094	0.330	N. A.

Table 5.59: Inter-specific sequence variability of the CSPs that belong to sequences in a clade that contains a sequence and more than one inter-specific otholog.

CSP sequences included	π	SD π	θ	SD θ	(Mean) K$_a$/K$_s$	SD K$_a$/K$_s$
LmigCSP11,SgreCSP11, (LmigCSP25)	0.190	0.095	0.190	0.135	0.152	0.067
LmigCSP12,SgreCSP12, (LmigCSP30)	0.363	0.182	0.363	0.258	0.378	0.194
LmigCSP15,SgreCSP15, (LmigCSP31)	0.217	0.108	0.217	0.154	0.385	0.066
LmigCSP16,SgreCSP16, (LmigCSP32)	0.119	0.060	0.119	0.085	0.337	0.355
LmigCSP23,SgreCSP23, (LmigCSP46)	0.266	0.133	0.266	0.189	0.281	0.248
LmigCSP2,SgreCSP35, (SgreCSP36)	0.182	0.091	0.182	0.130	0.172	0.149
LmigCSP9,SgreCSP9, (SgreCSP27)	0.204	0.102	0.204	0.145	0.116	0.043
LmigCSP10,SgreCSP10, (SgreCSP28)	0.240	0.120	0.240	0.171	0.242	0.109
LmigCSP20,SgreCSP20, (SgreCSP33)	0.302	0.151	0.302	0.215	0.153	0.141
LmigCSP22,SgreCSP22, (SgreCSP34)	0.317	0.159	0.317	0.225	0.234	0.115

Table 5.60: Inter-specific sequence variability of the CSPs that belong to all paired comparisons and umpaired comparisons (intra and inter-specific).

CSP sequences included	π	SD π	θ	SD θ	(Mean) K$_a$/K$_s$	SD K$_a$/K$_s$
All analyzed sequences	0.405	0.012	0.183	0.048	0.376	0.245
All remaining comparisons	0.360	0.093	0.333	0.167	0.659	0.479
Only *L. migratoria* remaining comparisons	0.462	0.231	0.462	0.328	0.736	0.524
Only *S. gregaria* remaining comparisons	0.315	0.129	0.307	0.185	0.595	0.396
All sequences	0.398	0.012	0.179	0.047	0.649	0.479
Only *L. migratoria* sequences	0.406	0.016	0.206	0.059	0.712	0.524
Only *S. gregaria* sequences	0.430	0.018	0.216	0.064	0.572	0.409

Table 5.61: Intra- and inter-specific sequence variability of the CSPs that belong to all possible comparisons (intra- and interspeciffic).

CSP sequences included	π	SD π	θ	SD θ	Pairwise K_a/K_s		
	Intra-specific				Intra-specific		Inter-Specific
					Locusta migratoria	*Schistocerca gregaria*	
LmigCSP11,SgreCSP11, (LmigCSP25)	0.124	0.062	0.124	0.088	0.514	N. A.	0.308
LmigCSP12,SgreCSP12, (LmigCSP30)	0.098	0.049	0.098	0.071	0.596	N. A.	0.317
LmigCSP15,SgreCSP15, (LmigCSP31)	0.069	0.034	0.069	0.050	0.326	N. A.	0.317
LmigCSP16,SgreCSP16, (LmigCSP32)	0.225	0.113	0.225	0.161	0.286	N. A.	0.056
LmigCSP23,SgreCSP23, (LmigCSP46)	0.045	0.022	0.045	0.045	0.633	N. A.	0.536
LmigCSP2,SgreCSP35, (SgreCSP36)	0.158	0.079	0.158	0.113	N. A.	0.301	0.358
LmigCSP9,SgreCSP9, (SgreCSP27)	0.177	0.088	0.177	0.126	N. A.	0.893	0.198
LmigCSP10,SgreCSP10, (SgreCSP28)	0.077	0.038	0.077	0.055	N. A.	0.133	0.223
LmigCSP20,SgreCSP20, (SgreCSP33)	0.141	0.071	0.141	0.101	N. A.	0.259	0.245
LmigCSP22,SgreCSP22, (SgreCSP34)	0.167	0.083	0.167	0.119	N. A.	0.171	0.364

Table 5.62: Intra- and inter-specific sequence variability of the CSPs that belong to the remaining inter-species clade subsets.

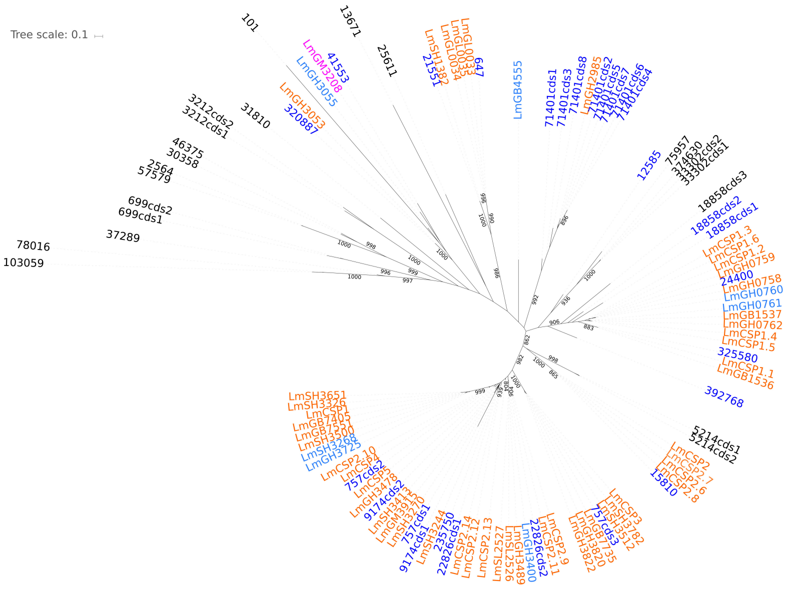

Figure 5.47: Maximum likelyhood phylogenetic tree of the nucleotide sequences of the *L. migratoria* CSP genes and ESTs. The CSPs whose sequences were identified in *L. migratoria* genome are in blue (with positive BLAST result against ESTs) or black (without positive BLAST result against ESTs), *L. migratoria* ESTs are in orange, loci derived from ESTs are in light blue and the only CSP EST for which no genomic locus has been identified is in pink. Only branch supports that are higher than 75 % are shown in their respective branches. The green bubbles mark the sequence clusters used for establishing the identity threshold for removal of redundant alleles from *S. gregaria* CSP transcripts.

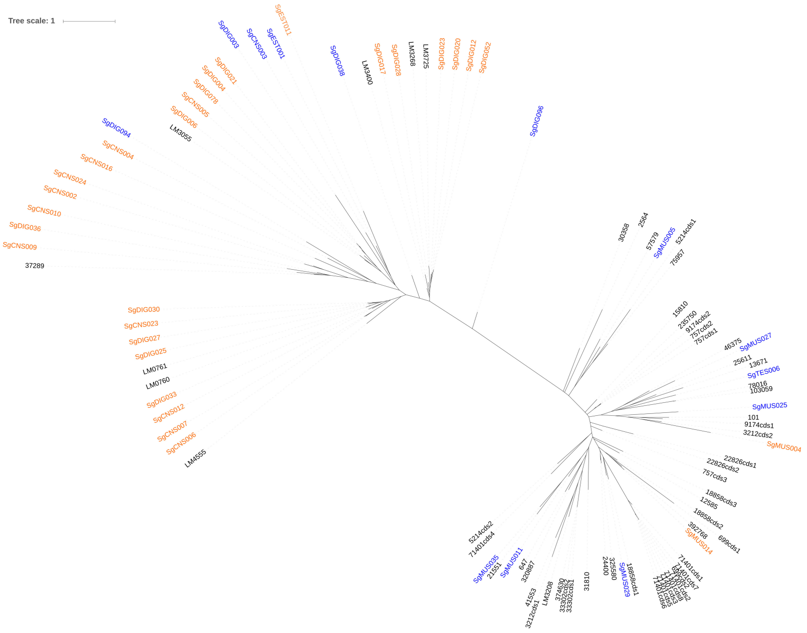

Figure 5.48: Maximum likelyhood phylogenetic tree of the nucleotide of exon 1 *L. migratoria* and *S. gregaria* CSPs. *L. migratoria* CSPs are in black and *S. gregaria* transcripts in orange. Incongruently placed *S. gregaria* exon sequences are in blue. The branch length scale is at the top of the trees. Only branch support values that are higherthan 75 % are shown in their respective branches.

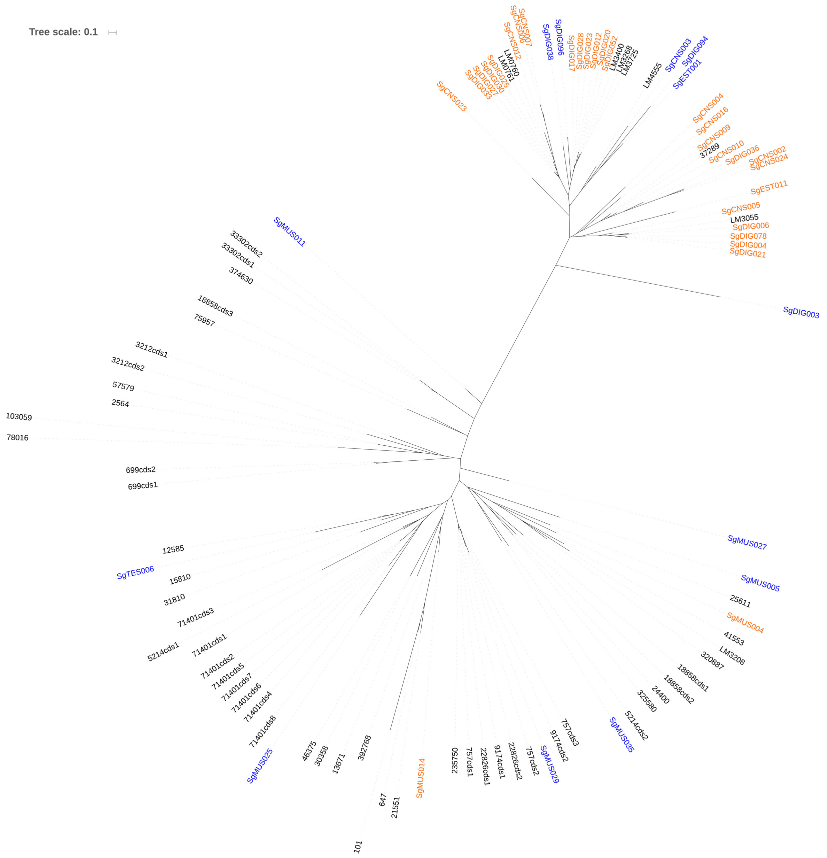

Figure 5.49: Maximum likelyhood phylogenetic tree of the nucleotide of exon 2 *L. migratoria* and *S. gregaria* CSPs. *L. migratoria* CSPs are in black and *S. gregaria* transcripts in orange. Incongruently placed *S. gregaria* exon sequences are in blue. The branch length scale is at the top of the trees. Only branch support values that are higher than 75 % are shown in their respective branches.

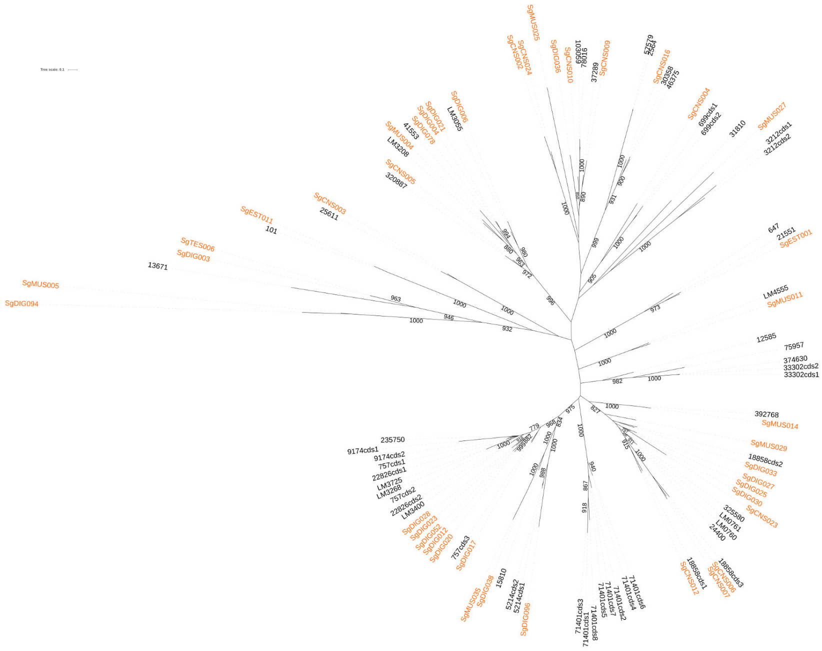

Figure 5.50: Maximum likelyhood phylogenetic tree of the nucleotide sequences of whole *L. migratoria* and *S. gregaria* CSPs. *L. migratoria* CSPs are in black and *S. gregaria* transcripts in orange. The branch length scale is at the top of the trees. Only branch support values that are higher than 75 % are shown in their respective branches.

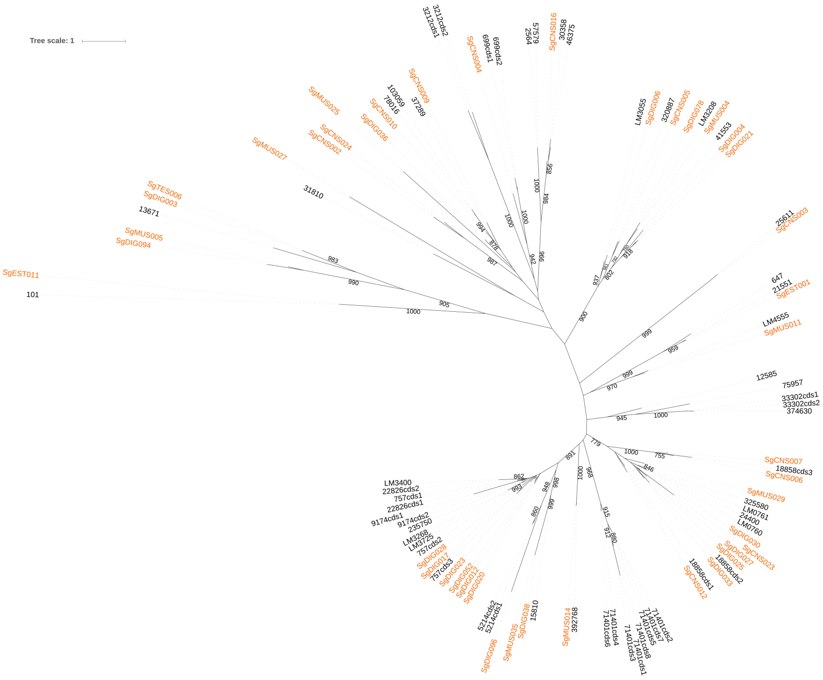

Figure 5.51: Maximum likelyhood phylogenetic tree of the amino acids sequences of whole *L. migratoria* and *S. gregaria* CSPs. *L. migratoria* CSPs are in black and *S. gregaria* transcripts in orange. The branch length scale is at the top of the trees. Only branch support values that are higher than 75 % are shown in their respective branches.

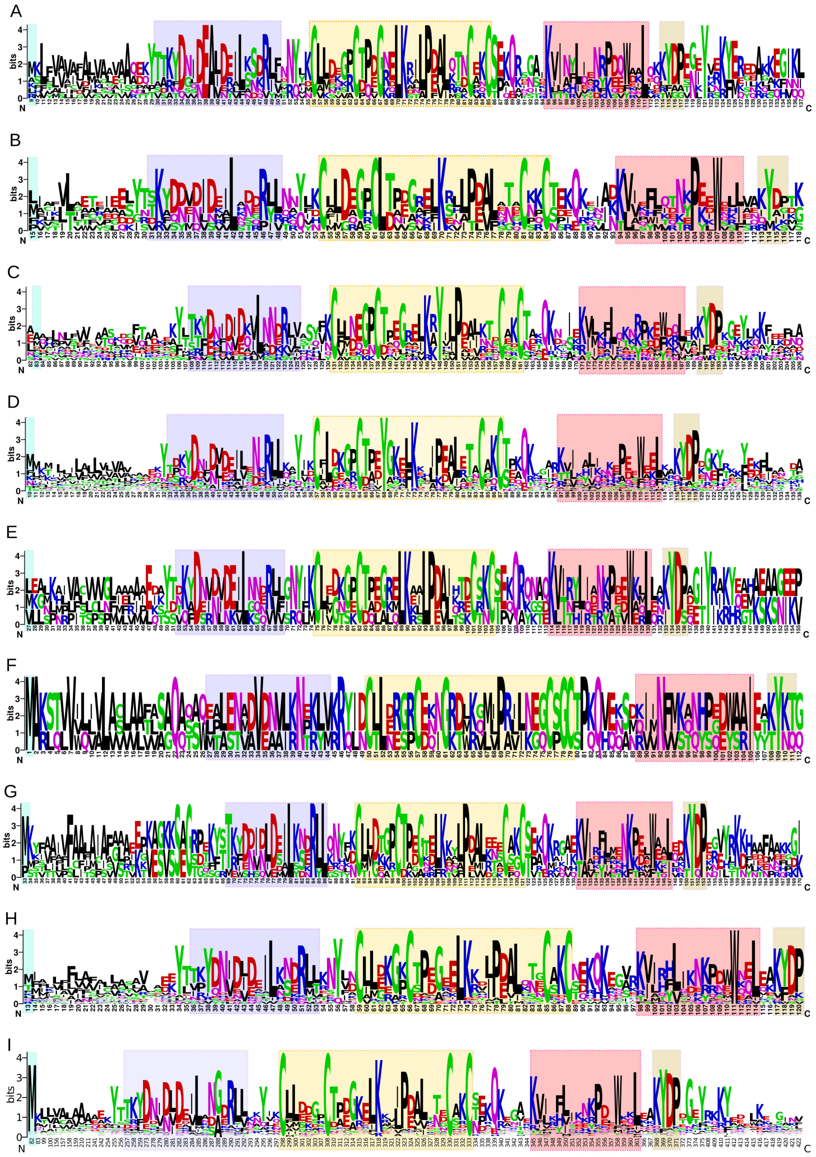

Figure 5.52: Sequence logo reflecting amino acid conservation along the positions of the alignment from *Anopheles gambiae* (A), *Apis mellifera* (B), *Acyrthosiphon pisum* (C), *Bombyx mori* (D), *Drosophila melanogaster* (E), *Daphnia pulex* (F), *Pediculus humanus* (G), *Tribolium castaneum* (H), and all these species plus *L. migratoria* and *S. gregaria* (I). The five conserved sections of the alignment (consensus initial metionine, signal peptide conserved region, cysteine box, leucine-isoleucine-valine-metionine conserved region and KYDP region) are boxed.

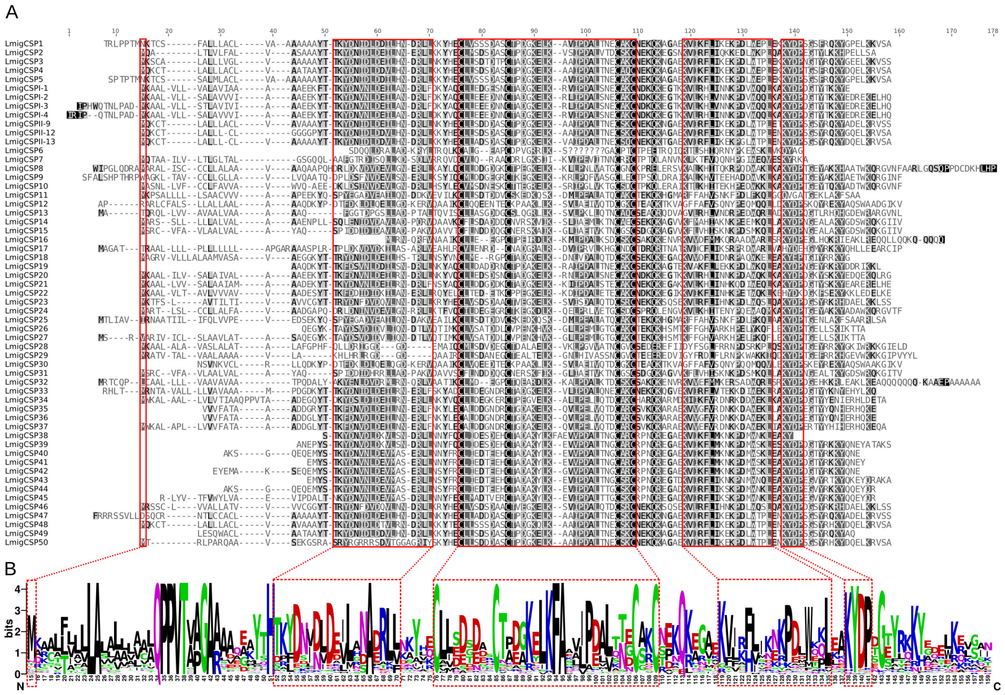

Figure 5.53: Alignment (A) and sequence logo (B) of *L. migratoria* CSP amino acid sequences. Similarity is proportionally represented in the alignment in greyscale hues, being white a non-conserved position and black a fully conserved position. The hight of the logo at each position reflects conservation of that position. Five conserved sections (consensus initial metionine, signal peptide conserved region, cysteine box, leucine-isoleucine-valine-metionine conserved region and KYDP region) are boxed in both figures.

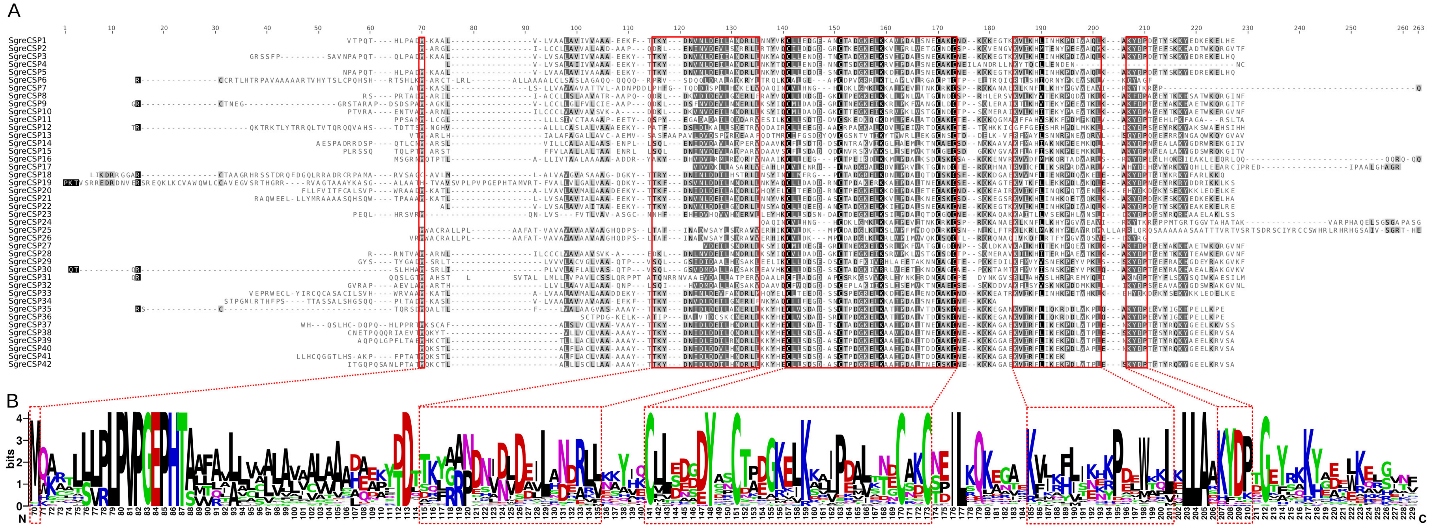

Figure 5.54: Alignment (A) and sequence logo (B) of *S. gregaria* CSP amino acid sequences. Similarity is proportionally represented in the alignment in greyscale hues, being white a non-conserved position and black a fully conserved position. The hight of the logo at each position reflects the conservation of that position. Five conserved sections (consensus initial metionine, signal peptide conserved region, cysteine box, leucine-isoleucine-valine-metionine conserved region and KYDP region) are boxed in both figures.

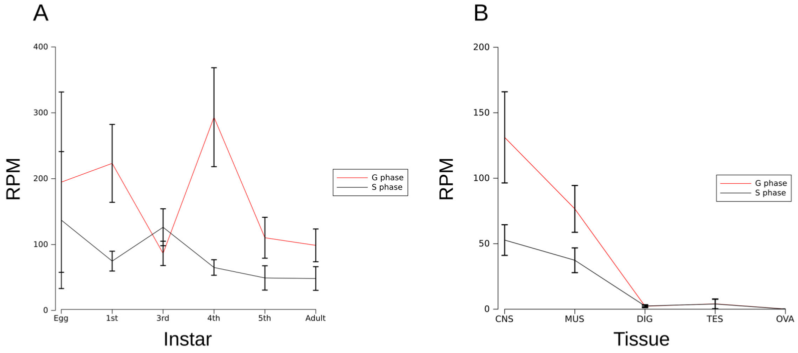

Figure 5.55: Distribution of the mean reads per million of total mapped reads (RPM) from all the CSPs identified in *L. migratoria* (A) and *S. gregaria* (B).

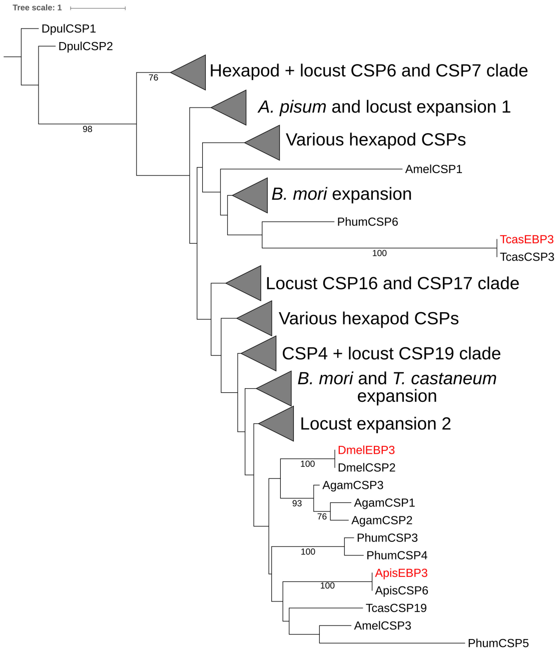

Figure 5.56: Maximum likelyhood phylogenetic tree of the amino acids sequences of locust CSPs and insect EBP3. Non-relevant branches were collapsed for clarity.

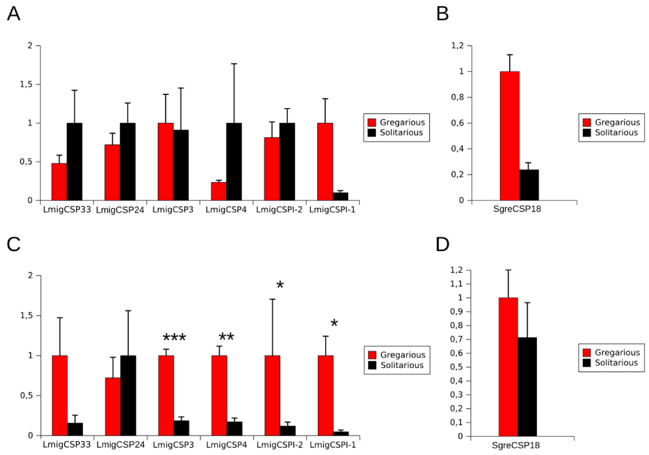

Figure 5.57: Standardized qPCR results of the expression of six *L. migratoria* and one *S. gregaria* CSP in adults (A and B) and numphs (C and D). The standardization consisted in dividing by the highest value of the expression levels, so the maximum value for each CSP sequence is 1. Black bars represent gregarious expression levels and grey bars represent solitarious expression values. t-tests were performed to obtain signification and, in case of significantly difference, asterisks are placed over the bars.

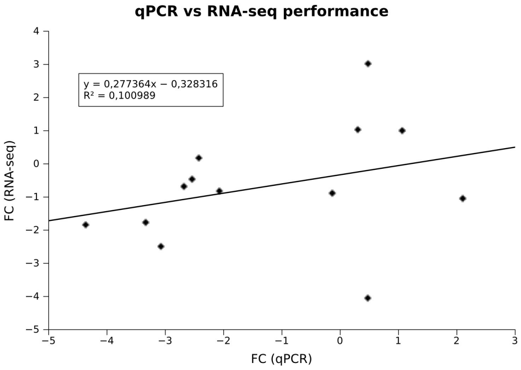

Figure 5.58: Comparison between the RNA-seq and qPCR results on the differential expression of seven CSPs between solitarious and gregarious *L. migratoria* adults (red) and nymphs (black). The fold change was calculated as detailed in general methodology. Pearson regression coefficients were $R^2 = 0.203$ for adults and $R^2 = 0.268$ for 4th instar nymph.

Discusión y perspectivas

Los resultados aportados por esta tesis representan un cambio de enfoque en el estudio del cambio de fase de *S. gregaria*, ya que en lugar de pretender localizar un disparador del cambio como se enfoca en la gran mayoría de trabajos sobre el fenómeno en esta especie, tratamos de visualizar las consecuencias moleculares del mismo a gran escala. Para ello adoptamos el análisis transcriptómico comparativo como estrategia. Procuramos, además, ser lógicos, metódicos y meticulosos para que el material sea no solamente adecuado para responder a nuestras preguntas sino que esté lo menos alterado y más representativo de la realidad como nos fue posible. Igualmente las técnicas y métodos usados fueron escogidos y aplicados tras un serio proceso de evaluación objetiva, optimización y diseño coherente y lógico. Hay que resaltar que desde que se planteó el proyecto que dio lugar a la presente tesis, las aportaciones sobre *L. migratoria* del productivo grupo de Dr. Le Kang en Pekín generaron una clara desigualdad entre el número de trabajos "ómicos" sobre esa especie y los hasta hace poco inexistentes sobre *S. gregaria.* La diferencia es tal que *L. migratoria* ya cuenta con un borrador de genoma de referencia ensamblado, además de varios estudios transcriptómicos (ver cuadro 1 de la Introducción de esta tesis). Nuestro trabajo de hecho es el único basado sobre secuenciación NGS en *S. gregaria* y sus datos permiten además recortar esta diferencia. Aunque *L. migratoria* tiene una distribución mundial más amplia, *S. gregaria* no deja de ser la especie de langosta que provoca más perjuicios a la agricultura (como se menciona en la introducción), y dada su importancia económica (sin restársela a *L. migratoria*), debería presentar al menos la misma cantidad de estudios a nivel "ómico" que *L. migratoria.* Por eso, esta discusión está enfocada tanto en los resultados de este trabajo han aportado como a los que se aportarán al campo del estudio del cambio de fase de las langostas en un futuro próximo.

Lo que aportamos ...

Para poder adaptar los resultados de los estudios “ómicos” al campo de la genómica funcional, primero necesitamos desarrollar una herramienta estandarizada y validada que cuantifique de forma precisa si los tratamientos aplicados a un grupo de langostas son efectivos a la hora de prevenir o activar el cambio de fase. Como hemos descrito en el **capítulo 1** de esta tesis, conseguimos modelos basados sobre el método tradicional de estudio en el cambio de fase (el índice P_{greg}) incluyendo un mayor número de variables morfológicas y comportamentales, para que los trabajos sean comparables entre estudios. Al introducir la novedad de mitigar el efecto del tamaño del animal pudimos, por primera vez, ofrecer una herramienta comprobada para su uso por distintos grupos de investigación. La herramienta que ofrecemos es de por sí fácil de usar, y al construir y ofrecer programas informáticos para su automatización logramos que sea al alcance de todo grupo que la pueda necesitar. A pesar de que el cambio de fase seguramente esté evolutivamente conservado dentro de la familia Acrididae, tuvimos que desarrollar modelos específicos para *S. gregaria* y para *L. migratoria*, lo cual aporta información sobre las diferencias inter-específicas en el cambio de fase. No obstante, mientras los modelos para *S. gregaria* han sido comprobados con mayor resolución dando lugar a predicciones más sólidas, los modelos que hemos construido para *L. migratoria* presentan peor capacidad de predicción por el mero hecho de haber utilizado menos poblaciones en estados de fase transitorios para validarlos. Con el tiempo, se debería actualizar este modelo para que los estudios en *L. migratoria* pudieran compararse no solo entre sí, sino con los de *S. gregaria.* Incluso es posible que al aplicar la metodología aportada en este capítulo, se puedan elaborar modelos para otras especies de langostas, como *Dociostaurus maroccanus*, *Calliptamus itallicus* o *Chortoicetes terminifera*, permitiendo así el estudio del comportamiento en otras especies que presenten cambio de fase.

Sin duda el **capítulo 2** fue la punta de lanza durante el desarrollo de esta tesis: no solamente sirvió para asentar los protocolos tanto moleculares como bioinformáticos a usar, sino que también supone el grueso de los datos más relevantes relacionados con el cambio de fase en *S. gregaria* aportados por esta tesis. Hay que hacer hincapié en que las poblaciones tanto gregaria como solitaria de *S. gregaria* que se utilizaron fueron representativas y que el manejo tanto de los individuos como del ARN extraído de ellos se llevó a cabo con todas las precauciones posibles para evitar tanto posibles cambios de expresión por estrés antes de la disección de tejidos como la degradación del material durante la extracción de ARN. A pesar de lo impactante que pueda parecer que más de veinte mil transcritos estén sobre-expresados en la fase gregaria, el polifenismo conlleva un cambio radical en la regulación genética de muchos procesos que intervienen tanto en su activación como en su mantenimiento, aunque éste pueda ser activado por factores puntuales (estimulación mecánica, visual, olfativa u hormonal). Teniendo en cuenta que estamos estudiando el proceso en adultos expuestos o no a estímulos gregarizantes, nos encontramos ante las consecuencias producidas por el cambio, por lo que es de esperar que una gran cantidad de genes estén afectados en un proceso tan complejo.

El estudio del transcriptoma del tubo digestivo del **capítulo 3** nos aporta información sobre los cambios de niveles de expresión de genes en otro tejido. Sirve asimismo también para comprobar la fiabilidad de los métodos y datos obtenidos. Además, sirve para obtener información comparativa a cerca de microorganismos que se encontraban en la flora intestinal de nuestros cultivos de langostas solitarias y gregarias. Como era de esperar, la expresión de genes para los microorganismos fue mayor en el tubo digestivo de los individuos gregarios, que presentaron más transcritos sobre-expresados pertenecientes a protozoos, hongos y bacterias. No obstante, siendo ese estudio transcriptómico y orientado a determinar diferencias entre tejidos de langostas solitarias y gregarias, los datos sobre diversidad microbiana que se presenta en el trabajo son aproximativos y necesitarían experimentos más precisos para poder validarlos. Esto se debe a que, entre otras cosas, nuestra estima de la abundancia se basó sobre etiquetas de los resultados BLAST y la abundancia de esas, lo que limita el estudio debido a la incapacidad de diferenciar tanto especies como su abundancia absoluta de forma precisa. Mediante un estudio metagenómico basado en secuenciación de amplicones de algún marcador filogenético (como el ARN ribosomal S16 o la citocromo C oxidasa I) se podría hacer una mejor estima de la diversidad de microorganismos y las diferencias de ésta entre los tubos digestivos de las dos fases. También queda por determinar si el protozoo detectado es realmente el apicomplejo *Gregarina niphandrodes*, ya que el que se ha descrito como parásito de cultivos de laboratorio es la especie *Gregarina garnhami*, de la cual no parece que existan secuencias en las bases de datos en línea. Si esto fuera así, tenemos cerca de dos mil secuencias probablemente pertenecientes a esta especie y además se podría investigar su potencial como posible herramienta de control biológico.

La información obtenida con los dos transcriptomas nos sirvió para poder localizar genes que compartieron el mismo patrón de expresión diferencial entre tejidos, expuestos en el **capítulo 4**. Aunque esperábamos más candidatos, la lista de genes con tendencias similares en los dos tejidos se acerca a los cien, comprendiendo genes con tendencia ya mencionada en la bibliografía (como el caso de la pacifastina 4, el inhibidor de serina proteasa 3 o la anexina IX) como otros genes que resaltamos por primera vez (como la fosfoenolpiruvato carboxikinasa). También logramos confirmar por homología la presencia de secuencias no anotadas en ambos tejidos, aunque el umbral de identidad utilizado es discutible, puesto que nos basamos en secuencias codificantes para obtenerlo, siendo el grado de conservación posiblemente distinto entre secuencias codificantes y otras secuencias reguladoras. La comparación de los resultados entre *S. gregaria* y *L. migratoria*, tanto experimental como bibliográfica, nos ha permitido confirmar genes con patrones de expresión similares en ambas especies, lo que permite inferir hasta qué punto está conservado evolutivamente el cambio de fase. La comparación de resultados de diversas fuentes logra consolidar los resultados de todos los trabajos involucrados, además enfocar futuros estudios hacia una lista más reducida de genes candidatos. Este estudio podría ser complementado por redes de co-expresión una vez que los datos del resto de tejidos estén analizados, e incluso añadir datos de tejidos de otras especies (como *L. migratoria*) para comprobar si las interacciones de genes se mantienen evolutivamente.

En el **capítulo 5** hacemos uso de la información generada por los transcriptomas de *S. gregaria* así como de información disponible para el genoma y transcriptomas de *L. migratoria* para identificar el número aproximado de una familia de proteínas en ambas especies. Aunque el número de CSPs presentado es elevado (57 y 42 para *L. migratoria* y *S. gregaria*, respectivamente), el tamaño del genoma de ambas especies es también gigantesco, lo que sugiere que puedan haber ocurrido eventos de duplicación que hayan derivado en la formación de nuevas copias y una posterior divergencia, como sugiere la proximidad con la que aparecen algunas de estas secuencias (incluso colocadas en tándem) en el genoma de *L. migratoria.* Las langostas parecen tener también mayor densidad de CSPs incluso que las hormigas—famosas por sus capacidades quimosensoriales. No obstante, hay que recordar que las secuencias han sido identificadas mediante herramientas bioinformáticas, y aunque logramos amplificar algunas secuencias mediante PCR e incluso cuantificarlas mediante qPCR, amplificar el resto de secuencias aportaría la mejor evidencia posible para ser confirmar esas secuencias que por el momento son putativas. Filogenéticamente, esta familia parece tener un comportamiento similar en taxones con un mayor número de CSPs que el ancestral, presentando expansiones específicas tanto a nivel de orden (ortópteros) como de especie. Incluyendo secuencias de otras especies de un mismo orden (como podrían haber sido las varias especies de hormiga con genoma secuenciado) podríamos comprobar si esta tendencia es algo normal. También es tranquilizador el hecho de que todas las CSPs homólogas entre especies presenten mayormente indicios de selección purificadora, manteniendo posiblemente la función específica de cada pareja de homólogos en ambas especies. También es importante la comparación de perfiles de expresión de las parejas homólogas, ya que siete de ellas presentaron un patrón de sobreexpresión hacia la fase gregaria, pudiendo estar potencialmente relacionadas con el cambio de fase. La confirmación definitiva de esta posible relación necesitaría experimentos funcionales que demuestren la función concreta de cada pareja de homólogos.

. . . y lo que aportaremos

El grupo de investigación, gracias al proyecto y a la línea de investigación en el que se ha desarrollado esta tesis, está generando muchos más datos que no forman parte de esta memoria. En consecuencia, se han abierto muchos nuevos interrogantes que quedan por resolver. Es cierto que para comprender la base molecular del cambio de fase el conjunto de diferencias entre secuencias codificantes es muy importante, pero seguramente las diferencias vayan un paso más, implicando a otros elementos de regulación como pueden ser la epigenética o los ARNs reguladores. No obstante, tener una gran cantidad de secuencias de referencia, validadas gracias a la comparación entre tejidos e incluso con otros trabajos ofrece un apoyo sólido a todo el campo del estudio molecular de las plagas de langostas.

Los transcriptomas analizados en esta tesis nos han dejado un alto número de secuencias sin anotación conocida, y un gran parte de ellas están presentes en ambos transcriptomas (o al menos presentan homología entre ellas). Algunas de ellas presentan marcos abiertos de lectura que pueden corresponder a secuencias codificantes todavía no caracterizadas. Otras de ellas pueden ser ARNs no codificantes de tamaño grande, lo que implicaría abrir un nuevo frente de investigación en el cambio de fase para descubrir qué papel tienen y cómo se comportan los ARNs reguladores durante este proceso. El grupo de investigación pretende confeccionar una base de datos de este y otros tipos de moléculas que podría servir para comprobar si son reales y para aportar más secuencias de este tipo a la comunidad científica para poder reforzar los algoritmos de predicción de ARNs no codificantes. Además, el grupo tiene proyectada la secuenciación de ARNs pequeños y micro-ARNs (siRNAs y miRNAs), dilatando aún más si cabe este campo.

Por supuesto, estos experimentos tienen componentes muy teóricos, pero también pueden enfocarse estudios de corte más aplicado. Gracias al ARN interferente, el grupo tiene proyectado provocar el silenciamiento de la expresión de genes con secuencia conocida y observar qué fenotipo resultante provoca. En dicha técnica el investigador principal del grupo tiene experiencia ya documentada [Cabrero et al., 2013, Ruiz-Estévez et al., 2014]. Además, con la herramienta de cálculo de P_{greg} que presentamos en esta tesis, se puede estandarizar el estudio de cambios de fenotipo a nivel de morfología y comportamiento para determinar si el tratamiento afecta al cambio de fase. Así pues, uno de los objetivos futuros del grupo consiste en el silenciamiento sistemático de genes que presenten patrones de expresión relacionados con la fase, analizando su fenotipo mediante el cálculo de P_{greg} para comprobar si hay efecto en el estadote fase. Recientemente también se ha logrado poner en funcionamiento la técnica que usa CRISPR/Cas9 para la edición in vivo de ADN genómico en *L. migratoria*, [Li et al., 2016] lo cual puede extender también su aplicación a *S. gregaria*, pese a no tener una referencia genómica ensamblada aún.

En síntesis, los resultados de esta tesis no solo asientan protocolos y herramientas para estudiar el cambio de fase en langostas, sino que ofrecen extensas listas de genes con los que realizar experimentos e identificar su papel en el cambio de fase. También ha servido para reforzar otros resultados que indican la implicación de grandes diferencias de expresión genética entre los estados o fases de los polifenismos en general (al igual que ocurre con otros polifenismos, como la determinación de castas en himenópteros eusociales). Finalmente, genera muchas más preguntas que van más allá de las diferencias de expresión y comienzan a adentrarse en las funciones propias de genes identificados como potencialmente implicados en el cambio de fase, con el consiguiente enfoque hacia la genómica funcional.

Conclusiones

1. Mientras que el sexo del individuo no pareció tener efecto significativo sobre las variables asociadas al cambio de fase, el tamaño sí afecta a las variables comportamentales asociadas con la velocidad de movimiento. Normalizar estas variables con el tamaño del individuo permitió homogeneizar la variabilidad y hacer comparables los datos extraídos a partir de animales de distinto tamaño.

2. Ofrecemos dos modelos comportamentales para *S. gregaria*: un modelo de trece variables (incluyendo variables morfométricas) para estudiar muestras distintas o para estudiar el efecto de un factor en el cambio de fase entre mudas de ninfas; y un modelo de diez variables (solo de comportamiento) recomendado para medir el efecto de un factor en el cambio de fase de una misma muestra de individuos (adultos o ninfas que no hayan mudado).

3. Las variables morfométricas como las de comportamiento presentan valores específicos de especie. Por lo tanto, no es posible utilizar modelos basados en una especie para predecir la probabilidad de ser gregario de otra. Debido a esta diferencia, se confeccionaron modelos específicos para *L. migratoria*. No obstante, los modelos conseguidos para esa última especie no fueron tan precisos como los confeccionados para *S. gregaria* debido a un tamaño de muestra bajo, por lo que aún quedan por refinar.

4. El sistema nervioso central de *S. gregaria* presenta una proporción elevada de transcritos diferencialmente expresados entre fases (cerca del 40 %), el 90 % de los cuales están sobre-expresados en la fase gregaria. Estoindica que el estado alterado y excitado de los individuos gregarios se asocia a una mayor actividad de la expresión génica.

5. Reconstruimos con detalle una cascada de eventos relacionados con la formación de la plaga, basándonos en la interpretación de los cambios en la expresión de genes entre el sistema nervioso central de individuos gregarios y solitarios.

6. Entre los transcritos y rutas génicas sobre-representadas en el sistema nervioso central de la fase gregaria destacamos los relacionados con la estructura neuronal, la señalización, los neurotransmisores y las catecolaminas. Igualmente destacamos los genes asociados con el sistema inmune, el estrés, el metabolismo, la detoxificación, la muerte celular y la expresión génica. La sobre-expresión de genes de esta última categoría es inevitable debido al alto número de genes afectados por el cambio.

7. El tubo digestivo de *S. gregaria* presenta apenas un 10 % de transcritos diferencialmente expresados, con casi tantos transcritos sobre-expresados tanto en fase gregaria como solitaria. Esto indica que este tejido está menos afectado por el cambio de fase si lo comparamos con el sistema nervioso central.

8. En el tubo digestivo de *S. gregaria* los transcritos de protozoos, hongos y bacterias presentan mayor abundancia y profundidad de secuenciación en la fase gregaria, lo que indica una mayor presencia de microorganismos en esta fase. *Gregarina niphandrodes* (Apicomplexa, Eugregarinida) presentó el mayor número de transcritos sobre-expresados pertenecientes a un microorganismo en el tubo digestivo en fase gregaria.

9. Debido a la mayor presencia de parásitos y al mayor grado de estrés presente en la fase gregaria, los procesos biológicos más representados en el tubo digestivo gregario están relacionados con la respuesta inmune, la respuesta al estrés y la apoptosis. Por el contrario, en el tubo digestivo solitario se sobre-expresan principalmente genes relacionados con la señalización para la hormona juvenil, con el mantenimiento de fibras musculares y con componentes de la matriz peritrófica. Esto nos indica que se invierte más en paliar el estrés y controlar a los patógenos en la fase gregaria y que en la fase solitaria se invierte más en funciones constitutivas. La mayoría de transcritos de enzimas digestivas no presentaron diferencias significativas entre fases, por lo que la función digestiva no parece afectada por el cambio de fase en nuestra colonia de langostas.

10. La cantidad de transcritos con resultado BLAST compartido entre los transcriptomas del sistema nervioso y el tubo digestivo (2.772) fue mucho menor que lo esperado debido tanto a las diferencias de expresión génica entre tejidos como a la asignación de mejores resultados BLAST pertenecientes a distintas especies para un mismo gen presente en ambos trancriptomas.

11. Una gran cantidad de secuencias no anotadas está presente en ambos transcriptomas. Éstas podrían ser pertenecientes a genes codificantes aún no descritos o a ARNs no codificantes. También hay secuencias no anotadas específicas de cada transcriptoma. En ambos casos hay secuencias con altos niveles de expresión, lo que apoya que sean secuencias reales.

12. Dadas las diferencias en el perfil de expresión génica entre transcriptomas y las limitaciones de las comparaciones basadas sobre análisis BLAST, el número de transcritos diferencialmente expresados en la misma fase en ambos transcriptomas fue muy bajo (menos de un centenar).

13. La concordancia entre nuestros datos y los descritos en la bibliografía es baja debido a las diferencias entre las técnicas, especies, estadíos de desarrollo y tejidos utilizados por esos distintos trabajos de investigación, incluido el nuestro. Aun así, pudimos confirmar la asociación de algunos genes con el cambio de fase en *S. gregaria* y en otras langostas.

14. Existen más de medio centenar de copias de genes para proteínas quimosensoras (CSPs) en el genoma de *L. migratoria*. Confirmamos la expresión de cerca de cincuenta de ellas en esta especie, y el número de CSPs no redundantes ensambladas en *S. gregaria* es también cercano a cincuenta. Encontramos que las duplicaciones génicas han contribuído al alto número de CSPs presentes en langostas, que es elevado comparado con otras especies.

15. La filogenia de las CSPs indica la presencia de siete clados específicos de ortópteros, siendo dos de ellos expansiones que engloban la inmensa mayoría de las secuencias. *S. gregaria* y *L. migratoria* presentan 21 parejas de CSPs homólogas entre las cuales predomina la selección purificadora. Sin embargo, la selección neutral parece ser la tendencia presente entre las secuencias homólogas intra-específicas.

16. Existen diferencias en la expresión de CSPs entre especies, fases, estadios y tejidos, siendo el cuarto estadío ninfal de *L. migratoria* y el tejido nervioso de *S. gregaria*, ambos en fase gregaria, los que más expresión de estas proteínas muestran.

17. Entre las CSPs diferencialmente expresadas, detectamos siete parejas de homólogos inter-específicos que comparten un patrón de expresión gregario en *L. migratoria* y *S. gregaria*. Esto indica que su función en el cambio de fase podría estar conservada evolutivamente.

Conclusions

1. While sex did not have significant effect on the variables associated with phase change, size did affect the movement-related behavioural variables. Normalization of these variables by size reduced the heterogeinity between measurements and allowed comparing results from animals of different size.

2. We offer two behavioural models for *S. gregaria*: One with thirteen variables (including morphometric) to perfomr comparative analyses on different samples or detecting the effect of a factor on phase change between molts of the same nymphs; the other with ten variables (only behavioural) is for detecting the effect of a factor on phase change of the same sample of adults or nymphs that have not molted during the experiment.

3. The morphometric and behavioural variables take species-specific values. Thus, prediction of the probability of being gregarious cannot be based on models that were developed for other species. We hence elaborated models that are specific to *L. migratoria.* However, these models were not as accurate as the ones that we developed for *S. gregaria*, due to small sample size, and better ones are still to be built and validated for that species.

4. *S. gregaria*'s central nervous system presents a high proportion of differentially expressed transcripts between phases (circa 40 %). Most of them (90 %) were over-expressed in gregarious phase. The excited and altered gregarious state is therefore associated with higher levels of gene expression.

5. We elaborated a detailed cascade of events related to plague formation based on our interpretation of the changes in the levels of gene expression between the gregarious and solitarious central nervous systems.

6. Among the transcripts and genetic pathways that are up-regulated in the gregarious phase, we highlight the ones involved in neural structure, signalling, neurotransmitters and catecholamines. In this category are also genes related to the immune system, stress response, metabolism, detoxification, cell death and gene expression. Over-expression of the latter is unavoidably necessary given the high number of genes associated to gregarious phase.

7. Only 10% of the transcripts assembled for the transcriptome of *S. gregaria*'s digestive tube were differentially expressed, with almost equal number in gregarious and solitarious phase. This tissue is therefore less affected by the phase change compared to the central nervous system.

8. Transcripts from protozoa, fungi and bacteria are more abundant and show higher sequencing coverage in the digestive tube of the gregarious phase, indicating a higher presence of microorganisms in that phase. The highest number of over-expressed transcripts belonging to a single microorganisms in the digestive tube of gregarious *S. gregaria* belonged to *Gregarina niphandordes* (Apicomplexa, Eugregarinida). This gregarine species might thus offer an additional option for biological control of the locusts.

9. Due to the higher presence of parasites and stressful stimuli, the most represented biological processes in the digestive tube of the gregarious phase are related to immune response, stress response and apoptosis. In contrast, most of the over-expressed transcripts in the digestive tube during the solitarious phase are related to juvenille hormone signalling, muscular fiber maintenance and peritrophic matrix components. This indicates higher investment in stress and pathogen control during the gregarious phase and in constitutive functions during the solitarious phase. Most of the transcripts relating to digestive functions were not differentially expressed, meaning that digestion is not affected by the phase change in our locust colony.

10. The amount of transcripts sharing a BLAST result between the transcriptomes from the central nervous system and the digestive tube was low (2,772). This is due to the inherent gene expression differences between tissues and the frequent assignation of best BLAST results from different species to the same sequence in different transcriptomes.

11. Both transcriptomes contain large numbers of unannotated sequences, some of which shared and others seem tissue-specific. The high coverage

of many of these unannotated sequences, in both cases, supports their genuine nature. They might belong to unknown coding sequences or to non-coding RNAs.

12. Given the differences in gene expression profiles between transcriptomes and the limitations the BLAST-based comparisons, the number of transcripts differentially expressed towards the same phase in both transcriptomes was low. Still, we report tens of genes whose expression is congruently altered in both transcriptomes.

13. The congruency between our data and those reported in previous works is lffected by the differences of techniques, species, developmental stages, and tissues used in each work. Still, we could confirm the association of several genes with the phase change either in *S. gregaria* only or locusts in general.

14. There are about sixty CSP gene copies in *L. migratoria*'s genome. The expression of about fifty of these is confirmed and the number of non-redundant CSP transcripts that we assembled for *S. gregaria* is also close to fifty. Gene duplications were found to be the contributors to the high number of CSPs in locusts compared to other species.

15. CSP phylogeny shows seven Orthoptera-specific clades, two of which being expansions that include the majority of the sequences. *S. gregaria* and *L. migratoria* share 21 orthologous CSP pairs, and while these orthologs seem to be under purifying selection, the homologous sequences within each genome seem to be under neutral selection.

16. In addition to the gene-dependent differences, there are differences in CSP expression both at the species, phase, developmental stage and tissue levels, being *L. migratoria*'s fourth nymphal stage, *S. gregaria*'s central nervous system and the gregarious phase the conditions with the highest CSP expression.

17. Among the CSPs that show significant differential expression between phases in each locust species, seven orthologous pairs share a gregarious over-expression pattern both in *S. gregaria* and *L. migratoria.* Their involvement in the development and/or maintenance of gregariousness is hence conserved in locusts.

Bibliografía

Federico Abascal, Rafael Zardoya, and David Posada. Prottest: selection of best-fit models of protein evolution. *Bioinformatics*, 21(9):2104–2105, 2005.

Edward P Abraham and Ernst Chain. An enzyme from bacteria able to destroy penicillin. *Nature*, 146(3713):837, 1940.

SA Adamo, JL Roberts, RH Easy, and NW Ross. Competition between immune function and lipid transport for the protein apolipophorin iii leads to stress-induced immunosuppression in crickets. *Journal of Experimental Biology*, 211(4): 531–538, 2008.

Shelley Anne Adamo. Stress responses sculpt the insect immune system, optimizing defense in an ever-changing world. *Developmental & Comparative Immunology*, 2016.

Katayoun Afshar, Pierre Gönczy, Stephen DiNardo, and Steven A Wasserman. fumble encodes a pantothenate kinase homolog required for proper mitosis and meiosis in *Drosophila melanogaster. Genetics*, 157(3):1267–1276, 2001.

Bogos Agianian, Paul A Tucker, Arie Schouten, Kevin Leonard, Belinda Bullard, and Piet Gros. Structure of a drosophila sigma class glutathione s-transferase reveals a novel active site topography suited for lipid peroxidation products. *Journal of molecular biology*, 326(1):151–165, 2003.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215 (3):403–410, 1990.

Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

Harindra E Amarasinghe, Eamonn B Mallon, and Swidbert R Ott. Socially induced behavioural plasticity precedes pronounced epigenetic differentiation in the cns of desert locusts. *bioRxiv*, page 018499, 2015.

SA Ament, RA Velarde, MH Kolodkin, D Moyse, and GE Robinson. Neuropeptide y-like signalling and nutritionally mediated gene expression and behaviour in the honey bee. *Insect molecular biology*, 20(3):335–345, 2011.

Simon Anders. Htseq: Analysing high-throughput sequencing data with python. *URL http://www-huber. embl. de/users/anders/HTSeq/doc/overview. html*, 2010.

Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, page btu638, 2014.

Birgit Andersen, Stina Lundgren, Doreen Dobritzsch, and Jure Piškur. A recruited protease is involved in catabolism of pyrimidines. *Journal of molecular biology*, 379(2):243–250, 2008.

Leah J Anderson, Kai Lin, Teresa Compton, and Brigitte Wiedmann. Inhibition of cyclophilins alters lipid trafficking and blocks hepatitis c virus secretion. *Virology journal*, 8(1):1, 2011.

S Andrews. Fastqc: A quality control tool for high throughput sequence data. *Reference Source*, 2010.

Sergio Angeli, Francesca Ceron, Andrea Scaloni, Maria Monti, Gaia Monteforti, Antonio Minnocci, Ruggero Petacchi, and Paolo Pelosi. Purification, structural characterization, cloning and immunocytochemical localization of chemoreception proteins from *Schistocerca gregaria*. *European Journal of Biochemistry*, 262(3): 745–754, 1999.

Michael L Anstey, Stephen M Rogers, Swidbert R Ott, Malcolm Burrows, and Stephen J Simpson. Serotonin mediates behavioral gregarization underlying swarm formation in desert locusts. *science*, 323 (5914):627–630, 2009.

L Aravind. The wwe domain: a common interaction module in protein ubiquitination and adp ribosylation. *Trends in biochemical sciences*, 26(5):273–275, 2001.

Elias SJ Arnér and Arne Holmgren. Physiological functions of thioredoxin and thioredoxin reductase. *European Journal of Biochemistry*, 267(20):6102–6109, 2000.

Anna Aspán, Martin Hall, and Kenneth Söderhäll. The effect of endogeneous proteinase inhibitors on the prophenoloxidase activating enzyme, a serine proteinase from crayfish haemocytes. *Insect Biochemistry*, 20(5):485–492, 1990.

A Ayali and MP Pener. Density-dependent phase polymorphism affects response to adipokinetic hormone in *Locusta*. *Comparative Biochemistry and Physiology Part A: Physiology*, 101(3):549–552, 1992.

Liesbeth Badisco, Jurgen Huybrechts, Gert Simonet, Heleen Verlinden, Elisabeth Marchal, Roger Huybrechts, Liliane Schoofs, Arnold De Loof, and Jozef Vanden Broeck. Transcriptome analysis of the desert locust central nervous system: production and annotation of a *Schistocerca gregaria* est database. *PloS one*, 2011a.

Liesbeth Badisco, Swidbert R Ott, Stephen M Rogers, Thomas Matheson, Dries Knapen, Lucia Vergauwen, Heleen Verlinden, Elisabeth Marchal, Matt RJ Sheehy, and Malcolm Burrows. Microarray-based transcriptomic analysis of differences between long-term gregarious and solitarious desert locusts. *PloS one*, 6(11):e28110, 2011b.

Chang Bai, Partha Sen, Kay Hofmann, Lei Ma, Mark Goebl, J Wade Harper, and Stephen J Elledge. Skp1 connects cell cycle regulators to the ubiquitin proteolysis machinery through a novel motif, the f-box. *Cell*, 86(2):263–274, 1996.

Mohammed Bakkali. Microevolution of cis-regulatory elements: an example from the pair-rule segmentation gene fushi tarazu in the *Drosophila melanogaster* subgroup. *PloS one*, 6(11):e27376, 2011.

Mohammed Bakkali. A bird's-eye view on the modern genetics workflow and its potential applicability to the locust problem. *Comptes rendus biologies*, 336(8): 375–383, 2013.

A Bocar Bal and Sidi Mohamed Sidati. Réduction des doses efficaces d'insecticides contre les larves de criquet pèlerin (*Schistocerca gregaria* forskål, 1775: Orthoptera, acrididae) par utilisation de quantités réduites de phénylacétonitrile. *Biotechnologie, Agronomie, Société et Environnement*, 17(4):572–579, 2013.

L Ban, A Scaloni, A Brandazza, S Angeli, L Zhang, Y Yan, and Paolo Pelosi. Chemosensory proteins of *Locusta migratoria*. *Insect molecular biology*, 12(2):125–134, 2003.

Liping Ban, Long Zhang, Yuhua Yan, and Paolo Pelosi. Binding properties of a locust's chemosensory protein. *Biochemical and biophysical research communications*, 293(1):50–54, 2002.

Raymond V Barbehenn. Roles of peritrophic membranes in protecting herbivorous insects from ingested plant allelochemicals. *Archives of insect biochemistry and physiology*, 47(2):86–99, 2001.

Andrew I Barnes and Michael T Siva-Jothy. Density–dependent prophylaxis in the mealworm beetle *Tenebrio molitor* l.(coleoptera: Tenebrionidae): cuticular melanization is an indicator of investment in immunity. *Proceedings of the Royal Society of London B: Biological Sciences*, 267(1439):177–182, 2000.

Sepideh Bazazi, Jerome Buhl, Joseph J Hale, Michael L Anstey, Gregory A Sword, Stephen J Simpson, and Iain D Couzin. Collective motion and cannibalism in locust migratory bands. *Current Biology*, 18(10):735–739, 2008.

Sepideh Bazazi, Frederic Bartumeus, Joseph J Hale, and Iain D Couzin. Intermittent motion in desert locusts: behavioural complexity in simple environments. *PLoS Comput Biol*, 8(5):e1002498, 2012.

Spencer T Behmer, Corlisa E Belt, and Martin S Shapiro. Variable rewards and discrimination ability in an insect herbivore: what and how does a hungry locust learn? *Journal of Experimental Biology*, 208(18): 3463–3473, 2005.

Jessica K Bell, Gregory ED Mullen, Cynthia A Leifer, Alessandra Mazzoni, David R Davies, and David M Segal. Leucine-rich repeats and pathogen recognition in toll-like receptors. *Trends in immunology*, 24(10):528–533, 2003.

Y Ben-Shahar, H-T Leung, WL Pak, MB Sokolowski, and GE Robinson. cgmp-dependent changes in phototaxis: a possible role for the foraging gene in honey bee division of labor. *Journal of Experimental Biology*, 206(14):2507–2515, 2003.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

Ronald Bentley and R Meganathan. Biosynthesis of vitamin k (menaquinone) in bacteria. *Microbiological reviews*, 46 (3):241, 1982.

J Benz and A Hofmann. Annexins: from structure to function. *Biological chemistry*, 378(3-4):177–183, 1996.

EA Bernays. A specialized region of the gastric caeca in the locust, *Schistocerca gregaria*. *Physiological Entomology*, 6(1): 1–6, 1981.

EA Bernays. The insect on the plant—a closer look. In *Proceedings 5th international symposium on insect-plant relationships. Pudoc, Wageningen*, pages 3–17, 1982.

EA Bernays and DJ Chamberlain. A study of tolerance of ingested tannin in *Schistocerca gregaria*. *Journal of Insect Physiology*, 26(6):415–420, 1980.

EA Bernays and AC Lewis. The effect of wilting on palatability of plants to *Schistocerca gregaria*, the desert locust. *Oecologia*, 70(1):132–135, 1986.

Elizabeth A Bernays and Reginald F Chapman. Plant secondary compounds and grasshoppers: beyond plant defenses. *Journal of Chemical Ecology*, 26(8):1773–1794, 2000.

Inanç Birol, Shaun D Jackman, Cydney B Nielsen, Jenny Q Qian, Richard Varhol, Greg Stazyk, Ryan D Morin, Yongjun Zhao, Martin Hirst, and Jacqueline E Schein. De novo transcriptome assembly with abyss. *Bioinformatics*, 25(21):2872–2877, 2009.

Bart Boerjan, Filip Sas, Ulrich R Ernst, Julie Tobback, Filip Lemière, Michiel B Vandegehuchte, Colin R Janssen, Liesbeth Badisco, Elisabeth Marchal, and Heleen Verlinden. Locust phase polyphenism: Does epigenetic precede endocrine regulation? *General and comparative endocrinology*, 173(1):120–128, 2011.

Ykelien L Boersma, Janet Newman, Timothy E Adams, Nathan Cowieson, Guy Krippner, Kiymet Bozaoglu, and Thomas S Peat. The structure of vanin 1: a key enzyme linking metabolic disease and inflammation. *Acta Crystallographica Section D: Biological Crystallography*, 70(12):3320–3329, 2014.

Jonathan Bohbot, Franck Sobrio, Philippe Lucas, and Patricia Nagnan-Le Meillour. Functional characterization of a new class of odorant-binding proteins in the moth-mamestra brassicae. *Biochemical and* *biophysical research communications*, 253 (2):489–494, 1998.

J Borycz, JA Borycz, A Kubow, V Lloyd, and IA Meinertzhagen. Drosophila abc transporter mutants white, brown and scarlet have altered contents and distribution of biogenic amines in the brain. *Journal of Experimental Biology*, 211(21): 3454–3466, 2008.

Jose A Botella, Julia K Ulschmid, Christoph Gruenewald, Christoph Moehle, Doris Kretzschmar, Katja Becker, and Stephan Schneuwly. The drosophila carbonyl reductase sniffer prevents oxidative stress-induced neurodegeneration. *Current biology*, 14(9):782–786, 2004.

Abdelghani Bouaichi, Peter Roessingh, and Stephen J Simpson. An analysis of the behavioural effects of crowding and re-isolation on solitary-reared adult desert locusts (*Schistocerca gregaria*) and their offspring. *Physiological Entomology*, 20 (3):199–208, 1995.

Loïc Briand, Nicharat Swasdipan, Claude Nespoulous, Valérie Bézirard, Florence Blon, Jean-Claude Huet Huet, Paul Ebert, and Jean-Claude Pernollet. Characterization of a chemosensory protein (asp3c) from honeybee (*Apis mellifera* l.) as a brood pheromone carrier. *European Journal of Biochemistry*, 269(18):4586–4596, 2002.

Jozef Vanden Broeck, Liliane Schoofs, Robert Huybrechts, and Arnold De Loof. http://titan.biotec.uiuc.edu/locust/, 2005.

James B Brown, Nathan Boley, Robert Eisman, Gemma E May, Marcus H Stoiber, Michael O Duff, Ben W Booth, Jiayu Wen, Soo Park, and Ana Maria Suzuki. Diversity and dynamics of the drosophila transcriptome. *Nature*, 2014.

Jerome Buhl, David JT Sumpter, Iain D Couzin, Joe J Hale, Emma Despland, ER Miller, and Steve J Simpson. From disorder to order in marching locusts. *Science*, 312(5778):1402–1406, 2006.

Geert Bultynck, Santeri Kiviluoto, Nadine Henke, Hristina Ivanova, Lars Schneider, Volodymyr Rybalchenko, Tomas Luyten, Koen Nuyts, Wim De Borggraeve, and Ilya Bezprozvanny. The c terminus of bax inhibitor-1 forms a ca2+-permeable channel pore. *Journal of Biological Chemistry*, 287(4):2544–2557, 2012.

THORSTEN Burmester. Evolution and function of the insect hexamerins. *European Journal of Entomology*, 96:213–226, 1999.

Josefa Cabrero, Mohammed Bakkali, Beatriz Navarro-Domínguez, Francisco J Ruíz-Ruano, Rubén Martín-Blázquez, María Dolores López-León, and Juan Pedro M Camacho. The ku70 dna-repair protein is involved in centromere function in a grasshopper species. *Chromosome research*, 21(4):393–406, 2013.

JP M Camacho, FJ Ruiz-Ruano, R Martín-Blázquez, MD López-León, J Cabrero, P Lorite, DC Cabral-de Mello, and M Bakkali. A step to the gigantic genome of the desert locust: chromosome sizes and repeated dnas. *Chromosoma*, 124(2): 263–275, 2015.

Valérie Campanacci, Audrey Lartigue, B Martin Hällberg, T Alwyn Jones, Marie-Therèse Giudici-Orticoni, Mariella Tegoni, and Christian Cambillau. Moth chemosensory protein exhibits drastic conformational changes and cooperativity on ligand binding. *Proceedings of the National Academy of Sciences*, 100(9): 5069–5074, 2003.

Elizabeth U Canning. A new eugregarine of locusts, *Gregarina garnhami* n. sp., parasitic in *Schistocerca gregaria* forsk. *The Journal of Protozoology*, 3(2):50–62, 1956.

Arianne J Cease, James J Elser, Eli P Fenichel, Joleen C Hadrich, Jon F Harrison, and Brian E Robinson. Living with locusts: Connecting soil nitrogen, locust outbreaks, livelihoods, and livestock markets. *BioScience*, 65(6):551–558, 2015.

R Cerritos. Insects as food: an ecological, social and economical approach. *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources*, 4(027):1–10, 2009.

Bing Chen, Shaoqin Li, Qiang Ren, Xiwen Tong, Xia Zhang, and Le Kang. Paternal epigenetic effects of population density on locust phase-related characteristics associated with heat-shock protein expression. *Molecular ecology*, 24(4):851–862, 2015.

Shuang Chen, Pengcheng Yang, Feng Jiang, Yuanyuan Wei, Zongyuan Ma, and Le Kang. De novo analysis of transcriptome dynamics in the migratory locust during the development of phase traits. *PloS one*, 5(12):e15633, 2010.

TH Chen, JR Brody, FE Romantsev, JG Yu, AE Kayler, E Voneiff, FP Kuhajda, and GR Pasternack. Structure of pp32, an acidic nuclear protein which inhibits oncogene-induced formation of transformed foci. *Molecular biology of the cell*, 7(12):2045–2056, 1996.

Thomas T Chen, Pierre Couble, Randa Abu-Hakima, and Gerard R Wyatt. Juvenile hormone-controlled vitellogenin synthesis in *Locusta migratoria* fat body: Hormonal induction in vivo. *Developmental biology*, 69(1):59–72, 1979.

Li E Cheng, Wei Song, Loren L Looger, Lily Yeh Jan, and Yuh Nung Jan. The role of the trp channel nompc in *Drosophila* larval and adult locomotion. *Neuron*, 67 (3):373–380, 2010.

Xavier Cheseto, Serge Philibert Kuate, David P Tchouassi, Mary Ndung'u, Peter EA Teal, and Baldwyn Torto. Potential of the desert locust schistocerca gregaria (orthoptera: Acrididae) as an unconventional source of dietary and therapeutic sterols. *PloS one*, 10(5):e0127171, 2015.

Kyung Tae Chung and Donald D Ourth. Purification and characterization of apolipophorin iii from immune hemolymph of heliothis virescens pupae. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 132 (2):505–514, 2002.

Sory Cisse, Saïd Ghaout, Ahmed Mazih, Ould Babah Ebbe, Mohamed Abdallahi, and Cyril Piou. Estimation of density threshold of gregarization of desert locust hoppers from field sampling in mauritania. *Entomologia Experimentalis et Applicata*, 156(2):136–148, 2015.

Ilse Claeys, Bert Breugelmans, Gert Simonet, Sofie Van Soest, Filip Sas, Arnold De Loof, and Jozef Vanden Broeck. Neuroparsin transcripts as molecular markers in the process of desert locust (*Schistocerca gregaria*) phase transition. *Biochemical and biophysical research communications*, 341(2):599–606, 2006.

Richard E Clopton, J Janovy Jr, and TJ Percival. Host stadium specificity in the gregarine assemblage parasitizing *Tenebrio molitor*. *The Journal of parasitology*, pages 334–337, 1992.

Brad S Coates, Douglas V Sumerford, Richard L Hellmich, and Leslie C Lewis. Mining an *Ostrinia nubilalis* midgut expressed sequence tag (est) library for candidate genes and single nucleotide polymorphisms (snps). *Insect molecular biology*, 17 (6):607–620, 2008.

Christopher L Colbert, Chai-Wan Kim, Young-Ah Moon, Lisa Henry, Maya Palnitkar, William B McKean, Kevin Fitzgerald, Johann Deisenhofer, Jay D Horton, and Hyock Joo Kwon. Crystal structure of spot 14, a modulator of fatty acid synthesis. *Proceedings of the National Academy of Sciences*, 107(44):18820–18825, 2010.

Ana Conesa and Stefan Götz. Blast2go: A comprehensive suite for functional analysis in plant genomics. *International journal of plant genomics*, 2008, 2008.

Ana Conesa, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles. Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676, 2005.

COPR. The locust and grasshopper agricultural manual. *Overseas Pest Research, London*, page 690, 1982.

Miguel Corona, Rodrigo A Velarde, Silvia Remolina, Adrienne Moran-Lauter, Ying Wang, Kimberly A Hughes, and Gene E Robinson. Vitellogenin, juvenile hormone, insulin signaling, and queen honey bee longevity. *Proceedings of the National Academy of Sciences*, 104(17):7128–7133, 2007.

Miguel Corona, Romain Libbrecht, Yannick Wurm, Oksana Riba-Grognuz, Romain A Studer, and Laurent Keller. Vitellogenin underwent subfunctionalization to acquire caste and behavioral specific expression in the harvester ant *Pogonomyrmex barbatus*. *PLoS Genet*, 9(8):e1003730, 2013.

PK Cottee, EA Bernays, and AJ Mordue. Comparisons of deterrency and toxicity of selected secondary plant compounds to an oligophagous and a polyphagous acridid. *Entomologia experimentalis et applicata*, 46(3):241–247, 1988.

Anna T Curtis, Masahiro Hori, Janell M Green, William J Wolfgang, Kiyoshi Hiruma, and Lynn M Riddiford. Ecdysteroid regulation of the onset of cuticular melanization in allatectomized and black mutant *Manduca sexta* larvae. *Journal of insect physiology*, 30(8):597–606, 1984.

Svetlana Cvejic, Zheng Zhu, Sarah J Felice, Yemiliya Berman, and Xin-Yun Huang. The endogenous ligand stunted of the gpcr methuselah extends lifespan in *Drosophila*. *Nature cell biology*, 6(6):540–546, 2004.

Phillip J Daborn, Christopher Lumb, Adrian Boey, Wayn Wong, and Philip Batterham. Evaluating the insecticide resistance potential of eight *Drosophila melanogaster* cytochrome p450 genes by transgenic over-expression. *Insect biochemistry and molecular biology*, 37(5):512–519, 2007.

Mads Daugaard, Mikkel Rohde, and Marja Jäättelä. The heat shock protein 70 family: Highly homologous proteins with overlapping and distinct functions. *FEBS letters*, 581(19):3702–3710, 2007.

Mary Ellen Davey and Frans J de Bruijn. A homologue of the tryptophan-rich sensory protein tspo and fixl regulate a novel nutrient deprivation-induced *Sinorhizobium meliloti* locus. *Applied and environmental microbiology*, 66(12):5353–5359, 2000.

PM Davey. Quantities of food eaten by the desert locust, *Schistocerca gregaria* (forsk.), in relation to growth. *Bulletin of entomological research*, 45(03):539–551, 1954.

Clarice de Azevedo Souza, Sung Soo Kim, Stefanie Koch, Lucie Kienow, Katja Schneider, Sarah M McKim, George W Haughn, Erich Kombrink, and Carl J Douglas. A novel fatty acyl-coa synthetase is required for pollen development and sporopollenin biosynthesis in *Arabidopsis*. *The Plant Cell*, 21(2):507–525, 2009.

Sujata De Chaudhuri, Pritha Ghosh, Nilendu Sarma, Papiya Majumdar, Tanmoy Jyoti Sau, Santanu Basu, Susanta Roychoudhury, Kunal Ray, and Ashok K Giri. Genetic variants associated with arsenic susceptibility: study of purine nucleoside phosphorylase, arsenic (+ 3) methyltransferase, and glutathione s-transferase omega genes. *Environmental health perspectives*, 116(4):501, 2008.

CAD De Kort and NA Granger. Regulation of jh titers: the relevance of degradative enzymes and binding proteins. *Archives of Insect Biochemistry and Physiology*, 33 (1):1–26, 1996.

JL De La Pompa, JR Garcia, and Alberto Ferrús. Genetic analysis of muscle development in *Drosophila melanogaster*. *Developmental biology*, 131(2):439–454, 1989.

Arnold De Loof, Jurgen Huybrechts, Marisa Geens, Tim Vandersmissen, Bart Boerjan, and Liliane Schoofs. Sexual differentiation in adult insects: male-specific cuticular yellowing in *Schistocerca gregaria* as a model for reevaluating some current (neuro) endocrine concepts. *Journal of insect physiology*, 56(8):919–925, 2010.

Veronique Delmas, David G Stokes, and Robert P Perry. A mammalian dna-binding protein that contains a chromodomain and an snf2/swi2-like helicase domain. *Proceedings of the National Academy of Sciences*, 90(6):2414–2418, 1993.

Isabelle Delon and Nicholas H Brown. The integrin adhesion complex changes its composition and function during morphogenesis of an epithelium. *Journal of cell science*, 122(23):4363–4374, 2009.

Arop Leek Deng, Baldwyn Torto, Ahmed Hassanali, and EE Ali. Effects of shifting to crowded or solitary conditions on pheromone release and morphometrics of the desert locust, *Schistocerca gregaria* (forskål)(orthoptera: Acrididae). *Journal of Insect Physiology*, 42(8):771–776, 1996.

Eric Dessaud, Danièle Salaün, Odile Gayet, and Marie Chabbert. Identification of lynx2, a novel member of the ly-6/neurotoxin superfamily, expressed in neuronal subpopulations during mouse development. *Molecular and Cellular Neuroscience*, 31(2):232–242, 2006.

Senne Dillen, Rik Verdonck, Sven Zels, Pieter Van Wielendaele, and Jozef Vanden Broeck. Identification of the short neuropeptide f precursor in the desert locust: evidence for an inhibitory role of snpf in the control of feeding. *Peptides*, 53:134–139, 2014.

RJ Dillon and AK Charnley. Invasion of the pathogenic fungus *Metarhizium anisopliae* through the guts of germfree desert locusts, *Schistocerca gregaria. Mycopathologia*, 96(1):59–66, 1986.

Rod Dillon and Keith Charnley. Mutualism between the desert locust *Schistocerca gregaria* and its gut microbiota. *Research in Microbiology*, 153(8):503–509, 2002.

Rod J Dillon, Chris T Vennard, and A Keith Charnley. Pheromones: Exploitation of gut bacteria in the locust. *Nature*, 403 (6772):851–851, 2000.

Zhiping Ding, Yucong Wen, Baojun Yang, Yixi Zhang, Shuhua Liu, Zewen Liu, and Zhaojun Han. Biochemical mechanisms of imidacloprid resistance in nilaparvata lugens: over-expression of cytochrome p450 cyp6ay1. *Insect biochemistry and molecular biology*, 43(11):1021–1027, 2013.

Robin E Dodson and David J Shapiro. Vigilin, a ubiquitous protein with 14 k homology domains, is the estrogen-inducible vitellogenin mrna 3'-untranslated region-binding protein. *Journal of Biological Chemistry*, 272(19):12249–12252, 1997.

Juliane C Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput dna sequencing. *Nucleic acids research*, 36(16):e105–e105, 2008.

Ying Dong and Markus Friedrich. Nymphal rnai: systemic rnai mediated gene knockdown in juvenile grasshopper. *BMC biotechnology*, 5(1):25, 2005.

Jeffery A Dusek, Hasan H Otu, Ann L Wohlhueter, Manoj Bhasin, Luiz F Zerbini, Marie G Joseph, Herbert Benson, and Towia A Libermann. Genomic counter-stress changes induced by the relaxation response. *PloS one*, 3(7):e2576, 2008.

Roman Dziarski. Peptidoglycan recognition proteins (pgrps). *Molecular immunology*, 40(12):877–886, 2004.

Benjamin A Eaton, Richard D Fetter, and Graeme W Davis. Dynactin is necessary for synapse stabilization. *Neuron*, 34(5): 729–741, 2002.

Robert C. Edgar. http://www.drive5.com/uclust, 2010.

Peggy E Ellis. Changes in the social aggregation of locust hoppers with changes in rearing conditions. *Animal Behaviour*, 11 (1):152–160, 1963.

Peggy E Ellis. Marching and colour in locust hoppers in relation to social factors. *Behaviour*, 23(3):177–191, 1964.

Philipp Engel and Nancy A Moran. The gut microbiota of insects–diversity in structure and function. *FEMS Microbiology Reviews*, 37(5):699–735, 2013.

Patamarerk Engsontia, Alan P Sanderson, Matthew Cobb, Kimberly KO Walden, Hugh M Robertson, and Stephen Brown. The red flour beetle's large nose: an expanded odorant receptor gene family in *Tribolium castaneum. Insect biochemistry and molecular biology*, 38(4):387–397, 2008.

Jorrit M Enserink and Richard D Kolodner. An overview of cdk1-controlled targets and processes. *Cell division*, 5(1):1, 2010.

Jay D Evans and Diana E Wheeler. Differential gene expression between developing queens and workers in the honey bee, *Apis mellifera*. *Proceedings of the National Academy of Sciences*, 96(10):5575–5580, 1999.

Cassandra Falckenhayn, Bart Boerjan, Günter Raddatz, Marcus Frohme, Liliane Schoofs, and Frank Lyko. Characterization of genome methylation patterns in the desert locust *Schistocerca gregaria*. *The Journal of experimental biology*, 216 (8):1423–1429, 2013.

Kimberly L Falk and Jonathan Gershenzon. The desert locust, *Schistocerca gregaria*, detoxifies the glucosinolates of schouwia purpurea by desulfation. *Journal of chemical ecology*, 33(8):1542–1555, 2007.

Min Fang, Zhirong Shen, Song Huang, Liping Zhao, She Chen, Tak W Mak, and Xiaodong Wang. The er udpase entpd5 promotes protein n-glycosylation, the warburg effect, and proliferation in the pten pathway. *Cell*, 143(5):711–724, 2010.

FAO. http://www.fao.org/edible-insects/en/, 2003a.

FAO. Workshop of use of green muscle® (*Metarhizium anisopliae* var *acridum*) and pan to control desert locust hopper bands, 2003b.

FAO. http://www.fao.org/ag/locusts/en/info/info/index.html, 2009.

Jacobus Christian Faure. Phase variation inthe army worm, laphygma exempta (walk.). *Science Bulletin. Department of Agriculture and Forestry, Union of South Africa*, (234), 1943.

YJ Feng, Y Ge, SQ Tan, KQ Zhang, R Ji, and WP Shi. Effect of paranosema locustae (microsporidia) on the behavioural phases of *Locusta migratoria* (orthoptera: Acrididae) in the laboratory. *Biocontrol Science and Technology*, 25(1):48–55, 2015.

Hans-Joerg Ferenz and Karsten Seidelmann. Pheromones in relation to aggregation and reproduction in desert locusts. *Physiological Entomology*, 28(1):11–18, 2003.

Rene Feyereisen. Insect p450 enzymes. *Annual review of entomology*, 44(1):507–533, 1999.

Richard H Ffrench-Constant, Thomas A Rocheleau, Jessica C Steichen, and Alison E Chalmers. A point mutation in a *Drosophila* gaba receptor confers insecticide resistance. *Nature*, 363(6428):449–451, 1993.

Ronald A Fisher. The design of experiments. 1935. *Oliver and Boyd, Edinburgh*, 1935.

RB Flavell, MD Bennett, JB Smith, and DB Smith. Genome size and the proportion of repeated nucleotide sequence dna in plants. *Biochemical genetics*, 12 (4):257–269, 1974.

Sylvain Forêt, Kevin W Wanner, and Ryszard Maleszka. Chemosensory proteins in the honey bee: Insights from the annotated genome, comparative analyses and expressional profiling. *Insect biochemistry and molecular biology*, 37(1):19–28, 2007.

DP Fox. The control of chiasma distribution in the locust, *Schistocerca gregaria* (forskål). *Chromosoma*, 43(3):289–328, 1973.

Frances V Fuller-Pace and Simak Ali. The dead box rna helicases p68 (ddx5) and p72 (ddx17): novel transcriptional coregulators. *Biochemical Society Transactions*, 36(4):609–612, 2008.

Yosuke Funato, Tatsuo Michiue, Makoto Asashima, and Hiroaki Miki. The thioredoxin-related redox-regulating protein nucleoredoxin inhibits wnt-β-catenin signalling through dishevelled. *Nature cell biology*, 8(5):501–508, 2006.

Edward Gaten, Stephen J Huston, Harold B Dowse, and Tom Matheson. Solitary and gregarious locusts differ in circadian rhythmicity of a visual output neuron. *Journal of Biological Rhythms*, 27(3):196–205, 2012.

Robert Gentleman, Ross Ihaka, and D Bates. The r project for statistical computing. *R home web site: http://www. r-project. org*, 1997.

Sylvia D Gillett. Social determinants of aggregation behaviour in adults of the desert locust. *Animal Behaviour*, 21(3): 599–606, 1973.

Sylvia D Gillett. Changes in the social behaviour of the desert locust, *Schistocerca gregaria*, in response to the gregarizing pheromone. *Animal Behaviour*, 23:494–503, 1975.

John I Glendinning. How do herbivorous insects cope with noxious secondary plant compounds in their diet? In *Proceedings of the 11th International Symposium on Insect-Plant Relationships*, pages 15–25. Springer, 2002.

Michael Gmachl, Christian Gieffers, Alexandre V Podtelejnikov, Matthias Mann, and Jan-Michael Peters. The ring-h2 finger protein apc11 and the e2 enzyme ubc4 are sufficient to ubiquitinate substrates of the anaphase-promoting complex. *Proceedings of the National Academy of Sciences*, 97(16):8973–8978, 2000.

MM Goddeeris, E Cook-Wiens, WJ Horton, H Wolf, JR Stoltzfus, M Borrusch, and MS Grotewiel. Delayed behavioural aging and altered mortality in *Drosophila*β integrin mutants. *Aging Cell*, 2(5):257–264, 2003.

Verena Goebeler, Daniela Ruhe, Volker Gerke, and Ursula Rescher. Atypical properties displayed by annexin a9, a novel member of the annexin family of ca^{2+} and lipid binding proteins. *FEBS letters*, 546(2-3): 359–364, 2003.

André Goffeau, Benoît De Hertogh, and Philippe Baret. Abc transporters. 2004.

Da-Ping Gong, Hui-jie Zhang, Ping Zhao, Ying Lin, Qing-You Xia, and Zhong-Huai Xiang. Identification and expression pattern of the chemosensory protein gene family in the silkworm, *Bombyx mori. Insect biochemistry and molecular biology*, 37(3):266–277, 2007.

SD Gordon, S Rogers, and J Windmill. Hearing differences of gregarious and solitary locusts (*Schistocerca gregaria*), an example of epigenetic effects. In *Front. Behav. Neurosci. Conference Abstract: Tenth International Congress of Neuroethology. doi: 10.3389/conf. fnbeh*, volume 122, 2012.

J Chad Gore and Coby Schal. Gene expression and tissue distribution of the major human allergen bla g 1 in the german cockroach, *Blattella germanica*a l.(dictyoptera: Blattellidae). *Journal of medical entomology*, 41(5):953–960, 2004.

Steven Gotham and Hojun Song. Nonswarming grasshoppers exhibit density-dependent phenotypic plasticity reminiscent of swarming locusts. *Journal of insect physiology*, 59(11):1151–1159, 2013.

Brenton R Graveley, Angela N Brooks, Joseph W Carlson, Michael O Duff, Jane M Landolin, Li Yang, Carlo G Artieri, Marijke J van Baren, Nathan Boley, and Benjamin W Booth. The developmental transcriptome of *Drosophila melanogaster. Nature*, 471(7339):473–479, 2011.

Lindsey J Gray, Gregory A Sword, Michael L Anstey, Fiona J Clissold, and Stephen J Simpson. Behavioural phase polyphenism in the australian plague locust (*Chortoicetes terminifera*). *Biology Letters*, 5(3): 306–309, 2009.

Dina N Greene, Tzintzuni Garcia, R Bryan Sutton, Kim M Gernert, Guy M Benian, and Andres F Oberhauser. Single-molecule force spectroscopy reveals a stepwise unfolding of caenorhabditis elegans giant protein kinase domains. *Biophysical journal*, 95(3):1360–1370, 2008.

T Ryan Gregory. Animal genome size database, 2001.

T Ryan Gregory. Synergy between sequence and size in large-scale genomics. *Nature Reviews Genetics*, 6(9):699–708, 2005.

Sam Griffiths-Jones, Simon Moxon, Mhairi Marshall, Ajay Khanna, Sean R Eddy, and Alex Bateman. Rfam: annotating non-coding rnas in complete genomes. *Nucleic acids research*, 33(suppl 1):D121–D124, 2005.

Patricia JTA Groenen, Karin B Merck, Wilfried W Jong, and Hans Bloemendal. Structure and modifications of the junior chaperone α-crystallin. *European Journal of Biochemistry*, 225(1):1–19, 1994.

Sebastián Guelman, Kenji Kozuka, Yifan Mao, Victoria Pham, Mark J Solloway, John Wang, Jiansheng Wu, Jennie R Lill, and Jiping Zha. The double-histone-acetyltransferase complex atac is essential for mammalian development. *Molecular and cellular biology*, 29(5):1176–1188, 2009.

Stéphane Guindon, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Systematic biology*, 59(3):307–321, 2010.

Su Guo and Kenneth J Kemphues. par-1, a gene required for establishing polarity in c. elegans embryos, encodes a putative ser/thr kinase that is asymmetrically distributed. *Cell*, 81(4):611–620, 1995.

Wei Guo, Xianhui Wang, Zongyuan Ma, Liang Xue, Jingyao Han, Dan Yu, and Le Kang. Csp and takeout genes modulate the switch between attraction and repulsion during behavioral phase change in the migratory locust. *PLoS Genet*, 7(2): e1001291, 2011.

Xiaojiao Guo, Zongyuan Ma, and Le Kang. Serotonin enhances solitariness in phase transition of the migratory locust. *Frontiers in behavioral neuroscience*, 7, 2013.

Vishwesha Guttal, Pawel Romanczuk, Stephen J Simpson, Gregory A Sword, and Iain D Couzin. Cannibalism can drive the evolution of behavioural phase polyphenism in locusts. *Ecology letters*, 15(10): 1158–1166, 2012.

Dirk Görlich, Siegfried Prehn, Enno Hartmann, Joachim Herz, Albrecht Otto, Regine Kraft, Martin Wiedmann, Siegne Knespel, Bernhard Dobberstein, and Tom A Rapoport. The signal sequence receptor has a second subunit and is part of a translocation complex in the endoplasmic reticulum as probed by bifunctional reagents. *The Journal of cell biology*, 111 (6):2283–2294, 1990.

Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, and Matthias Lieber. De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis. *Nature protocols*, 8(8):1494–1512, 2013.

Bernd F Hägele, Vicky Oag, Abdelghani Bouaïchi, Alan R McCaffery, and Stephen J Simpson. The role of female accessory glands in maternal inheritance of phase in the desert locust *Schistocerca gregaria*. *Journal of Insect Physiology*, 46 (3):275–280, 2000.

Ricci J Haines, Laura C Pendleton, and Duane C Eichler. Argininosuccinate synthase: at the center of arginine metabolism. *International journal of biochemistry and molecular biology*, 2(1):8, 2011.

Parvin Hakimi, Jianqi Yang, Gemma Casadesus, Duna Massillon, Fatima Tolentino-Silva, Colleen K Nye, Marco E Cabrera, David R Hagen, Christopher B Utter, and Yacoub Baghdy. Overexpression of the cytosolic form of phosphoenolpyruvate carboxykinase (gtp) in skeletal muscle repatterns energy metabolism in the mouse. *Journal of Biological Chemistry*, 282 (45):32844–32855, 2007.

Amel Ben Hamouda, Seiji Tanaka, Mohamed Habib Ben Hamouda, and Abderrahmen Bouain. Density-dependent phenotypic plasticity in body coloration and morphometry and its transgenerational changes in the migratory locust, *Locusta migratoria*. *Journal of Entomology and Nematology*, 3(7):105–116, 2011.

Richard W Hanson and Parvin Hakimi. Born to run; the story of the pepck-c mus mouse. *Biochimie*, 90(6):838–842, 2008.

Zheng-Bo He, Yue-Qing Cao, You-Ping Yin, Zhong-Kang Wang, Bin Chen, Guo-Xiong Peng, and Yu-Xian Xia. Role of hunchback in segment patterning of *Locusta migratoria* manilensis revealed by parental rnai. *Development, growth & differentiation*, 48(7):439–445, 2006.

Ashok N Hegde, Alfred L Goldberg, and James H Schwartz. Regulatory subunits of camp-dependent protein kinases are degraded after conjugation to ubiquitin:

a molecular mechanism underlying long-term synaptic plasticity. *Proceedings of the National Academy of Sciences*, 90 (16):7436–7440, 1993.

Yael Heifetz, Hillary Voet, and Shalom W Applebaum. Factors affecting behavioral phase transition in the desert locust, *Schistocerca gregaria* (forskål)(orthoptera: Acrididae). *Journal of Chemical Ecology*, 22(9):1717–1734, 1996.

Osnat Herzberg, John Moult, and MN James. Calcium binding to skeletal muscle troponin c and the regulation of muscle contraction. In *Ciba Foundation Symposium*, volume 122, pages 120–144, 1986.

Kiyoshi Hiruma and Lynn M Riddiford. The molecular mechanisms of cuticular melanization: the ecdysone cascade leading to dopa decarboxylase expression in manduca sexta. *Insect biochemistry and molecular biology*, 39(4):245–253, 2009.

Jules A Hoffmann. The immune response of drosophila. *Nature*, 426(6962):33–38, 2003.

Roger A Hoskins, Jane M Landolin, James B Brown, Jeremy E Sandler, Hazuki Takahashi, Timo Lassmann, Charles Yu, Benjamin W Booth, Dayu Zhang, and Kenneth H Wan. Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome research*, 21(2):182–192, 2011.

Bruno Hoste, SJ Simpson, S Tanaka, D-H Zhu, Arnold De Loof, and Michael Breuer. Effects of [his 7]-corazonin on the phase state of isolated-reared (solitarious) desert locusts, *Schistocerca gregaria*. *Journal of Insect Physiology*, 48(10):981–990, 2002.

Xiaoqiu Huang and Anup Madan. Cap3: A dna sequence assembly program. *Genome research*, 9(9):868–877, 1999.

S Hudault, J Guignot, and AL Servin. Escherichia coli strains colonising the gastrointestinal tract protect germfree mice againstsalmonella typhimuriuminfection. *Gut*, 49(1):47–55, 2001.

Joseph Hughes and Alfried P Vogler. Gene expression in the gut of keratin-feeding clothes moths (tineola) and keratin beetles (trox) revealed by subtracted cdna libraries. *Insect biochemistry and molecular biology*, 36(7):584–592, 2006.

Jun R Huh, Ian Foe, Israel Muro, Chun Hong Chen, Jae Hong Seol, Soon Ji Yoo, Ming Guo, Jin Mo Park, and Bruce A Hay. The drosophila inhibitor of apoptosis (iap) diap2 is dispensable for cell survival, required for the innate immune response to gram-negative bacterial infection, and can be negatively regulated by the reaper/hid/grim family of iap-binding apoptosis inducers. *Journal of Biological Chemistry*, 282(3):2056–2068, 2007.

Kamal M Ibrahim, Patricia Sourrouille, and Godfrey M Hewitt. Are recession populations of the desert locust (*Schistocerca gregaria*) remnants of past swarms? *Molecular Ecology*, 9(6):783–791, 2000.

HS Injeyan and SS Tobe. Phase polymorphism in *Schistocerca gregaria*: assessment of juvenile hormone synthesis in relation to vitellogenesis. *Journal of Insect Physiology*, 27(3):203–210, 1981.

M Saiful Islam, Peter Roessingh, Stephen J Simpson, and Alan R Mccaffery. Effects of population density experienced by parents during mating and oviposition on the phase of hatchling desert locusts, *Schistocerca gregaria*. *Proceedings of the Royal Society of London B: Biological Sciences*, 257(1348):93–98, 1994.

Yoichi Ito and Haruyuki Sonobe. The role of ecdysteroid 22-kinase in the accumulation of ecdysteroids in ovary of silkworm *Bombyx mori*. *Annals of the New York Academy of Sciences*, 1163(1):421–424, 2009.

Syun'iti Iwao. A new regression method for analyzing the aggregation pattern of animal populations. *Researches on Population Ecology*, 10(1):1–20, 1968.

Eva Jablonka and Marion J Lamb. *Epigenetic inheritance and evolution: the Lamarckian dimension*. Oxford University Press on Demand, 1999.

Eva Jablonka and Gal Raz. Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *The*

Quarterly review of biology, 84(2):131–176, 2009.

Emmanuelle Jacquin-Joly, Richard G Vogt, Marie-Christine François, and Patricia Nagnan-Le Meillour. Functional and expression pattern analysis of chemosensory proteins expressed in antennae and pheromonal gland of *Mamestra brassicae*. *Chemical senses*, 26(7):833–844, 2001.

Arumugam Jayakumar, Ming-Hong Tai, Wei-Yong Huang, Walid Al-Feel, Matthew Hsu, Lutfi Abu-Elheiga, Subrahmanyam S Chirala, and Salih J Wakil. Human fatty acid synthase: properties and molecular cloning. *Proceedings of the National Academy of Sciences*, 92(19):8695–8699, 1995.

Marek Jindra, Subba R Palli, and Lynn M Riddiford. The juvenile hormone signaling pathway in insect development. *Annual review of entomology*, 58:181–204, 2013.

Bruce Johnson. Wing polymorphism in aphids iii. the influence of the host plant. *Entomologia Experimentalis et Applicata*, 9(2):213–222, 1966.

Erica S Johnson, Ingrid Schwienhorst, R Jürgen Dohmen, and Günter Blobel. The ubiquitin-like protein smt3p is activated for conjugation to other proteins by an aos1p/uba2p heterodimer. *The EMBO journal*, 16(18):5509–5519, 1997.

H-H Kaatz, H-J Ferenz, B Langer, and RFA Moritz. Isolation and characterization of nine polymorphic microsatellite loci from the desert locust, *Schistocerca gregaria*. *Molecular Ecology Notes*, 7(6):1042–1044, 2007.

Yukiko Kabeya, Noboru Mizushima, Akitsugu Yamamoto, Satsuki Oshitani-Okamoto, Yoshinori Ohsumi, and Tamotsu Yoshimori. Lc3, gabarap and gate16 localize to autophagosomal membrane depending on form-ii formation. *Journal of cell science*, 117(13):2805–2812, 2004.

Bernhard Kadenbach, Susanne Arnold, Icksoo Lee, and Maik Hüttemann. The possible role of cytochrome c oxidase in stress-induced apoptosis and degenerative diseases. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1655:400–408, 2004.

Mohamed Habib Kane. *Cross-stage physiological effects of the Desert locust, Schistocerca gregaria, aggregation pheromones on their behaviour and susceptibility to control agents.* PhD thesis, 2012.

Le Kang, XiangYong Chen, Yan Zhou, Bo-Wan Liu, Wei Zheng, RuiQiang Li, Jun Wang, and Jun Yu. The analysis of large-scale gene expression correlated to the phase changes of the migratory locust. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51):17611–17615, 2004.

Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

Yohei Katoh, Brigitte Ritter, Thomas Gaffry, Francois Blondeau, Stefan Höning, and Peter S McPherson. The clavesin family, neuron-specific lipid-and clathrin-binding sec14 proteins regulating lysosomal morphology. *Journal of Biological Chemistry*, 284(40):27646–27654, 2009.

Tetyana Khomenko, Xiaoming Deng, Martin R Jadus, and Sandor Szabo. Effect of cysteamine on redox-sensitive thiol-containing proteins in the duodenal mucosa. *Biochemical and biophysical research communications*, 309(4):910–916, 2003.

Yong-Sik Kim, Ji-Hwan Ryu, Sung-Jun Han, Kun-Ho Choi, Ki-Bum Nam, In-Hwan Jang, Bruno Lemaitre, Paul T Brey, and Won-Jae Lee. Gram-negative bacteria-binding protein, a pattern recognition receptor for lipopolysaccharide and β-1, 3-glucan that mediates the signaling for the induction of innate immune genes in *Drosophila melanogaster* cells. *Journal of Biological Chemistry*, 275(42):32721–32727, 2000.

Aya Nomura Kitabayashi, Toshimitsu Arai, Takeo Kubo, and Shunji Natori. Molecular cloning of cdna for p10, a novel protein that increases in the regenerating legs of *Periplaneta americana* (american cockroach). *Insect biochemistry and molecular biology*, 28(10):785–790, 1998.

Josef T Kittler, Philippe Rostaing, Giampietro Schiavo, Jean-Marc Fritschy, Richard Olsen, Antoine Triller, and Stephen J Moss. The subcellular distribution of gabarap and its ability to interact with nsf suggest a role for this protein in the intracellular transport of gaba a receptors. *Molecular and Cellular Neuroscience*, 18 (1):13–25, 2001.

Bostjan Kobe and Johann Deisenhofer. Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. *Nature*, 366(6457):751, 1993.

Lei Kong, Yong Zhang, Zhi-Qiang Ye, Xiao-Qiao Liu, Shu-Qi Zhao, Liping Wei, and Ge Gao. Cpc: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic acids research*, 35(suppl 2):W345–W349, 2007.

Ingrid Chou Koo, Yamini M Ohol, Ping Wu, J Hiroshi Morisaki, Jeffery S Cox, and Eric J Brown. Role for lysosomal enzyme β-hexosaminidase in the control of mycobacteria infection. *Proceedings of the National Academy of Sciences*, 105 (2):710–715, 2008.

Maryla Krajewska, Lucy Xu, Wenjie Xu, Stan Krajewski, Christina L Kress, Jiankun Cui, Li Yang, Fumitoshi Irie, Yu Yamaguchi, and Stuart A Lipton. Endoplasmic reticulum protein bi-1 modulates unfolded protein response signaling and protects against stroke and traumatic brain injury. *Brain research*, 1370:227–237, 2011.

Maxwell M Krem and Enrico Di Cera. Molecular markers of serine protease evolution. *The EMBO journal*, 20(12):3036–3045, 2001.

Charli Kruse, Arnold Grünweller, K Dagmar Willkomm, Thomas Pfeiffer, K Roland Hartmann, and K Peter Müller. trna is entrapped in similar, but distinct, nuclear and cytoplasmic ribonucleoprotein complexes, both of which contain vigilin and elongation factor 1α. *Biochemical Journal*, 329(3):615–621, 1998.

Jonna Kulmuni and Heli Havukainen. Insights into the evolution of the csp gene family through the integration of evolutionary analysis and comparative protein modeling. *PloS one*, 2013.

Michael Kühl, Laird C Sheldahl, Maiyon Park, Jeffrey R Miller, and Randall T Moon. The wnt/ca 2+ pathway: a new vertebrate wnt signaling pathway takes shape. *Trends in genetics*, 16(7):279–283, 2000.

Sadhana Lal and Silvia Tabacchioni. Ecology and biotechnological potential of paenibacillus polymyxa: a minireview. *Indian Journal of Microbiology*, 49(1):2–10, 2009.

Ronald E Laliberte, David G Perregaux, Lise R Hoth, Philip J Rosner, Crystal K Jordan, Kevin M Peese, James F Eggler, Mark A Dombroski, Kieran F Geoghegan, and Christopher A Gabel. Glutathione s-transferase omega 1-1 is a target of cytokine release inhibitory drugs and may be responsible for their effect on interleukin-1β posttranslational processing. *Journal of Biological Chemistry*, 278(19):16567–16578, 2003.

Yaara Lancet and Reuven Dukas. Socially influenced behaviour and learning in locusts. *Ethology*, 118(3):302–310, 2012.

Minglin Lang, Lei Wang, Qiangwang Fan, Guiran Xiao, Xiaoxi Wang, Yi Zhong, and Bing Zhou. Genetic inhibition of solute-linked carrier 39 family transporter 1 ameliorates aβ pathology in a drosophila model of alzheimer's disease. *PLoS Genet*, 8(4):e1002683, 2012.

Alexandre V Latchininsky. Locusts and remote sensing: a review. *Journal of Applied Remote Sensing*, 7(1):075099–075099, 2013.

Jelica Lazarević and Milena Janković-Tomanić. Dietary and phylogenetic correlates of digestive trypsin activity in insect pests. *Entomologia Experimentalis et Applicata*, 157(2):123–151, 2015.

Si Quang Le and Olivier Gascuel. An improved general amino acid replacement matrix. *Molecular biology and evolution*, 25(7):1307–1320, 2008.

In Hye Lee, Yoshichika Kawai, Maria M Fergusson, Ilsa I Rovira, Alexander JR Bishop, Noboru Motoyama, Liu Cao, and Toren Finkel. Atg7 modulates p53 activity to regulate cell cycle and survival during metabolic stress. *Science*, 336(6078):225–228, 2012.

Kyu-Sun Lee, Kwan-Hee You, Jong-Kil Choo, Yong-Mahn Han, and Kweon Yu. Drosophila short neuropeptide f regulates food intake and body size. *Journal of Biological Chemistry*, 279(49):50781–50789, 2004.

Yun-Il Lee, Daniel Giovinazzo, Ho Chul Kang, Yunjong Lee, Jun Seop Jeong, Paschalis-Thomas Doulias, Zhi Xie, Jianfei Hu, Mehdi Ghasemi, and Harry Ischiropoulos. Protein microarray characterization of the s-nitrosoproteome. *Molecular & Cellular Proteomics*, 13(1):63–72, 2014.

Pierre Legrain, Christine Chapon, and Fr6d6rique Galisson. Interactions between prp9 and spp91 splicing factors identify a protein complex required in prespliceosome assembly. *Genes & development*, 7(7b):1390–1399, 1993.

Catherine Leonard, Norman A Ratcliffe, and Andrew F Rowley. The role of prophenoloxidase activation in non-self recognition and phagocytosis by insect blood cells. *Journal of Insect Physiology*, 31(10):789–799, 1985.

R Leo Lester, Constantin Grach, Meir Paul Pener, and Stephen J Simpson. Stimuli inducing gregarious colouration and behaviour in nymphs of *Schistocerca gregaria*. *Journal of Insect Physiology*, 51(7):737–747, 2005.

Ivica Letunic and Peer Bork. Interactive tree of life v2: online annotation and display of phylogenetic trees made easy. *Nucleic acids research*, page gkr201, 2011.

Elena A Levashina, Emma Langley, Clare Green, David Gubb, Michael Ashburner, Jules A Hoffmann, and Jean-Marc Reichhart. Constitutive activation of toll-mediated antifungal defense in serpin-deficient drosophila. *Science*, 285(5435):1917–1919, 1999.

Robin J Levin, Patricia L Boychuk, Colleen M Croniger, Jeffrey A Kazzaz, and Charles E Rozek. Structure and expression of a muscle specific gene which is adjacent to the *Drosophila* myosin heavy-chain gene and can encode a cytochrome b related protein. *Nucleic acids research*, 17(15):6349–6367, 1989.

Shoshana Levy and Tsipi Shoham. The tetraspanin web modulates immune-signalling complexes. *Nature Reviews Immunology*, 5(2):136–148, 2005.

Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

Liwu Li, Mats Ljungman, and Jack E Dixon. The human cdc14 phosphatases interact with and dephosphorylate the tumor suppressor protein p53. *Journal of Biological Chemistry*, 275(4):2410–2414, 2000.

Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

Yan Li, Jie Zhang, Dafeng Chen, Pengcheng Yang, Feng Jiang, Xianhui Wang, and Le Kang. Crispr/cas9 in locusts: Successful establishment of an olfactory deficiency line by targeting the mutagenesis of an odorant receptor co-receptor (orco). *Insect Biochemistry and Molecular Biology*, 79:27–35, 2016.

Zhao-Qun Li, Shuai Zhang, Jun-Yu Luo, Jing Zhu, Jin-Jie Cui, and Shuang-Lin Dong. Expression analysis and binding assays in the chemosensory protein gene family indicate multiple roles in helicoverpa armigera. *Journal of chemical ecology*, 41(5):473–485, 2015.

Zicai Liang, Lars Sottrup-Jensen, Anna Aspán, Martin Hall, and Kenneth Söderhäll. Pacifastin, a novel 155-kda heterodimeric proteinase inhibitor containing a unique transferrin chain. *Proceedings of the National Academy of Sciences*, 94(13):6682–6687, 1997.

Romain Libbrecht, Miguel Corona, Franziska Wende, Dihego O Azevedo, Jose E Serrão, and Laurent Keller. Interplay between insulin signaling, juvenile hormone, and vitellogenin regulates maternal effects on polyphenism in ants. *Proceedings of the National Academy of Sciences*, 110 (27):11050–11055, 2013.

Pablo Librado and Julio Rozas. Dnasp v5: a software for comprehensive analysis of dna polymorphism data. *Bioinformatics*, 25(11):1451–1452, 2009.

Petros Ligoxygakis, Nadege Pelte, Jules A Hoffmann, and Jean-Marc Reichhart. Activation of *Drosophila* toll during fungal infection by a blood serine protease. *Science*, 297(5578):114–116, 2002.

S Lindquist and EA Craig. The heat-shock proteins. *Annual review of genetics*, 22 (1):631–677, 1988.

Jun Liu, Mark W Albers, Chih-Ming Chen, Stuart L Schreiber, and Christopher T Walsh. Cloning, expression, and purification of human cyclophilin in *Escherichia coli* and assessment of the catalytic role of cysteines by site-directed mutagenesis. *Proceedings of the National Academy of Sciences*, 87(6):2304–2308, 1990.

Kenneth J Livak and Thomas D Schmittgen. Analysis of relative gene expression data using real-time quantitative pcr and the 2- $\delta\delta$ct method. *methods*, 25(4):402–408, 2001.

Jeffrey A Lockwood. Cannibalism in rangeland grasshoppers (orthoptera: Acrididae): attraction to cadavers. *Journal of the Kansas Entomological Society*, pages 379–387, 1988.

Arnold Loof, Ilse Claeys, Gert Simonet, Peter Verleyen, TIM Vandersmissen, Filip Sas, and Jurgen Huybrechts. Molecular markers of phase transition in locusts. *Insect Science*, 13(1):3–12, 2006.

NR Lovejoy, SP Mullen, GA Sword, RF Chapman, and RG Harrison. Ancient trans-atlantic flight explains locust biogeography: molecular phylogenetics of schistocerca. *Proceedings of the Royal Society of London B: Biological Sciences*, 273(1588):767–774, 2006.

C Lucas, R Kornfein, M Chakaborty-Chatterjee, J Schonfeld, N Geva, MB Sokolowski, and A Ayali. The locust foraging gene. *Archives of insect biochemistry and physiology*, 74(1):52–66, 2010.

Zongyuan Ma, Jun Yu, and Le Kang. Locustdb: a relational database for the transcriptome and biology of the migratory locust (*Locusta migratoria*). *BMC genomics*, 7(1):1, 2006.

Zongyuan Ma, Wei Guo, Xiaojiao Guo, Xianhui Wang, and Le Kang. Modulation of behavioral phase changes of the migratory locust by the catecholamine metabolic pathway. *Proceedings of the National Academy of Sciences*, 108(10):3882–3887, 2011.

Susan M Mackenzie, Michael R Brooker, Timothy R Gill, Graeme B Cox, Antony J Howells, and Gary D Ewart. Mutations in the white gene of *Drosophila melanogaster* affecting abc transporters that determine eye colouration. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1419(2):173–185, 1999.

Louise Madsen, Franziska Kriegenburg, Andrea Vala, Diana Best, Søren Prag, Kay Hofmann, Michael Seeger, Ian R Adams, and Rasmus Hartmann-Petersen. The tissue-specific rep8/ubxd6 tethers p97 to the endoplasmic reticulum membrane for degradation of misfolded proteins. *PloS one*, 6(9):e25061, 2011.

K Maeno, T Gotoh, and S Tanaka. Phase-related morphological changes induced by [his 7]-corazonin in two species of locusts, *Schistocerca gregaria* and *Locusta migratoria* (orthoptera: Acrididae). *Bulletin of entomological research*, 94(04): 349–357, 2004.

AM Mainguet, A Louveaux, G Sayed, and P Rollin. Ability of a generalist insect, *Schistocerca gregaria*, to overcome thioglucoside defense in desert plants: tolerance or adaptation? *Entomologia experimentalis et applicata*, 94(3):309–317, 2000.

J Maleszka, S Forêt, R Saint, and R Maleszka. Rnai-induced phenotypes suggest a novel role for a chemosensory protein csp5 in the development of embryonic integument in the honeybee (*Apis mellifera*). *Development genes and evolution*, 217(3):189–196, 2007.

Zulfiqar A Malik and Sumaira Amir. An apolipophorin iii protein from the hemolymph of desert locust, *Schistocerca gregaria*. *Applied biochemistry and biotechnology*, 165(7-8):1779–1788, 2011.

Bruno Maras, Donatella Barra, Silvestro Duprè, and Giuseppina Pitari. Is pantetheinase the actual identity of mouse and human vanin-1 proteins? *FEBS letters*, 461 (3):149–152, 1999.

Elaine R Mardis. Next-generation sequencing platforms. *Annual review of analytical chemistry*, 6:287–303, 2013.

Jeffrey A Martin and Zhong Wang. Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12(10):671–682, 2011.

Juliana R Martins, Francis MF Nunes, Alexandre S Cristino, Zilá LP Simões, and Márcia MG Bitondi. The four hexamerin genes in the honey bee: structure, molecular evolution and function deduced from expression patterns in queens, workers and drones. *BMC Molecular Biology*, 11(1):1, 2010.

Yasunari Matsuzaka, Koichi Okamoto, Tomotaka Mabuchi, Mariko Iizuka, Akira Ozawa, Akira Oka, Gen Tamiya, Jerzy K Kulski, and Hidetoshi Inoko. Identification, expression analysis and polymorphism of a novel rltpr gene encoding a rgd motif, tropomodulin domain and proline/leucine-rich regions. *Gene*, 343 (2):291–304, 2004.

John S Mattick and Igor V Makunin. Noncoding rna. *Human molecular genetics*, 15(suppl 1):R17–R29, 2006.

DONNA–MARIE McCafferty, JOHN S Mudgett, MARK G Swain, and PAUL Kubes. Inducible nitric oxide synthase plays a critical role in resolving intestinal inflammation. *Gastroenterology*, 112 (3):1022–1027, 1997.

ALAN R McCaffery, STEPHEN J Simpson, M Saiful Islam, and PETER Roessingh. A gregarizing factor present in the egg pod foam of the desert locust *Schistocerca gregaria*. *The Journal of experimental biology*, 201(3):347–363, 1998.

John D McCorvy and Bryan L Roth. Structure and function of serotonin g protein-coupled receptors. *Pharmacology & therapeutics*, 150:129–142, 2015.

Michael P McKenna, Daria S Hekmat-Scafe, Peter Gaines, and John R Carlson. Putative drosophila pheromone-binding proteins expressed in a subregion of the olfactory system. *Journal of Biological Chemistry*, 269(23):16340–16347, 1994.

S McNeely, R Beckmann, and AK Bence Lin. Chek again: revisiting the development of chk1 inhibitors for cancer therapy. *Pharmacology & therapeutics*, 142 (1):1–10, 2014.

Brian B McSpadden Gardener. Ecology of bacillus and paenibacillus spp. in agricultural systems. *Phytopathology*, 94(11): 1252–1258, 2004.

Annika Meinander, Christopher Runchel, Tencho Tenev, Li Chen, Chan-Hee Kim, Paulo S Ribeiro, Meike Broemer, Francois Leulier, Marketa Zvelebil, and Neal Silverman. Ubiquitylation of the initiator caspase dredd is required for innate immune signalling. *The EMBO journal*, 31 (12):2770–2783, 2012.

Michael L Metzker. Sequencing technologies—the next generation. *Nature reviews genetics*, 11(1):31–46, 2010.

Gabriel A Miller, M Saiful Islam, Timothy DW Claridge, Tim Dodgson, and Stephen J Simpson. Swarm formation in the desert locust *Schistocerca gregaria*: isolation and nmr analysis of the primary maternal gregarizing agent. *Journal of Experimental Biology*, 211(3):370–376, 2008.

Jason R Miller, Sergey Koren, and Granger Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95 (6):315–327, 2010.

Annamaria Mocciaro and Elmar Schiebel. Cdc14: a highly conserved family of phosphatases with non-conserved functions? *J Cell Sci*, 123(17):2867–2876, 2010.

Armin Philipp Moczek. Horn polyphenism in the beetle onthophagus taurus: larval diet quality and plasticity in parental investment determine adult body size and male horn morphology. *Behavioral Ecology*, 9(6):636–641, 1998.

Elagba HA Mohamed. Determination of nutritive value of the edible migratory locust *Locusta migratoria*, linnaeus, 1758 (orthoptera: Acrididae). 2016.

Peter Mombaerts. Seven-transmembrane proteins as odorant and chemosensory receptors. *Science*, 286(5440):707–711, 1999.

Hiroshi Moriuchi, Noriko Koda, Emiko Okuda-Ashitaka, Hiromi Daiyasu, Kensuke Ogasawara, Hiroyuki Toh, Seiji Ito, David F Woodward, and Kikuko Watanabe. Molecular characterization of a novel type of prostamide/prostaglandin f synthase, belonging to the thioredoxin-like superfamily. *Journal of biological chemistry*, 283(2):792–801, 2008.

Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

Ivan Nalvarte, Anastasios E Damdimopoulos, Joëlle Rüegg, and Giannis Spyrou. The expression and activity of thioredoxin reductase 1 splice variants v1 and v2 regulate the expression of genes associated with differentiation and adhesion. *Bioscience reports*, 35(6):e00269, 2015.

Renuka R Nayak, William E Bernal, Jessica W Lee, Michael J Kearns, and Vivian G Cheung. Stress-induced changes in gene interactions in human cells. *Nucleic acids research*, 42(3):1757–1771, 2014.

Vu Thuong Nguyen, Assane Ndoye, and Sergei A Grando. Pemphigus vulgaris antibody identifies pemphaxin a novel keratinocyte annexin-like molecule binding acetylcholine. *Journal of Biological Chemistry*, 275(38):29466–29476, 2000.

Makiya Nishikawa, Seiji Takemoto, and Yoshinobu Takakura. Heat shock protein derivatives for delivery of antigens to antigen presenting cells. *International journal of pharmaceutics*, 354(1):23–27, 2008.

Peter GN Njagi, B Torto, D Obeng-Ofori, and A Hassanali. Phase-independent responses to phase-specific aggregation pheromone in adult desert locusts, *Schistocerca gregaria* (orthoptera: Acrididae). *Physiological Entomology*, 21(2):131–137, 1996.

DJ Nolte. A pheromone for melanization of locusts. *Nature*, 200(4907):660–661, 1963.

DJ Nolte. Chiasma-induction and tyrosine metabolism in locusts. *Chromosoma*, 26 (3):287–297, 1969.

Audrey R Odom, Alke Stahlberg, Susan R Wente, and John D York. A role for nuclear inositol 1, 4, 5-trisphosphate kinase in transcriptional control. *Science*, 287 (5460):2026–2029, 2000.

Dorington O Ogoyi, Ellie O Osir, and Norah K Olembo. Lipophorin and apolipophorin-iii in solitary and gregarious phases of *Schistocerca gregaria*. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 112(3):441–449, 1995.

DGAB Oonincx and AFB Van der Poel. Effects of diet on the chemical composition of migratory locusts (*Locusta migratoria*). *Zoo biology*, 30(1):9–16, 2011.

Asako Otomo, Shinji Hadano, Takeya Okada, Hikaru Mizumura, Ryota Kunita, Hitoshi Nishijima, Junko Showguchi-Miyata, Yoshiko Yanagisawa, Eri Kohiki, and Etsuko Suga. Als2, a novel guanine nucleotide exchange factor for the small gtpase rab5, is implicated in endosomal dynamics. *Human molecular genetics*, 12 (14):1671–1687, 2003.

Swidbert R Ott and Stephen M Rogers. Gregarious desert locusts have substantially larger brains with altered proportions compared with the solitarious phase. *Proceedings of the Royal Society of*

London B: Biological Sciences, 277(1697): 3087–3096, 2010.

Swidbert R Ott, Heleen Verlinden, Stephen M Rogers, Caroline H Brighton, Pei Shan Quah, Rut K Vleugels, Rik Verdonck, and Jozef Vanden Broeck. Critical role for protein kinase a in the acquisition of gregarious behavior in the desert locust. *Proceedings of the National Academy of Sciences*, 109(7):E381–E387, 2012.

ML Pan and William H Telfer. Storage hexamer utilization in two lepidopterans: differences correlated with the timing of egg formation. *Journal of Insect Science*, 1(1):2, 2001.

Pedgley and David. Desert locust forecasting manual (volume 1 of 2). 1981.

JHF Pedra, A Brandt, R Westerman, N Lobo, H-M Li, J Romero-Severson, LL Murdock, and BR Pittendrigh. Transcriptome analysis of the cowpea weevil bruchid: identification of putative proteinases and α-amylases associated with food breakdown. *Insect molecular biology*, 12 (4):405–412, 2003.

Paolo Pelosi. Perireceptor events in olfaction. *Journal of neurobiology*, 30(1):3–19, 1996.

Paolo Pelosi and Rosario Maida. Odorant-binding proteins in insects. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 111 (3):503–514, 1995.

Paolo Pelosi, Mariantonietta Calvello, and Liping Ban. Diversity of odorant-binding proteins and chemosensory proteins in insects. *Chemical senses*, 30(suppl 1): i291–i292, 2005.

Meir Paul Pener and Stephen J Simpson. Locust phase polyphenism: an update. *Advances in Insect Physiology*, 36:1–272, 2009.

H Sofia Pereira and Marla B Sokolowski. Mutations in the larval foraging gene affect adult locomotory behavior after feeding in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences*, 90(11):5044–5046, 1993.

Michael W Pfaffl. A new mathematical model for relative quantification in real-time rt–pcr. *Nucleic acids research*, 29 (9):e45–e45, 2001.

Daniel Phaneuf, Y Labelle, D Berube, K Arden, W Cavenee, R Gagné, and RM Tanguay. Cloning and expression of the cdna encoding human fumarylacetoacetate hydrolase, the enzyme deficient in hereditary tyrosinemia: assignment of the gene to chromosome 15. *American journal of human genetics*, 48(3):525, 1991.

Jean-François Picimbon, Karen Dietrich, Heinz Breer, and Jürgen Krieger. Chemosensory proteins of *Locusta migratoria* (orthoptera: Acrididae). *Insect biochemistry and molecular biology*, 30(3):233–241, 2000.

Kristen L Pierce, Richard T Premont, and Robert J Lefkowitz. Seven-transmembrane receptors. *Nature reviews Molecular cell biology*, 3(9):639–650, 2002.

CW Pikielny, G Hasan, F Rouyer, and M Rosbash. Members of a family of drosophila putative odorant-binding proteins are expressed in different subsets of olfactory hairs. *Neuron*, 12(1):35–49, 1994.

Yong Ping and Susan Tsunoda. Inactivity-induced increase in nachrs upregulates shal k+ channels to stabilize synaptic potentials. *Nature neuroscience*, 15(1): 90–97, 2012.

Robert M Pitman. Transmitter substances in insects: a review. *Comparative and general pharmacology*, 2(7):347–371, 1971.

Marc S Pittman, Hilary C Robinson, and Robert K Poole. A bacterial glutathione transporter (escherichia coli cyddc) exports reductant to the periplasm. *Journal of Biological Chemistry*, 280(37):32254–32261, 2005.

Maaike S Pols and Judith Klumperman. Trafficking and function of the tetraspanin cd63. *Experimental cell research*, 315 (9):1584–1592, 2009.

Catherine J Potrikus and John A Breznak. Gut bacteria recycle uric acid nitrogen in termites: a strategy for nutrient conservation. *Proceedings of the National Academy of Sciences*, 78(7):4601–4605, 1981.

Yi Qin, Christopher Capaldo, Barry M Gumbiner, and Ian G Macara. The mammalian scribble polarity protein regulates epithelial cell adhesion and migration through e-cadherin. *The Journal of cell biology*, 171(6):1061–1071, 2005.

Mazibur M Rahman, Luc Vanden Bosch, Murshida Begum, Karsten Seidelmann, Michael Breuer, and Arnold De Loof. Phase-related 6-kda peptide titre in haemolymph of larvae and adult *schistocerca gregaria* and its role in yellow-protein synthesis. *Physiological Entomology*, 33 (2):123–128, 2008.

MM Rahman, L Vanden Bosch, Geert Baggerman, Elke Clynen, Korneel Hens, Bruno Hoste, Karen Meylaers, Tom Vercammen, Liliane Schoofs, and Arnold De Loof. Search for peptidic molecular markers in hemolymph of crowd-(gregarious) and isolated-reared (solitary) desert locusts, *Schistocerca gregaria*. *Peptides*, 23(11):1907–1914, 2002.

MM Rahman, Anick Vandingenen, M Begum, Michael Breuer, Arnold De Loof, and Roger Huybrechts. Search for phase specific genes in the brain of desert locust, *Schistocerca gregaria* (orthoptera: Acrididae) by differential display polymerase chain reaction. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, 135(2):221–228, 2003.

John M Rawls. Analysis of pyrimidine catabolism in *Drosophila melanogaster* using epistatic interactions with mutations of pyrimidine biosynthesis and β-alanine metabolism. *Genetics*, 172(3): 1665–1674, 2006.

B Regueiro, AMELUNXE. R, and S Grisolia. Synthesis and degradation of monohydroxytetra-hydronicotinamide adenine dinucleotide phosphate. *Physiological Chemistry and Physics*, 2(5): 445–&, 1970.

Sue Goo Rhee, Sang Won Kang, Woojin Jeong, Tong-Shin Chang, Kap-Seok Yang, and Hyun Ae Woo. Intracellular messenger function of hydrogen peroxide and its regulation by peroxiredoxins. *Current opinion in cell biology*, 17(2):183–189, 2005.

PG Richman and A Meister. Regulation of gamma-glutamyl-cysteine synthetase by nonallosteric feedback inhibition by glutathione. *Journal of Biological Chemistry*, 250(4):1422–1426, 1975.

Gordon Robertson, Jacqueline Schein, Readman Chiu, Richard Corbett, Matthew Field, Shaun D Jackman, Karen Mungall, Sam Lee, Hisanaga Mark Okada, and Jenny Q Qian. De novo assembly and analysis of rna-seq data. *Nature methods*, 7(11):909–912, 2010.

Matthew J Robinson, David Sancho, Emma C Slack, Salomé LeibundGut-Landmann, and Caetano Reis e Sousa. Myeloid c-type lectins in innate immunity. *Nature immunology*, 7(12): 1258–1265, 2006.

Peter Roessingh, Stephen J Simpson, and Samantha James. Analysis of phase-related changes in behaviour of desert locust nymphs. *Proceedings of the Royal Society of London B: Biological Sciences*, 252(1333):43–49, 1993.

Peter Roessingh, Abdelghani Bouaïchi, and Stephen J Simpson. Effects of sensory stimuli on the behavioural phase state of the desert locust, *Schistocerca gregaria*. *Journal of Insect Physiology*, 44(10):883–893, 1998.

Stephen M Rogers and Swidbert R Ott. Differential activation of serotonergic neurons during short-and long-term gregarization of desert locusts. *Proceedings of the Royal Society of London B: Biological Sciences*, 282(1800):20142062, 2015.

Stephen M Rogers, Thomas Matheson, Emma Despland, Timothy Dodgson, Malcolm Burrows, and Stephen J Simpson. Mechanosensory-induced behavioural gregarization in the desert locust *Schistocerca gregaria*. *Journal of Experimental Biology*, 206(22):3991–4002, 2003.

Stephen M Rogers, Darron A Cullen, Michael L Anstey, Malcolm Burrows, Emma Despland, Tim Dodgson, Tom Matheson, Swidbert R Ott, Katja Stettin, and Gregory A Sword. Rapid behavioural gregarization in the desert locust, *Schistocerca gregaria* entails synchronous changes in both activity and attraction to conspecifics. *Journal of insect physiology*, 65:9–26, 2014.

FJ Rohlf. tpsdig, version 1.40. *Department of Ecology and Evolution, State University of New York at Stony Brook*, 2004.

J-P Roussel. Modification des taux d'ecdysteroïdes selon la phase chez *Locusta migratoria* l. *Bulletin de la Société zoologique de France*, 118(4):367–373, 1993.

Mercedes Ruiz-Estévez, Mohammed Bakkali, Josefa Cabrero, Juan Pedro M Camacho, and María Dolores López-León. Hp1 knockdown is associated with abnormal condensation of almost all chromatin types in a grasshopper (*Eyprepocnemis plorans*). *Chromosome research*, 22(3):253–266, 2014.

A Sanchez-Gracia, FG Vieira, and J Rozas. Molecular evolution of the major chemosensory gene families in insects. *Heredity*, 103(3):208–216, 2009.

Howard A Schneiderman and Lawrence I Gilbert. Control of growth and development in insects. *Science*, 143(3604):325–333, 1964.

Frank Schnorrer, Cornelia Schönbauer, Christoph CH Langer, Georg Dietzl, Maria Novatchkova, Katharina Schernhuber, Michaela Fellner, Anna Azaryan, Martin Radolf, and Alexander Stark. Systematic genetic analysis of muscle morphogenesis and function in drosophila. *Nature*, 464 (7286):287–291, 2010.

Ralf R Schumann, Steven R Leong, Gail W Flaggs, Patrick W Gray, Samuel D Wright, John C Mathison, Peter S Tobias, and Richard J Ulevitch. Structure and function of lipopolysaccharide binding protein. *Science*, 249(4975):1429–1431, 1990.

Stephan C Schuster. Next-generation sequencing transforms today's biology. *Nature*, 200(8):16–18, 2007.

Ruchira Sen, Rhitoban Raychoudhury, Yunpeng Cai, Yijun Sun, Verena-Ulrike Lietze, Drion G Boucias, and Michael E Scharf. Differential impacts of juvenile hormone, soldier head extract and alternate caste phenotypes on host and symbiont transcriptome composition in the gut of the termite *Reticulitermes flavipes*. *BMC genomics*, 14(1):491, 2013.

Katharine J Sepp, Pengyu Hong, Sofia B Lizarraga, Judy S Liu, Luis A Mejia, Christopher A Walsh, and Norbert Perrimon. Identification of neural outgrowth genes using genome-wide rnai. *PLoS Genet*, 4 (7):e1000111, 2008.

MJP Shaw. Effects of population density on alienicolae of aphis fabae scop. *Annals of Applied Biology*, 65(2):205–212, 1970.

David Sheehan, Gerardene MEADE, and Vivienne M FOLEY. Structure, function and evolution of glutathione transferases: implications for classification of non-mammalian members of an ancient enzyme superfamily. *Biochemical Journal*, 360 (1):1–16, 2001.

Wangpeng Shi, Yang Guo, Chuan Xu, Shuqian Tan, Jing Miao, Yanjie Feng, Hong Zhao, Raymond J St Leger, and Weiguo Fang. Unveiling the mechanism by which microsporidian parasites prevent locust swarm behavior. *Proceedings of the National Academy of Sciences*, 111(4):1343–1348, 2014.

Roland J Siezen and Jack AM Leunissen. Subtilases: the superfamily of subtilisin-like serine proteases. *Protein Science*, 6 (3):501–523, 1997.

James P Simmer, Ruth E Kelly, Austin G Rinker, Barbara H Zimmermann, Joshua L Scully, Hyesook Kim, and David R Evans. Mammalian dihydroorotase: nucleotide sequence, peptide sequences, and evolution of the dihydroorotase domain of the multifunctional protein cad. *Proceedings of the National Academy of Sciences*, 87(1):174–178, 1990.

MSJ Simmonds and WM Blaney. Effects of rearing density on development and feeding behaviour in larvae of spodoptera exempta. *Journal of Insect Physiology*, 32 (12):1043–1053, 1986.

Gert Simonet, Bert Breugelmans, Paul Proost, Ilse Claeys, Jozef Van Damme, DE Arnold, and Jozef Vanden Broeck. Characterization of two novel pacifastin-like peptide precursor isoforms in the desert locust (*Schistocerca gregaria*): cdna cloning, functional analysis and real-time rt-pcr gene expression studies. *Biochemical Journal*, 388(1):281–289, 2005.

Jared T Simpson, Kim Wong, Shaun D Jackman, Jacqueline E Schein, Steven JM Jones, and Inanç Birol. Abyss: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–1123, 2009.

SJ Simpson, E Despland, BF Hägele, and T Dodgson. Gregarious behavior in desert locusts is evoked by touching their back legs. *Proceedings of the National Academy of Sciences*, 98(7):3895–3897, 2001.

SJ Simpson, D Raubenheimer, ST Behmer, A Whitworth, and GA Wright. A comparison of nutritional regulation in solitarious-and gregarious-phase nymphs of the desert locust *Schistocerca gregaria*. *Journal of Experimental Biology*, 205(1): 121–129, 2002.

Stephen J Simpson and Gregory A Sword. Locusts. *Current Biology*, 18(9):R364–R366, 2008.

Stephen J Simpson, ALAN McCaffery, and BERND F HAeGELE. A behavioural analysis of phase change in the desert locust. *Biological Reviews*, 74(4):461–480, 1999.

Patrício MV Simões, Jeremy E Niven, and Swidbert R Ott. Phenotypic transformation affects associative learning in the desert locust. *Current Biology*, 23(23):2407–2412, 2013.

Hilary A Snaith, Christopher G Armstrong, Yiquan Guo, Kim Kaiser, and PT Cohen. Deficiency of protein phosphatase 2a uncouples the nuclear and centrosome cycles and prevents attachment of microtubules to the kinetochore in drosophila microtubule star (mts) embryos. *Journal of Cell Science*, 109(13):3001–3012, 1996.

Jamie Snider, Guillaume Thibault, and Walid A Houry. The aaa+ superfamily of functionally diverse proteins. *Genome biology*, 9(4):1, 2008.

Hojun Song. Phylogenetic perspectives on the evolution of locust phase polyphenism. *Journal of Orthoptera Research*, 14 (2):235–245, 2005.

Hojun Song. Density-dependent phase polyphenism in nonmodel locusts: a minireview. *Psyche: A Journal of Entomology*, 2011, 2010.

Hojun Song and John W Wenzel. Phylogeny of bird-grasshopper subfamily cyrtacanthacridinae (orthoptera: Acrididae) and the evolution of locust phase polyphenism. *Cladistics*, 24(4):515–542, 2008.

Jornt Spit, Liesbeth Badisco, L Vergauwen, D Knapen, and Jozef Vanden Broeck. Microarray-based annotation of the gut transcriptome of the migratory locust, *Locusta migratoria*. *Insect Molecular Biology*, 2016.

David Stanley, Jon Miller, and Hasan Tunaz. Eicosanoid actions in insect immunity. *Journal of Innate Immunity*, 1(4):282–290, 2009.

Régis Stentz, Samantha Osborne, Nikki Horn, Arthur WH Li, Isabelle Hautefort, Roy Bongaerts, Marine Rouyer, Paul Bailey, Stephen B Shears, and Andrew M Hemmings. A bacterial homolog of a eukaryotic inositol phosphate signaling enzyme mediates cross-kingdom dialog in the mammalian gut. *Cell reports*, 6(4): 646–656, 2014.

Alexander Stephan, José María Mateos, Serguei V Kozlov, Paolo Cinelli, Andreas David Kistler, Stefan Hettwer, Thomas Rülicke, Peter Streit, Beat Kunz, and Peter Sonderegger. Neurotrypsin cleaves agrin locally at the synapse. *The FASEB Journal*, 22(6):1861–1873, 2008.

WJ Stower. The colour pattern of hoppers of the desert locust (*Schistocerca gregaria* forskal). *Anti-Locust Bull*, (32):75, 1959.

WJ Stower, DE Davies, and IB Jones. Morphometric studies of the desert locust, *Schistocerca gregaria* (forsk.). *The Journal of Animal Ecology*, pages 309–339, 1960.

Ryohei Sugahara, Shinjiro Saeki, Akiya Jouraku, Takahiro Shiotsuki, and Seiji Tanaka. Knockdown of the corazonin gene reveals its critical role in the control of gregarious characteristics in the desert locust. *Journal of insect physiology*, 79: 80–87, 2015.

Ryohei Sugahara, Seiji Tanaka, Akiya Jouraku, and Takahiro Shiotsuki. Two types of albino mutants in desert and migratory locusts are caused by gene defects in the same signaling pathway. *Gene*, 2017.

Ryohichi Sugimura and Linheng Li. Noncanonical wnt signaling in vertebrate development, stem cells, and diseases. *Birth Defects Research Part C: Embryo Today: Reviews*, 90(4):243–256, 2010.

Yukio Sugino, Hiroshi Teraoka, and Hideyo Shimono. Metabolism of deoxyribonucleotides i. purification and properties of deoxycytidine monophosphokinase of calf thymus. *Journal of Biological Chemistry*, 241(4):961–969, 1966.

Gregory A Sword and Stephen J Simpson. Is there an intraspecific role for density-dependent colour change in the desert locust? *Animal Behaviour*, 59(4):861–870, 2000.

Gregory A Sword, Stephen J Simpson, Ould Taleb M El Hadi, and Hans Wilps. Density–dependent aposematism in the desert locust. *Proceedings of the Royal Society of London B: Biological Sciences*, 267(1438): 63–68, 2000.

PM Symmons. A morphometric measure of phase in the desert locust, *Schistocerca gregaria* (forsk.). *Bulletin of Entomological Research*, 58(04):803–809, 1969.

Diego Sánchez, María D Ganfornina, Gabriel Gutiérrez, and Michael J Bastiani. Molecular characterization and phylogenetic relationships of a protein with potential oxygen-binding capabilities in the grasshopper embryo. a hemocyanin in insects? *Molecular biology and evolution*, 15 (4):415–426, 1998.

Kenneth Söderhäll and Lage Cerenius. Role of the prophenoloxidase-activating system in invertebrate immunity. *Current opinion in immunology*, 10(1):23–28, 1998.

Nobuhiro Takahashi, Toshiya Hayano, and Masanori Suzuki. Peptidyl-prolyl cis-trans isomerase is the cyclosporin a-binding protein cyclophilin. *Nature*, 337 (6206):473–475, 1989.

Seiji Tanaka and Yudai Nishide. Behavioral phase shift in nymphs of the desert locust, *Schistocerca gregaria*: special attention to attraction/avoidance behaviors and the role of serotonin. *Journal of insect physiology*, 59(1):101–112, 2013.

Seiji Tanaka and Shigemi Yagi. Evidence for the involvement of a neuropeptide in the control of body color in the desert locust, *Schistocerca gregaria*. *Japanese Journal of Applied Entomology and Zoology*, 65 (3):447–457, 1997.

Michael Tavaria, Tim Gabriele, Ismail Kola, and Robin L Anderson. A hitchhiker's guide to the human hsp70 family. *Cell stress & chaperones*, 1(1):23, 1996.

Amer I Tawfik. Hormonal control of the phase polyphenism of the desert locust: A review of current understanding. *Open Entomology Journal*, 6:22–41, 2012.

Amer I Tawfik and Frantisek Sehnal. A role for ecdysteroids in the phase polymorphism of the desert locust. *Physiological entomology*, 28(1):19–24, 2003.

Amer I Tawfik, Seiji Tanaka, Arnold De Loof, Liliane Schoofs, Geert Baggerman, Etienne Waelkens, Rita Derua, Yoram Milner, Yoram Yerushalmi, and M Paul Pener. Identification of the gregarization-associated dark-pigmentotropin in locusts through an albino mutant. *Proceedings of the National Academy of Sciences*, 96(12):7083–7087, 1999.

Amer I Tawfik, Roland Kellner, Klaus H Hoffmann, and Matthias W Lorenz. Purification, characterisation and titre of the haemolymph juvenile hormone binding proteins from *Schistocerca gregaria* and *Gryllus bimaculatus*. *Journal of insect physiology*, 52(3):255–268, 2006.

LR Taylor. Aggregation, variance and the mean. 1961.

Walter R Terra. The origin and functions of the insect peritrophic membrane and peritrophic gel. *Archives of insect biochemistry and physiology*, 47(2):47–61, 2001.

Gregory A Thompson and Alton Meister. Utilization of l-cystine by the gamma-glutamyl transpeptidase-gamma-glutamyl cyclotransferase pathway. *Proceedings of the National Academy of Sciences*, 72(6):1985–1988, 1975.

Travis Thomson, Niankun Liu, Alexey Arkov, Ruth Lehmann, and Paul Lasko. Isolation of new polar granule components in drosophila reveals p body and er associated proteins. *Mechanisms of development*, 125(9):865–873, 2008.

Jimmy R Thériault, Salamatu S Mambula, Tatsuya Sawamura, Mary Ann Stevenson, and Stuart K Calderwood. Extracellular hsp70 binding to surface receptors present on antigen presenting cells and endothelial/epithelial cells. *FEBS letters*, 579(9): 1951–1960, 2005.

Alisa Tietz, M Lindberg, and Eugene P Kennedy. A new pteridine-requiring enzyme system for the oxidation of glyceryl ethers. *Journal of Biological Chemistry*, 239(12): 4081–4090, 1964.

Jeffrey K Tong, Christian A Hassig, Gavin R Schnitzler, Robert E Kingston, and Stuart L Schreiber. Chromatin deacetylation by an atp-dependent nucleosome remodelling complex. *Nature*, 395(6705): 917–921, 1998.

Y Torrens, MC Montety, M Etr, JC Beaujouan, and J Glowinski. Tachykinin receptors of the nk1 type (substance p) coupled positively to phospholipase c on cortical astrocytes from the newborn mouse in primary culture. *Journal of neurochemistry*, 52(6):1913–1918, 1989.

Baldwyn Torto, Daniel Obeng-Ofori, Peter GN Njagi, Ahmed Hassanali, and Habert Amiani. Aggregation pheromone system of adult gregarious desert locust *Schistocerca gregaria* (forskal). *Journal of chemical ecology*, 20(7):1749–1762, 1994.

Thorsten Trimbuch and Christian Rosenmund. Should i stop or should i go? the role of complexin in neurotransmitter release. *Nature Reviews Neuroscience*, 17 (2):118–125, 2016.

Stephen C Trowell. High affinity juvenile hormone carrier proteins in the haemolymph of insects. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*, 103(4):795–807, 1992.

Victor I Tsetlin. Three-finger snake neurotoxins and ly6 proteins targeting nicotinic acetylcholine receptors: pharmacological tools and endogenous modulators. *Trends in pharmacological sciences*, 36 (2):109–123, 2015.

Neil D Tsutsui, Andrew V Suarez, Joseph C Spagna, and J Spencer Johnston. The evolution of genome size in ants. *BMC Evolutionary Biology*, 8(1):1, 2008.

Hasan Tunaz, Russell A Jurenka, and David W Stanley. Prostaglandin biosynthesis by fat body from true armyworms, pseudaletia unipuncta. *Insect biochemistry and molecular biology*, 31(4):435–444, 2001.

YKTT Uchida, K Izai, T Orii, and T Hashimoto. Novel fatty acid beta-oxidation enzymes in rat liver mitochondria. ii. purification and properties of enoyl-coenzyme a (coa) hydratase/3-hydroxyacyl-coa dehydrogenase/3-ketoacyl-coa thiolase trifunctional protein. *Journal of Biological Chemistry*, 267(2):1034–1041, 1992.

Rungrutai Udomsinprasert, Saengtong Pongjaroenkit, Jantana Wongsantichon, Aaron J Oakley, La-aied Prapanthadara, Matthew CJ Wilce, and Albert J Ketterman. Identification, characterization and structure of a new delta class glutathione transferase isoenzyme. *Biochemical Journal*, 388(3):763–771, 2005.

Boris Uvarov. *Grasshoppers and locusts. A handbook of general acridology. Volume 2. Behaviour, ecology, biogeography, population dynamics.* Centre for Overseas Pest Research., 1977.

Boris P Uvarov. A revision of the genus locusta, l.(= pachytylus, fieb.), with a new theory as to the periodicity and migrations of locusts. *Bulletin of entomological Research*, 12(02):135–163, 1921.

Matthias B Van Hiel, Pieter Van Wielendaele, Liesbet Temmerman, Sofie Van Soest, Kristel Vuerinckx, Roger Huybrechts, Jozef V Broeck, and Gert Simonet. Identification and validation of housekeeping genes in brains of the desert locust *Schistocerca gregaria* under different developmental conditions. *BMC molecular biology*, 10(1):56, 2009.

Jo Vandesompele, Katleen De Preter, Filip Pattyn, Bruce Poppe, Nadine Van Roy, Anne De Paepe, and Frank Speleman. Accurate normalization of real-time quantitative rt-pcr data by geometric averaging of multiple internal control genes. *Genome biology*, 3(7):1, 2002.

Jan A Veenstra. Isolation and structure of corazonin, a cardioactive peptide from the american cockroach. *Febs Letters*, 250(2): 231–234, 1989.

Filipe G Vieira and Julio Rozas. Comparative genomics of the odorant-binding and chemosensory protein gene families across the arthropoda: origin and evolutionary history of the chemosensory system. *Genome Biology and Evolution*, 3:476–490, 2011.

Vincenzo Viscosi and Andrea Cardini. Leaf morphology, taxonomy and geometric morphometrics: a simplified protocol for beginners. *PloS one*, 6(10):e25630, 2011.

Michel J Vos, Serena Carra, Bart Kanon, Floris Bosveld, Karin Klauke, Ody Sibon, and Harm H Kampinga. Specific protein homeostatic functions of small heat-shock proteins increase lifespan. *Aging cell*, 2015.

Z Waloff. upsurges and recessions of the desert locust plague; an historical survey. 1966.

Chong Wang, Yi Jiang, Jinming Ma, Huixian Wu, Daniel Wacker, Vsevolod Katritch, Gye Won Han, Wei Liu, Xi-Ping Huang, and Eyal Vardy. Structural basis for molecular recognition at serotonin receptors. *Science*, 340(6132):610–614, 2013.

Chutao Wang, Yueqing Cao, Zhongkang Wang, Youping Yin, Guoxiong Peng, Zhenlun Li, Hua Zhao, and Yuxian Xia. Differentially-expressed glycoproteins in *Locusta migratoria* hemolymph infected with *Metarhizium anisopliae*. *Journal of invertebrate pathology*, 96(3):230–236, 2007.

Qiang Wang, Xing Liu, Ye Cui, Yijun Tang, Wei Chen, Senlin Li, Huansha Yu, Youdong Pan, and Chen Wang. The e3 ubiquitin ligase amfr and insig1 bridge the activation of tbk1 kinase by modifying the adaptor sting. *Immunity*, 41(6):919–933, 2014a.

Xianhui Wang, Xiaodong Fang, Pengcheng Yang, Xuanting Jiang, Feng Jiang, Dejian Zhao, Bolei Li, Feng Cui, Jianing Wei, and Chuan Ma. The locust genome provides insight into swarm formation and long-distance flight. *Nature communications*, 5, 2014b.

Yanli Wang, Feng Jiang, Huimin Wang, Tianqi Song, Yuanyuan Wei, Meiling Yang, Jianzhen Zhang, and Le Kang. Evidence for the expression of abundant micrornas in the locust genome. *Scientific reports*, 5, 2015.

Ying Wang, Noushin Ghaffari, Charles D Johnson, Ulisses M Braga-Neto, Hui Wang, Rui Chen, and Huaijun Zhou. Evaluation of the coverage and depth of transcriptome by rna-seq in chickens. *BMC bioinformatics*, 12(10):S5, 2011.

Yuanyuan Wei, Shuang Chen, Pengcheng Yang, Zongyuan Ma, and Le Kang. Characterization and comparative profiling of the small rna transcriptomes in two phases of locust. *Genome biology*, 10(1):1, 2009.

William I Weis, Maureen E Taylor, and Kurt Drickamer. The c-type lectin superfamily in the immune system. *Immunological reviews*, 163(1):19–34, 1998.

Diana E Wheeler, Irina Tuchinskaya, Norman A Buck, and Bruce E Tabashnik. Hexameric storage proteins during metamorphosis and egg production in the diamondback moth, plutella xylostella (lepidoptera). *Journal of Insect Physiology*, 46 (6):951–958, 2000.

Charles W Whitfield, Yehuda Ben-Shahar, Charles Brillet, Isabelle Leoncini, Didier Crauser, Yves LeConte, Sandra Rodriguez-Zas, and Gene E Robinson. Genomic dissection of behavioral maturation in the honey bee. *Proceedings of the National Academy of Sciences*, 103(44): 16068–16075, 2006.

DW Whitman and AA Agrawal. What is phenotypic plasticity and why is it important? in 'phenotypic plasticity of insects'.(eds dw whitman and tn ananthakrishnan.) pp. 1–63, 2009.

BJR Whittle, J Lopez-Belmonte, and S Moncada. Regulation of gastric mucosal integrity by endogenous nitric oxide: interactions with prostanoids and sensory neuropeptides in the rat. *British journal of pharmacology*, 99(3):607–611, 1990.

Gabriele Wiesel, Sonja Tappermann, and August Dorn. Effects of juvenile hormone and juvenile hormone analogues on the phase behaviour of *Schistocerca gregaria* and *Locusta migratoria*. *Journal of insect physiology*, 42(4):385–395, 1996.

Martijn J Wilmer, Leo AJ Kluijtmans, Thea J Van Der Velden, Peter H Willems, Peter G Scheffer, Rosalinde Masereeuw, Leo A Monnens, Lambertus P Van den Heuvel, and Elena N Levtchenko. Cysteamine restores glutathione redox status in cultured cystinotic proximal tubular epithelial cells. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1812(6):643–651, 2011.

Patrick John Wilmore and Andrea Kay Brown. Molecular properties of orthopteran dna. *Chromosoma*, 51(4):337–345, 1975.

Kenneth Wilson, Sheena C Cotter, Andrew F Reeson, and Judith K Pell. Melanism and disease resistance in insects. *Ecology Letters*, 4(6):637–649, 2001.

Kenneth Wilson, Matthew B Thomas, Simon Blanford, Matthew Doggett, Stephen J Simpson, and Sarah L Moore. Coping with crowds: density-dependent disease resistance in desert locusts. *Proceedings of the National Academy of Sciences*, 99(8):5471–5475, 2002.

Rui Wu, Zeming Wu, Xianhui Wang, Pengcheng Yang, Dan Yu, Chunxia Zhao, Guowang Xu, and Le Kang. Metabolomic analysis reveals that carnitines are key regulatory metabolites in phase transition of the locusts. *Proceedings of the National Academy of Sciences*, 109(9):3259–3263, 2012.

Grith Bacher Wybrandt and Svend Olav Andersen. Purification and sequence determination of a yellow protein from sexually mature males of the desert locust, *Schistocerca gregaria*. *Insect biochemistry and molecular biology*, 31(12):1183–1189, 2001.

Inga Wójtowicz, Jadwiga Jabłońska, Monika Zmojdzian, Ouarda Taghli-Lamallem, Yoan Renaud, Guillaume Junion, Malgorzata Daczewska, Sven Huelsmann, Krzysztof Jagla, and Teresa Jagla. Drosophila small heat shock protein cryab ensures structural integrity of developing muscles, and proper muscle and heart performance. *Development*, 142(5):994–1005, 2015.

Yong Xia, Ah-Lim Tsai, Vladimir Berka, and Jay L Zweier. Superoxide generation from endothelial nitric-oxide synthase a ca2+/calmodulin-dependent and tetrahydrobiopterin regulatory process. *Journal of Biological Chemistry*, 273(40):25804–25808, 1998.

Chen Xiaoming, Feng Ying, Zhang Hong, and Chen Zhiyong. Review of the nutritive value of edible insects. *Forest insects as food: humans bite back*, page 85, 2010.

Chaoyong Xie, Jiao Yuan, Hui Li, Ming Li, Guoguang Zhao, Dechao Bu, Weimin Zhu, Wei Wu, Runsheng Chen, and Yi Zhao. Noncodev4: exploring the world of long non-coding rna genes. *Nucleic acids research*, 42(D1):D98–D103, 2014.

Ya-Long Xu, Peng He, Lan Zhang, Shao-Qing Fang, Shuang-Lin Dong, Yong-Jun Zhang, and Fei Li. Large-scale identification of odorant-binding proteins and chemosensory proteins from expressed sequence tags in insects. *BMC genomics*, 10 (1):1, 2009.

Meiling Yang, Yuanyuan Wei, Feng Jiang, Yanli Wang, Xiaojiao Guo, Jing He, and Le Kang. Microrna-133 inhibits behavioral aggregation by controlling dopamine synthesis in locusts. *PLoS Genet*, 10(2): e1004206, 2014.

Yuanxue Yang, Lixin Huang, Yunchao Wang, Yixi Zhang, Siqi Fang, and Zewen Liu. No cross-resistance between imidacloprid and pymetrozine in the brown planthopper: status and mechanisms. *Pesticide biochemistry and physiology*, 130:79–83, 2016.

Frank Yates. Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, 1(2):217–235, 1934.

Michael Zasloff. Antimicrobial peptides of multicellular organisms. *nature*, 415 (6870):389–395, 2002.

Agnieszka Zdybicka-Barabas, Sylwia Staczek, Paweł Mak, Krzysztof Skrzypiec, Ewaryst Mendyk, and Małgorzata Cytryńska. Synergistic action of galleria mellonella apolipophorin iii and lysozyme against gram-negative bacteria. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1828(6):1449–1456, 2013.

Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829, 2008.

Zhengyi Zhang, Zhi-Yu Peng, Kang Yi, Yanbing Cheng, and Yuxian Xia. Identification of representative genes of the central nervous system of the locust, *Locusta migratoria* manilensis by deep sequencing. *Journal of insect science*, 12(1):86, 2012.

Yuguang Zhao, Tomas Malinauskas, Karl Harlos, and E Yvonne Jones. Structural insights into the inhibition of wnt signaling by cancer antigen 5t4/wnt-activated inhibitory factor 1. *Structure*, 22(4):612–620, 2014.

Xuguo Zhou, Matthew R Tarver, and Michael E Scharf. Hexamerin-based regulation of juvenile hormone-dependent gene expression underlies phenotypic plasticity in a social insect. *Development*, 134(3): 601–610, 2007.

Yifei Zhu, Yang Wang, Maureen J Gorman, Haobo Jiang, and Michael R Kanost. Manduca sexta serpin-3 regulates prophenoloxidase activation in response to infection by inhibiting prophenoloxidase-activating proteinases. *Journal of Biological Chemistry*, 278(47):46556–46564, 2003.