Naturaleza y origen de los cromosomas B de *Locusta migratoria*

Universidad de Granada
Departamento de Genética

Francisco J. Ruiz-Ruano Campaña
Tesis Doctoral
Granada 2016

Editor: Universidad de Granada. Tesis Doctorales
Autor: Francisco Jesús Ruiz-Ruano Campaña
ISBN: 978-84-9125-984-8
URI: http://hdl.handle.net/10481/44087

Naturaleza y origen de los cromosomas B de *Locusta migratoria*

Memoria de Tesis Doctoral para optar al grado de Doctor
por la Universidad de Granada presentada por el licenciado
D. Francisco J. Ruiz-Ruano Campaña

Dirigida por el Doctor:

Dr. Juan Pedro Martínez Camacho

El doctorando Francisco J. Ruiz-Ruano Campaña y el director de la tesis Juan Pedro Martínez Camacho garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo dirección de su director de 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 publicaciones

Granada, a 19 de mayo de 2016

Director de la tesis	Doctorando
Fdo.: Dr. Juan Pedro Martínez Camacho	Fdo.: D. Francisco J. Ruiz-Ruano Campaña

El presente trabajo se ha desarrollado en el Grupo de Genética Evolutiva del Departamento de Genética de la Universidad de Granada. La investigación ha sido financiada por la Consejería de Economía, Innovación, Ciencia y Empleo de la Junta de Andalucía a través del Proyecto de Investigación de Excelencia P10-CVI-6649.

Durante la realización de este proyecto de Tesis Doctoral he disfrutado de una Beca Predoctoral de Formación de Personal Docente e Investigador (FPDI) de la Consejería de Economía, Innovación, Ciencia y Empleo de la Junta de Andalucía asociada al Proyecto de Investigación de Excelencia P10-CVI-6649.

Índice | Table of Contents

Resumen

El saltamontes *Locusta migratoria* es causante de una de las plagas más devastadoras sobre la agricultura: las plagas de langosta. Su genoma se caracteriza por tener uno de los sistemas de cromosomas B más ampliamente distribuidos de todos los estudiados en plantas y animales, puesto que se ha encontrado en poblaciones naturales de Asia, Africa, Australia y Europa. Con anterioridad al presente trabajo, se sabía que el cromosoma B de *L. migratoria* contiene genes para las histonas H3 y H4. La comparación de la secuencia de estos genes entre los cromosomas A y B ha sugerido que el cromosoma B probablemente derivó del autosoma 8, ya que éste es el único cromosoma A que muestra genes para las histonas H3 y H4. La divergencia en secuencia para estos genes entre cromosomas A y B proporcionó una primera estima de la edad mínima del cromosoma B (750.000 años). El objetivo general de esta tesis doctoral es profundizar en el conocimiento del origen y evolución de los cromosomas B de *L. migratoria* mediante el análisis en profundidad de su contenido molecular, tanto de ADN repetitivo como de genes codificadores de proteínas, que puede ser indicativo de su posible funcionalidad.

Para ello, llevamos a cabo secuenciación masiva de ADN genómico y ARN obtenidos de individuos portadores y no portadores de cromosomas B procedentes de poblaciones naturales localizadas en Padul (Granada) y Los Barrios (Cádiz), y complementamos esta información con secuencias de ADN y ARN obtenidas de individuos de China disponibles en las bases de datos públicas. Para probar las nuevas metodologías desarrolladas en el transcurso de esta tesis, también hemos llevado a cabo análisis genómicos en los saltamontes *Eumigus monticola, Eyprepocnemis plorans* y *Oedaleus decorus*.

En el primer capítulo, realizamos un estudio genómico y citogenético de los microsatélites en los cromosomas A y B de *L. migratoria* y *E. plorans*. Para ello, combinamos la información de su localización cromosómica, aportada por los mapeos físicos de una selección de microsatélites, con el análisis bioinformático de lecturas genómicas obtenidas por pirosecuenciación 454 para estimar la abundancia de todos los microsatélites y su asociación con otros elementos repetidos. Así hemos determinado que los microsatélites constituyen el 0.54 % y el 0.46 % de los genomas de *L. migratoria* y *E. plorans*, respectivamente. Están localizados tanto en regiones eucromáticas no codificantes, en su mayoría próximas a elementos transponibles, como en espaciadores de histonas y el espaciador intergénico del ADN ribosómico 45S. Dado que *L. migratoria* tiene un genoma más pequeño y rico en microsatélites que *E. plorans*, está claro que el contenido

en microsatélites no explica las diferencias en tamaño genómico entre estas dos especies, ni el tamaño gigante de los genomas de saltamontes. Los cromosomas B de ambas especies son heterocromáticos y resultaron estar empobrecidos en microsatélites. El cromosoma B de *L. migratoria* tiene más concentración de microsatélites en el tercio eucromático proximal, coincidiendo con la localización del cluster de histonas. Por otro lado, en el cromosoma B de *E. plorans* los microsatélites se localizan, sobre todo, en las bandas intersticiales DAPI^{-} y en el IGS del ARNr 45S.

El ADN satélite es uno de los principales componentes, aunque el gran desconocido, de los genomas eucarióticos. En el segundo capítulo, desarrollamos una serie de herramientas de análisis para descubrir el mayor número posible de ADNs satélite en el genoma de *L. migratoria*. Hasta el momento no se habían descrito secuencias de este tipo en esta especie mediante procedimientos *in vitro*, por lo que hemos desarrollado satMiner, un protocolo de análisis *in silico* para ensamblar y describir los ADNs satélites de un genoma. Proponemos el término satelitoma para denominar al conjunto de familias de ADN satélite presentes en el genoma. Concretamente, el satelitoma de *L. migratoria* consta de 62 familias de ADN satélite con longitudes de monómero entre 5 y 400 pb. El análisis en profundidad del satelitoma de esta especie, combinando el análisis bioinformático con el mapeo físico mediante FISH, nos ha llevado a proponer un modelo de evolución del ADN satélite, por el que éste se origina mediante duplicación segmental en una localización discreta del genoma, luego se disemina a otras localizaciones genómicas de cromosomas homólogos y no homólogos, y finalmente se clusteriza mediante amplificación local. Sólo después del último paso los ADNs satélites son visibles mediante FISH. En bacterias existen repeticiones en tándem variables en número (VNTRs) que muestran propiedades similares a los ADNs satélites no clusterizados de eucariotas, sugiriendo que muchas propiedades de estos elementos repetitivos en tándem son comunes a todos los seres vivos.

Para probar los procedimientos de análisis desarrollados en el capítulo anterior, hemos analizado también el satelitoma en individuos con y sin cromosomas B del saltamontes *Eumigus monticola*, una especie que también muestra cromosomas B mitóticamente inestables. En este tercer capítulo, utilizamos marcadores citogenéticos clásicos, como bandas C, tinción DAPI y FISH para el gen de la histona H3 y para el ADNr 5S, y a la vez analizamos el satelitoma de esta especie. Mientras que las familias génicas dieron información muy pobre con respecto al origen del cromosoma B, el satelitoma fue decisivo. Está compuesto por 27 familias de ADN satélite, con monómeros entre 5 y 325 pb. El octavo cromosoma en tamaño (S8, también denominado megamérico) es el cromosoma del genoma A que contiene más ADN satélite (13 familias), y el cromosoma B contiene seis de ellos, todos localizados en el tercio proximal del S8. Además, el cromosoma B contiene el ADNr 5S que también está presente en el S8, pero no lleva los genes de histonas de este último. Todo esto indica que el cromosoma B de esta especie se originó a partir del tercio proximal del autosoma S8, incluyendo el ADN satélite EmoSat11-122 pero no los genes para histonas. Finalmente, la formación *de novo* de dos ADNs satélites en el cromosoma B, que solo muestran señales de FISH en el cromosoma B, sugiere que la presencia de ADNs satélites específicos de un cromosoma B no apoyan necesariamente el origen interespecífico de éste.

En el cuarto capítulo abordamos el estudio del contenido en ADN repetitivo del cromosoma B de *L. migratoria*. Para ello, primero hemos realizado el ensamblaje y anotación de la familia de genes para histonas, que hasta el momento no había sido caracterizada en ortópteros. Esta familia génica tiene un tamaño total de 8.262 pb, con una estructura H1> <H3 H4> <H2A H2B> para los cinco genes. El espaciador 5 mostró algunas variantes estructurales características de los individuos con B. Una de ellas consistía en una serie de deleciones específicas del cromosoma B que permitieron diseñar un marcador molecular de la presencia del cromosoma B. Además, hemos determinado que los elementos repetitivos representan el 55 % del genoma 0B de *L. migratoria*. Restando *in silico* la abundancia de elementos repetitivos en las librerías genómicas con B y sin B hemos inferido que los cromosomas B están enriquecidos en genes para histonas y ADNsat, pero empobrecidos en elementos transponibles (ETs). Es notable que un solo ADN satellite (LmiSat02-176) representa el 70 % del ADN repetitivo del cromosoma B, y el análisis mediante FISH demostró que está localizado por toda la longitud del cromosoma B. Además, los experimentos de FISH demostraron la presencia de otros 6 ADNs satélites en el cromosoma B, estando todos ellos presentes en el autosoma S9. Esto apunta a que este cromosoma jugó un papel en el origen del cromosoma B, tal vez mediante una translocación con el cromosoma S8, el único cromosoma A que comparte genes para histonas con el B. El análisis *in silico* de la abundancia de ETs mostró que, a pesar de la escasez general de ETs, los cromosomas B están enriquecidos particularmente en algunos tipos. El mapeo mediante FISH para algunos de estos elementos mostró que están clusterizados en una región específica del cromosoma B, concretamente en la región heterocromática próxima al límite con la región eucromática. El uso de lecturas de pirosecuenciación 454 permitió ensamblar una secuencia quimérica de más de 17.000 pb que incluye fragmentos de 29 ETs distintos, lo que indica que esta región del cromosoma B es un sumidero de ETs. Finalmente, demostramos la expresión de algunos ETs portadores de SNPs específicos del B, sugiriendo que estas copias del cromosoma B están activas.

En el último capítulo, investigamos la presencia de genes codificadores

de proteínas en el cromosoma B de *L. migratoria*, desafiando la convención general que asume que los cromosomas B son elementos genéticamente inertes. Mediante análisis bioinformático de los genomas y transcriptomas con B y sin B, hemos localizado 24 genes en el cromosoma B, 12 de los cuales están completos. La comparación de la secuencia de estos genes en los genomas 0B y con B, usando como referencia los genes encontrados en *Oedaleus decorus*, que compartió el antecesor común más reciente con *L. migratoria* hace unos 23 millones de años, sugiere que la edad del cromosoma B está comprendida entre 1 y 4 millones de años. Pero lo más importante ha sido encontrar SNPs específicos del B en muchos genes, que indicaron que son transcritos activamente. Además, algunos de ellos podrían ser muy importantes para la transmisión del cromosoma B por estar involucrados en la regulación del ciclo celular. Uno de los genes más llamativos es *APC1*, que codifica para la subunidad mayor del complejo proteico promotor de la anafase, que promueve el paso de metafase a anafase durante la mitosis. Esto facilita que las células metafásicas donde sólo una de las cromátidas del B esté unida a las fibras del huso puedan progresar hacia anafase, debido al exceso de APC1, promoviendo así la no-disyunción mitótica del cromosoma B. Esto podría constituir la explicación para los dos mecanismos de acumulación conocidos para este cromosoma B. Además, el cromosoma B tiene activos otros dos genes que codifican para ubiquitin ligasas tipo E3, que pueden modificar postraduccionalmente la expresión de muchos genes de los cromosomas A. Esto demuestra que los cromosomas B son transcripcionalmente activos y que su contenido génico podría jugar un papel muy relevante en su éxito evolutivo. Esto los convierte en auténticos parásitos, capaces de manipular la expresión génica del genoma hospedador en su propio beneficio.

Summary

The grasshopper *Locusta migratoria* is responsible for one of the most devastating pests in agriculture: locust outbreaks. Its genome carries one of the most widespread B chromosome systems hitherto described in plants and animals, as they have been found in natural populations from Asia, Africa, Australia and Europe. Prior to this study, it was known that this B chromosome contains genes for H3 and H4 histone proteins. Sequence comparison of these genes between A and B chromosomes suggested that the B chromosome probably derived from autosome 8, because it is the only A chromosome carrying genes for H3 and H4 histones. Divergence between these gene sequences in A and B chromosomes provided a first estimate of a minimum age for the B chromosome (750,000 years). The most general objective of this thesis is to deepen the knowledge about the origin and evolution of B chromosomes in *L. migratoria*, through in-depth analysis of its molecular content for repetitive DNA and protein-coding genes, which might suggest their possible functionality.

For this purpose, we perform here next generation sequencing (NGS) of genomic DNA and RNA obtained from B-carrying and B-lacking individuals collected at two natural populations located in Padul (Granada) and Los Barrios (Cádiz), and complemented this information with that obtained from Chinese individuals from public databases. To test the new methodologies developed in the course of this thesis, we also performed genomic analyses in the grasshoppers *Eumigus monticola*, *Eyprepocnemis plorans* and *Oedaleus decorus*.

In the first chapter, we conducted a genomic and cytogenetic study of microsatellites on A and B chromosomes of *L. migratoria* and *E. plorans*. To do this, we combined information from chromosomal location, provided by the physical FISH mapping of selected microsatellites, together with the bioinformatic analysis of genomic reads obtained by 454 pyrosequencing, to estimate microsatellite abundance and their association with other repeat elements. In brief, microsatellites represent 0.54% and 0.46% of the *L. migratoria* and *E. plorans* genomes, respectively. They are located in noncoding euchromatic regions, mostly close to transposable elements, and in spacers of the histone gene cluster and the IGS of 45S rDNA. Since *L. migratoria* has a smaller and microsatellite-rich genome than *E. plorans*, it is clear that microsatellite content does not explain the differences in genome size between these two species, nor the huge size of gigantic grasshopper genomes. B chromosomes in both species are heterochromatic and impoverished in microsatellite content. The B chromosomes of *L. migratoria* show the highest concentration of microsatellites in the proximal euchromatic third, coinciding with the location of the histone gene cluster. On the other hand, the B chromosomes of *E. plorans* show microsatellites only in thin interstitial DAPI$^-$ bands and in the IGS of the 45S rDNA.

Satellite DNA (satDNA) is one of the major components, yet the great unknown, of eukaryotic genomes. In the second chapter, we develop a toolkit for throughput analysis of satDNA content in the genome of *L. migratoria*. So far, not a single satDNA had been found in this species through conventional *in vitro* methods, for which reason we have developed satMiner, an *in silico* analysis protocol to assemble and describe many satDNAs in a same genome. This has allowed us to define the satellitome, a term including all satDNA families in the genome. Specifically, the satellitome of *L. migratoria* consists of 62 satDNA families, with monomer length ranging between 5 and 400 bp. The throughput analysis of the satellitome in this species, combining the bioinformatic analysis with the physical mapping by FISH, has led us to propose a model for the evolution of satDNA, by which satDNAs are originated by segmental duplication in a given genome site, they then disseminate to other genomic locations on homologous and non-homologous chromosomes, and finally they become clustered by local amplification. Only after the last step, satDNAs are visible by FISH. Variable number tandem repeats (VNTR) in bacteria show similar properties to the non-clustered satDNAs in eukaryotes, suggesting that many satDNA properties are common to all living beings.

To test the analytical procedures developed in the former chapter, we have also analyzed the satellitome in individuals with and without B chromosomes of the grasshopper *Eumigus monticola*, a species also carrying mitotically unstable B chromosomes, like *L. migratoria*. In this third chapter, we use classical cytogenetic markers, such as C-banding, DAPI staining and FISH for histone H3 gene and 5S rDNA, and also analyze the satellitome of this species. Whereas the gene families gave very poor information about B chromosome origin, the satellitome was decisive. It is composed of 27 families of satDNA, with monomers ranging from 5 to 325 bp. The eighth autosome in size (S8, also called megameric) is the A chromosome carrying more satDNAs (13 families), and the B chromosome carries six of them, all located in the proximal third of the S8. In addition, the B chromosome contains 5S rDNA, which is also in the S8, but it does not carry the H3 histone genes carried by the latter. Taken together, this indicates that the B chromosome of this species originated from the proximal third of the S8 autosome, including the EmoSat11-122 satDNA but not the histone genes. Finally, the *de novo* formation of two satDNAs in the B chromosome, which only showed FISH signals on the B chromosome, calls attention on the fact that finding B-specific satDNA does not necessarily support the interespecific origin of B chromosomes.

In the fourth chapter, we analyze the content in repetitive DNA of the B chromosome in *L. migratoria*. To do this, we first performed the assembling and annotation of the histone gene family, which had not hitherto been characterized in Orthoptera. In this species, it showed a total size of 8,262 bp, with H1> <H3 H4> <H2A H2B> structure for the five genes and spacers. Spacer 5 showed several structural variants being specific to B-carrying individuals. One of them consisted of a series of deletions which allowed designing a molecular marker for B chromosome presence. In addition, we found that repetitive elements represent 55% of the *L. migratoria* 0B genome. By *in silico* substracting the abundance of repetitive elements in B-carrying and B-lacking genomic libraries, we inferred that B chromosomes are enriched in genes for histones and satDNA, but are impoverished in transposable elements (TEs). Remarkably, a single satDNA (LmiSat02-176) represents 70% of all repetitive DNA in the B chromosome, and FISH analysis showed that it is located across all B chromosome length. In addition, FISH experiments showed the presence of 6 other satDNAs on the B chromosome, all of which are also present in the S9 autosome. This suggests that this chromosome might have played a role in the origin of the B chromosome, perhaps through a translocation with chromosome S8, which is the only A chromosome sharing histone genes with the B. *In silico* analysis of TE abundance showed that B chromosomes are enriched in some particular TEs. FISH mapping for some of these TEs showed their clustering in a specific region of the B chromosome, i.e. an interstitial heterochromatic region close to the border with the euchromatic region. The use of 454 pyrosequencing reads allowed assembling a chimeric sequence of more than 17,000 bp including fragments of 29 different TEs, suggesting that this B chromosome region is a sink for TEs. We finally showed the transcription of some TEs carrying B-specific SNPs, suggesting that these B chromosome copies are active.

In the last chapter, we investigate the presence of protein-coding genes in the B chromosome of *L. migratoria*, to challenge the general convention that B chromosomes are genetically inert elements. By means of the bioinformatic analysis of genomes and transcriptomes of B-carrying and B-lacking individuals, we found 24 protein-coding genes in the B chromosome, 12 of which were complete. Sequence comparison of these genes in B-carrying and B-lacking genomes, using as reference the gene sequences found in *Oedaleus decorus*, a species sharing the most recent common ancestor with *L. migratoria* about 23 million years ago, suggested that the age of the B chromosome is between 1 and 4 million years. Most importantly, the finding of B-specific SNPs in many of these genes indicated that they are actively transcribed. In addition, some of them could be very important for B chromosome transmission due to their involvement in cell cycle reg-

ulation. One of the most striking genes is *APC1*, which encodes the largest subunit of the anaphase promoting complex, which promotes metaphase to anaphase progression during mitosis. This facilitates that metaphase cells where only one chromatid of the B chromosome is attached to the spindle fibers can progress to anaphase, due to the excess of APC1, thus promoting the mitotic nondisjunction of the B chromosome. This might constitute the basis for one of the two accumulation mechanisms known for this B chromosome. In addition, the B chromosome shows two active genes also encoding E3 ubiquitin ligases, which can post-translationally modify the expression of many A chromosome genes. This shows that B chromosomes are transcriptionally active and that their gene content may play a very relevant role in their evolutionary success. This become them into genuine parasites being able to manipulate gene expression in the host genome to their own benefit.

Introducción general y objetivos

Los cromosomas B

El genoma puede considerarse como el resultado de un mutualismo entre un conjunto de genes que contribuyen a originar individuos adaptados al ambiente. Pero, además de los genes cooperadores, los genomas de muchos organismos eucarióticos albergan también una enorme variedad de elementos genéticos cuya principal función es procurar su propia transmisión y multiplicarse en los genomas y en las poblaciones, generalmente a costa de otros genes y a pesar de ser, a menudo, deletéreos para el organismo. Éstos son elementos genéticos egoístas, que obtienen ventaja durante la transmisión al desobedecer las leyes mendelianas de la herencia. Entre ellos cabe destacar los transposones, los distorsionadores de la segregación, muchos factores citoplásmicos y los cromosomas B. Estos últimos fueron realmente los primeros elementos genéticos egoístas en ser descubiertos (Wilson, 1907), pero su naturaleza parasítica (Östergren, 1945) y egoísta (Jones, 1985) sólo fue reconocida muchos años después.

La invasión de los elementos genéticos egoístas, y los posibles efectos deletéreos que ello conlleva, genera el contexto para la selección, en el genoma hospedador, de variantes génicas que favorecen la resistencia al parásito, dando lugar a un conflicto intragenómico. En principio, cualquier sistema genético puede ser invadido por elementos genéticos egoístas, por lo que, durante los últimos años, ha ido obteniendo cada vez más apoyo la idea de que la existencia de elementos genéticos egoístas y el conflicto que crean puede ser una importante fuerza en la evolución de los sistemas genéticos, ampliándose continuamente el grupo de fenómenos que pueden ser considerados como el resultado de conflictos genéticos.

La reproducción sexual promueve la generación de conflictos genéticos porque la asociación de los alelos de cada locus es temporal y, dentro de ese proceso, la meiosis es un momento idóneo para la actuación de un gen o un cromosoma egoísta que promueva su propia transmisión asegurándose una mayor presencia en los gametos producidos por los individuos heterocigotos. A esta mayor tasa de transmisión se le llama, en general, distorsión de la segregación, y cuando la causa es claramente meiótica se le denomina impulso meiótico que ha sido observado en muchas especies.

La segregación desigual puede ocurrir porque el elemento genético egoísta inhabilita los gametos portadores de su alternativa alélica, tal como se ha demostrado en los sistemas mejor estudiados: Segregation distorter (*SD*) en *Drosophila melanogaster*, y el locus *t* en *Mus musculus*. Otra posibilidad es promover una conversión génica sesgada a favor de dicho elemento (ver Hurst *et al.* 1992). La manipulación del proceso meiótico es también una de las estrategias utilizadas por los cromosomas B, presentes en muchas especies de animales y plantas.

Los cromosomas B, también denominados accesorios o supernumerarios, son cromosomas adicionales y dispensables que no recombinan con los cromosomas A, por lo que puede decirse que siguen su propio camino evolutivo. Además, y ésta es la característica que les hace egoístas, los cromosomas B muestran comportamientos mitóticos y/o meióticos irregulares que suelen constituir la base de su acumulación en la línea germinal determinando modos de herencia no mendelianos que implican tasas de transmisión superiores a la de los cromosomas normales (0,5).

Los cromosomas B se han descrito en más de 1300 especies de plantas, en casi 500 especies de animales (para revisión, ver Jones & Rees 1982; Jones & Puertas 1993; Jones 1995; Camacho *et al.* 2004; Camacho 2005) y en varias especies de hongos (Mills *et al.*, 1990; Miao *et al.*, 1991a,b; Tzeng *et al.*, 1992; Geiser *et al.*, 1996). La frecuencia de un cromosoma B en poblaciones naturales depende de la interacción de un conjunto de factores direccionales (selección fenotípica e intragenómica), dispersivos (deriva genética) y contingentes (historia del polimorfismo).

Naturaleza

La mayoría de los cromosomas B son heterocromáticos y contienen principalmente ADN repetitivo, lo que sugiere que son elementos genéticamente inertes. Además, cuando se inyecta uridina tritiada a organismos vivos con B, ésta se incorpora en los cromosomas A pero no en los B, lo que sugiere que los cromosomas B no son transcripcionalmente activos (Fox *et al.*, 1974; Ishak *et al.*, 1991). Otra evidencia a favor de la baja actividad génica de los cromosomas B se ha obtenido recientemente al observar que los cromosomas B del saltamontes *Eyporepocnemis plorans* están hipoacetilados para la lisina 9 de la histona H3 durante toda la meiosis (Cabrero *et al.*, 2007). Sin embargo, existen algunos cromosomas B que muestran actividad transcripcional en el estado plumoso en la rana *Leiopelma hochstetteri* (Green, 1988) o en el estado politénico en el mosquito *Simulium juxtacrenobium* (Brockhouse *et al.*, 1989). En otros casos se ha demostrado que los cromosomas B son portadores de genes ribosómicos (para revisión, ver (Green, 1990; Beukeboom, 1994; Jones, 1995), aunque están casi siempre inactivos, e incluso existe un cromosoma B portador de un gen que confiere resistencia a un antibiótico, producido por la planta hospedadora, en el hongo *Nectria haematococca*, favoreciendo así su patogenicidad (Miao *et al.*, 1991a,b). Esto indica que no todos los cromosomas B son genéticamente inactivos. De hecho, se ha obtenido evidencia molecular de la transcripción de genes

ARNr presentes en los cromosomas B de la planta *Crepis capillaris* (Leach *et al.*, 2005).

En pocos caso se ha demostrado que la presencia de cromomosomas B está asociada con cambios de expresión génica. Éste es el caso de la planta *Scilla autumnalis* (Ruiz-Rejón *et al.*, 1980; Oliver *et al.*, 1982) y del roedor *Apodemus flavicollis* (Tanić *et al.*, 2005). En la planta *Scilla autumnalis* Ruiz-Rejón *et al.* (1980) y Oliver *et al.* (1982) demostraron la expresión diferencial de un gen para esterasas en los bulbos con B. En este mismo sentido, se ha obtenido evidencia de expresión génica diferencial, asociada a la presencia de cromosomas B, para tres genes en el roedor *Apodemus flavicollis* (Tanić *et al.*, 2005).

Los primeros análisis realizados durante los años setenta y ochenta demostraron que los cromosomas B contienen ADN esencialmente similar al de los cromosomas A (para revisión, ver Jones & Rees, 1982). Las investigaciones de los noventa han permitido aislar, clonar y secuenciar numerosos ADNs repetitivos localizados en los cromosomas B de varias especies; algunos de ellos son específicos de los cromosomas B mientras que otros son compartidos con los cromosomas A, especialmente el ADN ribosómico y los elementos transponibles. También se ha demostrado la presencia del proto-oncogen C-KIT en los cromosomas B de dos especies de cánidos (Graphodatsky *et al.*, 2005). Más recientemente, se ha demostrado que los transcritos de ARN ribosómico procedentes de un cromosoma B son funcionales (Ruiz-Estevez *et al.*, 2012), y se ha obtenido la primera evidencia de la transcripción de un gen para proteinas en un cromosoma B (Trifonov *et al.*, 2013). Además, la secuenciación masiva ha acelerado enormemente la investigación molecular de los cromosomas B, indicando que éstos son ricos en fragmentos génicos procedentes de los cromosomas A (Martis *et al.*, 2012; Valente *et al.*, 2014), y que muchos de ellos se transcriben (Banaei-Moghaddam *et al.*, 2013).

Origen

Como resultado de los estudios sobre la naturaleza molecular de los cromosomas B, se han propuesto dos teorías principales sobre su origen. La teoría más ampliamente aceptada sostiene que los cromosomas B se originan intraespecíficamente a partir de los cromosomas A (Jones & Rees, 1982), aunque existen casos de un posible origen interespecífico por hibridación. La mejor evidencia del origen intraespecífico de los cromosomas B es la existencia de familias de ADN repetitivo presentes tanto en los cromosomas A como en los B. Por ejemplo, todas las secuencias de ADN repetitivo aisladas por microdisección del cromosoma B de la planta *Crepis capillaris* están también presentes en los cromosomas A (Jamilena *et al.*, 1994, 1995). En el saltamontes *Eyprepocnemis plorans*, la ordenación de dos secuencias de ADN (un repetitivo de 180 pb y ADN ribosómico) respecto al centrómero sólo coincide con la del cromosoma X, lo que llevó a López-León *et al.* (1994) a proponer que, en esta especie, los cromosomas B derivaron de la región pericentromérica del X y la posterior amplificación de los dos tipos de secuencias allí contenidos. No obstante, nuestros estudios más recientes han rechazado esta hipótesis y apuntan hacia un origen autosómico de estos Bs (Teruel *et al.*, 2014). La formación de cromosomas B como resultado de hibridación interespecífica fue propuesta por Sapre & Deshpande (1987), y puede sospecharse a partir de la presencia de secuencias de ADN específicas de los cromosomas B que se encuentran en los cromosomas A de una especie afín. Este podría ser el caso de los cromosomas B de la avispa *Nasonia vitripennis* (Eickbush *et al.*, 1992). Se han obtenido evidencias directas del origen de algunos cromosomas B mediante hibridación interespecífica en el pez ginogenético *Poecilia formosa* (Schartl *et al.*, 1995) y en cruzamientos experimentales entre dos especies de avispas del género *Nasonia* (Perfectti & Werren, 2001).

Efectos

Existen numerosas evidencias de que los cromosomas B pueden afectar, tanto en plantas como en animales, a multitud de procesos celulares y fisiológicos. Los efectos raramente son apreciables en el fenotipo externo, con excepción de la planta *Haplopappus gracilis*, donde la presencia de un cromosoma B cambia el color de los aquenios (Jackson & Newmark, 1960), y el maíz donde las plantas con B muestran las hojas rayadas (Staub, 1987). Frecuentemente, los cromosomas B afectan negativamente a caracteres relacionados con la eficacia biológica, tales como el vigor, la fertilidad y la fecundidad. Estos efectos negativos sobre caracteres relacionados con la eficacia biológica de los individuos portadores confirman la naturaleza parasítica de los cromosomas B.

En algunos casos, los efectos de los cromosomas B pueden ser adscribibles directamente a los productos de genes presentes en ellos. Es el caso de los genes para la resistencia a la roya (enfermedad causada por un hongo) presentes en los Bs de *Avena sativa* (Dherawattana & Sadanaga, 1973) y los genes para resistencia a antibióticos en los Bs del hongo *Nectria haematococca* (Miao *et al.*, 1991a,b).

La presencia de cromosomas B provoca, en muchos casos, cambios en la frecuencia de quiasmas de los cromosomas A y, por tanto, la frecuencia de recombinación del hospedador. En la mayoría de los casos, la presencia de los cromosomas B está asociada con un incremento en la frecuencia de quiasmas, aunque se han descrito también casos de disminución y de ausencia de efecto (ver Jones & Rees 1982). Bell & Burt (1990) propusieron que los cromosomas B, como parásitos verticales de la línea germinal, inducen una respuesta adaptativa en el genoma hospedador caracterizada por el incremento en el número de quiasmas y, por tanto, en la variabilidad genética de su descendencia, alguna de la cual puede ser más resistente al parásito. Nuestro grupo obtuvo evidencias a favor de esta teoría (denominada de la recombinación inducible), ya que el incremento en la frecuencia de quiasmas es tanto mayor cuanta más acumulación tiene el cromosoma B (Camacho *et al.*, 2002).

Frecuentemente, los cromosomas B muestran tasas de transmisión netamente superiores a 0,5, es decir, poseen acumulación (ver Jones 1991). La acumulación puede ser premeiótica (como ocurre en *L. migratoria*), meiótica (en *Myrmeleotettix maculatus*) o postmeiótica (frecuente en plantas como el centeno), e incluso puede ser ameiótica (en Nasonia). En algunas especies, como el centeno (Puertas, 2002) y *L. migratoria* (Pardo *et al.*, 1994), se produce acumulación en los dos sexos. Pero no todos los cromosomas B muestran acumulación. Es el caso de los descritos en las plantas *Poa alpina*, *Poa trivialis*, *Centaurea scabiosa*, *Ranunculus acris*, *Allium schoenoprasum* y *Guizotia scabra* y el saltamontes *Eyprepocnemis plorans* en muchas poblaciones naturales (ver referencias en (Camacho *et al.*, 2000; Camacho, 2005). Como han demostrado nuestros estudios en *E. plorans*, los cromosomas B que no muestran acumulación pudieron haberla tenido anteriormente y perderla como consecuencia de su coevolución con los cromosomas A.

Evolución

Los dos modelos clásicos sobre la evolución de los cromosomas B, el modelo parasítico (Östergren 1945; también denominado egoísta; Jones 1985; Shaw & Hewitt 1990) y el modelo heterótico (White, 1977), asumen que la frecuencia de los cromosomas B está en equilibrio en las poblaciones actuales. Se diferencian en las fuerzas que consiguen ese equilibrio. Según el modelo egoísta, el equilibrio es el resultado de la acción de la acumulación del B (que afecta positivamente a su frecuencia) y de sus efectos perjudiciales sobre la eficacia biológica de los individuos portadores (que afectan negativamente a la frecuencia del B). El modelo heterótico, sin embargo, supone que el equilibrio es el resultado del efecto beneficioso de los Bs (que carecen de acumulación) sobre la eficacia biológica de los individuos portadores, cuando están en bajo número, y de su efecto negativo cuando están en números elevados. El primer modelo interpreta que los cromosomas B se comportan como parásitos, pero el heterótico propone que los cromosomas B son beneficiosos para los individuos que los albergan.

La gran mayoría de los sistemas de cromosomas B que se han analizado hasta ahora en profundidad son compatibles con el modelo parasítico (para revisión, ver Camacho *et al.* 2000). No obstante, también existen Bs que producen efectos beneficiosos sobre los individuos portadores. No cabe duda de que la resistencia a la roya conferida a la avena, o a la pisatina en *Nectria haematococca* representan un claro beneficio para los individuos con B en las poblaciones afectadas por esas enfermedades. Pero el único cromosoma B donde los análisis cuantitativos apuntan hacia su naturaleza heterótica es el del ajo silvestre *Allium schoenoprasum*, ya que carece de acumulación y se ha demostrado que las plantas con B sobreviven mejor en el hábitat natural que las plantas sin B en el desarrollo de semilla a plántula (Holmes & Bougourd, 1989).

A primera vista, podría parecer que todo cromosoma B que carezca de acumulación ha de ser necesariamente heterótico, ya que su permanencia en las poblaciones naturales sólo podría explicarse por selección a favor de los individuos portadores de pocos Bs. Pero eso no es necesariamente así, tal como han demostrado nuestras investigaciones sobre el polimorfismo para cromosomas B del saltamontes *Eyprepocnemis plorans*.

Las dos propiedades más notables de los Bs de *E. plorans* son la aparente ausencia de acumulación (López-León *et al.*, 1992) y de efectos sobre la eficacia biológica de los portadores (Lopez-Leon *et al.*, 1992; Camacho *et al.*, 1997). Este sistema de cromosomas B nos permitió desarrollar un tercer modelo de evolución de los cromosomas B, esencialmente derivado del modelo parasítico, y que propone que los cromosomas B pasan por etapas diferentes durante su coevolución con los cromosomas A, comenzando como elementos parásitos que muestran acumulación y que posteriormente son neutralizados por la evolución de genes supresores en los cromosomas A, lo que puede conducir a su eliminación o a su sustitución por un nuevo cromosoma B mutante que recupere la capacidad de acumulación (Camacho *et al.*, 1997; Herrera *et al.*, 1996; Zurita *et al.*, 1998; Bakkali *et al.*, 2002).

Este modelo ha proporcionado una nueva visión de los elementos genéticos egoístas, en general, puesto que pone de manifiesto los mecanismos de ataque y defensa que caracterizan la carrera de armamentos que estos elementos llevan a cabo con el genoma hospedador (Frank, 2000), y ha sido aplicado recientemente a otros sistemas de cromosomas B como, por ejemplo, los de maíz y centeno (Chiavarino *et al.*, 2001; Puertas, 2002; González-Sánchez *et al.*, 2003).

El sistema de cromosomas B de *Locusta migratoria*

De lo expuesto en el apartado anterior se deduce que los cromosomas B son parásitos genómicos en constante coevolución con el genoma estándar, un proceso que parece discurrir por caminos evolutivos diferentes en las dos especies objeto de nuestros estudios actuales. En *E. plorans*, los cromosomas B son mitóticamente estables (por lo que todas las células de un mismo individuo poseen el mismo número de Bs) y desarrollan un juego coevolutivo con los cromosomas A que ha llevado a la aparición, en ambos contendientes, de armas para la defensa y el ataque. Entre ellas, cabe destacar la acumulación de los Bs durante la meiosis femenina (Herrera *et al.*, 1996; Zurita *et al.*, 1998; Bakkali *et al.*, 2002), su elevada tasa de mutación, próxima a 0,5 % (López-León *et al.*, 1993; Bakkali & Camacho, 2004), y la capacidad de los cromosomas A para responder suprimiendo la acumulación de los Bs (Herrera *et al.*, 1996; Perfectti *et al.*, 2004). Lógicamente esto produce una variedad de situaciones diferentes entre poblaciones naturales, mostrando acumulación en algunas (Zurita *et al.*, 1998; Bakkali *et al.*, 2002) pero no en otras (López-León *et al.*, 1992; Camacho *et al.*, 2002). Dado que la etapa casi-neutra, donde la acumulación ha sido neutralizada, es muy larga, es más probable encontrar poblaciones donde los Bs no muestran acumulación, un hecho que coincide con nuestras observaciones.

En *L. migratoria*, sin embargo, el número de cromosomas B varía entre las células de un mismo individuo, es decir, son mitóticamente inestables. Esta inestabilidad es debida a la no disyunción mitótica de estos cromosomas durante el desarrollo embrionario que constituye un mecanismo de acumulación premeiótica en los machos de la especie (Nur, 1969; Cabrero *et al.*, 1984; Viseras *et al.*, 1990). Por esta razón, existe una gran variación intraindividual en el número de cromosomas B que portan las células de un individuo y parecen llevar a cabo un proceso coevolutivo más homogéneo entre poblaciones, dado que algunos aspectos, como la acumulación en los machos, son bastante similares en poblaciones muy alejadas como las españolas (Cabrero *et al.*, 1984) y las japonesas (Kayano, 1971). La existencia de acumulación en los dos sexos (Pardo *et al.*, 1994) y la elevada capacidad migratoria de esta especie explican la amplia distribución geográfica de los cromosomas B en *L. migratoria*, ya que se ha descrito su presencia en poblaciones naturales de Japón, (Itoh, 1934; Nur, 1969; Kayano, 1971), China (Hsiang, 1958), Mali (Dearn, 1974), Australia (King & John, 1980) y España (Cabrero *et al.*, 1984), es decir, en todos los continentes donde habita esta especie.

Los cromosomas B de *L. migratoria* están mucho menos estudiados que los de *E. plorans,* tanto en aspectos evolutivos como de su naturaleza y origen. La inestabilidad de los cromosomas B es ya patente en los embriones de 3 días, alcanzando su máximo al quinto día de desarrollo (Pardo *et al.*, 1995). En cruces controlados, sin embargo, se muestra que los cromosomas B también se acumulan por vía materna mediante la migración preferencial del cromosoma B hacia el oocito secundario durante la primera división meiótica (Pardo *et al.*, 1994). El mosaicismo resultante complica mucho el estudio de los efectos al nivel individual, aunque éstos pueden ser analizados al nivel celular. Esto ha mostrado patrones enigmáticos por los que la intensidad de los efectos depende de si los cromosomas B están en número par o impar para caracteres tales como la frecuencia de quiasmas, el número de NORs activas y la frecuencia de productos meióticos aberrantes durante la espermatogénesis (Camacho *et al.*, 2004). Los estudios de los efectos de los cromosomas B sugieren indirectamente que estos cromosomas podrían tener algún tipo de actividad génica.

Una forma de investigar este extremo es analizar la composición molecular de los cromosomas B e investigar la posible expresión de los genes o secuencias detectadas en ellos, así como la posible expresión diferencial de genes de los cromosomas A, asociada a la presencia de los cromosomas B.

El análisis citogenético de los cromosomas B de *L. migratoria* durante la espermatogénesis mostró que su mitad proximal tiene un carácter eucromático mientras que la mitad distal es heterocromática (Cabrero *et al.*, 1984). La técnica del bandeo C corrobora esta naturaleza heterocromática del extremo distal (Cabrero *et al.*, 1984), lo que es consistente con el patrón observado en las poblaciones españolas (Santos, 1980). El bandeo con fluorescencia también muestra que la zona distal de este cromosoma B es una región DAPI$^+$, lo que es usual en zonas heterocromáticas ricas en A+T, mientras que la región proximal es CMA3$^+$ que muestra que esta región es rica en ADN G+C (Camacho *et al.*, 1991).

Los primeros datos sobre la composición molecular del cromosoma B de *L. migratoria* se obtuvieron mediante la microdisección de este cromosoma realizada por Teruel *et al.* (2009), encontrando que los cromosomas B comparten al menos dos tipos de secuencias repetidas con los cromosomas A. Una de estas secuencias repetidas se localiza en la región distal y heterocrómatica del B y en las regiones pericentroméricas de la mitad de los cromosomas A, incluyendo el cromosoma X. El otro tipo de ADN repetido se localiza a lo largo de las regiones eucromáticas de todos los cromosomas A y en la zona eucromática proximal del cromosoma B, lo que sugiere que estas secuencias puedan corresponder con elementos móviles. Estos resultados apoyaban con fuerza el posible origen intraespecífico del cromosoma B de *L. migratoria*, pero no delimitaban cual de los cromosomas A sería el precursor del B.

El análisis, mediante FISH, de varios tipos de ADN repetitivo, ha permitido detectar la presencia de un bloque de genes para las histonas H3 y

H4 en la región proximal del cromosoma B. El análisis de la localización cromosómica de los genes para H3 y H4 en 35 especies de saltamontes, incluyendo *L. migratoria* (Cabrero *et al.*, 2009), ha revelado que en la mayoría de las especies existe un único bloque para H3-H4 localizado en el octavo autosoma en tamaño, y *L. migratoria* no es una excepción a este respecto. Por tanto, la presencia exclusiva de estos bloques de genes para H3 y H4 en los cromosomas 8 y B convierte al autosoma 8 en el mejor candidato para haber dado origen al cromosoma B en esta especie.

Por otra parte Teruel *et al.* (2010) amplificaron, mediante PCR con cebadores universales, el ADNr 45S y los genes para las histonas H3 y H4. Tanto el análisis mediante FISH como la ausencia de amplificación por PCR, indicaron que estos cromosomas B carecen de ADNr. En cambio, la amplificación por PCR de los genes para las histonas H3 y H4, mediante cebadores universales, a partir del ADN microdiseccionado de los cromosomas B, indicó la presencia de estas secuencias en el mismo. Una búsqueda en la base de datos NR mediante BLAST demostró que las secuencias del B eran similares a las de los genes para H3 y H4 de otros insectos, especialmente ortópteros. El análisis de la similitud entre las secuencias de las histonas H3 y H4 de los dos cromosomas corrobora este origen, y sugieren la inactividad de las copias del cromosoma B. Las estimas del tiempo de divergencia entre los genes de histonas de los cromosomas A y B indican, además, que los cromosomas B constituyen un polimorfismo antiguo en *L. migratoria* (más de 750.000 años) que demuestra la capacidad de los cromosomas B para mantenerse en las poblaciones naturales durante largos periodos de tiempo. Esta larga vida se debe, probablemente, a i) su carácter altamente egoísta, dado que muestra mecanismos de acumulación tanto en los machos como en las hembras, como dijimos anteriormente, y ii) a que estos cromosomas B no afectan severamente al nivel somático (Castro y col. 1998), y esto les hace más tolerables.

Objetivos

La hipótesis de partida de la presente tesis es que "los cromosomas B de *L. migratoria* derivaron del único cromosoma que contiene un bloque de genes para las histonas H3 y H4, es decir, del 8º autosoma en tamaño. Además, los cromosomas B parecen tener escasa actividad génica, y se espera que su efecto sobre la expresión génica de los cromosomas A sea pequeño". Sobre esta base, planteamos las siguientes preguntas:

Objetivos generales

1. Además de genes para histonas, ¿qué tipo de secuencias de ADN repetitivo se encuentran en los cromosomas B?

2. ¿Se originó el cromosoma B hace más de 750.000 años, como propusieron Teruel *et al.* (2010)?

3. ¿Existen genes codificadores de proteína en el cromosoma B?

4. ¿Muestran los cromosomas B actividad transcripcional para elementos repetitivos o para genes de proteínas?

Objetivos específicos

1. Estudiar la presencia, abundancia, diversidad y localización cromosómica de microsatélites en el genoma A y B de *L. migratoria* y de *E. plorans*, mediante aproximaciones genómicas y citogenéticas usando secuenciación masiva y FISH.

2. Desarrollar un protocolo de análisis bioinformático para conocer en profundidad el satelitoma de *L. migratoria*. Esto nos puede proporcionar una serie de marcadores que ayuden a desvelar el origen del B a partir de alguno de los cromosomas A. Además, disponer de una buena colección de ADN satélites puede ayudar a entender mejor su modo de evolución, que es muy importante para su utilización como marcadores genómicos.

3. Investigar el origen del cromosoma B mitóticamente inestable del saltamontes *Eumigus montícola*, mediante mapeo físico de varias familias génicas de ADN repetitivo y el análisis en profundidad del satelitoma.

4. Obtener la secuencia completa del cistrón de las histonas, que son los únicos elementos genómicos cuya presencia se conocía hasta ahora en el cromosoma B. La comparación de esta secuencia entre las copias autosómicas y las situadas en el cromosoma B nos proporcionará una visión más integral de los cambios producidos en las secuencias de los cromosomas A y B.

5. Identificar otros elementos de ADN repetitivo, además de los genes para histonas, que puedan ser informativos sobre la naturaleza molecular del cromosoma B, y puedan servir para probar la hipótesis sobre su origen a partir del autosoma 8. Si el cromosoma B se originó a partir de un único cromosoma A (el 8), estos cromosomas deberían

tener muchas secuencias de ADN en común. La comparación de esas secuencias permitirá arrojar más luz sobre la hipótesis del origen del cromosoma B.

6. Investigar la existencia de genes codificadores para proteínas en el cromosoma B. Conocer su estructura y comparar su secuencia entre las copias de los cromosomas A y B nos proporcionará una visión más integral de los cambios producidos en las secuencias del cromosoma B, una cuantificación más precisa de la magnitud de las diferencias acumuladas en las secuencias del cromosoma B y su A ancestral, y una mejor estimación del tiempo de divergencia entre estas secuencias y, en consecuencia, de la edad del cromosoma B.

7. Averiguar si los cromosomas B son genéticamente inertes, buscando SNPs específicos de los genomas con B y cuantificando su expresión.

Referencias

Bakkali M, Camacho JPM (2004) The B chromosome polymorphism of the grasshopper *Eyprepocnemis plorans* in North Africa: III. Mutation rate of B chromosomes. *Heredity*, **92**, 428–433.

Bakkali M, Perfectti F, Camacho JPM (2002) The B-chromosome polymorphism of the grasshopper *Eyprepocnemis plorans* in North Africa: II. Parasitic and neutralized B1 chromosomes. *Heredity*, **88**, 14–18.

Banaei-Moghaddam AM, Meier K, Karimi-Ashtiyani R, Houben A (2013) Formation and expression of pseudogenes on the B chromosome of rye. *The Plant Cell*, **25**, 2536–2544.

Bell G, Burt A (1990) B-chromosomes: germ-line parasites which induce changes in host recombination. *Parasitology*, **100**, S19–S26.

Beukeboom LW (1994) Bewildering Bs: an impression of the 1st B-chromosome conference. *Heredity*, **73**, 328–336.

Brockhouse C, Bass JAB, Feraday RM, Straus NA (1989) Supernumerary chromosome evolution in the *Simulium vernum* group (Diptera: Simuliidae). *Genome*, **32**, 516–521.

Cabrero J, López-León MD, Teruel M, Camacho JPM (2009) Chromosome mapping of H3 and H4 histone gene clusters in 35 species of acridid grasshoppers. *Chromosome research*, **17**, 397–404.

Cabrero J, Teruel M, Carmona FD, Jiménez R, Camacho JPM (2007) Histone H3 lysine 9 acetylation pattern suggests that X and B chromosomes are silenced during entire male meiosis in a grasshopper. *Cytogenetic and genome research*, **119**, 135–142.

Cabrero J, Viseras E, Camacho JPM (1984) The B-chromosomes of *Locusta migratoria* I. Detection of negative correlation between mean chiasma frequency and the rate of accumulation of the B's; a reanalysis of the available data about the transmission of these B-chromosomes. *Genetica*, **64**, 155–164.

Camacho JPM (2005) B chromosomes. *The Evolution of the Genome (ed. T. R. Gregory)*, pp. 223–286.

Camacho JPM, Bakkali M, Corral JM, *et al.* (2002) Host recombination is dependent on the degree of parasitism. *Proceedings of the Royal Society of London B: Biological Sciences*, **269**, 2173–2177.

Camacho JPM, Cabrero J, Viseras E, López-León MD, Navas-Castillo J, Alché JD (1991) G banding in two species of grasshopper and its relationship to C, N, and fluorescence banding techniques. *Genome*, **34**, 638–643.

Camacho JPM, López-León MD, Pardo MC, Cabrero J, Shaw MW (1997) Population dynamics of a selfish B chromosome neutralized by the standard genome in the grasshopper *Eyprepocnemis plorans*. *The American Naturalist*, **149**, 1030–1050.

Camacho JPM, Perfectti F, Teruel M, López-León MD, Cabrero J (2004) The odd-even effect in mitotically unstable B chromosomes in grasshoppers. *Cytogenetic and genome research*, **106**, 325–331.

Camacho JPM, Sharbel TF, Beukeboom LW (2000) B-chromosome evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **355**, 163–178.

Chiavarino AM, González-Sánchez M, Poggio L, Puertas MJ, Rosato M, Rosi P (2001) Is maize B chromosome preferential fertilization controlled by a single gene? *Heredity*, **86**, 743–748.

Dearn JM (1974) Phase transformation and chiasma frequency variation in locusts. *Chromosoma*, **45**, 321–338.

Dherawattana A, Sadanaga K (1973) Cytogenetics of a crown rust-resistant hexaploid oat with 42+ 2 fragment chromosomes. *Crop Science*, **13**, 591–594.

Eickbush DG, Eickbush TH, Werren JH (1992) Molecular characterization of repetitive DNA sequences from a B chromosome. *Chromosoma*, **101**, 575–583.

Fox DP, Hewitt GM, Hall DJ (1974) DNA replication and RNA transcription of euchromatic and heterochromatic chromosome regions during grasshopper meiosis. *Chromosoma*, **45**, 43–62.

Frank SA (2000) Polymorphism of attack and defense. *Trends in ecology & evolution*, **15**, 167–171.

Geiser DM, Arnold ML, Timberlake WE (1996) Wild chromosomal variants in *Aspergillus nidulans*. *Current genetics*, **29**, 293–300.

González-Sánchez M, Gonzalez-Gonzalez E, Molina F, Chiavarino AM, Rosato M, Puertas MJ (2003) One gene determines maize B chromosome accumulation by preferential fertilisation; another gene (s) determines their meiotic loss. *Heredity*, **90**, 122–129.

Graphodatsky AS, Kukekova AV, Yudkin DV, *et al.* (2005) The proto-oncogene C-KIT maps to canid B-chromosomes. *Chromosome Research*, **13**, 113–122.

Green DM (1988) Cytogenetics of the endemic New Zealand frog, *Leiopelma hochstetteri*: extraordinary supernumerary chromosome variation and a unique sex-chromosome system. *Chromosoma*, **97**, 55–70.

Green DM (1990) Muller's Ratchet and the evolution of supernumerary chromosomes. *Genome*, **33**, 818–824.

Herrera JA, López-León MD, Cabrero J, Shaw MW, Camacho JPM (1996) Evidence for B chromosome drive suppression in the grasshopper *Eyprepocnemis plorans*. *Heredity*, **76**.

Holmes DS, Bougourd SM (1989) B-chromosome selection in *Allium schoenoprasum*. I. Natural populations. *Heredity*, **63**, 83–87.

Hsiang W (1958) Cytological studies on migratory locust hybrid, *Locusta migratoria migratoria* L. *Locusta migratoria manilensis* Meyen. *Acta Zoologica Sinica*, **1**, 006.

Hurst GD, Hurst LD, Johnstone RA (1992) Intranuclear conflict and its role in evolution. *Trends in ecology & evolution*, **7**, 373–378.

Ishak B, Jaafar H, Maetz JL, Rumpler Y (1991) Absence of transcriptional activity of the B-chromosomes of *Apodemus peninsulae* during pachytene. *Chromosoma*, **100**, 278–281.

Itoh H (1934) Chromosomal variation in the spermatogenesis of a grasshopper, *Locusta danica*. *Jap. J. Genet*, **10**, 115–134.

Jackson RC, Newmark P (1960) Effects of supernumerary chromosomes on production of pigment in *Haplopappus gracilis*. *Science*, **132**, 1316–1317.

Jamilena M, Garrido-Ramos M, Rejon MR, Rejon CR, Parker JS (1995) Characterisation of repeated sequences from microdissected B chromosomes of *Crepis capillaris*. *Chromosoma*, **104**, 113–120.

Jamilena M, Rejón CR, Rejon MR (1994) A molecular analysis of the origin of the *Crepis capillaris* B chromosome. *Journal of Cell Science*, **107**, 703–708.

Jones RN (1985) Are B chromosomes selfish? In *The evolution of genome size (ed. T. Cavalier–Smith)*, vol. 30, pp. 397–425. Wiley, London.

Jones RN (1991) B-chromosome drive. *American Naturalist*, pp. 430–442.

Jones RN (1995) Tansley review no. 85. B chromosomes in plants. *New Phytologist*, pp. 411–434.

Jones RN, Puertas MJ (1993) The B-chromosomes of rye (*Secale cereale* L.). In *Frontiers in plant science research (K.K. Dhir y T.S. Sareen (eds.))*, pp. 81–112. Bhagwati Enterprises, Delhi.

Jones RN, Rees H (1982) *B Chromosomes*. Academic Press, New York.

Kayano H (1971) Accumulation of B chromosomes in the germ line of *Locusta migratoria*. *Heredity*.

King M, John B (1980) Regularities and restrictions governing C-band variation in acridoid grasshoppers. *Chromosoma*, **76**, 123–150.

Leach CR, Houben A, Field B, Pistrick K, Demidov D, Timmis JN (2005) Molecular evidence for transcription of genes on a B chromosome in *Crepis capillaris*. *Genetics*, **171**, 269–278.

López-León MD, Cabrero J, Camacho JPM, Cano MI, Santos JL (1992) A widespread B chromosome polymorphism maintained without apparent drive. *Evolution*, pp. 529–539.

López-León MD, Cabrero J, Pardo MC, Viseras E, Camacho JPM, Santos JL (1993) Generating high variability of B chromosomes in *Eyprepocnemis plorans* (grasshopper). *HEREDITY-LONDON-*, **71**, 352–352.

López-León MD, Neves N, Schwarzacher T, Heslop-Harrison JP, Hewitt GM, Camacho JPM (1994) Possible origin of a B chromosome deduced from its DNA composition using double FISH technique. *Chromosome Research*, **2**, 87–92.

Lopez-Leon MD, Pardo MC, Cabrero J, Camacho JPM (1992) Random mating and absence of sexual selection for B chromosomes in two natural populations of the grasshopper *Eyprepocnemis plorans*. *HEREDITY-LONDON-*, **69**, 558–558.

Martis MM, Klemme S, Banaei-Moghaddam AM, *et al.* (2012) Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proceedings of the National Academy of Sciences*, **109**, 13343–13346.

Miao VP, Covert SF, VanEtten HD (1991a) A fungal gene for antibiotic resistance on a dispensable ("B") chromosome. *Science*, **254**, 1773–1776.

Miao VP, Matthews DE, VanEtten HD (1991b) Identification and chromosomal locations of a family of cytochrome P-450 genes for pisatin detoxification in the fungus *Nectrla haematococca*. *Molecular and General Genetics MGG*, **226**, 214–223.

Mills D, McCluskey K, others (1990) Electrophoretic karyotypes of fungi: the new cytology. *Mol. Plant-Microbe Interact*, **3**, 351–357.

Nur U (1969) Mitotic instability leading to an accumulation of B-chromosomes in grasshoppers. *Chromosoma*, **27**, 1–19.

Oliver JL, Posse F, Martinez-Zapater JM, Enriquez AM, Ruiz-Rejón M (1982) B-Chromosomes and E-1 isozyme activity in mosaic bulbs of *Scilla autumnalis* (Liliaceae). *Chromosoma*, **85**, 399–403.

Östergren G (1945) Parasitic nature of extra fragment chromosomes. *Botaniska Notiser*, **2**, 157–163.

Pardo MC, López-León MD, Cabrero J, Camacho JPM (1994) Transmission analysis of mitotically unstable B chromosomes in *Locusta migratoria*. *Genome*, **37**, 1027–1034.

Pardo MC, López-León MD, Viseras E, Cabrero J, Camacho JPM (1995) Mitotic instability of B chromosomes during embryo development in *Locusta migratoria*. *Heredity*, **74**, 164–169.

Perfectti F, Corral JM, Mesa JA, *et al.* (2004) Rapid suppression of drive for a parasitic B chromosome. *Cytogenetic and genome research*, **106**, 338–343.

Perfectti F, Werren JH (2001) The interspecific origin of B chromosomes: experimental evidence. *Evolution*, **55**, 1069–1073.

Puertas MJ (2002) Nature and evolution of B chromosomes in plants: a non-coding but information-rich part of plant genomes. *Cytogenetic and genome research*, **96**, 198–205.

Ruiz-Estevez M, Lopez-Leon MD, Cabrero J, Camacho JPM (2012) B-chromosome ribosomal DNA is functional in the grasshopper *Eyprepocnemis plorans*. *PLoS One*, **7**, e36600.

Ruiz-Rejón M, Posse F, Oliver JL (1980) The B chromosome system of *Scilla autumnalis* (Liliaceae): effects at the isozyme level. *Chromosoma*, **79**, 341–348.

Santos JL (1980) *Variación de la heterocromatina constitutiva en el cariotipo de los Acridoidea y su efecto en el comportamiento cromosómico en meiosis*. Ph.D. thesis, Ph. D. thesis, Universidad Complutense de Madrid, Spain.

Sapre AB, Deshpande DS (1987) Origin of B chromosomes in *Coix* L. through spontaneous interspecific hybridization. *Journal of Heredity*, **78**, 191–196.

Schartl M, Nanda I, Schlupp I, *et al.* (1995) Incorporation of subgenomic amounts of DNA as compensation for mutational load in a gynogenetic fish. *Nature*, **373**, 68–71.

Shaw MW, Hewitt GM (1990) B chromosomes, selfish DNA and theoretical models: where next? In *Oxford Surveys in Evolutionary Biology (D. Futuyma y J. Antonovics (eds.))*, vol. 7, pp. 197–223. Oxford University Press.

Staub RW (1987) Leaf striping correlated with the presence of B chromosomes in maize. *Journal of Heredity*, **78**, 71–74.

Tanić N, Vujošević M, Dedović-Tanić N, Dimitrijević B (2005) Differential gene expression in yellow-necked mice *Apodemus flavicollis* (Rodentia, Mammalia) with and without B chromosomes. *Chromosoma*, **113**, 418–427.

Teruel M, Cabrero J, Montiel EE, Acosta MJ, Sánchez A, Camacho JPM (2009) Microdissection and chromosome painting of X and B chromosomes in *Locusta migratoria*. *Chromosome Research*, **17**, 11–18.

Teruel M, Cabrero J, Perfectti F, Camacho JPM (2010) B chromosome ancestry revealed by histone genes in the migratory locust. *Chromosoma*, **119**, 217–225.

Teruel M, Ruíz-Ruano FJ, Marchal JA, *et al.* (2014) Disparate molecular evolution of two types of repetitive DNAs in the genome of the grasshopper *Eyprepocnemis plorans*. *Heredity*, **112**, 531–542.

Trifonov VA, Dementyeva PV, Larkin DM, *et al.* (2013) Transcription of a protein-coding gene on B chromosomes of the Siberian roe deer (*Capreolus pygargus*). *BMC biology*, **11**, 1.

Tzeng TH, Lyngholm LK, Ford CF, Bronson CR (1992) A restriction fragment length polymorphism map and electrophoretic karyotype of the fungal maize pathogen *Cochliobolus heterostrophus*. *Genetics*, **130**, 81–96.

Valente GT, Conte MA, Fantinatti BE, *et al.* (2014) Origin and evolution of B chromosomes in the cichlid fish *Astatotilapia latifasciata* based on integrated genomic analyses. *Molecular biology and evolution*, p. msu148.

Viseras E, Camacho JPM, Cano MI, Santos JL (1990) Relationship between mitotic instability and accumulation of B chromosomes in males and females of *Locusta migratoria*. *Genome*, **33**, 23–29.

White MJD (1977) *Animal cytology and evolution*. CUP Archive.

Wilson EB (1907) The supernumerary chromosomes of Hemiptera. *Science*, **26**, 870–871.

Zurita S, Cabrero J, López-León MD, Camacho JPM (1998) Polymorphism regeneration for a neutralized selfish B chromosome. *Evolution*, pp. 274–277.

Material y métodos

Material

El material utilizado para la elaboración de este Tesis Doctoral ha consistido en machos de *L. migratoria* y otras especies de saltamontes. En el caso de *L. migratoria*, hemos muestreado en una población situada en Padul (Granada) y otras dos en Los Barrios (Cádiz), además de comprar animales en una tienda de mascotas. También hemos utilizado datos genómicos y transcriptómicos depositados en la base de datos de lecturas cortas (SRA) de individuos procedentes de China. También hemos utilizado material de otras especies de saltamontes como *Eyprepocnemis plorans*, procedentes de Torrox (Málaga) y Otívar (Granada), *E. monticola* en Hoya de la Mora (Granada) y de *Oedaleus decorus* en Capileira (Granada). Hemos utilizado material prodente de machos adultos, separando parte del testículo para análisis citogenético y el resto del cuerpo para análisis genómico y transcriptómico mediante secuenciación con las plataformas 454 e Illumina. También hemos obtenido embriones para análisis citogenéticos a partir de cultivos en nuestro laboratorio.

Métodos citogenéticos y moleculares

Fijación y caracterización de la presencia de cromosoma B

La fijación de testículo de saltamontes se ha realizado en una solución de etanol:ácido acético 3:1 preparada justo antes de su uso. El resto del cuerpo y testículo se fijó sumergiéndolo en nitrógeno líquido y almacenándolo a -80ºC. La presencia de cromosomas B se ha determinado realizando preparaciones en orceína lactopropiónica al 2 % por aplastamiento de varios folículos testiculares.

Los embriones para FISH se han obtenido a partir cultivos en nuestro laboratorio incubando la puesta a 28ºC durante 10 días para *L. migratoria* y 12 días para *E. plorans*. Los huevos fueron disecciondos en solución salina para insectos. Los embriones se fijaron según el protocolo descrito en Camacho *et al.* (1991).

Extracción de ADN y ARN

En caso de partes del cuerpo con cutícula congelamos el material con nitrógeno líquido para disgregarlo con un mortero de porcelana. Para tejidos blandos, como es el caso de testículos, disgregamos con un pistilo para eppedorfs de 1,5 mL el material en tampón de suspensión del kit de extracción. Para las extracciones de ADN genómico utilizamos el kit "Gen Elute Mamalian Genomic DNA Miniprep" (Sigma-Aldrich) siguiendo el protocolo del fabricante. Para las extracciones de ARN total de pata utilizamos el kit "Real Total RNA Spin Plus" (Durviz), mientras que para extracciones de ARN total de testículo utilizamos el kit "RNeasy Lipid Tissue" (Qiagen). Medimos la concentración del ADN y ARN extraidos mediante el espectrofotómetro "Infinite 200 NanoQuant" (Tecan). Además comprobamos la integridad del ADN corriendo 2 μL de muestra en un gel de agarosa al 1 %, mientras que para el ARN corrimos 2 μL en un gel MOPS con agua DEPC. Una vez comprobada la buena concentración y la calidad de las muestras enviamos al menos 1 ng a 50 ng/μL para su secuenciación.

Amplificación por PCR

Realizamos las reacciones de PCR en un volumen total de 25 μL en soliciones compuestas de tampón de PCR 1x, 2 mM de $MgCl_2$, 200 μM de dNTPs, 0.4 μM de cada cebador, 10 ng de ADN y 1 unidad de Taq polimerasa (MBL002). Los programas de PCR utilizados para elementos lineales comienzan con un paso de desnaturalización inicial a 95ºC durante 5 minutos, seguido de 30 ciclos a 94ºC durante 30 segundos, 55-60-65ºC como temperaturas de hibridación y 30 segundos de extensión a 72ºC, terminando con un paso de extensión final a 72ºC durante 7 minutos. Para secuencias repetidas en tándem con monómero mayor a 50 bp utilizamos 35 ciclos con la mismas temperaturas, pero con 20 segundos de desnaturalización, 40 segundos de hibridación y 20 segundos de extensión, reduciendo el tiempo de hibridación a 10 segundos en el caso de monómeros más cortos de 50 pb. En caso de que fuese necesario reamplificar el producto de PCR recortamos la banda del gel, la estrujamos en un cuadrado de parafilm y repetimos la PCR con 0,5 μL de la solución resultante. Los productos de PCR resultantes fueron utilizados en algunos casos para comprobar los ensamblajes realizados *in silico* y, en otros casos, para generar sondas para FISH.

Hibridación *in situ* a partir de producto de PCR

Los productos de PCR fueron marcados por "nick translation" con 2,5 unidades de ADN polimerasa I/DNasa I (Invitrogen), siguiendo el protocolo estándar, para ser utilizados como sondas para hibribridación *in situ*. Marcamos las sondas con tetrametilrodamina-5-dUTP o fluoresceina-11-dUTP (Roche), y realizamos el mapeo físico siguiendo el protocolo de Cabrero *et al.* (2003).

Hibridación *in situ* de microsatélites

Las preparaciones de embriones se realizaron según el protocolo descrito por Camacho *et al.* (1991). Se deshidrataron con etanol en series de 70 %, 90 % y 100 % de 3, 3 y 5 minutos respectivamente, y se dejaron secar. Posteriormente recibieron un pretratamiento con RNasa, se fijaron con paraformaldehído al 4 % y se deshidrataron con etanol en series antes de dejarse secar. Entre los dos tratamientos, las preparaciones fueron lavadas en 2xSSC como está previamente descrito en Cuadrado & Jouve (2007). Como sonda utilizamos oligonucleótidos sintéticos marcados con biotina en ambos extremos (Roche Applied Science). Realizamos la desnaturalización de cromosomas y sondas, así como la hibridación *in situ* según el protocolo descrito por Cuadrado *et al.* (2000). Brevemente, consiste en preparar una mezcla de hibridación añadiendo un 50 % de formamida desionizada, un 10 % de dextrán sulfato, 2x SSC, 0,1 % de SDS (dodecilsulfato sódico), 2 ppm de sonda de microsatelite. Para el lavado posterior a la hibridación, sumergimos las preparacioes en 4xSSC/0.2 %Tween-20 y las agitamos durante 10 minutos a temperatura ambiente. Para detectar la biotina incubamos las muestras en Estreptavidina-Cy3 (Sigma) con BSA al 5 % (p/v) durante una hora a 37ºC. Antes de la tinción con DAPI (4 ′,6-diamino-2-fenilindol), enjuagamos las muestras con 4xSSC/0.2 %Tween-20 a temperatura ambiente. Montamos las preparaciones en medio Vectashield antifading (Serva).

Métodos bioinformáticos

La mayoría de scripts desarrollados para esta tesis doctoral pueden obtenerse en mi repositorio de Github (https://github.com/fjruizruano/ngs-protocols) y protocolos de análisis más detallados en mi blog "A Little bioinformatician" (https://littlebioinformatician.wordpress.com). A continuación expongo una relación resumida de los métodos bioinformáticos más relevantes utilizados en esta tesis.

Ensamblaje a partir de lecturas de secuenciación de alto rendimiento

Ensamblaje y anotación a partir de lecturas 454

Realizamos una identificación *de novo* y modelado de secuencias repetitivas utilizando RepeatModeler en las lecturas 454. Este programa integra los análisis de RECON y RepeatScout (Benson, 1999; Bao & Eddy, 2002) además de anotar utilizando la base de datos RepBase (Jurka *et al.*, 2005).

Ensamblaje con una secuencia de referencia

Utilizando el programa MITObim (Hahn *et al.*, 2013) a partir de una secuencia de referencia conseguimos ensamblar la secuencia del elemento repetido con nuestros datos. En otras ocasiones sirvió para obtener una secuencia consenso mayoritario.

Clustering y ensamblaje

El programa RepeatExplorer (Novák *et al.*, 2013) agrupa lecturas con al menos un 80 % de identidad en un clúster. Posteriormente representa en un grafo las conexiones entre lecturas, lo cual nos puede dar una idea sobre la estructura de ese elemento en el genoma. Finalmente ensambla estas lecturas generando contigs que anota con Repbase y opcionalmente con una base de datos generada por nosotros.

Ensamblaje de ADNs satélites

Con el protocolo estándar de RepeatExplorer podemos caracterizar secuencias repetidas, incluido el ADN satélite. Sin embargo, está limitado por el número de lecturas capaz de utilizar, llegando un momento donde requiere mucho tiempo computacional. Por este motivo desarrollamos un procedimiento dentro de las herramientas de satMiner para filtrar utilizando DeconSeq los elementos repetidos que ensambla RepeatExplorer, pudiendo de nuevo ensamblar con este último para sacar nuevas secuencias de ADNs satélites.

Ensamblaje y anotación de transcriptoma

Generamos un transcriptoma *de novo* a partir de lecturas Illumina. En primer lugar concatenamos todos los ficheros FASTQ para realizar una normalización *in silico* con una máxima cobertura de 50x. Los ficheros FASTQ con lecturas selecionadas fueron utilizados para realizar un ensamblaje *de novo* mediante Trinity (Haas *et al.*, 2013) con las opciones por defecto. Posteriormente predecimos ORFs de 100 o más aminoácidos con Transdecoder (Haas *et al.*, 2013) y eliminamos redundancias con CDHit-EST (Li & Godzik, 2006). Realizamos la anotación de los contigs con Trinotate (https://trinotate.github.io) utilizando las bases de datos SWISS-PROT (Boeckmann *et al.*, 2003). Realizamos anotación funcional utilizando grupos de proteínas eucariotas Ortólogas (KOG) buscando la secuencia de proteína en el servidor WebMGA (http://weizhong-lab.ucsd.edu/metagenomic-analysis/ server/kog).

Análisis de la abundancia, diversidad y estructura de un elemento

Identificación de microsatélites

Utilizamos un programa escrito en lenguaje Perl desarrollado por Meglécz *et al.* (2012) aportado por los autores. Como criterios de búsqueda consideramos microsatélites prefectos de mono- (al menos 12 repeticiones), di-, tri-, tetra- y hexanucleótidos (al menos 5 repeticiones). Consideramos el mismo motivo de microsatélite aquellos que resultaban ser permutaciones circulares y/o secuencias reverso complementarias.

Abundancia y divergencia de elementos repetidos

Utilizando las secuencias de referencia alineamos las lecturas Illumina con el programa RepeatMasker (Smit *et al.*, 2013) utilizando Cross_match como motor de búsqueda. Lanzamos RepeatMakser con opción "-a" para generar un fichero *.align que contiene información sobre el número de nucleótidos alineados para cada secuencia de referencia, así como la divergencia con respecto a la referencia. Podemos estimar la abundancia total y divergencia promedio. Además podemos generar un paisaje repetitivo ("repeat landscape") con el script distribuído con RepeatMasker llamado "calcDivergenceFromAlign.pl". Consiste en representar gráficamente la abundancia para cada rango de divergencia para el conjunto de las secuencias repetidas, así podemos tener una idea general de la abundancia y diversidad de ciertos elementos.

Cobertura a lo largo de un elemento

En ocasiones es interesante comprobar si la cobertura a lo largo de una secuencia está alterada con respecto a lo esperado. Para comprobar esto, podemos realizar mapeos de lecturas a un conjunto de secuencias de referencia con SSAHA2 (Ning *et al.*, 2001). Después podemos visualizar el resultado con IGV (Thorvaldsdóttir *et al.*, 2013) y comprobar así si existe variación de cobertura anómala a lo largo del contig.

Estudios de SNPs

Hemos desarrollado un script para detectar la presencia de SNPs característicos de las librerías con B a partir de los mapeos mencionados en el apartado anterior. Para SNPs próximos, a una distancia de la mitad de una lectura Illumina, aproximadamente 50 nucleótidos, estudiamos la frecuencia de haplotipos extrayendo lecturas mapeadas que mapeasen en los SNPs seleccionados.

Referencias

Bao Z, Eddy SR (2002) Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome research*, **12**, 1269–1276.

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*, **27**, 573.

Boeckmann B, Bairoch A, Apweiler R, *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research*, **31**, 365–370.

Cabrero J, Bakkali M, Bugrov A, *et al.* (2003) Multiregional origin of B chromosomes in the grasshopper *Eyprepocnemis plorans*. *Chromosoma*, **112**, 207–211.

Camacho JPM, Cabrero J, Viseras E, López-León MD, Navas-Castillo J, Alché JD (1991) G banding in two species of grasshopper and its relationship to C, N, and fluorescence banding techniques. *Genome*, **34**, 638–643.

Cuadrado A, Jouve N (2007) Similarities in the chromosomal distribution of AG and AC repeats within and between *Drosophila*, human and barley chromosomes. *Cytogenetic and genome research*, **119**, 91–99.

Cuadrado A, Schwarzacher T, Jouve N (2000) Identification of different chromatin classes in wheat using in situ hybridization with simple sequence repeat oligonucleotides. *Theoretical and Applied Genetics*, **101**, 711–717.

Haas BJ, Papanicolaou A, Yassour M, *et al.* (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, **8**, 1494–1512.

Hahn C, Bachmann L, Chevreux B (2013) Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Research*, **41**, e129–e129.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*, **110**, 462–467.

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Meglécz E, Nève G, Biffin E, Gardner MG (2012) Breakdown of phylogenetic signal: a survey of microsatellite densities in 454 shotgun sequences from 154 non model eukaryote species. *PLoS One*, **7**, e40861.

Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. *Genome research*, **11**, 1725–1729.

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792–793.

Smit AFA, Hubley R, Green P (2013) RepeatMasker Open-4.0.

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, **14**, 178–192.

Chapter 1. Next generation sequencing and FISH reveal uneven and non random microsatellite distribution in two grasshopper genomes

Francisco J. Ruiz-Ruano[1], Ángeles Cuadrado[2], Eugenia E. Montiel[1], Juan Pedro M Camacho[1] and María Dolores López-León[1]

[1]Departamento de Genética, Universidad de Granada
[2]Departamento de Biomedicina y Biotecnología, Universidad de Alcalá

Published in *Chromosoma*, **124**, 221-234 (2015).
http://doi.org/10.1007/s00412-014-0492-7

Abstract. Simple sequence repeats (SSRs), also known as microsatellites, are one of the prominent DNA sequences shaping the repeated fraction of eukaryotic genomes. In spite of their profuse use as molecular markers for a variety of genetic and evolutionary studies, their genomic location, distribution and function are not yet well understood. Here we report the first thorough joint analysis of microsatellite motifs at both genomic and chromosomal levels in animal species, by a combination of 454 sequencing and *in situ* hybridization (FISH) techniques performed on two grasshopper species. The *in silico* analysis of the 454 reads suggested that microsatellite expansion is not driving size increase of these genomes, as SSR abundance was higher in the species showing the smallest genome. However, the two species showed the same uneven and nonrandom location of SSRs, with clear predominance of dinucleotide motifs and association with several types of repetitive elements, mostly histone-genes spacers, ribosomal DNA intergenic spacers (IGS), and transposable elements (TEs). The FISH analysis showed a dispersed chromosome distribution of microsatellite motifs in euchromatic regions, in coincidence with chromosome location patterns previously observed for many mobile elements in these species. However, some SSR motifs were clustered, especially those located in the histone gene cluster.

Introduction

Microsatellite DNA (or SSRs: simple sequence repeats) is an abundant class of repetitive DNA composed of tandemly arranged short repeated motifs of 1 to 6 bp and being widely distributed in plant and animal genomes (Tautz & Renz, 1984; Tóth *et al.*, 2000; Zane *et al.*, 2002; Cuadrado & Jouve, 2007a, 2011). Studies using microsatellites have been mainly focused on their use as polymorphic markers in population genetics, genetic diversity or kinship contexts (reviewed in: Schlötterer & Pemberton, 1998; Schlötterer & Goldstein, 1999; Blondin *et al.*, 2013. This is particularly true in insects, where reports based on SSR molecular variation are abundant (Insuan *et al.*, 2007; Augustinos *et al.*, 2011; Manrique-Poyato *et al.*, 2013). In recent years, microsatellite genomic analysis has been powered by the development of high throughput next-generation sequencing (NGS) technologies, based on massive sequencing approaches which enable a rapid, low-cost and low time-consuming way to characterize microsatellites (Malausa *et al.*, 2011; Iquebal *et al.*, 2013). This technology is nowdays particularly interesting for non-model species (Meglécz *et al.*, 2012; Hunter & Hart, 2013; Schoebel *et al.*, 2013), as it allows a global description of repetitive DNAs, including the microsatellite fraction.

Much less is known about microsatellite chromosomal distribution, even though its knowledge is critical for addressing key issues such as chromosome origin, organization, structure, function and evolution. Moreover, precise chromosomal identification may be based on specific SSR distribution pattern, as is the case in barley where chromosome distribution of $(AAG)_5$ and $(ACT)_5$ microsatellite repeats allows easy chromosome identification and detection of structural chromosome rearrangements (e.g. translocations) (Carmona *et al.*, 2013). *Drosophila* is one of the few animals where microsatellites have been profusely analyzed at chromosomal level (Cuadrado & Jouve, 2007b; Santos *et al.*, 2010; Cuadrado & Jouve, 2011).

In silico mining approaches, based on genome database screening or flow-sorted chromosomes, can be useful for establishing SSR abundance and chromosomal assignment (Tóth *et al.*, 2000; Basset *et al.*, 2006; Shi *et al.*, 2014). However, physical mapping by fluorescent *in situ* hybridization (FISH) constitutes a quicker and easier strategy for ascertaining the abundance and chromosomal distribution of microsatellite and other repeated sequences, thus providing a means for determining, in detail, the precise chromosomal regions in which repeats are clustered, and this is the approach of choice in non-model species. The development of NGS technology has highlighted the need for combined genomic and chromosomal analyses for mapping new characterized sequences, and confirming *de novo* genome assemblies, but these combined studies are still scarce (Soltis *et al.*, 2013; Kejnovskỳ *et al.*, 2013).

Different lines of evidence have demonstrated that microsatellite distribution is non random. They are frequently located in heterochromatic regions (Lohe *et al.*, 1993; Cuadrado & Jouve, 2011) and sometimes are differentially accumulated on specific chromosomes (e.g. sex chromosomes) (Kubat *et al.*, 2008; Poltronieri *et al.*, 2013). In other cases, SSRs colonize the euchromatin (Pardue *et al.*, 1987; Cuadrado & Jouve, 2007b, 2010) suggesting a possible functional role for them. SSR accumulation in specialized chromosome regions, like centromeres (Areshchenkova & Ganal, 1999; Cuadrado & Jouve, 2007a), telomeres (Hatanaka *et al.*, 2003) or nucleolar organizing regions (Nanda *et al.*, 1991; Cuadrado & Jouve, 2007b) has also been reported. Basically, each species or group of species seems to show preferential accumulation of some specific SSR motifs with a particular chromosomal distribution (Tóth *et al.*, 2000).

Comparative studies have pointed out the idea that SSR frequency and stretch length are correlated with genome size (Hancock, 2002; Tóth *et al.*, 2000), although this relationship is not universal (Lim *et al.*, 2004; Ustinova *et al.*, 2006; Pannebakker *et al.*, 2010). The grasshoppers *Eyprepocnemis plorans* and *Locusta migratoria* have large genomes of 11.11 and 6.3 Gb, respectively (Ruiz-Ruano *et al.*, 2011), with high prevalence of repetitive DNA, such as satellite DNA or transposable elements (TEs) (López-León *et al.*, 1994; Montiel *et al.*, 2012; Wang *et al.*, 2014; Chapter 2, 4). This makes them good candidates to harbour microsatellite DNA sequences. Up to know, only two studies had analyzed microsatellites in these two species: a FISH analysis showing the presence of interspersed AG microsatellite motifs in the euchromatic regions of *E. plorans* chromosomes (Cuadrado & Jouve, 2010), and a molecular study of eight polymorphic microsatellite loci in *L. migratoria* (Zhang *et al.*, 2003). Recently, Wang *et al.* (2014) have published a draft 6.5 Gb genome sequence of *L. migratoria*, but nothing is said about the presence and abundance of microsatellites among the different repetitive DNA families they found. These two grasshopper species have a chromosome complement consisting of 11 autosome pairs and an X0♂/XX♀ sex chromosome system, with constitutive heterochromatin regions mainly located in pericentromeric regions (Camacho *et al.*, 1991). Both species harbour supernumerary (B) chromosomes, i.e. dispensable chromosomes being mostly heterochromatic and behaving as parasitic elements (Camacho, 2005). In *E. plorans*, there is a complex system of B chromosomes with more than 50 B variants which are mainly composed of a 180 bp satellite DNA and ribosomal DNA (López-León *et al.*, 1994; Cabrero *et al.*, 1999), although they also contain a small fraction of transposable DNA (Montiel *et al.*, 2012) and a microsatellite motif (Muñoz-Pajares *et al.*, 2011). Likewise, *L. migratoria* B chromosomes contain TEs (Chapter 4) and genes for the H3 and H4 histones (Teruel *et al.*, 2010). Their predominantly heterochromatic nature makes it likely the presence of microsatellites in them.

In view of the recently highlighted need of microsatellite studies combining genomic and chromosomal approaches, we perform here NGS and FISH analyses of microsatellites in two grasshopper species showing different genome size (*E. plorans* and *L. migratoria*), with the aim of getting thorough genomic information on microsatellite presence, abundance, diversity and chromosome location.

Materials and methods

Experimental materials and chromosome preparations

Adult males and females of *E. plorans* and *L. migratoria* were collected at Torrox (Málaga, Spain) and Otívar (Granada, Spain) in the first species and at Padul (Granada) population in the second species. Two *E. plorans* males, one from Torrox and another from Otívar populations with two and three B chromosomes, respectively, and one *L. migratoria* male carrying B chromosomes were used for 454 sequencing experiments. The presence of B chromosomes was determined in 2% lactopropionic orcein squash preparations of two testicular follicles.

Several other adult specimens were crossed in the laboratory to obtain embryos for FISH purposes. After egg pod incubation at 28ºC for twelve (*E. plorans*) or ten (*L. migratoria*) days, embryos were obtained by dissection of eggs in insect saline solution. Embryo fixation for cytological analysis and chromosome preparations were made as described in Camacho *et al.* (1991) and 18 embryos from each species were used for FISH experiments. For genomic DNA isolation and massive genome sequencing, an adult male body from each of the three populations sampled was frozen in liquid nitrogen and stored at -80ºC until use.

Fluorescent *in situ* hybridization (FISH)

Chromosome preparations were dehydrated in a series of 70%, 90% and 100% ethanol for 3, 3 and 5 min, respectively, and air dried. Then they were pre-treated with RNase, fixed with 4% paraformaldehyde and dehydrated in an ethanol series before air drying. Between each two treatments, the slides were washed in 2xSSC as previously reported (Cuadrado & Jouve, 2007b).

A total of 15 different mono-, di-, tri- and tetra-nucleotide microsatellite motifs were physically mapped using synthetic oligonucleotides labelled with biotin at both ends (Roche Applied Science) as probes (Table 1.1).

Table 1.1: Genomic abundance (%), chromosomal distribution and colocalization with the histone cistron of SSR motifs observed in *E. plorans* by cytogenetic (FISH) and molecular (454 sequencing) approaches.

	FISH			454 sequencing	
SSR probe	DP[1]	CS[2]/CL[3]	MCH[4]	Abundance[5]	AH[6]
A	+	2d	+	10.06	1.90
AC	+	2d	+	47.80	85.50
AG	+	9i		12.27	5.87
AAC	+	2d	+	4.56	1.15
AAG	+	2d Xi 10i	+	1.43	0.18
GATA	+	1p 2d Xi 10i	+	0.03	0
GACA	+	2d	+	0.13	0.01
AAT	-	-		4.18	0.47
AGG	-	-		0.41	0.08
ACT	-	-		0.92	0.04
CAT	-	-		1.82	0.08
CAC	-	-		0.55	0.07
ACG	-	-		0.19	0
CAG	-	-		1.01	0.01
GCC	-	-		0.75	0

[1] DP: Dispersed pattern of the FISH signal
[2] CS: Chromosome size (autosomes are numbered from 1 to 11 in order of decreasing length)
[3] CL: Chromosome location of SSR clusters in respect to centromere location (d= distal; i= interstitial; p= proximal)
[4] MCH: Microsatellite motifs colocalizing with the histone cistron
[5] Abundance: SSR abundance (%) expressed as relative coverage (nucleotides/Mb library)
[6] AH: Percent of SSR loci associated with the histone cistron
+: presence of FISH signal; -: absence of FISH signal
Nomenclature correspondence between microsatellite probes and 454 sequencing microsatellite data: GATA/AGAT; GACA/ACAG; CAT/ATC ; CAC/ACC; CAG/AGC; GCC/CCG

The histone H3 gene probe was obtained from cloned DNA fragments described in Cabrero *et al.* (2009) and was labeled with fluorescein 11-dUTP (Roche) by nick translation using 2.5 units of DNA polymerase I/DNase I (Invitrogen) in a 50 μL reaction following standard protocol.

Chromosome and probe denaturation and *in situ* hybridization were performed as described by Cuadrado *et al.* (2000). In brief, the hybridization mixture was prepared by adding 50% deionized formamide, 10% dextran sulphate, 2x SSC, 0.1% SDS (sodium dodecyl sulphate), 2 pm of the microsatellite probes and 50 ng of the histone H3 probe. For post-hybridization washing, slides were immersed in 4xSSC/0.2%Tween-20 and agitated for 10 min at room temperature (RT). The detection of biotin was performed by incubating the slides in streptavidin-Cy3 (Sigma) in 5% (w/v) BSA for 1h at 37ºC. Before DAPI (4',6-diamidino-2-phenylindole) staining, slides were rinsed for 10 min in 4xSSC/0.2%Tween-20 at RT. They were mounted in Vectashield antifading medium (Serva) and examinated with a Zeiss Axiophot microscope. The images captured from each filter were recorded separately using a CCD camera (Nikon DS) and the resulting digital images were processed using Adobe Photoshop.

Genomic DNA isolation and 454 sequencing

Genomic DNA was extracted from frozen grasshopper bodies using the Gen Elute Mammalian Genomic DNA Miniprep Kit (Sigma-Aldrich) protocol. DNA quantity and quality were assesed with a Tecan's Infinite 200 NanoQuant and a 1% agarose gel electrophoresis. Pyrosequencing assays were performed in a 454 Genome Sequencer FLX Plus platform by Macrogen Inc: 3/8 plate for *E. plorans* (1/8 Torrox and 2/8 Otívar) and 2/8 plate for *L. migratoria*.

Analysis of 454 sequence reads

Low quality regions of raw reads were trimmed with the Roche's 454 Data Analysis Software. Microsatellite identification was performed using a Perl script developed by Meglécz *et al.* (2012) for microsatellite array searching, kindly provided by the authors. As search criteria, we considered perfect microsatellites of mononucleotide (with at least 12 repetitions), dinucleotide, trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide (with a minimum of 5 repetitions). We chose these criteria in order to get results being comparable with most previous microsatellite studies. Repeat motifs being circular permutations and/or reverse complement of each other were pooled together as a same class. Microsatellite content was determined on the basis of three parameters: a) microsatellite frequency, expressed as the number of microsatellite loci per Mb of DNA sequence, b) microsatellite coverage, i.e. the number of microsatellite nucleotides per Mb of DNA, and c) relative microsatellite abundance, expressed either as microsatellite composition per repeat unit length or microsatellite composition per motif sequence, both of them given as percentage.

To identify non-microsatellite repeats associated with microsatellites, we annotated the reads containing microsatellites. First, we performed a *de novo* identification and modelling of repetitive sequences separately in each library with the RepeatModeler pipeline, which uses the RECON and Repeat Scout (Benson, 1999; Bao & Eddy, 2002) programs with the RepBase database version 20130422 (Jurka *et al.*, 2005). In addition, we included the sequence of the histone and ribosomal RNA (rRNA) gene clusters, including spacers, assembled with RepeatExplorer (Novák *et al.*, 2013) in both species (Chapter 4; Ruiz-Ruano *et al.*, in preparation). The resulting file was used as a database to annotate all the reads (containing microsatellites or not) by means of RepeatMasker (Smit *et al.*, 2013). Then we used a custom Python script to count the number of repeats per microsatellite motif and annotation, with the aim of getting an indication of microsatellite abundance in the vicinity of other non-microsatellites repeats. For this purpose, the script uses the table of masked sequences provided by RepeatMasker for the microsatellite-containing reads, and a table, built by us, containing the number of repeats found per read for each SSR motif.

The possible association between microsatellite and non-microsatellite repeats can be investigated by contingency chi-square tests applied to the number of reads annotated for a given non-microsatellite repeat containing or lacking microsatellites, but this inevitably leads to an underestimate. This is due to the fact that read length (ca. 700 nt) was lower than the length of most non-microsatellite repeat DNA types found (shown in Tables 1.S1 and 1.S2). This means that a proportion of the reads obtained from a microsatellite-associated element will lack microsatellites thus being included in the microsatellite-free class. For instance, if a repetitive element is 3500 nt long and is associated with a microsatellite in the 3′ end, we can observe only 20% (i.e. 700/3500) of its reads being associated with the microsatellite, whereas the remaining 80% of the reads would be considered as microsatellite-free reads. This flaw is difficult to solve in the absence of a full genome sequence, which is completely absent in *E. plorans* and only in a draft public annotation in *L. migratoria* (Wang *et al.*, 2014).

With this caution in mind, we wrote another custom Python script to count the number of reads found in the SSR-carrying and SSR-lacking read groups for every annotated non-SSR repeat found in the RepeatMasker table mentioned above. We then applied contingency chi square tests and calculated the odds ratio (OR) for each kind of repetitive element. We then applied the Bonferroni method for multiple tests and only considered as significant associations those showing $P<0.05$ and $OR>1$. The inherent flaw of this approach (see above) runs against finding significant associations, thus suggesting that it is actually very conservative and will only reveal the strongest associations.

Results

FISH maping of SSRs in *E. plorans*

Seven out of the 15 microsatellite motifs analyzed yielded signals on *E. plorans* chromosomes after FISH assay. These were $(A)_{20}$, $(AC)_{10}$, $(AG)_{10}$, $(AAC)_5$, $(AAG)_5$, $(GATA)_5$ and $(GACA)_5$, (Table 1.1). No signal was however observed for the remaining eight SSR motifs, all being trinucleotide motifs: $(AAT)_5$, $(AGG)_5$, $(ACT)_5$, $(CAT)_5$, $(CAC)_5$, $(ACG)_5$, $(CAG)_5$ and $(GCC)_5$ (Table 1.1). All mono, di, and tetranucleotide microsatellites studied, as well as the two trinucleotide ones, $(AAC)_5$ and $(AAG)_5$, showed a dispersed but non random distribution on all chromosomes, appearing scattered over euchromatin regions covering nearly the entire chromosome lengths except small paracentromeric regions (Fig. 1.1), where heterochromatin is located in this species (Camacho *et al.*, 1991). Dinucleotides and trinucletides yielded the strongest and the weakest hybridization signals, respectively. Additionally, six out of the seven microsatellites that were detectable by FISH, namely $(A)_{20}$, $(AC)_{10}$, $(AAC)_5$, $(AAG)_5$, $(GATA)_5$ and $(GACA)_5$, showed clustered hybridization (i.e. signals occurring closely together in a chromosome region yielding band-like hybridization pattern) on a distal region of chromosome 2 (Fig. 1.1a, b, d, f, g, h), the exception being $(AG)_5$. Double FISH with DNA probes for SSRs and H3 histone genes showed that this distal region in the chromosome 2 also contains the cluster for histone genes in this species (Fig. 1.1e, i).

Other microsatellite clustered sites appeared for dinucleotide $(AG)_5$, trinucleotide $(AAG)_5$ and tetranucleotide $(GATA)_5$ repeats, with AG stretches being interstitially (i.e. with an intermediate chromosome position) located on chromosome 9 and AAG repeats located interstitially on chromosomes X and 10, yielding discrete bands (Fig. 1.1c, f). GATA repeats yielded a band-like pattern that can be observed proximally (i.e. close to the centromeric region) on chromosome 1 and interstitially on chromosomes X and 10. Variation among individuals for the presence of these bands can be observed (Figs. 1.1h and 1.S8a). Different hybridization patterns are summarized in Table 1.1.

The AC microsatellite motif is present in the B chromosome of *E. plorans*, mapping at the distal region, where the ribosomal DNA (rDNA) is located (Cabrero *et al.*, 1999), and also at the two interstitial $DAPI^-$ bands located between the three heterochromatic $DAPI^+$ blocks which characterize this B variant (Fig. 1.1b). The remaining SSR motifs appear to be absent from the B chromosome, although we cannot discard the presence of few undetected copies located at $DAPI^-$ bands due to the general high level of chromatin condensation in the B chromosome.

NGS analysis of microsatellites in the *E. plorans* genome

To ascertain the abundance, types and diversity of microsatellite motifs in the *E. plorans* genome, 454 sequencing was performed on genomic DNA from a B-carrying individual. 454 runs yielded 243.93 Mb of data (380,394 reads) composed of reads with a N50 value of 768 bp and, after trimming off low quality regions, the N50 was 765 bp, making a total of 243.17 Mb

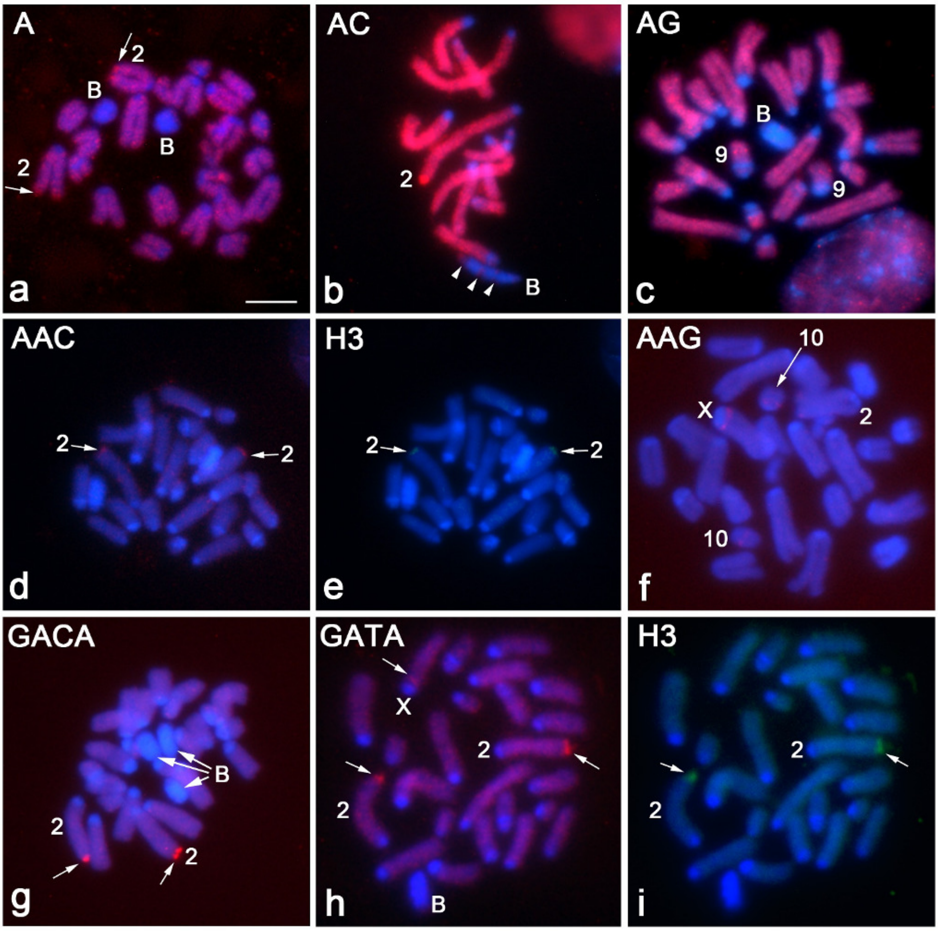


Figure 1.1: FISH mapping of microsatellite motifs in *E. plorans*. Embryo mitotic metaphase cells of *E. plorans* showing DAPI staining plus FISH for A, AC, AG, AAG, GACA microsatellite motifs (a-c, f, g) and double FISH for AAC (d,e) and GATA (h ,i) microsatellites (red) and H3 gene (green). Note the absence of FISH signals on the paracentromeric regions of all chromosomes. Arrows point to clustered hybridization signals on different chromosomes. Note the colocalization of microsatellites and H3 genes in the distal region of chromosome 2 (d,e h,i).and the presence of three interstitial AC carrying bands (arrowheads) in the B chromosome, located between the three heterochromatic DAPI+ blocks. B= B chromosome. Bar = 5 μm.

implying about 0.022x genome coverage (NCBI's SRA database: accesion numbers SRR1200829 and SRR1200835).

Out of the 243.17 Mb of nucleotides obtained, 1,121,693 (0.46%) nucleotides corresponded to microsatellite repeats (586,676) meeting our criteria specifications. We found a microsatellite frequency of 240 loci per Mb of DNA and total microsatellite coverage of 4,612 nucleotides per Mb of

DNA sequenced, the average length of microsatellite loci being 10.03 repeats. The estimated mean distance among loci was 4,158 nucleotides. We identified 148 different microsatellite motifs in the 454 library sequences. They were classified on the basis of motif length and their abundance was expressed as relative coverage, 12.75% of them being mononucleotides, 66.19% dinucleotides, 15.81% trinucleotides, 2.95% tetranucletides, 1.97% pentanucleotides and 0.33% hexanucleotides. Therefore, there was a prevalence of mono-di-and trinucleotide repeats, with dinucleotides being the most abundant microsatellite repeats (Tables 1.2 and 1.S3). In fact, 64% of the 58,478 microsatellite loci found were dinucleotides, whereas only 17% were mononucleotides and 15% were trinucleotides, the remaining classes showing frequencies lower than 3%.

Table 1.2: Microsatellite coverage and relative abundance per motif length in *E. plorans* and *L. migratoria*

	E. plorans		*L. migratoria*	
Motif length class	Nucleotides per Mb library	Relative motif abundance (%)	Nucleotides per Mb library	Relative motif abundance (%)
mononucleotides	588	12.75	1192	21.91
dinucleotides	3052	66.19	3349	61.56
trinucleotides	729	15.81	680	12.50
tetranucleotides	136	2.95	174	3.20
pentanucleotides	91	1.97	38	0.70
hexanucleotides	15	0.33	7	0.13

The mean number of repeats per locus was higher for mono-nucleotide loci (14.15) followed by that for dinucleotides (9.88) (Table 1.3). This is not surprising since the search criterion for mononucleotide motifs was showing at least 12 repeats, whereas it was 5 for all other motifs. The longest microsatellite stretch found was a 356 bp AC microsatellite (Table 1.S4).

Table 1.3: Number and length of microsatellite loci per motif-size in the *E. plorans* and *L. migratoria* genomes.

	E. plorans			*L. migratoria*		
Motif length class	NR1[1]	NL[2]	MNRL[3]	NR[1]	NL[2]	MRL[3]
mononucleotides	143090	9900	14.45	237773	15517	15.32
dinucleotides	371081	37572	9.88	333980	32688	10.22
trinucleotides	59100	8727	6.77	45231	6571	6.88
tetranucleotides	8298	1446	5.74	8699	1489	5.84
pentanucleotides	4447	726	6.13	1523	199	7.65
hexanucleotides	619	107	5.79	248	45	5.51

[1] NR: Number of repeats

[2] NL: number of loci

[3] MNRL: Mean number of repeats per locus

Table 1.4: Microsatellite abundance (%) per motif length and sequence in the *E. plorans* and *L. migratoria* genomes.

Motif length class	Motif sequence	*E. plorans*	*L. migratoria*
mononucleotides	A	78.89	89.02
	C	21.11	10.98
dinucleotides	AC	72.24	70.64
	AG	18.55	19.47
	AT	8.33	8.44
	CG	0.88	1.44
trinucleotides	AAC	28.83	26.04
	AAG	9.02	5.65
	AAT	26.42	37.57
	ACC	3.46	4.39
	ACG	1.22	1.17
	ACT	5.83	4.51
	AGC	6.39	3.79
	AGG	2.60	3.29
	ATC	11.49	10.48
	CCG	4.74	3.12

The analysis of relative abundance of the different microsatellite motif sequences was based on the number of repeats found for every motif, and was expressed as percentage. The predominant motifs for the three most frequent classes were A (78.89%) within mononucleotides, AC (72.24%) within dinucleotides, and AAC (28.83%) and AAT (26.42%) within trinucleotides (Tables 1.4 and 1.S3). The most abundant SSRs were detected by FISH, with stronger signals for those with higher relative abundance in the 454 library, e.g. AC (47.80%), AG (12.27%) and A (10.06%). In addition, three SSR motifs (AAG, GATA and GACA) showing 454 coverage lower than 1.5% were also apparent by FISH (Tables 1.1 and 1.S3; Figs. 1.1 and 1.S1a). However, all trinucleotide motifs that had failed to show hybridization signals by FISH were found after the 454 sequencing (Tables 1.1 and 1.S3), but they were very scarce in the *E. plorans* genome, representing only 4.18 % (AAT) or less of microsatellite coverage.

The assembly with RepeatModeler yielded 2,760 contigs in *E. plorans*, summing up 1,448,555 nt (N50= 708 nt). About 31% of the these contigs were annotated (Table 1.S5). Since the spacers of the histone gene cluster and the IGS region of the 45S rDNA) were not annotated by ReapeatModeler, we assembled them with RepeatExplorer and added the resulting contigs to the database. All reads were then annotated through sequence search in the database. 22,949 out of the 47,763 reads containing microsatellite motifs were annotated for non-microsatellite repeats. As Table 1.5 shows, most SSR repeats (84.7%) were located close to transposable elements (TEs), with slightly higher representation of class I retrotransposons (45.1%) over class II DNA transposons (39.6%). LINEs were prevalent among retrotransposons whereas Tc/ mariner family elements were the most frequent transposon. The remaining repetitive DNAs found nearby microsatellites included histone and ribosomal RNA gene spacers, satellite DNA, and others (Tables 1.5 and 1.S1).

Table 1.5: Number and relative abundance (%) of microsatellite repeats in the reads annotated following the RepBase criterion for the analysis of abundance of microsatellite repeats in the proximity of other repetitive DNA.

	E. plorans			*L. migratoria*		
Repeat DNA	MR[1]	MRA[2] (%)	MFM[3]	MR[1]	MRA[2] (%)	MFM[3]
Clase I (Retrotransposon)						
LINEs	152326	35.94	AC	203569	43.48	AC
LTRs	15069	3.56	AC	24824	5.30	A
SINE	23711	5.59	AC	45973	9.81	A
Total	191106	45.09		274366	58.59	
Clase II (DNATransposons)						
HATs	45750	10.79	AC	16990	3.63	A
Tc/mariner	79060	18.65	AC	106631	22.77	AC
Other transposons	31770	7.50	AC	12154	6.12	AC
Helitron	11084	2.62	AC	18690	3.99	AC
Total	167664	39.56		170951	36.51	
Other repeated sequences						
RNA	138	0.03	A	–	–	
rDNA	2189	0.52	A	2151	0.46	A
rDNA/IGS	8574	2.02	ATC	3992	0.85	A
snRNA	–	–		464	0.1	A
Histones genes spacers	53546	12.63	AC	5177	1.11	AC
Satellites	18	0.01	AG	9056	1.93	A
Simple repeats	589	0.14	CCG	2080	0.44	A
Total	65054	15.35		22920	4.89	

[1] MR: Number of microsatellite repeats
[2] MRA: Microsatellite repeats abundance expressed as percentage
[3] MFM: Most frequent microsatellite motif

The statistical analysis of the 454 annotated reads for the association between microsatellite and non-microsatellite repeats revealed significant association of SSRs with histone gene spacers, rDNA IGS sequences and several kinds of mobile elements, including LTR and non-LTR retrotransposons and DNA transposons, such as LTR/ERVK, LINE/R1, hAT-Ac, Sola or Helitron, the strongest association being shown with histone gene spacers. In the absence of Bonferroni correction, additional associations were observed with other mobile sequences as SINE/tRNA-RTE, TcMar-Tc2 or Merlin DNA trasposons (Table 1.S6).

As mentioned above, FISH also revealed the colocalization of six diferents motifs (A, AC, AAC, AAG, GATA and GACA) with histone genes in the distal region of chromosome 2 (Table 1.1; Fig. 1.1e, i). These mono-di-trinucleotides and tetranucleotide GACA were also found in the vicinityof histone gene spacers in the 454 reads. As expected, the 454 analysis showed other motifs being close to histone genes spacers (Tables 1.1 and 1.S1).

FISH mapping of SSRs in *Locusta migratoria*

Seven SSR motifs, i.e. $(A)_{20}$, $(AC)_{10}$, $(AG)_{10}$, $(AAG)_5$, $(ACT)_5$, $(GATA)_5$ and $(GACA)_5$, hybridized on *L. migratoria* chromosomes (Table 1.6), exhibiting a dispersed FISH pattern similar to that described above for *E. plorans*. Scattered hybridization signals were seen over the euchromatic regions of all chromosomes, comprising nearly the entire chromosome length, or else yielding some additional bands on different chromosomes. Pericentromeric regions, which are heterochromatic, were apparently devoid of all microsatellite motifs analyzed (Fig. 1.2). The $(A)_{20}$ mononucleotide and the $(AC)_{10}$ and $(AG)_{10}$ dinucleotide repeats were scattered over all euchromatic regions (Fig. 1.2a-c). In addition, the two latter SSRs hybridized in specific regions, i.e. interstitially on the chromosomes 2, 8, 9 and 11 and distally on the chromosome 10, in the case of $(AC)_{10}$ repeats (Fig. 1.2b), and interstitially on chromosomes X and 8 and proximally on chromosomes 3 and 8, in the case of $(AG)_{10}$ (Fig. 1.2c). All eight SSRs that failed to show FISH signals with *L. migratoria* chromosomes were trinucleotide repeats ($(AAT)_5$, $(AAC)_5$, $(AGG)_5$, $(CAT)_5$, $(CAC)_5$, $(ACG)_5$, $(CAG)_5$, $(GCC)_5$), in almost complete coincidence with *E. plorans*, the only exceptions being $(AAC)_5$ and $(ACT)_5$ which were observed only in *E. plorans* and *L. migratoria*, respectively. The two trinucleotides that gave positive hybridization signals in *L. migratoria*, $(AAG)_5$ and $(ACT)_5$, showed a dispersed distribution across euchromatin, but were also clustered in specific regions, revealing $(AAG)_5$ repeat bands that can be observed located proximally and interstitially on chromosome 8, proximally on chromosome 3 and interstitially on the X, 9 and 11 chromosomes (Figs. 1.2d and 1.S1b). Likewise, $(ACT)_5$ repeats yielded interstitial bands on chromosomes 2, 8 and 9, and a distal one on chromosome 10 (Fig. 1.2f). Tetranucleotide $(GATA)_5$ repeats were clustered proximally on chromosome 3 and interstitially on chromosomes 9 and 10 (Fig. 1.2i). Finally, $(GACA)_5$ was located in proximal clusters on chromosomes 1 and 3 and interstitially on chromosomes 8, 9 and 10 (Fig. 1.2g). Some variation for the presence of these clustered motifs was observed (Figs. 1.2 and 1.S1b).

Double FISH performed with SSRs and histone H3 gene probes showed that the $(AC)_{10}$, $(AAG)_5$, $(ACT)_5$ and $(GACA)_5$ motifs co-located with the H3 histone genes, whereas $(AG)_{10}$ mapped close to them, in an interstitial position on the chromosome 8 (Fig. 1.2c, e, h). B chromosomes in this

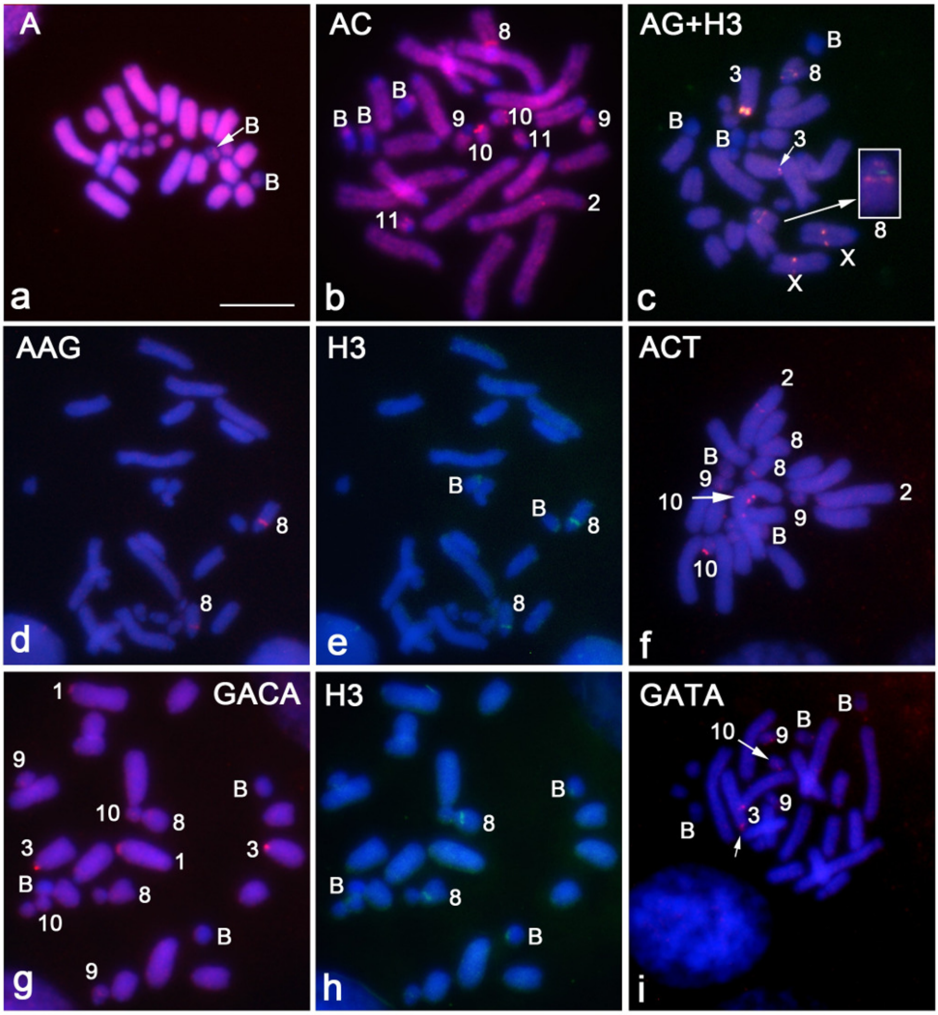


Figure 1.2: FISH mapping of microsatellite motifs in *L. migratoria*. Embryo mitotic metaphase cells of *L. migratoria* showing DAPI staining plus FISH for A, AC, ACT, GATA microsatellite motifs (a,b, f, i) and double FISH for AG, AAG and GACA microsatellites motifs (red) and H3 gene (green) (c-e, g, h). Note the absence of FISH signals on small paracentromeric regions of all chromosomes as well as the presence of clustered hybridization signals for AC, AG, AAG, ACT, GACA and GATA motifs on different chromosomes (b-d, f, g, i). Also note the colocalization of the AAG and GACA microsatellites with H3 genes in an interstitial region of chromosome 8, and the nearby location of the AG microsatellite (c-e, g, h). B = B chromosome. Bar = 5 μm.

species contained A, AC, ACT and GACA motifs dispersed over its proximal DAPI$^+$ third (Fig. 1.2). Other motifs could also be present but at low copy number, making them undetectable by FISH in the highly condensed B chromosomes.

NGS analysis of microsatellites in the *L. migratoria* genome

A total of 292,352 reads were obtained from the 454 genomic DNA sequencing. They showed N50 of 839 bp implying a total of 199.60 Mb and, after trimming off low quality regions, the N50 was 838 bp summing up 199.45 Mb of data in the library, implying about 0.032x genome coverage (NCBI's SRA data base: accession number SRR1200889).

The analysis of the 454 genomic DNA library revealed the presence of 627,459 microsatellite repeats with 99 different microsatellite motifs comprising a total of 1,085,325 microsatellite nucleotides (i.e. 0.54% of total sequences) meeting our selection criterion. Total microsatellite coverage was 5,441 nucleotides per Mb of DNA, with 283 microsatellite loci per Mb of DNA and average length of 11.1 repeats per locus. The estimated density was one microsatellite locus every 3,530 nucleotides of DNA within the *L. migratoria* genome.

Among the 99 different microsatellite motifs, we found 21.91% mononucleotide, 61.56% dinucleotide, 12.50% trinucleotide, 3.20% tetranucleotide, 0.70% pentanucleotide and 0.13% hexanucleotide motifs relative abundance, in high resemblance with *E. plorans* (Tables 1.2 and 1.S3). This way, dinucleotides are the most abundant SSR motifs. In fact, 57.85% of the 56,509 microsatellite loci found were dinucleotides and only 27.5% were mononucleotides and 11.63% trinucleotides. The remaining motifs constituted a small percentage (3.07%) of loci. Likewise in *E. plorans*, mononucleotide loci contained the highest mean number of repeats (15.32) followed by dinucleotides (10.22), but the longest microsatellite locus in *L. migratoria* was a 600 bp pentanucleotide AACCT motif (Tables 1.3 and 1.S7), which was almost twice longer than that found in *E. plorans* (see above).

Within each of the four most frequent motif classes, A repeats were the most frequent mononucleotide (89.02%), AC was the predominant dinucleotide (70.64%), AAT (37.57%) and AAC (26.04%) were the most frequent trinucleotides, and AAAT was the most prevalent tetranucletoide (57.86%) motifs (Tables 1.4 and 1.S3). The most abundant SSRs in the 454 library, e.g. AC (43.48%), A (19.50%) and AG (11.98%), also showed strong FISH signals (Fig. 1.2). In adition, two microsatellite motifs (GATA and GACA) yielded hybridization signals despite their low abundance (Fig. 1.2; Table 1.6). However, all eight trinucleotide motifs that had failed to show hybridization signals by FISH were present in the 454 library (Tables 1.6 and

Table 1.6: Genomic abundance (%), chromosomal distribution and colocalization with the histone cistron of SSR motifs observed in *L. migratoria* by cytogenetic (FISH) and molecular (454 sequencing) approaches.

	FISH			454 sequencing	
SSR probe	DP[1]	CS[2]/CL[3]	MCH[4]	Abundance[5]	AH[7]
A	+			19.50	26.83
AC	+	2i 8i 9i 10d 11i	+	43.48	51.09
AG	+	Xi 3p 8i p		11.98	5.35
AAG	+	Xi 3p 8i p 9i 11i	+	0.71	0.52
ACT	+	2i 8i 9i 10d	+	0.56	0
GATA	+	3p 9i 10i		0.02	0
GACA	+	1p 3p 8i 9i, 10i	+	0.17	0
AAT	-	-		4.70	0.43
AAC	-	-		3.26	0.68
AGG	-	-		0.41	3.42
CAT	-	-		1.31	0
CAC	-	-		0.55	0
ACG	-	-		0.15	2.01
CAG	-	-		0.47	2.07
GCC	-	-		0.39	0

[1] DP: Dispersed pattern of the FISH signal
[2] CS: Chromosome size (autosomes are numbered from 1 to 11 in order of decreasing length)
[3] CL: Chromosome location of SSR clusters in respect to centromere location (d= distal; i= interstitial; p= proximal)
[4] MCH: Microsatellite motifs colocalizing with the histone cistron
[5] Abundance: SSR abundance (%) expressed as relative coverage (nucleotides/Mb library)
[6] AH: Percent of SSR loci associated with the histone cistron
+: presence of FISH signal; -: absence of FISH signal
Nomenclature correspondence between microsatellite probes and 454 sequencing microsatellite data: GATA/AGAT; GACA/ACAG; CAT/ATC ; CAC/ACC; CAG/AGC; GCC/CCG

1.S3), although at very low frequency, representing 4.7% (AAT) or less of all microsatellites found.

The assembly with RepeatModeler yielded 1,904 contigs in *L. migratoria*, summing up 1,040,826 nt (N50= 703 nt). About 33% of the these contigs were annotated (Table 1.S5). After adding the histone spacers and IGS regions assembled by RepeatExplorer, we annotated all reads. 22,316 out of the 44,416 reads containing microsatellite motifs were annotated for other repetitive elements. Most SSR repeats (95.1%) were found close to transposable elements (TEs) (Tables 1.5 and 1.S2): 58.6% near retrotransposons (especially LINEs) and 36.5% in the vicinity of transposons, the most frequent being the Tc/mariner family. Other repeated sequences, e.g. histone and rRNA genes spacers, satellite DNA and snRNA, were also found near microsatellite repeats, but at much lower frequency (a mere collective 4.9%) (Tables 1.5 and 1.S2).

The statistical analysis of the association between microsatellite and non-microsatellite repeats in the *L. migratoria* 454 library showed significant association with histone genes spacers, snRNA and some retrotransposons and DNA transposons such as LINE/I-Nimb, LINE/Penelope Helitron or Tc Mariner-Tc2, as well as a satellite DNA (Table 1.S8). In the absence of the Bonferroni correction, we also found association with other TEs such as Sola elements or SINE/tRNA-Lys elements.

As mentioned above, the FISH analysis showed the co-localization of AC, AAG, ACT and GACA motifs with histone genes in an interstitial region of chromosome 8, whereas (AG) was close to them (Table 1.6; Fig. 1.2c, e, h). Only two of these motifs (AC, AAG) were found in the vicinity of histone gene spacers in the 454 analysis. Likewise in *E. plorans*, the 454 analysis in *L. migratoria* detected more microsatellite-histone associations than the FISH analysis (Tables 1.6 and 1.S2).

Discussion

The two grasshopper species analyzed here, *E. plorans* and *L. migratoria*, are non-model species with very limited genomic information about microsatellite occurrence, abundance and chromosome location (Zhang *et al.*, 2003; Cuadrado & Jouve, 2010; Muñoz-Pajares *et al.*, 2011). NGS technology is rapidly increasing genomic microsatellite characterization in model and non-model species, but studies addressing their characterization at both genomic and chromosomal levels have not virtually been done (Kejnovskỳ *et al.*, 2013) . We have performed this combined approach in two grasshopper species with genomes of different size, that in *E. plorans* (11.11 Gb) being almost double that in *L. migratoria* (6.3 Gb). The 454 NGS and FISH results have shown that, in contrast to the frequently reported species-specific distribution pattern (for review, see Tóth *et al.* 2000; Sharma *et al.* 2007), both species show a strikingly similar genome organization and composition regarding to their microsatellite abundance, motif-types distribution and chromosome location. The repetitive DNA sequences found in the neighborhood of microsatellites are also coincident in both species, with high prevalence of TEs, histone gene spacers and ribosomal DNA IGS.

In fact, the association analysis showed that, in both species, these were the repetitive DNAs showing the strongest association with microsatellites, presumably because their smaller length (compared to TEs) and their interstitial location within these spacers increase the likelihood of finding microsatellite-containing reads in them.

The similarities found between the two species analyzed are remarkable taking into account that microsatellite abundance varies extensively among Arthropoda species (Meglécz *et al.*, 2012) and even between closely related species, as shown in *Drosophila* (Ross *et al.*, 2003; Pannebakker *et al.*, 2010). Meglécz *et al.* (2012) analyzed 12 insect species and found maximum microsatellite coverage of 10,467 nucleotides per Mb of library in *Necterosoma penicilatus* (Coleoptera) and a minimum of 1,183 nucleotides per Mb in *Ischnura heterosticta* (Odonata). Compared to these insect groups, the two species analyzed here show intermediate microsatellite coverage (4,612-5,441). The fact that both species belong to two phylogenetically distant subfamilies (Eyprepocnemidinae and Oedipodinae) within the family Acrididae raise the possibility that their similar microsatellite abundance and composition is due to similar mechanisms of origin and evolution of these genomic elements in grasshoppers, rather than to conservation of a common genomic structure.

Assuming that the 454 library obtained is representative of the whole genome (Martin *et al.*, 2010), we can infer microsatellite genomic contents of 0.46% in *E. plorans* and 0.54% in *L. migratoria*. This finding is consistent with the well established idea that although microsatellites are present in all vertebrate and invertebrate species studied so far, invertebrate genomes are less prone to accumulate them (Tóth *et al.*, 2000; Li *et al.*, 2002), although the genomes of some invertebrates actually show a remarkable microsatellite content (Meglécz *et al.*, 2012). Microsatellite abundance in grasshoppers was thus lower than those found in the human (>1.5%) (Baltimore, 2001; Sharma *et al.*, 2007), mouse (2%) (Sharma *et al.*, 2007) and snake (2.8%) Castoe *et al.* (2012) genomes, and were within the ranges observed in arthropod genomes (1% or lower) (Pannebakker *et al.*, 2010).

Although it is commonly accepted that microsatellite abundance increases with genome size, many exceptions have been reported in animals and plants (Ustinova *et al.*, 2006; Pannebakker *et al.*, 2010). Our present results have also shown an inverse relationship between genome size and microsatellite content, as the species showing the highest microsatellite abundance (0.54% in *L. migratoria* vs. 0.46% in *E. plorans*) showed the smallest genome (6.3 Gb vs. 11.11 Gb, respectively; see Ruiz-Ruano *et al.*, 2011). In addition, the frequency of microsatellite loci per Mb (283 and 240, respectively) was lower than that reported in hymenopteran species with much smaller genomes, such as the bee *Apis mellifera* with 308 microsatellite loci per Mb in its 264.1 Mb genome (Pannebakker *et al.*, 2010). Furthermore, in fungi, Lim *et al.* (2004) reported that microsatellite content was not proportional to genome size in 14 species analyzed, in consistency with a similar observation by Sharma *et al.* (2007) and Schoebel *et al.* (2013) in a number of different eukaryote species. An inverse relationship between genome size and microsatellite frequency has also been observed by Morgante *et al.* (2002) in plants.

The different microsatellite motif-length classes showed uneven distribution in the genomes of *E. plorans* and *L. migratoria*, with dinucleotides representing more than 60% of the microsatellite genome coverage, and mono- and trinucletoides being the next motifs in abundance. Relative abundance of microsatellite motif-length classes shows considerable variation among species (Sharma *et al.*, 2007), but dinucleotides and mononucleotides are the prevalent types in most cases (Ross *et al.*, 2003; Sharma *et al.*, 2007; Meglécz *et al.*, 2012; Kejnovský *et al.*, 2013; Sawaya *et al.*, 2013; Schoebel *et al.*, 2013). However, tri-, tetra- and pentanucleotides have also been reported as predominant SSR classes in several species of Diptera (Pannebakker *et al.*, 2010) and snakes (Castoe *et al.*, 2012). Interestingly, Sharma *et al.* (2007) observed a prevalence of trinucleotide motifs in species with low repetitive DNA content. The genomes of *E. plorans* and *L. migratoria* contain high amounts of repetitive DNA, especially transposable elements and satellite DNA (López-León *et al.*, 1994; Montiel *et al.*, 2012; Wang *et al.*, 2014; Chapters 2 and 4), which may explain the higher abundance of dinucleotide and mononucleotide motifs, since these two classes are more abundant in non coding regions whereas trinucleotide repeats are especially frequent in coding regions and ESTs (Li *et al.*, 2002; Morgante *et al.*, 2002).

Microsatellite locus size (10.03 and 11.10 repeats in *E. plorans* and *L. migratoria*, respectively) and distance between loci (4.16 kb in *E. plorans*, and 3.53 kb in *L. migratoria*, on average) also support SSR association with non coding sequences because exonic regions use to bear microsatellite repeats at low density (3 or less units at about 40 kb distance) (Li *et al.*, 2002, 2011). However, the presence of more distant microsatellite loci with less than 5 repeats, located in coding regions, would have gone unnoticed with our search criterion. Notwithstanding, our FISH experiments revealed that mono-, di-, tri- and tetranucleotide microsatellites showed a dispersed distribution over euchromatic regions in all chromosomes, covering nearly their entire length and being absent from the heterochromatic pericentromeric regions, in consistency with Guo *et al.* (2009) claiming that microsatellite content is lower in centromeric and pericentromeric regions. This same dispersed-euchromatic pattern of microsatellite distribution has been previously observed in the grasshopper *Abracris flavolineata* (Milani & Cabral-de Mello, 2014) and for AG motif in *E. plorans* (Cuadrado & Jouve, 2010). Interestingly, this euchromatic chromosomal distribution of microsatellites in *E. plorans* was remarkably coincident with that of TEs (Montiel *et al.*, 2012), and the same is also observed in *L. migratoria* (not shown). This is strongly supported by our finding of high percentage (84.7% in *E. plorans* and 95.1% in *L. migratoria*) of microsatellite repeats found in the neighborhood of several types of transposons and retrotransposons (Tables 1.5, 1.S1 and 1.S2). Therefore, apart from other mechanisms inducing microsatellite propagation, such as unequal crossing over, DNA recombination and DNA repair or slippage (Dover 1993, McMurray 1995; Hancock 1996), the expansion of microsatellites observed in *E. plorans* and *L. migratoria* chromosomes may be a result of their close localization and association with TEs. Hence, the movement of some transposons, such as MITEs, has been suggested to be responsible for the proliferation of hitchhiking microsatellites in several insect species (for review see Coates *et al.* 2011). Accordingly, Primmer *et al.* (1997) have pointed out the lack of association between microsatellites and SINES/LINES to explain the low frequency of microsatellites in avian genomes. Alternatively, microsatellites could be good targets for transposons insertion (Kejnovskỳ *et al.*, 2013). All these observations are consistent with our results, as the grasshopper species with higher microsatellite coverage (*L. migratoria*) also showed higher microsatellite repeats abundance in the vicinity of transposons in the annotated 454 reads.

Other non coding sequences where microsatellites could be located in *E. plorans* and *L. migratoria* euchromatin are pseudogenes, introns, intergenic spacers and untranslated 3′ and 5′UTR sequences. Indeed, gene spacers (e.g. in rDNA and histone clusters) were strongly associated with microsatellites in the 454 reads of the two species analyzed here (Tables 1.5, 1.S1, 1.S2, 1.S6 and 1.S8). This probably diminish their functional effects thus making them more tolerable despite residing within functional genes.

In any case, we cannot rule out that trinucleotide motifs, representing 15.81% (*E. plorans*) and 12.50% (*L. migratoria*) of total microsatellite coverage, could predominate in coding regions, as observed in other species (Tóth *et al.*, 2000; Li *et al.*, 2002; Morgante *et al.*, 2002) although, in general, microsatellites are comparatively less frequent in exonic regions than in introns or non coding regions, except in some cases like *Populus*, where exonic microsatellites are threefold more frequent than intronic ones (Li *et al.*, 2011). Finally, it is also conceivable that some of the abundant dinucleotide motifs could be located in coding regions in *E. plorans* and *L. migratoria* euchromatin, since recent evidence has shown that long exonic dinucleotide SSRs are submitted to strong selective constraints (Haasl & Payseur, 2014).

Taken together, our results suggest that, in both grasshopper species, i) most microsatellites reside into euchromatic non coding regions, most of them placed in the vicinity of TEs, whereas others are located within histone-gene spacers and rDNA IGS, and ii) their abundance is not high enough to explain the huge size of grasshopper genomes.

The relative distribution of different motif sequences within each motif-lenght class was also rather similar in the *E. plorans* and *L. migratoria* genomes. The predominance of A motifs among mononucleotides, and AC and AG motifs among dinucleotides, in both species, is coincident with the general pattern observed in animal genomes. Indeed, A predominates in animals and plants (Meglécz *et al.*, 2012) and AC and AG seem to be the most common dinucleotide motif in animals whereas AG and AT are predominant in plant genomes (Powell *et al.*, 1996; Ross *et al.*, 2003; Meglécz *et al.*, 2012). Likewise, AAC and AAT are prevalent among trinucleotide motifs in *E. plorans* and *L. migratoria*, in consistency with other invertebrate genomes (Pannebakker *et al.*, 2010; Meglécz *et al.*, 2012). The remaining motifs (tetra-, penta- and hexanucleotides) are scarce and thus probably play a minor role in genome architecture and evolution.

Our 454 NGS and FISH results were roughly consistent in each of the two species analyzed here, with the most abundant microsatellites (AC, AG and A) yielding the brighter FISH signals. Remarkably, all SSRs that did not yield any hybridization signals in *E. plorans* (AAT, AGG, ACT, CAT, CAC, ACG, CAG, GCC) and *L. migratoria* (AAT, AAC, AGG, CAT, CAC, ACG, CAG, GCC) were trinucleotide motifs, with a high level of coincidence between species (seven out of eight are identical). All these trinucleotides were actually present in the *E. plorans* and *L. migratoria* 454 libraries, but at low frequency (see Tables 1.1, 1.6 and 1.S3). By contrast, in both species, GACA and GATA tetranucleotides yielded conspicuous FISH signals (Fig. 1.1g, h and 1.2g, i) despite their scarcity in the 454 libraries (Tables 1.1 and 1.6). This might be explained by the possible existence of non canonical microsatellite loci where GATA or GACA repeats are intermixed with other sequences which allows detection by FISH but prevents detection in the 454 libraries, with the search criterion employed.

Some microsatellites showed clustered distribution in the *E. plorans* and *L. migratoria* chromosomes, yielding band-like hybridization patterns. Clustered microsatellite distribution has been previously reported in several species (Cuadrado *et al.*, 2000; Ananiev *et al.*, 2005; Santos *et al.*, 2010). Remarkably, in *E. plorans*, a distal band on chromosome 2 is shared by nearly all SSR motifs, highlighting this chromosome region as enriched in different microsatellites (Fig. 1.1; Table 1.1). In *L. migratoria*, clustered SSRs locate predominantly on interstitial regions of the chromosomes 8 and 9 (Fig. 1.2; Table 1.6). In addition, we observed variation for the presence of some other clustered SSRs, as reported in maize (Ananiev *et al.*, 2005). It has been argued that microsatellite clustering might suggest a functional role (Santos *et al.*, 2010). In fact, the distal microsatellite cluster in chromosome 2 of *E. plorans* and the interstitial one in chromosome 8 of *L. migratoria* colocalizes with histone H3 genes and at least one of the H3-colocating motifs (AC) has been shown to modulate gene expression in the ascomycete Podospora anserina (Khashnobish *et al.*, 1999) and is one of the most common motifs in transcription start site in human promoters (Sawaya *et al.*, 2013). Indeed, AC is the most predominant motif being found associated with histone genes spacers in both species analyzed (Table 1.5). In addition to AC, other microsatellites showed FISH bands co-locating with the histone cluster (A, AAC, AAG, GATA and GACA in *E. plorans*, and, AAG, ACT and GACA in *L. migratoria*) and most of them (excepting GATA in *E. plorans* and ACT, GACA in *L. migratoria*) were found in the 454 reads with microsatellites (Tables 1.1, 1.6, 1.S1, 1.S2, 1.S6 and 1.S8). These latter molecularly undetected motifs presumably belong to minor infrequent alleles differing from the consensus sequence going unnoticed by the 454 analysis but still being detectable by FISH.

We have not observed a preferential accumulation of microsatellites on the X chromosome, despite the growing number of reports on motifs differentially accumulated on sexual chromosomes of animals and plants (Kejnovskỳ *et al.*, 2013; Poltronieri *et al.*, 2013). This suggests a minor role for microsatellites in sexual chromosome differentation and evolution in both grasshopper species.

B chromosomes in *E. plorans* and *L. migratoria* are mainly heterochromatic, and they are rather poor in microsatellites, in consistency with the scarcity observed in the pericentromeric heterochromatic regions of A chromosomes in both species. In *E. plorans* B chromosomes, SSR presence was restricted to the distal region and the small interstitial DAPI$^-$ bands located between the DAPI$^+$ heterochromatic blocks, which include several types of satellite DNA (López-León *et al.*, 1994; Martín-Peciña *et al.*, in preparation). In consistency with these results, Muñoz-Pajares *et al.* (2011) reported the presence of a $(GATTA)_6$ microsatellite within a SCAR (sequence-characterized amplified region) located in the external spacers of the 45S rDNA which is distally located on the B chromosome. This GATTA/ AATCT motif has been found in our 454 data but at very low frequency (0.14%). In *L. migratoria*, B chromosomes show higher abundance of SSRs in the proximal euchromatic third, thus showing similar microsatellite compartmentalization as A chromosomes. Likewise, microsatellite clusters have been detected on the long arm of maize B chromosome (Ananiev *et al.*, 2005) and also on the B chromosomes in *Crepis capillaris* (Jamilena *et al.*, 1994) and rye (Langdon *et al.*, 2000). The B chromosome of the grasshopper *Abracris flavolineata* is also enriched in microsatellites repeats (Milani & Cabral-de Mello, 2014). Although microsatellites in B chromosomes are unlikely related with gene regulation, since B chromosomes are mostly silenced, we cannot rule out that they could have something to do with the high mutability of *E. plorans* B chromosomes since this species shows the highest number of B variants ever found for a B chromosome system (López-León *et al.*, 1994; Bakkali & Camacho, 2004). Unveiling the precise role of transposable elements in shaping microsatellite abundance and chromosome distribution is an interesting prospect for future research.

Acknowledgements

This study was supported by grants from the Spanish Ministerio de Ciencia y Tecnología (CGL2009-11917) and Plan Andaluz de Investigacion (CVI-6649), and was partially performed by FEDER funds. E.E. Montiel was supported by a Junta de Andalucía fellowship

References

Ananiev EV, Chamberlin MA, Klaiber J, Svitashev S (2005) Microsatellite megatracts in the maize (*Zea mays* L.) genome. *Genome*, **48**, 1061–1069.

Areshchenkova T, Ganal MW (1999) Long tomato microsatellites are predominantly associated with centromeric regions. *Genome*, **42**, 536–544.

Augustinos AA, Asimakopoulou AK, Papadopoulos NT, Bourtzis K (2011) Cross-amplified microsatellites in the European cherry fly, *Rhagoletis cerasi*: medium polymorphic–highly informative markers. *Bulletin of entomological research*, **101**, 45–52.

Bakkali M, Camacho JPM (2004) The B chromosome polymorphism of the grasshopper *Eyprepocnemis plorans* in North Africa: III. Mutation rate of B chromosomes. *Heredity*, **92**, 428–433.

Baltimore D (2001) Our genome unveiled. *Nature*, **409**, 814–816.

Bao Z, Eddy SR (2002) Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome research*, **12**, 1269–1276.

Basset P, Yannic G, Yang F, *et al.* (2006) Chromosome localization of microsatellite markers in the shrews of the *Sorex araneus* group. *Chromosome Research*, **14**, 253–262.

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*, **27**, 573.

Blondin L, Badisco L, Foucart A, *et al.* (2013) Characterization and comparison of microsatellite markers derived from genomic and expressed libraries for the desert locust. *Journal of Applied Entomology*, **137**, 673–683.

Cabrero J, López-León MD, Bakkali M, Camacho JPM (1999) Common origin of B chromosome variants in the grasshopper *Eyprepocnemis plorans*. *Heredity*, **83**, 435–439.

Cabrero J, López-León MD, Teruel M, Camacho JPM (2009) Chromosome mapping of H3 and H4 histone gene clusters in 35 species of acridid grasshoppers. *Chromosome research*, **17**, 397–404.

Camacho JPM (2005) B chromosomes. *The Evolution of the Genome (ed. T. R. Gregory)*, pp. 223–286.

Camacho JPM, Cabrero J, Viseras E, López-León MD, Navas-Castillo J, Alché JD (1991) G banding in two species of grasshopper and its relationship to C, N, and fluorescence banding techniques. *Genome*, **34**, 638–643.

Carmona A, Friero E, de Bustos A, Jouve N, Cuadrado A (2013) Cytogenetic diversity of SSR motifs within and between *Hordeum* species carrying the H genome: *H. vulgare* L. and *H. bulbosum* L. *Theoretical and applied genetics*, **126**, 949–961.

Castoe TA, Streicher JW, Meik JM, *et al.* (2012) Thousands of microsatellite loci from the venomous coralsnake *Micrurus fulvius* and variability of select loci across populations and related species. *Molecular ecology resources*, **12**, 1105–1113.

Coates BS, Kroemer JA, Sumerford DV, Hellmich RL (2011) A novel class of miniature inverted repeat transposable elements (MITEs) that contain hitchhiking (GTCY)n microsatellites. *Insect molecular biology*, **20**, 15–27.

Cuadrado A, Jouve N (2007a) The nonrandom distribution of long clusters of all possible classes of trinucleotide repeats in barley chromosomes. *Chromosome Research*, **15**, 711–720.

Cuadrado A, Jouve N (2007b) Similarities in the chromosomal distribution of AG and AC repeats within and between *Drosophila*, human and barley chromosomes. *Cytogenetic and genome research*, **119**, 91–99.

Cuadrado Á, Jouve N (2010) Chromosomal detection of simple sequence repeats (SSRs) using nondenaturing FISH (ND-FISH). *Chromosoma*, **119**, 495–503.

Cuadrado Á, Jouve N (2011) Novel simple sequence repeats (SSRs) detected by ND-FISH in heterochromatin of *Drosophila melanogaster*. *BMC genomics*, **12**, 1.

Cuadrado A, Schwarzacher T, Jouve N (2000) Identification of different chromatin classes in wheat using in situ hybridization with simple sequence repeat oligonucleotides. *Theoretical and Applied Genetics*, **101**, 711–717.

Guo WJ, Ling J, Li P (2009) Consensus features of microsatellite distribution: microsatellite contents are universally correlated with recombination rates and are preferentially depressed by centromeres in multicellular eukaryotic genomes. *Genomics*, **93**, 323–331.

Haasl RJ, Payseur BA (2014) Remarkable selective constraints on exonic dinucleotide repeats. *Evolution*, **68**, 2737–2744.

Hancock JM (2002) Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. *Genetica*, **115**, 93–103.

Hatanaka T, Henrique-Silva F, Galetti Jr PM (2003) A polymorphic, telomeric-like sequence microsatellite in the Neotropical fish *Prochilodus*. *Cytogenetic and genome research*, **98**, 308–310.

Hunter ME, Hart KM (2013) Rapid microsatellite marker development using next generation pyrosequencing to inform invasive burmese python—*Python molurus bivittatus*—management. *International journal of molecular sciences*, **14**, 4793–4804.

Insuan S, Deowanish S, Klinbunga S, Sittipraneed S, Sylvester HA, Wongsiri S (2007) Genetic differentiation of the giant honey bee (*Apis dorsata*) in Thailand analyzed by mitochondrial genes and microsatellites. *Biochemical Genetics*, **45**, 345–361.

Iquebal MA, Arora V, Verma N, Rai A, Kumar D, others (2013) First whole genome based microsatellite DNA marker database of tomato for mapping and variety identification. *BMC Plant Biology*, **13**, 197.

Jamilena M, Rejón CR, Rejon MR (1994) A molecular analysis of the origin of the *Crepis capillaris* B chromosome. *Journal of Cell Science*, **107**, 703–708.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*, **110**, 462–467.

Kejnovskỳ E, Michalovova M, Steflova P, *et al.* (2013) Expansion of microsatellites on evolutionary young Y chromosome. *PLoS One*, **8**, e45519.

Khashnobish A, Hamann A, Osiewacz HD (1999) Modulation of gene expression by (CA) n microsatellites in the filamentous ascomycete *Podospora anserina*. *Applied microbiology and biotechnology*, **52**, 191–195.

Kubat Z, Hobza R, Vyskot B, Kejnovsky E (2008) Microsatellite accumulation on the Y chromosome in *Silene latifolia*. *Genome*, **51**, 350–356.

Langdon T, Seago C, Jones RN, *et al.* (2000) *De novo* evolution of satellite DNA on the rye B chromosome. *Genetics*, **154**, 869–884.

Li S, Yin T, Wang M, Tuskan GA (2011) Characterization of microsatellites in the coding regions of the *Populus* genome. *Molecular breeding*, **27**, 59–66.

Li YC, Korol AB, Fahima T, Beiles A, Nevo E (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular ecology*, **11**, 2453–2465.

Lim S, Notley-McRobb L, Lim M, Carter DA (2004) A comparison of the nature and abundance of microsatellites in 14 fungal genomes. *Fungal Genetics and Biology*, **41**, 1025–1036.

Lohe AR, Hilliker AJ, Roberts PA (1993) Mapping simple repeated DNA sequences in heterochromatin of *Drosophila melanogaster*. *Genetics*, **134**, 1149–1174.

López-León MD, Neves N, Schwarzacher T, Heslop-Harrison JP, Hewitt GM, Camacho JPM (1994) Possible origin of a B chromosome deduced from its DNA composition using double FISH technique. *Chromosome Research*, **2**, 87–92.

Malausa T, Gilles A, Meglecz E, *et al.* (2011) High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries. *Molecular Ecology Resources*, **11**, 638–644.

Manrique-Poyato MI, López-León MD, Gómez R, Perfectti F, Camacho JPM (2013) Population genetic structure of the grasshopper *Eyprepocnemis plorans* in the south and east of the Iberian Peninsula. *PloS one*, **8**, e59041.

Martin JF, Pech N, Meglécz E, *et al.* (2010) Representativeness of microsatellite distributions in genomes, as revealed by 454 GS-FLX Titanium pyrosequencing. *BMC genomics*, **11**, 560.

Meglécz E, Nève G, Biffin E, Gardner MG (2012) Breakdown of phylogenetic signal: a survey of microsatellite densities in 454 shotgun sequences from 154 non model eukaryote species. *PLoS One*, **7**, e40861.

Milani D, Cabral-de Mello DC (2014) Microsatellite organization in the grasshopper *Abracris flavolineata* (Orthoptera: Acrididae) revealed by FISH mapping: remarkable spreading in the A and B chromosomes. *PloS one*, **9**, e97956.

Montiel EE, Cabrero J, Camacho JPM, López-León MD (2012) Gypsy, RTE and Mariner transposable elements populate *Eyprepocnemis plorans* genome. *Genetica*, **140**, 365–374.

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature genetics*, **30**, 194–200.

Muñoz-Pajares AJ, Martínez-Rodríguez L, Teruel M, Cabrero J, Camacho JPM, Perfectti F (2011) A single, recent origin of the accessory B chromosome of the grasshopper *Eyprepocnemis plorans*. *Genetics*, **187**, 853–863.

Nanda I, Zischler H, Epplen C, Guttenbach M, Schmid M (1991) Chromosomal organization of simple repeated DNA sequences. *Electrophoresis*, **12**, 193–203.

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792–793.

Pannebakker BA, Niehuis O, Hedley A, Gadau J, Shuker DM (2010) The distribution of microsatellites in the *Nasonia parasitoid* wasp genome. *Insect molecular biology*, **19**, 91–98.

Pardue ML, Lowenhaupt K, Rich A, Nordheim A (1987) (dC-dA) n.(dG-dT) n sequences have evolutionarily conserved chromosomal locations in *Drosophila* with implications for roles in chromosome structure and function. *The EMBO journal*, **6**, 1781.

Poltronieri J, Marquioni V, Bertollo LAC, *et al.* (2013) Comparative chromosomal mapping of microsatellites in *Leporinus* species (Characiformes, Anostomidae): Unequal accumulation on the W chromosomes. *Cytogenetic and genome research*, **142**, 40–45.

Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. *Trends in plant science*, **1**, 215–222.

Primmer CR, Raudsepp T, Chowdhary BP, Møller AP, Ellegren H (1997) Low frequency of microsatellites in the avian genome. *Genome Research*, **7**, 471–482.

Ross CL, Dyer KA, Erez T, Miller SJ, Jaenike J, Markow TA (2003) Rapid divergence of microsatellite abundance among species of *Drosophila*. *Molecular Biology and Evolution*, **20**, 1143–1157.

Ruiz-Ruano FJ, Ruiz-Estévez M, Rodríguez-Pérez J, López-Pino JL, Cabrero J, Camacho JPM (2011) DNA amount of X and B chromosomes in the grasshoppers *Eyprepocnemis plorans* and *Locusta migratoria*. *Cytogenetic and genome research*, **134**, 120–126.

Santos J, Serra L, Solé E, Pascual M (2010) FISH mapping of microsatellite loci from *Drosophila subobscura* and its comparison to related species. *Chromosome research*, **18**, 213–226.

Sawaya S, Bagshaw A, Buschiazzo E, *et al.* (2013) Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PloS one*, **8**, e54710.

Schlötterer C, Goldstein DB (1999) *Microsatellites: evolution and applications*. Oxford University Press, Oxford.

Schlötterer C, Pemberton J (1998) The use of microsatellites for genetic analysis of natural populations—a critical review. In *Molecular approaches to ecology and evolution (DeSalle R, Schierwater B (eds))*, pp. 71–86. Birkhaäuser, Berlin.

Schoebel CN, Brodbeck S, Buehler D, *et al.* (2013) Lessons learned from microsatellite development for nonmodel organisms using 454 pyrosequencing. *Journal of Evolutionary Biology*, **26**, 600–611.

Sharma PC, Grover A, Kahl G (2007) Mining microsatellites in eukaryotic genomes. *Trends in biotechnology*, **25**, 490–498.

Shi J, Huang S, Zhan J, *et al.* (2014) Genome-wide microsatellite characterization and marker development in the sequenced *Brassica* crop species. *DNA research*, **21**, 53–68.

Smit AFA, Hubley R, Green P (2013) RepeatMasker Open-4.0.

Soltis DE, Gitzendanner MA, Stull G, *et al.* (2013) The potential of genomics in plant systematics. *Taxon*, **62**, 886–898.

Tautz D, Renz M (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic acids research*, **12**, 4127–4138.

Teruel M, Cabrero J, Perfectti F, Camacho JPM (2010) B chromosome ancestry revealed by histone genes in the migratory locust. *Chromosoma*, **119**, 217–225.

Tóth G, Gáspári Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome research*, **10**, 967–981.

Ustinova J, Achmann R, Cremer S, Mayer F (2006) Long repeats in a huge genome: microsatellite loci in the grasshopper *Chorthippus biguttulus*. *Journal of molecular evolution*, **62**, 158–167.

Wang X, Fang X, Yang P, *et al.* (2014) The locust genome provides insight into swarm formation and long-distance flight. *Nature communications*, **5**.

Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. *Molecular ecology*, **11**, 1–16.

Zhang DX, Yan LN, Ji YJ, Kang LE, Hewitt GM, Huang ZS (2003) Isolation, characterization and cross-species amplification of eight microsatellite DNA loci in the migratory locust (*Locusta migratoria*). *Molecular Ecology Notes*, **3**, 483–486.

Supplementary Information

Supplementary Figures

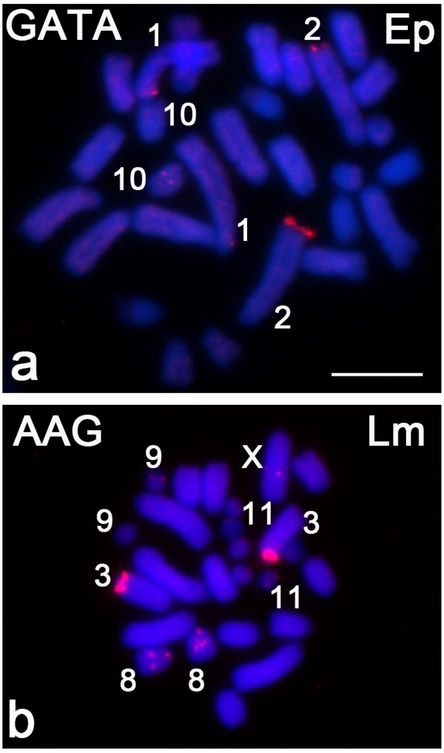


Figure 1.S1: Embryo mitotic metaphase cells of *E. plorans* (a) and *L. migratoria* (b) showing DAPI staining + FISH. Note the variation observed for hybridization clustered sites (see also Fig. 1 and 2). Bar = 5 μm.

Supplementary Tables

Supplementary tables for this chapter are available in https://dx.doi.org/10.6084/m9.figshare.3255556.

Table 1.S1: Number of microsatellite repeats found in the vicinity of non-microsatellite repeats in the *E. plorans* 454 library (given separately for each microsatellite motif). Note the predominance of transposable elements.

Table 1.S2: Number of microsatellite repeats found in the vicinity of non-microsatellite repeats in the *L. migratoria* 454 library (given separately for each microsatellite motif). Note the predominance of transposable elements.

Table 1.S3: Number of repeats per SSR motif sequence and their relative abundance in the *E. plorans* and *L. migratoria* reads present in the two 454 libraries analyzed.

Table 1.S4: Number of microsatellite loci with different number of repeats (expressed per microsatellite motif) found in the *E. plorans* 454 library.

Table 1.S5: Number of contigs contained in each RepeatModeler database.

Table 1.S6: Association between non-microsatellite repeats and microsatellites performed on SSR-containing and SSR-lacking reads in the *E. plorans* 454 library (with Bonferroni correction). Significant associations are highlighted in red.

Table 1.S7: Number of microsatellite loci with different number of repeats (expressed per microsatellite motif) found in the *L. migratoria* 454 library.

Table 1.S8: Association between non-microsatellite repeats and microsatellites performed on SSR-containing and SSR-lacking reads in the *L. migratoria* 454 library (without Bonferroni correction). Significant associations are highlighted in red, and significant associations lost after applying the Bonferroni correction are highlighted in bold-type letter.

Chapter 2. High-throughput analysis of the satellitome illuminates satellite DNA evolution

Francisco J. Ruiz-Ruano, María Dolores López-León, Josefa Cabrero and Juan Pedro M. Camacho

Departamento de Genética, Universidad de Granada

Abstract. Satellite DNA (satDNA) is a major component yet the great unknown of eukaryote genomes and clearly underrepresented in genome sequencing projects. Here we show the high-throughput analysis of satDNA content in the migratory locust by means of the bioinformatic analysis of Illumina reads with the RepeatExplorer and RepeatMasker programs. This unveiled 62 satDNA families and we propose the term "satellitome" for the whole collection of different satDNA families in a genome. The finding that satDNAs were present in many contigs of the migratory locust draft genome indicates that they show many genomic locations invisible by fluorescent *in situ* hybridization (FISH). The cytological pattern of five satDNAs showing common descent (belonging to the SF3 superfamily) suggests that non-clustered satDNAs can become into clustered through local amplification at any of the many genomic loci resulting from previous dissemination of short satDNA arrays. The fact that all kinds of satDNA (micro- mini- and satellites) can show the non-clustered and clustered states suggests that all these elements are mostly similar, except for repeat length. Finally, the presence of VNTRs in bacteria, showing similar properties to non-clustered satDNAs in eukaryotes, suggests that this kind of tandem repeats show common properties in all living beings.

Introduction

Eukaryote genomes are plenty of repetitive elements including transposable elements (TEs), tandem repeats, segmental duplications, ribosomal DNA, multi-copy gene families, pseudogenes, etc. which, collectively, constitute the repeatome (Kim *et al.*, 2014). Satellite DNA consists of a single sequence tandemly repeated many times, in contrast to tandemly repeated genes (e.g. ribosomal RNA and histone genes) where the repeating unit consists of several different DNA sequences (i.e. genes and spacers). Satellite DNA has been classified into microsatellites, minisatellites and satellites, with no complete consensus about the precise length limits (Tautz & Renz, 1984; Richard *et al.*, 2008). Although satellite DNA has traditionally been considered to be junk DNA, some possible functions have been suggested during last years. One of the most accepted functional roles for satDNA is its implication in centromeric function (Plohl *et al.*, 2014), but other possible functional roles have also been suggested in relation with heterochromatin formation through the siRNA pathway (Lee *et al.*, 2006; Usakin *et al.*, 2007).

The name "satellite DNA" is historical since this kind of repetitive DNA was discovered as a small peak in the CsCl ultracentrifugation profile (Kit, 1961). Today this technique is not performed to search for satDNA, since it was replaced by other techniques such as DNA renaturation kinetics (Britten *et al.*, 1974), restriction digestion and electrophoresis yielding a ladder pattern (Singer, 1982) and, most recently, by the bioinformatic analysis of a huge collection of short DNA sequences yielded by Next Generation Sequencing (NGS) (Novák *et al.*, 2013). Anyway, the term "satellite DNA" is still useful because it is simple, descriptive and profusely used in the literature. On this basis, we are proposing here the name "satellitome" for the whole collection of satDNAs in a genome.

The recent publication of a draft genome of the migratory locust (*Locusta migratoria*) represents a milestone as it is the largest animal genome hitherto sequenced (Wang *et al.*, 2014). There is no doubt that it has provided excellent information for performing genomic work in other insects even though annotation is not complete. However, as in other sequenced genomes, information about the repetitive components of the genome is rather scarce, especially for satDNA. We have recently reported microsatellite content in *L. migratoria* at both genomic and cytogenetic levels (Chapter 1), but the search for satDNAs through the classical restriction endonuclease digestion and electrophoresis approach failed in this species (MD López-León and P Lorite, personal communication). Up to now, only 21 satDNAs have been reported in 12 orthopteran species, most of them grasshoppers (Table 2.S1).

Recently, the use of NGS and new bioinformatic tools like RepeatExplorer (Novák *et al.*, 2013) has allowed the high-throughput detection of repetitive DNA, including satellite DNA, the most extreme case being the plant *Luzula elegans* with 37 satDNA families, 20 of which were analyzed by FISH (Heckmann *et al.*, 2013). Here we perform the high-throughput analysis of the satellitome from the information contained in Illumina reads obtained from two individuals of the migratory locust. By means of stepwise clustering of repetitive DNA (Novák *et al.*, 2013) intermingled by substraction of the repetitive elements found in previous steps, we found that the satellitome of *L. migratoria* consists of, at least, 62 different satDNAs with monomer size ranging between 5 and 400 bp. This procedure allowed detection of many poorly abundant satDNAs which would have gone unnoticed through conventional methods. The physical mapping of 59 of them by FISH showed three types of chromosome distribution, with clear predominance of chromosome-specific satDNAs. Finally, this broad catalog of different satDNAs families allowed an analysis for general features which provided new insights on the origin and evolution of this part of the repeatome.

Materials and Methods

Materials

We collected males and females of *Locusta migratoria* at Padul (Granada) in the South of the Iberian Peninsula. Due to the extremely high frequency of supernumerary (B) chromosomes in Spanish field populations (Cabrero *et al.*, 1984), it is very difficult to find B-lacking individuals. For this reason, we obtained males and females from a pet shop whose laboratory culture lacks B chromosomes. We crossed a B-carrying male from Padul with a B-lacking female from the culture, and the male offspring was analyzed cytologically to choose one B-lacking individual, following protocols in Cabrero *et al.* (1984). We then extracted genomic DNA (gDNA) with the GenElute Mamalian Genomic DNA Miniprep kit (Sigma) and sequenced the gDNA library (insert size= 226±81 bp) in the Illumina HiSeq2000 platform yielding about 6 Gb data of 2x101 nt paired-end reads, ~1x coverage for the gDNA [SRA: SRR2911427]. We also used gDNA Illumina reads (2x100 nt) stored in SRA from the *L. migratoria* Chinese individual used for the genome assembling performed by Wang *et al.* (2014) [SRA: SRR764583], randomly selecting the same number of reads as for the Spanish gDNA library. To better characterize the Spanish and Chinese genomes used in this study, we assembled their full mitogenome with MITObim (Hahn *et al.*, 2013) and built a maximum-likelihood tree with PhyML v3 (Guindon *et al.*, 2010) also including the sequences used by Ma *et al.* (2012) [GenBank: JN858148-JN858212, GenBank: NC_011114-NC011115] to define the Northern and Southern lineages (NL and SL, respectively) in this species. As Fig. 2.S6 shows, the Spanish genome was grouped with the populations from the SL, and the Chinese genome was included within the NL.

To test the performace of our satDNA mining protocol in comparison with a typical RepeatExplorer run, we applied it to a gDNA Illumina library (2x101 nt) from *Luzula elegans* [SRA: ERR149838] previously analyzed (Heckmann *et al.*, 2013).

Bioinformatic analysis

We developed satMiner, a toolkit for mining and analyzing satDNA. Scripts and running instructions are freely available in GitHub (https://github.com /fjruizruano/satminer). The satMiner protocol consists of the high-throughput extraction of satDNA sequences and their subsequent analyses. For satDNA mining, we performed a protocol for assembly and identification of satDNA families based on the RepeatExplorer software (Novák *et al.*, 2013). Since the number of reads for a RepeatExplorer run is computationally restricted to a few millions, and in order to identify as many satDNA families as possible, our protocol included filtering out reads showing high similarity with previously known sequences (Fig. 2.1).

The first step consists in discarding low quality reads with Trimmomatic (Bolger *et al.*, 2014), by removing adapters and selecting read pairs with all their nucleotides, i.e. 2x100 or 2x101, with Q>20, using the options "ILLUMINACLIP: TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN: [100/101]". We randomly selected 2x250,000 reads with SeqTK (https://github.com/ lh3/seqtk) and run RepeatExplorer with default options and a custom database of repeated sequences, in addition to Repbase v20.10 (Bao *et al.*, 2015), last accessed October 28, 2015. We manually selected the clusters with spherical or ring-shaped structure and density values (i. e. the mean number of links per read) being higher than 0.1. For each cluster we chose the contigs showing the highest coverage and generated a dotplot with Geneious v4.8 (Drummond *et al.*, 2009). If we detected tandem structure, we split the contigs in monomers to align them and generate a consensus monomer for each contig. We then chose a new collection of reads and those that matched previously detected satDNAs were filtered out with the DeconSeq v0.4.3 software (Schmieder & Edwards, 2011), with default options, before a new RepeatExplorer run was performed. We used satDNA dimers as reference and, in case of

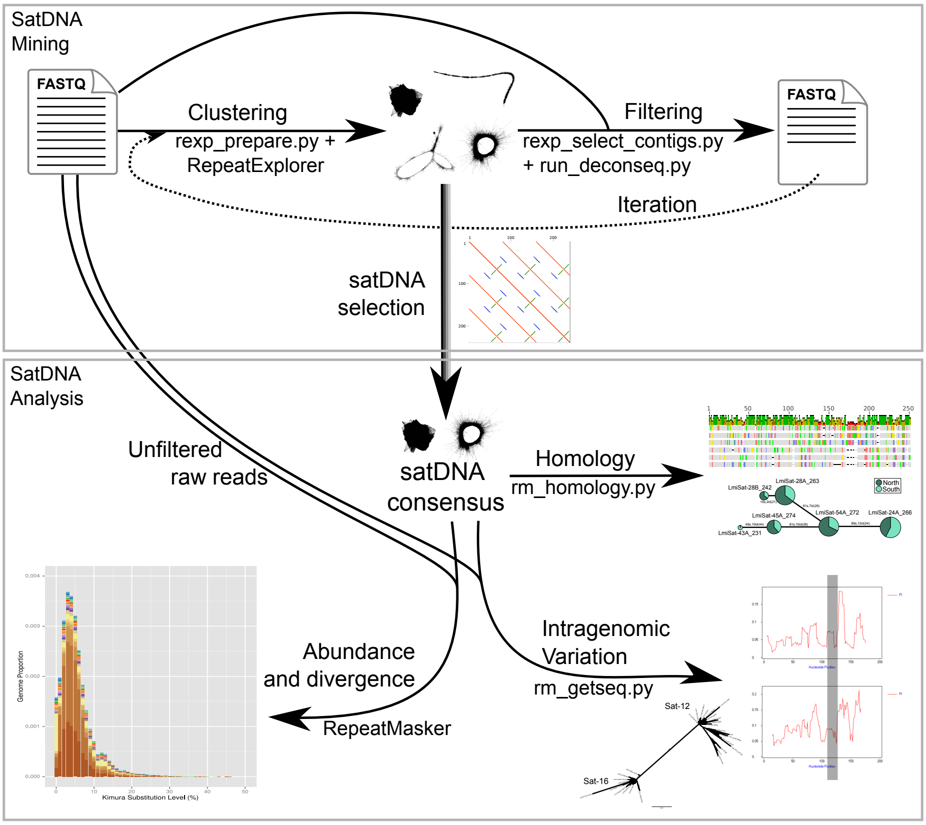


Figure 2.1: Pipeline for satDNA analysis. The mining step starts with raw reads and a typical clustering with RepeatExplorer. This yields linear, spherical or ring-shaped clusters, the two latter types most likely being satDNAs. Each of these clusters is then split into monomers to search for a consensus satDNA sequence. The assembled sequences, and those showing homology with those included in Repbase and a custom database, were used to filter a new set of raw reads before performing a new RepeatExplorer run. Several clustering and filtering steps were performed until no new satDNA appeared. This increased the number of reads analyzed by RepeatExplorer without greatly increasing computing requirements. The satDNA collection obtained is then analyzed for different features such as homology between different consensus sequences and their intragenomic diversity, and a repeat landscape is built.

dimers shorter than 200 bp, we concatenated as many monomers as needed to surpass this length. The mismatched reads were then assembled in a new run of RepeatExplorer to search for the presence of satDNAs being poorly represented in the crude reads but detectable in the filtered ones. This procedure increased very much the number of analyzed reads without dramatically increasing computational effort. Therefore, we run RepeatExplorer with 2x500,000 filtered reads, searched for new satDNAs and filter them out. We repeated this process two more times adding 2x500,000 reads in each iteration, until no new satDNA was detected by RepeatExplorer. We mined satDNAs following the same steps in parallel for the gDNA libraries from the Northern and Southern lineages.

For satellite DNA sequence analysis, we compared the consensus sequences of all satDNAs found in order to investigate possible homology between some of them. For this purpose, we aligned each satDNA against the whole satDNA catalog with RepeatMasker v4.0.5 (Smit *et al.*, 2013), using the Cross_match search engine, recording all matches between satDNAs. When sequences showed less than 80% of identity we considered them as different satDNA families sharing a same superfamily. Sequences showing identity higher than 80% were considered variants of the same family, and those showing identity higher than 95% were considered the same variant. We numbered satDNA families in order of decreasing abundance in the Southern lineage individual [GenBank:KU056702-KU056808]. We built a minimum spanning tree for DNA sequences in each superfamily with Arlequin v3.5 (Excoffier & Lischer, 2010), considering each indel position as a single change and representing the relative abundance among Southern and Northern individuals.

We used RepeatMasker (Smit *et al.*, 2013) with "-s" option to estimate abundance and divergence for each satDNA variant in gDNA libraries. We selected 2x5 millions of paired reads where all nucleotides met quality criteria applied for the satMiner protocol. Abundance estimates provided by RepeatMasker showed highly significant positive correlation with those yielded by RepeatExplorer in both the Southern (Spearman rs= 0.84, N= 15, P= 0.000074) and Northern (rs= 0.97, N= 17, P< 0.000001) lineages. Compared to RepeatExplorer, RepeatMasker has the advantage of working on a much higher number of reads, with reasonable computing times, and it can simultaneously estimate the abundance of all satDNA variants previously collected, whereas several runs of RepeatExplorer are necessary to obtain the whole collection of satDNAs, using different reads thus making it difficult normalization, especially for rare variants. We estimated the average divergence generating a repeat landscape considering distances from the sequences applying the Kimura 2-parameter model with the script calcDivergenceFromAlign.pl within the RepeatMasker suite (Smit *et al.*, 2013). In the resulting output, we calculated the weighed mean divergence for each satDNA family, considering all variants. Additionally, we estimated each satDNA family abundance as the sum of nucleotides for all variants. We normalized for the number of selected nucleotides and represented abundance as percentage of the library, i.e., genome proportion.

We estimated the frequency of random occurrence for some short satDNA monomers. For this purpose, we generated 1,000 Tb, i.e., ~159 genomes, shuffling nucleotides with the uShuffle program (Jiang *et al.*, 2008) preserving the dinucleotide frequencies of the assembled genome of *L. migratoria* (Wang *et al.*, 2014), accession number AVCP000000000. In addition, we analyzed the abundance of some satDNA families in these artificial genomes by using RepeatMasker (Smit *et al.*, 2013). For each satDNA family, we scored the number of contigs where at least 200 bp were present, and also scored the number of nucleotides aligning in each contig.

Primer design and PCR

We tested the reliability of the satDNAs found and synthtesized FISH DNA-probes by PCR amplification of all satDNA families. For this purpose, we aligned each satDNA monomers to get a consensus sequence and selected the most conserved region to design primers in opposite orientation ensuring to minimize the distance between them or even overlapping them up to 3 bp at the 5′ end, when necessary (Table 2.S8, Fig. 2.S7a). For this purpose, we used the Primer3 software (Untergasser *et al.*, 2012) with an optimal melting temperature of 60ºC. Alternatively, for monomers shorter than 50 bp, we designed primers manually with a similar melting temperature and with the less stable extensive dimers predicted by the software PerlPrimer (Marshall, 2004) (Fig. 2.S7b). For families with monomer longer than 50 bp, we performed PCR amplification with a starting denaturation step of 95ºC during 5 min, 35 cycles with 94ºC during 20 s, with 55 to 65ºC as annealing temperature during 40 s and 72ºC during 20 s and a final extension step of 7 min. We checked the resulting products in a 2% agarose gel to see the typical ladder pattern of tandem repeats (Fig. 2.S7a). We trimmed the band of the monomer for the annealing temperature with less smear and extracted the DNA squeezing it in a parafilm square. We reamplified 0.5 μL of the resulting solution. For satDNAs shorter than 50 bp, we reduced the time of annealing to 10 s in order to get longer amplicons. This PCR displayed a smear (Fig. 2.S7b). We performed a reamplification using 0.2 μL of the previous PCR product. We purified all PCR products using the GenElute PCR Clean Up kit (Sigma). We only got success for 59 satDNAs, 25 of which were Sanger sequenced and the reliability of the PCR product was confirmed.

Physical mapping

All these PCR products were labeled by nick translation with 2.5 units of DNA polymerase I/DNase I (Invitrogen), following the standard protocol, to be used as DNA probes for fluorescent *in situ* hybridization (FISH). Mapping of satDNAs was performed following the protocol described in Cabrero *et al.* (2003). FISH probes were labeled with tetramethylrhodamine-5-dUTP (satDNAs) or fluorescein-12-dUTP (rDNA and histone H3 genes) from Roche.

L. migratoria chromosomes are all acrocentric and the autosomes can be classified into three size groups: long (L1 and L2), medium (M3-M8) and short (S9-S11). The X chromosome is the third element in size. Previous research has shown that, in this species, the 45S ribosomal DNA (rDNA) is distally located on L2 and M6, and interstitially on S9 chromosomes (Fox & Santos, 1985), whereas a single histone gene cluster is interstitially located on M8 (Cabrero *et al.*, 2009). We employed these two markers to perform double FISH with selected satDNAs located on M chromosomes scarcely differing in size, to identify the satDNA-carrying chromosome. In addition, it is known that the L1 autosome carries the U1 snRNA gene cluster (Anjos *et al.*, 2015), but we differentiated this chromosome from L2 because the latter carries a distal cluster of rDNA. We distinguished three types of satDNA localization: proximal to centromere in any chromosome arm (p), interstitial in the long arm (i) and distal to centromere in the long arm (d).

Statistical analyses

Statistical analyses included non parametric Spearman rank correlation, Wilcoxon matched pairs test, Wilcoxon one-sample test and Mann-Whitney test, all of them performed with Statistica software. Contingency chi-square tests were performed with the RXC program (George Carmody, University of Ottawa, Canada) by a Monte Carlo approach to calculate statistical significance, with 5,000 permutations. When multiple tests were performed, the resulting probability was corrected by the sequential Bonferroni method (Holm, 1979), represented here as Pb.

Results

High-troughput search for satDNAs

In the first run of RepeatExplorer (RE), performed in parallel on Illumina reads from the SL and NL individuals, we found 26 and 21 satDNAs, respectively. We then selected all available reads of each lineage which, after DeconSeq filtering, lacked homology with all satDNAs and other RE clustered sequences previously found in both lineages. A new RE run detected 11 new satDNAs in SL and 9 in NL. After a new step of filtering and RE analysis, we found 2 new satDNAs in SL and 5 in NL. However, the next filtering and RE step failed to show any new satDNA in both lineages, for which reason we stopped this iterative process. At the end, we found 39 satDNAs in the SL individual and 35 in the NL individual. As a whole, these analyses revealed the existence of 62 different satDNA families, 27 of which were assembled in SL, 23 in NL and 12 in both lineages. Subsequent analyses with RepeatMasker to score satDNA abundance and divergence, revealed that 59 out of the 62 satDNAs were present in both lineages, whereas two of them (LmiSat31-8 and LmiSat43-231) were not found in the NL individual, and one (LmiSat62-23) was not found in the SL individual (Table 2.1). The analysis of variation within these 62 satDNA families showed the presence of 107 sequence variants (i.e. 1-7 per family), defined by monomer length and sequence (Tables 2.1 and 2.S2, and Fig. 2.S1). Collectively, all 62 satDNAs represent about 2.39% of the Southern genome and 2.74% of the Northern one (Table 2.1 and Fig. 2.S2). This low amount of satDNA is consistent with the low amount of constitutive heterochromatin revealed by C-banding in this species (Camacho *et al.*, 1991).

The high number of different satDNA families found in the genome of the migratory locust and the plant *L. elegans* (Heckmann *et al.*, 2013) indicates that eukaryote genomes usually contain a high diversity of satDNA families. During next years, huge amounts of new satDNAs are expected to be uncovered using NGS approaches. We therefore suggest the following simple nomenclature rules to help managing this new information: satDNA name should begin with species abbreviation in Repbase (e.g. Lmi for *Locusta migratoria*) followed by the term "Sat", a catalog number in order of decreasing abundance (according to the first genome analyzed), fol-

Table 2.1: Abundance, divergence, presence in the NL draft genome and chromosome location in SL for 62 satDNA families in *Locusta migratoria*. Length (nt), A+T content (%), number of variants (V), abundance (% of the genome), divergence (%), number of contigs found in the draft genome of *Locusta migratoria*[17], maximum number of repeats per contig (MNRPC), chromosome location (in the Southern lineage) and clustering pattern of all 62 satDNA families and superfamilies (SF). In each family, length and A+T content are given for the most abundant variant. Divergence per family is expressed as percentage of Kimura divergence. Chromosome location was analyzed by FISH in a Spanish population. SL= Southern lineage, NL= Northern lineage. Chromosome locations: t= telomeric, p= proximal to centromere, i= interstitial, d= distal. Chromosome distribution patterns: c= clustered, nc= non-clustered, m = mixed. When a satDNA showed two loci in a same chromosome, their locations were indicated separated by a comma. Totals at the bottom do not include LmiSat07-5 (the telomeric repeat).►

					Abundance		Divergence		Genome (NL)[21]	
SF	SatDNA family	Length	A+T	V	SL	NL	SL	NL	Contigs	MNRPC
1	LmiSat01-193	193	59.59	5	0.98225	0.6903	4.67	5.07	332	15.3
	LmiSat02-176	176	53.41	1	0.47509	0.9996	5.32	5.38	12931	100.0
	LmiSat03-195	195	58.97	6	0.29481	0.2305	5.42	5.96	1003	205.8
	LmiSat04-18	18	50.00	2	0.06194	0.0816	7.2	7.23	108	155.9
	LmiSat05-400	400	51.25	1	0.05431	0.0483	4.65	5.04	91	3.3
	LmiSat06-185	185	59.46	4	0.0541	0.0700	4.76	5.28	274	42.3
	LmiSat07-5-tel	5	60.00	1	0.04438	0.1611	1.75	6.12	57	2867.8
	LmiSat08-168	168	57.74	1	0.03737	0.0467	4.96	4.91	327	28.4
	LmiSat09-181	181	60.22	5	0.02944	0.0072	5.38	7.42	45	60.3
	LmiSat10-9	9	55.56	2	0.02269	0.0290	11.79	11.42	267	242.9
	LmiSat11-37	37	62.16	7	0.01873	0.0069	7.75	8.12	317	105.7
2	LmiSat12-273	273	56.41	3	0.01836	0.0113	3.5	5.29	23	16.0
1	LmiSat13-259	259	57.53	5	0.01697	0.0115	4.38	6.25	137	27.4
	LmiSat14-216	216	51.85	4	0.01426	0.0091	5.39	8.79	70	40.0
	LmiSat15-190	190	55.26	1	0.01426	0.0166	4.09	4.5	212	9.0
2	LmiSat16-278	278	62.59	1	0.0139	0.0082	2.49	3.01	17	9.1
	LmiSat17-75	75	57.33	1	0.01177	0.0033	5.79	6.66	112	7.2
	LmiSat18-210	210	60.48	1	0.01121	0.0267	6.33	4.59	6	2.3
	LmiSat19-89	89	60.67	1	0.01058	0.0034	3.82	6.44	10	4.3
	LmiSat20-15	15	53.33	1	0.01032	0.0201	12.71	14.15	190	255.9
	LmiSat21-38	38	50	1	0.01013	0.0019	2.85	2.91	7	19.7
	LmiSat22-17	17	58.82	1	0.01000	0.0092	10.81	10.28	182	425.5
	LmiSat23-223	223	61.43	1	0.00927	0.0106	4.42	5.73	18	9.5
3	LmiSat24-266	266	56.39	1	0.00895	0.0066	2.06	5.14	51	3.6
	LmiSat25-219	219	39.73	2	0.00834	0.0105	5.88	8.2	21	5.3
4	LmiSat26-240	240	66.2	2	0.00809	0.00436	7.44	9.25	33	8.1
	LmiSat27-57	57	47.37	1	0.0079	0.0103	8.99	9.66	333	326.1
3	LmiSat28-263	263	57.41	2	0.00768	0.0139	1.79	2.22	91	11.5
	LmiSat29-68	68	58.82	1	0.00719	0.0019	9.36	14.48	46	88.9
	LmiSat30-138	138	40.58	1	0.0068	0.0055	5.74	9.03	8	1.7
5	LmiSat31-8	8	50	3	0.00668	–	3.86	–	23	83.0
	LmiSat32-261	261	51.72	1	0.00631	0.0056	5.98	9.18	37	12.3
	LmiSat33-21	21	47.62	1	0.00627	0.0039	7.77	8.35	30	179.0
	LmiSat34-299	299	61.87	1	0.00622	0.0048	6.81	7.39	406	3.4
	LmiSat35-228	228	55.7	1	0.00597	0.0053	2.43	4.64	25	18.3
	LmiSat36-15	15	60	2	0.00585	0.0093	16.88	15.12	279	301.9
4	LmiSat37-238	238	66	1	0.00544	0.0022	6.53	6.52	111	37.4
	LmiSat38-42	42	64.29	1	0.00511	0.0046	14.56	14.94	106	692.0
	LmiSat39-53	53	32.08	1	0.00503	0.0013	6.79	9.17	14	119.0
	LmiSat40-148	148	67.57	1	0.00459	0.0023	2.35	3.05	20	3.7
	LmiSat41-180	180	61.67	1	0.00455	0.0058	3.38	2.14	4	5.5
	LmiSat42-127	127	51.18	1	0.00447	0.0012	2.02	4.60	2	1.6
3	LmiSat43-231	231	53.68	1	0.0044	–	0.68	–	44	2.7
	LmiSat44-17	17	29.41	1	0.00428	0.0005	11.45	11.3	7	53.4
3	LmiSat45-274	274	54.01	1	0.0042	0.0066	8.2	7.22	152	12.4
	LmiSat46-353	353	59.77	1	0.00407	0.0071	15.49	11.38	1799	2.0
	LmiSat47-41	41	41.46	1	0.00369	0.0058	12.46	13.22	48	394.0
	LmiSat48-220	220	58.18	1	0.00366	0.0011	3.8	7.74	18	3.1
	LmiSat49-47	47	42.55	1	0.00362	0.0113	6.24	6.70	127	281.8
5	LmiSat50-16	16	56.25	2	0.00331	0.0169	8.31	8.24	54	64.3
4	LmiSat51-241	241	63.9	1	0.00294	0.0058	7.32	3.97	33	137.5
	LmiSat52-143	143	51.75	1	0.00257	0.0076	22.15	14.01	1796	3.4
	LmiSat53-47	47	40.43	1	0.00248	0.0190	3.16	5.20	9	23.4
3	LmiSat54-272	272	56.25	1	0.00244	0.0051	4.55	4.15	164	51.3
	LmiSat55-90	90	35.56	1	0.00164	0.0074	15.62	8.57	4	3.3
	LmiSat56-19	19	52.63	4	0.00083	0.0067	5.09	4.31	15	97.2
	LmiSat57-230	230	63.04	1	0.00052	0.0047	18.21	3.40	212	25.0
	LmiSat58-86	86	41.86	1	0.00008	0.0127	5.99	3.12	10	4.4
5	LmiSat59-16	16	43.75	3	0.00004	0.0049	18.23	14.54	13	13.3
	LmiSat60-255	255	52.94	1	0.00004	0.0053	1.03	0.99	0	0.0
	LmiSat61-63	63	42.86	1	0.00002	0.0062	14.99	4.60	1	11.0
	LmiSat62-23	23	43.48	1	–	0.0045		4.57	1	9.0

	Chromosome location (SL)												
SatDNA family	L1	L2	X	M3	M4	M5	M6	M7	M8	S9	S10	S11	Pattern
LmiSat01-193	p	p	p	p	p	p	p	p	p	p	p	p	c
LmiSat02-176			p			p		p		p	p	p	c
LmiSat03-195	p			p									c
LmiSat04-18										i,d		i	c
LmiSat05-400										i,d		p	c
LmiSat06-185		p		p					p	p		p	c
LmiSat07-5-tel	t	t	t	t	t	t	t	t	t	t	t	t	c
LmiSat08-168			p										m
LmiSat09-181									p				c
LmiSat10-9										p		p	c
LmiSat11-37				p									c
LmiSat12-273		d											c
LmiSat13-259					p								c
LmiSat14-216				i,i								i	c
LmiSat15-190								p					c
LmiSat16-278		d											c
LmiSat17-75				p									c
LmiSat18-210								p					c
LmiSat19-89		p											c
LmiSat20-15													nc
LmiSat21-38										i,d			m
LmiSat22-17								i					c
LmiSat23-223		d					d			i			c
LmiSat24-266													nc
LmiSat25-219		d											c
LmiSat26-240												i	c
LmiSat27-57													nc
LmiSat28-263		i,i											c
LmiSat29-68				p									c
LmiSat30-138									p				c
LmiSat31-8										p	p		c
LmiSat32-261		d											c
LmiSat33-21										i			c
LmiSat34-299		p											c
LmiSat35-228													nc
LmiSat36-15													nc
LmiSat37-238	i											p	c
LmiSat38-42												i	c
LmiSat39-53										i			c
LmiSat40-148										d			c
LmiSat41-180									i				c
LmiSat42-127						p							c
LmiSat43-231									i				m
LmiSat44-17										i			c
LmiSat45-274	p,i		p	p									c
LmiSat46-353													–
LmiSat47-41					p								c
LmiSat48-220													nc
LmiSat49-47				p									c
LmiSat50-16										i			c
LmiSat51-241		i											c
LmiSat52-143													–
LmiSat53-47										i			c
LmiSat54-272			p	i			i	p	d				m
LmiSat55-90													nc
LmiSat56-19							p			i			m
LmiSat57-230													–
LmiSat58-86													nc
LmiSat59-16													nc
LmiSat60-255													nc
LmiSat61-63													nc
LmiSat62-23								p					c
Total p	3	4	5	8	3	3	2	6	4	5	3	6	52
Total i	2	3	0	3	0	0	1	1	2	10	0	4	26
Total d	0	5	0	0	0	0	1	0	1	4	0	0	11
Total loci	5	12	5	11	3	3	4	7	7	19	3	10	89
satDNAs	4	11	5	10	3	3	4	7	7	16	3	10	83

lowed by consensus monomer length. For instance, the most abundant satDNA in the Spanish genome of *L. migratoria* would read LmiSat01-193. The catalog number would allow differentiating two satDNAs coinciding in length. If, in the future, additional satDNA families were found in other populations of the same species, they should be numbered subsequently to the last one described in previous work. Optionally, if a function is assigned to a satDNA, a reference to it could be added at the end of the name. For instance, since we know that LmiSat07-5 in *L. migratoria* is the telomeric DNA repeat (Frydrychová *et al.*, 2004), we could name it LmiSat07-5-tel.

SatDNA abundance was very similar in both genomes, but divergence showed a tendency to be higher in the Northern genome (Supplementary Results 2.S1). To test the reliability of the satDNAs found, we designed primers in opposite orientation for all of them and PCR amplified 59 of them on genomic DNA from Spanish specimens, belonging to the Southern lineage. The three exceptions (LmiSat46-353, LmiSat52-143 and LmiSat57-230) were rare satDNAs which had been found by RepeatExplorer only in the NL genome, whereas RepeatMasker detected them also in the SL individual from the Padul population. However, PCR failed to amplify them in four different SL individuals from the Cádiz population, suggesting population differences for the presence of these rare satDNAs. In addition, LmiSat07-5-tel corresponded with the telomeric DNA repeat (TTAGG) (Frydrychová *et al.*, 2004), and was excluded from subsequent analyses because of its known function. Therefore, we will work here with the remaining 58 satDNAs.

The 58 satDNAs showed high variation for monomer length (8-400 bp) and A+T content (29.4-67.6%) (Table 2.1). Monomer length showed a bimodal distribution, with a 37 bp gap (between 90 and 127 bp) dividing the 58 satDNAs into two groups, one including 26 short satDNAs (8-90 bp) and the other comprising 32 long satDNAs (127-400 bp). The 37 bp gap in monomer length appears to be an oddity of the *L. migratoria* genome, as we have not found such a long gap in *L. elegans* or other grasshopper species (Ruiz-Ruano et al., unpublished). Long satDNAs showed higher A+T content and lower divergence than short ones, and the latter show a very high tendency to arise from G+C-rich genomic regions (Supplementary Results 2.S2).

Short and long satDNAs show similar patterns of chromosomal location

We performed single FISH analysis for all 58 satDNAs and also double FISH combining a satDNA probe with rDNA or histone gene probes, when needed for accurate identification of the satDNA-carrying chromosomes.

Both short (Fig. 2.2) and long (Fig. 2.3) satDNAs showed three main patterns at cytological level: clustered at specific chromosome regions (c), non-clustered (nc) and a mixed pattern (m) (Table 2.1). Depending on satDNA abundance, the non-clustered pattern can go from complete absence of FISH signal to general chromosome brightness above background. The mixed pattern includes both large and very small clusters. The frequencies of c, nc and m patterns did not differ significantly between the two length classes (Supplementary Results 2.S3).

As a whole, the 47 clustered satDNAs (excluding telomeric DNA) showed 89 chromosomal clusters per haploid genome, i.e. 1.89 per satDNA and 7.42 per chromosome pair, on average. Most of them were proximal (52), whereas only 26 were interstitial and 11 distal, with a similar distribution between short and long satDNAs (Supplementary Results 2.S3).

With the exception of the telomeric repeat, short satDNAs were clustered on only 1 or 2 chromosome pairs, whereas clustered long satDNAs were found on 1, 2, 3, 5, 6 or all 12 chromosome pairs, the latter condition being found only for LmiSat01-193, which was located proximal to the centromeric region in all chromosomes, with clusters in the eight shortest chromosome pairs (M4-S11) being larger than those in the four longer chromosomes (L1, L2, X and M3) (Fig. 2.3g).

The most frequent pattern, in both short and long satDNAs, was the presence of a large cluster in a single chromosome pair, as was the case for 15 short and 18 long satDNAs (see Table 2.S3, and some examples in Fig. 2.2b,d,e and Fig. 2.3e,f), with LmiSat21-38 and LmiSat28-263 showing two clusters in the same chromosome. One satDNA (LmiSat23-223) showed the same location as 45S rDNA in this species. The ideogram in Fig. 2.4 summarizes the location of all satDNAs.

Excluding the two only satDNAs which were present in all chromosomes, i.e. LmiSat01-193 and LmiSat07-5-tel, the remaining 46 families (19 short and 27 long) of clustered satDNAs (including those showing the mixed pattern) were irregularly distributed among the different chromosomes, with four chromosomes lacking short satDNAs (L1, X, M5 and M8) but all chromosomes carrying one or more different long satDNAs, in addition to LmiSat01-193 (Table 2.1). Remarkably, the S9 chromosome was the only chromosome carrying more short (10) than long (6) satDNAs.

Only 14 satDNAs (4 short and 10 long) showed clusters in more than one chromosome pair, and this allows testing the equilocality of satDNA distribution. As Table 2.S4 shows, short and long satDNAs displayed similar equilocality indices (0.63 and 0.65, respectively) thus reinforcing their similarities in chromosome distribution pattern.

The high number of different satDNAs described here is very useful for chromosome identification in *L. migratoria*, as 15 short and 18 long sat-

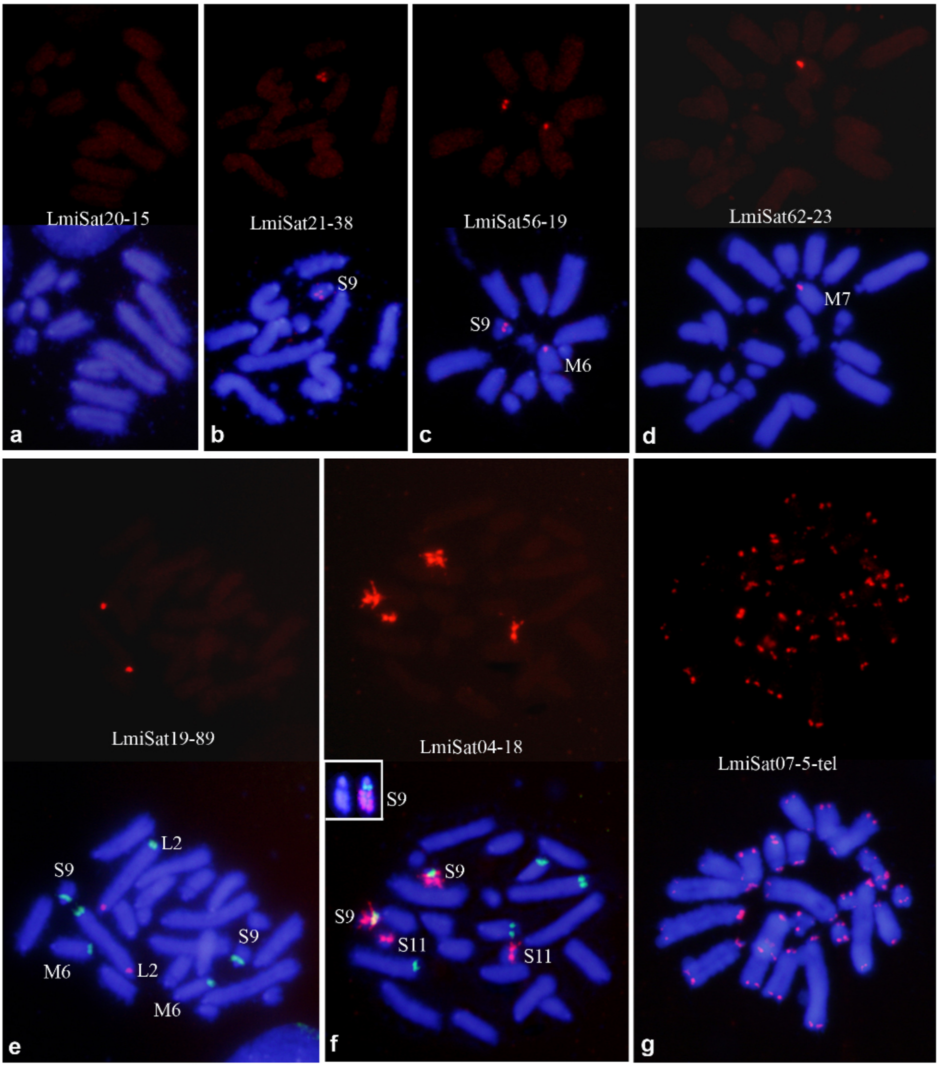


Figure 2.2: Physical mapping of seven of the short satDNAs found in *Locusta migratoria*, showing the three patterns of chromosome distribution observed: non-clustered (a), clustered (d-g) and mixed (b and c). a-c show haploid mitotic metaphase cells from haplodiploid embryos, whereas d-g show diploid cells from normal embryos. Each cell is shown in red color for satDNA FISH (upper panel) and merged with DAPI (lower panel). In e and f, double FISH was performed to distinguish whether the sat-carrying chromosome was L2 instead of L1 (e) and whether S9 carried LmiSat04-18 in addition to rDNA (shown in green color) (f). Inset in F shows the S9 chromosome stained with DAPI, on the left, and submitted to double FISH for LmiSat04-18 (red) and rDNA (green), on the right, which was selected from another cell showing lower chromosome condensation. Note the presence of three about similar sized satDNA blocks located in interstitial and distal regions of the S9 chromosome. In g, note that LmiSat07-5-tel shows the typical pattern of telomeric repeats.

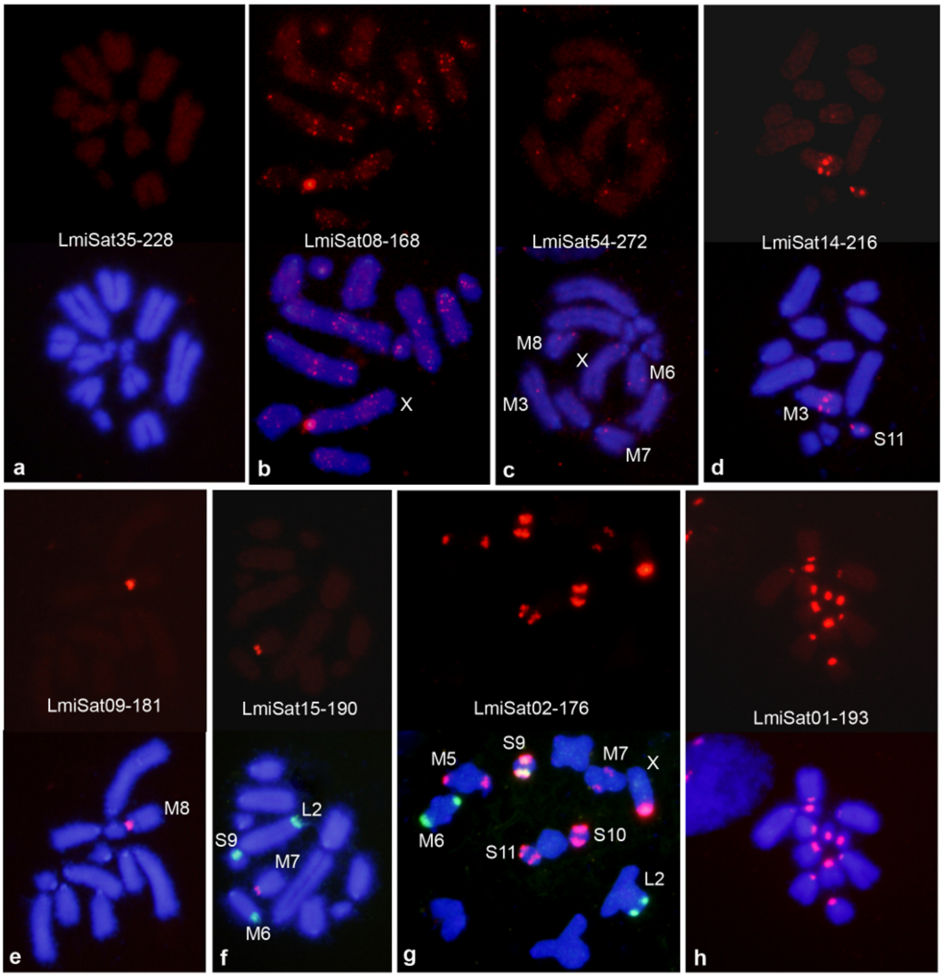


Figure 2.3: Physical mapping of eight of the long satDNAs found in *Locusta migratoria*, showing the three patterns of chromosome distribution observed: non-clustered (a), clustered (d-h) and mixed (b and c). All cells showed here (except that in g) are mitotic metaphase haploid cells from haplo-diploid embryos obtained in our laboratory. The cell in g is at meiotic metaphase I and was obtained from an adult male. Each cell is shown in red color for satDNA FISH (upper panel) and merged with DAPI (lower panel). In f and g, double FISH was performed to distinguish whether the sat-carrying chromosome was M6 (harboring rDNA shown in green) or any other medium-sized chromosome. Note in h the presence of LmiSat01-193 in the pericentromeric regions of all chromosomes.

DNAs were chromosome-specific markers allowing the direct identification of 9 out of the 12 chromosome pairs, the only exceptions being L1, M6 and S10 (Fig. 2.4 and Table 2.S3). However, these three chromosome pairs can indirectly be identified through their satDNA content pattern, since L1 is the only L-chromosome carrying LmiSat03-195, LmiSat37-238 and LmiSat45-274, M6 is the only M-chromosome carrying LmiSat56-19 and 45S rDNA, and S10 can be identified because it lacks the chromosome-specific satDNAs present in the two similar-sized autosomes (S9 and S11) (e.g. LmiSat04-18, LmiSat05-400 and LmiSat06-185).

A search for the 62 satDNA sequences in the draft genome of *L. migratoria*(Wang *et al.*, 2014) revealed that most of them were present in a surprisingly high number of contigs, with very high differences among satDNA families (Table 2.1), this variation being positively correlated with abundance (Spearman rank correlation: rs= 0.46, N= 58, P= 0.00026). Remarkably, clustered satDNAs showed no significant difference in the number of contigs compared with non-clustered ones (Mann-Whitney test: U= 198, P= 0.23), suggesting that both types of satDNAs are similarly scattered throughout the genome. Therefore, in addition to the large arrays present in the clusters revealed by FISH, clustered satDNAs show many short arrays at many loci across the genome.

Homologies between satDNAs define five superfamilies

A comparison of DNA sequence between the 58 monomer families revealed the existence of similarity between some of them, which allowed defining five superfamilies (Table 2.1). As shown in Fig. 2.S3, superfamily 1 (SF1) includes two long satDNA families: LmiSat01-193, located in pericentromeric regions of all chromosomes, and LmiSat13-259 located only in the M4 chromosome, thus being a case of local derivation of LmiSat13-259 from LmiSat01-193. SF2 includes LmiSat12-273 and LmiSat16-278 both distally located on the L2 chromosome, thus showing satDNA divergence without movement to non-homologous chromosomes. SF3 is composed of five different long satDNA families showing all patterns of chromosome location, thus illustrating how long satDNAs may evolve through sequence diversification and changes in chromosome location patterns (Table 2.1 and Fig. 2.5). SF4 includes three long satDNA families (LmiSat26-240, LmiSat37-238 and LmiSat51-241) interstitially located on different chromosomes (S11, L1 and L2, respectively), thus providing evidence for clustering on different non-homologous chromosomes. Finally, SF5 included three short satDNAs (LmiSat31-8, LmiSat50-16 and LmiSat59-16) showing different location patterns, but the reliability of this superfamily is doubtful (see Fig. 2.S4 and Table 2.S5).

Homology with other repeated sequences

We found seven satDNA families with homology to sequences from Orthoptera contained in Repbase (Table 2.S6 and Fig. 2.S5). LmiSat06-

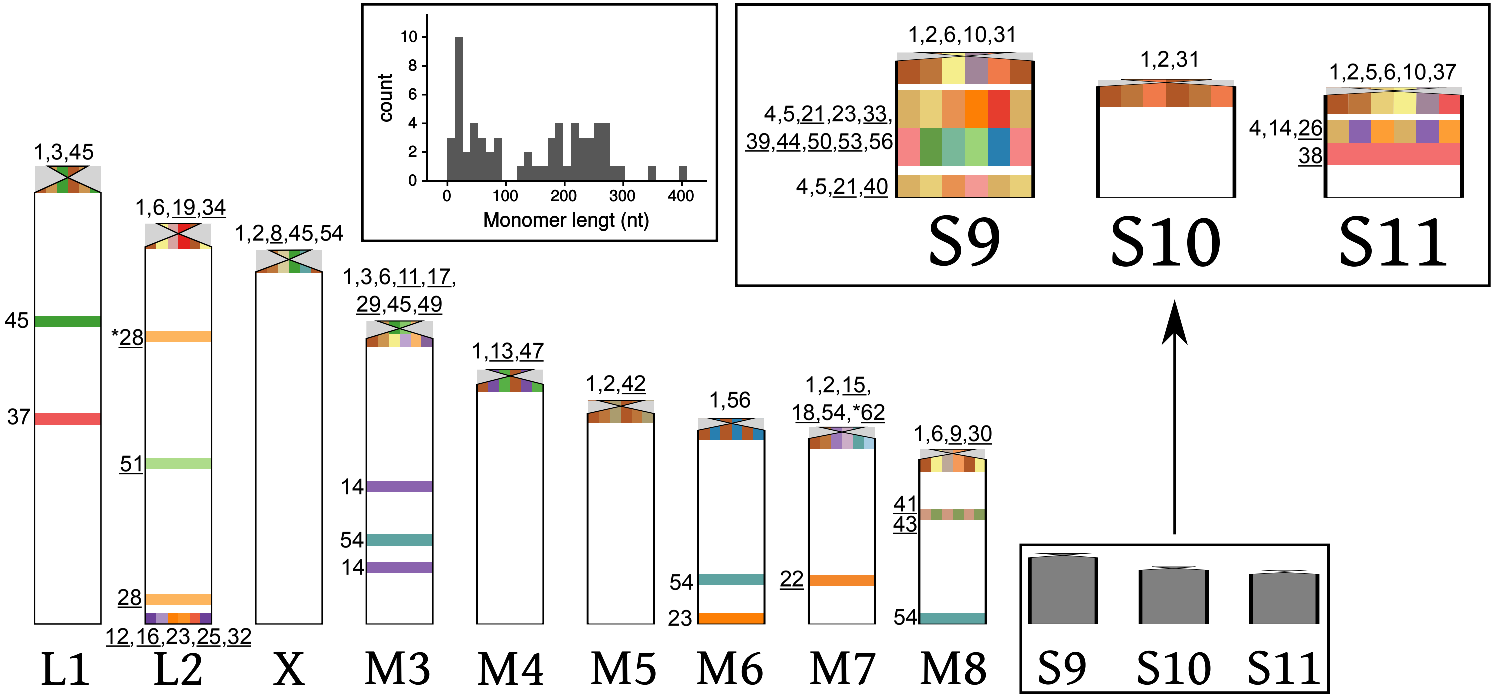


Figure 2.4: Ideogram showing chromosome location of satDNA clusters mapped by FISH. SatDNAs are noted here only by the catalog number, which is underlined in the case of chromosome-specific families. Polymorphic loci are indicated by an asterisk. Pericentromeric light-grey areas represent constitutive heterochromatin. The inset on the left shows a histogram of monomer lengths for the 62 satDNA families. Note the gap between 90 and 127 bp.

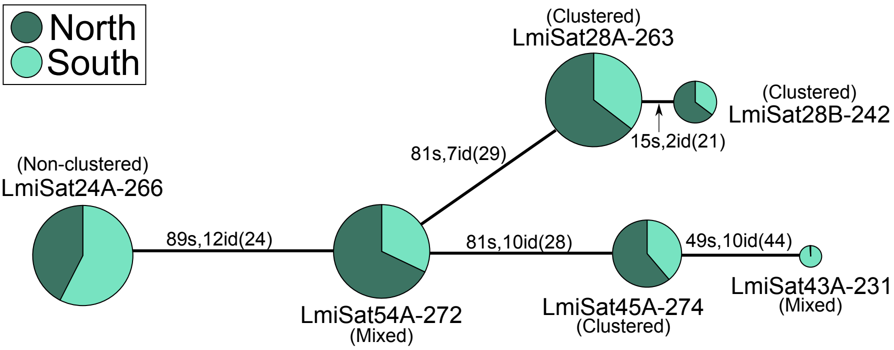


Figure 2.5: Minimum spanning tree for SF3 superfamily. The link size between haplotypes is proportional to the number of substitutions (s) and indels (id). In brackets, it is indicated the sum of nucleotides involved in the indels. SF3 was composed of six sequences corresponding to five different satDNAs, with lengths ranging between 231 and 274 bp. Note that they constitute a heterogeneous collection of satDNAs showing common descent and displaying all patterns of chromosome location and thus illustrating how long satDNAs may evolve by changing sequence and chromosome location patterns (see Table 2.1).

185 showed homology with a satDNA previously described in the grasshopper *Caledia captiva* (Arnold *et al.*, 1986), whereas the six remaining matches in Repbase were with transposable elements (TEs). LmiSat02-176 showed homology with the 5′-end of a Helitron lineage. Two long satDNAs (LmiSat15-190 and LmiSat34-299) showed homology with the CDS of TEs type Gypsy and Polinton, respectively. Likewise, LmiSat29-68 andLmiSat55-90 aligned with a region outside the CDS of two different hAT transposons, and LmiSat19-89 with a DNA transposon described in *L. migratoria*. In addition, LmiSat07-5-tel is the telomeric DNA repeat conserved in the majority of insects (Frydrychová *et al.*, 2004). Finally, LmiSat11-37 showed high variation for the number of repeats for a GA microsatellite, for which reason this satDNA showed the highest divergence (56%) and number of variants (7). No other satDNA carrying microsatellites was found. Taken together, these results suggest the possibility that some satDNAs in *L. migratoria* originated from TEs, as in other organisms (Plohl *et al.*, 2008; Meštrović *et al.*, 2015).

Discussion

The 62 satDNA families of *L. migratoria* reported here constitute the highest number of satDNA families ever found in a non-model species. The closest case was the 37 satDNAs reported in the plant *Luzula elegans* within a normal run of RepeatExplorer yielding 291 major repeat clusters with genome proportions of at least 0.01% (Heckmann *et al.*, 2013). Remarkably, the application of our filtering approach to the Illumina reads deposited by Heckmann *et al.* (2013) in SRA uncovered 85 satDNA families (grouped into 5 superfamilies), with genome proportions of 0.00035% or higher (Table 2.S7). This indicates that our approach improves significantly the bioinformatic analysis for satDNA characterization with RepeatExplorer, by being able to find satDNAs showing 28-fold lower abundance. By performing several successive filtering steps and searches with RepeatExplorer, in each step subtracting those repetitive elements found in previous steps, the chance of finding other poorly represented satDNAs is substantially increased. In *L. migratoria*, the use of genomic reads from two distant populations has also been very useful, allowing detection of satDNAs with abundance as low as 0.00002%. Anyway, it is still conceivable the existence of other less abundant satDNA families which have gone unnoticed with our methodology. Likewise, other individuals from the same or a different population could harbour other satDNA variants or families.

The high-throughput analysis of the satellitome in *L. migratoria* has unveiled several interesting properties of this kind of tandem repeats:

1) The "library" hypothesis (Fry & Salser, 1977) predicts that related species share an ancestral set of different conserved satellite DNA families which may be differentially amplified in each species due to stochastic mechanisms of concerted evolution (Meštrović *et al.*, 2006). The Northern and Southern lineages of *L. migratoria* have shown very similar satellitome catalogs, with only slight differences indicating differential amplification between individuals and/or populations. The intraspecific library shown by the *L. migratoria* satellitome is not composed of completely independent satDNAs, as some of them show similarities enough to constitute five superfamilies. Remarkable conservation was displayed by LmiSat06-185, which showed 72.2% similarity with a satDNA described in *Caledia captiva* (Acridinae subfamily) (Arnold *et al.*, 1986), a species sharing the most recent common ancestor with *L. migratoria* (Oedipodinae subfamily) about 47 million years ago25. SatDNA conservatism has been reported in several organisms, such as beetles genus *Palorus* (Mestrović *et al.*, 1998), the human alpha-satellite DNA (which is highly conserved in chicken and zebrafish (Li & Kirby, 2003)) and satDNAs in some plants (Garrido-Ramos, 2015), the most extreme case being the persistence of a satDNA for 540 million years in bivalve mollusks (Plohl *et al.*, 2010). The satellitome opens new avenues to test the library hypothesis at several phylogenetic levels, and library catalogs will be known in unsuspected detail thanks to the NGS techniques.

2) Short and long satDNAs showed the same three patterns of chro-

mosome location (non-clustered, clustered or mixed), and similar equilocal distribution across non-homologous chromosomes. In consistency with previous observations on minisatellites (Charlesworth *et al.*, 1994), the short satDNAs observed in *L. migratoria* tend to show high G+C content and sequence divergence, the latter being especially apparent when they are interspersed into euchromatin.

3) The observed equilocality for short and long clustered satDNAs indicates that heterochromatin equilocality (John *et al.*, 1985) (i.e. the tendency to occupy similar location on non-homologous chromosomes) is actually based on satDNA equilocality, and this pattern may be facilitated by telomere reunion at first meiotic prophase bouquet (Mravinac & Plohl, 2010) which, in the case of acrocentric chromosomes, also implies the reunion of centromeres. Remarkably, short and long satDNAs showed very similar tendency to equilocality.

4) Satellite DNA is frequently located into heterochromatin, and this feature is used to define this kind of DNA. In *L. migratoria*, constitutive heterochromatin is restricted to small pericentromeric regions (Camacho *et al.*, 1991), which thus include the 52 pericentromeric clusters found for 26 satDNAs. However, the 26 interstitial (for 21 satDNAs) and 11 distal (for 10 satDNAs) clusters are outside constitutive heterochromatin in this species. Therefore, we conclude that satellite DNA is also contained into euchromatic regions, in consistency with recent findings in *Drosophila* (Kuhn *et al.*, 2012) and *Tribolium castaneum* (Feliciello *et al.*, 2015).

5) The high-throughput analysis of the satellitome has been highly informative on satellite DNA evolution. Our present results suggest that all previously defined types of satellite DNA (Richard *et al.*, 2008) (microsatellites, minisatellites and satellites) show similarities at genomic and cytological levels. We have found here satDNAs with monomer length reaching the domains of typical microsatellites, such as the 5 bp telomeric repeat in *L. migratoria* or several satDNAs in *L. elegans* showing monomer lengths of only 4 or 6 bp (Table 2.S7). Likewise, about half of the satDNAs found in *L. migratoria* showed monomer lengths like those defining minisatellites (<100 bp). Remarkably, satDNAs of any length can be clustered or non-clustered at cytological level. Examples of clustered microsatellites can be found in the literature (Ruiz-Ruano *et al.*, 2015), and our Figs. 1 and 2 show that satDNAs between 5 and 400 bp show the same cytological patterns irrespectively of monomer length.

6) The combination of monomer length and number of repeats per locus define array size per locus (ASPL), which actually constitutes the interface between the genomic and cytological levels. Those satDNAs showing ASPL below FISH detection threshold (i.e. about 1.5 kb Schwarzacher *et al.* 2000), will be non-clustered at cytological level, even though they

can be relatively abundant in the genome. Of course, reaching the minimum ASPL to be cytologically observed as a clustered genomic element is more difficult for short satDNAs (especially microsatellites), as many more repeats per locus are necessary (this explains the paucity of clustered microsatellites in the literature). Even long satDNAs can fail to be clustered if ASPL is below 1.5 kbp, but they would become into clustered ones if ASPL would grow above the former threshold at a single genomic locus. For instance, the non-clustered LmiSat24-266 shares the SF3 superfamily with four clustered satellites, two showing the mixed pattern (LmiSat43-231 and LmiSat54-272) and two being clustered (LmiSat28-263 and LmiSat45-274) (see Table 2.1). A minimum spanning tree of this superfamily suggested a changing dynamics of clustering pattern during evolution (Fig. 2.5).

Taken together, the former considerations lead us to suggest a model for satellite DNA evolution (Fig. 2.6). The first proposals about *de novo* formation of tandem repeats included the joint action of mutation and unequal crossing-over (Smith, 1976), but other mechanisms, such as slippage replication and/or rolling circle amplification, have also been proposed, with most probable implication of the former in the case of short repeats and the latter in the case of long repeats (Charlesworth *et al.*, 1994). Evidences are also accumulating about rolling-circle replication implication in the amplification of satDNAs (Smith, 1976; Navrátilová *et al.*, 2008), as this mechanism could actually disseminate intragenomically a *de novo* duplicated segment through replication and reinsertion at other genomic locations.

After intragenomic dissemination, many small arrays of a given satDNA will be scattered across multiple genomic locations. Our analysis of the *L. migratoria* draft genome has shown that the immense majority of the 62 satDNAs were contained in many different contigs, suggesting that either most satDNA arrays contain a variety of interspersed sequences or, most likely, they show many different genomic locations. This is also valid for the 33 chromosome-specific clustered satDNAs. The local amplification of short arrays pre-existing at different loci would explain the patterns observed in SF3 and SF4 (see Table 2.1). In *D. melanogaster*, the 1.688 satellite shows long arrays in the heterochromatin of chromosomes 2, 3 and X, but it is also found as short arrays (1-5 repeats) in the euchromatin of the same chromosomes (Kuhn *et al.*, 2012). Likewise, large blocks of the Responder satellite are found in the pericentromeric heterochromatin of chromosome 2 in *D. melanogaster*, but small blocks are also present in the euchromatin (Larracuente, 2014). Therefore, a same satellite DNA can be present as short arrays at many cytologically invisible genomic locations and also as long arrays at discrete clusters revealed by FISH.

Remarkably, satellite DNA sequences also exist in bacteria, as variable

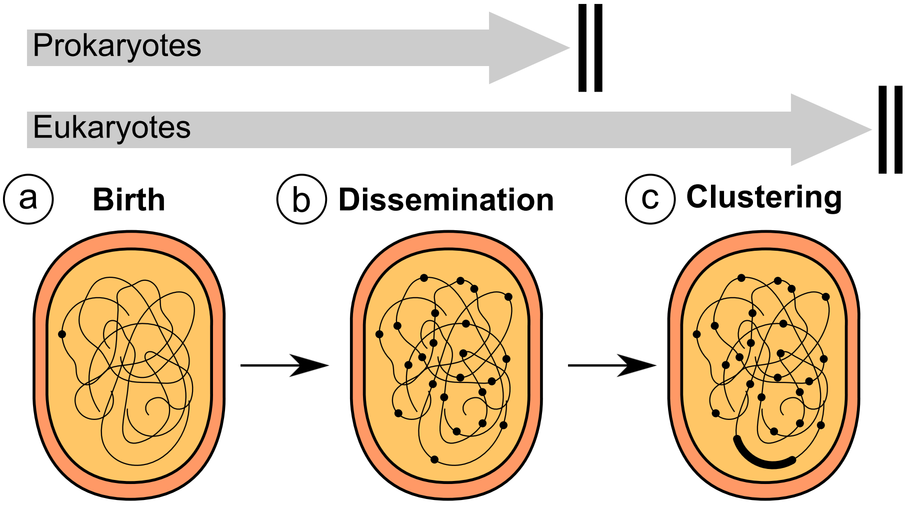


Figure 2.6: Hypothesis on satellite DNA evolution, based on the fact that all kinds of elements can be found clustered or non-clustered, at cytological level, and show many common satellitome properties. The birth of a satellite DNA implies a *de novo* duplication of a genomic sequence of two or more bp. This can occurs, for instance, by means of replication slippage in the case of short satDNAs or rolling circle replication in the case of long ones. This gives rise to a short array (<1.5 kb, i.e. the sensitivity FISH threshold) at a single genomic location (small dot in a). This short array is then disseminated throughout the genome by unknown mechanisms, although transposable elements or rolling circle replication and reinsertion elsewhere might be good candidates (b). All satDNAs remain at this stage in prokaryote species, where genomic constraints and natural selection (represented by double vertical bars) pose rigid limits to satDNA accumulation, and some of them remain this way in eukaryotes appearing as non-clustered satDNAs (b). In eukaryotes, however, any of the short arrays can undergo local amplification surpassing the 1.5 kb thus becoming into a clustered satDNA and being visible by FISH (c). The fact that all clustered satDNAs found in the *L. migratoria* genome were found in many different contigs of the assembled genome provides strong support to the hypothesis that dissemination precedes clustering. Local amplification implies rapid increase in array size and could take place, for instance, by unequal crossing over. Based on our simulation of random genomes of the *L. migratoria* size, satDNA arrays of 15 bp or less can appear by chance at many genomic locations (>15) (Table 2.S9). For this reason, all microsatellites and the shortest minisatellites can start their life-cycle at stage b. Of course, further research is necessary to unveil the precise mechanisms involved in reaching each stage, as those included here are only suggestions based on current literature.

number tandem repeats (VNTRs) have been reported in several species. For instance, *Bacillus anthracis* shows VNTRs with repeat size ranging from 2 to 36 bp and array size from 1 to 23 repeats (Keim *et al.*, 2000), *Salmonella enterica* subsp. *enterica* shows VNTRs with monomer size between 6 bp and 189 bp, with array size of 4-15 repeats (Lindstedt *et al.*, 2003), and *Streptococcus uberus* shows VNTRs from 12 to 208 bp monomer length and 2-5 repeats per array (Gilbert *et al.*, 2006). The high similarity of repeat size between bacteria and the animal and plant species analyzed here, and the resemblance of the short arrays length and dissemination pattern, suggest that satellite DNA is a common phenomenon to prokaryotes and eukaryotes. The only difference lies in maximum array size, which is much more limited in bacteria. SatDNA clustering appears to be a eukaryotic innovation presumably facilitated by their large genomes, but total amount of satDNA is likely limited by genomic constraints and natural selection, as in prokaryotes. The fact that 48 satDNAs in *L. migratoria* are clustered, and only 11 are non-clustered, might suggest that clustering is a dead end for satDNA evolution. We suggest that the reverse pathway is conceivable through the action of natural selection when SSRT amounts become a burden. Of course satellitome analysis in other species will throw much light on this subject.

Finally, the equilocal distribution of different clustered satDNA families within a same eukaryotic genome needs an explanation. Certainly, the presence of short arrays acting as seeds at many genomic locations may facilitate contagious equilocal satDNA amplification through unequal crossing over during meiotic bouquet, since this kind of recombination requires the presence of at least short arrays of the same satDNA in different non-homologous chromosomes, and previous dissemination provides them. This is an interesting prospect for future research.

Acknowledgements

We thank Beatriz Navarro-Dominguez and José L. Oliver for their help with the filtering step and genome simulations, respectively, Manuel Garrido-Ramos for the critical reading of the manuscript, and David Martinez for English grammar corrections, and two anonymous reviewers for their constructive criticism. This research was funded by the Spanish Junta de Andalucia (Plan Andaluz de Investigacion: CVI-6649), and was partially performed by FEDER funds.

References

Anjos A, Ruiz-Ruano FJ, Camacho JPM, *et al.* (2015) U1 snDNA clusters in grasshoppers: chromosomal dynamics and genomic organization. *Heredity*, **114**, 207–219.

Arnold ML, Appels R, Shaw DD (1986) The heterochromatin of grasshoppers from the *Caledia captiva* species complex. I. Sequence evolution and conservation in a highly repeated DNA family. *Molecular biology and evolution*, **3**, 29–43.

Arnold ML, Shaw DD (1985) The heterochromatin of grasshoppers from the *Caledia captiva* species complex. *Chromosoma*, **93**, 183–190.

Bachmann L, Venanzetti F, Sbordoni V (1994) Characterization of a species-specific satellite DNA family of *Dolichopoda schiavazzii* (Orthoptera, Rhaphidophoridae) cave crickets. *Journal of molecular evolution*, **39**, 274–281.

Bachmann L, Venanzetti F, Sbordoni V (1996) Tandemly repeated satellite DNA of *Dolichopoda schiavazzii*: A test for models on the evolution of highly repetitive DNA. *Journal of molecular evolution*, **43**, 135–144.

Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 1.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, p. btu170.

Britten RJ, Graham DE, Neufeld BR (1974) Analysis of repeating DNA sequences by reassociation. *Methods in enzymology*, **29**, 363–418.

Cabrero J, Bakkali M, Bugrov A, *et al.* (2003) Multiregional origin of B chromosomes in the grasshopper *Eyprepocnemis plorans*. *Chromosoma*, **112**, 207–211.

Cabrero J, López-León MD, Teruel M, Camacho JPM (2009) Chromosome mapping of H3 and H4 histone gene clusters in 35 species of acridid grasshoppers. *Chromosome research*, **17**, 397–404.

Cabrero J, Viseras E, Camacho JPM (1984) The B-chromosomes of *Locusta migratoria* I. Detection of negative correlation between mean chiasma frequency and the rate of accumulation of the B's; a reanalysis of the available data about the transmission of these B-chromosomes. *Genetica*, **64**, 155–164.

Camacho JM, Ruiz-Ruano FJ, Martín-Blázquez R, *et al.* (2015) A step to the gigantic genome of the desert locust: chromosome sizes and repeated DNAs. *Chromosoma*, **124**, 263–275.

Camacho JPM, Cabrero J, Viseras E, López-León MD, Navas-Castillo J, Alché JD (1991) G banding in two species of grasshopper and its relationship to C, N, and fluorescence banding techniques. *Genome*, **34**, 638–643.

Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive DNA in eukaryotes.

Dennis ES, Peacock WJ, White MJD, Appels R, Contreras N (1981) Cytogenetics of the parthenogenetic grasshopper *Warramaba virgo* and its bisexual relatives. *Chromosoma*, **82**, 453–469.

Drummond AJ, Ashton B, Cheung M, *et al.* (2009) Geneious v. 4.8. 5 Biomatters Ltd. *Aukland, New Zealand*.

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources*, **10**, 564–567.

Feliciello I, Akrap I, Brajković J, Zlatar I, Ugarković \j (2015) Satellite DNA as a driver of population divergence in the red flour beetle *Tribolium castaneum*. *Genome biology and evolution*, **7**, 228–239.

Fox DP, Santos JL (1985) N-bands and nucleolus expression in *Schistocerca gregaria* and *Locusta migratoria*. *Heredity*, **54**, 333–341.

Fry K, Salser W (1977) Nucleotide sequences of HS-α satellite DNA from kangaroo rat Dipodomys ordii and characterization of similar sequences in other rodents. *Cell*, **12**, 1069–1084.

Frydrychová R, Grossmann P, Trubac P, Vítková M, Marec Fe (2004) Phylogenetic distribution of TTAGG telomeric repeats in insects. *Genome*, **47**, 163–178.

Garrido-Ramos MA (2015) Satellite DNA in Plants: More than Just Rubbish. *Cytogenetic and genome research*, **146**, 153–170.

Gilbert FB, Fromageau A, Lamoureux J, Poutrel B (2006) Evaluation of tandem repeats for MLVA typing of *Streptococcus uberis* isolated from bovine mastitis. *BMC veterinary research*, **2**, 1.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, **59**, 307–321.

Hahn C, Bachmann L, Chevreux B (2013) Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Research*, **41**, e129–e129.

Heckmann S, Macas J, Kumke K, *et al.* (2013) The holocentric species *Luzula elegans* shows interplay between centromere and large-scale genome organization. *The Plant Journal*, **73**, 555–565.

Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70.

Jiang M, Anderson J, Gillespie J, Mayne M (2008) uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC bioinformatics*, **9**, 1.

John B, Appels R, Contreras N (1986) Population cytogenetics of *Atractomorpha similis*. *Chromosoma*, **94**, 45–58.

John B, King M, Schweizer D, Mendelak M (1985) Equilocality of heterochromatin distribution and heterochromatin heterogeneity in acridid grasshoppers. *Chromosoma*, **91**, 185–200.

Keim P, Price LB, Klevytska AM, *et al.* (2000) Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. *Journal of Bacteriology*, **182**, 2928–2936.

Kim YB, Oh JH, McIver LJ, *et al.* (2014) Divergence of *Drosophila melanogaster* repeatomes in response to a sharp microclimate contrast in Evolution Canyon, Israel. *Proceedings of the National Academy of Sciences*, **111**, 10630–10635.

Kit S (1961) Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. *Journal of molecular biology*, **3**, 711IN1–716IN2.

Kuhn GC, Küttler H, Moreira-Filho O, Heslop-Harrison JS (2012) The 1.688 repetitive DNA of *Drosophila*: concerted evolution at different genomic scales and association with genes. *Molecular biology and evolution*, **29**, 7–11.

Larracuente AM (2014) The organization and evolution of the Responder satellite in species of the *Drosophila melanogaster* group: dynamic evolution of a target of meiotic drive. *BMC evolutionary biology*, **14**, 233.

Lee HR, Neumann P, Macas J, Jiang J (2006) Transcription and evolutionary dynamics of the centromeric satellite repeat CentO in rice. *Molecular biology and evolution*, **23**, 2505–2520.

Li YX, Kirby ML (2003) Coordinated and conserved expression of alphoid repeat and alphoid repeat-tagged coding sequences. *Developmental dynamics*, **228**, 72–81.

Lindstedt BA, Heir E, Gjernes E, Kapperud G (2003) DNA fingerprinting of *Salmonella enterica* subsp. *enterica* serovar Typhimurium with emphasis on phage type DT104 based on variable number of tandem repeat loci. *Journal of Clinical Microbiology*, **41**, 1469–1479.

López-León MD, Vázquez P, Hewitt GM, Camacho JPM (1995) Cloning and sequence analysis of an extremely homogeneous tandemly repeated DNA in the grasshopper *Eyprepocnemis plorans*. *Heredity*, **75**, 370–375.

Ma C, Yang P, Jiang F, *et al.* (2012) Mitochondrial genomes reveal the global phylogeography and dispersal routes of the migratory locust. *Molecular Ecology*, **21**, 4344–4358.

Marshall OJ (2004) PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. *Bioinformatics*, **20**, 2471–2472.

Meštrović N, Castagnone-Sereno P, Plohl M (2006) Interplay of selective pressure and stochastic events directs evolution of the MEL172 satellite DNA library in root-knot nematodes. *Molecular biology and evolution*, **23**, 2316–2325.

Meštrović N, Mravinac B, Pavlek M, Vojvoda-Zeljko T, Šatović E, Plohl M (2015) Structural and functional liaisons between transposable elements and satellite DNAs. *Chromosome Research*, **23**, 583–596.

Mestrović N, Plohl M, Mravinac B, Ugarković D (1998) Evolution of satellite DNAs from the genus *Palorus*–experimental evidence for the" library" hypothesis. *Molecular biology and evolution*, **15**, 1062–1068.

Mravinac B, Plohl M (2010) Parallelism in evolution of highly repetitive DNAs in sibling species. *Molecular biology and evolution*, **27**, 1857–1867.

Navrátilová A, Koblížková A, Macas J (2008) Survey of extrachromosomal circular DNA derived from plant satellite repeats. *BMC plant biology*, **8**, 1.

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792–793.

Pita M, ZABAL-AGUIRRE M, Arroyo F, Gosálvez J, LÓPEZ-FERNÁNDEZ C, De La Torre J (2008) *Arcyptera fusca* and *Arcyptera tornosi* repetitive DNA families: whole-comparative genomic hybridization (W-CGH) as a novel approach to the study of satellite DNA libraries. *Journal of evolutionary biology*, **21**, 352–361.

Plohl M, Luchetti A, Meštrović N, Mantovani B (2008) Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero) chromatin. *Gene*, **409**, 72–82.

Plohl M, Meštrović N, Mravinac B (2014) Centromere identity from the DNA point of view. *Chromosoma*, **123**, 313–325.

Plohl M, Petrović V, Luchetti A, *et al.* (2010) Long-term conservation vs high sequence divergence: the case of an extraordinarily old satellite DNA in bivalve mollusks. *Heredity*, **104**, 543–551.

Rafferty JA, Fletcher HL (1992) Sequence analysis of a family of highly repeated DNA units in *Stauroderus scalaris* (Orthoptera). *International journal of genome research*.

Richard GF, Kerrest A, Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Reviews*, **72**, 686–727.

Rodríguez Iñigo E, Fernández-Calvín B, Capel J, La Vega D, García C (1996) Equilocality and heterogeneity of constitutive heterochromatin: *in situ* localization of two families of highly repetitive DNA in *Dociostaurus genei* (Orthoptera). *Heredity*, **76**.

Ruiz-Ruano FJ, Cuadrado Á, Montiel EE, Camacho JPM, López-León MD (2015) Next generation sequencing and FISH reveal uneven and nonrandom microsatellite distribution in two grasshopper genomes. *Chromosoma*, **124**, 221–234.

Schmieder R, Edwards R (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS one*, **6**, e17288.

Schwarzacher T, Heslop-Harrison P, others (2000) *Practical* in situ *Hybridization.* BIOS Scientific Publishers Ltd.

Singer MF (1982) Highly repeated sequences in mammalian genomes. *International review of cytology*, **76**, 67–112.

Smit AFA, Hubley R, Green P (2013) RepeatMasker Open-4.0.

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. *Science*, **191**, 528–535.

Tautz D, Renz M (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic acids research*, **12**, 4127–4138.

Untergasser A, Cutcutache I, Koressaar T, *et al.* (2012) Primer3—new capabilities and interfaces. *Nucleic acids research*, **40**, e115–e115.

Usakin L, Abad J, Vagin VV, de Pablos B, Villasante A, Gvozdev VA (2007) Transcription of the 1.688 satellite DNA family is under the control of RNA interference machinery in *Drosophila melanogaster* ovaries. *Genetics*, **176**, 1343–1349.

Wang X, Fang X, Yang P, *et al.* (2014) The locust genome provides insight into swarm formation and long-distance flight. *Nature communications*, **5**.

Yoshimura A, Nakata A, Kuro-o M, Obara Y, Ando Y (2005) Molecular cytogenetic characterization and chromosomal distribution of the satellite DNA in the genome of *Oxya hyla intricata* (Orthoptera: Catantopidae). *Cytogenetic and genome research*, **112**, 160–165.

Yoshimura A, Nakata A, Mito T, Noji S (2006) The characteristics of karyotype and telomeric satellite DNA sequences in the cricket, *Gryllus bimaculatus* (Orthoptera, Gryllidae). *Cytogenetic and genome research*, **112**, 329–336.

Supplementary Information

Supplementary Results

Results S1 | Northern and Southern lineage genomes show very similar satellitome content

A comparison of satDNA abundance between the Southern and Northern genomes showed good general agreement in abundance (Spearman rank correlation: rS= 0.50, N= 58, t= 4.34, P= 0.000059; Wilcoxon matched pairs test: T= 807, P= 0.71). Likewise, satDNA divergence in the 55 satDNAs found in both genomes showed significant positive correlation (rs= 0.76, t= 8.45, P< 0.000001), but it showed a significant tendency to be higher in the Northern genome (T= 472.5, P= 0.013). It is necessary to bear in mind that our analyses were made in a single individual per lineage, thus being intragenomic but not population estimates. Therefore, we cannot rule out that a given satDNA being absent in one of the two genomes analyzed might actually be present in other individuals from the same lineage. For instance, LmiSat62-23 was not bioinformatically found in the Southern genome, but it was observed by FISH in a different individual belonging to this same lineage (see Table 1).

Results S2 | Monomer length variation

The 58 satDNAs showed high variation for monomer length (8-400 nt) and A+T content (29.4-67.6%) (Table 1), two parameters showing significant positive correlation (rS= 0.35, t= 2.8, P= 0.006) thus suggesting that longer satDNAs tend to show higher A+T content and shorter ones tend to be G+C rich. Monomer length distribution showed a bimodal distribution, with a 37 nt gap (between 90 and 127 nt) dividing the 58 satDNAs into two groups, one including 26 short satDNAs (8-90 nt) and the other comprising 32 long satDNAs (127-400 nt). A comparison of A+T content between both groups showed significantly higher A+T content in the long satDNAs (Mann-Whitney test: U= 214.5, P= 0.0016), confirming the tendency suggested by the Spearman rank correlation above. Remarkably, the A+T average for the 58 satDNAs (53.44%) was significantly lower than that in the *L. migratoria* genome (59.32%) (Wilcoxon one-sample test: T= 290, P= 0.000012). The same bias was apparent when compared with Wilmore and Brown's[1] estimate of 58.37% A+T for the whole genome of this species (T= 392, P= 0.0003). This suggests that satDNAs in this species tend to arise from G+C-rich regions, which is more evident for short (mean= 49.17%; T= 22, N= 26, P= 0.000097) than long (mean= 56.91%; T= 160, N= 32, P= 0.052) ones. In addition, short satDNAs showed higher divergence than long ones (Southern genome: U= 132, P= 0.000016; Northern genome: U= 193.5, P= 0.0014). Taken together, these results indicate that short satDNAs show higher divergence and G+C content than long ones.

Results S3 | Chromosome location

The frequencies of clustered, non-clustered and mixed patterns (17, 7 and 2, respectively, for short satDNAs, and 25, 4 and 3, respectively, for long ones) did not differ significantly between the two length classes (RxC with 50,000 replicates: P= 0.343, SE= 0.006).

The total number of proximal, interstitial and distal loci did not differ significantly between short and long satDNAs (RxC: P= 0.170, SE= 0.006).

Supplementary Figures

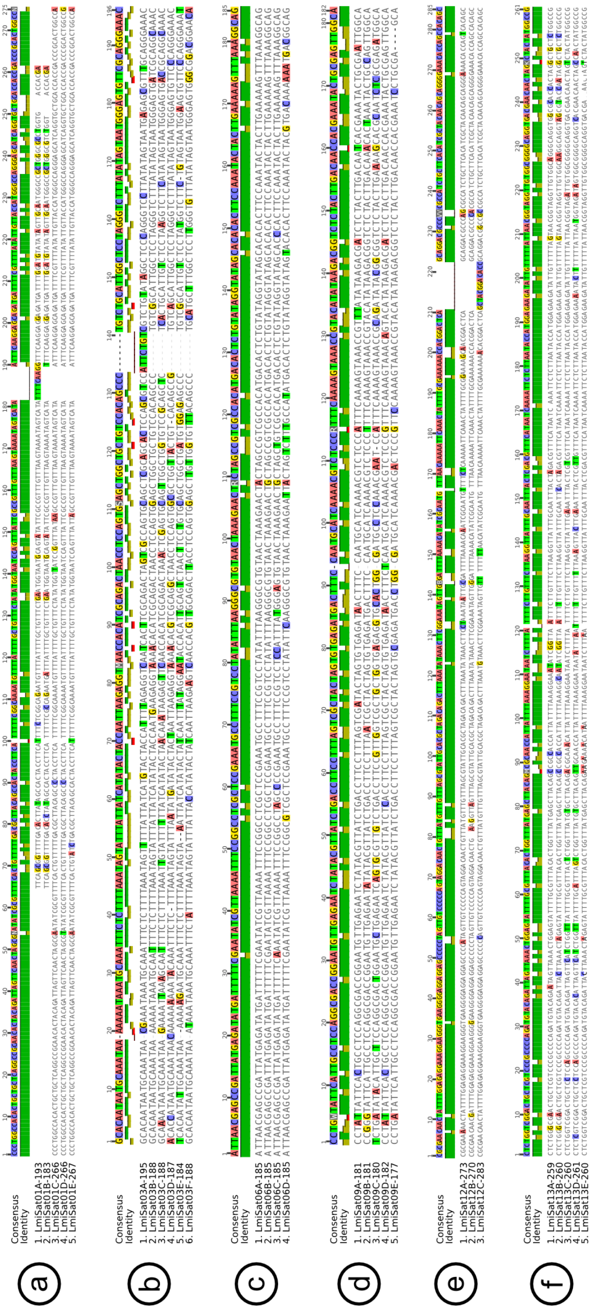


Figure 2.S1: Alignments between the different variants found for several long (**a-f**) satDNA families. [Continues on next page]

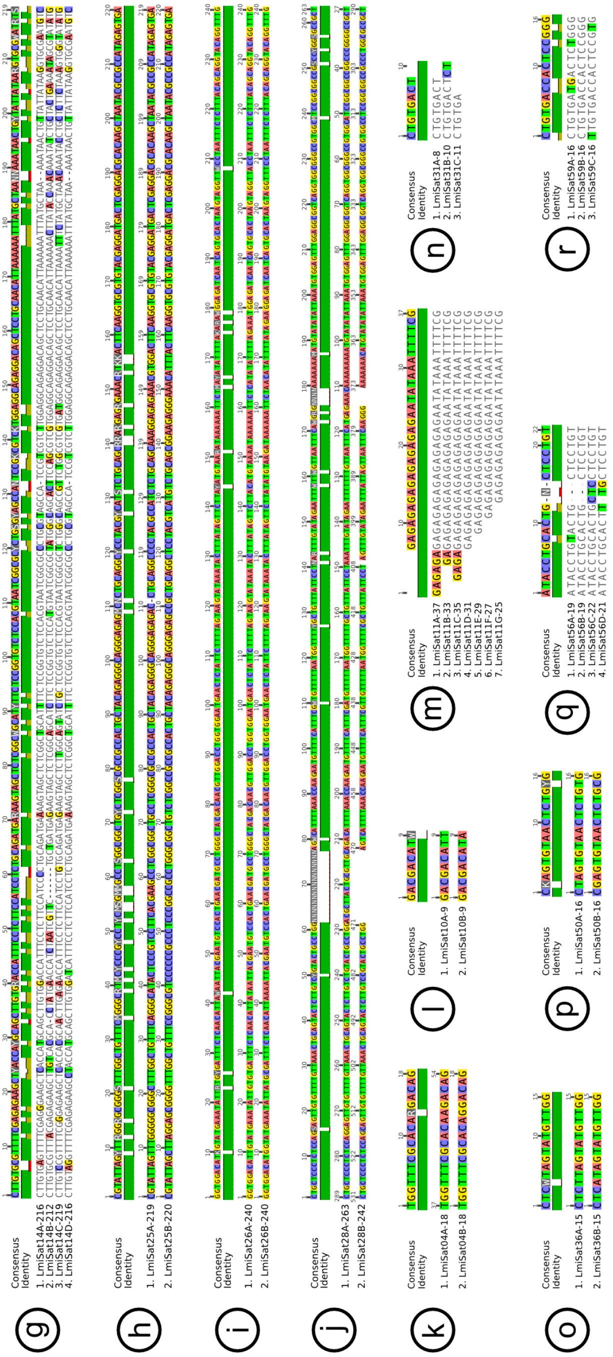


Figure 2.S1: [Continued from next page] Alignments between the different variants found for several long (**g-j**) and short (**k-r**) satDNA families.

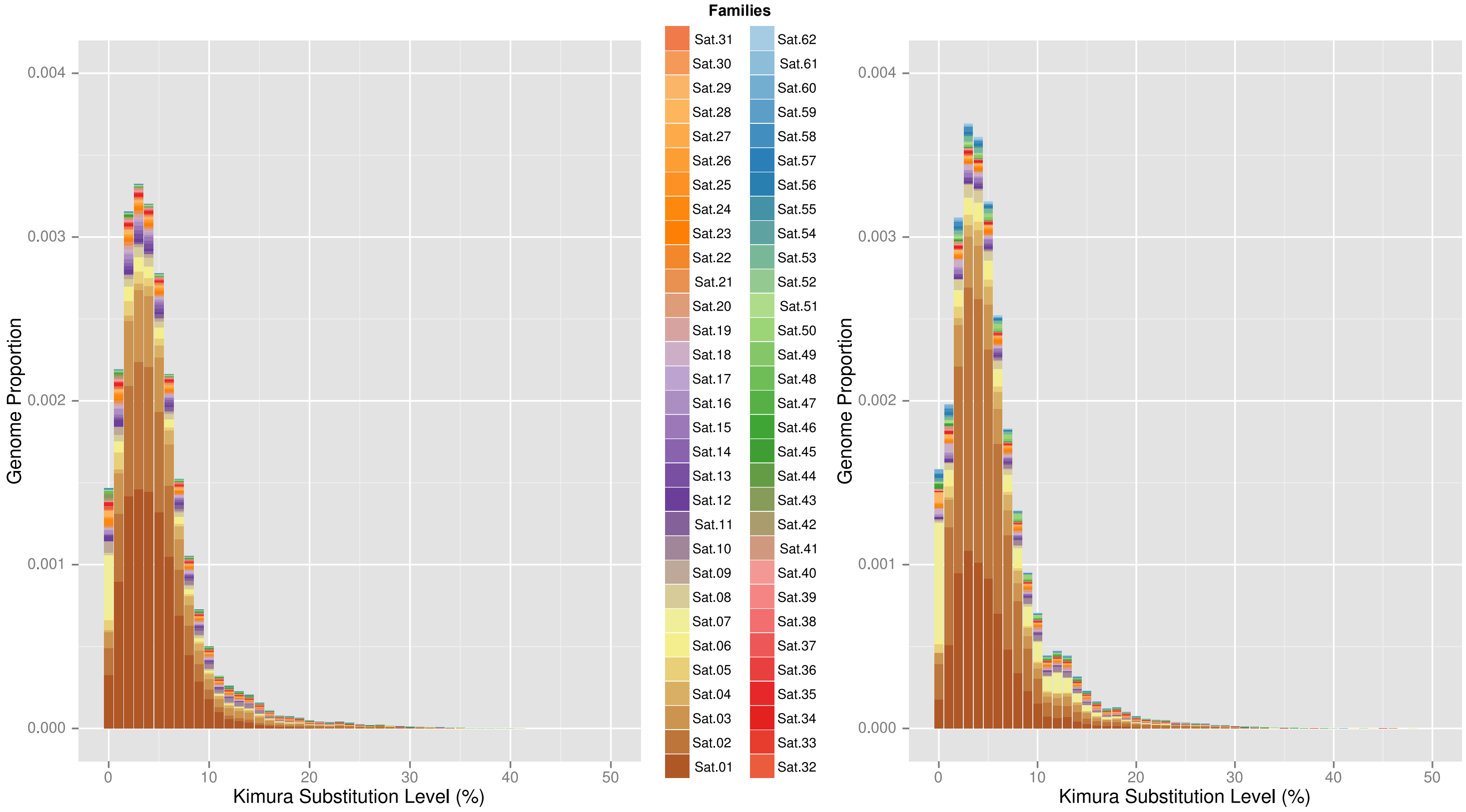


Figure 2.S2: Repeat landscapes for the 62 satDNA families in the individuals analyzed from the Southern (left) and Northern (right) lineages. Note that both lineages show a similar collection of satDNAs with only slight variations in abundance.

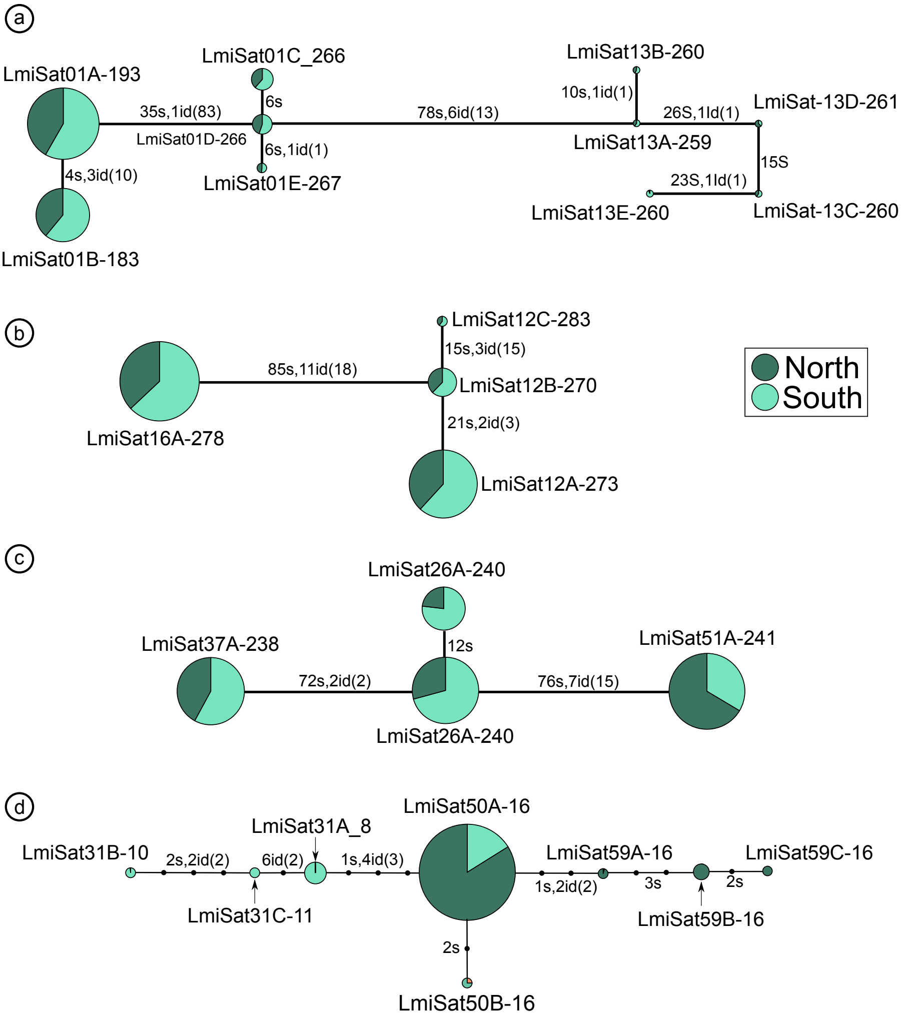


Figure 2.S3: Minimum spanning trees for superfamilies 1, 2, 4 and 5 (**a**-**d**). In **a**-**c**, link size between haplotypes is proportional to the number of substitutions (s) and indels (id) (in **d**, links are also indicated as mutational steps). In brackets is indicated the sum of nucleotides involved in the indels. [Legend continues in the next page]

Figure 2.S3 [Continuation]: **a**) Superfamily 1 (SF1) includes five sequence variants for LmiSat01-193 (three showing lengths about 1.5 fold the two remainder) and five for LmiSat13-259 (showing lengths similar to those of the longest LmiSat01-193 variants). On the basis of abundance, the ancestral monomer for this superfamily might be about 180-190 nt long (LmiSat01A-193 and LmiSat01B-183 variants), and the remaining variants in SF1, which are about 260 nt long, arose through a 83 nt insertion. Both satDNA families locate pericentromerically, but LmiSat01-193 was on all chromosomes and LmiSat13-259 was only on M4 (Table 1), suggesting that LmiSat13-259 arose from LmiSat01-193 in the M4 chromosome. **b**) Superfamily 2 (SF2) includes three sequence variants for LmiSat12-273 (270-283 nt) and one for LmiSat16-278 (278 nt). The exclusive presence of these two satDNAs at a coincident distal location in the L2 chromosome suggests that SF2 arose in this chromosome and has not moved to other non-homologous chromosomes. This case illustrates how the differential accumulation between variants give rise to new satDNA families when similarity decreases beyond the 80% criterion. **c**) Superfamily 4 (SF4) includes two variants of LmiSat26-240 (240 nt) and a single variant of LmiSat37-238 and LmiSat51-241. All three satDNA families were interstitially located but on different chromosomes: S11, L1 and L2, respectively, with LmiSat37-238 showing a second cluster proximally located on S11. SF4 thus reflects how satDNAs move between non-homologous chromosomes. **d**) Superfamily 5 (SF5) included three short satDNA (LmiSat31-8, LmiSat50-16 and LmiSat59-16) showing different location patterns: LmiSat31-8 is pericentromeric on S9 and S10, LmiSat50-16 is interstitial on S9, and LmiSat59-16 is non-clustered. Sequence alignment suggests that LmiSat49-16 and LmiSat58-16 families could have arisen from LmiSat31-8 through duplication (Supplementary Fig. S4). However, a minimum spanning tree for these three families suggests that LmiSat50A-16 (which is abundant in both lineages) is the ancestral variant, and that LmiSat31-8 emerged in the Southern genome and LmiSat59-16 in the Northern one. In addition, the fact that simulated genomes of *L. migratoria* would contain, by chance, more than 200,000 copies of DNA motives identical to the three LmiSat31-8 variants (Supplementary Table S5), together with its exclusive presence in the Southern genome, suggests the possibility that this extremely short satDNA arose independently from the two other SF5 members in the Southern lineage. Likewise, LmiSat50-16 and LmiSat59-16 could represent a case of derivation of LmiSat59-16 from LmiSat50-16 in the Northern lineage, but the fact that simulated genomes included 6 and 4 copies, respectively, for both (Supplementary Table S5), and their different patterns of chromosomal location (clustered and non-clustered, respectively) throw some doubts on this possibility. Therefore, the reliability of SF5 needs additional analysis.

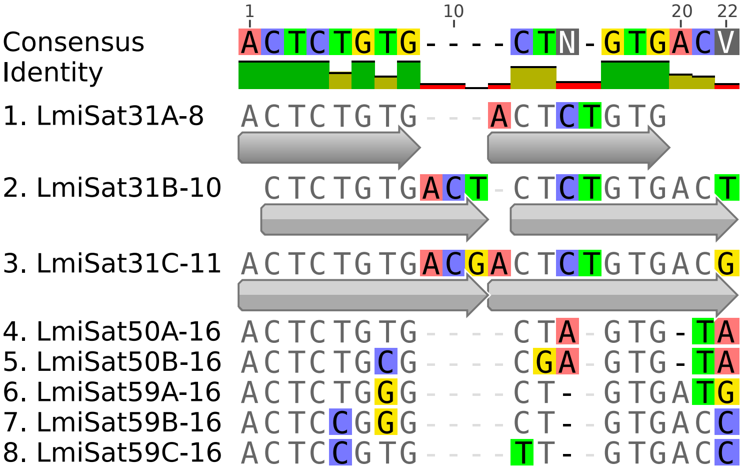


Figure 2.S4: Alignment of LmiSat31-8 dimers and LmiSat50-16 and LmiSat59-16 dimers, all belonging to superfamily 5, showing how the two latter families could have derived from a dimer for the former satDNA.

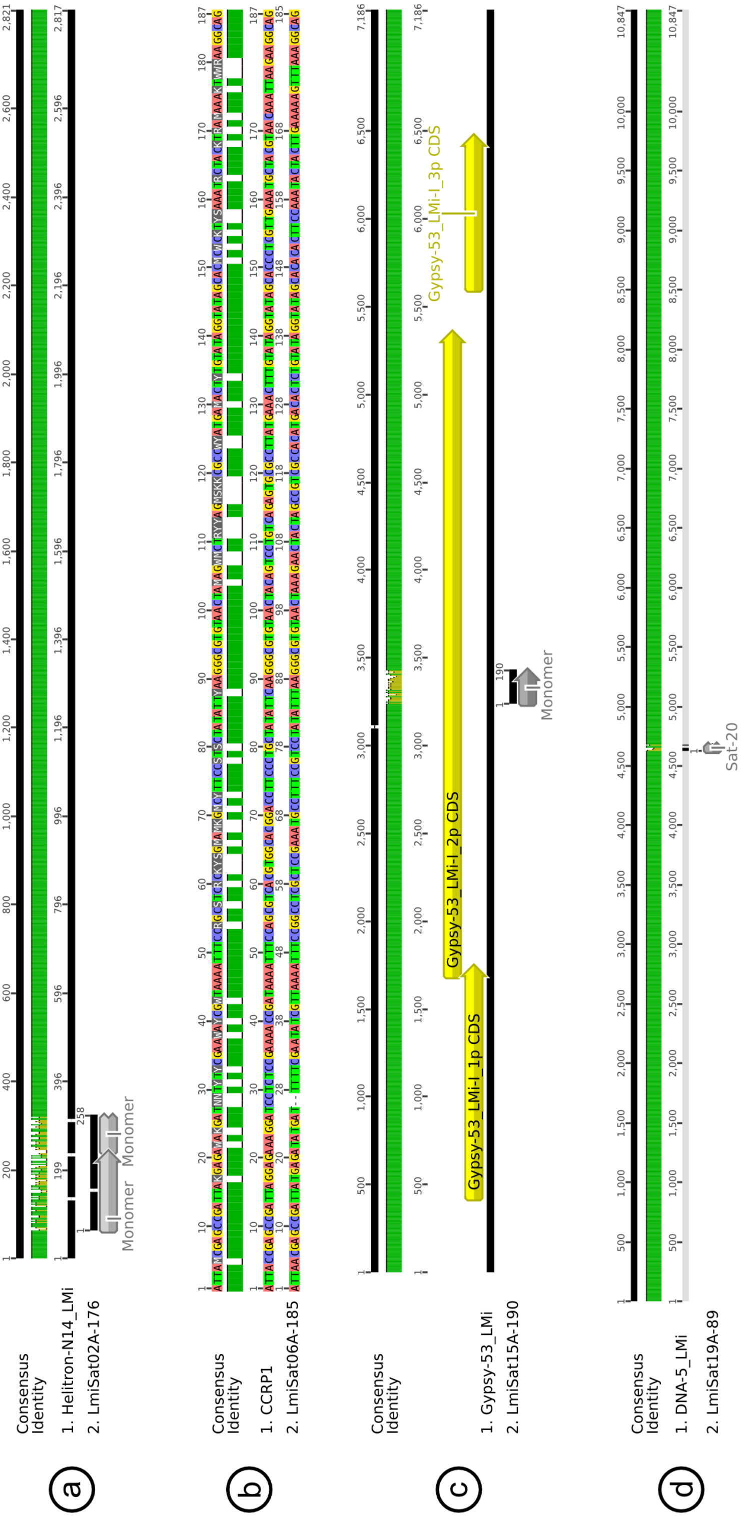


Figure 2.S5: Alignments of all satDNAs matching with Repbase entries. [Continues on next page]

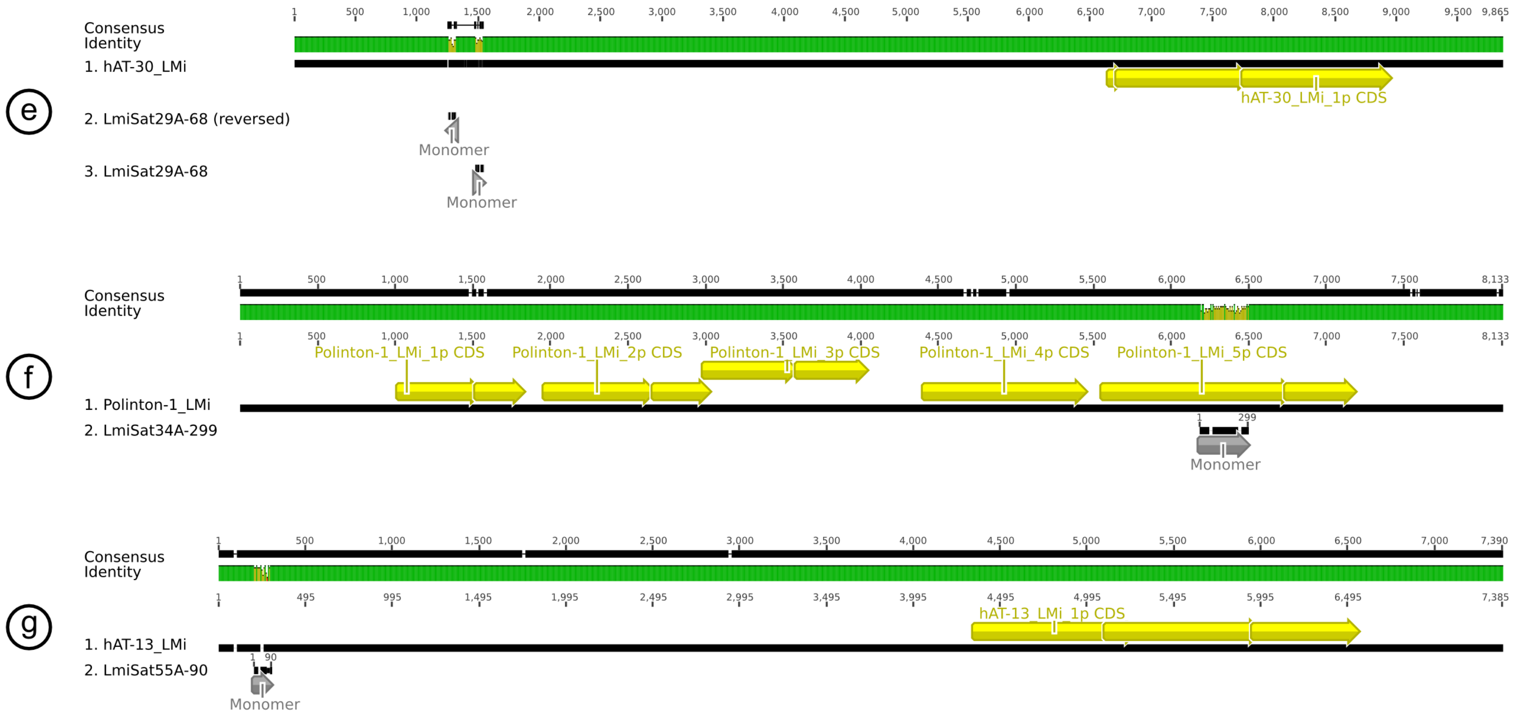


Figure 2.S5: [Continued from previous page] Alignments of all satDNAs matching with Repbase entries.

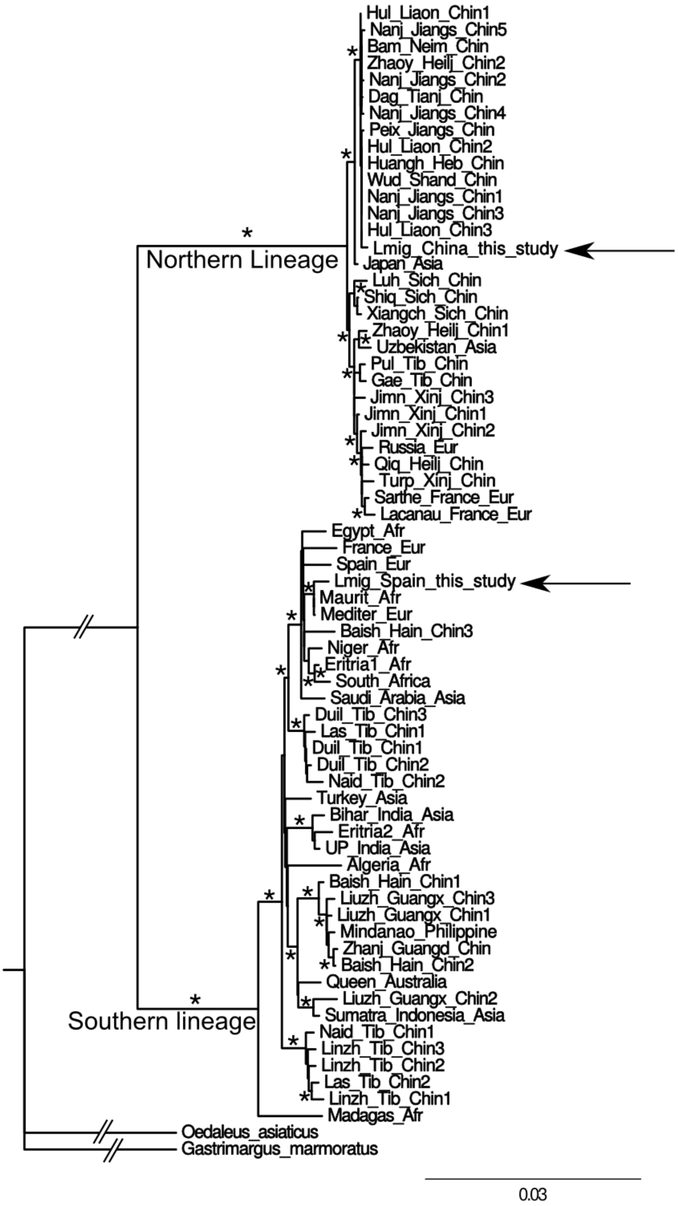


Figure 2.S6: Maximum likelihood phylogeny for full mitogenomes reported by Ma *et al.* (2012) in *L. migratoria* and those assembled by us from the same Illumina reads used to search for satDNAs in this study, from a Spanish and a Chinese individuals (arrows). Asterisks indicate branch supports higher than 90%. Note that the Spanish individual clustered with Southern mitogenomes whereas the Chinese one corresponds to the Nothern lineage.

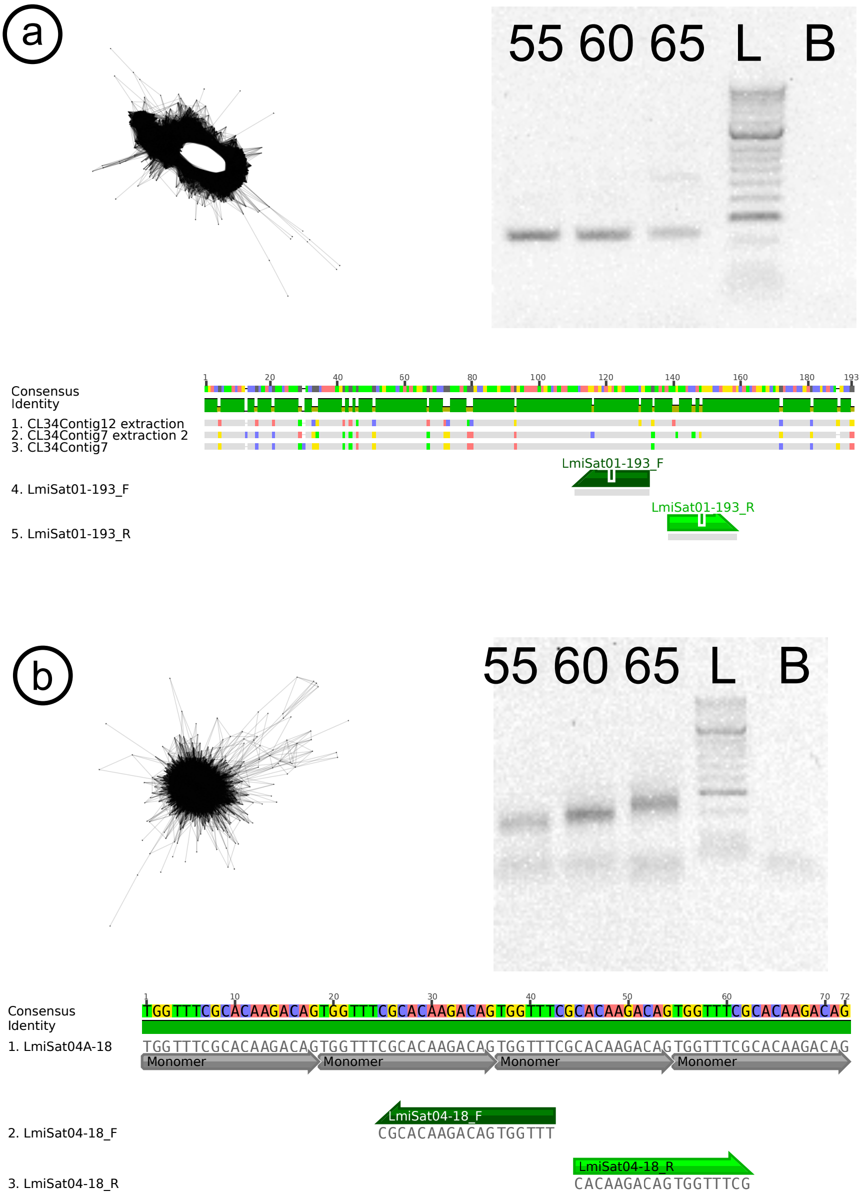


Figure 2.S7: Primer design and PCR amplification for long (**a**) and short (**b**) satDNA. Note that long satDNAs (e.g. LmiSat01-193 shown here) show ring-shaped RepeatExplorer cluster graphs because read length is lower than monomer length. We designed divergent primer pairs, with nearby 5′ ends, and tested them at 55, 60 and 65ºC annealing temperature. Dimer amplification was manifested at the highest temperature (**a**). Short satDNAs (e.g. LmiSat04-18 shown here) show spherical RepeatExplorer cluster graphs because monomer length is lower than read length. We designed divergent primers with the less stable extensive dimers. We obtained a delimited smear showing higher size with increasing annealing temperature (**b**).

Supplementary Tables

Table 2.S1: SatDNA families reported in Orthoptera before this study. NGS: Next-Generation Sequencing. W-CGH: Whole-Comparative genomic hybridization.

Species	Source	Name	nt	Method	Characteristics
Warramaba virgo	Dennis *et al.* (1981)	–	–	CoT	–
Atractomorpha similis	John *et al.* (1986)	537bp	537	Restriction (TaqI)	–
Caledia captiva	Arnold *et al.* (1986)	168bp	168	Restriction (TaqI)	Interstitial and distal
	Arnold & Shaw (1985)	144bp	144	Restriction (TaqI)	Paracentromeric
Stauroderus scalaris	Rafferty & Fletcher (1992)	168bp	168	CoT	Not determined (probably distal)
Dociostaurus genei	Rodríguez Iñigo *et al.* (1996)	DgT2	160	Restriction (TaqI)	Centromeric C-bands in each chromosome of the complement
	Rodríguez Iñigo *et al.* (1996)	DgA3	217	Restriction (AluI)	Distal C-bands present in most of the autosomal pairs
Dolichopoda spp.	Bachmann *et al.* (1994)	pDoP102	102	Restriction (PstI)	Species specific for *D. schiavazzii*, 30% of the genome
	Bachmann *et al.* (1996)	pDsPv400	~400	Restriction (PvuII)	Species specific for *D. schiavazzii*
	Bachmann *et al.* (1996)	pDoP500	~500	Restriction (PstI)	Probably present in all *Dolichopoda* species
Eyprepocnemis plorans	López-León *et al.* (1995)	180bp	180	Restriction (DraI)	Paracentromeric and B chromosome
Oxya hyla intricata	Yoshimura *et al.* (2005)	169bp	169	Restriction (HaeIII)	C-bands of the short arms of most of the chromosomes and species-specific
	Yoshimura *et al.* (2005)	204bp	204	Restriction (HaeIII)	Centromeric in three chromosome pairs and specific of *O. hyla intricata*
Gryllus bimaculatus	Yoshimura *et al.* (2006)	GBH535	535	Restriction (HindIII)	Conserved in *Gryllus* species. Derived from a common ancestral sequence
	Yoshimura *et al.* (2006)	GBH542	542	Restriction (HindIII)	Species-specific
Arcyptera fusca and *Arcyptera tornosi*	Pita *et al.* (2008)	EcoRV-390CEN	390	W-CGH	Centromeric
	Pita *et al.* (2008)	Sau3A-419CEN	419	W-CGH	Centromeric
	Pita *et al.* (2008)	Sau3A-197TEL	197	W-CGH	Heterochromatic distal regions
Schistocerca gregaria	Camacho *et al.* (2015)	SG1	171	NGS	Pericentromeric regions of complement
	Camacho *et al.* (2015)	SG2-alpha	352	NGS and Rest. (HindIII)	Distal C-bands of the three shortest chromosomes
	Camacho *et al.* (2015)	SG3	170	NGS	Interstitially in chromosome S10

Table 2.S2: Length (bp), A+T content (%), abundance (% of the genome), number of repeats calculated as "[abundance x genome size (6.3 Gb)]/repeat length", and divergence (%) for all satDNA variants found in the gDNA libraries analyzed from Southern (SL) and Northern (NL) lineages.

			Abundance (%)		Repeats		Divergence (%)	
Variant	Length	G+C	SL	NL	SL	NL	SL	NL
LmiSat01A	193	40.41	0.39467	0.28298	128830	92373	5.21	5.52
LmiSat01B	183	39.89	0.30928	0.19809	106473	68194	3.89	4.05
LmiSat01C	266	43.98	0.12652	0.07954	29966	18840	2.89	3.11
LmiSat01D	266	44.36	0.10516	0.08471	24907	20063	6.89	7.62
LmiSat01E	267	43.45	0.04661	0.04494	10998	10605	4.61	5.23
LmiSat02A	176	46.59	0.47509	0.99959	170059	357809	5.32	5.38
LmiSat03A	195	41.03	0.21447	0.17144	69290	55390	4.25	4.68
LmiSat03B	188	39.36	0.02336	0.02088	7828	6997	11.97	12.89
LmiSat03C	188	37.23	0.02219	0.01716	7437	5750	6.53	7.60
LmiSat03D	187	37.97	0.01625	0.01140	5475	3841	10.27	10.64
LmiSat03E	184	36.41	0.00963	0.00397	3296	1361	4.89	4.78
LmiSat03F	188	37.23	0.00891	0.00564	2987	1889	5.08	4.81
LmiSat04A	18	50.00	0.05540	0.07290	193895	255147	6.66	6.80
LmiSat04B	18	55.56	0.00654	0.00869	22877	30408	11.82	10.87
LmiSat05A	400	48.75	0.05431	0.04827	8553	7603	4.65	5.04
LmiSat06A	185	40.54	0.01845	0.01558	6284	5307	4.62	5.59
LmiSat06B	185	40.00	0.01759	0.02093	5991	7128	5.49	5.29
LmiSat06C	185	40.54	0.01625	0.02317	5535	7889	3.38	4.23
LmiSat06D	185	38.92	0.00180	0.01034	612	3522	10.93	7.14
LmiSat07A	5	40.00	0.04438	0.16113	559196	2030244	1.75	6.12
LmiSat08A	168	42.26	0.03737	0.04669	14015	17510	4.96	4.91
LmiSat09A	181	39.78	0.01405	0.00252	4892	876	1.07	1.53
LmiSat09B	181	41.99	0.00554	0.00179	1929	622	6.03	6.27
LmiSat09C	180	50.56	0.00419	0.00070	1468	245	14.57	21.06
LmiSat09D	182	41.76	0.00408	0.00127	1411	439	7.17	8.40
LmiSat09E	177	46.89	0.00157	0.00098	560	348	10.98	12.200
LmiSat10A	9	44.44	0.02052	0.02700	143648	189028	11.62	11.23
LmiSat10B	9	44.44	0.00217	0.00200	15172	14028	13.27	13.88
LmiSat11A	37	37.84	0.00651	0.00222	11082	3784	7.69	7.46
LmiSat11B	33	36.36	0.00342	0.00090	6538	1709	7.81	8.34
LmiSat11C	35	37.14	0.00315	0.00150	5677	2695	8.30	8.25
LmiSat11D	31	35.48	0.00301	0.00125	6114	2535	7.64	8.17
LmiSat11E	29	34.48	0.00145	0.00052	3157	1128	6.98	9.11
LmiSat11F	27	33.33	0.00090	0.00034	2104	789	7.81	8.52
LmiSat11G	25	32.00	0.00028	0.00021	705	529	7.63	10.03
LmiSat12A	273	43.59	0.01170	0.00723	2701	1669	2.19	3.64

Continued on next page

Continued from previous page

			Abundance (%)		Repeats		Divergence (%)	
Variant	Length	G+C	SL	NL	SL	NL	SL	NL
LmiSat12B	270	46.67	0.00494	0.00297	1152	694	5.77	7.23
LmiSat12C	283	46.29	0.00172	0.00112	382	250	5.52	10.50
LmiSat13A	259	42.47	0.00723	0.00488	1758	1188	4.01	6.40
LmiSat13B	260	40.77	0.00596	0.00443	1444	1074	4.44	5.20
LmiSat13C	260	37.31	0.00292	0.00196	707	475	5.81	7.36
LmiSat13D	261	34.87	0.00073	0.00005	177	12	1.78	21.24
LmiSat13E	260	37.31	0.00013	0.00022	32	52	6.04	9.97
LmiSat14A	216	48.15	0.00584	0.00468	1703	1365	5.11	9.69
LmiSat14B	212	48.11	0.00441	0.00257	1312	763	5.81	7.34
LmiSat14C	219	49.77	0.00241	0.00137	693	393	6.47	7.79
LmiSat14D	216	46.76	0.00160	0.00048	467	140	3.62	10.65
LmiSat15A	190	44.74	0.01426	0.01660	4727	5504	4.09	4.50
LmiSat16A	278	37.41	0.01390	0.00817	3149	1851	2.49	3.01
LmiSat17A	75	42.67	0.01177	0.00335	9891	2810	5.79	6.66
LmiSat18A	210	39.52	0.01121	0.02669	3362	8008	6.33	4.59
LmiSat19A	89	39.33	0.01058	0.00342	7486	2423	3.82	6.44
LmiSat20A	15	46.67	0.01032	0.02015	43324	84621	12.71	14.15
LmiSat21A	38	50.00	0.01013	0.00194	16790	3222	2.85	2.91
LmiSat22A	17	41.18	0.01000	0.00923	37056	34220	10.81	10.28
LmiSat23A	223	38.57	0.00927	0.01061	2618	2998	4.42	5.73
LmiSat24A	266	43.61	0.00895	0.00656	2120	1553	2.06	5.14
LmiSat25A	219	60.27	0.00558	0.00675	1605	1943	7.48	9.35
LmiSat25B	220	62.73	0.00276	0.00374	791	1070	2.79	6.11
LmiSat26A	240	33.48	0.00544	0.00224	1434	591	6.53	6.52
LmiSat26B	240	34.17	0.00359	0.00108	941	284	3.78	4.07
LmiSat27A	57	52.63	0.00790	0.01029	8729	11377	8.99	9.66
LmiSat28A	263	42.59	0.00532	0.00962	1275	2303	1.23	1.62
LmiSat28B	242	44.21	0.00236	0.00429	614	1117	2.94	3.57
LmiSat29A	68	41.18	0.00719	0.00193	6659	1786	9.36	14.48
LmiSat30A	138	59.42	0.00680	0.00550	3102	2511	5.74	9.03
LmiSat31A	8	50.00	0.00427	0.00001	33647	79	3.25	40.01
LmiSat31B	10	50.00	0.00161	0.00002	10122	116	5.76	22.57
LmiSat31C	11	54.55	0.00080	–	4567	–	3.55	–
LmiSat32A	261	48.28	0.00631	0.00565	1523	1363	5.98	9.18
LmiSat33A	21	52.38	0.00627	0.00394	18820	11817	7.77	8.35
LmiSat34A	299	38.13	0.00622	0.00475	1312	1001	6.81	7.39
LmiSat35A	228	44.30	0.00597	0.00529	1649	1463	2.43	4.64
LmiSat36A	15	40.00	0.00367	0.00603	15423	25322	16.84	15.39
LmiSat36B	15	40.00	0.00218	0.00331	9168	13920	16.94	14.65
LmiSat37A	238	34.03	0.00451	0.00328	1193	867	10.12	10.85
LmiSat38A	42	35.71	0.00511	0.00463	7668	6949	14.56	14.94
LmiSat39A	53	67.92	0.00503	0.00130	5984	1551	6.79	9.17
LmiSat40A	148	32.43	0.00459	0.00229	1954	975	2.35	3.05
LmiSat41A	180	38.33	0.00455	0.00579	1592	2026	3.38	2.14

Continued on next page

Continued from previous page

			Abundance (%)		Repeats		Divergence (%)	
Variant	Length	G+C	SL	NL	SL	NL	SL	NL
LmiSat42A	127	48.82	0.00447	0.00123	2218	610	2.02	4.60
LmiSat43A	231	46.32	0.00440	0.00003	1199	8	0.68	9.57
LmiSat44A	17	70.59	0.00428	0.00050	15869	1843	11.45	11.30
LmiSat45A	274	45.99	0.00420	0.00657	966	1510	8.20	7.22
LmiSat46A	353	40.23	0.00407	0.00710	727	1267	15.49	11.38
LmiSat47A	41	58.54	0.00369	0.00580	5675	8909	12.46	13.22
LmiSat48A	220	41.82	0.00366	0.00112	1048	322	3.80	7.74
LmiSat49A	47	57.45	0.00362	0.01129	4859	15133	6.24	6.70
LmiSat50A	16	43.75	0.00311	0.01631	12239	64229	8.27	8.23
LmiSat50B	16	56.25	0.00020	0.00059	780	2332	8.90	8.52
LmiSat51A	241	36.10	0.00294	0.00583	769	1524	7.32	3.97
LmiSat52A	143	48.25	0.00257	0.00758	1134	3340	22.15	14.01
LmiSat53A	47	59.57	0.00248	0.01904	3328	25520	3.16	5.20
LmiSat54A	272	43.75	0.00244	0.00512	565	1187	4.55	4.15
LmiSat55A	90	64.44	0.00164	0.00740	1147	5182	15.62	8.57
LmiSat56A	19	47.37	0.00047	0.00153	1558	5063	4.86	4.25
LmiSat56B	19	52.63	0.00029	0.00305	970	10103	5.31	4.74
LmiSat56C	22	54.55	0.00007	0.00108	202	3102	5.53	4.11
LmiSat56D	21	47.62	–	0.00102	–	3054	–	3.34
LmiSat57A	230	36.96	0.00052	0.00470	142	1286	18.21	3.40
LmiSat58A	86	58.14	0.00008	0.01273	56	9327	5.99	3.12
LmiSat59A	16	56.25	0.00004	0.00101	175	3978	18.23	15.88
LmiSat59B	16	68.75	–	0.00337	–	13254	–	14.39
LmiSat59C	16	56.25	–	0.00054	–	2136	–	13.02
LmiSat60A	255	47.06	0.00004	0.00527	10	1302	1.03	0.99
LmiSat61A	63	57.14	0.00002	0.00617	21	6171	14.99	4.60
LmiSat62A	23	56.52	–	0.00450	–	12338	–	4.57

Table 2.S3: Number of chromosome-specific short and long satDNA families. Note that L2 and S9 chromosomes showed the highest number of exclusive satDNAs.

	L1	L2	X	M3	M4	M5	M6	M7	M8	S9	S10	S11	Total
Short satDNAs	0	1	0	4	1	0	0	2	0	6	0	1	15
Long satDNAs	0	7	1	0	1	1	0	2	4	1	0	1	18
Total	0	8	1	4	2	1	0	4	4	7	0	2	33

Table 2.S4: Calculation of the equilocality index (EI) for short and long satDNAs. Only 4 short and 10 long satDNAs showed loci in more than one chromosome pair, and this allows testing the equilocality of satDNA distribution. Among the four short satDNAs, LmiSat56-19 showed a proximal cluster on the M6 chromosome and an interstitial one on S9, thus showing absence of equilocal distribution (equilocality index: EI= 0). By contrast, LmiSat10-9 and LmiSat31-8 showed one proximal cluster on two different chromosomes thus displaying full equilocal distribution (EI= 1). Finally, LmiSat04-18 showed interstitial and distal locations on S9 and interstitial on S11. Out of the two possible pairwise comparisons (i.e. S9i with S11i, and S9d with S11i) only the first one was equilocal, so that EI= 0.5 in this case. The average EI for the four short satDNAs was thus 0.63. In the case of long satDNAs, six of them showed full equilocality (LmiSat01-193, LmiSat02-176, LmiSat03-195, LmiSat06-185, LmiSat14-216 and LmiSat45-274), two showed absence of equilocality (LmiSat05-400 and LmiSat37-238) and two showed intermediate situations: LmiSat23-223 was distally located on two chromosome pairs and interstitially on another pair, so that only one out the three possible pairwise comparisons was equilocal (EI= 1/3). On the other hand, LmiSat54-272 was proximally located on two chromosome pairs, interstitially on two others and distally on another pair. Therefore, only two out of the ten possible pairwise comparisons were equilocal (EI= 0.2). On average, the ten long satDNAs showed 0.65 equilocality index, which is very similar to that calculated for short satDNAs.

		Chromosome no.												
satDNA family	Length (nt)	1	2	X	3	4	5	6	7	8	9	10	11	EI
LmiSat04-18	18										id		i	0.50
LmiSat10-9	9										p		p	1
LmiSat31-8	8										p	p		1
LmiSat56-19	19							p			i			0
Short														**0.63**
LmiSat01-193	193	p	p	p	p	p	p	p	p	p	p	p	p	1
LmiSat02-176	176			p			p		p		p	p	p	1
LmiSat03-195	195	p			p									1
LmiSat05-400	400										id		p	0
LmiSat06-185	185		p		p					p	p		p	1
LmiSat14-216	216				i,i								i	1
LmiSat23-223	223		d					d			i			0.33
LmiSat37-238	238	i											p	0
LmiSat45-274	274	p,i		p	p									1
LmiSat54-272	272			p	i			i	p	d				0.2
Long														**0.65**

Table 2.S6: Homology of *L. migratoria* satDNAs with other Orthoptera sequences in Repbase. All the matches are transposons described in *L. migratoria*, except the CCRP1 satDNA from *Caledia captiva*.

SatDNA family	RepBase
LmiSat02-176	Helitron-N14_LMi
LmiSat06-185	CCRP1
LmiSat15-190	Gypsy-53_LMi-I
LmiSat19-89	DNA-5_LMi
LmiSat29-68	hAT-30_LMi
LmiSat34-299	Polinton-1_LMi
LmiSat55-90	hAT-13_LMi

Table 2.S5: Extremely short satDNAs can arise by chance in the gigantic genome of *L. migratoria*. This table shows the number of occurrences found for satDNA variants, belonging to Superfamily 5, in 159 genomes randomly generated *in silico*, and searched for as monomers and dimers. Note that the LmiSat31A-8 monomer showed very high likelihood of arising by chance, and it even appeared three times as a dimer. For longer variants of this satDNA (B and C), however, dimers were not observed in the random genomes. The two other satDNAs (LmiSat50-16 and LmiSat59-16) were barely found as monomers but not as dimers.

SatDNA family	Sequence	Length (nt)	Occurrences	Number per genome
LmiSat31A-8	CTGTGACT	8	31518697	198568
LmiSat31B-10	CTGTGACTCT	10	1769538	11148
LmiSat31C-11	CTGTGACGACT	11	375521	2366
LmiSat50A-16	CTAGTGTAACTCTGTG	16	549	3
LmiSat50B-16	CGAGTGTAACTCTGCG	16	211	1
LmiSat59A-16	CTGTGATGACTCTGGG	16	244	2
LmiSat59B-16	CTGTGACCACTCCGGG	16	155	1
LmiSat59C-16	TTGTGACCACTCCGTG	16	406	3
LmiSat31A-8 dimer	CTGTGACTCTGTGACT	16	471	3
LmiSat31B-8 dimer	CTGTGACTCTCTGTGACTCT	20	0	0
LmiSat31C-8 dimer	CTGTGACGACTCTGTGACGACT	22	0	0
LmiSat50A-16 dimer	CTAGTGTAACTCTGTGCTAGTGTAACTCTGTG	32	0	0
LmiSat50B-16 dimer	CGAGTGTAACTCTGCGCGAGTGTAACTCTGCG	32	0	0
LmiSat59A-16 dimer	CTGTGATGACTCTGGGCTGTGATGACTCTGGG	32	0	0
LmiSat59B-16 dimer	CTGTGACCACTCCGGGCTGTGACCACTCCGGG	32	0	0
LmiSat59C-16 dimer	TTGTGACCACTCCGTGTTGTGACCACTCCGTG	32	0	0

Table 2.S7: Characterization of the *Luzula elegans* satellitome. Length (nt), A+T content (%), number of variants (V), abundance (% of the genome), divergence (%) and equivalency with satDNA analyzed by Heckmann *et al.*, 2013.

SF	satDNA family	Length	A+T	V	Abundance	Divergence	Heckmann *et al.*, 2013
	LelSat01-43	43	51.16	1	2.48845	7.18	CL21Contig28_X, CL9Contig39_X
2	LelSat02-4	4	75.00	1	1.50044	2.96	CL72Contig1
	LelSat03-150	150	62.00	1	1.18933	8.54	CL7Contig1
1	LelSat04-228	228	66.67	1	1.01380	6.36	–
	LelSat05-56	56	66.07	1	0.64527	9.43	CL11Contig68_X
1	LelSat06-359	359	64.62	2	0.59017	5.88	CL4Contig63
	LelSat07-42	42	47.62	1	0.52823	7.83	CL27Contig80_X
	LelSat08-41	41	41.46	1	0.44139	9.59	CL89Contig6
1	LelSat09-189	189	64.02	1	0.36094	6.14	CL4Contig269_X
	LelSat10-6	6	66.67	1	0.35370	8.58	CL36Contig19_X
	LelSat11-161	161	62.11	3	0.34232	4.72	CL17Contig4
	LelSat12-609	609	72.58	1	0.33025	8.13	CL18Contig96
	LelSat13-6	6	83.33	1	0.31573	3.32	CL25Contig1_X
	LelSat14-68	68	45.59	1	0.27338	5.72	–
	LelSat15-51	51	56.86	1	0.22059	8.20	CL22Contig21
4	LelSat16-179	179	49.72	2	0.20617	7.22	–
	LelSat17-137	137	61.31	1	0.20455	8.94	CL38Contig36
	LelSat18-57	57	71.93	2	0.19652	4.87	CL23Contig24_X
3	LelSat19-189	189	67.72	1	0.12367	4.35	CL43Contig13
	LelSat20-173	173	82.66	2	0.10765	7.17	CL16Contig7
	LelSat21-392	392	70.66	1	0.10619	3.71	CL28Contig19
	LelSat22-374	374	80.48	1	0.10563	7.59	–
	LelSat23-195	195	81.54	2	0.10466	4.20	CL16Contig6
5	LelSat24-344	344	69.48	1	0.10101	2.97	–
	LelSat25-726	726	68.87	1	0.08733	5.56	CL28Contig14
3	LelSat26-141	141	64.54	1	0.08230	7.47	–
5	LelSat27-203	203	79.31	1	0.07539	3.22	–
	LelSat28-89	89	69.66	1	0.07401	6.23	CL63Contig1
	LelSat29-66	66	39.39	1	0.06857	8.32	–
	LelSat30-42	42	61.90	1	0.05742	8.28	–
	LelSat31-45	45	46.67	1	0.04366	9.63	–
	LelSat32-180	180	70.00	1	0.04090	14.07	CL99Contig6
	LelSat33-82	82	79.27	1	0.03252	6.19	–
	LelSat34-33	33	81.82	1	0.03184	6.62	CL109Contig15
	LelSat35-45	45	64.44	1	0.03128	7.78	–
	LelSat36-37	37	48.65	2	0.03091	14.14	–
	LelSat37-42	42	57.14	1	0.03069	5.04	–
	LelSat38-177	177	75.71	1	0.03050	6.38	–
	LelSat39-43	43	65.12	1	0.03001	5.15	–
	LelSat40-99	99	54.55	1	0.02503	13.17	–
	LelSat41-541	541	63.22	1	0.02077	5.96	–
3	LelSat42-89	89	60.67	1	0.01730	4.02	–
	LelSat43-108	108	58.33	1	0.01696	5.47	–
	LelSat44-6	6	66.67	2	0.01619	7.92	–
	LelSat45-64	64	62.50	1	0.01439	5.58	–
3	LelSat46-37	37	62.16	1	0.01357	13.70	–
2	LelSat47-6	6	83.33	2	0.01323	4.59	–
3	LelSat48-228	228	65.35	1	0.01267	11.63	–

Continued on next page

Continued from previous page

SF	SatDNA family	Length	A+T	V	Abundance	Divergence	Heckmann *et al.*, 2013
	LelSat49-218	218	71.10	1	0.01218	3.94	–
	LelSat50-58	58	53.45	1	0.01180	8.02	–
4	LelSat51-213	213	50.23	1	0.01067	11.27	–
	LelSat52-42	42	69.05	1	0.01032	6.86	–
	LelSat53-107	107	48.60	1	0.01011	10.01	–
	LelSat54-113	113	74.34	2	0.01000	4.31	–
	LelSat55-7-tel	7	57.14	1	0.00913	7.52	–
3	LelSat56-76	76	65.79	1	0.00905	15.89	–
	LelSat57-137	137	57.66	1	0.00683	12.68	–
	LelSat58-82	82	29.27	1	0.00618	11.93	–
	LelSat59-107	107	80.37	1	0.00610	9.67	–
	LelSat60-18	18	72.22	2	0.00574	9.25	–
	LelSat61-21	21	61.90	1	0.00562	7.67	–
	LelSat62-66	66	56.06	1	0.00521	3.68	–
3	LelSat63-129	129	60.47	1	0.00514	8.23	–
	LelSat64-108	108	67.59	1	0.00502	11.29	–
	LelSat65-196	196	51.02	1	0.00458	6.93	–
	LelSat66-141	141	48.23	1	0.00435	5.93	–
	LelSat67-261	261	63.98	1	0.00412	10.17	–
	LelSat68-6	6	66.67	2	0.00401	6.26	–
	LelSat69-30	30	36.67	1	0.00281	12.73	–
	LelSat70-129	129	63.57	1	0.00259	7.13	–
	LelSat71-232	232	69.40	1	0.00254	10.70	–
	LelSat72-39	39	41.03	1	0.00254	8.45	–
	LelSat73-62	62	74.19	1	0.00246	9.48	–
	LelSat74-186	186	61.83	2	0.00244	6.91	–
	LelSat75-141	141	63.12	2	0.00240	7.65	–
	LelSat76-60	60	31.67	1	0.00240	9.28	–
	LelSat77-55	55	49.09	1	0.00238	5.87	–
5	LelSat78-77	77	81.82	1	0.00220	9.53	–
	LelSat79-56	56	66.07	1	0.00182	4.57	–
	LelSat80-309	309	54.37	1	0.00156	2.95	–
	LelSat81-23	23	43.48	1	0.00149	15.38	–
	LelSat82-30	30	33.33	1	0.00137	9.16	–
	LelSat83-166	166	60.24	1	0.00109	8.62	–
	LelSat84-82	82	71.95	1	0.00035	5.93	–
	LelSat85-115	115	50.43	1	0.00035	13.42	–
	Total	–	–	100	12.92641	–	–

Table 2.S8: Primers designed in this study to amplify each satDNA family.

SatDNA family	Forward	Reverse
LmiSat01-193	ACGAAAATCATCTGCTCCTTGA	TTGTTACCATGGGCCAGGGA
LmiSat02-176	GCCATCTTCCTGCACCTCCTCCT	CGTGTCTCCTGTAGCGTGAGTGG
LmiSat03-195	GCACTCCAGCGTCCATTCTGTCG	GCGAGCTGCACTGGCGACTA
LmiSat04-18	AAACCACTGTCTTGTGCG	CACAAGACAGTGGTTTCG
LmiSat05-400	TCCCATCGTTCCAAATTCACCC	CGCCAGGAGGCACGAAAG
LmiSat06-185	AGCCGTCGCCACATGACACT	CATTTCGGAGCGAGGCCGGA
LmiSat07-5	GGTTAGGTTAGGTTAGGTTA	TAACCTAACCTAACCTAACC
LmiSat08-168	ACCCCACTTTCAAGAAATTTAATTCT	GTGCTGCCAGTGGGTGCA
LmiSat09-181	TTCTCAACATTCCGGTCGCC	CGTTATCTGACCTTCCTTTAGTCG
LmiSat10-9	CGTCAATGTCGTCAATGTCG	GACATTGACGACATTGACGA
LmiSat11-37	CTCTCTCTCTCCGAAAATTTATATTC	AGAGAGAGAGAGAGAGAGAGA
LmiSat12-273	AGCGATGTGAAGCAGATGGC	GAAAACACCAGTCACAGCCG
LmiSat13-259	CCTTGCCACAACCTACCGTT	GCGTACCAATAGGCTGCTCT
LmiSat14-216	AGAAAATGCAGCCGAGAGCT	GGTGTCTCCACGTAATCGGC
LmiSat15-190	TGCCAATAGAAGAGCATGCAG	GCAGGGTCTGGAAATGTTCTGA
LmiSat16-278	TAGTTGCCCATTTACGGGCA	CCTCCTCCCCTTACACCCTG
LmiSat17-75	TGGTAAGAAGGGTCAAGTACAGGT	CACTACATTCTCAATAGTGAGCCT
LmiSat18-210	GAGCTGCTGGAGGCAACG	TCGTACAGCCCCTCCCTCTAT
LmiSat19-89	AGGAGAAGTAATTAAGCAATGCA	CCTACTACGTGTTGTCGAGC
LmiSat20-15	GGCAAGTATGCTTGTGGC	GCCACAAGCATACTTGCC
LmiSat21-38	GCCTCACTGCTGAGCTTTTGTATACG	CAGTGAGGCAACGCCAGGTAAC
LmiSat22-17	GGGAAAAACGCAGATATGGG	CCCATATCTGCGTTTTTCCC
LmiSat23-223	TAGTCTGCAGTGGCCAGGTG	TGCCTCTGCCCTCACTAGTC
LmiSat24-266	CTGGCACCGTCCACCCACC	CTCCAGAAGCGGCGGCTGG
LmiSat25-219	TGCGTCCTCGAGTCATCCTCG	GCACAAGCTAATACGCCGCCA
LmiSat26-240	CGTTCAGTGGACATTCGTAA	ACGATGCCTGGGCTACGAC
LmiSat27-57	TGGCGGGCCGTGGCATCC	ACCTGACCGCCTCCAAACTCCA
LmiSat28-263	CGCTTGAGTGCCGTTCTTCAGGT	CGCCCGAAACTAGCATGTATATGTGT
LmiSat29-68	GTGGCTGCGGCTAGACTGGC	CACGGCATCAGCGCAGCG
LmiSat30-138	TCACAGAGTCACAGAGTCACAGAG	CTCTGTGACTCTGTGACTCTGTGA
LmiSat31-8	CTCGCCCAACGTAGACTACAGC	GAGGAGCCGCACAGAGCGG
LmiSat32-261	CCGTACACGCTTAGCGAATCTCCG	CGGAGATTCGCTAAGCGTGTACGG
LmiSat33-21	GGTGTCTCCAGCTGAACAGATG	AGATTCATCATACTTGATTTTCAAACA
LmiSat34-299	TCCACCCTTTGTTTCATTGGAGT	GAAATAAAAGCAACAACTAAAAACAAC
LmiSat35-228	CCAACATACTATGAGCCAACATACTA	TAGTATGTTGGCTCATAGTATGTTGG
LmiSat36-15	ATTACGTCTATAAGATTACGAAA	ACGGCGCCAGAGATAATTTCG
LmiSat37-238	GCACTGTCATCCGATAATTAGGT	GCACTAATTCGAACATCTAATTTTTCT
LmiSat38-42	CGTCTATAAGATTACGAAATTATCTCT	TTACGGCGCCAGAGATAA
LmiSat39-53	TGGGAGAGGCGTGTGGAGGC	CACTGCTCGGCGACTGGCC
LmiSat40-148	ACCGTCACCACCAGTAGAGG	GGAGCCATTTCTGAAACAACCC
LmiSat41-180	ACTGAAAATAGGAAAATCCAGAGCCTC	AGTGTTTCAGGGATGTGTGTACTACA
LmiSat42-127	GCAGCATCGGTCTTCCTCTCTTTCG	GCACTCACCTCGGAAACTTCCACA
LmiSat43-231	CCAATGCAGACAACTGAAGGCAAC	TGGTATAGACGCTTTCCGGCGT
LmiSat44-17	CAGCCCTTCTGGACGGCC	CCGTCCAGAAGGGCTGGC
LmiSat45-274	ACGGAGGAGGTCATGTTTGCTGG	ACAAACGGCACTGAGCTTCCGA
LmiSat46-353	CAAATGGTACGTCACACATAAAATGGT	TTTTAACGTCATGCGCCTTCAC
LmiSat47-41	GACAGCAGTGGAATGCGCAGC	TCTCCACTCCTCCACAAACGC
LmiSat48-220	AGCACCACAGCGCTACATTT	GCTGCAAAACACAGTGGTCTG
LmiSat49-47	CCCCCTCTCCTTCTATACCACACG	GGGAAGCGGAGAAGGCAGGA
LmiSat50-16	AGCACAGAGTTACACTAGCAC	GTGCTAGTGTAACTCTGTGCT
LmiSat51-241	GCCCAGAGGAGCGTCAAGTGG	ACTGCGACGTTGGACCTGGA
LmiSat52-143	TCTGAGGCTGAACAGGCTGCC	ACACCGTCAAGCAAATGCAGCA
LmiSat53-47	CTCGCTGCTGAACAGAGCCA	AGCAACTTCACCAGCAGCGC
LmiSat54-272	TACAGGAGGCCGGCGGCAG	CAGCGCGCACCTCCCTCCTC
LmiSat55-90	GGCACACACAGTGGCGAGGG	GCCGCCGTGTTCAGCAGAGA
LmiSat56-19	CTCCTGTATACCTGCACTG	CAGTGCAGGTATACAGGAG
LmiSat57-230	TGCTACTCCACATAAAGATCGTGAG	TCTTCTTATGTTACTGTTCTGAGGCA
LmiSat58-86	TGCTGCCTTACAGCGTTGCG	AGGAGGGAAAGGGGCGTGAAC
LmiSat59-16	TCACAGCCCGGAGTGGTCACA	TGTGACCACTCCGGGCTGTGA
LmiSat60-255	GCAGCAGGATGAGCAAGGACGG	GCGGTGAAGAACTCTCCCCTGG
LmiSat61-63	GGGACGTGTGCTGTTATCAGTGGG	CCCTACCTGCAGCGTAACCAAGC
LmiSat62-23	AGGCAGCGAGGGCTCTGTTC	AGCCCTCGCTGCCTTATGAA

Table 2.S9: Frequency of repeats of different lengths observed in the simulated *L. migratoria* genomes. Note that sequences of 15 bp or less are present 16 or more times, indicating that many copies can independently arise by chance. We analyzed ~159 genomes randomly generated *in silico* and searched for a random sequence successively adding a nucleotide, preserving the genomic dinucleotide frequency.

Sequence	Length (nt)	Occurences	Number per genome
ATACAAGC	8	33632104	211882
ATACAAGCT	9	9622569	60622
ATACAAGCTT	10	3159365	19904
ATACAAGCTTA	11	813956	5128
ATACAAGCTTAA	12	266972	1682
ATACAAGCTTAAC	13	58277	367
ATACAAGCTTAACC	14	12023	76
ATACAAGCTTAACCC	15	2465	16
ATACAAGCTTAACCCG	16	451	3
ATACAAGCTTAACCCGT	17	157	1
ATACAAGCTTAACCCGTC	18	32	0
ATACAAGCTTAACCCGTCA	19	8	0
ATACAAGCTTAACCCGTCAT	20	4	0
ATACAAGCTTAACCCGTCATG	21	1	0
ATACAAGCTTAACCCGTCATGG	22	0	0
ATACAAGCTTAACCCGTCATGGT	23	0	0
ATACAAGCTTAACCCGTCATGGTA	24	0	0

Chapter 3. Satellitome analysis reveals the origin of a supernumerary (B) chromosome

Francisco J. Ruiz-Ruano, Josefa Cabrero, María Dolores López-León and Juan Pedro M. Camacho

Departamento de Genética, Universidad de Granada

Abstract. B chromosomes are supernumerary genomic elements frequently showing a parasitic nature. They most likely derived from the standard (A) chromosomes, but their dispensability freed their DNA sequences to evolve fast thus making it difficult to uncover their ancestry. Here we demonstrate that high-throughput analysis of the satellitome in the grasshopper *Eumigus monticola*, has been decisive in ascertaining the ancestry of a B chromosome. The satellitome in this species consists of 27 satDNA families, with monomer length between 5 and 325 nt, and A+T content between 42.9 and 83.3%. 20 of these satDNAs were clustered on one or more chromosomes. The A chromosome carrying the highest number of satDNA families was the megameric S8 (13 families), six of which were also present in the B chromosome, and three of these were exclusive of the S8 and B chromosomes. No other A chromosome showed this characteristic. This points to the S8 autosome as the B chromosome ancestor. The absence in the B chromosome of the H3 histone gene cluster (located interstitially on S8) and three satDNA families (located distally on S8) allowed delimiting the origin of the B chromosome to the proximal third of S8, through a breakpoint between EmoSat11-122 and the H3 cluster. The B chromosome carried two satDNAs (EmoSat26-41 and EmoSat27-102) that did not map by FISH on the A chromosomes. Interestingly, bioinformatic analysis revealed their presence in the A chromosomes, at very low abundance. Only EmoSat26-41 was tandemly repeated in the B-lacking genome, suggesting that it arose in the A chromosomes but was massively amplified in the B chromosome. However, EmoSat27-102 most likely arose in the B chromosome from an RTE element residing in both B-carrying and B-lacking individuals. Taken together, all results point to the intraspecific origin of this B chromosome, from the S8 autosome, and call attention on the fact that finding B-specific DNA sequences does not necessarily support the interspecific origin of B chromosomes.

Introduction

The satellitome is the catalog of satellite DNAs (satDNAs) contained in a genome (Chapter 2). Thorough knowledge of the satellitome thus opens the gates of the satDNA library (Fry & Salser, 1977) for intra- and interspecific analyses which might illuminate new evolutionary pathways for these repetitive elements. For instance, satellitome analysis in the Northern and Southern evolutionary lineages of the migratory locust, has revealed a total of 62 satDNA families, almost all coexisting in both lineages and showing only slight differences in abundance (Chapter 2), in consistency with the library hypothesis devised at interspecific level (Fry & Salser, 1977). Satellitome analysis in a group of species sharing their last common ancestor at different times, would provide estimates of average lifespan of satDNAs in different types of organisms. Alternatively, the intraspecific analysis of satellitome chromosomal distribution can be very informative about the spatial and temporal relationships between the chromosomes of a same genome, thus providing new approaches to analyze genomic compartmentalization based on satDNA chromosome localization, the relationships between homologous and/or non-homologous chromosomes during interphase, and also the origin of some genomic elements, such as B chromosomes, which is the focus of the present research.

Supernumerary (B) chromosomes are a dispensable part of the genome in about 15% of eukaryote organisms, but they are still mostly unknown at molecular and functional levels. It is believed that they derived from standard genomic elements (i.e. the A chromosomes) from the same (intraspecific origin) or a different (interspecific origin) species (for review, see Camacho, 2005), in the latter case including both their origin as a by-product during interspecific hybridization (Perfectti & Werren, 2001) or through interspecies introgression (Tosta *et al.*, 2014). The dispensability of B chromosomes explains why their DNA sequences evolve at high rate, which makes it difficult to ascertain their intragenomic ancestry, i.e. from which A chromosome (from the same or a different species) they derived, and this task gets more difficult for old B chromosomes. Up to now, reliable conclusions on B chromosome ancestry are actually scarce. Table 3.S1 shows 18 plant and animal species where B chromosome origin has been discerned in higher or lower detail (16 of them reported in the last nine years). Most of these B chromosomes (14) were reported as derived from the host genome (i.e. intraspecifically), in nine cases with a putative A chromosome ancestor, and only four B chromosomes arose interspecifically (in wasps, fish and bee).

Here we show that a B chromosome in the grasshopper *Eumigus monticola* arose from the proximal region of the S8 autosome, as deduced from satellitome analysis, as it shared a much higher number of satDNAs with S8 than with any other A chromosome. In addition, the presence of two specific satDNAs in the B chromosome can be explained intraspecifically since seeds for them were already present in the B-lacking genome.

Materials and methods

Two males and one female of the grasshopper *Eumigus monticola* were collected at Hoya de la Mora (HM) (37.092277N, -3.387084W, 2,505 m asl) in Sierra Nevada (Granada, Spain) in 2011, and two males in 2013. In addition, five males were collected in 2013 by the road to HM population, at 1.5 Km distance (37.103590N, -3.397716W, 2,394 m asl). All males showed the characteristic incurved apical valves of the penis shown in Cabrero *et al.* (1985).

Males were anesthetized with ethyl acetate vapors before dissection to fix testes in 3:1 ethanol:acetic acid and store them at 4ºC, for cytological analysis. Body remains were frozen in liquid nitrogen and stored at -80ºC, to extract DNA for molecular analysis.

The presence of B chromosomes was analyzed by the C-banding technique in all collected individuals. A first approach to analyzing the content of B chromosomes was performed by FISH for three repetitive gene families, namely 45S and 5S ribosomal DNA (rDNA) and the H3 histone gene, following the protocols described in Camacho *et al.* (2014).

We extracted genomic DNA (gDNA) from the hind legs of each individual, with the GenElute Mamalian Genomic DNA Miniprep kit (Sigma). Illumina HiSeq 2000 sequencing was performed on gDNA of one B-carrying and one B-lacking individuals, each yielding ~5 Gbp of 2x101bp reads. We deposited the 0B and +B genomic libraries in the SRA database wih accession numbers SRR3000673 and SRR3000773, respectively.

We first performed a typical RepeatExplorer (Novák *et al.*, 2013) clustering on 2x125,000 reads combined from the B-carrying and B-lacking genomic libraries. This allowed the detection of 14 satDNAs but, to uncover as many satDNAs as possible, we followed the protocol suggested by Ruiz-Ruano *et al.* (Chapter 2) by using the satMiner toolkit. Briefly, we performed a quality trimming with Trimomatic (Bolger *et al.*, 2014) and then a clustering with RepeatExplorer of a selection of 2x250,000 reads. We then selected clusters with typical satDNA structure, i.e., spherical or ring-shaped and searched for those assembled contigs showing tandem repeat structure with Geneious v4.8 (Drummond *et al.*, 2009). We then filtered out reads showing homology with the previously found clusters, with DeconSeq (Schmieder & Edwards, 2011), and performed RepeatExplorer to a sample of the remaining reads. We applied this pipeline four times duplicating the number of reads until no more satDNA clusters appeared. We searched for homology between the monomers found, and grouped them into a same sequence variant if identity was higher than 95%, within a same family if was higher than 80%, and within a same superfamily if identity was between 50% and 80%. Abundance and divergence for each variant was determined with RepeatMasker (Smit *et al.*, 2013), with the Cross_match search engine, and we assigned a catalog number to satDNA families, in order of decreasing abundance, following Ruiz-Ruano *et al.* (Chapter 2). Each satDNA family was named following the Ruiz-Ruano *et al.* (Chapter 2) criterion. The assembled sequences were deposited in GenBank with accession numbers KU315340-KU315381.

We searched for homology with other repetitive sequences in RepBase (Bao *et al.*, 2015). We built a minimum spanning tree for DNA sequences in each superfamily with Arlequin v3.5 (Excoffier & Lischer, 2010), considering each indel position as a single change and representing the relative abundance in 0B and +B individuals.

To detect satDNA families possibly residing in the B chromosome, we calculated log2 of the quotient between B-carrying and B-lacking abundances, and interpreted it as a measure of the fold change (FC) in abundance due to the presence of B chromosomes.

For satDNA families which appeared to be B-specific, we performed additional analyses to investigate their possible presence in the B-lacking genome. We first selected pairs of reads, in each library separately, showing homology with these two satDNAs, by using BLAT (Kent, 2002). This step is implemented in a custom script (https://github.com/fjruizruano/ngs-protocols/blob/master/mapping_blat_gs.py). We then selected 2x2500 of the reads selected from the B-lacking genome to run a RepeatExplorer clustering using a custom database for annotating the sequences of all assembled satDNAs. To assemble other repeated elements showing homology with the B-specific satDNAs, but lacking a tandem structure, we used the Roche's 454 Newbler assembler with the reads selected from the 0B library. We analyzed the contigs showing some homology with the satDNA, selected the corresponding reads, and assembled them until the contig could not grow longer. To analyze the genomic structure of these contigs, we finally assembled the selected reads with RepeatExplorer and mapped then with SSAHA2 software (Ning *et al.*, 2001) to score coverage along the contig.

All satDNAs found bioinformatically were amplified by PCR using divergent primers (Table 3.S2), as in Ruiz-Ruano *et al.* (Chapter 2), and they were sequenced by the Sanger method to verify their reliability. PCR product for each satDNA was labeled by nick translation using 2.5 units of DNA Polymerase I/DNase I (Invitrogen) to generate DNA probes to determine their chromosome location by fluorescent *in situ* hybridization (FISH), following the protocols described in Cabrero *et al.* (2003). FISH probes were labeled with tetramethylrhodamine-5-dUTP (satDNAs, H3 histone and rDNA 5S genes) or fluorescein-12-dUTP (rDNA 45S genes) from Roche.

To build an ideogram reflecting actual chromosome sizes, we quantified chromosomal area in 5 DAPI-stained metaphase I cells by the pyFIA software (Ruiz-Ruano *et al.*, 2011) and expressed it relative to total chromosome area.

Statistical analysis was performed by the non-parametric Mann-Whitney test, with the software Statistica 6.0. Contingency tests were performed with the RXC program (George Carmody, University of Ottawa, Canada) by a Monte Carlo approach to calculate statistical significance, with 5,000 permutations.

Results

B chromosomes carry 5S rRNA genes

The chromosome complement of *E. monticola* consists of nine autosome pairs and one or two X chromosomes in males or females, respectively. Autosomes can be classified into three size groups: four long (L1-L4), three medium-sized (M5-M7) and two short (S8 and S9) chromosomes, the S8 chromosome behaving as the megameric bivalent during first meiotic prophase thus showing positive heteropycnosis (Cabrero *et al.*, 1985). Two males from the HM population, one collected in 2011 and other in 2013, carried a mitotically unstable B chromosome, showing intraindividual variation in number. C-banding showed the presence of pericentromeric dark bands on all A chromosomes, but not on B chromosomes, although the latter were almost as dark as the pericentromeric bands in A chromosomes (Fig. 3.1a). By contrast, only the X chromosome and M5-S9 autosomes showed pericentromeric DAPI$^+$ bands, whereas all four L chromosomes and the B chromosome lacked them (Fig. 3.1b).

FISH analysis showed that 45S rDNA is located on autosomes L2 and L3 (Fig. 3.1c), in consistency with the silver staining pattern reported by Cabrero *et al.* (1985) in specimens from Campos de Otero (Sierra Nevada, Granada, Spain). H3 histone genes were located only on autosome S8 (Fig. 3.1c), whereas 5S rRNA genes were found on L1-L4, S8-S9, X and B chromosomes (Fig. 3.1d). The high number of A chromosomes carrying 5S rRNA genes made this marker useless to investigate B chromosome ori-

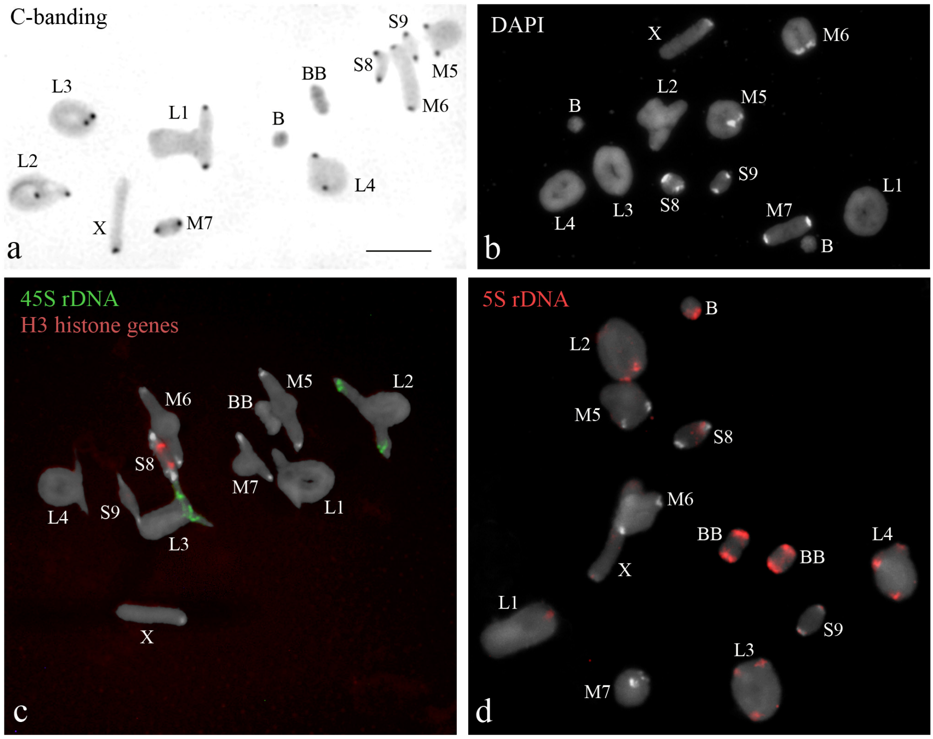


Figure 3.1: C-Banding (a), DAPI fluorescence (b), FISH for 45S rDNA (green) and histone H3 genes (red) (c), and FISH for 5S rDNA (d) in meiotic first metaphase cells of *E. monticola*. In (a), note the presence of C-bands on the pericentromeric regions of all A chromosomes and the dark grey color of the B chromosomes compared to the euchromatin of A chromosomes. In (b) note the absence of DAPI$^+$ bands on the L1-L4 chromosomes. In (c) note the presence of 45S rDNA close to pericentromeric regions of L2 and L3 chromosomes, and of H3 genes on an interstitial region of the S8 chromosome. In (d) note the presence of clusters for 5S RNA genes on L1-L4 and S8 chromosomes, and also on the B chromosomes.

gin, for which reason we analyzed the satellitome in one B-carrying and one B-lacking males.

Satellitome analysis

In the two genomes analyzed, as a whole, we found 41 satDNA sequences grouped into 27 families, with monomer lengths from 5 to 325 nt (mean= 115.5, SD= 99.5), and A+T content from 42.9 to 83.3% (mean= 59.6, SD= 10.08) (Tables 3.1 and 3.S3). We numbered them in order of decreasing abundance in the B-lacking genome to build the satellitome catalog in this species according to Ruiz-Ruano *et al.* (Chapter 2). The most abundant satDNA (EmoSat01-325) represented 0.53% of the B-lacking and 0.55% of the B-carrying genomes. The seventh most abundant satDNA (EmoSat07-5) was the telomeric repeat conserved in insects Frydrychová *et al.* (2004), representing 0.06% of the genome, a very similar figure to that reported in the migratory locust (Chapter 2). The least abundant satDNAs were EmoSat27-102 in the B-lacking genome (0.00032%) and EmoSat25-184 in the B-carrying genome (0.00423%) (Table 3.1).

Table 3.1: Characteristics, divergence and abundance for the satDNA families in *E. monticola*.

		length		Diverg. (%)		Abundance (%)		log2	
SF	SatDNA family	(nt)	G+C	0B	+B	0B	+B	(+B/0B)	+B-0B
	EmoSat01-325	325	37.5	5.76	6.01	0.52623	0.54959	0.06	0.02336
1	EmoSat02-89	89	25.8	6.11	6.26	0.28885	0.26365	-0.13	-0.02520
	EmoSat03-304	304	55.9	12.26	12.22	0.24134	0.24188	0.00	0.00054
	EmoSat04-60	60	16.7	12.13	12.24	0.17223	0.16315	-0.08	-0.00908
2	EmoSat05-16	16	56.3	17.39	17.43	0.10019	0.10598	0.08	0.00579
1	EmoSat06-87	87	26.4	10.13	10.15	0.06739	0.05906	-0.19	-0.00833
	EmoSat07-5	5	53.1	1.44	1.34	0.05976	0.06761	0.18	0.00785
	EmoSat08-41	41	41.5	11.45	11.48	0.04889	0.04831	-0.02	-0.00058
	EmoSat09-14	14	50.0	12.71	12.99	0.04739	0.03547	-0.42	-0.01192
	EmoSat10-302	302	36.4	8.74	8.43	0.03447	0.03579	0.05	0.00132
4	EmoSat11-122	122	36.1	6.10	7.34	0.03090	0.05049	**0.71**	**0.01959**
4	EmoSat12-122	122	35.2	6.38	8.08	0.02630	0.01441	-0.87	-0.01189
	EmoSat13-24	24	41.7	13.65	13.59	0.02517	0.02596	0.04	0.00079
	EmoSat14-24	24	37.5	13.88	13.72	0.01442	0.00953	-0.60	-0.00489
3	EmoSat15-207	207	37.2	8.28	8.49	0.01418	0.01949	0.46	0.00531
3	EmoSat16-208	208	42.3	5.90	5.95	0.01398	0.01275	-0.13	-0.00123
	EmoSat17-97	97	33.3	5.87	6.00	0.01358	0.01447	0.09	0.00089
2	EmoSat18-7	7	57.1	16.17	16.29	0.01299	0.01525	0.23	0.00226
3	EmoSat19-203	203	38.9	9	8.31	0.01205	0.01055	-0.19	-0.00150
3	EmoSat20-199	199	39.7	4.6	4.74	0.00912	0.01229	0.43	0.00317
3	EmoSat21-208	208	38.9	6.25	6.30	0.00832	0.00891	0.10	0.00059
	EmoSat22-12	12	56.4	8.1	9.35	0.00732	0.01802	**1.30**	**0.01070**
	EmoSat23-14	14	28.6	9.98	9.86	0.00588	0.00717	0.29	0.00129
	EmoSat24-101	101	34.7	13.6	13.26	0.00498	0.00501	0.01	0.00003
	EmoSat25-184	184	47.8	10.71	10.27	0.00472	0.00423	-0.16	-0.00049
	EmoSat26-41	41	41.5	16.83	6.22	0.00089	0.09836	**6.78**	**0.09747**
	EmoSat27-102	102	44.1	10.33	2.33	0.00032	0.01697	**5.73**	**0.01665**
	Total					1.79186	1.91435		0.12249

Sequence comparison between the 27 satDNA families revealed the existence of four superfamilies (SF), indicating the common ancestry of some of them, i.e. EmoSat02-89 and EmoSat06-87 (SF1), EmoSat05-16 and EmoSat18-7 (SF2), EmoSat15-207, EmoSat16-208, EmoSat19-203, EmoSat20-199 and EmoSat21-208 (SF3), and EmoSat11-122 and EmoSat12-122 (SF4).

Average satDNA family divergence in the B-carrying genome (9.21%)

was lower than that in the B-lacking one (9.77%), presumably as a result of satDNA amplification in the B chromosome (see below). The minimum divergence was shown by the telomeric repeat (1.44% and 1.34% in the B-lacking and B-carrying genomes, respectively) whereas the maximum divergence was shown by Emosat05-16 (17.39% and 17.43%). These figures are rather similar to those found in the migratory locust (Chapter 2).

We found homology of some satDNAs with previously known repeated elements in Repbase. EmoSat03-304 and EmoSat25-184 showed homology with the *L. migratoria* elements BEL-2_LMi-I and Helitron-N18_LMi, respectively, both laying outside TEs CDS region (Fig. 3.S2)

Fold change (FC) in abundance and difference between +B and 0B (diff) abundance due to the presence of B chromosomes showed that at least EmoSat11-122 (FC= 0.71, diff= 0.01959), EmoSat22-12 (FC= 1.3, diff= 0.01070%), EmoSat26-41 (FC= 6.8, diff= 0.09747%) and EmoSat27-102 (FC= 5.7, diff= 0.01665%) are abundant in the B chromosomes (Table 3.1 and Fig. 3.S1).

Location of satDNAs on standard (A) chromosomes

We designed primers for PCR amplification of 26 satDNAs (excluding the telomeric repeat), and all except four (EmoSat04-60, EmoSat13-24, EmoSat14-24 and EmoSat23-14) were successfully amplified. Note that three of them were actually short satDNAs and the longest showed long stretches of adenines difficulting its specific PCR amplification (Fig. 3.S1). The PCR products of the 22 satDNas amplified were then labeled to build probes for FISH mapping on chromosomes, which indicated that 19 of them were clustered on one or more chromosomes (Fig. 3.2a-d) whereas the three remaining were non-clustered, i.e interspersed across the whole genome (Fig. 3.2e). In the A chromosomes, we observed 64 satDNA clusters, most of them being proximal to the centromeric region (46) and very few being distal (11) or interstitial (7) (Table 3.2). Only one satDNA (EmoSat08-41) was present in all ten A chromosome pairs, and its pericentromeric location suggests its possible involvement in centromeric function (Fig. 3.2a) despite the fact that it was only the eighth satDNA in abundance. In total, nine out of the 17 clustered satDNAs on A chromosomes were located on two or more chromosome pairs, the eight remaining being chromosome specific (Table 3.22). Those satDNAs being located on two or more chromosomes show an interesting distribution pattern, as most of them were limited to M, S and X chromosomes (Fig. 3.2b,c), the only exception being EmoSat03-304 which was located on the four L chromosomes, the X chromosome and the S9 pair (Fig. 3.2d). We calculated an equilocality index for these nine satDNas as the proportion of clusters showing the same location (proximal, interstitial or distal) on non-homologous chromosomes, indicating that 95% of satDNA clusters are equilocal in this species.

The four satDNA superfamilies showed different patterns of chromosomal location (p, i or d) and chromosomes carrying them (Table 3.S4). SF1 was restricted to pericentromeric regions of six M, S and X chromosomes, with only slight differences between the two satDNA members of this superfamily, since EmoSat02-89 (5 clusters) was not found on S9 and EmoSat06-87 (4 clusters) was absent from M5 and X (Table 3.2). A Minimum spanning tree (MST) for this superfamily suggested that the EmoSat06-87 variant could be ancestral for this superfamily (Fig. 3.4a, 3.2), and that the accumulation of mutational changes gave rise to other variants of this family (B and C) and also to the Emosat02-89 family with its three variants (A-C). The high abundance of this latter family (the second most abundant one), and its low divergence (6%), point to a recent massive amplification of this family on the M, S and X chromosomes.

The SF2 superfamily showed exactly the same location for EmoSat05-16 and EmoSat18-7, namely pericentromeric regions of all six M, S and X chromosomes (Table 3.2). MST for this superfamily showed that EmoSat05-16 could have derived from two monomers of EmoSat18-7 with a GT replication slippage in one of them making it growing up to 16 nt (Fig. 3.4b). The presence of these two satDNA families in the same M, S, X and B chromosomes suggests that they precede B chromosome origin.

In high contrast, the five satDNAs included in SF3 showed higher differences between them since two of them (EmoSat16-208 and EmoSat20-199) were chromosome-specific but on different chromosomes (M5 and M7, respectively), whereas the three remaining satDNAs showed interstitial and distal clusters on two (EmoSat19-203), four (EmoSat15-207) or five (EmoSat21-208) chromosomes, mostly on M and S chromosomes, and two exceptional clusters interstitial were on L1 and L3 chromosomes (Table 3.2). The MST for this satDNA superfamily was less informative as it includes five families with about similar abundance, and it is difficult to infer their evolutionary relationships (Fig. 3.4c). Anyway, this is a complex superfamily mostly located on distal and interstitial clusters on M and S autosomes, probably being the oldest of the four superfamilies.

Finally, the SF4 superfamily included the EmoSat11-122 and EmoSat12-122 families located only in the S8 autosome (one interstitial cluster) and the B chromosome (Table 3.2). Both families show interesting differences in chromosome location pattern, as the former is highly abundant in the B chromosome, and scarce in S8, and the latter shows a slightly reversed pattern (Fig. 3.2f). A MST for this superfamily supported the differential amplification of the EmoSat11B-123 variant in the B chromosome and EmoSat11A-122 in the S8 chromosome (Fig. 3.4d).

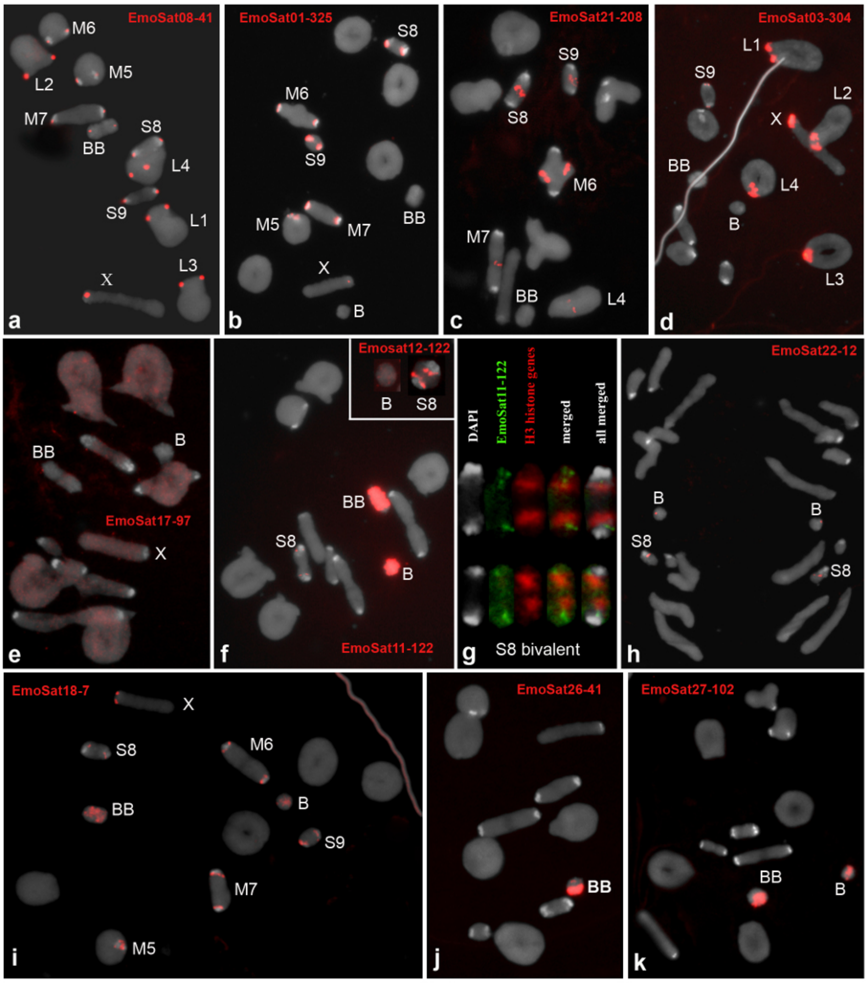


Figure 3.2: FISH for satDNAs found in the *E. monticola* genome on meiotic first metaphase cells (a-g, and i-k) and second anaphase cells (h). a) EmoSat08-41 is present in the pericentromeric regions of all A and B chromosomes. b) EmoSat01-325 is located on pericentromeric regions of M, S and X chromosomes but is absent from L and B ones. c) EmoSat21-208 is interstitially located on the L4 chromosome and distally on M6, M7, S8 and S9. d) EmoSat03-304 is located on pericentromeric regions of L and X chromosomes. e) EmoSat17-97 did not show clusters on A or B chromosomes. f) EmoSat11-122 showed a small cluster on the S8 chromosome, which is proximal in respect to the histone gene cluster (g), and occupies almost all B chromosome length, whereas EmoSat12-122 showed a conspicuous cluster on S8 but was very scarce on the B chromosome (inset in f). h) EmoSat22-12 is interstitially located on the S8 chromosome and pericentromerically on the B chromosome. i) EmoSat18-7 is pericentromerically located on the M, S and X chromosomes and shows conspicuous FISH signal in about half of B chromosome length. j and k) EmoSat26-41 and EmoSat27-102 showed FISH signal only on the B chromosome. B= B chromosome univalent, BB= B chromosome bivalent.

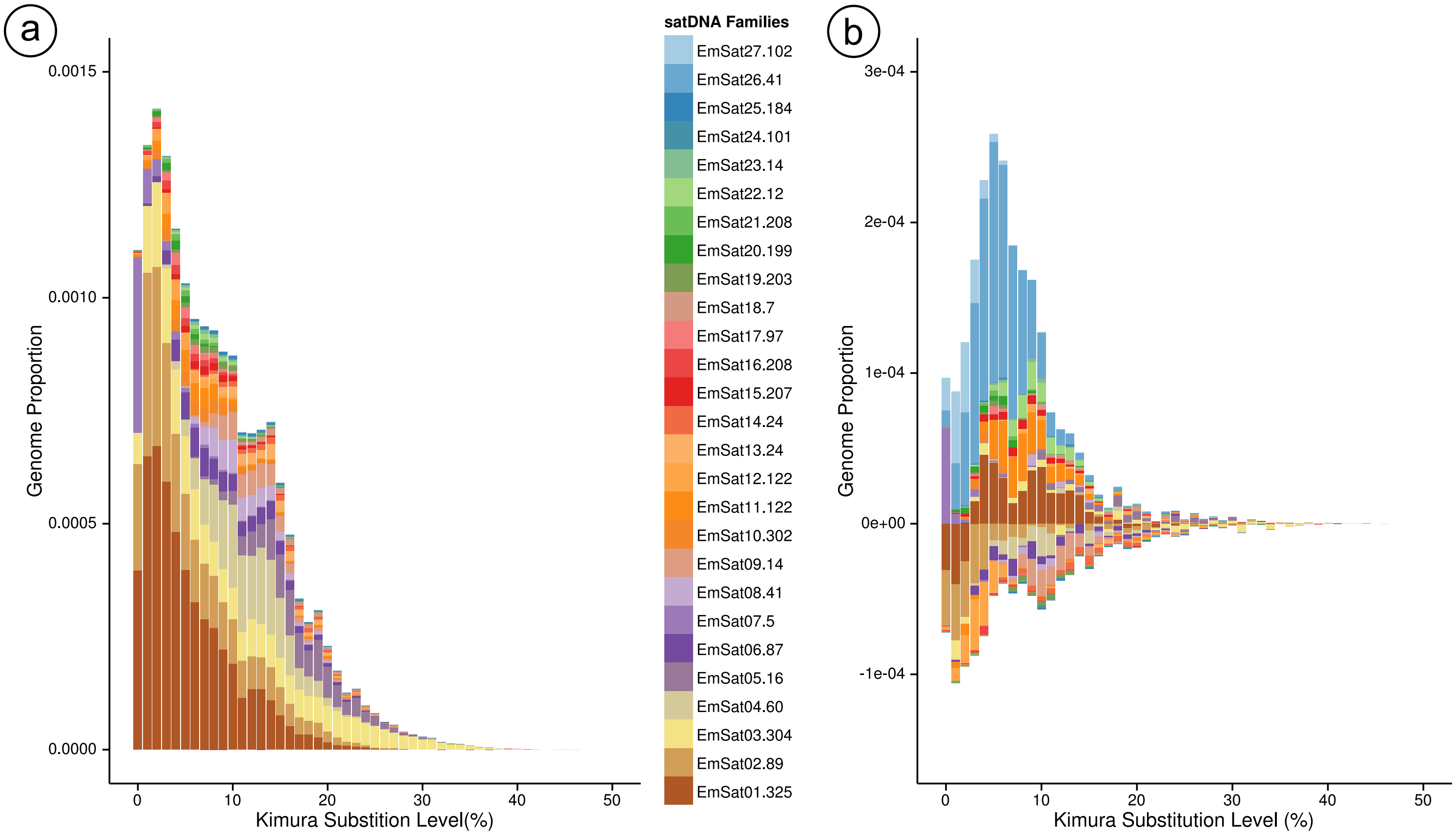


Figure 3.3: Repeat Landscape for the satDNAs of the 0B individual **(a)**. Repeat landscape representing the subtraction of the 0B counts to the +B counts for the satDNAs **(b)**.

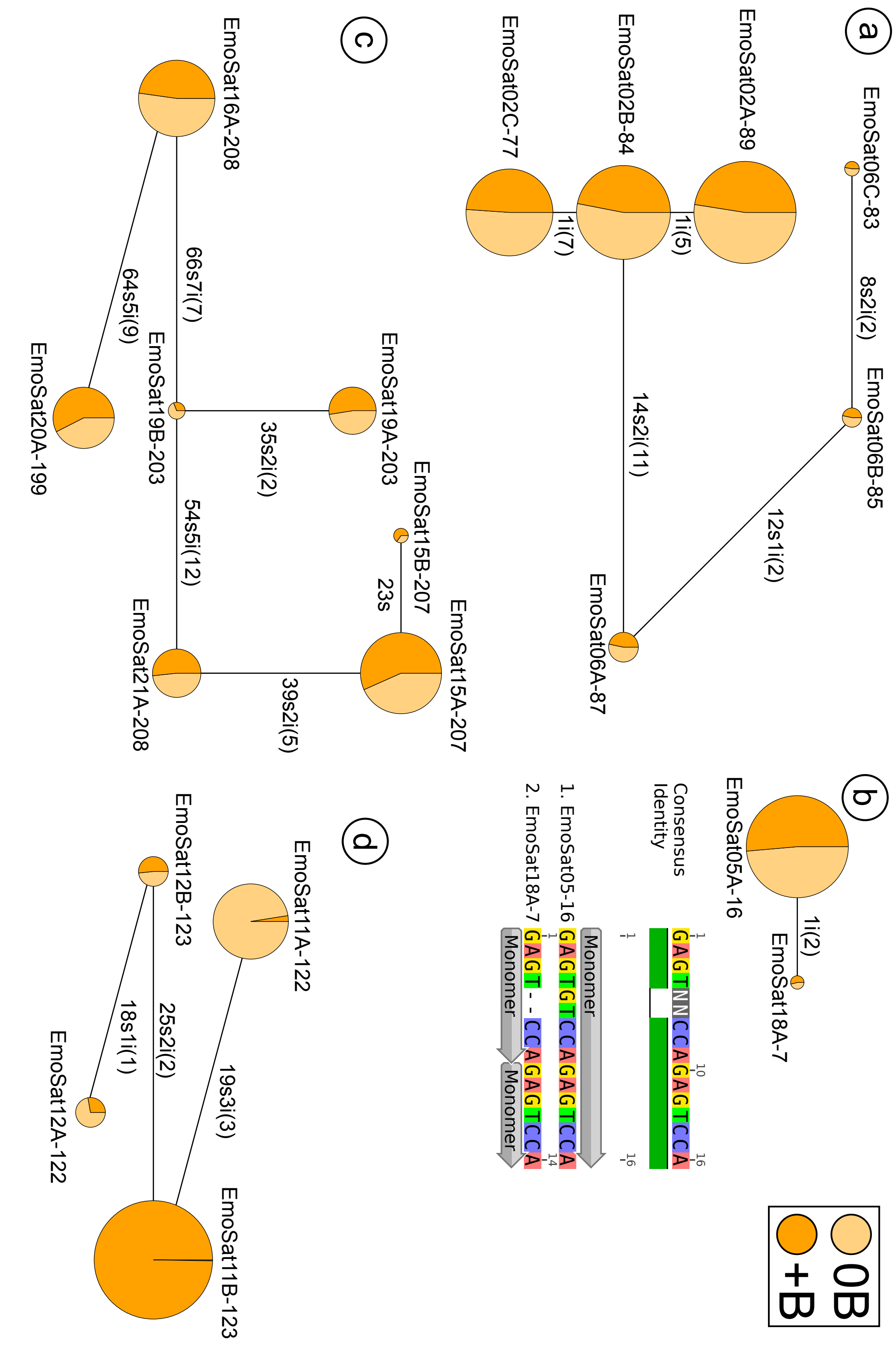


Figure 3.4: Minimum spaning trees for the four satDNA superfamilies 1-4 (a-d).

Table 3.2: Location of the cytogenetic markers used in this study including classical markers and satDNAs.

		Chromosome										
	Repetitive	L1	L2	L3	L4	M5	M6	M7	S8	S9	X	B
	BC	p	p	p	p	p	p	p	p	p	p	
	DAPI					p	p	p	p	p	p	
	45S rDNA		p		i							
	5S rDNA	p	p	pd	pd				p	p	p	p
	H3								i			
SF	SatDNA family	L1	L2	L3	L4	M5	M6	M7	S8	S9	X	B
	EmoSat01-325					p	p	p	p	p	p	
1	EmoSat02-89					p	p	p	p		p	
	EmoSat03-304	p	p	p	p					p	p	
	EmoSat04-60											
2	EmoSat05-16					p	p	p	p	p	p	s
1	EmoSat06-87						p	p	p	p		
	EmoSat07-5											
	EmoSat08-41	p	p	p	p	p	p	p	p	p	p	p
	EmoSat09-14					p			p			
	EmoSat10-302			p								
4	EmoSat11-122								i			pid
4	EmoSat12-122								i			s
	EmoSat13-24											
	EmoSat14-24											
3	EmoSat15-207	i						d	d	d		
3	EmoSat16-208					d						
	EmoSat17-97											
2	EmoSat18-7					p	p	p	p	p	p	pi
3	EmoSat19-203								id	id		
3	EmoSat20-199							d				
3	EmoSat21-208				i		d	d	d	d		
	EmoSat22-12								i			p
	EmoSat23-14											
	EmoSat24-101											
	EmoSat25-184											
	EmoSat26-41											pid
	EmoSat27-102											i
	p	2	2	3	2	6	6	6	7	6	6	5
	i	1	0	0	1	0	0	0	4	1	0	4
	d	0	0	0	0	1	1	3	3	3	0	2
	total	3	2	3	3	7	7	9	14	10	6	11
	sats	3	2	3	3	6	8	9	13	9	6	8
	shared with B	1	1	1	1	3	3	3	6	3	3	–

The different A chromosomes within the *E. monticola* genome differ in the number of satDNAs (excluding the telomeric repeat), with the four L autosomes carrying only 2 or 3 different satDNAs, in sharp contrast with the M, S and X chromosomes carrying 6-13 satDNAs. The S8 autosome was, by far, the chromosome carrying the highest number of satDNAs, with seven proximal, four interstitial and three distal clusters (belonging to 13 different satDNA families) along its entire length (Fig. 3.5).

SatDNAs in the B chromosome

Eight satDNAs were found in the B chromosome, namely EmoSat05-16, EmoSat08-41, EmoSat11-122, EmoSat12-122, EmoSat18-7, EmoSat22-12, EmoSat26-41 and EmoSat27-102. Remarkably, these 8 satDNAs showed significantly shorter monomer length (mean= 57.9, SD= 49.6) than the 11 satDNAs found only on A chromosomes (mean= 195.1, SD= 98.2) (Mann-Whitney test: U= 12, P= 0.008).

The satDNAs being most informative for B chromosome origin were EmoSat11-122, EmoSat12-122 and EmoSat22-12, as they were restricted to the S8 and B chromosomes, suggesting B chromosome origin from the S8 autosome.

Two satDNAs (EmoSat26-41 and EmoSat27-102) did not map on A chromosomes submitted to FISH (Table 3.2) thus appearing to be B-specific. However, they were found bioinformatically in the B-lacking individual, at very low abundances (0.00089% and 0.00032%, respectively) (Table 3.1). Interestingly, bioinformatic analysis in the B-lacking genome revealed the presence of four reads containing tandem monomers of EmoSat26-41 (Fig. 3.6), suggesting that this DNA is tandemly repeated in the A chromosomes, although at very low abundance, but it has massively been amplified in the B chromosome. On the other hand, no tandem repeats were found for EmoSat27-102 in the 0B genome, but we found a contig, annotated as RTE, whose DNA sequence includes this monomer sequence. Remarkably, in the B-carrying genome, this same contig included this satDNA, suggesting that it has most likely arisen in the B chromosome (Fig. 3.7).

Discussion

Genomic insights from the satellitome

The satellitome in *E. monticola* consists of 27 satellite DNAs, less than half than in the migratory locust, where Ruiz-Ruano *et al.* (Chapter 2) found 62. These two species show other remarkable satellitome differences in chromosome distribution patterns (clustered or non-clustered in *E. monticola* but with a third mixed pattern in *L. migratoria*) and the proportion of chromosome-specific satDNAs (one third and about half, respectively).

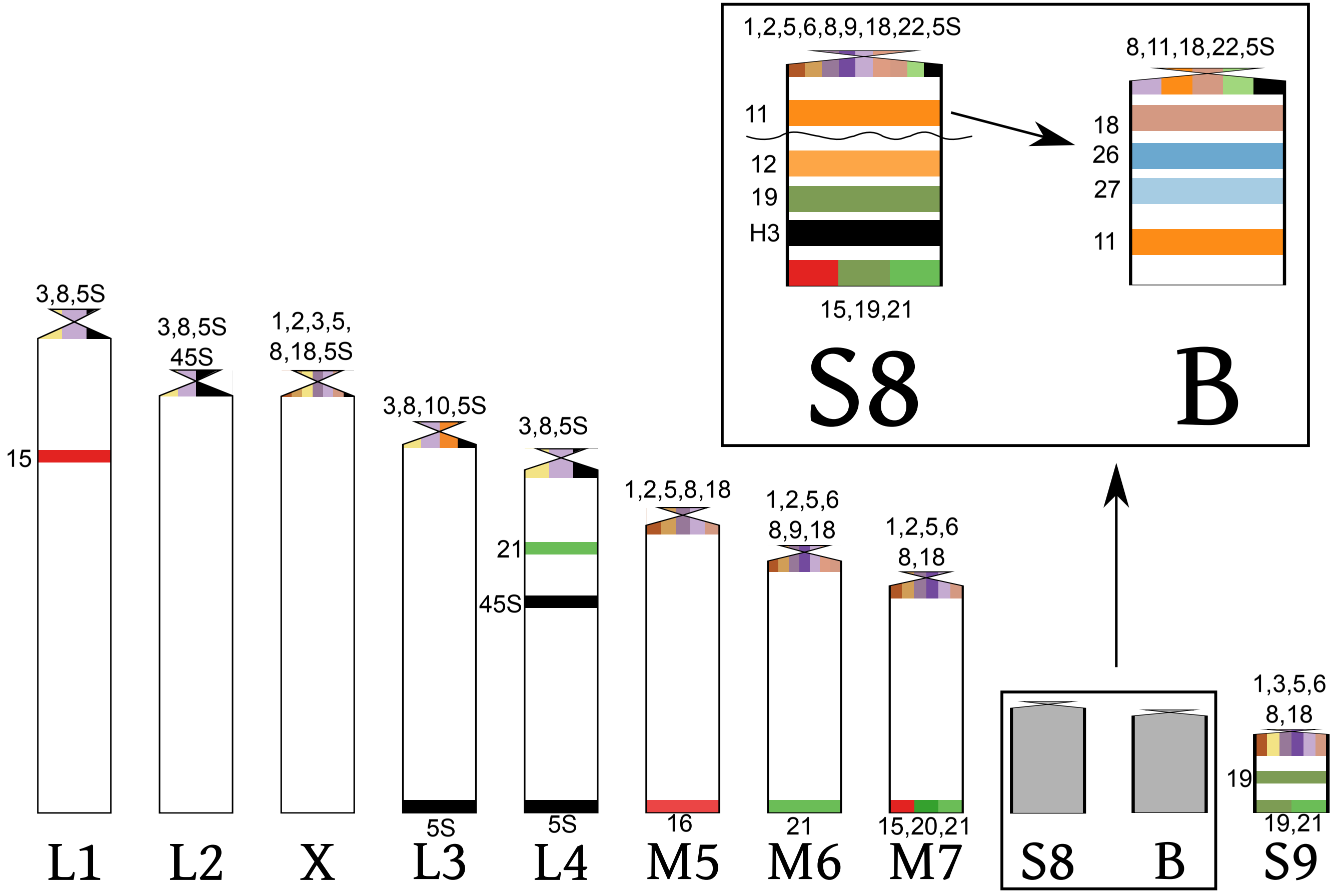


Figure 3.5: Ideogram representing the chromosomal location for each clustered satDNA and the H3 histone, 5S rDNA and 45S rDNA genes.

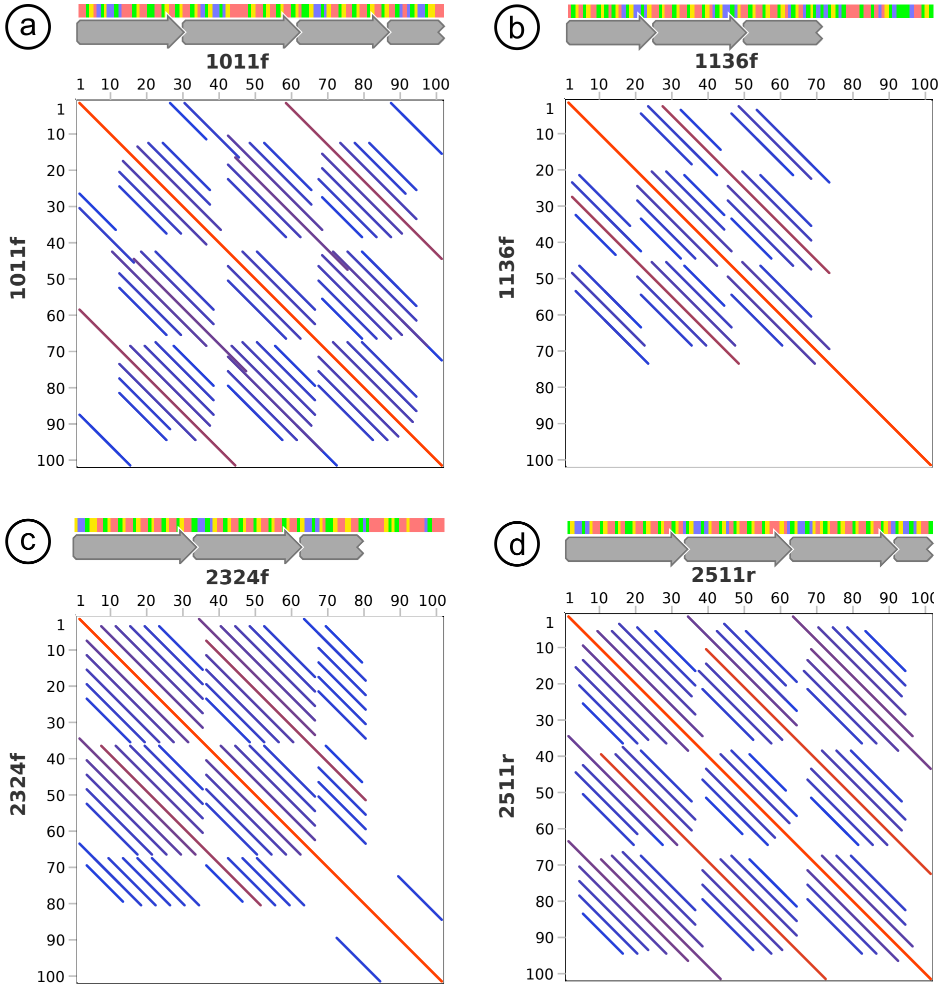


Figure 3.6: The satDNA EmoSat26-41 shows tandem repeats in the A chromosomes. We found four reads in the 0B individual with homology to this satDNA with tandem repeats. Note in the dotplots that two of them are completely covered by repetitions (A and D) and the two remaining (B and C) show repetitions in only a part of the read. We represent the location of the repetitions above the dotplots.

However, the most remarkable difference is the absence of homology between the satellitomes in both species, as they did not shared a single satellite family. This indicates that the satDNA library (Fry & Salser, 1977) might have completely diverged between the Acrididae (*L. migratoria*) and Pamphagidae (*E. monticola*) families of grasshoppers, which shared their most recent common ancestor about 100 mya (Song *et al.*, 2015), and indicates that the evolutionary lifespan of satDNA in grasshoppers is usually lower than this figure.

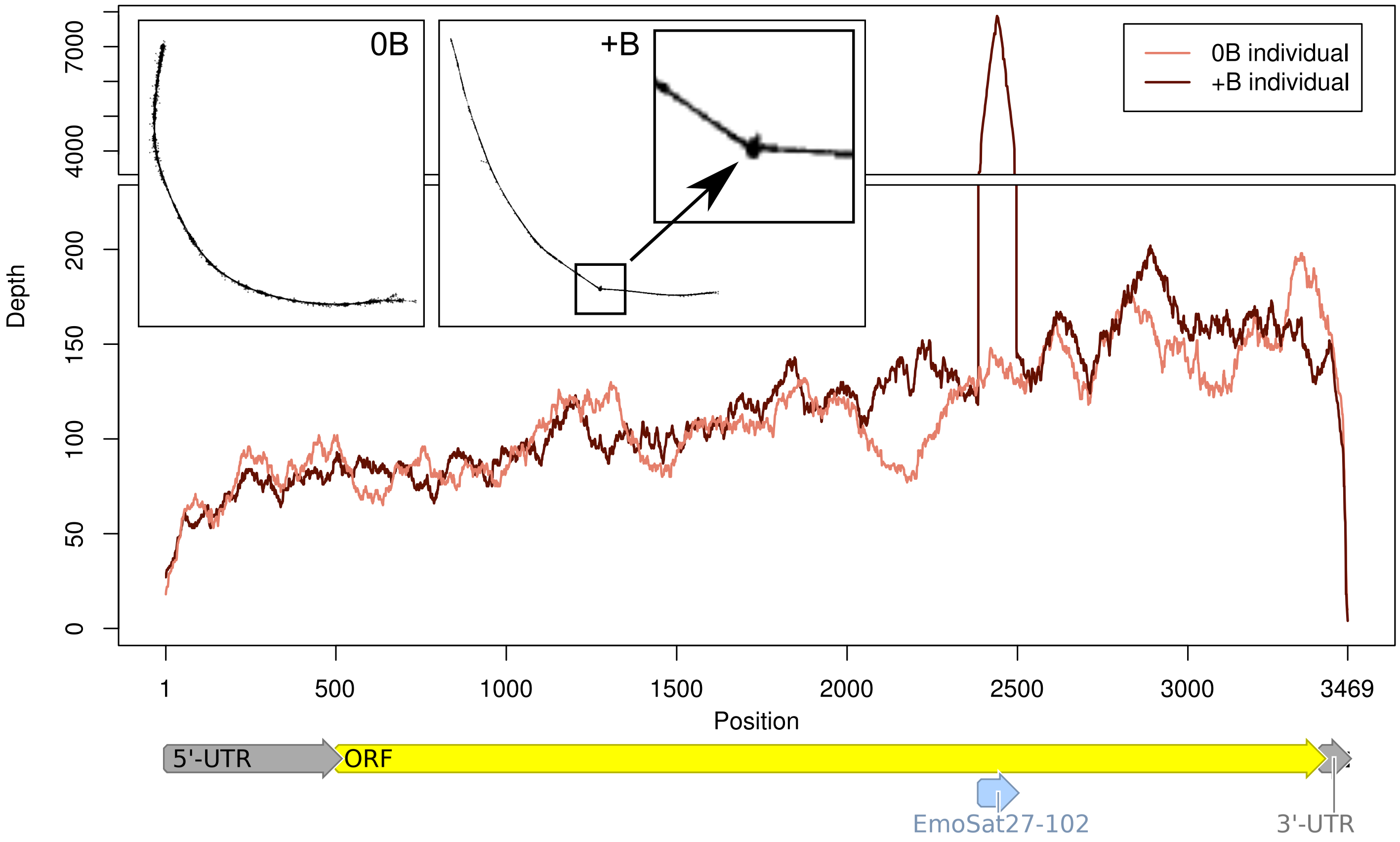


Figure 3.7: The satDNA EmoSat27-102 is derived from a RTE element and B-specific. Here we show the coverage per position of this RTE element in the 0B (soft red line) and +B (dark red line) individuals. Note an increase of ~40 times for the +B individual in the region homologous with EmoSat27-102. We selected reads homologous to this RTE element and clustered with RepeatExplorer separately in both individuals. We got a linear graph for the two individuals, but in the B-carrying one the graph shows a spherical structure coincident with the satDNA region (inset). We do not found tandem repeats for EmoSat27-102 in the 0B individual.

In *E. monticola*, EmoSat08-41 was the only satDNA being present in all ten A chromosome pairs (and also in the B chromosome), and its pericentromeric location suggests its possible involvement in centromeric function. The fact that it was only the eight satDNA in order of decreasing abundance, calls attention on the usual assumption that the most abundant satDNA in a genome plays a centromeric role (Melters *et al.*, 2013).

The satDNA showing the highest A+T content was EmoSat04-60 (83.3%), but the long stretches of adenines in its DNA sequence impeded its PCR amplification (Fig. 3.S1). Out of the remaining satDNAs, the two belonging to SF1 (EmoSat02-89 and EmoSat06-87) showed the highest A+T content (74.2% and 73.6%, respectively). Remarkably, the joint pattern of chromosome distribution for these two satDNAs (pericentromeric on M, S and X chromosomes) was coincident with the DAPI$^+$ pattern (see Fig. 3.1 and Table 3.2). This suggests that SF1 satDNAs are responsible for the DAPI$^+$ pattern in these chromosomes. A similar distribution pattern was shown for EmoSat01-325, with 62.5% A+T content. Therefore, the possibility exists that this A+T content is enough to yield the DAPI$^+$ pattern. However, the fact that the EmoSat10-302 and EmoSat15-207 sequences show 63.6% and 62.8% A+T, respectively, and are located on two L chromosomes lacking a DAPI$^+$ pattern, strongly suggests that, in *E. monticola*, the DAPI$^+$ pattern is due to the presence of the two SF1 satDNAs. In the migratory locust, however, all chromosomes show DAPI$^+$ bands in the pericentromeric region, and the only satDNA showing this location pattern (LmiSat01-193) exhibited only 59.59% A+T content (Chapter 2). This suggests either that A+T biases as low as 60% can yield a DAPI$^+$ pattern in some species, or else that the LmiSat01-193 is not responsible for the DAPI$^+$ pattern in *L. migratoria*. Knowledge of the satellitome in other species will allow, in next years, determining the minimum threshold for the A+T bias characterizing the DAPI$^+$ pattern.

According to the DAPI pattern, the A chromosomes in *E. monticola* are compartmentalized into two groups, with the M, S and X chromosomes showing DAPI$^+$ pericentromeric bands which are absent from the L chromosomes. This resembles the bimodal karyotype of the domestic pig, where all 12 acrocentric chromosomes carry DAPI$^+$ pericentromeric heterochromatin which is absent from the 24 biarmed autosomes and the X chromosome (Schwarzacher *et al.*, 1984), and it has been shown that these differences are due to the presence of different families of satDNA (Jantsch *et al.*, 1990). Our present results demonstrate that the DAPI$^+$ compartmentalization in *E. monticola* is also due to different satDNA content (for SF1, see above). However, the satellitome in this species also demonstrates compartmentalization for several other satDNA families being present in M, S and X chromosomes but not in L ones (e.g. EmoSat01-325, EmoSat05-16 and EmoSat18-7). This suggests a general tendency of these two groups of chromosomes to exchange different satDNAs and, perhaps, other repetitive elements.

SatDNAs in *E. monticola* show an extremely high tendency to equilocal distribution between non-homologous chromosomes (95%), much higher than the 64% calculated in the *Locusta migratoria* satellitome (Chapter 2). Chromosomes show characteristic polarized arrangements during interphase (Rabl orientation) and meiosis (bouquet) which join together all centromeres and/or telomeres in a same cell region (Carlton & Cande, 2002), and it has been suggested that these polarized arrangements can facilitate the equilocal non-clustered of satDNAs between non-homologous chromosomes (Jantsch *et al.*, 1990; Žinić *et al.*, 2000; Mravinac & Plohl, 2010).

The S8 autosome in *E. monticola* behaves as the megameric bivalent during spermatogenesis, so that it is facultatively heterochromatic and is thus highly condensed during first meiotic prophase, as in other grasshoppers (Corey, 1938). Remarkably, in *E. monticola* and the migratory locust, the megameric bivalent is the A chromosome carrying the highest number of different satDNAs, on which basis Ruiz-Ruano *et al.* (Chapter 2) suggested a possible relationship between satDNA content in this chromosome and its facultative heterochromatinization. Our present results support this conclusion.

SatDNA content illuminates B chromosome origin

The A chromosome sharing more satDNAs with the B chromosome was S8, which carried six out of the eight satDNAs contained in the B chromosome whereas none of the remaining A chromosomes carried more than three of the B-satDNAs, and all of them were also present in the S8 autosome (Table 3.2). Remarkably, three satDNAs shared with the B chromosome (EmoSat05-16, EmoSat08-41 and EmoSat18-7) are located in the pericentromeric region of S8 (Table 3.2), whereas the three remainder (EmoSat11-122, EmoSat12-122 and EmoSat22-12) are interstitially located next to the proximal DAPI$^+$ band in this A chromosome (Fig. 3.2f-h) and are exclusive of the S8 and B chromosomes (Table 3.2). The location of these three satDNAs restricted to S8 and B is strong evidence for B origin from S8. Interestingly, the B chromosome carries much higher amounts of EmoSat11-122 than EmoSat12-122, both belonging to SF4, whereas the S8 chromosome shows the opposite pattern (Fig. 3.2f), indicating that the amounts of both satDNAs have changed in these two chromosomes since B origin.

In addition to sharing these six S8 proximal-interstitial satDNAs, a S8-derived B chromosome could carry other S8-satDNAs. However, the three distal satDNAs (EmoSat15-207, EmoSat19-203 and EmoSat21-208) (Table

3.2) are absent from the B chromosome, indicating that the B chromosome derived from a partial S8 chromosome not including distal regions. Double FISH for the EmoSat11-122 satDNA and H3 histone genes (the latter being interstitially located on the S8 autosome but absent from the B chromosome) showed that, in the S8 chromosome, EmoSat11-122 is closer to the pericentromeric DAPI$^+$ band than H3 genes (Fig.3.22g), thus delimiting B chromosome origin to the proximal third of S8 including this satDNA but not the H3 histone genes. This explains the presence, in the B chromosome, of the six S8 proximal-interstitial satDNAs, as well as the 5S rDNA proximally located on S8, but not the interstitial H3 histone genes or the distal SF3 satDNAs.

Although this hypothesis sounds logical, the absence in the B of four satDNAs being present in the proximal region of S8 (EmoSat01-325, EmoSat02-89, EmoSat06-87 and EmoSat09-14) demands an explanation. Three out of the eight satDNAs contained in the B chromosome (EmoSat05-16, EmoSat11-122 and EmoSat18-7) were conspicuously more abundant in the B than in the S8 autosome (Fig. 3.2f). This demonstrates that satDNA abundance has changed in the B chromosome, since its origin, through the massive amplification of these three satDNAs. In addition, FISH mapping suggested that the B chromosome carries two exclusive satDNA families (EmoSat26-41 and EmoSat27-102) (Fig. 3.2i) which did not map on A chromosomes. Remarkably, the bioinformatic analysis showed that these two satDNAs were actually present in the B-lacking individuals, suggesting that they might be present in the S8, at small amounts, and that they have been massively amplified in the B chromosome. In fact, one of them (EmoSat26-41) is already tandemly arranged in the 0B genome but its abundance is below the FISH threshold, whereas the other most likely arose in the B chromosome itself from an RTE element, providing a nice example on how satDNA can emerge from TEs (see other examples in Meštrović *et al.* 2015).

The differential amplification of these five satDNAs in the B chromosome illustrates how B chromosomes enrich very much in repetitive DNA after their origin (Banaei-Moghaddam *et al.*, 2015). A parallel decrease in the relative proportion of the four S8 pericentric satDNAs which are apparently absent from the B chromosome (i.e. EmoSat01-325, EmoSat02-89, EmoSat06-87 and EmoSat09-14), or even their lost, would explain the difficulty in finding them in the B chromosome. The absence of DAPI$^+$ bands in the B chromosome, which in this species are associated with the presence of two of the former satDNAs belonging to the SF1 superfamily (EmoSat02-89 and EmoSat06-87), would appear to be consistent with B origin preceding SF1 origin, or else SF1 lost after B origin. The nucleotidic substitution rate for different satDNAs has been estimated to be 0,2% per Myr in Cetacea

(Arnason *et al.*, 1992), 3% in the Drosophila obscura group (Bachmann & Sperlich, 1993), and 0.5-1% in Bovids (Nijman & Lenstra, 2001). Bearing in mind that EmoSat02-89 and EmoSat06-87 showed 16% divergence between them, and assuming 1% substitution rate, we can estimate an approximate age of 16 Myr for the SF1 superfamily, which makes unlikely B origin prior to SF1, since the few cases of B chromosome age hitherto dated, in several organisms, indicate that B chromosome life span is usually lower than 3 Myr (Hewitt & Ruscoe, 1971; Lamb *et al.*, 2007; Teruel *et al.*, 2010; Martis *et al.*, 2012). On this basis, we consider that SF1 preceded B origin and that this superfamily is very scarce or absent in the B chromosome due to partial or full loss.

The finding of B-specific DNA sequences has been interpreted in the past as evidence for the interspecific origin of B chromosomes (for review, see Camacho (2005)). We could thus be tempted to reach this conclusion in *E. monticola* on the basis of the two B-specific satDNAs found in the B chromosome. This flightless grasshopper lives in Sierra Nevada (Granada, Spain) from 2,000 to 3,400 m asl, partly coinciding with *E. rubioi* in its distribution at the highest altitude. This might have provided ample opportunities for interspecific hybridization and B chromosome origin, but these two species show clear differences in penis morphology (Cabrero *et al.*, 1985) which could make mating inviable. Alternatively, the B chromosome could have arisen after intraspecific mating between individuals from two populations previously isolated geographically during pleistocene glaciations. Previous description of chromosomal races in *E. monticola* (Cabrero *et al.*, 1985) supports this possibility. Postglacial hybridization between diversified populations might have facilitated B chromosome origin as a byproduct of chromosome instability due to genetic differences accumulated during the isolation period. The high coincidence in satDNA content between S8 and B suggests that this latter hypothesis is more likely than the interspecific hybridization.

Our present results suggest that EmoSat26-41 and EmoSat27-102 could have arisen intrapopulationally through differential amplification and *de novo* origin from a TE, respectively. Similar de novo origin of B-specific satDNA families, from TEs, has previously been reported in rye (Langdon *et al.*, 2000), where NGS analysis has shown intraspecific B chromosome origin (Martis *et al.*, 2012). In *L. migratoria*, no B-specific satDNA has been formed (Chapter 4) even though the B chromosome is quite old (1-4 mya). Therefore, the origin of B-specific satDNA appears to be contingent in each species.

Taken together, all former considerations call attention on the possibility that B chromosomes carrying B-specific DNA sequences can have arisen intraspecifically. In fact, the intraspecific mode of origin appears to be much more frequent than the interspecific one (see Table 3.S1), and it is also more parsimonious since the latter implies breaking interspecies reproductive isolating barriers, which could be feasible only in certain types of eukaryote organisms. The interspecific mode of B chromosome origin has been directly witnessed in interspecific controlled crosses in *Nasonia* (Perfectti & Werren, 2001). Of course, situations of poor coadaptation, like the former, can be a source of chromosome instability leading to B chromosome origin, but our present results also suggest that about similar situations can occur intraspecifically. Therefore, we suggest that the interspecific explanation, except direct experimental demonstration, should be the last resort when no evidence for intraspecific origin is found.

References

Arnason U, Gretarsdottir S, Widegren B (1992) Mysticete (baleen whale) relationships based upon the sequence of the common cetacean DNA satellite. *Molecular biology and evolution*, **9**, 1018–1028.

Bachmann L, Sperlich D (1993) Gradual evolution of a specific satellite DNA family in *Drosophila ambigua*, *D. tristis*, and *D. obscura*. *Molecular biology and evolution*, **10**, 647–659.

Banaei-Moghaddam AM, Martis MM, Macas J, *et al.* (2015) Genes on B chromosomes: old questions revisited with new tools. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, **1849**, 64–70.

Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 1.

Bauerly E, Hughes SE, Vietti DR, Miller DE, McDowell W, Hawley RS (2014) Discovery of supernumerary B chromosomes in *Drosophila melanogaster*. *Genetics*, **196**, 1007–1016.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, p. btu170.

Cabrero J, Bakkali M, Bugrov A, *et al.* (2003) Multiregional origin of B chromosomes in the grasshopper *Eyprepocnemis plorans*. *Chromosoma*, **112**, 207–211.

Cabrero J, Camacho JPM, Pascual F (1985) Cytotaxonomic studies on pamphagids genus *Eumigus*. Detection of two chromosomal races in *E. monticol*a (Rambur)(Insecta, Orthoptera). *Caryologia*, **38**, 1–12.

Camacho JPM (2005) B chromosomes. *The Evolution of the Genome (ed. T. R. Gregory)*, pp. 223–286.

Camacho JPM, Cabrero J, López-León MD, Cabral-de Mello DC, Ruiz-Ruano FJ (2014) Grasshoppers (Orthoptera). In *Protocols for Cytogenetic Mapping of Arthropod Genomes (ed. I. V. Sharakhov)*, pp. 381–438. CRC Press.

Carlton PM, Cande WZ (2002) Telomeres act autonomously in maize to organize the meiotic bouquet from a semipolarized chromosome orientation. *The Journal of cell biology*, **157**, 231–242.

Menezes-de Carvalho NZ, Palacios-Gimenez OM, Milani D, Cabral-de Mello DC (2015) High similarity of U2 snDNA sequence between A and B chromosomes in the grasshopper *Abracris flavolineata*. *Molecular Genetics and Genomics*, **290**, 1787–1792.

Corey HI (1938) Heteropycnotic elements of orthopteran chromosomes. *Arch. Biol*, **49**, 159–172.

Drummond AJ, Ashton B, Cheung M, *et al.* (2009) Geneious v. 4.8. 5 Biomatters Ltd. *Aukland, New Zealand*.

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources*, **10**, 564–567.

Fry K, Salser W (1977) Nucleotide sequences of HS-α satellite DNA from kangaroo rat Dipodomys ordii and characterization of similar sequences in other rodents. *Cell*, **12**, 1069–1084.

Frydrychová R, Grossmann P, Trubac P, Vítková M, Marec Fe (2004) Phylogenetic distribution of TTAGG telomeric repeats in insects. *Genome*, **47**, 163–178.

Gruber SL, Diniz D, Sobrinho-Scudeler PE, Foresti F, Haddad CFB, Kasahara S (2014) Possible interspecific origin of the B chromosome of *Hypsiboas albopunctatus* (Spix, 1824)(Anura, Hylidae), revealed by microdissection, chromosome painting, and reverse hybridisation. *Comparative cytogenetics*, **8**, 185.

Hewitt G, Ruscoe C (1971) Changes in microclimate correlated with a cline for B-chromosomes in the grasshopper *Myrmeleotettix maculatus* (Thunb.)(Orthoptera: Acrididae). *The Journal of Animal Ecology*, pp. 753–765.

Jantsch M, Hamilton B, Mayr B, Schweizer D (1990) Meiotic chromosome behaviour reflects levels of sequence divergence in *Sus scrofa domestica* satellite DNA. *Chromosoma*, **99**, 330–335.

Kent WJ (2002) BLAT—the BLAST-like alignment tool. *Genome research*, **12**, 656–664.

Kour G, Kour B, Kaul S, Dhar MK (2009) Genetic and epigenetic instability of amplification-prone sequences of a novel B chromosome induced by tissue culture in *Plantago lagopus* L. *Plant cell reports*, **28**, 1857–1867.

Lamb JC, Riddle NC, Cheng YM, Theuri J, Birchler JA (2007) Localization and transcription of a retrotransposon-derived element on the maize B chromosome. *Chromosome research*, **15**, 383–398.

Langdon T, Seago C, Jones RN, *et al.* (2000) *De novo* evolution of satellite DNA on the rye B chromosome. *Genetics*, **154**, 869–884.

Marschner S, Meister A, Blattner FR, Houben A (2007) Evolution and function of B chromosome 45s rDNA sequences in Brachycome dichromosomatica. *Genome*, **50**, 638–644.

Martis MM, Klemme S, Banaei-Moghaddam AM, *et al.* (2012) Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proceedings of the National Academy of Sciences*, **109**, 13343–13346.

McAllister BF, Werren JH (1997) Hybrid origin of a B chromosome (PSR) in the parasitic wasp *Nasonia vitripennis*. *Chromosoma*, **106**, 243–253.

Melters DP, Bradnam KR, Young HA, *et al.* (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol*, **14**, R10.

Meštrović N, Mravinac B, Pavlek M, Vojvoda-Zeljko T, Šatović E, Plohl M (2015) Structural and functional liaisons between transposable elements and satellite DNAs. *Chromosome Research*, **23**, 583–596.

Mravinac B, Plohl M (2010) Parallelism in evolution of highly repetitive DNAs in sibling species. *Molecular biology and evolution*, **27**, 1857–1867.

Nijman IJ, Lenstra JA (2001) Mutation and recombination in cattle satellite DNA: a feedback model for the evolution of satellite DNA repeats. *Journal of Molecular Evolution*, **52**, 361–371.

Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. *Genome research*, **11**, 1725–1729.

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792–793.

Pansonato-Alves JC, Serrano ÉA, Utsunomia R, *et al.* (2014) Single origin of sex chromosomes and multiple origins of B chromosomes in fish genus *Characidium*. *PloS one*, **9**, e107169.

Perfectti F, Werren JH (2001) The interspecific origin of B chromosomes: experimental evidence. *Evolution*, **55**, 1069–1073.

Ruiz-Ruano FJ, Ruiz-Estévez M, Rodríguez-Pérez J, López-Pino JL, Cabrero J, Camacho JPM (2011) DNA amount of X and B chromosomes in the grasshoppers *Eyprepocnemis plorans* and *Locusta migratoria*. *Cytogenetic and genome research*, **134**, 120–126.

Santos LP, Castro JP, Francisco CM, *et al.* (2013) Cytogenetic analysis in the neotropical fish *Astyanax goyacensis* Eigenmann, 1908 (Characidae, incertae sedis): karyotype description and occurrence of B microchromosomes. *Molecular cytogenetics*, **6**, 48.

Schmieder R, Edwards R (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS one*, **6**, e17288.

Schwarzacher T, Mayr B, Schweizer D (1984) Heterochromatin and nucleolus-organizer-region behaviour at male pachytene of *Sus scrofa domestica*. *Chromosoma*, **91**, 12–19.

Sharbel TF, Green DM, Houben A (1998) B-chromosome origin in the endemic New Zealand frog *Leiopelma hochstetteri* through sex chromosome devolution. *Genome*, **41**, 14–22.

Silva DMZdA, Pansonato-Alves JC, Utsunomia R, *et al.* (2014) Delimiting the Origin of a B Chromosome by FISH Mapping, Chromosome Painting and DNA Sequence Analysis in Astyanax paranae (Teleostei, Characiformes). *PLOS ONE*, **9**, e94896.

Smit AFA, Hubley R, Green P (2013) RepeatMasker Open-4.0.

Song H, Amédégnato C, Cigliano MM, *et al.* (2015) 300 million years of diversification: elucidating the patterns of orthopteran evolution based on comprehensive taxon and gene sampling. *Cladistics*, **31**, 621–651.

Teruel M, Cabrero J, Perfectti F, Camacho JPM (2010) B chromosome ancestry revealed by histone genes in the migratory locust. *Chromosoma*, **119**, 217–225.

Teruel M, Ruíz-Ruano FJ, Marchal JA, *et al.* (2014) Disparate molecular evolution of two types of repetitive DNAs in the genome of the grasshopper *Eyprepocnemis plorans*. *Heredity*, **112**, 531–542.

Tosta VC, Marthe JB, Tavares MG, *et al.* (2014) Possible Introgression of B Chromosomes between Bee Species (Genus *Partamona*). *Cytogenetic and genome research*, **144**, 217–223.

Van Vugt JJ, de Jong H, Stouthamer R (2009) The origin of a selfish B chromosome triggering paternal sex ratio in the parasitoid wasp *Trichogramma kaykai*. *Proceedings of the Royal Society of London B: Biological Sciences*, **276**, 4149–4154.

Zhou Q, Zhu Hm, Huang Qf, *et al.* (2012) Deciphering neo-sex and B chromosome evolution by the draft genome of *Drosophila albomicans*. *BMC genomics*, **13**, 109.

Žinić SD, Ugarković D, Cornudella L, Plohl M (2000) A novel interspersed type of organization of satellite DNAs in *Tribolium madens* heterochromatin. *Chromosome Research*, **8**, 201–212.

Supplementary Information

Supplementary Figures

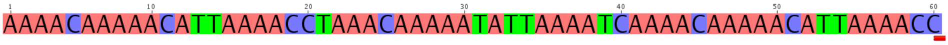

1
10
20
30
40
50
60
AAAACAAAAACATTAAAACCTAAACAAAAATATTAAAATCAAAACAAAAACATTAAAACC

Figure 3.S1: Consensus sequence of the satDNA EmoSat04-60.

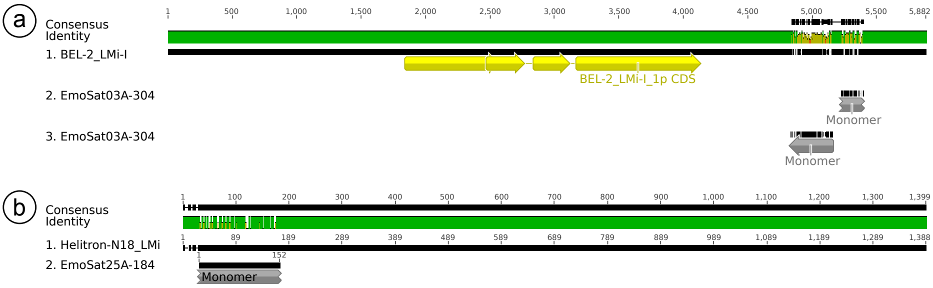


Figure 3.S2: RepBase homology

Supplementary Tables

Table 3.S1: B chromosome origin in other publications.

Organism	Species	Type of origin	Evidence	Ancestry	Year	Reference
Wasp	*Nasonia vitripennis*	interspecific	NATE		1997	McAllister & Werren (1997)
Frog	*Leiopelma hochstetteri*	intraspecific	microdissection	W	1998	Sharbel *et al.* (1998)
Plant	*Brachycome dichromosomatica*	**intraspecific**	ITS2		2007	Marschner *et al.* (2007)
Plant	*Plantago lagopus*	**intraspecific**	rDNA		2009	Kour *et al.* (2009)
Wasp	*Trichogramma kaykai*	interspecific	ITS2		2009	Van Vugt *et al.* (2009)
Grasshopper	*Locusta migratoria*	**intraspecific**	DNA seq	A8	2010	Teruel *et al.* (2010)
Fruit fly	*Drosophila albomicans*	**intraspecific**	NGS		2012	Zhou *et al.* (2012)
Plant	*Secale cereale*	**intraspecific**	NGS	3R and 7R	2012	Martis *et al.* (2012)
Fish	*Astyanax goyacensis*	**intraspecific**	CP		2013	Santos *et al.* (2013)
Fish	*Astyanax paranae*	**intraspecific**	CP		2014	Silva *et al.* (2014)
Fish	*Characidium gomesi*	**intraspecific**	CP	sex chrom	2014	Pansonato-Alves *et al.* (2014)
Fish	*Characidium pterostictum*	**intraspecific**	CP	sex chrom	2014	Pansonato-Alves *et al.* (2014)
Fish	*Characidium oiticicai*	**intraspecific**	CP	sex chrom	2014	Pansonato-Alves *et al.* (2014)
Fruit fly	*Drosophila melanogaster*	**intraspecific**	DNA seq	IV	2014	Bauerly *et al.* (2014)
Grasshopper	*Eyprepocnemis plorans*	**intraspecific**	DNA seq	A11	2014	Teruel *et al.* (2014)
Fish	*Hypsiboas albopunctatus*	interspecific	CP		2014	Gruber *et al.* (2014)
Bee	*Partamona helleri*	interspecific	SCAR		2014	Tosta *et al.* (2014)
Grasshopper	*Abracris flavolineata*	**intraspecific**	U2	A1	2015	Menezes-de Carvalho *et al.* (2015)
Grasshopper	*Eumigus monticola*	**intraspecific**	NGS	A8		This study

Table 3.S2: Primers designed for this work to amplified the different satDNA families.

SatDNA family	F	R
EmoSat01-325	CACCCTGGGCCATATCAACA	AGTCTCTTTATTCTTTGGTGTTCA
EmoSat02-89	TAATTCAGATCTAGTACGACCTACGAA	AATCGTCGTAGTGGTACAGCAT
EmoSat03-304	ACTATGTGTCCACGTTCGGC	AGACCTGAGCAAAACGCTCG
EmoSat04-60	TGTTTTGGTTTTAATGTTTTTGTTTTG	CAAAACAAAAACATTAAAACCAAAACA
EmoSat05-16	GGACTCTGGACACTCTGGACTCTGG	CCAGAGTCCAGAGTGTCCAGAGTCC
EmoSat05-16-B	CTACCGAGAGTCATAGTTACTCC	GTAACTATGACTCTCGGTAGAGG
EmoSat06-87	TAACTAGTGCTTGACAAATACGACT	AAGAAAGGGGAAGATAGTCGT
EmoSat06-87-B	TCCCCTTTCTTATTTAGGGCAGA	TCGTCGTATTTGTCCTGCACT
EmoSat08-41	TCTAAGTGCCAAAACGACGC	GGAACAAGCAAAAAGCGTCG
EmoSat09-14	GGCGACAGTTTACTGGCG	CGCCAGTAAACTGTCGCC
EmoSat09-14-B	AAACTGGCGCCAGTAAAC	GTTTACTGGCGCCAGTTT
EmoSat10-302	GCGATGTATTTCCTTGAGTGCC	CGCTCCAGAAAAGAAGGTAAGACA
EmoSat11-122	CAAATGATGGTTCCCTGCGC	TGGCTTTAAAAGAGGTATTTCC
EmoSat12-122	CCACATGCACAGGCACTACT	TTTCCTACATTTTCAACTCTGGCTC
EmoSat13-24	TCATAGTTACTCCTCTACTGAGAGTCA	TGACTCTCAGTAGAGGAGTAACTATGA
EmoSat14-24	TCTTCTCTCCATTCGATTTCGATATC	GATATCGAAATCGAATGGAGAGAAGA
EmoSat14-24-B	AGAGAAGATATCGAAATCGAATGG	TTCGATTTCGATATCTTCTCTCC
EmoSat15-207	GGGAGGGGAATTAAGATAGCAA	CCCTCTGACTACTGAGCAGC
EmoSat16-208	TGGGGGAGGGCTTGATTCTT	TCCACATGGTCGTTGTCTCG
EmoSat17-97	CCCCTTGGTGCTTTCAAATTTTT	GGGACTTCCGAGCTTAACT
EmoSat18-7	TCTGGACTCTGGACTC	GAGTCCAGAGTCCAGA
EmoSat19-203	GCCACTGAGCTGGATTAACTATAAG	GGCTCCTTCTGCTTGCCATGC
EmoSat20-199	TTTTAAGCTAGCACTCTAACTTTGA	TACTGCAGGGATGCTCCCAT
EmoSat21-208	GAATCATCCCAGCATTACATTTAAG	AACAGCTGAATGTTGTCTTGTGG
EmoSat22-12	CCAACAATGTGTCCAACAATGTGTCC	GGACACATTGTTGGACACATTGTTGG
EmoSat23-14	AAACGAAAGTTCATAAACGAAAGTTCA	TGAACTTTCGTTTATGAACTTTCGTTT
EmoSat24-101	ACCATACAATATGAATACTGACAGAGC	ACGAGGTAGGTAGGCAACATG
EmoSat25-184	AGTGGTTTGCAGCAGGTGGTT	TGGTTGTAGTTGTAGATAGCCCTTGGT
EmoSat26-41	ATGAATGAATGACTGCCTC	TCATTCATTCATTCATTCATTC
EmoSat27-102	TCATCAGCGTGCTGCAACTC	ATGCAGTAGTGGCTCAATCGG

Table 3.S3: Characteristics, divergence and abundance for each subfamily.

				Divergence (%)		Abundance (%)		log2	
SF	SatDNA subfamily	length	G+C	0B	+B	0B	+B	(+B/0B)	+B-0B
	EmoSat01A-325	325	37.5	5.76	6.01	0.5262	0.5496	0.0627	0.0234
1	EmoSat02A-89	89	25.8	6.08	6.1	0.1049	0.0949	-0.1440	-0.0100
1	EmoSat02B-84	84	27.4	6.91	7.1	0.0974	0.0862	-0.1754	-0.0112
1	EmoSat02C-77	77	29.9	4.59	4.98	0.0866	0.0825	-0.0699	-0.0041
	EmoSat03A-304	304	55.9	12.26	12.22	0.2413	0.2419	0.0032	0.0006
	EmoSat04A-60	60	16.7	12.13	12.24	0.1722	0.1631	-0.0781	-0.0091
2	EmoSat05A-16	16	56.3	17.39	17.43	0.1002	0.1060	0.0809	0.0058
1	EmoSat06A-87	87	26.4	11.41	11.34	0.0313	0.0274	-0.1948	-0.0039
1	EmoSat06B-85	85	28.2	9.74	9.92	0.0207	0.0180	-0.2018	-0.0027
1	EmoSat06C-83	83	28.9	7.97	7.99	0.0154	0.0137	-0.1664	-0.0017
	EmoSat07A-5	5	53.1	1.44	1.34	0.0598	0.0676	0.1780	0.0078
	EmoSat08A-41	41	41.5	11.45	11.48	0.0489	0.0483	-0.0170	-0.0006
	EmoSat09A-14	14	50.0	12.43	12.79	0.0205	0.0145	-0.5028	-0.0060
	EmoSat09B-14	14	57.1	12.79	13.16	0.0112	0.0085	-0.3939	-0.0027
	EmoSat09C-14	14	50.0	12.82	12.87	0.0107	0.0085	-0.3314	-0.0022
	EmoSat09D-14	14	57.1	13.4	13.61	0.0050	0.0040	-0.3301	-0.0010
	EmoSat10A-302	302	36.4	8.74	8.43	0.0345	0.0358	0.0542	0.0013
4	EmoSat11A-122	122	36.1	6.05	16.23	0.0308	0.0008	-5.3269	-0.0300
4	EmoSat11B-123	123	33.3	18.17	7.19	0.0001	0.0497	9.2101	0.0496
4	EmoSat12A-122	122	35.2	4.7	6.78	0.0201	0.0077	-1.3772	-0.0124
4	EmoSat12B-123	123	36.6	11.72	9.57	0.0062	0.0067	0.1023	0.0005
	EmoSat13A-24	24	41.7	13.65	13.59	0.0252	0.0260	0.0442	0.0008
	EmoSat14A-24	24	37.5	13.88	13.72	0.0144	0.0095	-0.5981	-0.0049
3	EmoSat15A-207	207	37.2	8.66	8.96	0.0124	0.0161	0.3820	0.0037
3	EmoSat15B-207	207	37.2	5.69	6.23	0.0018	0.0034	0.8955	0.0016
3	EmoSat16A-208	208	42.3	5.9	5.95	0.0140	0.0127	-0.1333	-0.0013
	EmoSat17A-97	97	33.3	5.87	6	0.0136	0.0145	0.0909	0.0009
2	EmoSat18A-7	7	57.1	16.17	16.29	0.0130	0.0153	0.2313	0.0023
3	EmoSat19A-203	203	38.9	8.63	7.89	0.0078	0.0087	0.1468	0.0009
3	EmoSat19B-203	203	37.9	9.68	10.26	0.0042	0.0019	-1.1749	-0.0023
3	EmoSat20A-199	199	39.7	4.6	4.74	0.0091	0.0123	0.4297	0.0032
3	EmoSat21A-208	208	38.9	6.25	6.3	0.0083	0.0089	0.0981	0.0006
	EmoSat22A-12	12	56.4	8.1	9.35	0.0073	0.0180	1.2995	0.0107
	EmoSat23A-14	14	28.6	10.78	10.74	0.0043	0.0050	0.2358	0.0007
	EmoSat23B-18	18	27.8	7.81	7.71	0.0016	0.0021	0.4096	0.0005
	EmoSat24A-101	101	34.7	13.6	13.26	0.0050	0.0050	0.0093	0.0000
	EmoSat25A-184	184	47.8	10.71	10.27	0.0047	0.0042	-0.1561	-0.0005
	EmoSat26A-41	41	41.5	17.37	7.88	0.0004	0.0219	5.6550	0.0215
	EmoSat26B-29	29	37.9	16.3	5.59	0.0004	0.0666	7.3242	0.0662
	EmoSat26C-17	17	47.1	16.36	6.82	0.0000	0.0098	7.8227	0.0098
	EmoSat27A-102	102	44.1	10.33	2.33	0.0003	0.0170	5.7266	0.0167
	Total					1.7919	1.9143		0.1224

Table 3.S4: Location of the satDNAs families grouped in superfamilies.

		Chromosome										
Superfamily	SatDNA family	L1	L2	L3	L4	M5	M6	M7	S8	S9	X	B
1	EmoSat02-89					p	p	p	p		p	
1	EmoSat06-87						p	p	p	p		
2	EmoSat05-16					p	p	p	p	p	p	s
2	EmoSat18-7					p	p	p	p	p	p	pi
3	EmoSat15-207	i						d	d	d		
3	EmoSat16-208					d						
3	EmoSat19-203								id	id		
3	EmoSat20-199							d				
3	EmoSat21-208				i		d	d	d	d		
4	EmoSat11-122								i			pid
4	EmoSat12-122								i			s

Capítulo 4. Contenido en ADN repetitivo de los cromosomas B de *Locusta migratoria*

Francisco J. Ruiz-Ruano, Josefa Cabrero, María Dolores López-León y
Juan Pedro M. Camacho

Departamento de Genética, Universidad de Granada

Resumen. Los cromosomas B son elementos dispensables en los genomas eucarióticos y, como tales, son un lugar ideal para la acumulación de ADN repetitivo. Aquí analizamos el contenido en ADN repetitivo de los cromosomas B de *Locusta migratoria*, mediante secuenciación masiva y análisis bioinformático, combinados con el mapeo físico de algunos elementos mediante hibridación in situ fluorescente. La comparación del contenido genómico en elementos de ADN repetitivo entre individuos con y sin cromosomas B ha revelado una composición de los cromosomas B muy diferente a la de los cromosomas A, ya que están muy enriquecidos en ADN satélite, representando uno de ellos (LmiSat02-176) más del 70 % del ADN repetitivo del B. También tienen genes para histonas, pero están muy empobrecidos en elementos transponibles. Estos últimos forman una quimera muy compleja, de la que hemos reconstruido más de 17 kb, y que incluye decenas de transposones diferentes, en su gran mayoría incompletos. A pesar de que estos elementos son minoritarios en el B, esta quimera de transposones sugiere la incorporación frecuente de transposones, unos sobre otros, en una misma región del B, convirtiendo a éste en un autentico sumidero de transposones. Por otra parte, el cromosoma B contiene 7 ADNs satélites visibles por FISH, y el único cromosoma A que contiene todos ellos es el cromosoma 9. Esto apunta a este cromosoma, además del cromosoma 8 (que es el único que tiene genes para histonas, como el cromosoma B), como los posibles cromosomas A antecesores del B, quizás como resultado de una translocación entre estos dos cromosomas.

Introducción

Durante muchos años se ha pensado que los cromosomas B sólo contienen ADN repetitivo, principalmente transposones, ADN satélite y algunas familias génicas, especialmente ADN ribosómico (para revisión, ver Camacho *et al.* 2000; Camacho 2005). Posteriormente se han descrito otras familias génicas en los cromosomas B, tales como los genes para las histonas (Teruel *et al.*, 2010; Oliveira *et al.*, 2011; Silva *et al.*, 2014; Utsunomia *et al.*, 2016), el ADNr 5S (Oliveira *et al.*, 2011; Kour *et al.*, 2013; Xie *et al.*, 2014; Jang *et al.*, 2015) y el ADN nuclear pequeño U2 (Bueno *et al.*, 2013; Menezes-de Carvalho *et al.*, 2015). Igualmente, se han encontrado abundantes ADNs satélites (Poletto *et al.*, 2010; Peng & Cheng, 2011; Klemme *et al.*, 2013; Bauerly *et al.*, 2014; Jang *et al.*, 2015; Utsunomia *et al.*, 2016) y transposones (Martis *et al.*, 2012; Klemme *et al.*, 2013; Houben *et al.*, 2014; Kour *et al.*, 2013; Valente *et al.*, 2014; Huang *et al.*, 2016). La abundancia de elementos repetidos ha dificultado encontrar genes para proteínas en los cromosomas B, pero la secuenciación masiva está demostrando que los genes para proteínas parecen ser un componente importante de los cromosomas B, tanto estructural- como funcionalmente (ver Capítulo 5).

Una de las aproximaciones más recientes para identificar elementos repetidos en los cromosomas B ha sido posible gracias a la introducción de las técnicas de secuenciación masiva, que han sido aplicadas en especies tales como el hongo *Alternaria arborescens* (Hu *et al.*, 2012), el centeno (Martis *et al.*, 2012; Klemme *et al.*, 2013), la mosca *Drosophila albomicans* (Zhou *et al.*, 2012), los peces *Astatotilapia latifasciata* (Valente *et al.*, 2014) y *Moenkhausia sanctaefilomenae* (Utsunomia *et al.*, 2016) y el saltamontes *Eumigus monticola* (ver Capítulo 3).

En este trabajo realizamos un análisis en profundidad del contenido en ADN repetitivo del cromosoma B de *Locusta migratoria*, mediante análisis de secuenciaciones genómicas a baja cobertura de varios individuos con y sin cromosomas B. Las comparaciones genómicas entre estas dos clases de individuos han permitido identificar y cuantificar los diferentes tipos de ADNs repetidos que son más abundantes en los cromosomas B. Hemos estudiado, por separado, tres componentes del repitoma: el clúster de los genes de histonas, el satelitoma y los elementos transponibles, analizando su abundancia, diversidad y organización en el cromosoma B, mediante análisis bioinformático, y mapeo físico mediante FISH.

Material y métodos

Material y extracción de ADN

Capturamos individuos en dos poblaciones de Los Barrios (Cádiz) y una de Padul (Granada). Todos los individuos de Padul eran portadores de cromosomas B, por lo que cruzamos varios de estos machos con hembras sin cromosomas B, obtenidas en una tienda de mascotas, y obtuvimos descendientes con y sin cromosomas B (que, en adelante, identificaremos como machos de Padul). Cuando los machos de la primera generación llegaron a adultos, fijamos un testículo en una solución de etanol-ácido acético 3:1, para los estudios citogenéticos que permiten determinar la presencia de cromosomas B, y congelamos el resto del cuerpo en nitrógeno líquido antes de almacenarlo en un congelador a -80ºC hasta su uso. Dos de estos machos (uno 0B y el otro con 2B) fueron seleccionados para secuenciación Illumina de ADN (a partir de una pata) y ARN (a partir del resto del cuerpo). Además, seleccionamos 6 machos de Cádiz, dos 0B y cuatro +B, para estudios genómicos y transcriptómicos similares. En estos individuos también extrajimos ADN de una pata saltadora, pero el ARN fue extraido, por separado, de un testículo y de la otra pata. El ADNg y el ARN de los dos machos de Padul fueron secuenciados mediante la plataforma Illumina HiSeq2000, obteniendo lecturas pareadas de 101 nucleótidos. En el caso de los seis individuos de Cádiz, sin embargo, utilizamos la plataforma Illumina HiSeq2500, que nos proporcionó lecturas pareadas de 125 nucleótidos. Además, hemos utilizado lecturas 454 de ADNg de un individuo con cromosoma B de Padul disponible en la base de datos SRA, con número de acceso SRR1200889, y de dos individuos de la especie *Eyprepocnemis plorans*, con números de acceso SRR1200829 y SRR1200835.

En total, secuenciamos 22 librerías mediante Illumina: dos de ADNg y dos de ARN a partir de los machos de Padul, y seis de ADNg más 12 de ARN de los machos de Cádiz. Y utilizamos 3 librerías 454 de ADNg provenientes de un individuo de *L. migratoria* +B de Padul y de dos individuos de *E. plorans*.

Obtención de las secuencias de referencia

Como secuencias de referencia, utilizamos las 107 secuencias de ADN satélite, pertenecientes a 62 familias (ver Capítulo 2), y las de los 1128 elementos transponibles incluidos en RepBase v20.10 (Bao *et al.*, 2015) (último acceso el 28 de Octubre de 2015). La secuencia del cluster de histonas no había sido caracterizada previamente en ningún ortóptero. Por tanto, para conseguir la secuencia en *L. migratoria*, buscamos genes para histonas mediante BLASTN (Altschul *et al.*, 1990) en el genoma ensamblado de *L. migratoria* (Wang *et al.*, 2014) con número de acceso AVCP000000000. Sin embargo, no localizamos ningún contig que contuviera todos los genes de histonas en un cluster, por lo que realizamos un clustering y ensamblaje de las lecturas 454 de ADNg de un individuo de *L. migratoria* con cromosoma B, utilizando el programa RepeatExplorer (Novák *et al.*, 2013). Además, utilizamos los resultados de este análisis para intentar localizar otras familias de secuencias que pudieran estar localizadas en los cromosomas B, ya que RepeatExplorer permite comparar la abundancia de cualquier cluster de ADN repetitivo entre librerías 0B y +B. Una vez obtenida la secuencia del cluster de histonas en *L. migratoria*, anotamos los diferentes genes haciendo una búsqueda con BLASTX (Altschul *et al.*, 1990) en la base de datos NR del NCBI. Adicionalmente, para caracterizar las regiones reguladoras conservadas, también ensamblamos el cluster de las histonas en *E. plorans* utilizando las lecturas 454 de ADNg, también mediante el programa RepeatExplorer (Novák *et al.*, 2013). Después alineamos el cluster de ambas especies con Geneious v4.8 (Drummond *et al.*, 2009) y buscamos regiones conservadas en los espaciadores. Anotamos las cajas TATA, la horquilla 3' y la región rica en purina. Finalmente, confirmamos el orden de los genes mediante PCR con cebadores diseñados para amplificar los espaciadores utilizando Primer3 (Untergasser *et al.*, 2012) (ver Fig. 4.S1).

Análisis de abundancia y divergencia

Para estimar la abundancia de los diferentes componentes del cluster de las histonas, del satelitoma y de los elementos transponibles, alineamos 10 millones de lecturas Illumina de cada una de las librerías de ADNg con las secuencias de referencia mediante RepeatMasker (Smit *et al.*, 2013). Además de la abundancia, el programa proporciona una estima de la divergencia. Esto nos permite representar ambos parámetros como un paisaje de repetitivo ("repeat landscape"). La comparación de abundancia y divergencia entre familias de elementos repetidos entre las librerias 0B y +B fue visualizada mediante paisajes sustractivos de ADN repetitivo, resultantes de restar los valores del paisaje 0B a los datos del paisaje +B. Estos paisajes muestran una representación global de la abundancia y divergencia de las secuencias sobre- y sub-representadas en los genomas con B, que podemos atribuir a la presencia de los cromosomas B. Finalmente obtuvimos una estima de la frecuencia de cada elemento de ADN repetitivo en el cromosoma B a partir de la diferencia entre la cobertura genómica promedio en los cuatro individuos con B y los dos sin B de Cádiz.

Búsqueda de variaciones en secuencia específicas del cromosoma B

Primero seleccionamos las secuencias repetitivas que eran claramente más abundantes en los genomas +B, en comparación con los 0B, y según las referencias en el caso de los elementos transponibles. En el caso del cistrón de las histonas, además de la secuencia consenso, buscamos contigs candidatos a ser alelos característicos del cromosoma B en los clusters de RepeatExplorer, con la ayuda del programa SeqGrapheR (Novák *et al.*, 2010).

Sobre estas secuencias buscamos dos tipos de variación específica del cromosoma B, tanto a nivel de secuencia mediante polimorfismos de sitio único (SNP, del inglés “Single Nucleotide Polymorphism”), como estructural mediante inserciones, deleciones o inversiones. Para ello, seleccionamos lecturas Illumina con homología con las secuencias de referencia mediante el algoritmo BLAT (Kent, 2002). Por un lado, mapeamos con SSAHA2 (Ning *et al.*, 2001) y seleccionamos SNPs con un nucleótido de referencia fijado en los individuos 0B y con un nucleótido alternativo exclusivo de los individuos +B. Estimamos la frecuencia de las variantes específicas de los B en ADNg y RNA, y calculamos la intensidad de expresión (IE) como el cociente, en cada individuo, entre su frecuencia en ARN y ADNg.

Búsqueda de variaciones estructurales específicas del cromosoma B

En primer lugar, visualizamos los mapeos anteriormente efectuados con SSAHA2 mediante IGV (Thorvaldsdóttir *et al.*, 2013) buscando variación de cobertura a lo largo de los contigs que pudiera indicar variación estructural. Para comparar la estructura de este tipo de regiones en una librería 0B y otra +B de Cádiz, seleccionamos 2500 parejas de lecturas homólogas, por librería y por región, y realizamos separadamente un clustering con RepeatExplorer.

Cuando las variaciones estructurales que eran características de los individuos con B, y por tanto del B, eran complejas, intentamos aumentar la secuencia conocida de los contigs, hacia ambos lados, buscando lecturas 454 de ADNg homólogas a regiones próximas a los extremos de entre 300 y 400 nucleótidos mediante el programa BLASTN (Altschul *et al.*, 1990). Posteriormente las ensamblamos con el software Geneious v4.8 (Drummond *et al.*, 2009) y repetimos el proceso varias veces hasta que ya no pudimos alargar más el contig por ambos extremos.

Amplificación por PCR y FISH

Los cebadores diseñados para este trabajo se muestran en la Tabla 4.S1. El programa de PCR utilizado para amplificar las regiones comienza con un paso de desnaturalización inicial a 95ºC durante 5 minutos, seguido de 30 ciclos a 94ºC durante 30 segundos, a 55-60-65ºC como temperaturas de hibridación durante 30 segundos y a 72ºC durante 30 segundos, terminando con un paso de extensión final a 72ºC durante de 7 minutos. Los productos de PCR utilizados para FISH fueron marcados mediante "nick translation" con 2,5 unidades de ADN polimerasa I/DNAsa I (Invitrogen), siguiendo el protocolo estándar. El mapeo físico de estas sondas se realizó mediante hibridación in situ fluorescente (FISH) siguiendo el protocolo descrito en Cabrero et al. (2003). Las sondas se marcaron con tetrametilrodamina-5-dUTP o fluoresceína-12-dUTP de Roche.

Resultados

Más de la mitad del genoma de *L. migratoria* es ADN repetitivo

El análisis de abundancia global de ADN repetitivo en el genoma de *L. migratoria* indica que éste supone aproximadamente el 55 % del genoma (Tabla 4.S2). Las principales diferencias entre las librerías +B y 0B, a nivel global, consisten en que las +B muestran sobrerrepresentación de genes para histonas y, sobre todo, de ADN satélite, así como una infrarrepresentación acusada de elementos transponibles (Fig. 4.1). Esto sugiere que el cromosoma B está enriquecido en histonas y ADN satélite, pero está empobrecido en elementos transponibles (Fig. 4.1).

Primera descripción del cluster completo de los genes para histonas en ortópteros

En el ensamblaje generado por RepeatExplorer usando las lecturas de ADNg de *L. migratoria*, encontramos un cluster que contenía genes para histonas. Su estructura circular sugería que podía incluir la unidad repetitiva completa (Fig. 4.2). Re-ensamblamos los contigs resultantes y observamos que el final era igual que el principio de las secuencias, confirmando así que habíamos obtenido la unidad completa. Anotamos los genes de histonas y encontramos que el orden y orientación del cluster de las histonas en *L. migratoria* es H1> <H3 H4> <H2A H2B> (Fig. 4.S1), así como en *E. plorans*. A continuación verificamos el orden de los genes mediante PCR con cebadores anclados en las regiones conservadas de los genes para amplificar los espaciadores. Todas las combinaciones dieron una banda del

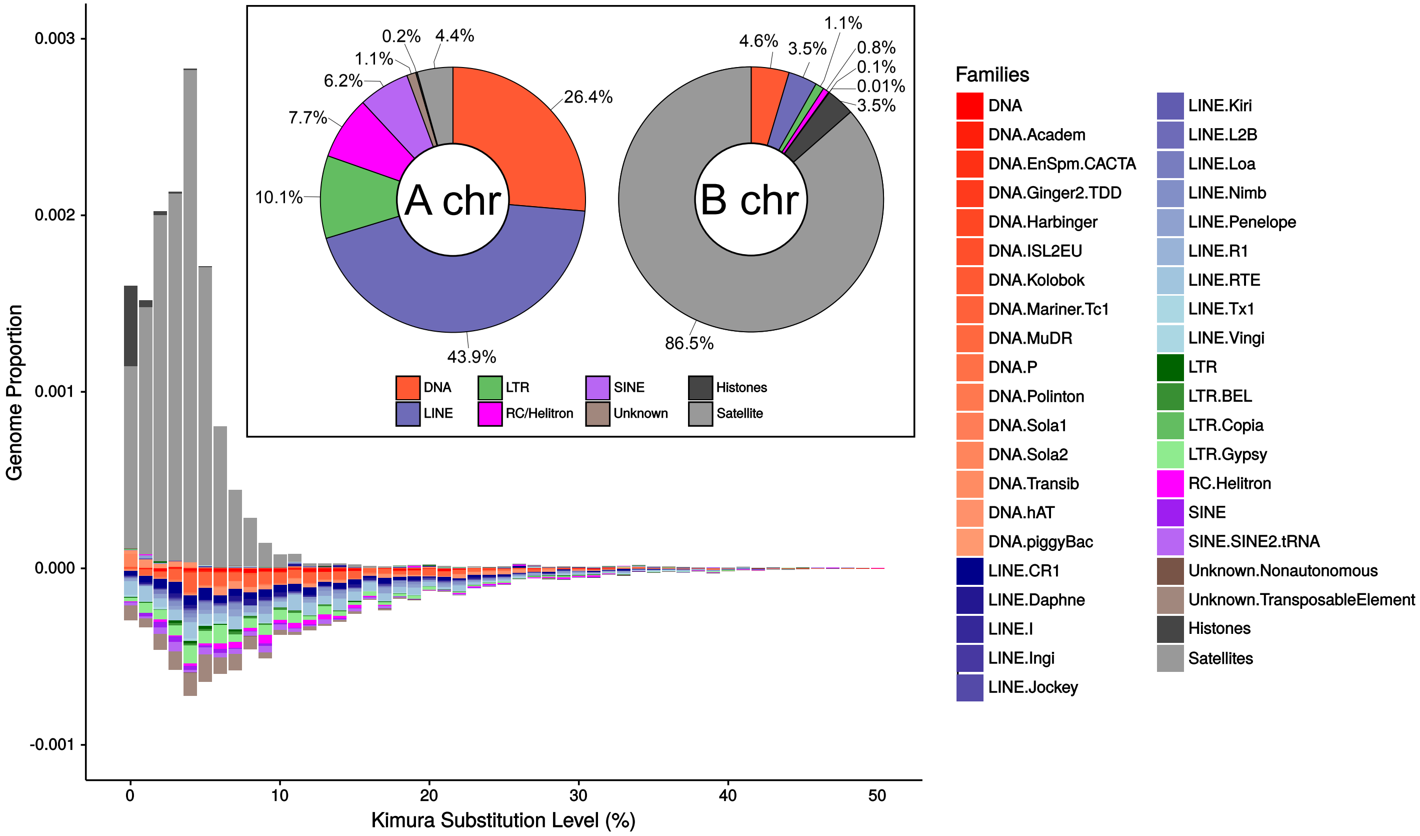


Figura 4.1: Paisaje de repetitivo resultante de restar los conteos promedio de las librerías de ADNg 0B a los de las librerías +B de Cádiz. Destaca la acumulación de ADN satélite e histonas, así como el empobrecimiento del B en elementos transponibles. En el recuadro representamos la frecuencia relativa de elementos en los cromosomas A, obtenida a partir del promedio en las librerías 0B, y la frecuencia relativa de elementos repetidos en el cromosoma B, obtenida como la diferencia entre el promedio +B y 0B.

tamaño esperado, confirmando el orden observado mediante RepeatExplorer. La longitud promedio del cluster de las histonas resultó ser 8.262 pb y 9.888 pb para *L. migratoria* y *E. plorans*, respectivamente (Tabla 4.S3). El alineamiento de la secuencia de los clusters de ambas especies nos permitió inferir la presencia de regiones conservadas en los espaciadores que corresponden a las cajas TATA, las horquillas 3′ y las regiones ricas en purina (Fig. 4.S1). Mediante FISH, mapeamos el gen H3 y el espaciador 1, y ambos mostraron señales de hibridación exclusivamente en el autosoma 8 y el cromosoma B, en coincidencia con lo observado previamente por Teruel *et al.* (2010) para los genes H3 y H4 (Fig. 4.S2). Esto demuestra que los cromosomas B contienen clusters de histonas incluyen tanto los genes como los espaciadores.

La variación en un espaciador de histonas permite diseñar un marcador específico de los individuos con cromosomas B

En el grafo generado por RepeatExplorer con las secuencias de los individuos con B, observamos varias regiones con bifurcaciones que probablemente correspondían a diferentes variantes en secuencia (Fig. 4.2a). Con el programa SeqGrapheR identificamos los contigs correspondientes a estas bifurcaciones y comparamos sus secuencias. Esto nos permitió localizar una region del espaciador 5 en la que dos contigs diferían para varias inserciones y deleciones (Fig. 4.2b). Para averiguar si uno de los contigs era específico de las histonas presentes en el B, diseñamos cebadores en la región conservada de forma que diera lugar a productos de diferente longitud en cada contig. La longitud esperada era de 320 pb para el contig 7 (con menor cobertura en +B) y de 232 pb para el contig 26 (con mayor cobertura). Amplificamos mediante PCR en ADNg de individuos portadores de cromosoma B, y otros sin él, y observamos la amplificación de la banda más grande en todos los individuos, tuvieran cromosomas B o no, pero el amplificado más pequeño sólo aparecía en los individuos portadores de cromosomas B (Fig. 4.2c). Esto demostró que el amplificado de menor tamaño es exclusivo de los cromosomas B, por lo que estos cebadores podrían ser utilizados como marcador molecular de la presencia de cromosomas B. Para probarlo, realizamos la amplificación por PCR con estos cebadores en 20 individuos adultos capturados en Cádiz, previamente estudiandos citogenénticamente para visualizar la presencia de cromosomas B. Los resultados demostraron que sólo los dos individuos sin cromosoma B mostraron el patrón de una banda mientras que los 18 restantes mostraron el patrón de dos bandas (resultados no mostrados).

Para intentar localizar variación en secuencia de ADN en los genes de histonas, que pudiera ser característica del cromosoma B, mapeamos lectu-

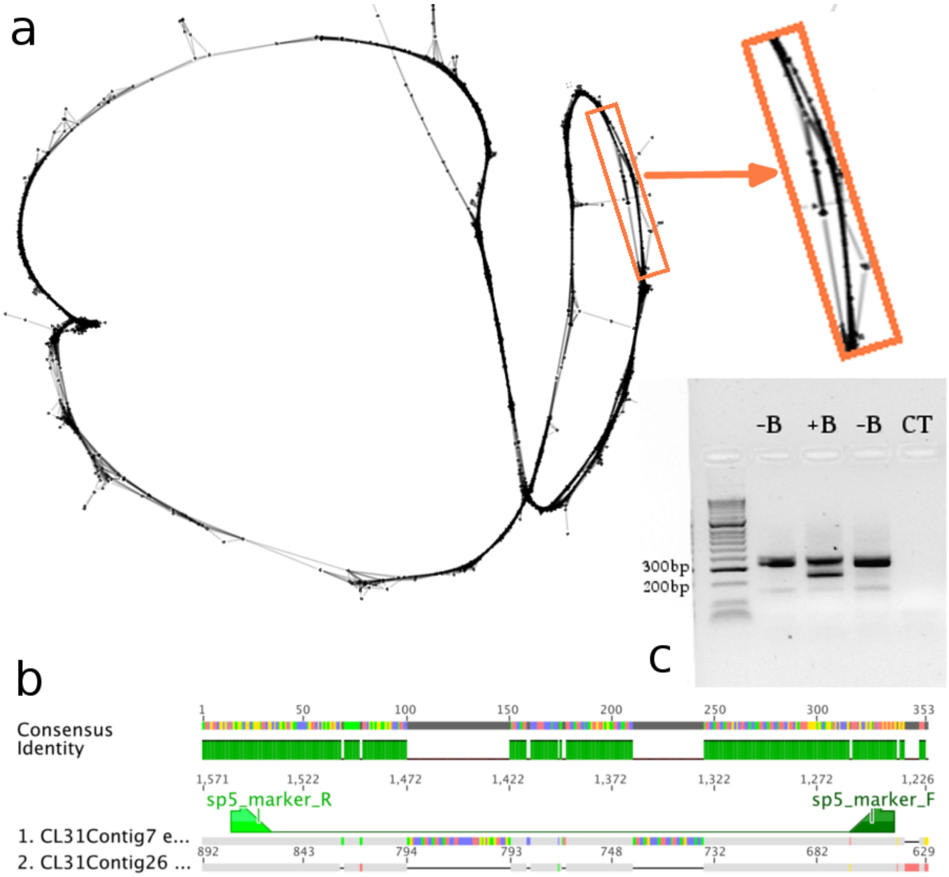


Figura 4.2: Generación de un marcador molecular del cromosoma B a partir del clúster de las histonas. a) Grafo de RepeatExplorer que contiene el cluster completo de las histonas. Su forma circular indica que el cistrón se encuentra repetido en tándem, y las bifurcaciones de líneas sugieren la existencia de varios alelos. A partir de la región ampliada del espaciador 5 hemos desarrollado un marcador molecular de la presencia del cromosoma B. b) Los dos contigs del espaciador 5 generados por RepeatExplorer presentan diferentes longitudes en la región acotada por los puntos de anclaje de los cebadores. c) Gel de agarosa al 1,5 % donde se muestra el patrón de amplificación de una banda para los individuos 0B y de dos bandas para los individuos +B.

ras Illumina de ADNg y ARN en individuos con y sin B, seleccionando las lecturas con BLAT y las mapeamos utilizando SSAHA2 contra la secuencia ensamblada del cluster de *L. migratoria*. No obstante, no encontramos ningún SNP en los genes que fuese característico de todos los individuos con B.

Una gran proporción del contenido del cromosoma B es un ADN satélite

El análisis comparativo de la abundancia de ADN satélite entre individuos +B y 0B sugiere la presencia de 37 familias de ADN satélite en el cromosoma B (Tabla 4.S4), siendo LmiSat02-176 el más abundante, con mucha diferencia, ya que representa el 85 % de todo el contenido en ADN satélite del cromosoma B, y éste representa el 87 % de todo su ADN repetitivo, por lo que LmiSat02-176 representa, aproximadamente, el 74 % del ADN repetitivo del B (Tabla 4.S5 y Fig. 4.S3). Los demás ADN satélites mostraron abundancias menores al 5 %. El mapeo físico mediante FISH confirmó que LmiSat02-176 es muy abundante en el cromosoma B, ya que la señal de FISH abarcaba todo el cromosoma B. Además, otros 6 ADN satélites mostraron clústers en el cromosoma B: LmiSat06-185 mostraba un cluster grande en la mitad distal del B, LmiSat07-5-tel, que es el ADN telomérico, mostraba clusters muy conspicuos en los telómeros del B, LmiSat53-143 forma un clúster en la región proximal del B, y LmiSat05-400 muestra un clúster muy pequeño cerca de la región distal del B (Fig. 4.3). Otros dos ADNs satélites (LmiSat01-193 y LmiSat04-18) mostraban señales de FISH muy débiles cerca de la región centromérica del B, y el primero mostraba además un leve punteado en la región heterocromática distal del cromosoma B (Fig. 4.3). El único cromosoma A que muestra clusters para estos 7 ADNs satélites es el autosoma 9 (ver Capítulo 2).

Una quimera de elementos transponibles en el cromosoma B

A diferencia de los ADNs satélites, existe por lo general una subrrepresentación de elementos transponibles en las librerías con B (Fig. 4.1). En conjunto, los elementos transponibles representan un 10 % del ADN repetitivo de los cromosomas B de *L. migratoria* (Tabla 4.S6 y Fig. 4.1), con mayor frecuencia de elementos de Clase II que de Clase I. Para probar este déficit de elementos transponibles en el B, escogimos dos elementos al azar (Daphne-8_LMi y Penelope-52_LM) y generamos sondas para FISH, que mostraron la escasa presencia de Daphne en todo el B, y de Penélope en la mitad eucromática del B (Fig. 4.S4).

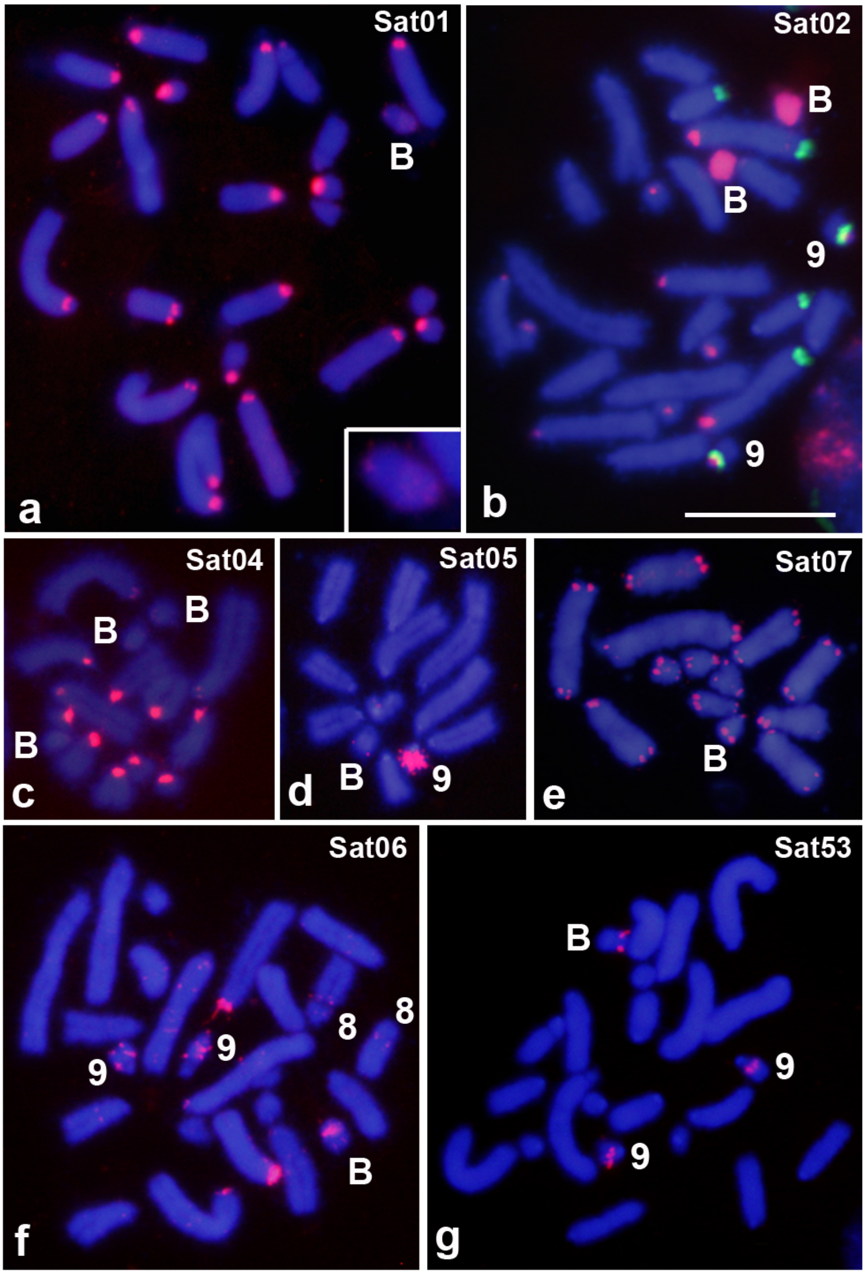


Figura 4.3: Hibridación *in situ* fluorescente (FISH) para los 7 ADN satélites visualizados en los cromosomas B. LmiSat01-193 muestra un pequeñísimo clúster en la región pericentromérica del cromosoma B, así como un leve punteado en la mitad heterocromática (distal) del mismo. LmiSat02-176 muestra señales de FISH a todo lo largo del cromosoma B. LmiSat04-18 y LmiSat05-400 muestran una señal muy débil en la región pericentromérica del B, LmiSat-07-5-tel muestra señales en ambos telómeros, LmiSat06-185 muestra un cluster grande en la mitad heterocromática del B, y LmiSat53-47 muestra un cluster en la región pericentromérica del B. Observe que los 7 ADN satélites se encuentran también en el cromosoma 9.

Por otro lado, realizamos conteos con RepeatMasker para cada secuencia, por separado, y contamos el número de nucleótidos alineados. Al disponerlos en orden decreciente, en función de la diferencia en abundancia entre +B y 0B, observamos que los transposones de DNA (clase II) son los elementos transponibles más abundantes en los cromosomas B, seguidos de varios tipos de LINEs, varios tipos de elementos LTR, Helitron y SINEs, en este orden (Tabla 4.S4, 4.S6 y Fig. 4.1).

Para conocer la estructura de estos elementos en los cromosomas A y B, primero realizamos un análisis bioinformático de cobertura que reveló que algunos elementos muestran ciertas regiones con mayor cobertura que otras, sugiriendo la posibilidad de amplificación diferencial entre regiones (Fig. 4.4, 4.S5-4.S11). Para averiguar si estas amplificaciones han dado lugar a regiones cromosómicas visibles mediante FISH, seleccionamos varias regiones de estos elementos y diseñamos cebadores para amplificarlas y generar sondas para FISH. Concretamente, seleccionamos los elementos DNA2-4_LMi, hAT-4_LMi, Kolobok-4_LMi, Helitron-N17B_LMi, Sola2-3N1_LMi, Penelope-42_LMi, Penelope-59_LMi y Tx1-1_LMi. El análisis mediante FISH reveló que la mayoría de los elementos analizados se localizan en una región intersticial del cromosoma B, concretamente en la zona heterocromática (DAPI$^+$) más próxima a la región eucromática del B (Fig. 4.5). En el caso de Penelope-42_LMi, Penelope-59_LMi y Tx1-1_LMi, amplificamos una región de alta cobertura y también otra región de baja cobertura. Mientras que en Penelope-42_LMi la región de baja cobertura también mapeó en la región intersticial del B, aunque con menor intensidad que la región de alta cobertura, en el caso de Penelope-59_LMi y Tx1-1_LMi la región de baja cobertura no mostró señal de FISH en el B, indicando que estas regiones están ausentes o poco representadas en el cromosoma B.

Con las regiones de alta cobertura de Penelope-59_LMi y Tx1-1_LMi realizamos un análisis adicional para intentar conocer mejor la estructura de la región intersticial del cromosoma B donde se acumulan tantos elementos transponibles diferentes. Para cada región de alta cobertura buscamos lecturas homólogas en las librerías 0B y +B, por separado, para realizar sobre ellas una clusterización y ensamblaje con RepeatExplorer. Para ambas regiones, los grafos en la librería 0B mostraban un patrón lineal, típico de los elementos transponibles, cuyo contig contenía las regiones del elemento transponible, flanqueantes a la región seleccionada (Fig. 4.4 y 4.S5). En el caso de la librería +B, el grafo mostraba dos estructuras lineales unidas por su región central (Fig. 4.4 y 4.S5). Una de ellas corresponde a la anteriormente encontrada en la librería 0B, mientras que la otra contiene la región de alta cobertura flanqueada por dos fragmentos de unos elementos transponibles distintos. Esta diferencia en los grafos entre librerías sugie-

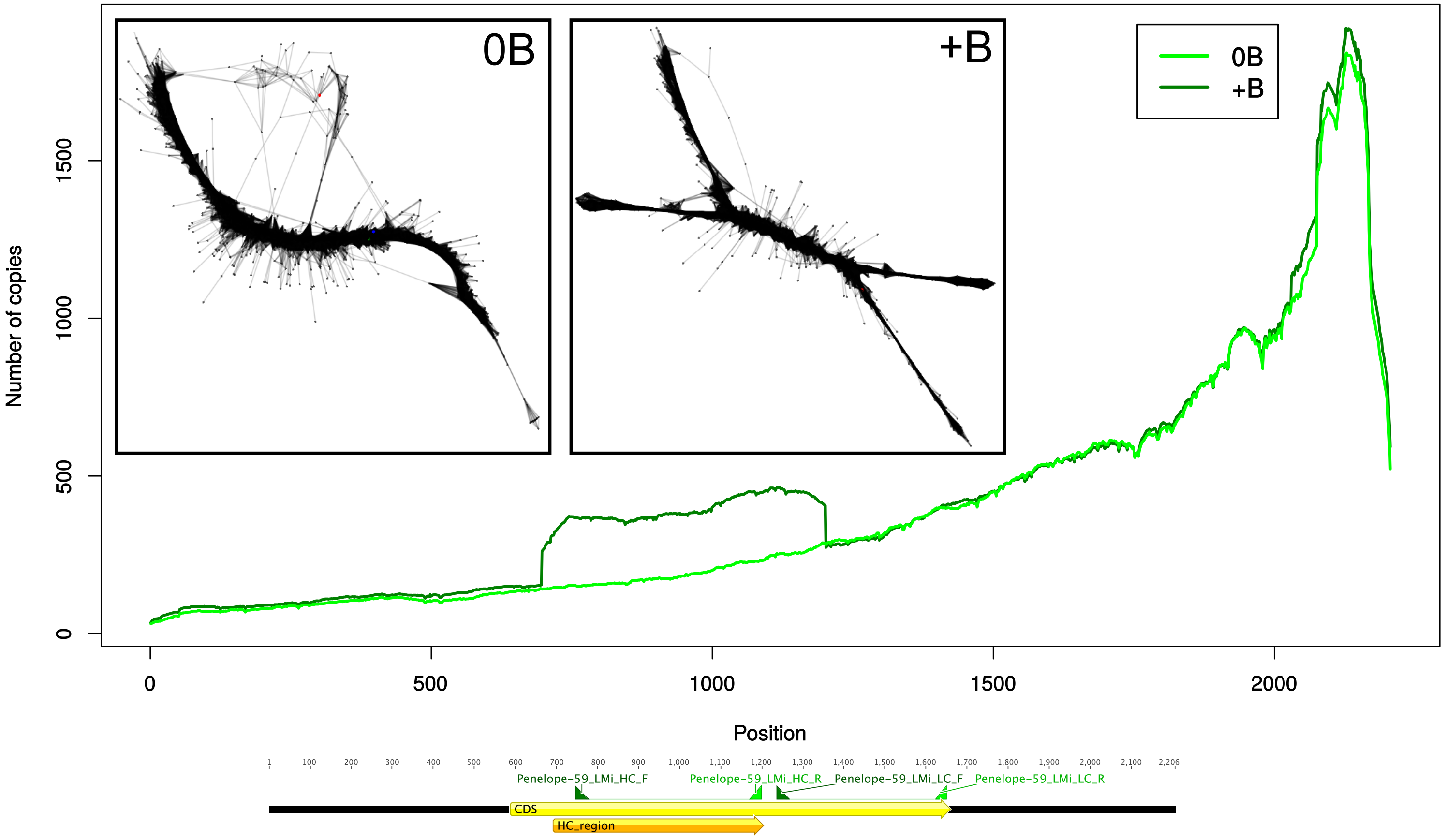


Figura 4.4: Cobertura promedio de las librerías de ADNg 0B y +B de Cádiz a lo largo del elemento Penelope-59_LMi. Destaca la presencia de una región de la CDS con una cobertura más elevada en las librerías +B con respecto a las 0B. En los recuadros se muestra la diferencia de un clusterizado con RepeatExplorer cuando seleccionamos lecturas Illumina de ADNg homólogas a esta región en un individuo 0B y otro +B de Cádiz. Destaca la estructura lineal del grafo en la librería 0B, mientras que en la librería +B la estructura es en forma de aspa con dos líneas, una del elemento transponible y otra de la quimera que comparte la región de alta cobertura en las librerías +B.

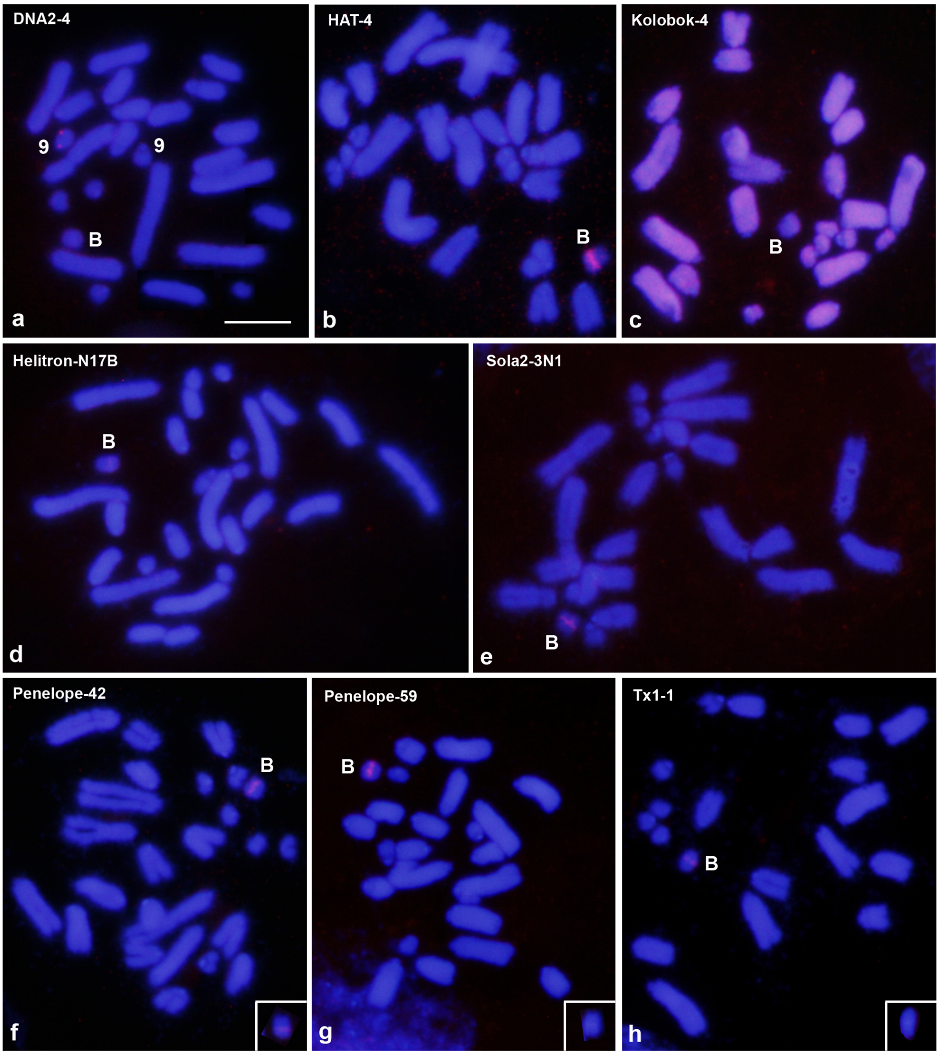


Figura 4.5: FISH para elementos transponibles sobrerrepresentados en el cromosoma B. Observe que DNA2-4_LMi forma un pequeño clúster en uno de los cromosomas S9, pero no el cromosoma B (a), hAT-4_LMi (b) se encuentra concentrado en una región intersticial donde también se encuentran Helitron-N17B_LMi (d), Sola2-3N1_LMi (e), Penelope-42_LMi (f), Penelope-59_LMi (g) y Tx1-1_LMi (h), mientras que Kolobok-4_LMi está disperso por todos los cromosomas A y en el B es más abundante en la mitad heterocromática (c). Los recuadros en f-h representan cromosomas B sometidos a FISH con las sondas de las regiones de baja cobertura de Penelope-42_LMi, Penelope-59_LMi y Tx1-1_LMi. Observe que en Penelope-42_LMi la región de baja cobertura también mapeó en la región intersticial del B, con menor intensidad que la región de alta cobertura, mientras que en Penelope-59_LMi y Tx1-1_LMi la región de baja cobertura no mostró señal de FISH en el B, indicando que estas regiones están ausentes o poco representadas en el cromosoma B.

re que estas secuencias quiméricas son exclusivas de los individuos con cromosoma B. Para extender la secuencia conocida de estos dos contigs característicos del cromosoma B, continuamos el proceso descrito anteriormente, de forma reiterativa, utilizando las lecturas de 454, y conseguimos llegar a ensamblar un contig de 17.327 nucleótidos que contenía fragmentos de 29 elementos transponibles, en distintas orientaciones (Fig. 4.6).

Algunos elementos transponibles del cromosoma B están activos

La búsqueda de SNPs en las librerías de ADNg y ARN de los individuos con B y sin B mostró la existencia de una serie de SNPs con una variante de referencia (Ref), fijada en los individuos 0B, y una variante alternativa (Alt) exclusiva de los individuos con B. Primero calculamos la frecuencia que representaba la variante Alt en los genomas y transcriptomas de cada individuo, y luego calculamos la intensidad de expresión (IE) de cada variante Alt como el cociente entre su frecuencia en transcriptoma y genoma del mismo individuo. El resultado más sobresaliente fue que los cuatro SNPs encontrados en el elemento Sola1-3 indicaban la sobreexpresión de este elemento en los testículos de todos los machos con B de Cádiz, y con algo menos de intensidad en los transcriptomas de pata, estando también sobreexpresado en el individuo con B de Padul (Figura 4.7 y Tabla 4.S7). Con menor intensidad, también observamos actividad transcripcional para el elemento Tx1-1, pero sólo en testículo de los machos de Cádiz, no encontrando actividad ni en pata de los machos de Cádiz ni en el macho de Padul.

Discusión

Nuestros resultados sugieren que los cromosomas B de *L. migratoria* están compuestos mayoritariamente por ADN repetitivo, en consistencia con lo observado en otras especies. No obstante, a pesar de estar compuestos de los mismos elementos que los cromosomas A, la proporción de los distintos tipos es muy diferente. La diferencia más notable se observa para el ADN satélite, ya que en los cromosomas A (es decir, en un genoma 0B) las 62 familias de ADN satélite representan sólo el 2,39 % del genoma en el linaje sur de la especie (ver capítulo 2), mientras que el ADN satélite representa más del 80 % del ADN repetitivo del B, y sólo uno de ellos, LmiSat02-176, supone más del 70 % del repetitivo del B. De hecho, en el promedio de los 4 individuos con B de Cádiz observamos 592.738 repeticiones de este ADN satélite mientras que los 2 individuos 0B tenían sólo 177.411, por lo que estimamos que unas 415.326 repeticiones estaban en los cromosomas B, es

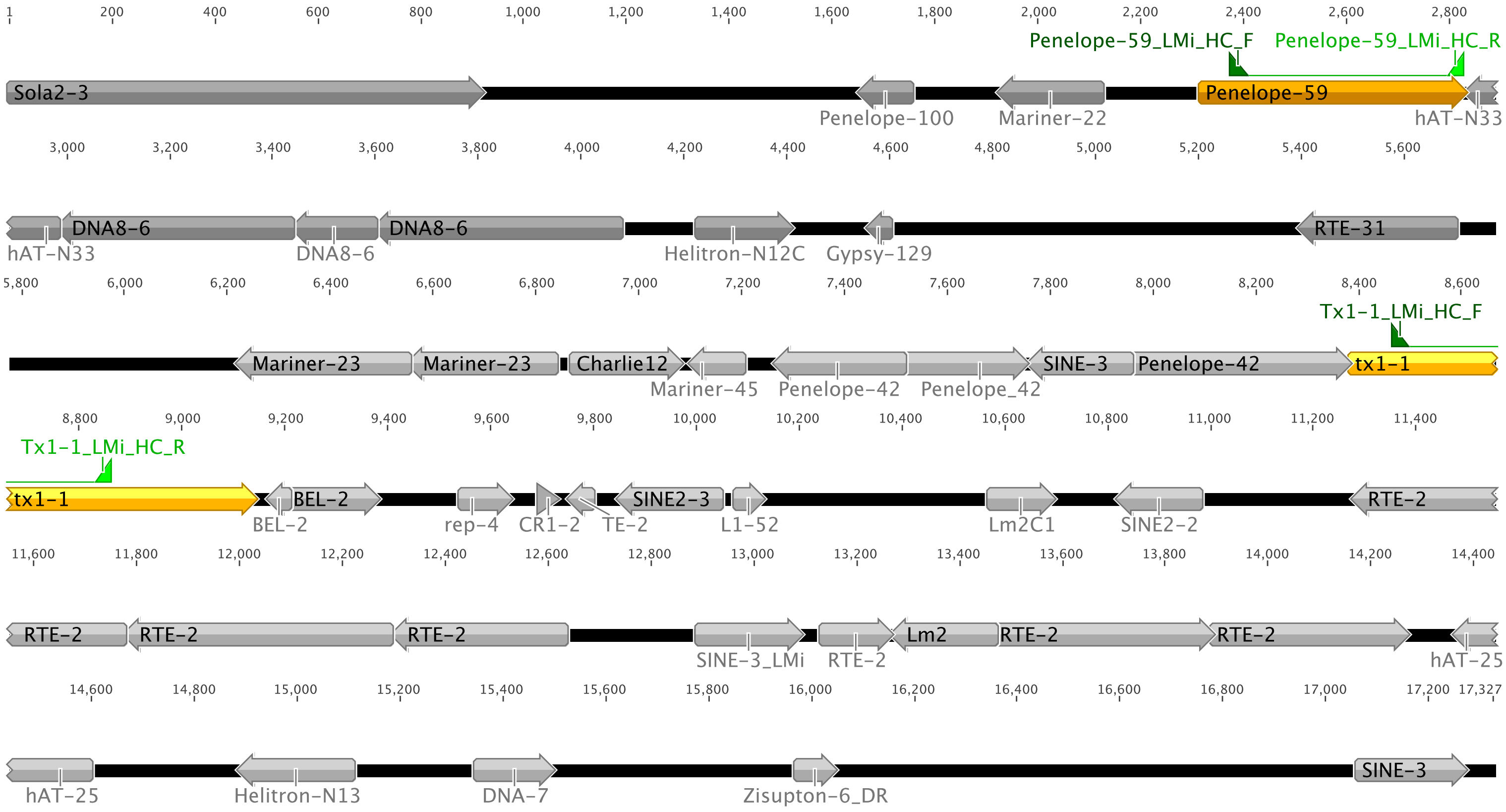


Figura 4.6: Estructura de la quimera extendida a partir de las regiones con más cobertura de los elementos Penelope-59_LMi y Tx1-1_LMi, utilizando lecturas 454 obtenidas a partir ADN genómico de un individuo +B . Consta de 17.327 pb con regiones homólogas a 29 elementos transponibles pertenecientes a 18 familias diferentes.

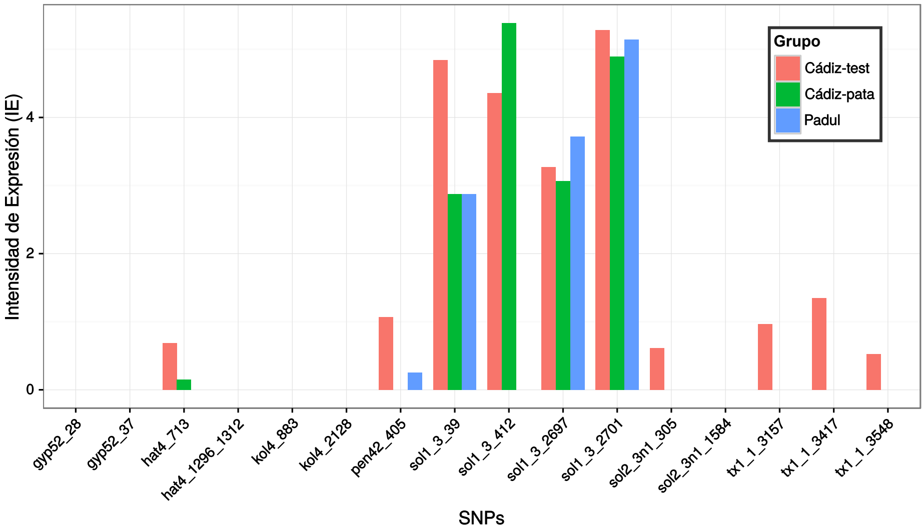


Figura 4.7: Intensidad de expresión (IE) para SNPs específicos del cromosoma B.

decir, éstos contenían más del doble de repeticiones que todos los cromosomas A en conjunto. La enorme amplificación de este ADN satélite en el cromosoma B ha podido ser facilitada por la elevada frecuencia con que los cromosomas B de *L. migratoria* forman bivalentes y quiasmas durante la diplotene meiótica, la mayoría de los cuales persisten hasta metafase I (Cabrero *et al.*, 1984), posiblemente mediante sobrecruzamiento desigual.

La enorme prevalencia del ADN satélite LmiSat02-176 en el cromosoma B de *L. migratoria*, y su presencia en el cromosoma X, indican que este ADN satélite es el principal responsable de los patrones de "chromosome painting" descritos por Teruel *et al.* (2009), ya que las sondas obtenidas por estos autores, mediante microdisección de los cromosomas X y B, dieron esencialmente el mismo patrón, marcando el cromosoma B en toda su longitud, además de las regiones pericentroméricas de algunos cromosomas A, incluyendo el cromosoma X. Este patrón coincide con el mostrado por LmiSat02-176 en los cromosomas A (ver capítulo 2) y en los cromosomas B (ver Fig. 4.3), y esto indica que aquellas sondas de "chromosome painting" estaban muy enriquecidas en LmiSat02-176.

Describimos aquí, por primera vez, la secuencia completa del clúster de los genes para histonas en un ortóptero. El orden y orientación de los genes en *L. migratoria* coinciden con los descritos previamente en el mosquito *Chironomus thummi* (Hankeln & Schmidt, 1991) y la mosca *Drosophila hydei* (Kremer & Hennig, 1990), pero con la histona H1 en orientación contraria en este último caso. Nuestro interés en esta familia génica se debe a su presencia en el cromosoma B de *L. migratoria*, demostrada anteriormente por Teruel *et al.* (2010). Nuestros resultados confirman la presencia de genes para histonas en el cromosoma B, representando el 3,5 % del ADN repetitivo del B. Nuestros análisis bioinformáticos indican que el clúster de las histonas está completo en el cromosoma B, y que existen casi 400 copias en el B mientras que hay casi 800 en los cromosomas A. De hecho, la señal de FISH observada por Teruel *et al.* (2010) en el B era más pequeña que la del cromosoma 8, que es el único cromosoma A que lleva estos genes, y nuestro presente análisis de FISH indica que tanto para los genes como para los espaciadores, la señal en los cromosomas A es más grande que la de los cromosomas B (ver Fig. 4.S2). Por tanto, los cromosomas B contienen clusters completos para histonas, que incluyen tanto los genes como los espaciadores, aunque en número algo menor que los del cromosoma 8.

El bajo contenido de los cromosomas B de *L. migratoria* en elementos transponibles contrasta con la idea general de que estos cromosomas, al ser dispensables, constituyen un paraíso donde los elementos transponibles pueden proliferar sin afectar a la eficacia biológica de los individuos portadores (para revisiones, ver Camacho *et al.*, 2000; Camacho, 2005. Sin embargo, la escasa abundancia de TEs coincide con la pobre presencia de estos elementos observada mediante FISH en los cromosomas B del saltamontes *Eyprepocnemis plorans* (Montiel *et al.*, 2012), y es confirmada por nuestros análisis de FISH para elementos como Daphne y Penelope en *L. migratoria* (ver Fig. 4.S4). Como muestran nuestros análisis bioinformáticos sobre el contenido de ADN repetitivo del cromosoma B de *L. migratoria*, los elementos transponibles son, en realidad, un componente muy minoritario del ADN repetitivo del B, que no es representativo de su papel mayoritario en los cromosomas A. La razón de este desajuste, teniendo en cuenta que los cromosomas B muy probablemente derivaron de los A (Teruel *et al.*, 2010), estriba en su elevado contenido en ADN satélite. Como hemos comentado antes, LmiSat02-176 representa, por sí solo, más del 70 % de todo el ADN repetitivo del B, y está distribuido por todo él, tanto en la mitad proximal eucromática como en la mitad distal heterocromática.

A pesar de la baja proporción de elementos transponibles en el cromosoma B de *L. migratoria*, es interesante reseñar la acumulación de una enorme variedad de elementos diferentes en una región específica del B, concretamente en la parte de la región heterocromática que es adyacente a la región eucromática del B. Esta banda intersticial es aparente para los transposones hAT-4_LMi, Helitron-N17B_LMi, Sola2-3N1_LMi, Penelope-42_LMi, Penelope-59_LMi y Tx1-1_LMi. El análisis bioinformático nos ha permitido reconstruir una quimera de transposones que es exclusiva de los individuos portadores de cromosomas B y que podría estar localizada en esta región intersticial del B, dado que la quimera incluye 4 de estos elementos visualizados por FISH en esa región (Sola2-3N1_LMi, Penelope-

42_LMi, Penelope-59_LMi y Tx1-1_LMi). La gran mayoría de los elementos transponibles incluidos en la quimera estan incompletos, y unos interrumpen a otros, sugiriendo que ésta es una región del B que resulta dispensable para él y permite la acumulación de transposones sin afectar a la viabilidad del cromosoma B. La quimera reconstruida comprendía 17.327 pb incluyendo secuencias que muestran homología con 29 elementos transponibles diferentes pertenecientes a 18 familias diferentes (Tabla 4.S8 y Fig. 4.6), sólo uno de los cuales (Lm2) estaba completo. Este maremágnum de transposones incompletos sugiere un escenario de incorporación frecuente de transposones en ciertas regiones del B, donde unos transposones se insertan sobre otros preexistentes, inactivándose sucesivamente unos a otros. Esto convierte al cromosoma B en un verdadero sumidero evolutivo de transposones, tal como se ha propuesto para los retrotransposones R2 en el ADN ribosómico del cromosoma B de *E. plorans* (Montiel *et al.*, 2014).

En próximas investigaciones analizaremos si esta quimera de transposones está realmente localizada en el cromosoma B y si es exclusiva de los individuos con B, por lo que podría servir para diseñar marcadores moleculares específicos de la presencia del cromosoma B. Para ello realizaremos experimentos de PCR con cebadores diseñados para probar el orden inferido bioinformáticamente, y diseñaremos sondas FISH para visualizar las regiones del B donde se encuentra esta quimera.

El contenido en ADN repetitivo de los cromosomas B de *L. migratoria*, desvelado en esta investigación, tiene importantes implicaciones sobre el origen del cromosoma B. La única hipótesis anterior al respecto fue propuesta por nosotros en base a la comparación de las secuencias de los genes para las histonas H3 y H4 obtenidas de los cromosomas B microdiseccionados y de individuos 0B, por tanto de los cromosomas A (Teruel *et al.*, 2010). La presencia exclusiva de estos genes en el cromosoma 8 y en el B apuntó la posibilidad de que el B derivó a partir de este cromosoma A, y la divergencia observada entre las secuencias de los genes para H3 y H4, obtenidas a partir de los cromosomas A y B, sugirió que el cromosoma B se originó hace más de 750.000 años (Teruel *et al.*, 2010). Nuestros resultados de la secuenciación masiva de genomas 0B y +B, así como de transcriptomas de ambos tipos, han permitido descubrir 24 genes presentes en el cromosoma B que codifican para proteínas, la mitad de los cuales están aparentemente completos (ver Capítulo 5). La divergencia observada en la secuencia de estos genes, al comparar los individuos con B y sin B con respecto a *Oedaleus decorus*, ha indicado que la edad del B está entre 1 y 4 millones de años. Esta elevada edad del B dificulta mucho identificar el cromosoma A del que se originó el B, porque durante un tiempo evolutivo tan largo el B puede haber experimentado muchos cambios en secuencia y estructura dotándole de características irreconocibles en los cromosomas A. Un posible ejemplo es la quimera de transposones, que parece ser exclusiva del B.

Nuestros análisis actuales sobre el cluster de las histonas son consistentes con el origen del B a partir del cromosoma 8, e indican que los cromosomas B contienen unas 400 copias del clúster completo para histonas, aproximadamente la mitad que un cromosoma 8. Esto indica que el B no adquirió todas las copias del 8 en su origen, o bien que el número de copias en el cromosoma 8 y en el B ha cambiado durante este millón de años. Por el contrario, los resultados sobre el contenido en ADN satélite de los cromosomas B indican la presencia masiva del ADN satélite LmiSat02-176 a lo largo de todo el B, representando más del 70 % del ADN repetitivo del B. Además, el cromosoma B parece contener otros 36 ADNs satélites en cantidades muy inferiores (<5 %), en una estructura genómica que recuerda mucho la del satelitoma de los cromosomas A (véase el capítulo 2), ya que sólo seis de ellos mostraban clústers visibles mediante FISH, es decir, de tamaños superiores a 1,5 kb (Schwarzacher *et al.*, 2000). Las pequeñas cadenas de ADN satélites que no superan este umbral podrían explicar por qué los ADN satélites que forman un cluster en una sola pareja cromosómica forman parte de cientos de contigs diferentes del genoma secuenciado por Wang *et al.* (2014), y fueron interpretadas por nosotros como el resultado de la diseminación de los ADN satélites entre cromosomas no homólogos (ver Capítulo 2). Sobre esta base, sería de esperar que cada cromosoma A particular (y también el cromosoma B) albergara decenas de ADN satélites invisibles mediante FISH.

La información que los ADNs satélites proporcionan sobre el origen del cromosoma B no es consistente con el origen exclusivo del B a partir del cromosoma 8, ya que apunta hacia la posibilidad de que el cromosoma 9 tuviese algo que ver en el origen del B, al ser el único cromosoma A que contiene clusterizados los mismos 7 ADNs satélites que hemos visualizado sobre el B. Aunque la fuerza de esta prueba queda un poco debilitada por la diseminación intragenómica de los ADN satélites, que posibilita la llegada ADN satéllites al B después de su origen, pensamos que, con la información disponible, lo más parsimonioso es proponer que tanto el cromosoma 8 como el 9 pudieron ser los cromosomas A que dieron origen al cromosoma B de *L. migratoria*. Una posibilidad es que el B resultara de una translocación recíproca entre los cromosomas 8 y 9. En el centeno, la secuenciación masiva ha demostrado que los cromosomas B se originaron a partir de dos cromosomas A (Martis *et al.*, 2012).

Aunque hemos acotado mucho el problema del origen del cromosoma B, sin embargo, ésta es realmente una cuestión difícil de resolver, debido a la enorme edad de este cromosoma B. Sin embargo, creemos que la respuesta final vendrá dada por el contenido del B en genes para proteínas (ver Capítulo 5), ya que éstos constituyen un bloque de genes que se espera que estén ligados en el cromosoma A de origen, un aspecto sobre el que enfocaremos nuestras próximas investigaciones.

Referencias

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology*, **215**, 403–410.

Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 1.

Bauerly E, Hughes SE, Vietti DR, Miller DE, McDowell W, Hawley RS (2014) Discovery of supernumerary B chromosomes in *Drosophila melanogaster*. *Genetics*, **196**, 1007–1016.

Bueno D, Palacios-Gimenez OM, Cabral-de Mello DC (2013) Chromosomal mapping of repetitive DNAs in the grasshopper *Abracris flavolineata* reveal possible ancestry of the B chromosome and H3 histone spreading. *PLoS One*, **8**, e66532.

Cabrero J, Viseras E, Camacho JPM (1984) The B-chromosomes of *Locusta migratoria* I. Detection of negative correlation between mean chiasma frequency and the rate of accumulation of the B's; a reanalysis of the available data about the transmission of these B-chromosomes. *Genetica*, **64**, 155–164.

Camacho JPM (2005) B chromosomes. *The Evolution of the Genome (ed. T. R. Gregory)*, pp. 223–286.

Camacho JPM, Sharbel TF, Beukeboom LW (2000) B-chromosome evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **355**, 163–178.

Menezes-de Carvalho NZ, Palacios-Gimenez OM, Milani D, Cabral-de Mello DC (2015) High similarity of U2 snDNA sequence between A and B chromosomes in the grasshopper *Abracris flavolineata*. *Molecular Genetics and Genomics*, **290**, 1787–1792.

Drummond AJ, Ashton B, Cheung M, *et al.* (2009) Geneious v. 4.8. 5 Biomatters Ltd. *Aukland, New Zealand*.

Hankeln T, Schmidt ER (1991) The organization, localization and nucleotide sequence of the histone genes of the midge *Chironomus thummi*. *Chromosoma*, **101**, 25–31.

Houben A, Banaei-Moghaddam AM, Klemme S, Timmis JN (2014) Evolution and biology of supernumerary B chromosomes. *Cellular and Molecular Life Sciences*, **71**, 467–478.

Hu J, Chen C, Peever T, Dang H, Lawrence C, Mitchell T (2012) Genomic characterization of the conditionally dispensable chromosome in *Alternaria arborescens* provides evidence for horizontal gene transfer. *BMC genomics*, **13**, 171.

Huang W, Du Y, Zhao X, Jin W (2016) B chromosome contains active genes and impacts the transcription of A chromosomes in maize (*Zea mays* L.). *BMC Plant Biology*, **16**, 1.

Jang TS, Parker JS, Weiss-Schneeweiss H (2015) Structural polymorphisms and distinct genomic composition suggest recurrent origin and ongoing evolution of B chromosomes in the *Prospero autumnale* complex (Hyacinthaceae). *New Phytologist*.

Kent WJ (2002) BLAT—the BLAST-like alignment tool. *Genome research*, **12**, 656–664.

Klemme S, Banaei-Moghaddam AM, Macas J, Wicker T, Novák P, Houben A (2013) High-copy sequences reveal distinct evolution of the rye B chromosome. *New Phytologist*, **199**, 550–558.

Kour G, Kaul S, Dhar MK (2013) Molecular characterization of repetitive DNA sequences from B chromosome in *Plantago lagopus* L. *Cytogenetic and genome research*, **142**, 121–128.

Kremer H, Hennig W (1990) Isolation and characterization of a *Drosophila hydei* histone DNA repeat unit. *Nucleic acids research*, **18**, 1573–1586.

Martis MM, Klemme S, Banaei-Moghaddam AM, *et al.* (2012) Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proceedings of the National Academy of Sciences*, **109**, 13343–13346.

Montiel EE, Cabrero J, Camacho JPM, López-León MD (2012) Gypsy, RTE and Mariner transposable elements populate *Eyprepocnemis plorans* genome. *Genetica*, **140**, 365–374.

Montiel EE, Cabrero J, Ruiz-Estévez M, *et al.* (2014) Preferential occupancy of R2 retroelements on the B chromosomes of the grasshopper *Eyprepocnemis plorans*. *PloS one*, **9**, e91820.

Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. *Genome research*, **11**, 1725–1729.

Novák P, Neumann P, Macas J (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC bioinformatics*, **11**, 1.

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792–793.

Oliveira NL, Cabral-de Mello DC, Rocha MF, Loreto V, Martins C, Moura RC (2011) Chromosomal mapping of rDNAs and H3 histone sequences in the grasshopper *Rhammatocerus brasiliensis* (Acrididae, Gomphocerinae): extensive chromosomal dispersion and co-localization of 5s rDNA/H3 histone clusters in the A complement and B chromosome. *Molecular cytogenetics*, **4**, 1.

Peng SF, Cheng YM (2011) Characterization of satellite CentC repeats from heterochromatic regions on the long arm of maize B-chromosome. *Chromosome research*, **19**, 183–191.

Poletto AB, Ferreira IA, Martins C (2010) The B chromosomes of the African cichlid fish *Haplochromis obliquidens* harbour 18s rRNA gene copies. *BMC genetics*, **11**, 1.

Schwarzacher T, Heslop-Harrison P, others (2000) *Practical* in situ *Hybridization.* BIOS Scientific Publishers Ltd.

Silva DMZdA, Pansonato-Alves JC, Utsunomia R, *et al.* (2014) Delimiting the Origin of a B Chromosome by FISH Mapping, Chromosome Painting and DNA Sequence Analysis in Astyanax paranae (Teleostei, Characiformes). *PLOS ONE*, **9**, e94896.

Smit AFA, Hubley R, Green P (2013) RepeatMasker Open-4.0.

Teruel M, Cabrero J, Montiel EE, Acosta MJ, Sánchez A, Camacho JPM (2009) Microdissection and chromosome painting of X and B chromosomes in *Locusta migratoria*. *Chromosome Research*, **17**, 11–18.

Teruel M, Cabrero J, Perfectti F, Camacho JPM (2010) B chromosome ancestry revealed by histone genes in the migratory locust. *Chromosoma*, **119**, 217–225.

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, **14**, 178–192.

Untergasser A, Cutcutache I, Koressaar T, *et al.* (2012) Primer3—new capabilities and interfaces. *Nucleic acids research*, **40**, e115–e115.

Utsunomia R, de Andrade Silva DMZ, Ruiz-Ruano FJ, *et al.* (2016) Uncovering the ancestry of B chromosomes in *Moenkhausia sanctaefilomenae* (Teleostei, Characidae). *PloS one*, **11**, e0150573.

Valente GT, Conte MA, Fantinatti BE, *et al.* (2014) Origin and evolution of B chromosomes in the cichlid fish *Astatotilapia latifasciata* based on integrated genomic analyses. *Molecular biology and evolution*, p. msu148.

Wang X, Fang X, Yang P, *et al.* (2014) The locust genome provides insight into swarm formation and long-distance flight. *Nature communications*, **5**.

Xie S, Marasek-Ciolakowska A, Ramanna MS, Arens P, Visser RG, van Tuyl JM (2014) Characterization of B chromosomes in *Lilium* hybrids through GISH and FISH. *Plant Systematics and Evolution*, **300**, 1771–1777.

Zhou Q, Zhu Hm, Huang Qf, *et al.* (2012) Deciphering neo-sex and B chromosome evolution by the draft genome of *Drosophila albomicans*. *BMC genomics*, **13**, 109.

Información suplementaria

Figuras Suplementarias

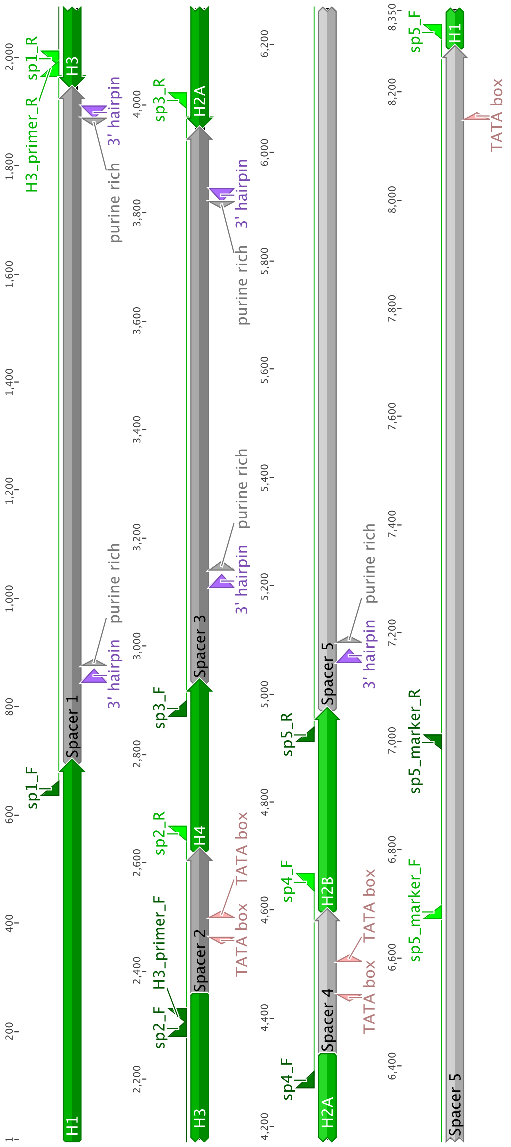


Figura 4.S1: Anotación del cistrón de las histonas de *L. migratoria*. Consta de 8.262 pb y contiene los cinco genes de histonas canónicas en el siguiente orden y orientación: H1> <H3 H4> <H2A H2B>. Incluimos también la localización de las cajas TATA, las horquillas 3′ y las regiones ricas en purina.

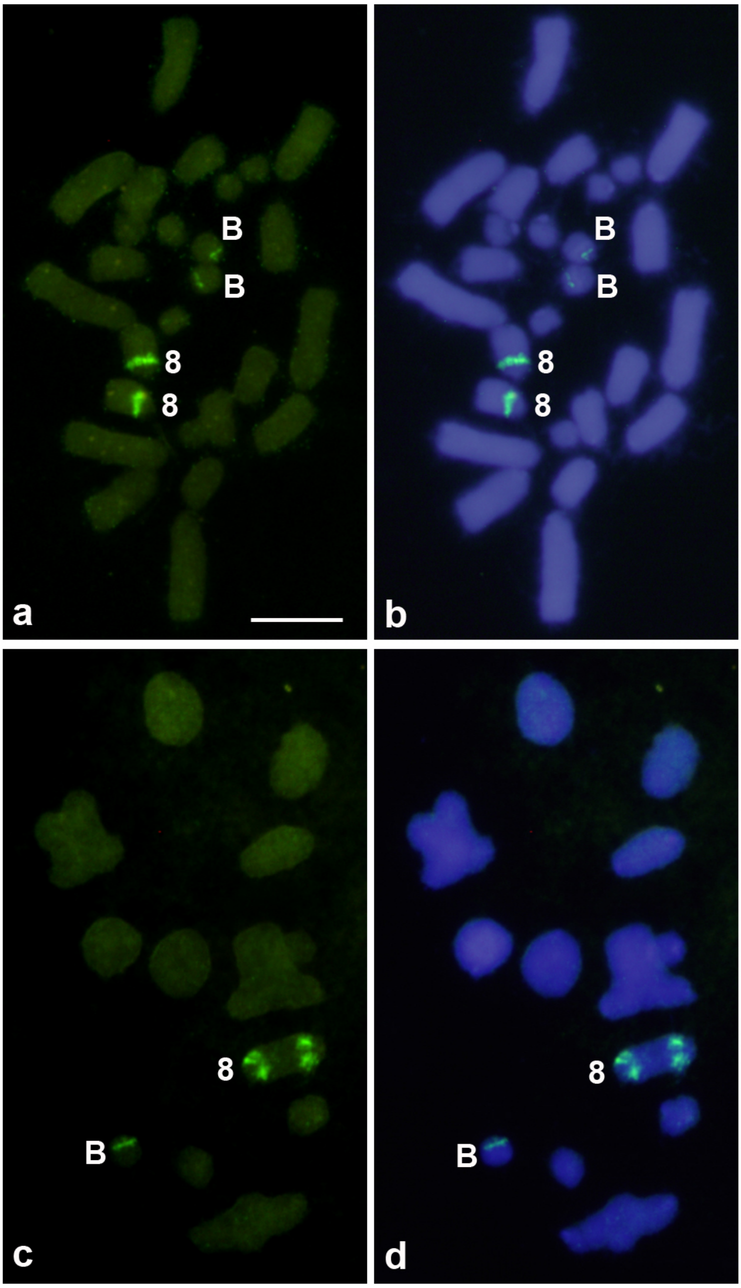


Figura 4.S2: Mapeo físico mediante FISH para el gen H3 (a y b) y el espaciador 1 (c y d) del cluster de genes para histonas, mediante cebadores diseñados sobre regiones conservadas en *L. migratoria* y *E. plorans*. Las mismas células sometidas a FISH (a y c) son mostradas, a la vez, con tinción DAPI (b y d).

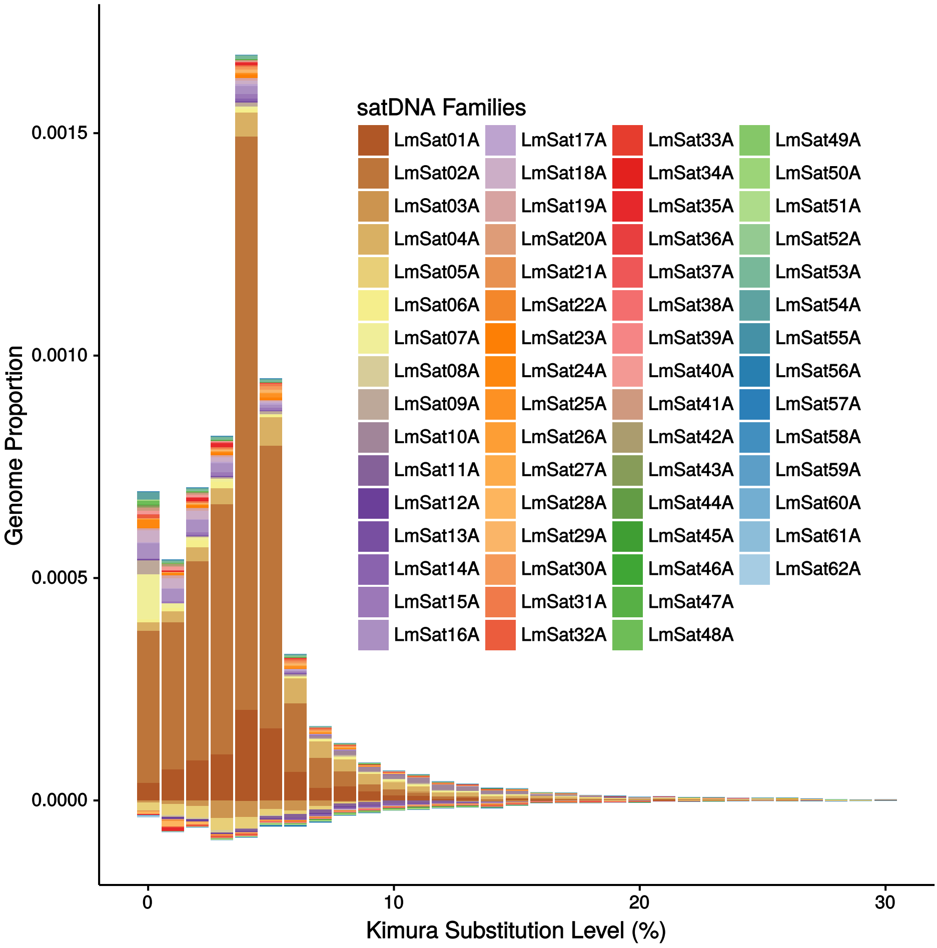


Figura 4.S3: Paisaje sustractivo de abundancia de ADN satélites en el cromosoma B, obtenido restando a la frecuencia promedio de las librerías +B de ADN genómico de Cádiz la frecuencia promedio de las librerías 0B. Se observa una clara sobrerrepresentación de ADN satélite en el B, sobre todo para el LmiSat02-176.

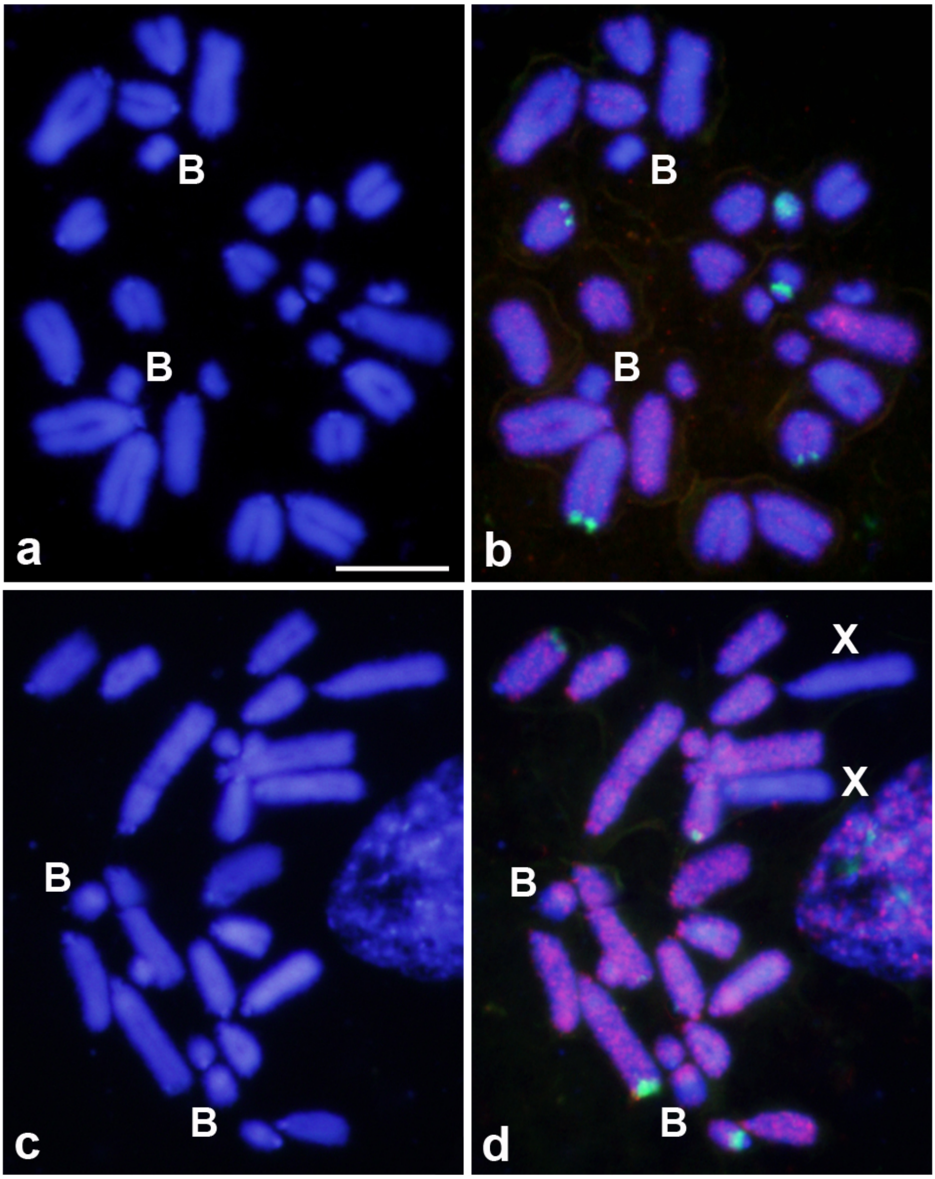


Figura 4.S4: Mapeo físico de dos elementos transponibles seleccionados al azar: Daphne-8_LMi (a y b) y Penelope-52_LMi (c y d). Observe que el primero es escaso en todo el cromosoma B, y el segundo es abundante en su mitad heterocromática.

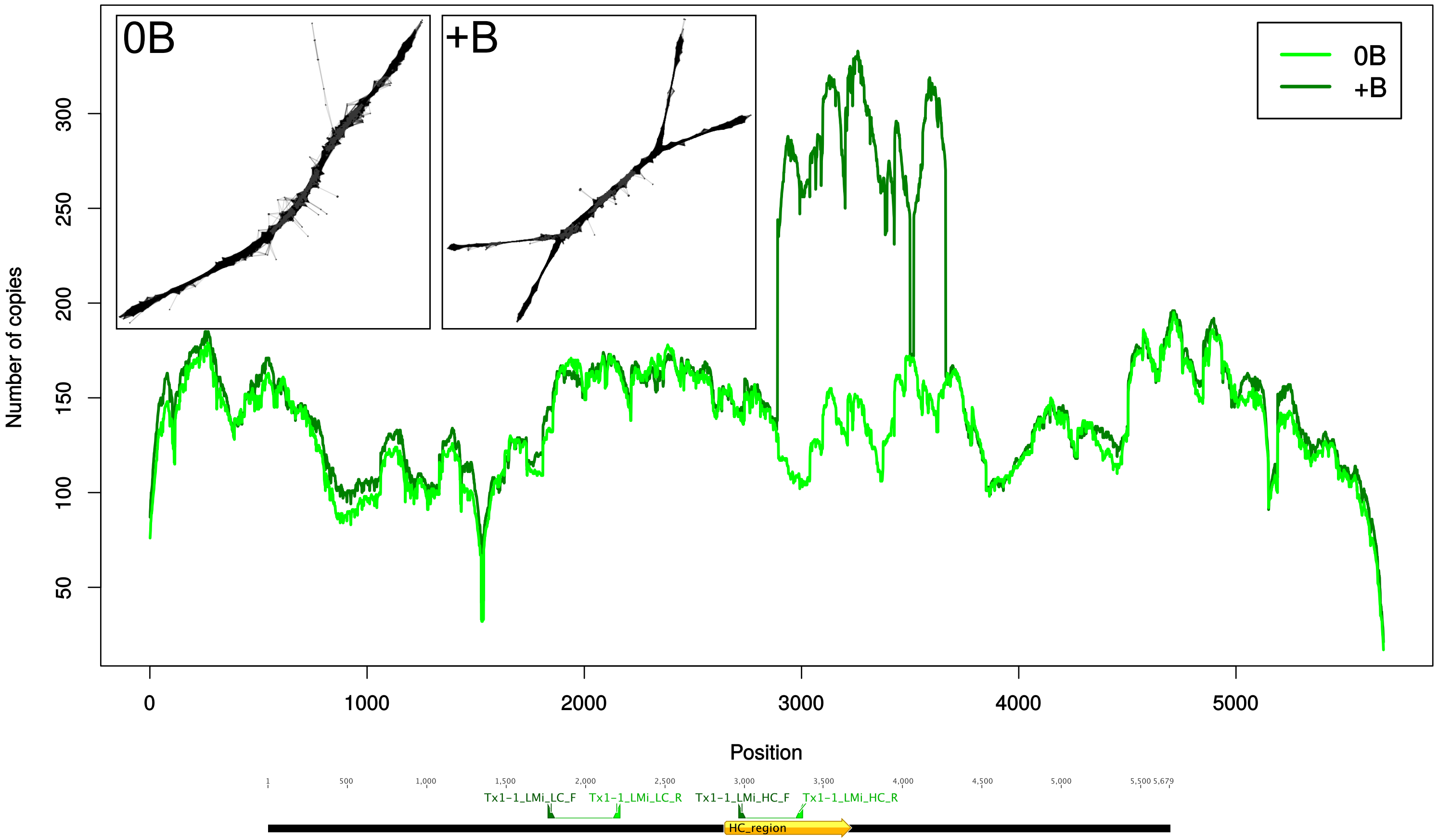


Figura 4.S5: Cobertura promedio de las librerías de ADN genómico 0B y +B de Cádiz, a lo largo del elemento Tx1-1_LMi. Destaca la presencia de una región con cobertura más elevada en las librerías +B con respecto a las 0B. Diseñamos dos parejas de cebadores anclados dentro y fuera de esta región para generar sondas de FISH, que mostraron la presencia de la región sobrerrepresentada en la región intersticial del cromosoma B, y la ausencia de la otra región (ver Figura 4.5h). Los recuadros muestran los grafos resultantes de clusterizar con RepeatExplorer las lecturas Illumina de ADN genómico que eran homólogas a la región de alta cobertura en un individuo 0B y otro +B de Cádiz. Destaca la estructura lineal del grafo en la librería 0B, y la forma de aspa en la librería +B, correspondiendo una de las líneas al elemento transponible y la otra a la quimera del cromosoma B.

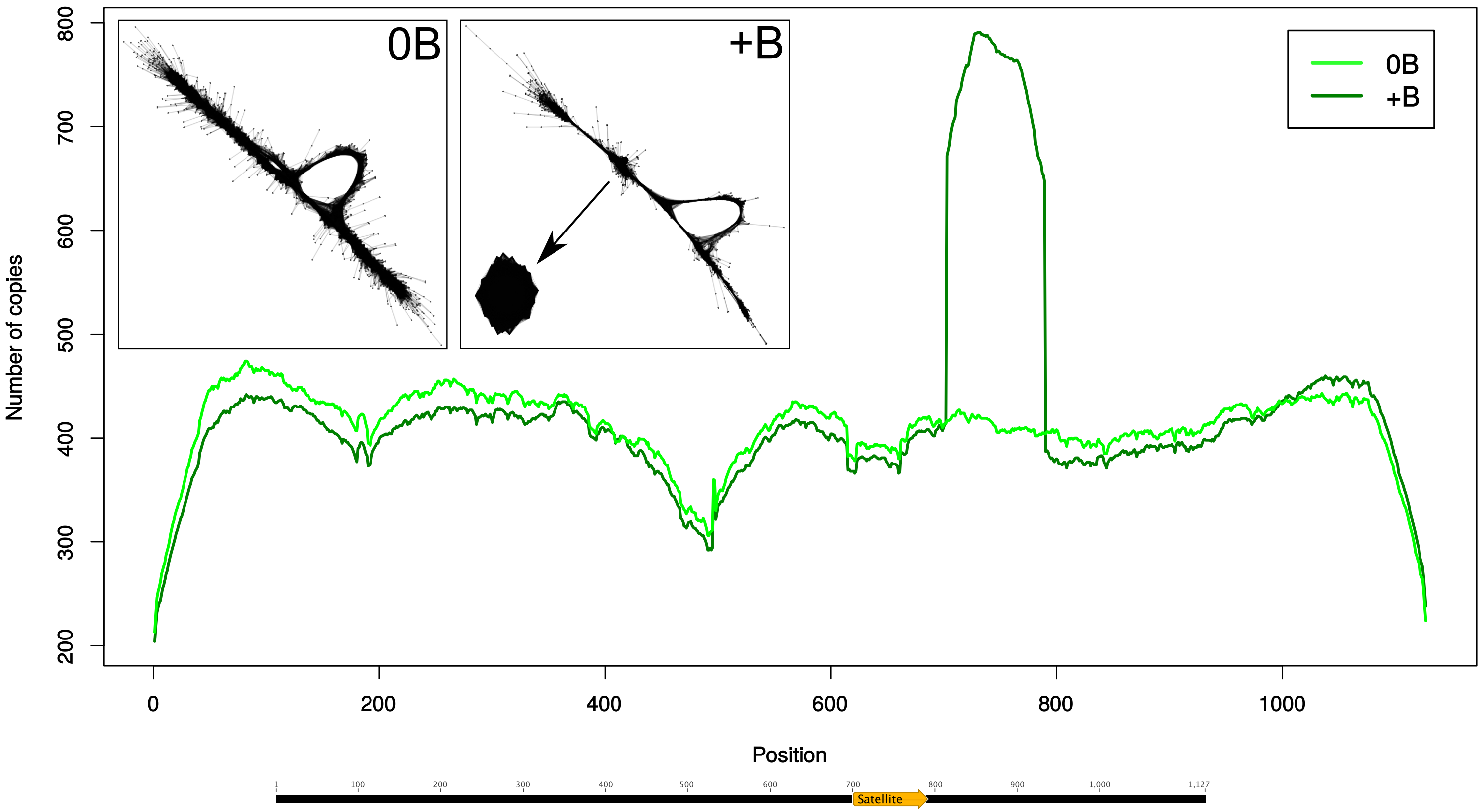


Figura 4.S6: Cobertura promedio de las librerías de ADN genómico 0B y +B de Cádiz, a lo largo del elemento DNA2-4_LMi. Note la presencia de una región con cobertura más elevada en las librerías +B que en las 0B. Los recuadros muestran los clusterizados realizados mediante RepeatExplorer con lecturas Illumina de ADN genómico homólogas a esta región en un individuo 0B y otro +B de Cádiz. Esta región de mayor cobertura se encuentra repetida en tándem.

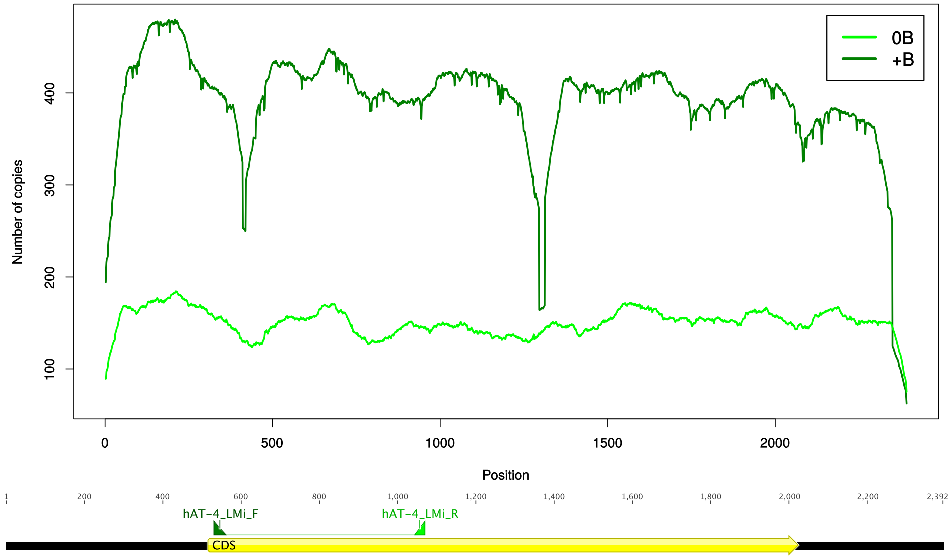


Figura 4.S7: Cobertura promedio de las librerías de ADN genómico 0B y +B de Cádiz, a lo largo del elemento hAT-4_LMi. Destaca la mayor cobertura en las librerías +B prácticamente a lo largo de todo elemento. Existen dos pequeñas regiones donde baja drásticamente la cobertura en las librerías +B que corresponden con deleciones.

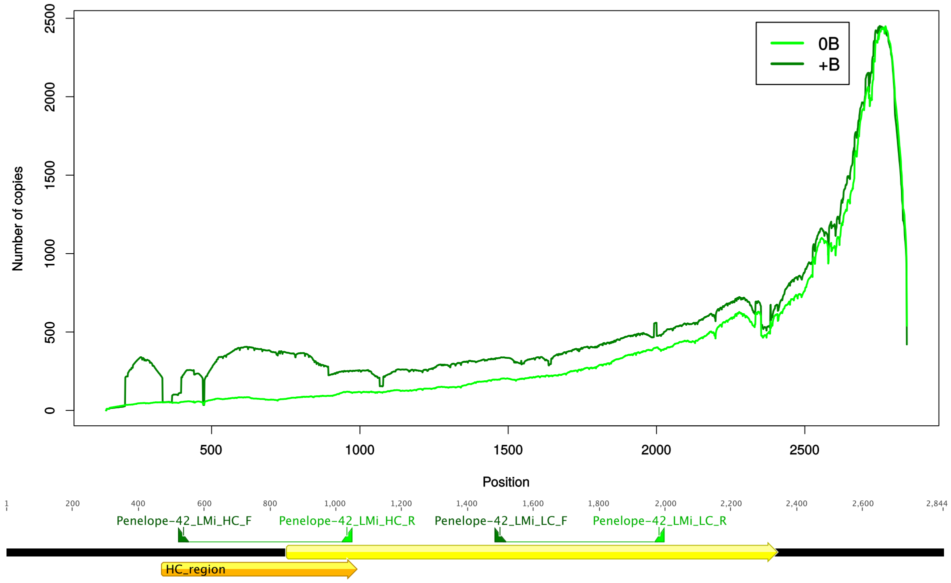


Figura 4.S8: Cobertura promedio de las librerías de ADN genómico 0B y +B de Cádiz, a lo largo del elemento Penelope-42_LMi. Destaca la mayor cobertura en las librerías +B prácticamente a lo largo de todo el elemento

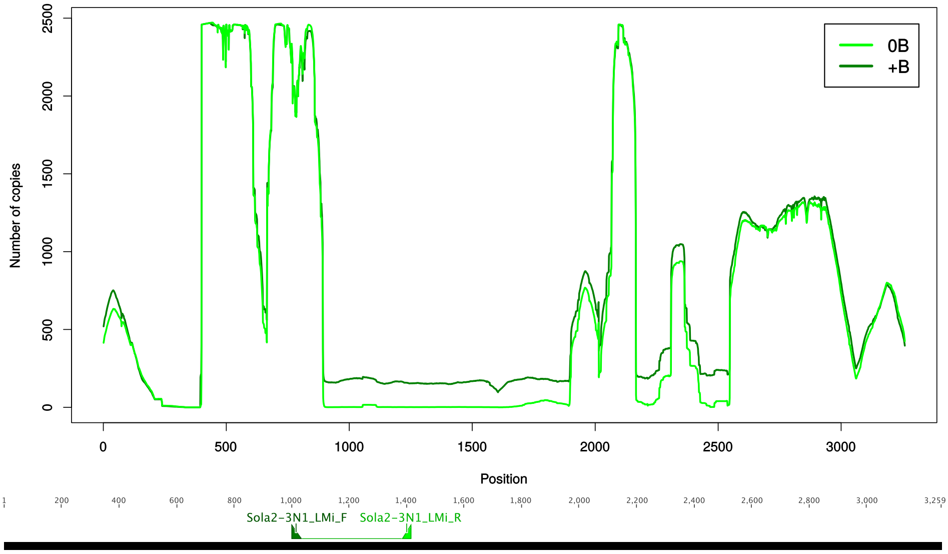


Figura 4.S9: Cobertura promedio de las librerías de ADN genómico 0B y +B de Cádiz, a lo largo del elemento Sola2-3N1_LMi. Destaca la mayor cobertura en las librerías +B entre principalmente entre las posiciones 900 y 1800, y además las posiciones 2200 y 2600.

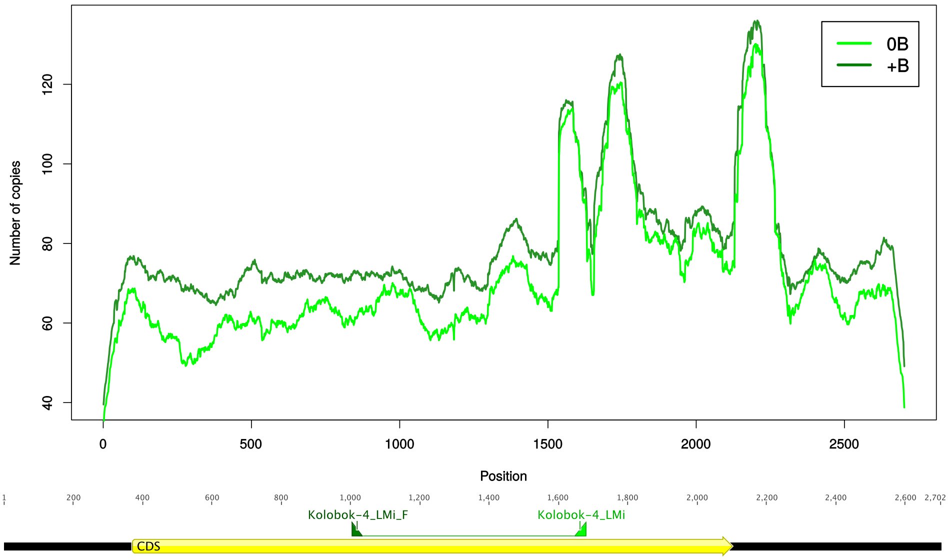


Figura 4.S10: Cobertura promedio de las librerías de ADN genómico 0B y +B de Cádiz, a lo largo del elemento Sola2-3N1_LMi. Destaca una mayor cobertura a lo largo de todo el elemento en las librerías +B.

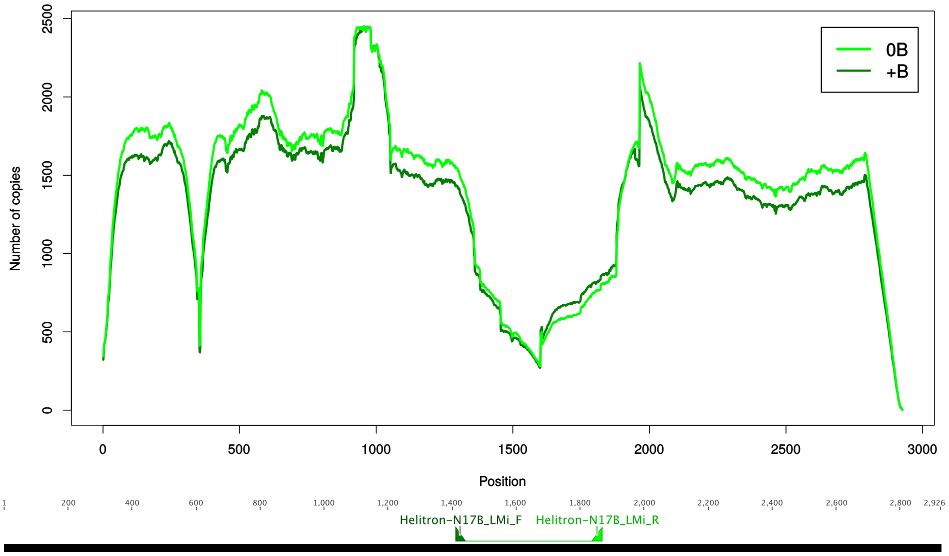


Figura 4.S11: Cobertura promedio de las librerías de ADN genómico 0B y +B de Cádiz, a lo largo del elemento Helitron-N17B_LMi. Destaca la mayor cobertura en las librerías +B en una pequeña región entre las posiciones 1600 y 1800.

Tablas suplementarias

Tabla 4.S1: Cebadores de histonas y elementos transponibles diseñados para este estudio. Los cebadores utizados para generar sondas de FISH para ADN satélites se encuentran en la Tabla 2.S8.

Pareja de cebadores	F	R
Histones_Lmig_SP1	CGCCCCAAGAAGGCGACGAC	AGCGCGTCACCATCATGCCC
Histones_Lmig_SP2	CTTGGTGGCGAGCTGTTTGCG	CCCTTGCCGAGCCCCTTTCC
Histones_Lmig_SP3	CATGGACGTGGTGTACGCCCT	GCGGTGTCCTGCCCAACATCC
Histones_Lmig_SP4	CTGGAGCGGGACTTTGACTTTCCC	TTCGAGATATTCTTCTGCGCCTTGCC
Histones_Lmig_SP5	GGCCGCAGCAGTCTCGTTCG	GGCCGCAGCAGTCTCGTTCG
Histones_Lmig_SP5_marker	CTCTCTTGTACGCGGCGTTT	GTGGTCGGCTGGGTAACG
Histones_Lmig_H3	AAGTCCACCGGCGGAAAGGC	CGCGCCAGCTGGATGTCCTT
Daphne-7_LMi	AGACCTATTTCTATGCCATCGG	GCAACATCCTTCACTACCAAGC
Penelope-52_LMi	TTCCTTTCACGCCGACCAC	TGCGGGATGTGAAAGCTGTTAT
DNA2-4_LMi_sat	TGGAAGAAATCCATGCATGTGTCACG	CCACACATGTCCTTTTGACATTTCG
hAT-4_LMi	GCTCAGACAGCGTGAAGTTGCG	TCTCCAGGCCCATCAACTTT
Helitron-N17B_LMi	AGATTCCAAGTTACCGCACA	ACAAACGTCAGACCAAATTGACT
Kolobok-4_LMi	TGGGACATGGCAGAAGCGTGG	TGGCAGGGAGTGATGGTGTGTGT
Penelope-42_LMi_HC	AGCCCAGAAAGCAATGAAAGCC	TCTTCTATTAGCTGCTTGCTGT
Penelope-42_LMi_LC	ACAGTGGAACATGAAAACAATGGCA	AGATGCCTGAGTGTGTGTGT
Penelope-59_LMi_HC	TGAAGACTTCCATCAAAGGTTTAACA	TTTGTCCCCAGTATGGCAGT
Penelope-59_LMi_LC	AGGCTTCTACACTAACAACACT	GGCCAAGGTGTTATGGAAAT
Sola2-3N1_LMi	AACGTTGACCTGTTTGCTGT	TGGAGCGTTGCAGTGCGTGA
Tx1-1_LMi_HC	TGCTTCGAGCGACACTTCTGACA	TGAAGCAAAACCCGAGCAACT
Tx1-1_LMi_LC	ACAATCATACAGTATAGCCACCGT	CGGCATCTGATCGTTTGCGGA

Tabla 4.S2: Abundancia de histonas, ADN satélites y elementos repetidos en el ADNg de los individuos de Cádiz.

	01_0B	04_0B	02_PB	03_PB	05_PB	06_PB	Average 0B	Average PB	Cádiz log2(PB/0B)
DNA	0.0301	0.0304	0.0300	0.0304	0.0300	0.0303	0.0302	0.0302	-0.0006
DNA/Academ	0.0023	0.0024	0.0024	0.0023	0.0024	0.0023	0.0024	0.0023	-0.0005
DNA/EnSpm-CACTA	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0402
DNA/Ginger2-TDD	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	-0.0525
DNA/Harbinger	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	-0.0278
DNA/ISL2EU	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0280
DNA/Kolobok	0.0021	0.0021	0.0020	0.0020	0.0021	0.0022	0.0021	0.0021	-0.0134
DNA/Mariner-Tc1	0.0607	0.0609	0.0608	0.0602	0.0597	0.0600	0.0608	0.0602	-0.0043
DNA/MuDR	0.0008	0.0008	0.0007	0.0007	0.0007	0.0007	0.0008	0.0007	-0.0225
DNA/P	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0329
DNA/Polinton	0.0033	0.0035	0.0033	0.0034	0.0033	0.0033	0.0034	0.0033	-0.0110
DNA/Sola1	0.0008	0.0008	0.0008	0.0008	0.0009	0.0008	0.0008	0.0008	0.0229
DNA/Sola2	0.0019	0.0019	0.0020	0.0020	0.0020	0.0020	0.0019	0.0020	0.0149
DNA/Transib	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0005	0.0004	-0.0764
DNA/hAT	0.0362	0.0360	0.0360	0.0361	0.0356	0.0353	0.0361	0.0357	-0.0044
DNA/piggyBac	0.0044	0.0043	0.0042	0.0044	0.0044	0.0044	0.0043	0.0043	-0.0023
LINE/CR1	0.0582	0.0584	0.0572	0.0574	0.0572	0.0570	0.0583	0.0572	-0.0085
LINE/Daphne	0.0188	0.0193	0.0189	0.0186	0.0190	0.0186	0.0191	0.0188	-0.0068
LINE/I	0.0010	0.0011	0.0011	0.0011	0.0010	0.0010	0.0011	0.0011	-0.0027
LINE/Ingi	0.0003	0.0002	0.0003	0.0003	0.0003	0.0003	0.0002	0.0003	0.0537
LINE/Jockey	0.0039	0.0037	0.0039	0.0035	0.0037	0.0036	0.0038	0.0037	-0.0138
LINE/Kiri	0.0012	0.0012	0.0011	0.0012	0.0011	0.0012	0.0012	0.0012	-0.0033
LINE/L2B	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	-0.0433
LINE/Loa	0.0049	0.0052	0.0048	0.0049	0.0050	0.0048	0.0050	0.0049	-0.0152
LINE/Nimb	0.0294	0.0292	0.0296	0.0284	0.0292	0.0293	0.0293	0.0291	-0.0021
LINE/Penelope	0.0410	0.0415	0.0414	0.0410	0.0407	0.0403	0.0412	0.0409	-0.0041
LINE/R1	0.0007	0.0007	0.0006	0.0007	0.0007	0.0007	0.0007	0.0007	0.0136
LINE/RTE	0.0728	0.0735	0.0719	0.0717	0.0722	0.0714	0.0731	0.0718	-0.0080
LINE/Tx1	0.0048	0.0048	0.0047	0.0045	0.0047	0.0048	0.0048	0.0047	-0.0080
LINE/Vingi	0.0038	0.0037	0.0037	0.0035	0.0037	0.0036	0.0037	0.0036	-0.0139
LTR	0.0059	0.0058	0.0057	0.0059	0.0057	0.0058	0.0058	0.0058	-0.0047
LTR/BEL	0.0058	0.0059	0.0059	0.0059	0.0057	0.0059	0.0059	0.0058	-0.0018
LTR/Copia	0.0010	0.0010	0.0009	0.0009	0.0009	0.0009	0.0010	0.0009	-0.0422
LTR/Gypsy	0.0426	0.0431	0.0419	0.0412	0.0415	0.0419	0.0428	0.0416	-0.0124
RC/Helitron	0.0421	0.0424	0.0418	0.0428	0.0419	0.0416	0.0423	0.0420	-0.0028
SINE	0.0126	0.0123	0.0124	0.0124	0.0122	0.0124	0.0124	0.0123	-0.0031
SINE/SINE2-tRNA	0.0216	0.0219	0.0214	0.0211	0.0212	0.0213	0.0218	0.0212	-0.0103
SatDNAs	0.0240	0.0241	0.0354	0.0323	0.0354	0.0438	0.0241	0.0367	0.1839
Unknown/Nonautonomous	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0070
Unknown/TransposableElement	0.0058	0.0061	0.0051	0.0048	0.0051	0.0044	0.0060	0.0048	-0.0918
Histones	0.0011	0.0009	0.0017	0.0013	0.0016	0.0015	0.0010	0.0015	0.1815
Total	0.5475	0.5507	0.5552	0.5494	0.5525	0.5588	0.5491	0.5540	0.0643

Tabla 4.S3: Tamaño de los genes de histonas en *L. migratoria* y *E. plorans*.

		length	
Region	Orientation	*L. migratoria*	*E. plorans*
H1	+	699	711
Spacer 1		1244	1761
H3	-	411	411
Spacer 2		269	252
H4	+	312	312
Spacer 3		1021	1574
H2A	-	375	375
Spacer 4		268	273
H2B	+	372	372
Spacer 5		3311	3848
Total		8282	9889

Tabla 4.S4: Abundancia de todas las secuencias de referencia utilizadas en este estudio ordenadas por diferencia de frecuencia entre librerías +B y 0B.

This table can be downloaded in https://dx.doi.org/10.6084/m9.figshare.3255556

Tabla 4.S5: Diferencia de abundancia entre liberarías +B y 0B de ADNg para las distintas familias de ADNs satélites.

Family	+B-0B	B Proportion	Family	+B-0B	B Proportion
LmSat02A	0.0116562	0.7364671	LmSat35A	0.0000242	0.0015288
LmSat01A	0.0005160	0.0326014	LmSat54A	0.0000240	0.0015160
LmSat04A	0.0004732	0.0298974	LmSat48A	0.0000198	0.0012500
LmSat16A	0.0001706	0.0107791	LmSat41A	0.0000192	0.0012151
LmSat18A	0.0001024	0.0064680	LmSat31A	0.0000183	0.0011544
LmSat09A	0.0000983	0.0062109	LmSat42A	0.0000148	0.0009323
LmSat10A	0.0000704	0.0044455	LmSat55A	0.0000136	0.0008616
LmSat06A	0.0000677	0.0042802	LmSat25A	0.0000121	0.0007672
LmSat53A	0.0000432	0.0027296	LmSat45A	0.0000089	0.0005644
LmSat19A	0.0000385	0.0024340	LmSat59A	0.0000077	0.0004837
LmSat15A	0.0000368	0.0023221	LmSat13A	0.0000075	0.0004709
LmSat40A	0.0000338	0.0021370	LmSat49A	0.0000039	0.0002455
LmSat33A	0.0000334	0.0021088	LmSat46A	0.0000009	0.0000579
LmSat29A	0.0000324	0.0020473	LmSat43A	0.0000008	0.0000529
LmSat23A	0.0000281	0.0017781	LmSat52A	0.0000008	0.0000485
LmSat17A	0.0000280	0.0017713	LmSat26A	0.0000007	0.0000465
LmSat07A	0.0000257	0.0016262	LmSat27A	0.0000007	0.0000461
LmSat24A	0.0000253	0.0015976	LmSat58A	0.0000004	0.0000252
LmSat30A	0.0000250	0.0015778	**Total**	**0.0136833**	**0.8645467**

Tabla 4.S6: Diferencia de abundancia entre liberarías +B y 0B de ADNg para las distintos grupos de elementos repetidos.

Type of element	+B-0B	B Proportion
SatDNAs	0.0136833	0.8645500
Histones	0.0005521	0.0348800
DNA/hAT	0.0003117	0.0196900
LINE/Penelope	0.0002092	0.0132200
LTR/Gypsy	0.0001280	0.0080900
RC/Helitron	0.0001201	0.0075900
LINE/Nimb	0.0001200	0.0075800
DNA/Sola2	0.0000995	0.0062900
DNA/Mariner	0.0000813	0.0051400
LINE/Daphne	0.0000698	0.0044100
DNA/DNA8	0.0000569	0.0036000
DNA/piggyBac	0.0000538	0.0034000
LINE/RTE-8	0.0000474	0.0030000
LINE/Vingi-5	0.0000407	0.0025700
LINE/Tx1	0.0000374	0.0023600
LTR/Copia-6	0.0000326	0.0020600
DNA/DNA9	0.0000316	0.0020000
DNA/DNA3	0.0000235	0.0014800
DNA/Sola1	0.0000189	0.0011900
DNA/Academ	0.0000164	0.0010400
LINE/Jockey	0.0000145	0.0009200
LINE/I	0.0000138	0.0008700
SINE/SINE	0.0000114	0.0007200
DNA/Kolobok	0.0000113	0.0007100
DNA/DNA2	0.0000085	0.0005300
DNA/Polinton	0.0000084	0.0005300
DNA/MuDR	0.0000073	0.0004600
LTR/LTR	0.0000055	0.0003500
LINE/Kiri	0.0000025	0.0001600
Unknown	0.0000022	0.0001400
DNA/P	0.0000016	0.0001000
DNA/ISL2EU	0.0000015	0.0000900
DNA/Harbinger	0.0000014	0.0000900
LINE/Ingi	0.0000014	0.0000900
LTR/BEL	0.0000010	0.0000700
DNA/Ginger2	0.0000003	0.0000200
LINE/L2	0.0000002	0.0000100
Total	**0.0158300**	**1.0000000**

Tabla 4.S7: Estudio de la intensidad de transcripción para varios TEs, inferida de la presencia de SNPs específicos de las librerías de ADNg y ARN con B.

This table can be downloaded in https://dx.doi.org/10.6084/m9.figshare.3255556

Tabla 4.S8: Anotación de la quimera presente en el cromosoma B. d = direct orientation, c = complementary orientation.

Quimera position			Element position						
From	To	Name	From	To	Class	Dir	Sim	Pos / Mm:Ts	Score
1	919	Sola2-3_LMi	2780	3681	DNA/Sola/Sola2	d	0.9362	2.3	6861
1655	1754	Penelope-100_LMi	2335	2435	NonLTR/Penelope	c	0.6931	1.8125	294
1926	2124	Mariner-22_LMi	879	1086	DNA/Mariner	c	0.6158	1.6744	288
2318	2830	Penelope-59_LMi	689	1201	NonLTR/Penelope	d	0.9708	1.875	4271
2833	2981	hAT-N33_LMi	1036	1185	DNA/hAT	c	0.9533	2	1185
2982	3438	DNA8-6_LMi	1006	1462	DNA	c	0.9168	2.1111	3363
3440	3599	DNA8-6_LMi	621	800	DNA	c	0.9317	1.6667	1159
3600	4076	DNA8-6_LMi	45	544	DNA	c	0.9272	2.2143	3551
4228	4408	Helitron-N12C_LMi	24	204	DNA/Helitron	d	0.7956	2.2857	778
4559	4600	Gypsy-129_SBi-LTR	360	401	LTR/Gypsy	c	0.8095	1.6	234
5399	5701	RTE-31_LMi	2677	2991	NonLTR/RTE	c	0.6343	1.623	375
6220	6553	Mariner-23_LMi	936	1269	DNA/Mariner	c	0.9222	1.3684	2594
6554	6838	Mariner-23_LMi	1	307	DNA/Mariner	c	0.9373	2.2857	2132
6872	7082	Charlie12	1797	2011	DNA/hAT	d	0.7656	2.0909	810
7101	7204	MARINER45_CB	321	430	DNA/Mariner	c	0.7404	1.9091	290
7266	7515	Penelope-42_LMi	2582	2843	NonLTR/Penelope	c	0.8976	1.5	1712
7516	7753	Penelope-42_LMi	223	481	NonLTR/Penelope	d	0.9185	2.5	1382
7759	7957	SINE-3_LMi	33	230	NonLTR/SINE	c	0.9347	2.4	1427
7962	8382	Penelope-42_LMi	474	893	NonLTR/Penelope	d	0.9454	1.5714	3379
8386	9141	Tx1-1_LMi	2890	3663	NonLTR/Tx1	d	0.9459	1.7273	5978
9169	9207	BEL-2_LMi-I	5314	5352	LTR/BEL	c	0.8462	1.2	269
9208	9380	BEL-2_LMi-I	4845	5017	LTR/BEL	d	0.7989	1.5	983
9542	9637	REP-4_LMi	401	496	Interspersed_Repeat	d	0.8750	4	567
9695	9729	CR1-2_DF	1576	1612	NonLTR/CR1	d	0.8889	1.5	213
9752	9796	TE-2_LMi	318	365	Interspersed_Repeat	c	0.8511	1.6667	268
9850	10046	SINE2-3_LMi	69	249	NonLTR/SINE/SINE2	c	0.7540	1.6957	673
10077	10129	L1-52_NN	2235	2287	NonLTR/L1	d	0.7547	1.8571	249
10571	10693	Lm2C1	62	199	NonLTR/SINE/SINE2	d	0.9274	1.6	887
10819	10980	SINE2-2_LMi	64	234	NonLTR/SINE/SINE2	c	0.8012	1.7059	743
11278	11776	RTE-2_LMi	2778	3285	NonLTR/RTE	c	0.9340	1.3913	4073
11777	12294	RTE-2_LMi	2132	2649	NonLTR/RTE	c	0.9208	1.5769	4123
12299	12633	RTE-2_LMi	2	336	NonLTR/RTE	c	0.9373	1.5	2816
12892	13092	SINE-3_LMi	33	229	NonLTR/SINE	d	0.9447	1.2857	1464
13132	13267	RTE-2_LMi	2012	2147	NonLTR/RTE	d	0.9632	2.5	1174
13269	13469	Lm2	2	203	NonLTR/SINE/SINE2	c	0.9455	3	1564
13470	13892	RTE-2_LMi	2136	2558	NonLTR/RTE	d	0.9362	1.8	3479
13893	14271	RTE-2_LMi	2911	3285	NonLTR/RTE	d	0.9468	1.7273	3140
14364	14597	hAT-25_LMi	2047	2307	DNA/hAT	c	0.7676	1.6667	647
14887	15108	Helitron-N13_LMi	1552	1791	DNA/Helitron	c	0.7411	2.0769	685
15348	15496	DNA-7_LMi	2729	2871	DNA	d	0.7591	1.625	406
15970	16045	Zisupton-6_DR	13618	13690	DNA/Zisupton	d	0.8310	4	249
17062	17271	SINE-3_LMi	21	230	NonLTR/SINE	c	0.9336	1.8333	1491

Chapter 5. A genetically well equipped parasitic chromosome

Francisco J. Ruiz-Ruano and Juan Pedro M. Camacho

Departamento de Genética, Universidad de Granada

In many species we see new fragment chromosomes appearing from time to time. Most of these probably come to nothing. (Darlington & Upcott, 1941)

Abstract. Supernumerary (B) chromosomes are regarded as beneficial (heterotic) or harmful (parasitic) genomic elements despite being incongruously considered genetically inactive. Here we show the presence of 24 protein-coding genes in the B chromosome of the migratory locust, about half of which showed a complete coding region. Sequence divergence between the A and B chromosome gene copies suggested that the age of this B chromosome is between 1 and 4 million years, and the finding of B-specific single nucleotide polymorphisms (SNPs) for many of these genes demonstrated the activity of B chromosome genes. The long age and the worldwide spread of B chromosomes in this species are based on their transmissional advantage through drive mechanisms (mitotic non-disjunction and meiotic non-Mendelian segregation). Remarkably, three of the B chromosome active genes code for E3 ubiquitin ligases involved in regulation of cell division, which is the main arena where B chromosome destiny is played. The *APC1* gene codes for the largest subunit of the Anaphase Promoting Complex, which regulates the metaphase-anaphase transition during mitosis and also meiotic resumption in oocytes. An excess of APC1 protein could pave the way to B chromosome non-disjunction by driving cells to anaphase even though only one of the B chromatids is linked to microtubules. *MDM2* codes for the Murine Double Minute 2 protein which is the main negative regulator of the tumor suppressor p53. *MDM2* activity in the B chromosome of *Locusta migratoria* might decrease p53 surveillance allowing a certain degree of genomic instability caused by the B chromosome. Finally, *RBBP6* codes for the retinoblastoma binding protein 6, which plays a role as a scaffold protein promoting the assembly of the p53/TP53-MDM2 complex. Taken together, these findings demonstrate that B chromosomes are not genetically inert elements. On the contrary, they behave as true parasites eliciting a true transcriptional arms race with the host A chromosomes.

Introduction

After 110 years since they were uncovered (Wilson, 1907), B chromosomes continue being a mysterious part of eukaryote genomes. On the basis of their highly condensed state, highly contrasting with standard (A) chromosomes, B chromosomes have been viewed for the most part of this time as merely genetically inert passengers of eukaryote genomes (Rhoades & McClintock, 1935). This view was supported by Randolph (1941) after showing experimentally that the B chromosomes of maize are not essential for normal growth and reproduction, and are not beneficial for the individuals carrying them. This inert view of B chromosomes was soon criticized by those who viewed B chromosomes as beneficial (Darlington & Upcott, 1941) or parasitic (Östergren, 1945). The logic of these alternative views is that B chromosome effects could be mediated through the expression of genes contained in them. It is thus surprising that whereas the beneficial (heterotic) and parasitic views of B chromosomes have been accepted, with predominance of the parasitic one, their genetic inertness has also remained in the literature as a universal feature of B chromosomes (see, for instance, Battaglia 1964; Jones & Rees 1982; Camacho *et al.* 2000) even though some authors appeared to be reluctant to view them as a large piece of inert chromatin simply acting as a "spanner in the works" (Harvey & Hewitt, 1979), and others viewed them as "active and selfish chromosomes with high degree of autonomy and genetic sophistication" (Jones, 1995).

Eleven years ago, three findings suggested that B chromosomes are not inert elements, namely the first molecular evidence of gene activity on B chromosomes (Leach *et al.*, 2005), the first finding of protein-coding genes in them (Graphodatsky *et al.*, 2005) and the differential expression of three genes in presence of B chromosomes Tanić *et al.* 2005. Evidence for transcription in B chromosomes was later shown by Carchilan *et al.* (2009) and Ruiz-Estevez *et al.* (2012) provided the first indication that B-transcripts are functional, as those individuals carrying ribosomal RNA transcripts with an adenine insertion being specific to B chromosomes also carried a nucleolus attached to the B chromosome. The first evidence for transcription of a protein-coding gene on B chromosomes was reported by Trifonov *et al.* (2013).

The arrival of next generation sequencing (NGS) has drastically accelerated B chromosome research during last years. The pioneering work by Martis *et al.* (2012) revealed that B chromosomes in rye are rich in gene fragments that represent copies of A chromosome genes. In this same species, Banaei-Moghaddam *et al.* (2013) compared these gene-like fragments of the B chromosome with their ancestral A-located counterparts, and confirmed their A chromosome origin and also the pseudogenization of many

B-located gene-like fragments, 15% of which were transcribed in a tissue-type and genotype-specific manner. Likewise, Valente *et al.* (2014) found that a B chromosome in fish contained thousands of gene-like sequences, most of them being fragmented but a few remaining largely intact, with at least three of them being transcriptionally active. Recently, Navarro-Domínguez et al. (submitted) have found evidence for the presence of nine protein-coding genes in the B chromosome of the grasshopper *Eyprepocnemis plorans*, five of which were active in B-carrying individuals, including three which were apparently pseudogenic. Taken together, these results indicate that B chromosomes are not as inert as previously thought, since they contain protein-coding genes which are actively transcribed even in the case of being pseudogenic.

All these evidences are challenging the view that B chromosomes are genetically inert elements lacking functional genes (Banaei-Moghaddam *et al.*, 2015). Anyway, the extent to which B chromosomes express their genetic content is still unknown. Lin *et al.* (2014) characterized B-chromosome-related transcripts in maize and concluded that the maize B chromosome harbours few transcriptionally active sequences, and that it might influence transcription in A chromosomes. Likewise, B chromosome rRNA genes in *E. plorans* actually contribute insignificant amounts of rRNA (Ruiz-Estévez *et al.*, 2014). Recently, Huang *et al.* (2016) have found that the expression of maize A chromosome genes is influenced by the presence of B chromosomes, and that four upregulated genes are actually present in the B chromosomes. Here we develop NGS approaches to unveil protein-coding genes in B chromosomes, by means of genomic and transcriptomic analyses. At least 24 genes were found in the B chromosome of the migratory locust, 12 of which were complete and 10 were active. Remarkably, three genes coding for E3 ubiquitin ligases were involved in functions being very useful for lowering host defenses and thus promoting B chromosome invasion and maintenance.

Materials and Methods

Materials

We collected 20 males of *Locusta migratoria* in two natural populations close to Los Barrios (Cádiz, Spain), in Finca El Patrón (36.20685N, -5.46481W) and 15 in Puente de Hierro (36.19251, -5.55131), and 10 males in Padul (Granada, Spain). After anaesthesia, we fixed one testis in 3:1 ethanol-acetic acid, for cytological analysis, and the other testis and body remains were separately frozen in liquid nitrogen for DNA and RNA extraction. We determined the presence of B chromosomes by squashing two testis follicles in 2% lacto-propionic orcein and visualizing primary spermatocytes at first meiotic prophase or metaphase. Only two males from Cádiz and one from Padul lacked B chromosomes. We selected these three 0B males and five B-carrying males (four from Cádiz and one from Padul) for genomic analysis. In the six males from Cádiz, we extracted RNA from the testis and one of the hind legs, and DNA from the other hind leg. In the males from Padul, we used testes for cytological analysis and DNA extraction and body remains for RNA extraction. The six individuals from Cádiz yielded 12 RNA-Seq paired-end libraries by means of the Illumina HiSeq2000 platform (2x101 nt read length), as well as 6 genomic DNA (gDNA) paired-end libraries by the Illumina HiSeq2500 platform (2x125 nt). The two males from Padul yielded one RNA-Seq and one gDNA libraries by means of Illumina HiSeq 2000 (2x101 nt). In addition we generated one RNA-Seq Illumina HiSeq2000 (2x101 nt) from testis of the *L. migratoria* closest relative species in Spain, *Oedaleus decorus*, which shared the most recent common ancestor about 22.81 mya (Song *et al.*, 2015). We collected this specimen at Capileira (Granada) (36.961348 N, -3.342403 W).

De novo transcriptome assembly, mapping and selection

We generated a *de novo* transcriptome, and it was used as a reference to map the genomic reads. For this purpose, we concatenated the 12 RNA-Seq libraries and performed an *in silico* normalization with 50x maximum coverage. This selected ~68 million reads which were assembled using Trinity (Haas *et al.*, 2013) with default options. Since we were only interested in knowing the presence of protein coding genes in the *L. migratoria* B chromosome, we extracted the CDSs being longer than 100 aminoacids, using the Transdecoder software (Haas *et al.*, 2013) and then reduced redundancy with CDHit-EST (Li & Godzik, 2006) with local alignment and the greedy algorithm, and grouped those sequences showing 80% or higher similarity in at least 80% of length (options -M 0 -aS 0.8 -c 0.8 -G 0 -g 1). For the *O. decorus* RNA-Seq library we performed the same *de novo* transcritome assembly protocol.

We mapped the genomic and transcriptomic reads against the reference transcriptome using SSAHA2 (Ning *et al.*, 2001). This software allows mapping reads showing high variation in respect to the reference, and it accepts partial read mappings. This is crucial for our purpose, since the sequences used for reference lack introns, which are however present in the genomic reads. We accepted mappings with at least 40 nt, and counted the reads in the BAM files by means of a custom script. We expressed CDS abundance in each library as the number of copies per haploid genome using the formula:

$$Number\,of\,copies = \frac{mapped\,reads \times read\,length \times genome\,size\,(nt)}{CDS\,length \times library\,size\,(nt)}$$

We separately estimated mean abundance in both 0B and +B genomic libraries. We then filtered out CDSs according to abundance. We excluded highly represented CDSs, because they were candidates to be repeated sequences, by selecting CDSs with a mean copy number lower than 4 in the genomic 0B libraries. In addition, we excluded CDSs with copy number being lower than 0.5, in both genomic 0B and +B libraries, as they could correspond with assembly errors. We then calculated the fold change (FC) in CDS abundance due to B chromosome presence as log2 of the quotient between copy numbers per haploid genome estimated in B-carrying and B-lacking gDNAs from the 6 males from Cádiz and the 2 from Padul. We annotated CDSs showing over-representation in +B genomes from Cádiz individuals using the Trinotate pipeline (https://trinotate.github.io/) and the SWISS-PROT database (Boeckmann *et al.*, 2003).

We estimated the expression level of the annotated CDSs in the 12 RNA-Seq libraries by calculating gene expression as RPKM following this equation:

$$RPKM = \frac{10^9 \times mapped\,reads\,contig}{total\,mapped\,reads \times contig\,length\,(nt)}$$

Sequence analysis for selected CDSs

For the genes being over-represented in B-carrying genomes, we performed additional analyses. We first checked if the CDS was complete in the contig. If it was incomplete, we used an additional Trinity transcriptome assembly with the two 0B libraries from Cádiz, to search for this CDS with the BLASTN algorithm (Altschul *et al.*, 1990). On the contig with the longest CDS, we mapped with SSAHA2 genomic and transcriptomic libraries to estimate the average coverage per position along the contig, also calculating the standard deviation, in 0B and +B libraries. For this purpose, we used the pysamstats software integrated in a custom script.

Exon limits were found using the Exonerate software (Slater & Birney, 2005) and aligning with the *L. migratoria* genome assembled by Wang *et al.* (2014). For genes showing high coverage for part of an exon only, we analyzed the possible tandem-repeated structure by selecting reads showing homology with this region, and clustering and assembling them with RepeatExplorer (Novák *et al.*, 2013). We also searched for B-specific SNPs in gDNA and RNA-Seq libraries with a custom script. We then searched for fixed positions in the 0B gDNA and RNA libraries showing an alternative nucleotide in the +B gDNA libraries. In addition we scored, in the Illumina reads, the abundance of haplotypes defined by SNPs separated by less than 100 nt (one read length).

We also calculated the coverage per nucleotide position in gDNA and RNA. In gDNA we estimated the mean number of copies by normalizing in respect to coverage in the genomic 0B libraries and library size. For RNA-Seq libraries we normalized in respect to library size.

Finally, we performed functional gene annotation using Eukaryotic Orthologous Groups of proteins (KOG), by searching the predicted protein sequence in the WebMGA server (http://weizhong-lab.ucsd.edu/metagenomic-analysis/server/kog).

Results

Searching for protein-coding genes in the B chromosome

De novo transcriptome assembly of the 12 RNA libraries from Cádiz (testis and hind leg from six males) yielded 523,445 contigs (N50= 792 nt), 108,517 of which included CDSs higher than 300 nt (N50= 891 nt). After clustering (to reduce redundancy) we got a final reference assembly including 49,476 CDSs (N50= 975 nt). On this *de novo* transcriptome reference, we separately mapped 8 gDNA libraries (6 from Cádiz and 2 from Padul) as well as 14 RNA libraries (the 12 from Cádiz plus 2 from Padul), scoring the number of reads from each library mapped on each CDS, and expressed the normalized results as number of gene copies per haploid genome, in the case of gDNA, and as RPKM for RNA reads. To discard poorly assembled CDSs or including repetitive DNA, we selected the 27,313 CDSs showing less than 4 or more than 0.5 copies in the 0B genome (Fig. 5.1).

Selection of candidate genes to reside in the B chromosome was done based on FC, following two criteria: i) all genes showing FC> 2 in the Cádiz libraries (i.e. four-fold copies in the B-carrying genomes) were selected, and ii) those genes showing FC> 1 in Cádiz and Padul were also selected. The high consistency in FC between Cádiz and Padul gDNAs (Fig. 5.2) provides strong support to the genes inferred. The few exceptions failing to be over-represented in the B-carrying male from Padul (i.e., *OR92*, *SV2*, *CL065*, *LSAMP* and *TRY1*) were actually genes showing irregular coverage, the four latter including satellite DNAs being only present in B-carrying individuals, i.e. presumably in the B chromosome. The absence of these satellites in some B-carrying individuals from Cadiz, and the fact that low

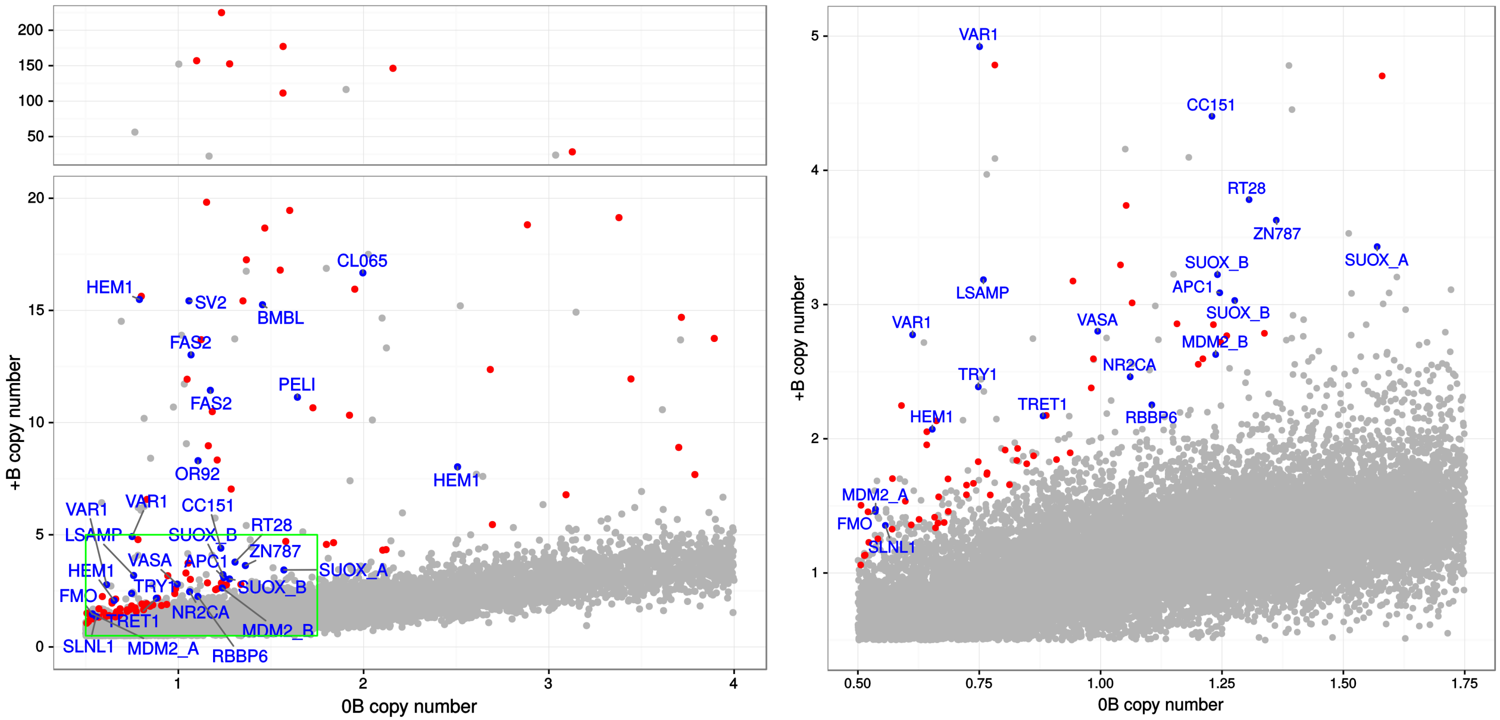


Figure 5.1: Scatterplot representing the mean copy number for 27,313 CDSs in 0B and +B individuals from the Cádiz population (lower panel) and enlargement of the region included in the green box (right panel). Candidate genes to be found in the B chromosome are remarked as blue dots with gene names in blue letters, whereas transposable elements are indicated as red dots.

Table 5.1: Candidate genes to be present in the B. GC: log2(+B/0B) of gDNA from Cádiz. GP: log2(+B/0B) of gDNA from Padul. RTC: log2(+B/0B) of testis RNA-Seq from Cádiz. RLC: log2(+B/0B) of leg RNA-Seq from Cádiz. RP: log2(+B/0B) of body RNA-Seq from Padul. They are sorted by GC.

		CDS	log2(+B/0B)				
Annotation	Acronym	length	GC	GP	RTC	RLC	RP
B-cell receptor CD22	*HEM1*	300	4.29	5.79	5.58	1.39	5.43
SV2	*SV2*	315	3.87	1.22	-0.93	-0.58	3.17
FAS2_SCHAM	*FAS2*	864	3.61	5.83	4.67	4.87	6.39
BMBL_DANRE	*BMBL*	816	3.39	7.73	5.48	3.93	4.92
CL065_HUMAN	*CL065*	468	3.06	-0.37	0.46	-0.03	-0.36
OR43A_DROME	*OR92*	564	2.91	1.03	3.13	-1.32	0.00
PELI_DROME	*PELI*	1179	2.76	4.21	2.60	2.66	3.26
Mitoc. ribos. VAR1	*VAR1*	351	2.71	4.10	4.94	3.32	2.00
LSAMP_MOUSE	*LSAMP*	480	2.07	0.22	0.32	-1.47	1.00
CC151	*CC151*	1650	1.84	2.57	3.14	0.01	1.03
TRY1_ANOGA	*TRY1*	510	1.67	-0.66	0.29	0.00	0.00
RT28_MOUSE	*RT28*	543	1.53	2.50	0.57	0.06	-0.57
Vasa	*VASA*	324	1.49	5.15	2.04	1.80	2.32
E3 ubiq-prot lig Mdm2	*MDM2_A*	510	1.46	2.72	1.44	1.61	1.67
GSOX5_ARATH	*FMO*	1248	1.44	3.35	1.95	1.33	2.30
ZN787_MOUSE	*ZSC22*	651	1.41	1.70	2.34	1.79	3.14
SUOX_DROME	*SUOX_A*	351	1.38	1.51	4.39	1.39	4.95
APC1_HUMAN	*APC1*	5874	1.31	2.48	2.23	1.24	0.60
TRET1_POLVA	*TRET1*	1119	1.30	2.00	0.43	1.52	2.81
SLNL1_HUMAN	*SLNL1*	360	1.28	1.44	1.70	0.00	0.00
NR2CA_DANRE	*NR2CA*	420	1.21	0.87	0.01	0.12	-0.68
SUOX_DROME	*SUOX_B*	777	1.13	2.45	0.77	-0.30	0.02
E3 ubiq-prot lig MDM2	*MDM2_B*	1176	1.09	2.15	1.38	1.41	2.68
E3 ubiq-prot lig RBBP6	*RBBP6*	336	1.03	1.07	4.04	1.28	0.79

coverage regions of these genes showed similar abundances in B-carrying and B-lacking individuals, suggest that these satellites are of recent origin and are not present in all B chromosomes.

Annotation of these 24 CDSs indicated that all were protein-coding genes (Table 5.1). We analyzed the sequence of these 24 gene transcripts trying to infer as much of their length as possible, thus including some UTR regions. For this purpose we generated a *de novo* transcriptome with the four RNA libraries from the two 0B individuals from Cádiz and searched for transcripts being homologous to the selected contigs by BLAST (Altschul *et al.*, 1990). This provided full length transcripts for 17 genes (Table 5.2).

Using these 24 gene transcripts as reference, we separately mapped the reads from each of the 8 gDNA libraries to examine coverage variation per position along transcript length (Table 5.3; see values per individual in Table 5.S1). The gDNA mappings revealed two types of B chromosome genes

Table 5.2: Basic characteristics for the candidate genes to be present in the B. They are sorted by type of presence in the B. Number of SNPs in the CDS is indicated. We consider deletions (d) as a single mutation and indicated if they generated a stop codon. We found the same pattern for all the genes with exception of five genes. In these latter genes we indicate the number of individuals where the mentioned pattern was found. UC = Uniform coverage. IC-tr = Irregular Coverage truncated. IC-ra = Irregular Coverage-regional amplification. IC-sat = Irregular Coverage-satellite.

	Distribution		CDS	Filter 2		Filter 3 (Alt/Ref > 25%)		
Acronym	in B copies	CDS complete?	length	#SNPs	Ratio	#Syn	#NonSyn	NonSyn/Syn
PELI	UC	Complete	1206	8	0.66%	0	0	–
VAR1	UC	Complete	1596	24	1.50%	0	0	–
CC151	UC	Complete	1650	5	0.30%	0	0	–
MDM2_A	UC	Lacks 5′ end	510	1	0.20%	0	0	–
FMO	UC	Complete	1350	4	0.30%	1	1	1.00
ZSC22	UC	Complete	633	13	2.05%	2	2 (1d-STOP)	1.00
SUOX_A	UC	Complete	777	15	1.28%	2	8	4.00
APC1	UC	Complete	5907	34	0.58%	8	1	0.13
NR2CA	UC	Complete	420	3	0.71%	0	1	–
SUOX_B	UC	Complete	1689	7	0.90%	1	0	0.00
MDM2_B	UC	Complete	1198	9	0.75%	0	1	–
RBBP6	UC	Lacks 5′ and 3′ end	336	5	1.49%	1	2	2.00
FAS2	IC-tr	Complete	2595	22	0.85%	10	2 (1d)	0.20
RT28	IC-tr	Complete	543	4	0.74%	1	0	0.00
VASA	IC-tr	complete	1800	38	2.11%	9	4 (3d-STOP)	0.75
TRET1	IC-tr	Lacks 5′ and 3′ end	1121	8	0.71%	1	5	5.00
SLNL1	IC-tr	Complete	1059	19	1.79%	8	6	0.75
HEM1	IC-ra	Lacks 5′ end	4185	153	3.66%	2	6	3.00
BMBL	IC-tr	Lacks 3′ end	1341	72	5.37%	1	0	0.00
OR92	IC-tr (3 ind)	Complete	438	68	1.,53%	3	1	0.33
SV2	IC-sat (3 ind)	Complete	1605	57	3.55%	0	2	–
CL065	IC-sat (2 ind)	Complete	468	18	3.85%	0	0	–
TRY1	IC-sat (1 ind)	Lacks 3′ end	514	47	9.14%	6	2	0.33
LSAMP	IC-sat (2 ind)	Lacks 5′ end	867	11	1.27%	1	2	2.00

(Table 5.3). Twelve genes showed uniform coverage (UC) manifested by low standard deviation (SD) of coverage per site, which we interpret as full genes (Fig. 5.3). The remaining 12 genes showed irregular coverage (IC) (thus high SD), with three subtypes: five genes being apparently truncated (IC-tr genes), three showing differential regional amplification (IC-ra) and four genes including part of their DNA sequence tandemly repeated yielding a satellite DNA into it (IC-sat) (see the patterns for all genes in Fig. 5.S2-5.S24).

In spite of low coverage, our approach has uncovered the presence of 24 protein-coding genes in the B chromosome of *L. migratoria*. Of course, this is by no means the complete set of B chromosome genes, and future

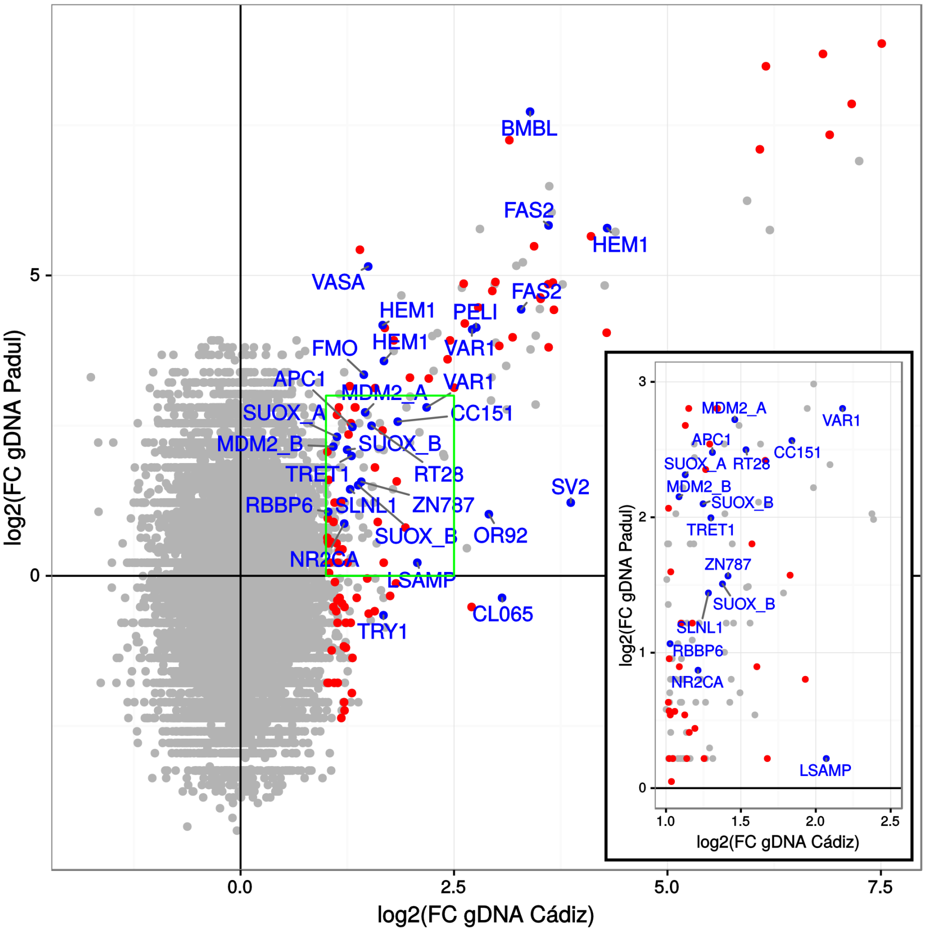


Figure 5.2: Comparison of FC in gDNA from Cádiz and gDNA from Padul. B chromosome genes are remarked as blue dots and transposable elements are indicated as red dots.

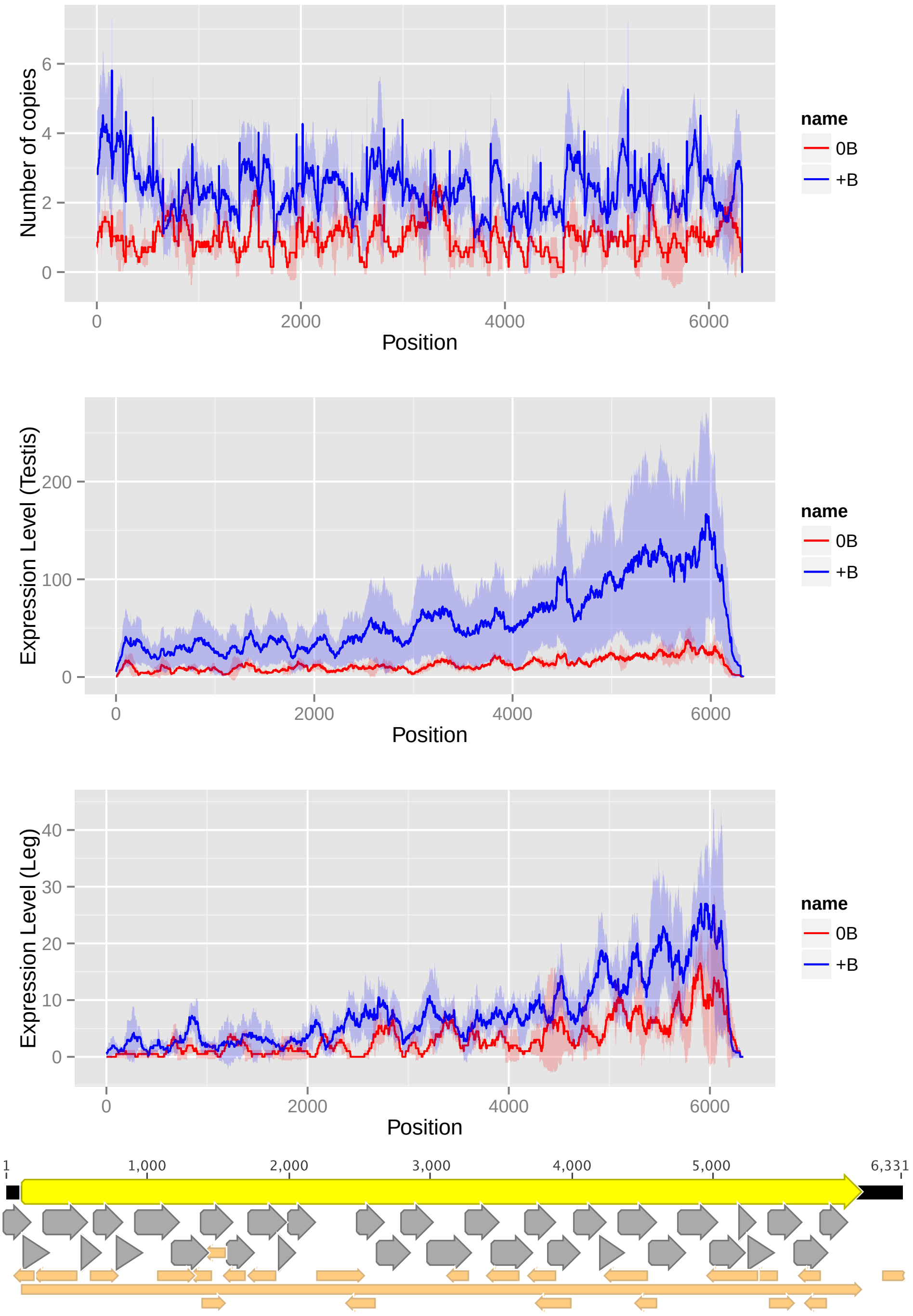


Figure 5.3: Coverage for the *APC1* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

Table 5.3: Genomic fold changes for candidate genes to be present in the B. L2FC= Log2(+B/0B). LC = Low coverage region. HC = High coverage region. UC = Uniform coverage. IC-tr = Irregular Coverage truncated. IC-ra = Irregular Coverage-regional amplification. IC-sat = Irregular Coverage-satellite.

		LC					HC				
	Distrib.	0B		+B			0B		+B		
Acronym	in B	mean	sd	mean	sd	L2FC	mean	sd	mean	sd	L2FC
PELI	UC	–	–	–	–	–	1.27	0.64	7.04	1.92	2.47
VAR1	UC	–	–	–	–	–	1.34	2.03	12.57	2.19	3.23
CC151	UC	–	–	–	–	–	0.90	0.44	3.37	1.17	1.90
MDM2_A	UC	–	–	–	–	–	0.50	0.32	1.30	0.69	1.38
FMO	UC	–	–	–	–	–	0.62	0.46	1.79	0.78	1.53
ZSC22	UC	–	–	–	–	–	1.76	1.12	3.94	1.51	1.16
SUOX_A	UC	–	–	–	–	–	1.05	0.49	2.31	0.94	1.14
APC1	UC	–	–	–	–	–	0.98	0.56	2.38	1.00	1.28
NR2CA	UC	–	–	–	–	–	0.97	1.01	1.76	1.22	0.86
SUOX_B	UC	–	–	–	–	–	0.95	0.73	2.40	0.87	1.34
MDM2_B	UC	–	–	–	–	–	0.91	0.57	1.97	0.96	1.11
RBBP6	UC	–	–	–	–	–	0.88	0.32	1.62	0.83	0.88
FAS2	IC-tr	0.81	0.56	0.95	0.61	0.23	1.16	0.68	13.32	2.27	3.52
RT28	IC-tr	1.03	0.38	1.38	0.51	0.42	0.77	0.38	5.81	1.45	2.92
VASA	IC-tr	1.13	0.59	1.09	0.57	-0.05	1.95	1.13	2.86	1.22	0.55
TRET1	IC-tr	0.44	0.40	0.89	0.62	1.02	1.31	0.56	3.20	0.96	1.29
SLNL1	IC-tr	0.93	0.66	0.94	0.63	0.02	1.02	0.38	2.47	0.94	1.28
HEM1	IC-ra	0.97	0.47	1.63	0.89	0.75	1.81	1.80	231.84	230.86	7.00
BMBL	IC-tr	0.98	0.56	3.80	1.78	1.96	2.01	0.59	36.01	7.83	4.16
OR92	IC-tr	0.85	0.55	1.09	0.62	0.36	0.32	0.48	33.44	25.62	6.71
SV2	IC-sat	1.18	0.73	1.20	0.85	0.02	2.89	2.47	51.53	39.35	4.16
CL065	IC-sat	1.42	0.93	1.16	0.83	-0.29	2.08	0.44	107.08	5.84	5.69
LSAMP	IC-sat	1.21	0.94	0.88	1.15	-0.46	2.16	0.58	29.81	1.85	3.79
TRY1	IC-sat	0.87	0.50	0.53	1.43	-0.72	0.54	0.49	12.62	0.78	4.55

research with higher coverage will surely uncover some more. Anyway, among these 24 genes, there are some promising candidates to be genes increasing B chromosome fitness in terms of success in cell divisions, which is the main arena where B chromosome destiny is decided. These 24 protein-coding genes show a wide variety of GO functions presumably reflecting those from the A chromosome region from which the B originated (see also KOG functions in Table 5.S2 for a more synthetic summary of gene functions).

B chromosome genes were over-represented in the transcriptomes

A first indication of gene expression in the B chromosome was given by the +B/0B RNA fold change (i.e. log2 of the quotient between copy num-

Table 5.4: Transcription of B-chromosome genes measured by the fold change between RPKM values in +B and 0B transcriptomes. L2FC= Log2(+B/0B). LC = Low coverage region. HC = High coverage region. UC = Uniform coverage. IC-tr = Irregular Coverage truncated. IC-ra = Irregular Coverage-regional amplification. IC-sat = Irregular Coverage-satellite.

		RNA testis						RNA leg					
	Distrib.	LC			HC			LC			HC		
Acronym	in B	0B	+B	L2FC	0B	+B	L2FC	0B	+B	L2FC	0B	+B	L2FC
PELI	UC	–	–	–	15.42	97.77	2.66	–	–	–	58.28	62.43	0.10
VAR1	UC	–	–	–	16.58	536.09	5.01	–	–	–	1.97	36.19	4.20
CC151	UC	–	–	–	112.32	973.42	3.12	–	–	–	2.64	2.68	0.02
MDM2_A	UC	–	–	–	53.16	146.78	1.47	–	–	–	7.30	19.25	1.40
FMO	UC	–	–	–	6.08	24.46	2.01	–	–	–	8.19	13.45	0.72
ZSC22	UC	–	–	–	72.54	347.47	2.26	–	–	–	9.85	28.11	1.51
SUOX_A	UC	–	–	–	33.30	130.98	1.98	–	–	–	20.09	75.20	1.90
APC1	UC	–	–	–	12.68	59.07	2.22	–	–	–	3.32	7.65	1.20
NR2CA	UC	–	–	–	47.74	45.30	-0.08	–	–	–	6.77	7.31	0.11
SUOX_B	UC	–	–	–	25.09	44.60	0.83	–	–	–	12.77	12.54	-0.03
MDM2_B	UC	–	–	–	33.10	85.71	1.37	–	–	–	3.41	8.91	1.39
RBBP6	UC	–	–	–	0.00	4.70	–	–	–	–	4.46	10.58	1.25
FAS2	IC-tr	47.91	30.02	-0.67	92.65	735.98	2.99	1.26	1.62	0.36	12.78	86.24	2.75
RT28	IC-tr	63.99	88.58	0.47	31.19	53.72	0.78	51.20	55.02	0.10	25.09	29.65	0.24
VASA	IC-tr	59.63	127.17	1.09	90.61	333.23	1.88	35.67	14.27	-1.32	49.51	160.89	1.70
TRET1	IC-tr	20.34	19.53	-0.06	13.17	33.99	1.37	0.00	0.43	–	0.98	1.18	0.27
SLNL1	IC-tr	6.40	15.44	1.27	8.93	15.49	0.79	0.00	0.01	–	0.30	15.49	5.69
HEM1	IC-ra	10.98	20.42	0.90	126.19	442.44	1.81	65.96	74.19	0.17	425.88	40.72	-3.39
BMBL	IC-tr	30.65	48.41	0.66	442.58	744.17	0.75	24.30	8.72	-1.48	409.47	120.97	-1.76
OR92	IC-tr	0.00	1.96	–	0.00	2.65	–	0.19	0.00	–	0.00	0.00	–
SV2	IC-sat	13.14	6.09	-1.11	40.74	17.20	-1.24	0.15	0.02	-2.91	0.73	0.00	–
CL065	IC-sat	244.41	318.58	0.38	199.39	273.89	0.46	90.66	79.39	-0.19	71.97	68.68	-0.07
LSAMP	IC-sat	1.46	0.80	-0.87	5.73	4.55	-0.33	4.43	2.29	-0.95	3.44	0.00	–
TRY1	IC-sat	0.85	1.02	0.26	0.94	2.53	1.43	0.00	0.00	–	0.00	0.00	–

bers in +B and 0B libraries) observed in the testis and hind leg transcriptomes from Cádiz. This showed that most contigs being over-represented in the B-carrying gDNA libraries were also over-represented in the transcriptomes of the same individuals (Fig. 5.4 and 5.S1). In fact, 16 and 14 out of the 24 genes showed expression fold change higher than 1 in testis and leg, respectively (Table 5.4). Remarkably, 10 of these genes showed upregulation in testis (*PELI*, *VAR1*, *CC151*, *MDM2_A*, *FMO*, *ZSC22*, *SUOX_A*, *APC1*, *MDM2_B*, *RBBP6*), but only seven of them showed upregulation in hind leg (*VAR1*, *MDM2_A*, *ZSC22*, *SUOX_A*, *APC1*, *MDM2_B*, *RBBP6*) (see Fig. 5.3 and 5.S2-5.S12).

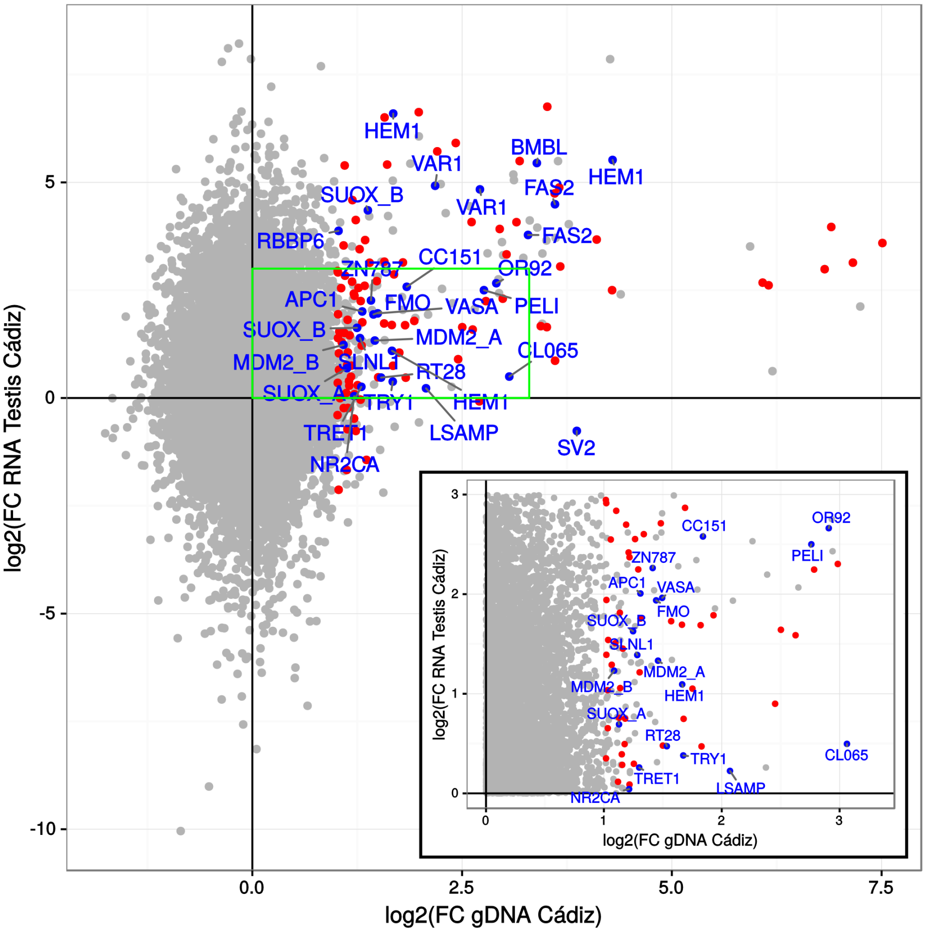


Figure 5.4: Comparison of FC in gDNA and RNA from testis libraries.B chromosome genes are remarked as blue dots and transposable elements are indicated as red dots.

Sequence variations specific to B chromosome genes

To be sure that the observed gene up-regulations are due to the expression of the B chromosome gene copies, and not to up-regulation of A chromosome ones, we searched for DNA sequence variations being specific to B chromosomes, by comparing the 0B and +B libraries from the Cádiz population. Specifically, we searched for nucleotide differences in the B-carrying libraries which were completely absent in the 0B gDNA and RNA libraries and showed a frequency in the B-carrying gDNA being 25% or higher. For this purpose, we first mapped 0B and +B gDNA reads against each gene sequence to find all nucleotide sites per gene differing in 0B and +B and showing two or more read counts in the +B gDNA but zero counts in the 0B gDNA. This rendered 1,279 nucleotide positions showing variation in the 24 genes as a whole. We then eliminated those positions showing one or more read counts in the 0B RNA of testis and hind leg, which left 1,058 positions, 645 of which were in CDSs whereas the remaining 413 were in UTR regions.

At each position, we considered as reference (Ref) the nucleotide being present in the 0B, and alternative (Alt) being present only in +B gDNA or RNA. Next we applied another filter to select those Alt sequence variations showing a frequency higher than 25% in the B-carrying genomes. This high threshold was chosen to increase the likelihood of managing nucleotide variation specific to the B chromosome, and was based on the fact that we had B-carrying libraries from four individuals. We finally found 208 single nucleotide polymorphisms (SNPs) showing high representation in the B-carrying gDNAs, so that we could reasonably believe that the Alt variant resided in the B chromosome. They were found in the following genes: *PELI* (4 SNPs), *VAR1* (3), *FMO* (2), *ZSC22* (32), *SUOX_A* (30) *APC1* (12), *NRC2A* (1), *SUOX_B* (3), *MDM2_B* (1), RBBP6 (3), FAS2 (25), RT28 (2), VASA (16), TRET1 (6), *SLNL1* (26), *HEM1* (8), *BMBL* (1), *OR92* (6), *SV2* (14), *TRY1* (8) and *LSAMP* (5). The high number of SNPs in the B chromosome genes suggests that B chromosomes in *L. migratoria* are quite old and their DNA sequence has been independently evolving for long from that of the A chromosome copies. In fact, divergence per nucleotide site (ps) was significantly lower in the 12 genes showing uniform coverage (thus being putatively functional) than in the 12 showing irregular coverage (thus being putatively pseudogenic) (Mann-Whitney test: P= 0.0018).

We also compared the divergence between the A and B chromosome gene sequences using as external reference the transcriptome sequences in *O. decorus*. Bearing in mind that *L. migratoria* and *O. decorus* shared their most recent common ancestor about 22.81 Mya (Song et al. 2015), we calculated separately the divergence accumulated in the A chromosome

Table 5.5: Estimation of B chromosome age from sequence divergence between A and B chromosomes, using *O. decorus* as outgroup. L2FC= Log2(+B/0B). LC = Low coverage region. HC = High coverage region.

Genes with uniform coverage				Genes with irregular coverage			
	Divergence		Time		Divergence		Time
Gene	0B/+B	0B/Odec	(mya)	Gene	0B/+B	0B/Odec	(mya)
PELI	0.0066	0.0577	1.48	*FAS2*	0.0085	0.0705	1.51
VAR1	0.0150	0.0689	2.26	*RT28*	0.0074	0.0460	2.17
CC151	0.0030	0.0364	1.04	*VASA*	0.0211	0.0682	4.82
MDM2_A	0.0020	–	–	*TRET1*	0.0071	0.0497	1.91
FMO	0.0030	0.0407	0.74	*SLNL1*	0.0179	0.0831	1.67
ZSC22	0.0079	0.0900	2.15	*HEM1*	0.0366	0.0602	3.26
SUOX_A	0.0193	–	–	*BMBL*	0.0537	0.1223	6.99
APC1	0.0058	0.0584	1.09	*OR92*	0.1553	–	–
NR2CA	0.0071	0.0500	0.81	*SV2*	0.0355	0.2361	6.59
SUOX_B	0.0041	0.0503	0.94	*CL065*	0.0385	0.0897	7.98
MDM2_B	0.0075	0.0999	0.42	*LSAMP*	0.0914	–	–
RBBP6	0.0149	–	–	*TRY1*	0.0127	0.0823	1.32
Average	–	–	**1.21**	**Average**	–	–	**3.82**
SD	–	–	**0.63**	**SD**	–	–	**2.56**

and B chromosome lineages, the latter being almost two orders of magnitude higher (0.000258±0.010268 and 0.015482±0.018353, respectively, for mean and SD) (Wilcoxon test: P= 0.0003) (see Table 5.5). This suggests that most changes in these genes have taken place in the B chromosome sequences, thus giving support to Teruel at al. (2010) assumption to calculate B chromosome age. We also got an estimate of B chromosome age from each of the 19 genes found in *O. decorus*, by the expression [(A_B divergence)*0.5*22.81]/(0B_*L. migratoria* vs. *O. decorus* divergence). Estimates were very different between genes (Mean= 2.59 my, SD= 2.29), and were clearly lower for the nine genes showing uniform coverage pattern (1.21±0.63) than for the ten genes showing irregular coverage patterns (3.82±2.56) (Mann-Whitney test: P= 0.0043).

Activity of B chromosome genes

To analyze the activity of B chromosome genes, we searched for SNPs being located, in each gene, at distances lower than Illumina read length (i.e. <100 nt). This allowed defining 2-4 haplotypes of closely linked SNPs for each of 14 genes and analyzing read counts in gDNA and RNA of B-carrying individuals (Table 5.S3). We calculated genomic haplotype frequency as the proportion of read counts found in gDNA of +B individuals, and transcriptomic haplotype frequency as that found in RNA of the same individuals. We then calculated an estimate of the intensity of gene expression (ExI) as the quotient between RNA and gDNA read counts for each haplotype, and them analyzed the consistency between the different haplotype regions scored within each gene. As Table 5.S3 shows, there was a generally high consistency between haplotype regions within genes, with 12 genes (*PELI*, *ZSC22*, *SUOXA*, *APC1*, *RPPB6*, *FAS2*, *TRET1*, *SLNL1*, *HEM1*, *SV2*, LSAMP and TRY1) showing evidence for active transcription of the B chromosome copies in the testis transcriptome, half of which (*RPPB6*, *TRET1*, *SLNL1*, *HEM1*, *SV2* and *TRY1*) failed to show expression in the hind leg, and only two genes (*VASA* and *OR92*) failing to show evidence for transcription of the B chromosome copies (Table 5.S1).

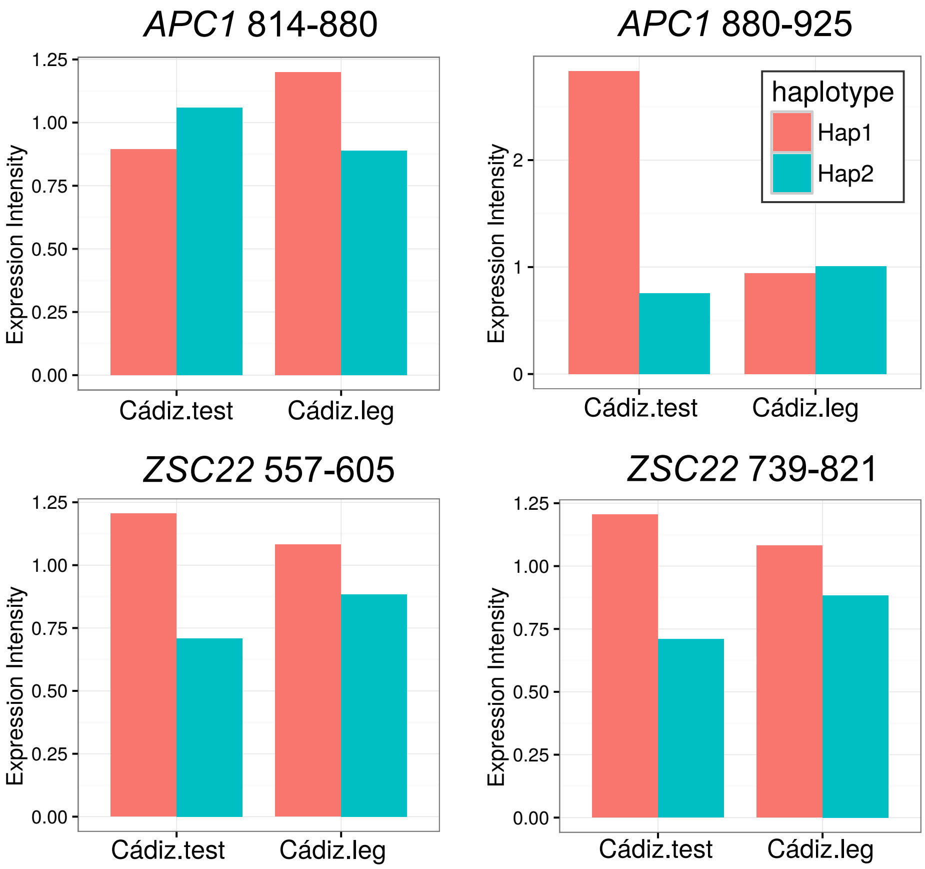


Figure 5.5: Expression Intensity (ExI) for pairs of B-specific SNPs in *APC1* and *ZSC22* genes. All the SNPs are in the CDS of the gene with exception of the nucleotide 821 in the *ZSC22* gene which is in the 3′-UTR. Note that haplotype 2 was exclusive of B-carrying individuals.

Discussion

Our results have shown that the B chromosome of *L. migratoria* contains at least 24 protein-coding genes, 12 of which appear to be complete and 10 of these are active (*PELI*, *VAR1*, *CC151*, *MDM2_A*, *FMO*, *ZSC22*, *SUOX_A*, *APC1*, *MDM2_B* and *RBBP6*), half of which were confirmed by the analysis of B-specific linked SNPs. In addition, four B chromosome genes showing irregular coverage, thus being putatively pseudogenic (*VASA*, *TRET1* and *SLNL1* and *HEM1*), were also active. These results corroborate recent findings of active protein-coding genes in the B chromosomes of the Siberian roe deer (Trifonov *et al.*, 2013) and the grasshopper *E. plorans* (Navarro-Dominguez *et al.*, submitted), and strongly run against the idea that B chromosomes are genetically inert. On the contrary, the present results unveil the character of B chromosomes as true intragenomic parasites being able to yield transcripts for some protein-coding genes thus establishing a transcriptomic cross-talk with the standard (host) chromosomes. Although many of the B chromosome genes in *L. migratoria* are pseudogenized, in consistency with previous findings in rye (Banaei-Moghaddam *et al.*, 2013), fish (Valente *et al.*, 2014) and the grasshopper *E. plorans* (Navarro-Dominguez *et al.*, submitted), almost half of them appeared to be complete and active in the migratory locust. Remarkably, some of them might provide interesting functions for a parasitic B chromosome (see below).

The high variation found among the 24 B chromosome genes in their divergence from the A chromosome copies (see Table 1.5) can partly be due to differences in functional constraint, especially for genes being complete in the B chromosome and showing transcriptional activity. Even higher variation in divergence was shown by the genes presumably being pseudogenized, perhaps due to differences in degeneration time, with divergence being lower for recently degenerated genes and higher for genes being pseudogenized for long. These high differences between genes make it difficult to get accurate estimates of B chromosome age, but allow delimiting an approximate age for the B chromosome in *L. migratoria* between 1 and 4 my. The lower limit is very close to the 0.75 my calculated by Teruel *et al.* (2010) by comparing DNA sequence of H3 and H4 histone genes between A and B chromosomes, and thus corroborates that these B chromosomes are quite old, in resemblance to the 2 my age of B chromosomes in maize (Lamb *et al.*, 2007).

During their long stay within the *L. migratoria* genome, B chromosomes have spread across natural populations in Asia (Itoh, 1934; Hsiang, 1958; Nur, 1969; Kayano, 1971), Africa (Dearn, 1974), Australia (King & John, 1980) and Europe (Cabrero *et al.*, 1984). The secret for such an extraordinary evolutionary success is unveiled by our present results. B chromosomes in

L. migratoria contain active protein-coding genes serving functions potentially being very useful for prospering by manipulating gene expression in the host genome. Of course, the most interesting genes are those being apparently complete and active in the B chromosome (*PELI*, *VAR1*, *CC151*, *MDM2_A*, *FMO*, *ZSC22*, *SUOX_A*, *APC1*, *MDM2_B* and *RBBP6*).

The most remarkable gene is, no doubt, *APC1* which codes for the largest subunit of the Anaphase Promoting Complex or Cyclosome (APC/C), a cell cycle E3 ubiquitin ligase regulating the metaphase-anaphase transition during mitosis (Jörgensen *et al.*, 2001; Peters, 2002, 2006), but it also plays a role during meiosis (Harper *et al.*, 2002; Barford, 2011). During mitosis, APC/C activation is tightly controlled by the spindle assembly checkpoint (SAC) which, in presence of kinetochores being unattached to microtubules, generates the mitotic checkpoint complex (MCC) which inhibits APC/C activation until all chromosomes are properly aligned to the metaphase plate (Kaisari *et al.*, 2016; Wild *et al.*, 2016). When this condition is met, APC/C is activated and anaphase begins. Remarkably, testis transcriptomes of B carrying males in *L. migratoria* showed upregulation of the APC1 gene, and there was about twice B-specific transcripts as A-specific ones for this gene (see Table 5.S3), reflecting their higher abundance in the B-carrying genomes (Table 5.1). This excess of *APC1* transcripts, if translated and yields functional protein, might lower MCC surveillance for kinetochores unattached to microtubules and this might facilitate mitotic non-disjuntion of the two B chromatids to a same pole, when only one of the two B sister kinetochores is attached to microtubules. Mitotic non-disjuction of B chromosomes in *L. migratoria* was cytologically visualized in embryos (Pardo *et al.*, 1995).

In grasshoppers, meiotic resumption of oocytes occurs upon fertilization during egg laying Henriques-Gil *et al.* (1986). Therefore, the second drive mechanism in *L. migratoria*, which takes place through preferential migration of the B chromosome to the viable pole during the first or second meiotic divisions of oogenesis (Pardo *et al.*, 1994), operates while the *APC1* gene of the B chromosome might have a chance to manipulate the course of the cell division where B chromosome destiny is into question, an interesting prospect for future research.

Other interesting genes are *MDM2_A* and *MDM2_B*, which are two sequence variants of the Murine Double Minute 2 (*MDM2*) protein which is an E3 ubiquitin ligase being the main negative regulator of the tumor suppressor p53. The latter protein is called the "guardian of genome" because of its ability for preserving cell genomic integrity under stressed conditions, by regulating genes involved in DNA repair, metabolism, cell cycle arrest, apoptosis and senescence (Urso *et al.*, 2016). In fact, the *MDM2* gene is overexpressed in many cancers (see Urso *et al.*, 2016). The presence of active *MDM2* gene copies in the B chromosomes of *L. migratoria* might indulge B-carrying cells showing B-caused genomic instability of entering into the apoptotic pathway, due to p53 repression by MDM2 expression in the B chromosome. Of course, this would only be possible if the B transcripts were translated to yield functional MDM2 protein.

Remarkably, the B chromosome carries another gene coding for E3 ubiquitin ligase, i.e. retinoblastoma binding protein 6 (RBBP6). It plays a role as a scaffold protein promoting the assembly of the p53/TP53-MDM2 complex, resulting in increase of MDM2-mediated ubiquitination and degradation of p53/TP53 (Chibi *et al.*, 2008).

The meaning of the remaining active B chromosome genes is less apparent, as it is very clear that they code for functions being useful for B chromosomes. For instance, PELI codes for a scaffold protein involved in the Toll signaling pathway, VAR1 and SUOX_A code for mitochondrial proteins, CC151 is a ciliary protein required for motile cilia function, FMO catalyzes glucosinolate reactions and ZSC22 codes for the zinc finger and SCAN domain-containing protein 22 and seems to be involved in transcriptional regulation. Anyway, the transcripts of these genes could influence the expression of genes sharing the same gene expression networks.

We believe that B chromosome genes could have derived, as a joint block, from a same A chromosome region, but only a few genes were actually crucial for B chromosome invasion and maintenance (*APC1*, *MDM2* and *RBBP6*), the three genes remarkably coding for E3 ubiquitin ligases. Linkage analysis for the B chromosome genes in B-lacking individuals might unveil whether B chromosome genes are actually linked in the A genome, and this could allow inferring which A chromosome was its ancestor. The expected progression towards a more complete version of the *L. migratoria* genome will probably help to infer B chromosome ancestry on the basis of its gene content.

B chromosome genes appear to show higher activity in the testis than the hind leg, as only half of genes showing activity in testis are actually active in leg. This is consistent with the idea that B chromosomes, as vertically transmitted parasites, are expected to be more aggressive during transmission (i.e. on the germ line) and less harmful on somatic tissues (to be less damaging to the host) (Muñoz *et al.*, 1998).

Taken together, the finding of active genes serving B chromosome interest represents a very important change in the evolutionary meaning of B chromosomes, since they cannot be anymore considered genetically inert elements but, on the contrary, they are true parasites, at intragenomic level, which elicit a true transcriptional arms race with the host A chromosomes.

The fact that the few B chromosomes where gene content is being envisaged carry active genes with functions related with cell division, i.e. the main arena determining B chromosome fitness, suggests that B chromosomes being established for long in natural populations are well equipped with gene contents granting the drive mechanisms needed for initial invasion (see Camacho *et al.*, 1997). Recent research has shown the presence of several genes coding for functions related with chromosome segregation in the fish *Astatotilapia latifasciata* (Valente *et al.*, 2014), and the presence of B-specific transcript variants for three of the genes in a relative species where B chromosomes have not been visualized (*Pundamilia nyererei*). In maize, four protein-coding genes were visualized in the B chromosome by FISH (Huang *et al.*, 2016), and in the grasshopper *E. plorans* (Navarro-Dominguez *et al.*, submitted), five protein-coding genes were found to be located in the B chromosome and up-regulated in B-carrying individuals. Three genes were truncated and showed upregulation only for the gene region being present in the B chromosome, thus clearly showing evidence for B chromosome transcriptional activity. Most remarkably, two complete genes (*CIP2A* and *KIF20A*) coded for functions being highly relevant for B chromosome transmission, such as chromosome segregation and stability, cytokinesis and microtubule assembly at metaphase-anaphase transition. Our present results in *L. migratoria* strongly support the idea that B chromosomes are not genetically inert, and we provide here evidence for the active expression of B chromosome gene variants, based on individual SNPs and haplotypes of linked SNPs. Remarkably, B-carrying individuals showed similar expression intensity for the A and B variants in most genes. Three of these active genes (*APC1*, *MDM2* and *RBBP6*) are important regulatory devices in transcriptional nets which can serve B chromosome interest in manipulating host genome expression to lowering defenses at crucial stages, such as the p53 regulation pathway or the mitotic checkpoint, rendering host genomes being more permissive with B chromosome nondisjunction or with genome instability caused by its presence or activity. As anticipated by Darlington & Upcott (1941), "new extra chromosomes appear from time to time in many species, but most of them come to nothing". The above results suggest the possibility that only those being genetically well equipped become into true B chromosomes.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology*, **215**, 403–410.

Banaei-Moghaddam AM, Martis MM, Macas J, *et al.* (2015) Genes on B chromosomes: old questions revisited with new tools. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, **1849**, 64–70.

Banaei-Moghaddam AM, Meier K, Karimi-Ashtiyani R, Houben A (2013) Formation and expression of pseudogenes on the B chromosome of rye. *The Plant Cell*, **25**, 2536–2544.

Barford D (2011) Structural insights into anaphase-promoting complex function and mechanism. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **366**, 3605–3624.

Battaglia E (1964) Cytogenetics of B-chromosomes. *Caryologia*, **17**, 245–299.

Boeckmann B, Bairoch A, Apweiler R, *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research*, **31**, 365–370.

Cabrero J, Viseras E, Camacho JPM (1984) The B-chromosomes of *Locusta migratoria* I. Detection of negative correlation between mean chiasma frequency and the rate of accumulation of the B's; a reanalysis of the available data about the transmission of these B-chromosomes. *Genetica*, **64**, 155–164.

Camacho JPM, López-León MD, Pardo MC, Cabrero J, Shaw MW (1997) Population dynamics of a selfish B chromosome neutralized by the standard genome in the grasshopper *Eyprepocnemis plorans*. *The American Naturalist*, **149**, 1030–1050.

Camacho JPM, Sharbel TF, Beukeboom LW (2000) B-chromosome evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **355**, 163–178.

Carchilan M, Kumke K, Mikolajewski S, Houben A (2009) Rye B chromosomes are weakly transcribed and might alter the transcriptional activity of A chromosome sequences. *Chromosoma*, **118**, 607–616.

Chibi M, Meyer M, Skepu A, Rees DJG, Moolman-Smook JC, Pugh DJ (2008) RBBP6 interacts with multifunctional protein YB-1 through its RING finger domain, leading to ubiquitination and proteosomal degradation of YB-1. *Journal of molecular biology*, **384**, 908–916.

Darlington CD, Upcott MB (1941) The activity of inert chromosomes in *Zea Mays*. *Journal of Genetics*, **41**, 275–296.

Dearn JM (1974) Phase transformation and chiasma frequency variation in locusts. *Chromosoma*, **45**, 321–338.

Graphodatsky AS, Kukekova AV, Yudkin DV, *et al.* (2005) The proto-oncogene C-KIT maps to canid B-chromosomes. *Chromosome Research*, **13**, 113–122.

Haas BJ, Papanicolaou A, Yassour M, *et al.* (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, **8**, 1494–1512.

Harper JW, Burton JL, Solomon MJ (2002) The anaphase-promoting complex: it's not just for mitosis any more. *Genes & development*, **16**, 2179–2206.

Harvey AW, Hewitt GM (1979) B chromosomes slow development in a grasshopper. *Heredity*, **42**, 397–401.

Henriques-Gil N, Jones GH, Cano MI, Arana P, Santos JL (1986) Female meiosis during oocyte maturation in *Eyprepocnemis plorans* (Orthoptera: Acrididae). *Canadian journal of genetics and cytology*, **28**, 84–87.

Hsiang W (1958) Cytological studies on migratory locust hybrid, *Locusta migratoria migratoria* L. *Locusta migratoria manilensis* Meyen. *Acta Zoologica Sinica*, **1**, 006.

Huang W, Du Y, Zhao X, Jin W (2016) B chromosome contains active genes and impacts the transcription of A chromosomes in maize (*Zea mays* L.). *BMC Plant Biology*, **16**, 1.

Itoh H (1934) Chromosomal variation in the spermatogenesis of a grasshopper, *Locusta danica*. *Jap. J. Genet*, **10**, 115–134.

Jones RN (1995) Tansley review no. 85. B chromosomes in plants. *New Phytologist*, pp. 411–434.

Jones RN, Rees H (1982) *B Chromosomes*. Academic Press, New York.

Jörgensen PM, Gräslund S, Betz R, St\a ahl S, Larsson C, Höög C (2001) Characterisation of the human APC1, the largest subunit of the anaphase-promoting complex. *Gene*, **262**, 51–59.

Kaisari S, Sitry-Shevah D, Miniowitz-Shemtov S, Hershko A (2016) Intermediates in the assembly of mitotic checkpoint complexes and their role in the regulation of the anaphase-promoting complex. *Proceedings of the National Academy of Sciences*, p. 201524551.

Kayano H (1971) Accumulation of B chromosomes in the germ line of *Locusta migratoria*. *Heredity*.

King M, John B (1980) Regularities and restrictions governing C-band variation in acridoid grasshoppers. *Chromosoma*, **76**, 123–150.

Lamb JC, Riddle NC, Cheng YM, Theuri J, Birchler JA (2007) Localization and transcription of a retrotransposon-derived element on the maize B chromosome. *Chromosome research*, **15**, 383–398.

Leach CR, Houben A, Field B, Pistrick K, Demidov D, Timmis JN (2005) Molecular evidence for transcription of genes on a B chromosome in *Crepis capillaris*. *Genetics*, **171**, 269–278.

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Lin HZ, Lin WD, Lin CY, Peng SF, Cheng YM (2014) Characterization of maize B-chromosome-related transcripts isolated via cDNA-AFLP. *Chromosoma*, **123**, 597–607.

Martis MM, Klemme S, Banaei-Moghaddam AM, *et al.* (2012) Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proceedings of the National Academy of Sciences*, **109**, 13343–13346.

Muñoz E, Perfectti F, Martín-Alganza Á, Camacho JPM (1998) Parallel effects of a B chromosome and a mite that decrease female fitness in the grasshopper *Eyprepocnemis plorans*. *Proceedings of the Royal Society of London B: Biological Sciences*, **265**, 1903–1909.

Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. *Genome research*, **11**, 1725–1729.

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792–793.

Nur U (1969) Mitotic instability leading to an accumulation of B-chromosomes in grasshoppers. *Chromosoma*, **27**, 1–19.

Östergren G (1945) Parasitic nature of extra fragment chromosomes. *Botaniska Notiser*, **2**, 157–163.

Pardo MC, López-León MD, Cabrero J, Camacho JPM (1994) Transmission analysis of mitotically unstable B chromosomes in *Locusta migratoria*. *Genome*, **37**, 1027–1034.

Pardo MC, López-León MD, Viseras E, Cabrero J, Camacho JPM (1995) Mitotic instability of B chromosomes during embryo development in *Locusta migratoria*. *Heredity*, **74**, 164–169.

Peters JM (2002) The anaphase-promoting complex: proteolysis in mitosis and beyond. *Molecular cell*, **9**, 931–943.

Peters JM (2006) The anaphase promoting complex/cyclosome: a machine designed to destroy. *Nature reviews Molecular cell biology*, **7**, 644–656.

Randolph LF (1941) Genetic characteristics of the B chromosomes in maize. *Genetics*, **26**, 608–631.

Rhoades MM, McClintock B (1935) The cytogenetics of maize. *The Botanical Review*, **1**, 292–325.

Ruiz-Estévez M, Badisco L, Broeck JV, *et al.* (2014) B chromosomes showing active ribosomal RNA genes contribute insignificant amounts of rRNA in the grasshopper *Eyprepocnemis plorans*. *Molecular genetics and genomics*, **289**, 1209–1216.

Ruiz-Estevez M, Lopez-Leon MD, Cabrero J, Camacho JPM (2012) B-chromosome ribosomal DNA is functional in the grasshopper *Eyprepocnemis plorans*. *PLoS One*, **7**, e36600.

Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, **6**, 31.

Song H, Amédégnato C, Cigliano MM, *et al.* (2015) 300 million years of diversification: elucidating the patterns of orthopteran evolution based on comprehensive taxon and gene sampling. *Cladistics*, **31**, 621–651.

Tanić N, Vujošević M, Dedović-Tanić N, Dimitrijević B (2005) Differential gene expression in yellow-necked mice *Apodemus flavicollis* (Rodentia, Mammalia) with and without B chromosomes. *Chromosoma*, **113**, 418–427.

Teruel M, Cabrero J, Perfectti F, Camacho JPM (2010) B chromosome ancestry revealed by histone genes in the migratory locust. *Chromosoma*, **119**, 217–225.

Trifonov VA, Dementyeva PV, Larkin DM, *et al.* (2013) Transcription of a protein-coding gene on B chromosomes of the Siberian roe deer (*Capreolus pygargus*). *BMC biology*, **11**, 1.

Urso L, Calabrese F, Favaretto A, Conte P, Pasello G (2016) Critical review about MDM2 in cancer: Possible role in malignant mesothelioma and implications for treatment. *Critical Reviews in Oncology/Hematology*, **97**, 220–230.

Valente GT, Conte MA, Fantinatti BE, *et al.* (2014) Origin and evolution of B chromosomes in the cichlid fish *Astatotilapia latifasciata* based on integrated genomic analyses. *Molecular biology and evolution*, p. msu148.

Wang X, Fang X, Yang P, *et al.* (2014) The locust genome provides insight into swarm formation and long-distance flight. *Nature communications*, **5**.

Wild T, Larsen MSY, Narita T, Schou J, Nilsson J, Choudhary C (2016) The Spindle Assembly Checkpoint Is Not Essential for Viability of Human Cells with Genetically Lowered APC/C Activity. *Cell reports*, **14**, 1829–1840.

Wilson EB (1907) The supernumerary chromosomes of Hemiptera. *Science*, **26**, 870–871.

Supplementary Information

Supplementary Figures

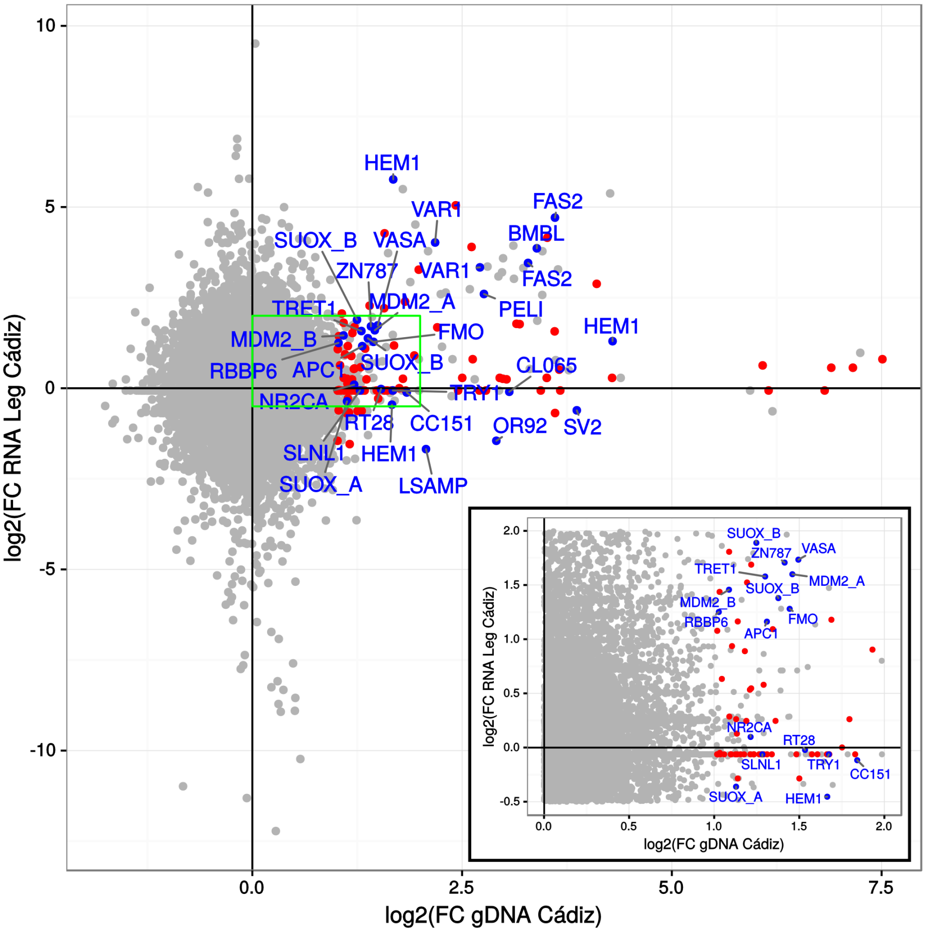


Figure 5.S1: Comparison of FC in gDNA and RNA from leg libraries. B chromosome genes are remarked as blue dots and transposable elements are indicated as red dots.

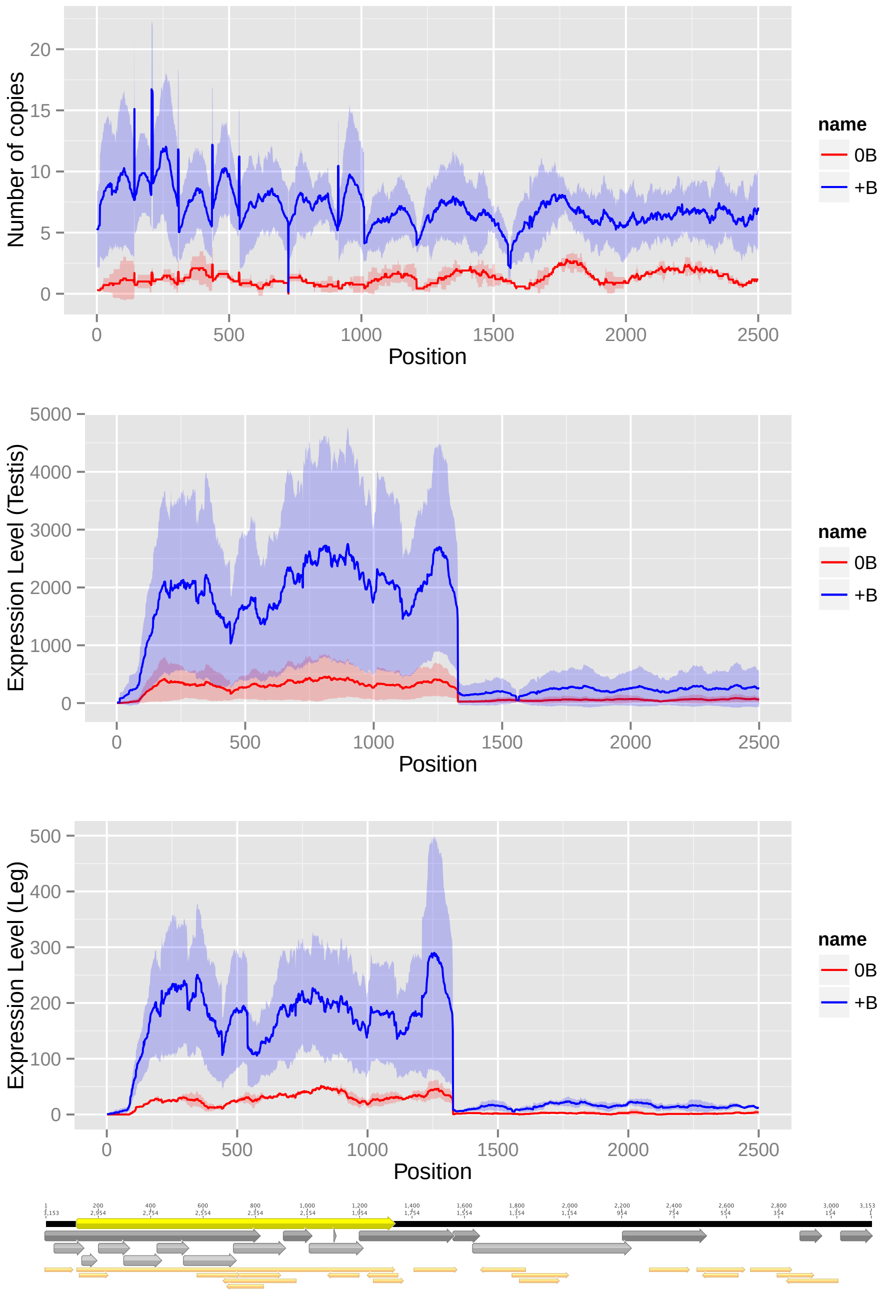


Figure 5.S2: Coverage for the *PELI* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

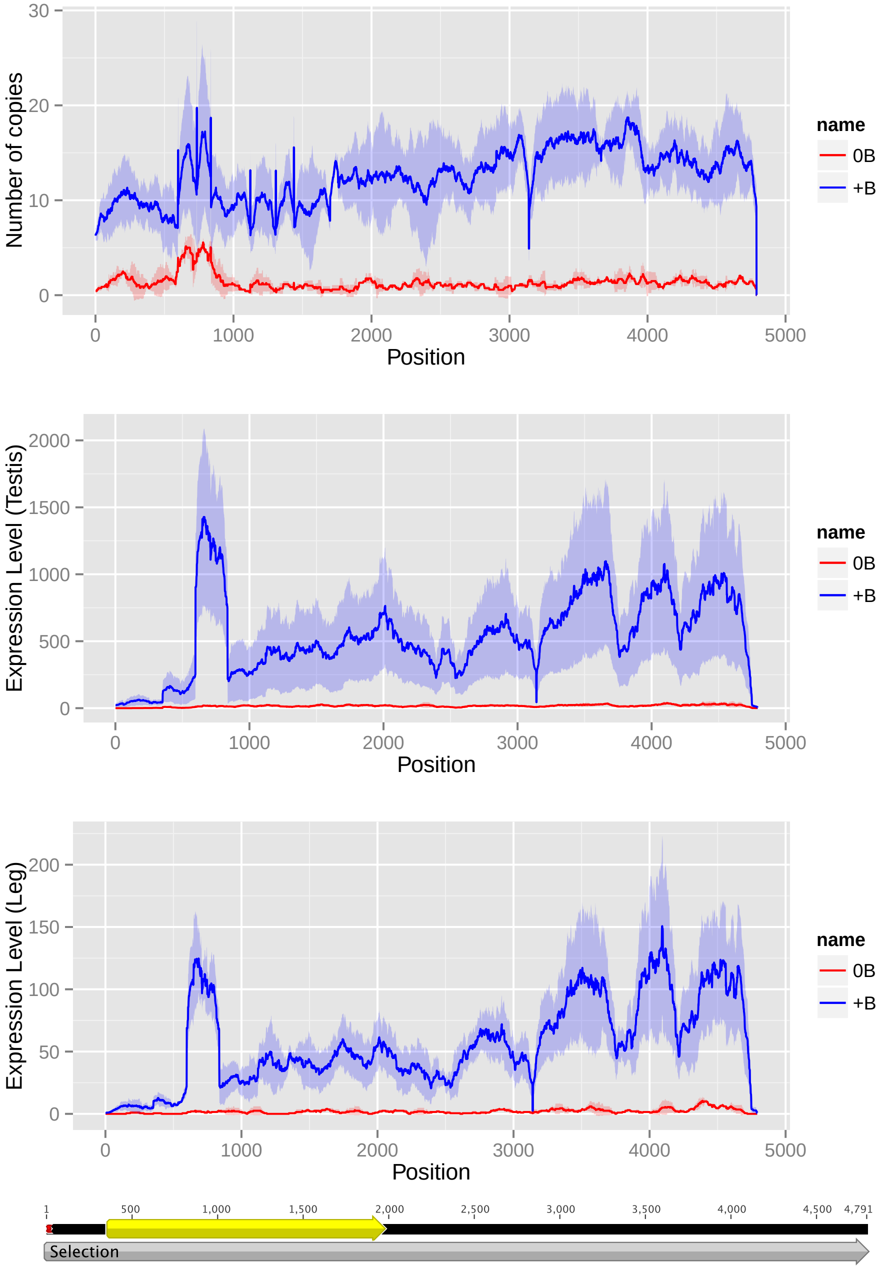


Figure 5.S3: Coverage for the *VAR1* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

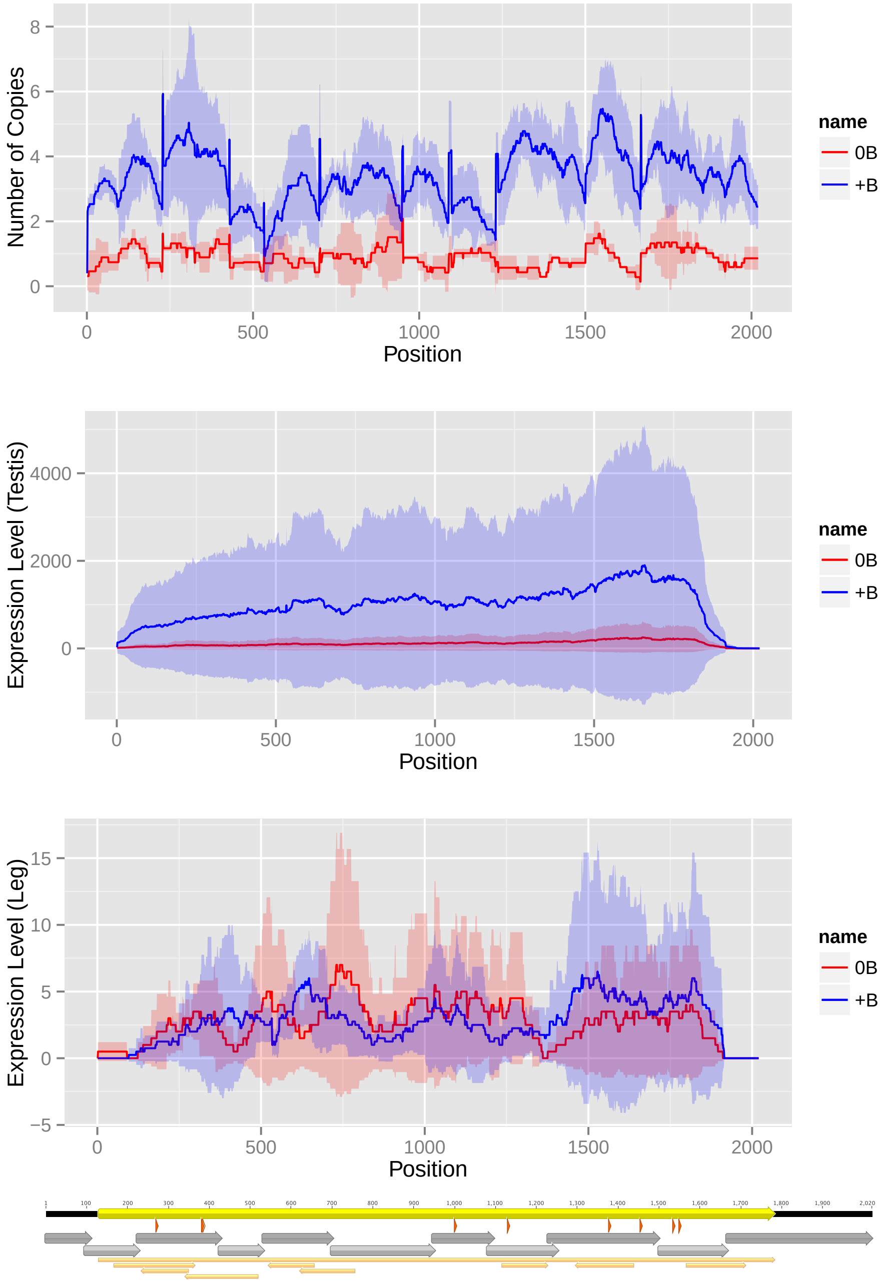


Figure 5.S4: Coverage for the *CC151* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

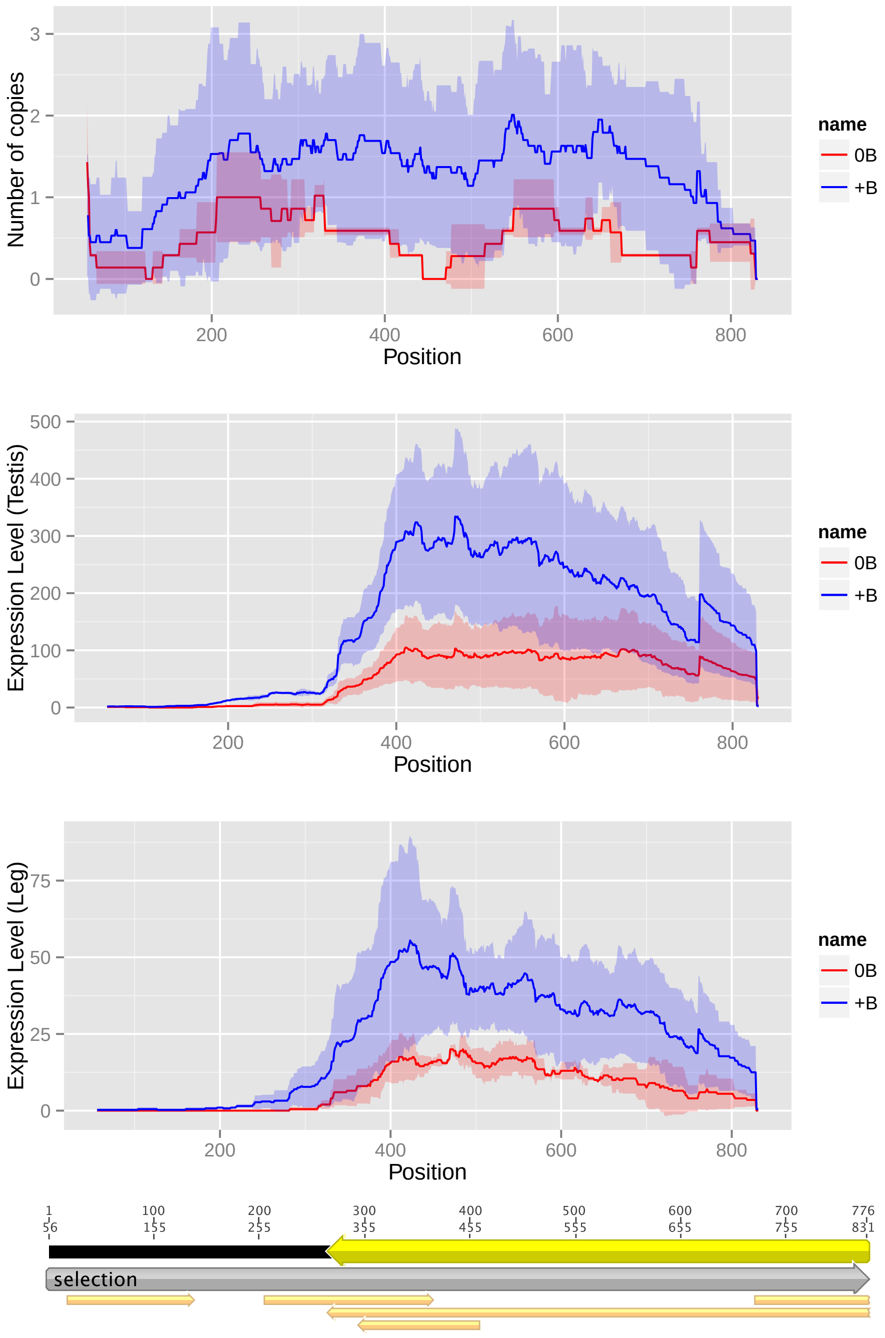


Figure 5.S5: Coverage for the *MDM2_A* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

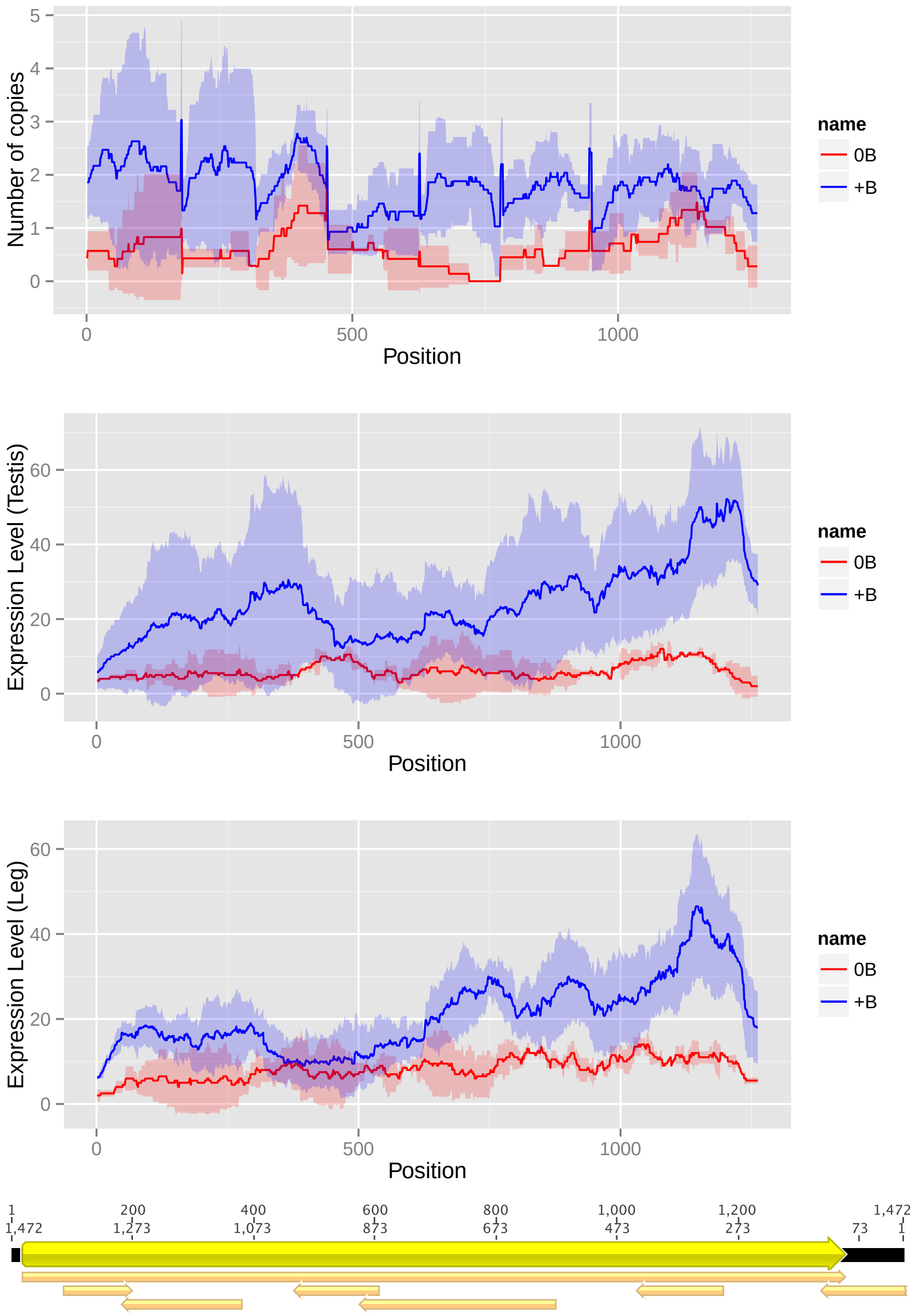


Figure 5.S6: Coverage for the *FMO* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

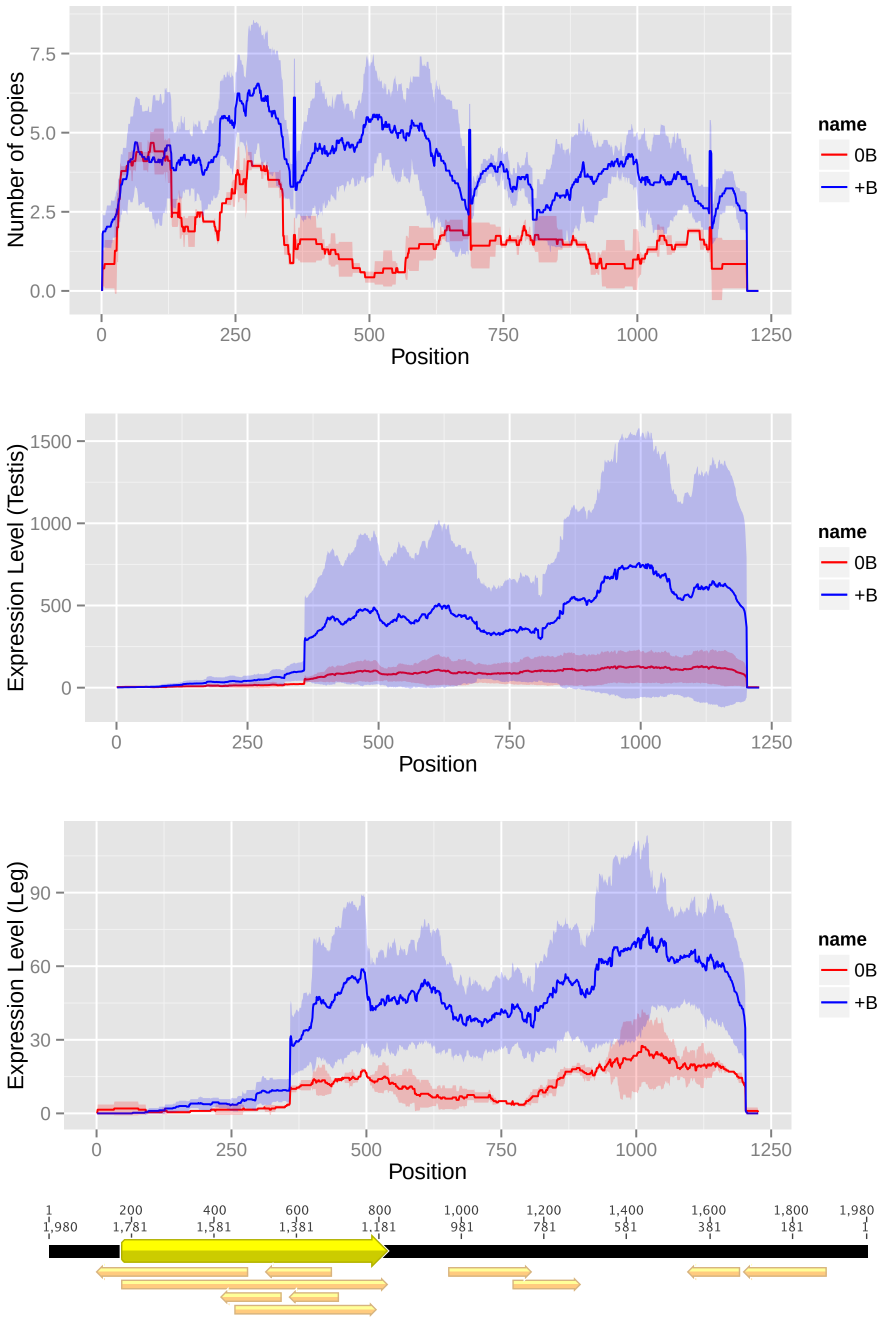


Figure 5.S7: Coverage for the *ZSC22* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

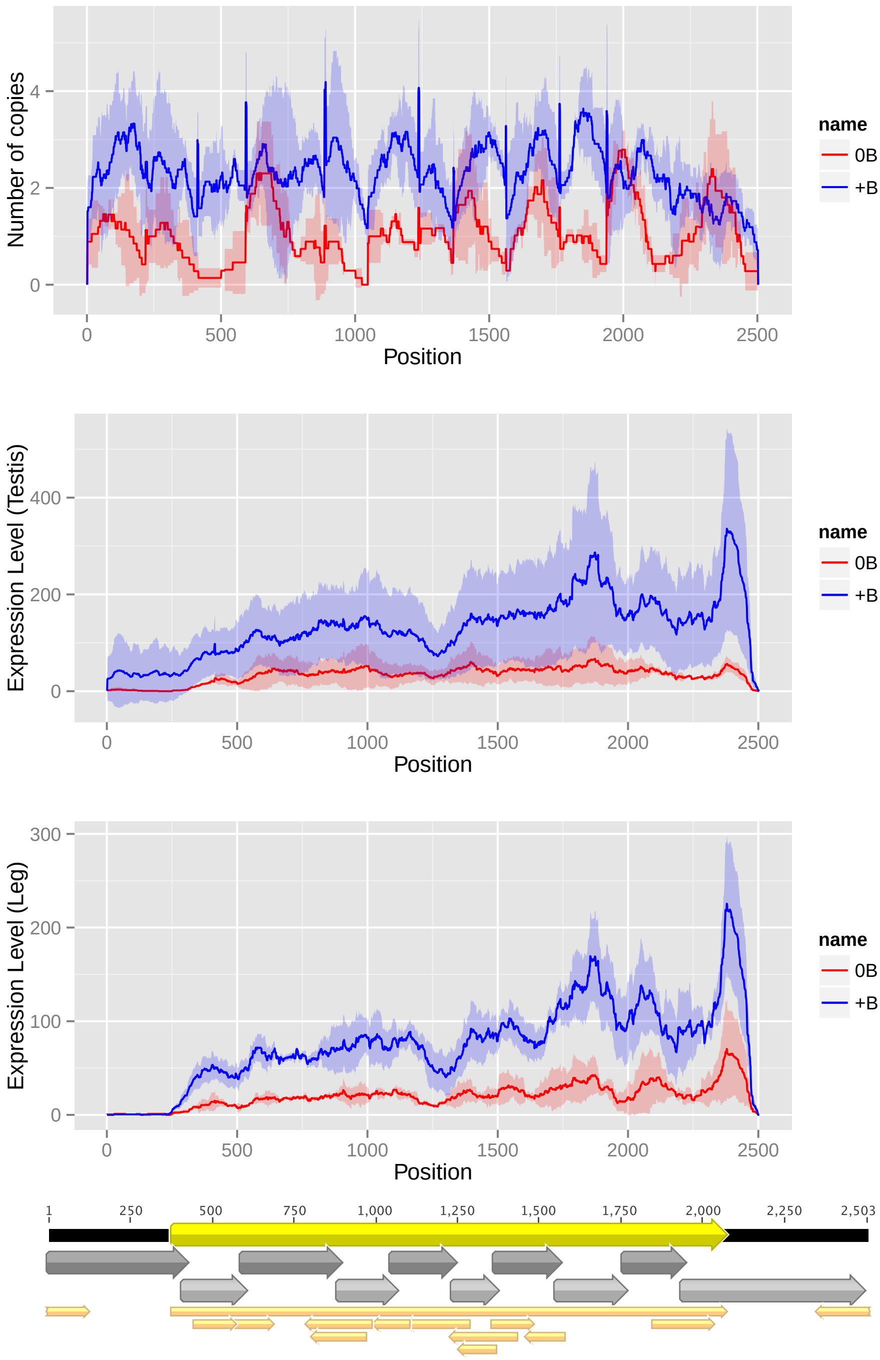


Figure 5.S8: Coverage for the *SUOX_A* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

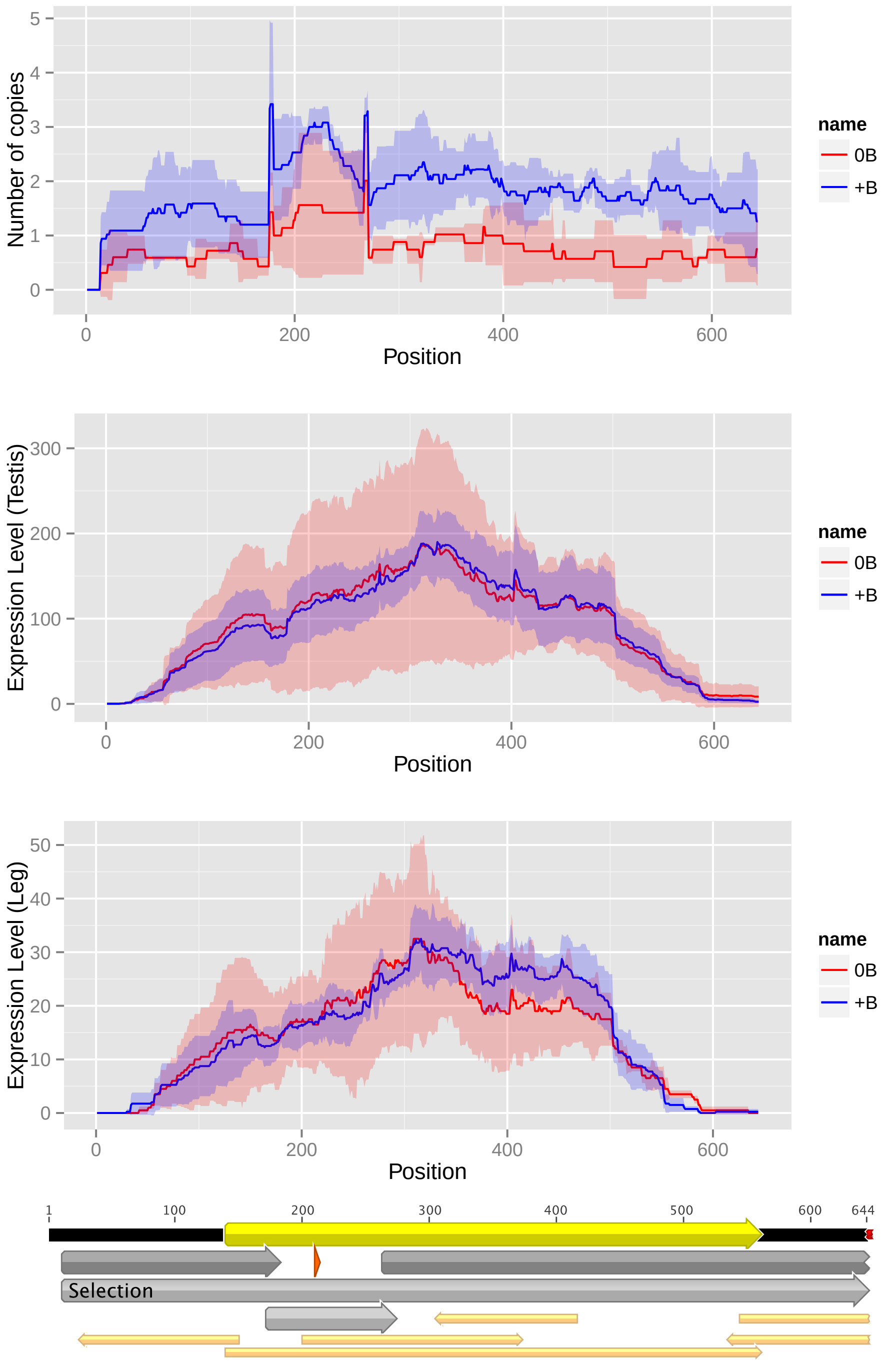


Figure 5.S9: Coverage for the *NR2CA* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals. NR2CA.

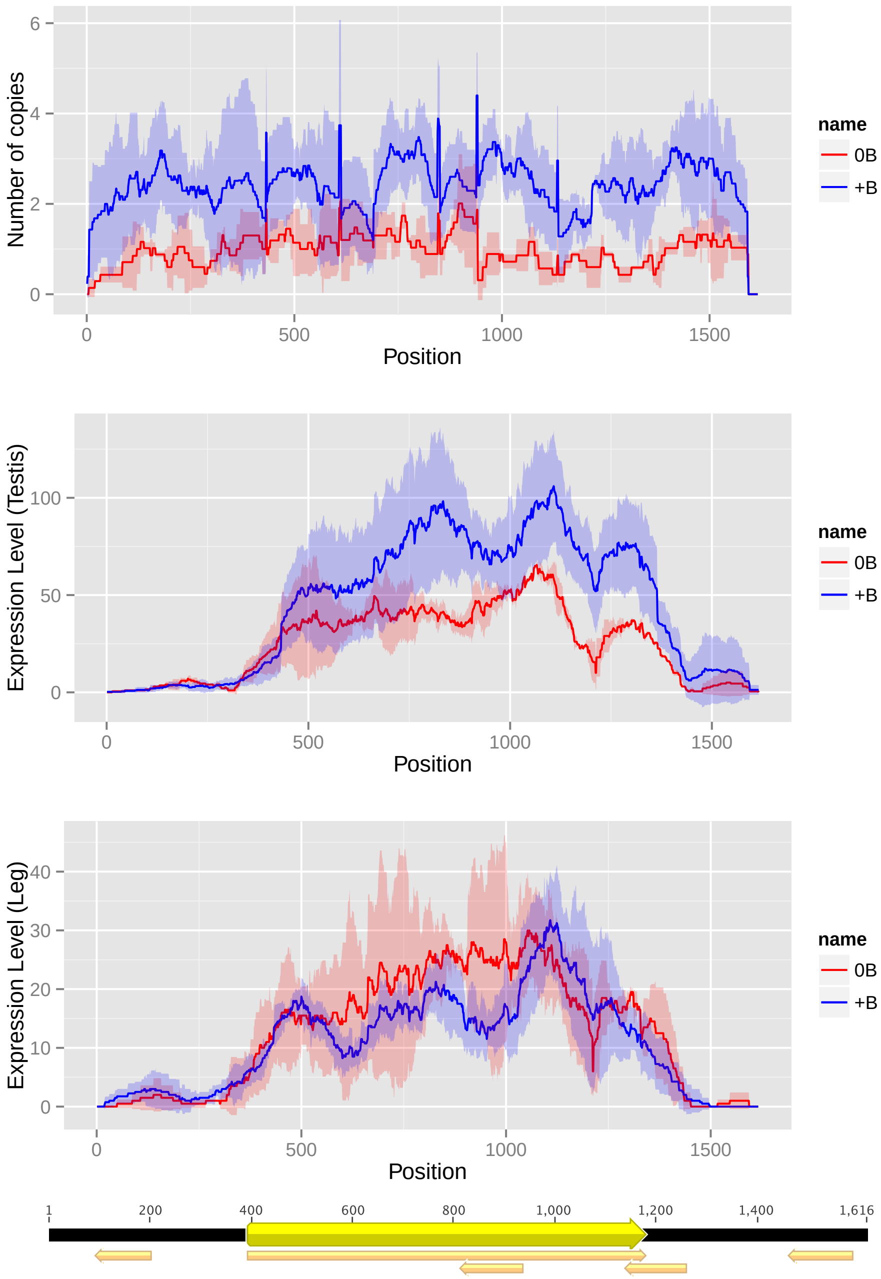


Figure 5.S10: Coverage for the *SUOX_B* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals. SUOX_B.

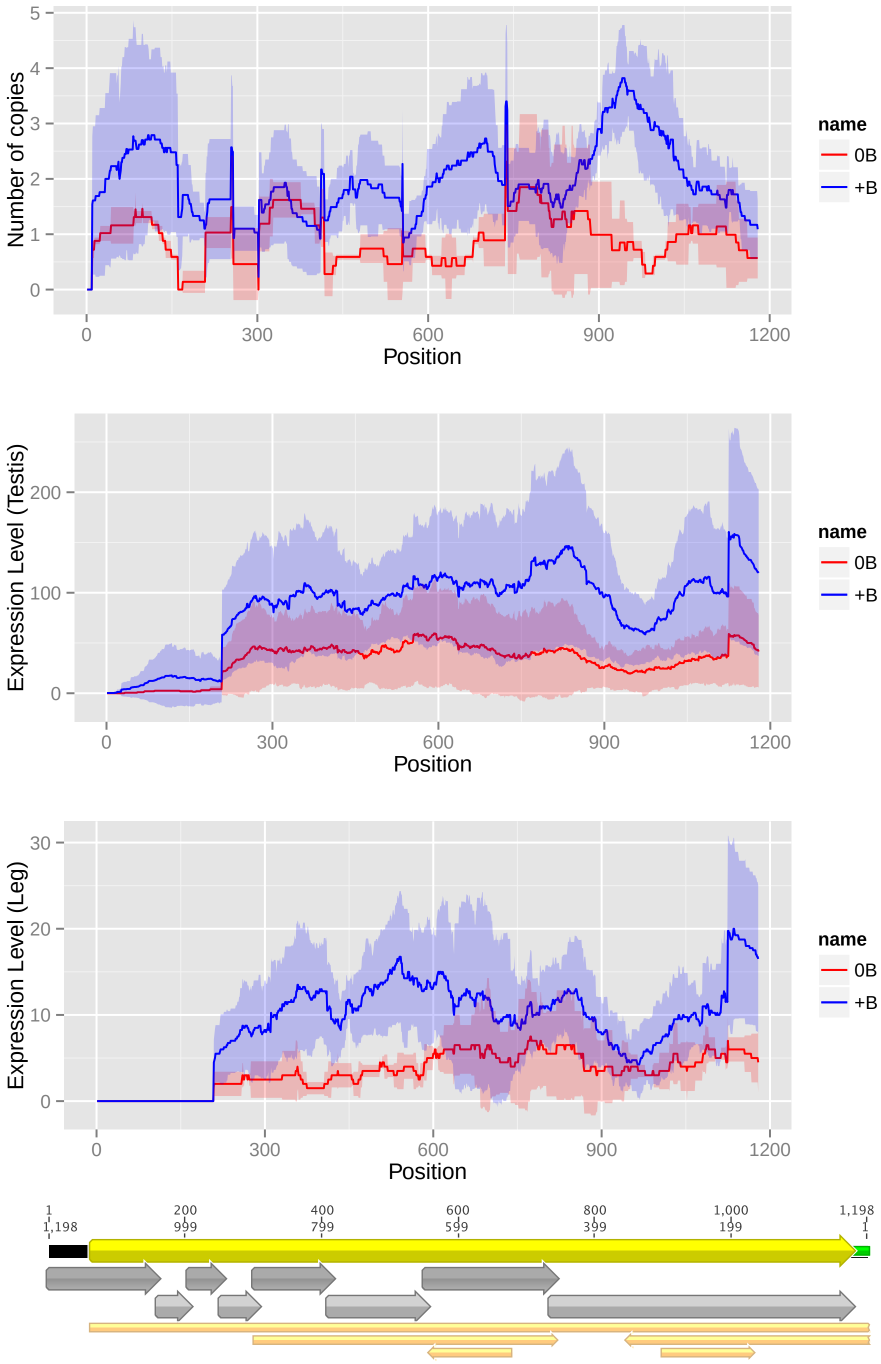


Figure 5.S11: Coverage for the *MDM2_B* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

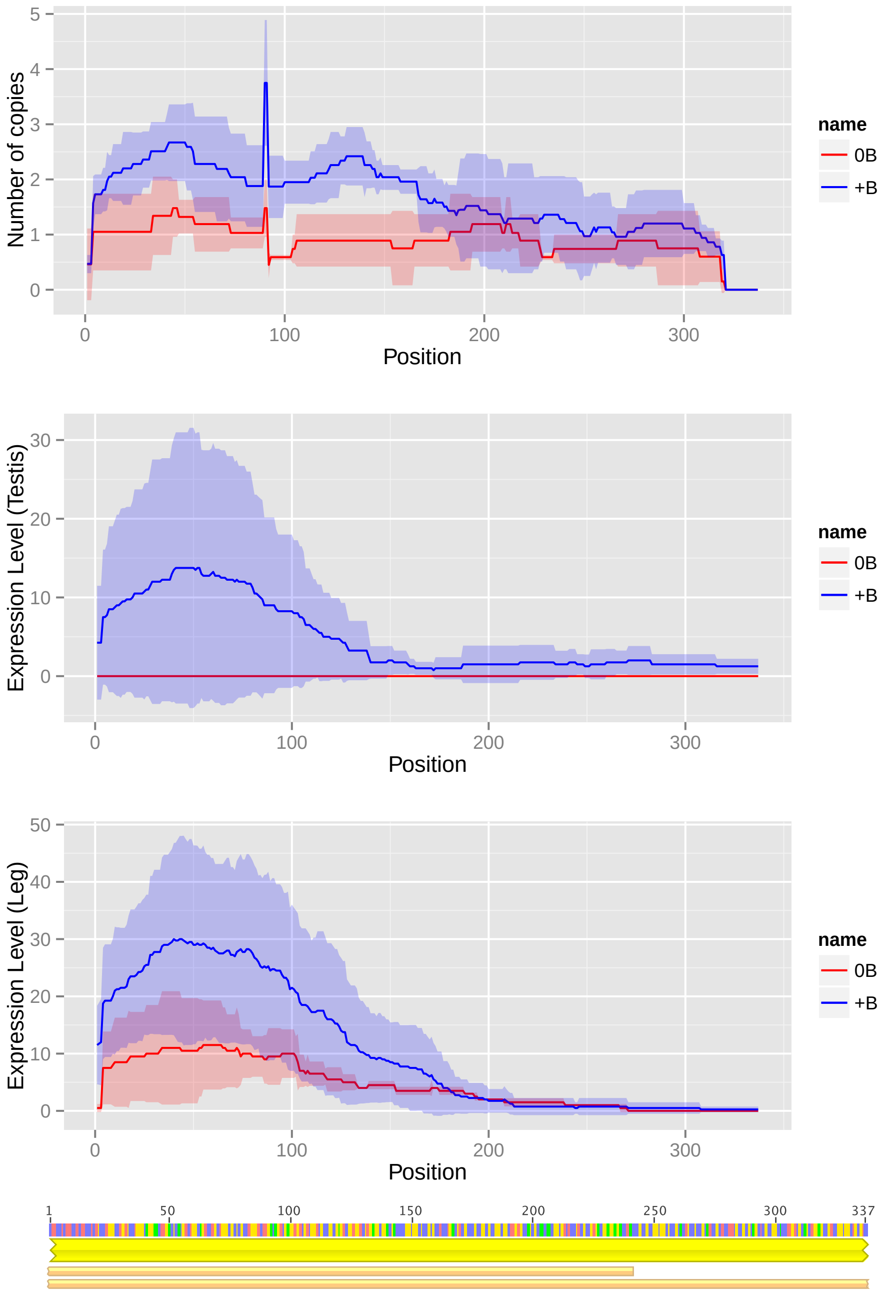


Figure 5.S12: Coverage for the *RPPB6* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

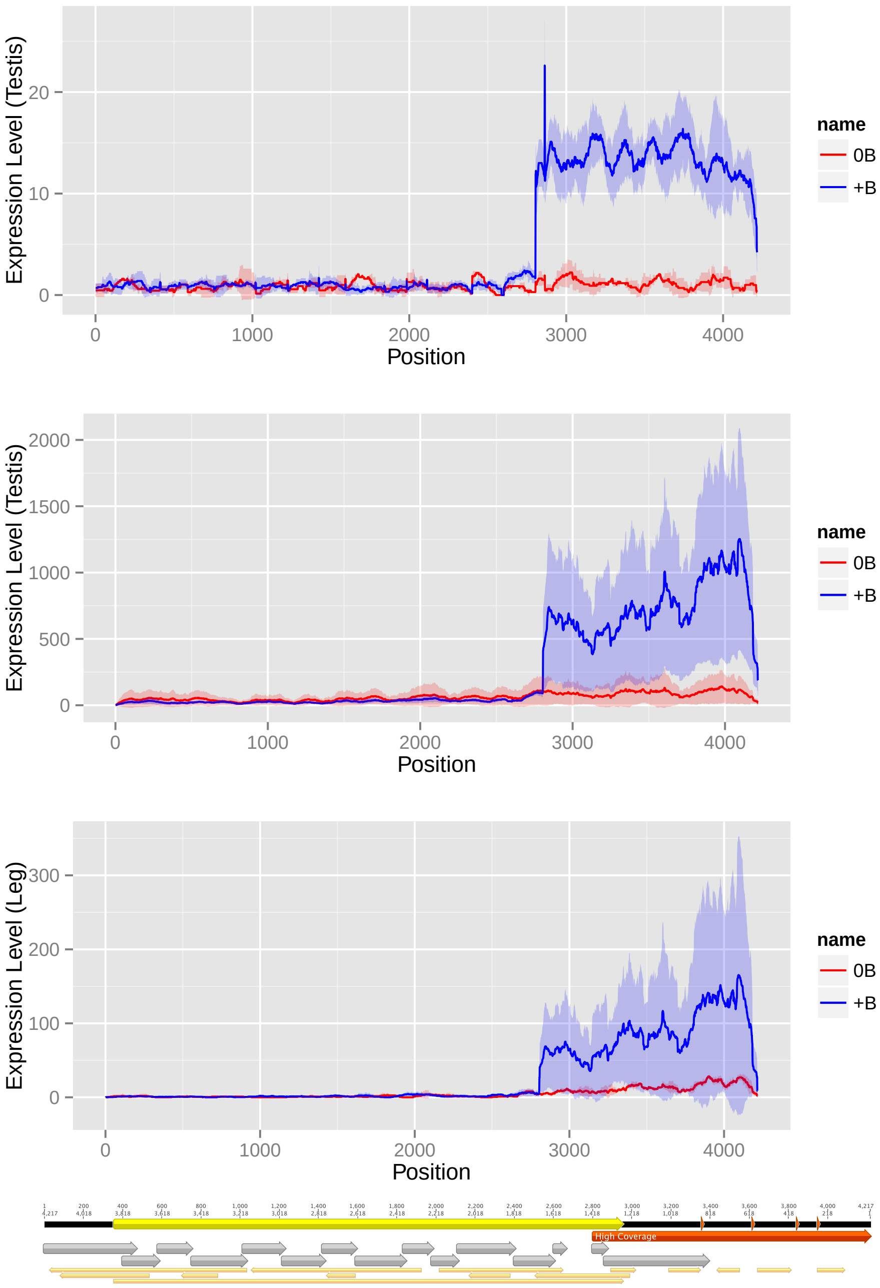


Figure 5.S13: Coverage for the *FAS2* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

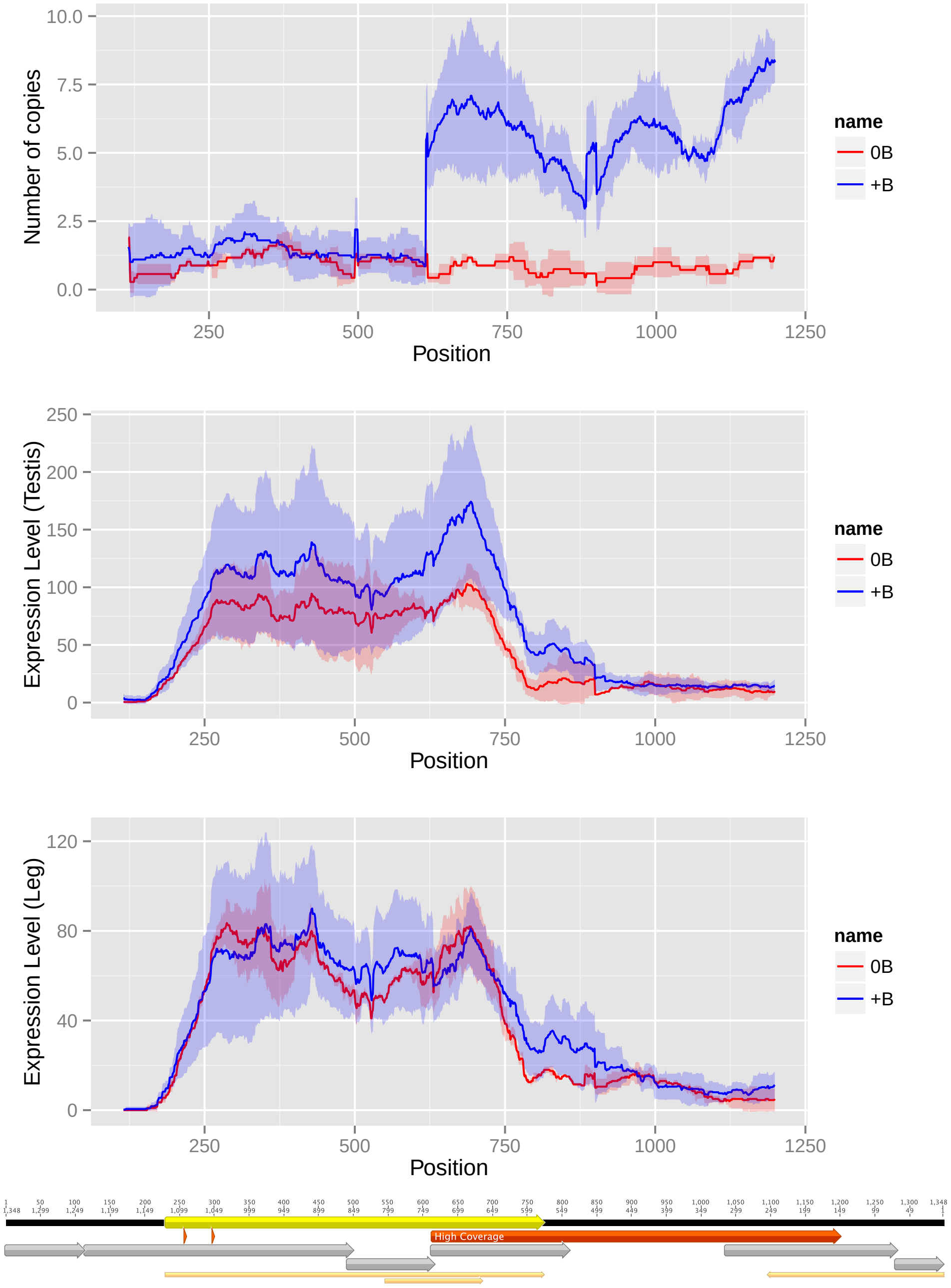


Figure 5.S14: Coverage for the *RT28* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

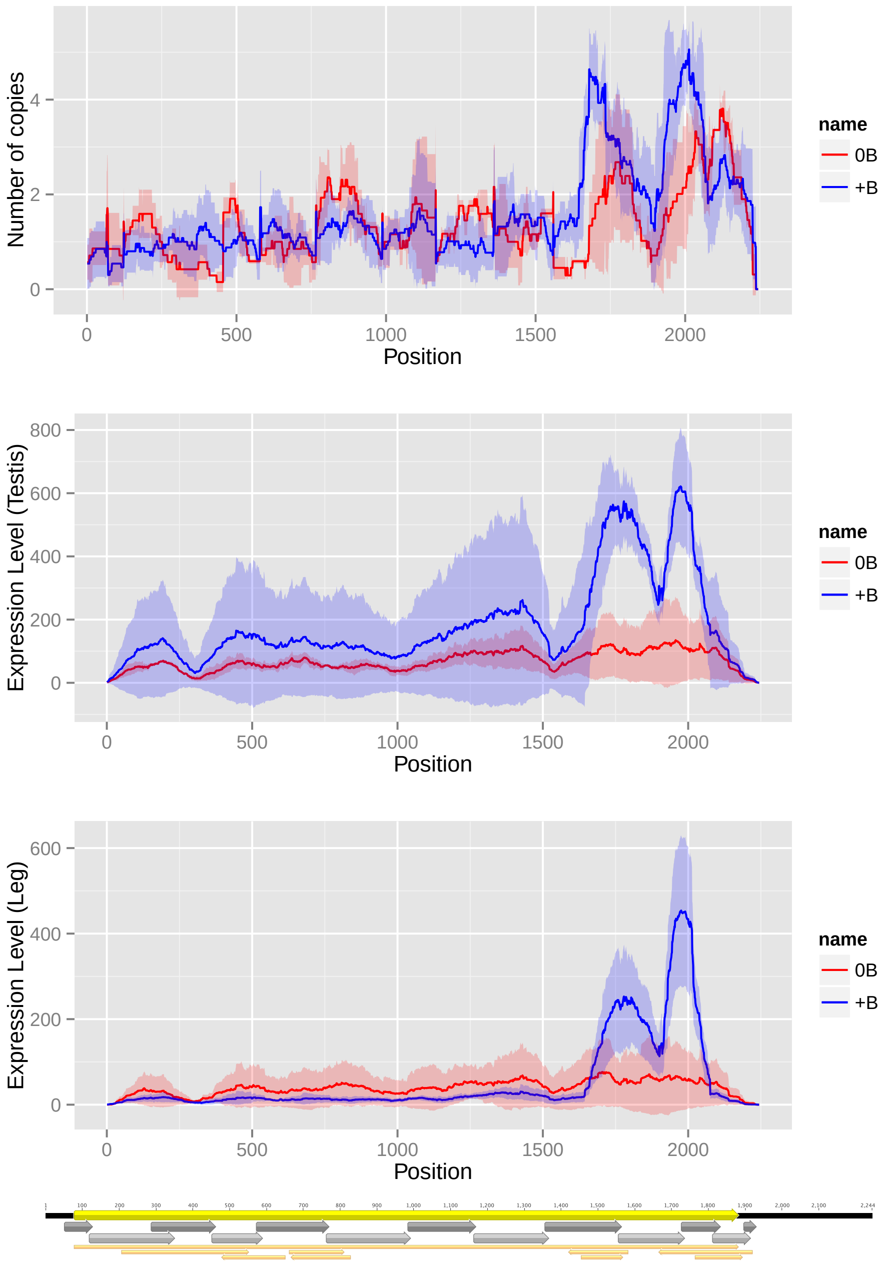


Figure 5.S15: Coverage for the *VASA* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

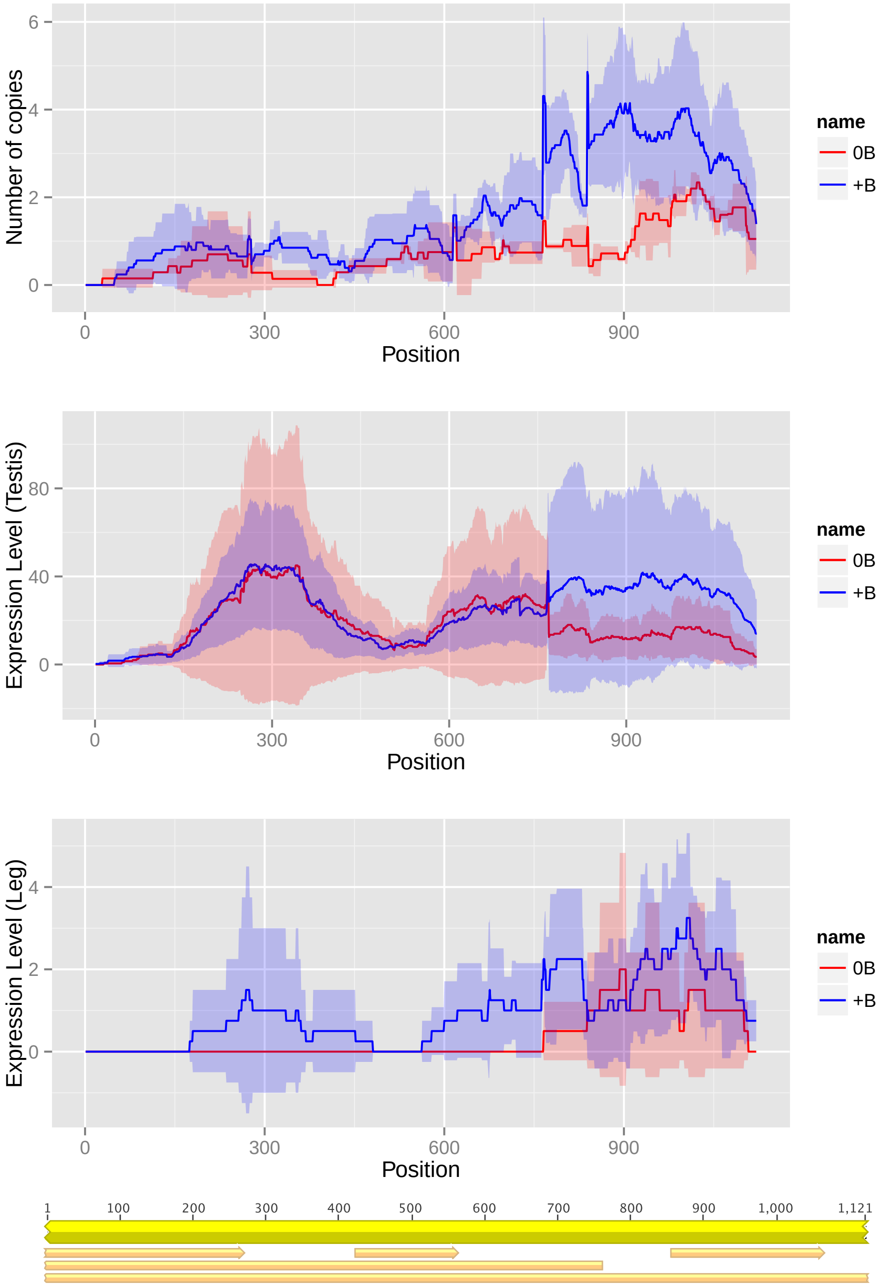


Figure 5.S16: Coverage for the *TRET1* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

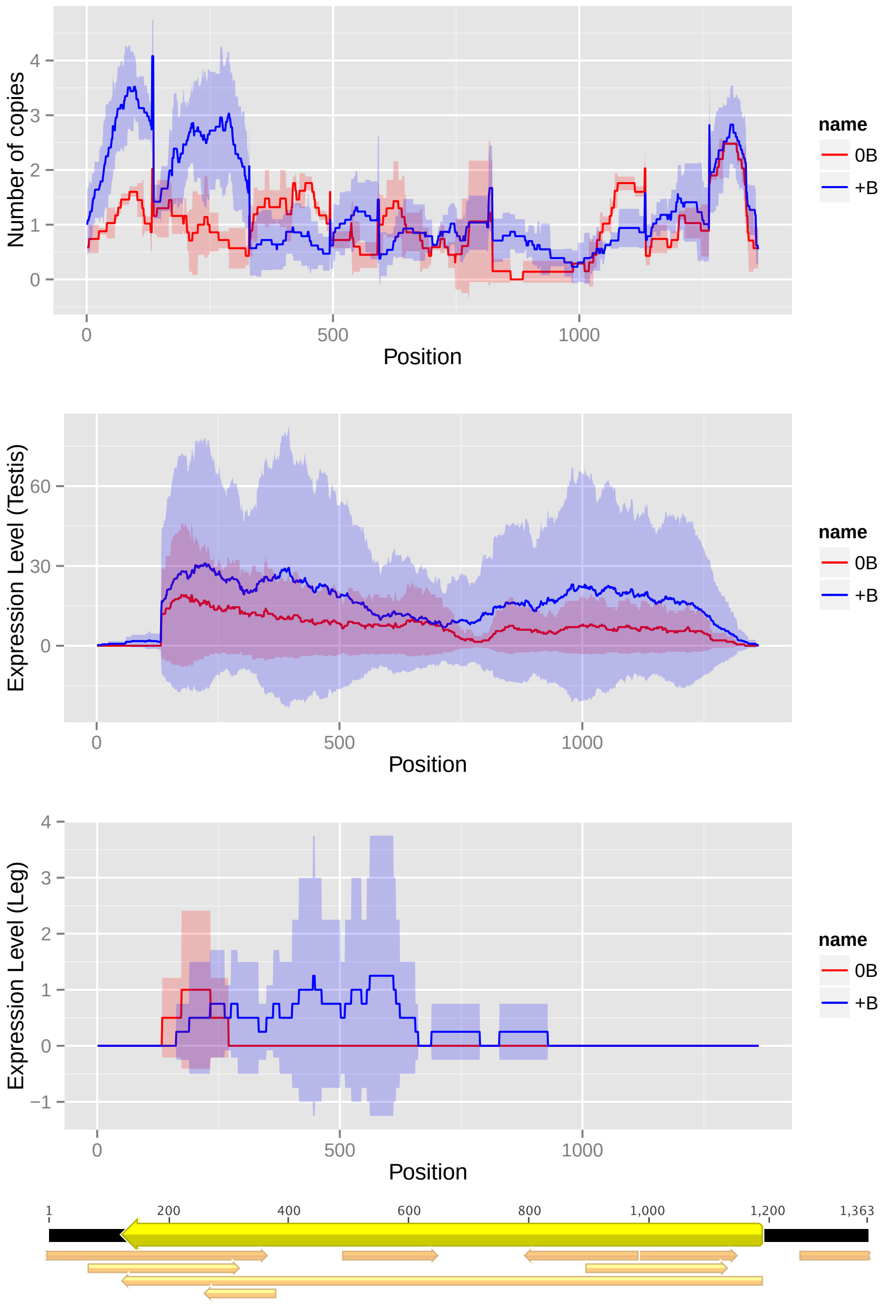


Figure 5.S17: Coverage for the *SLNL1* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

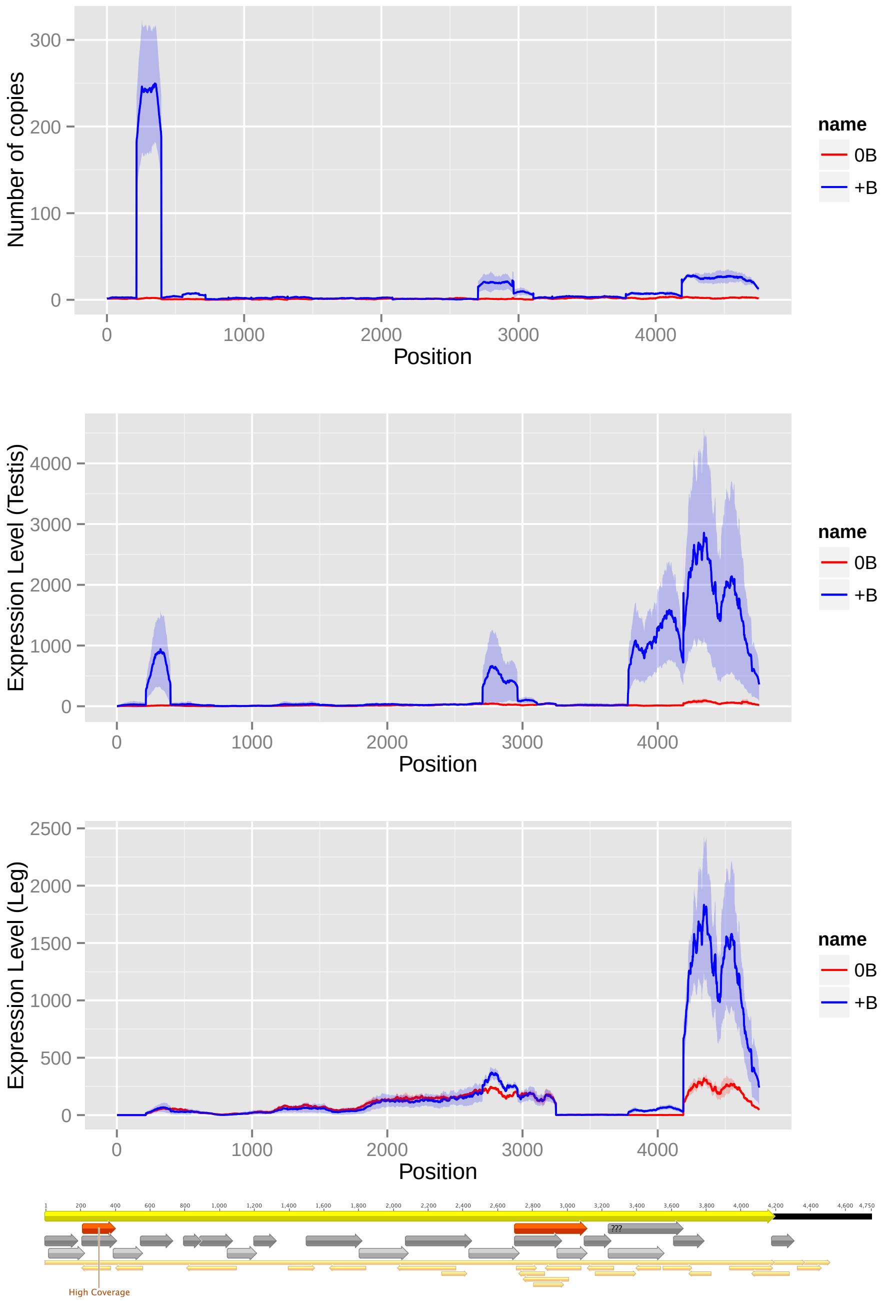


Figure 5.S18: Coverage for the *HEM1* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

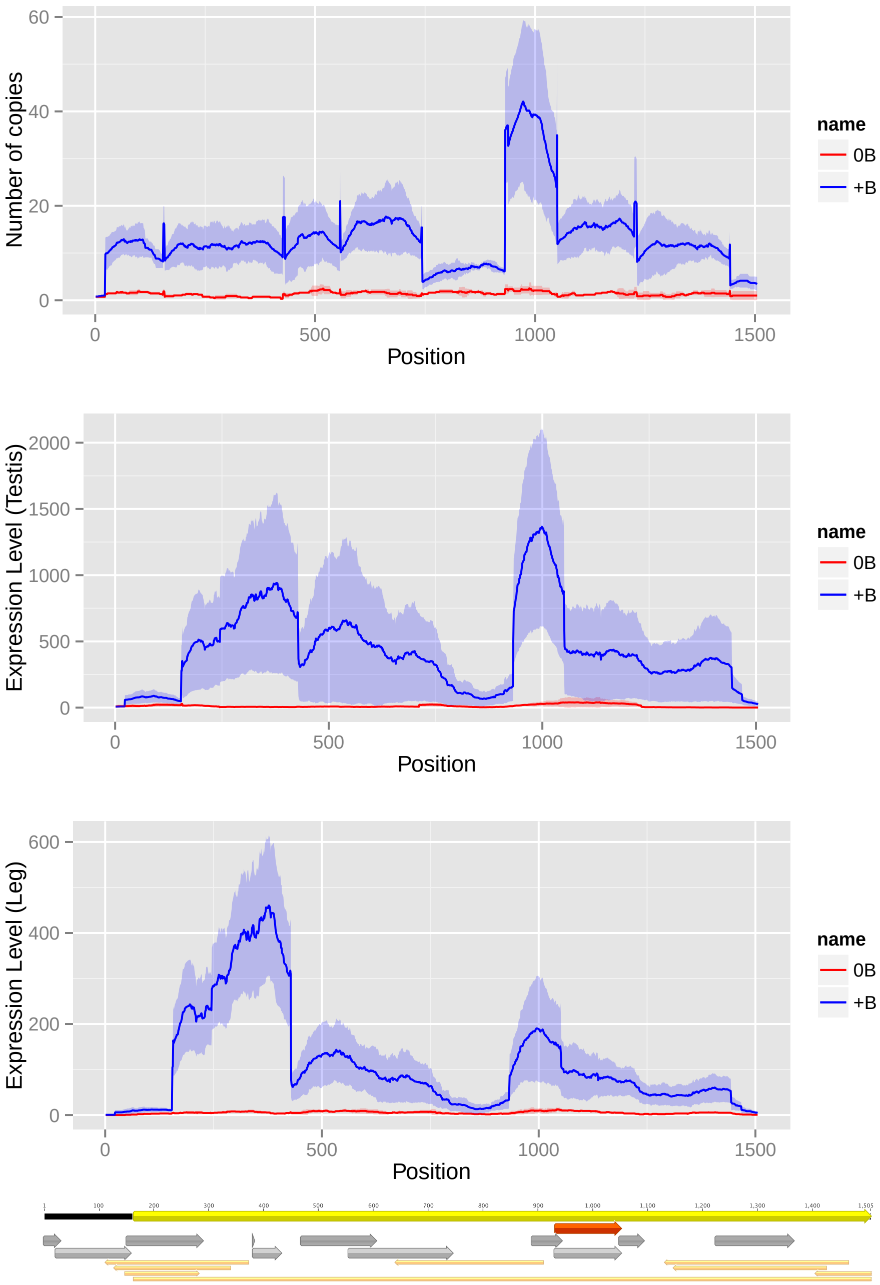


Figure 5.S19: Coverage for the *BMBL* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

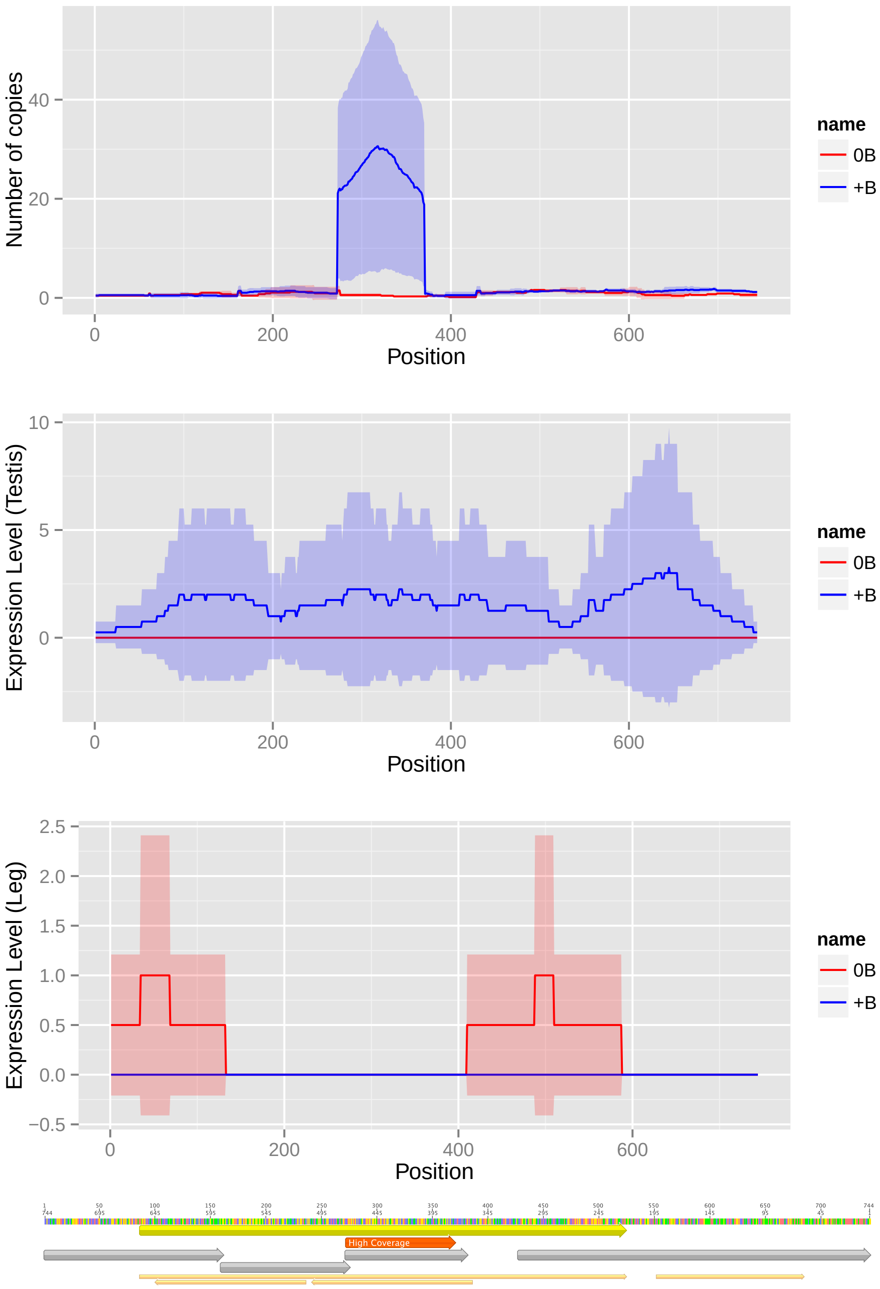


Figure 5.S20: Coverage for the *OR92* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

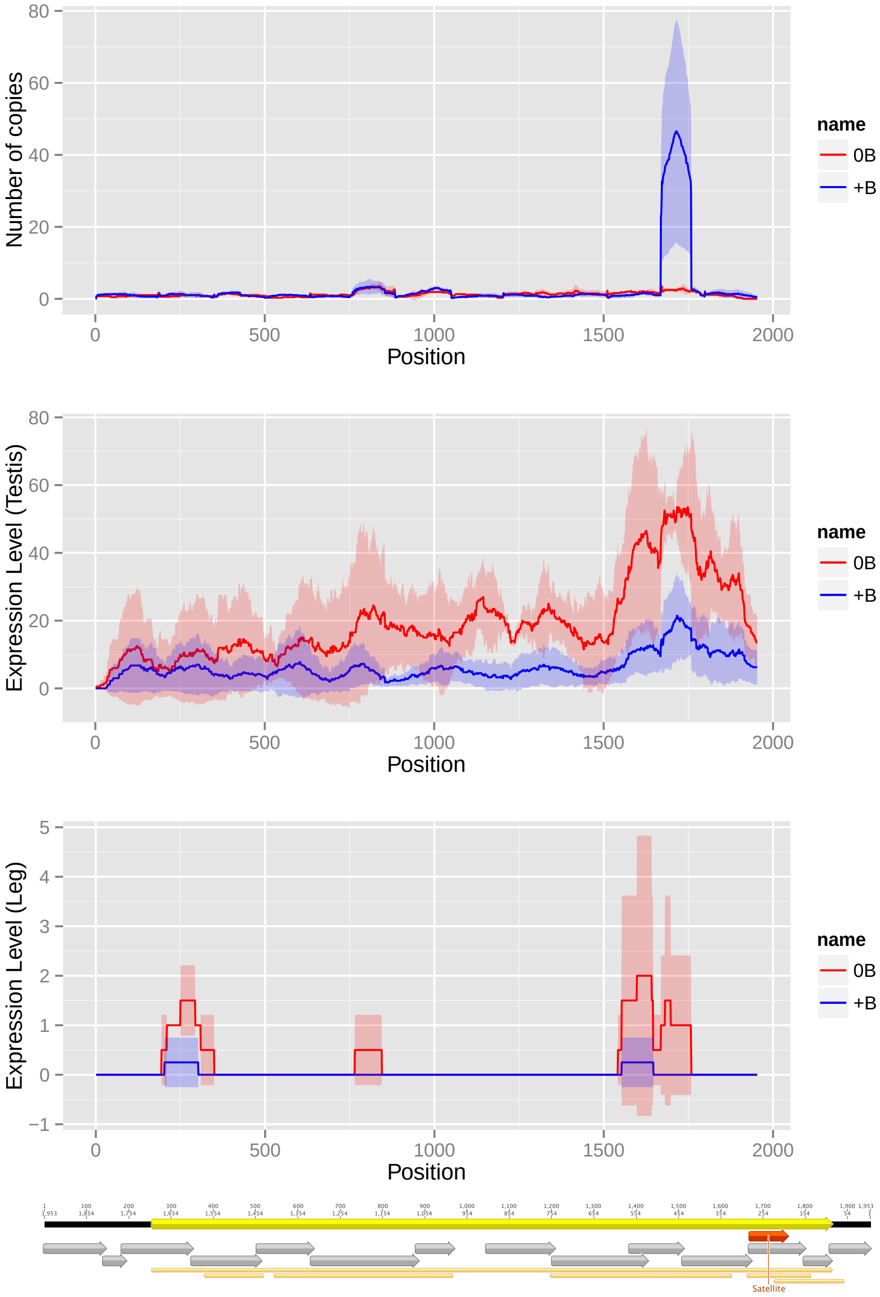


Figure 5.S21: Coverage for the *SV2* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

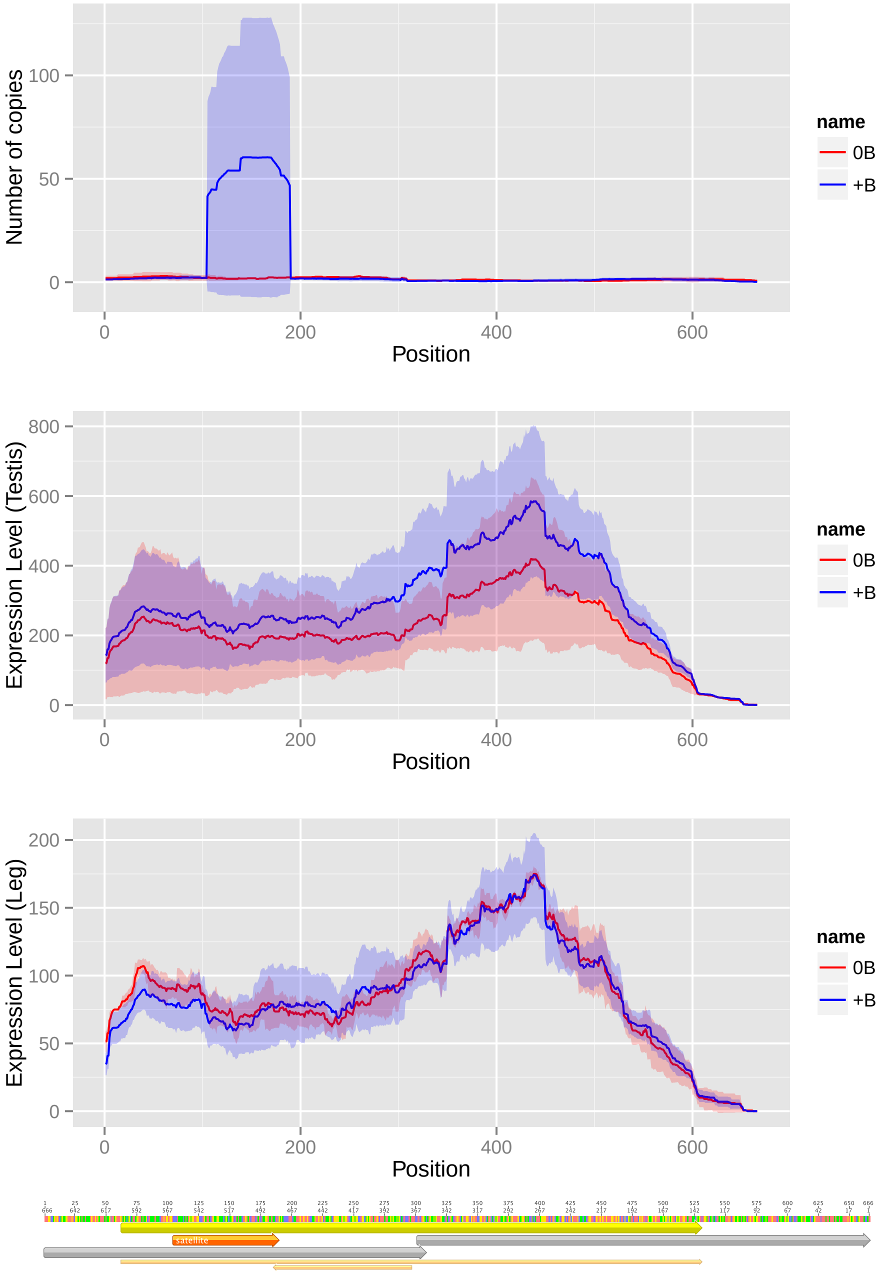


Figure 5.S22: Coverage for the *CL065* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

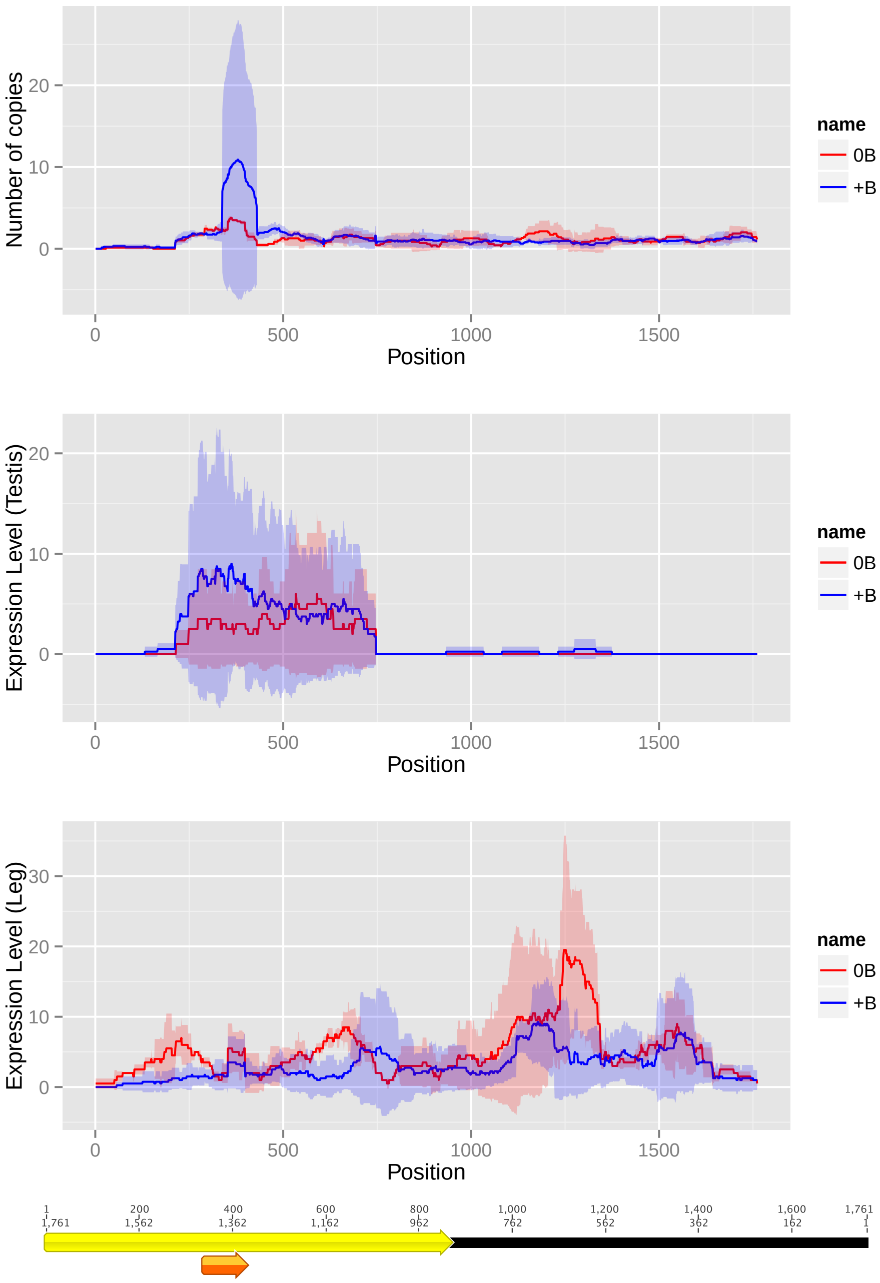


Figure 5.S23: Coverage for the *LSAMP* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals.

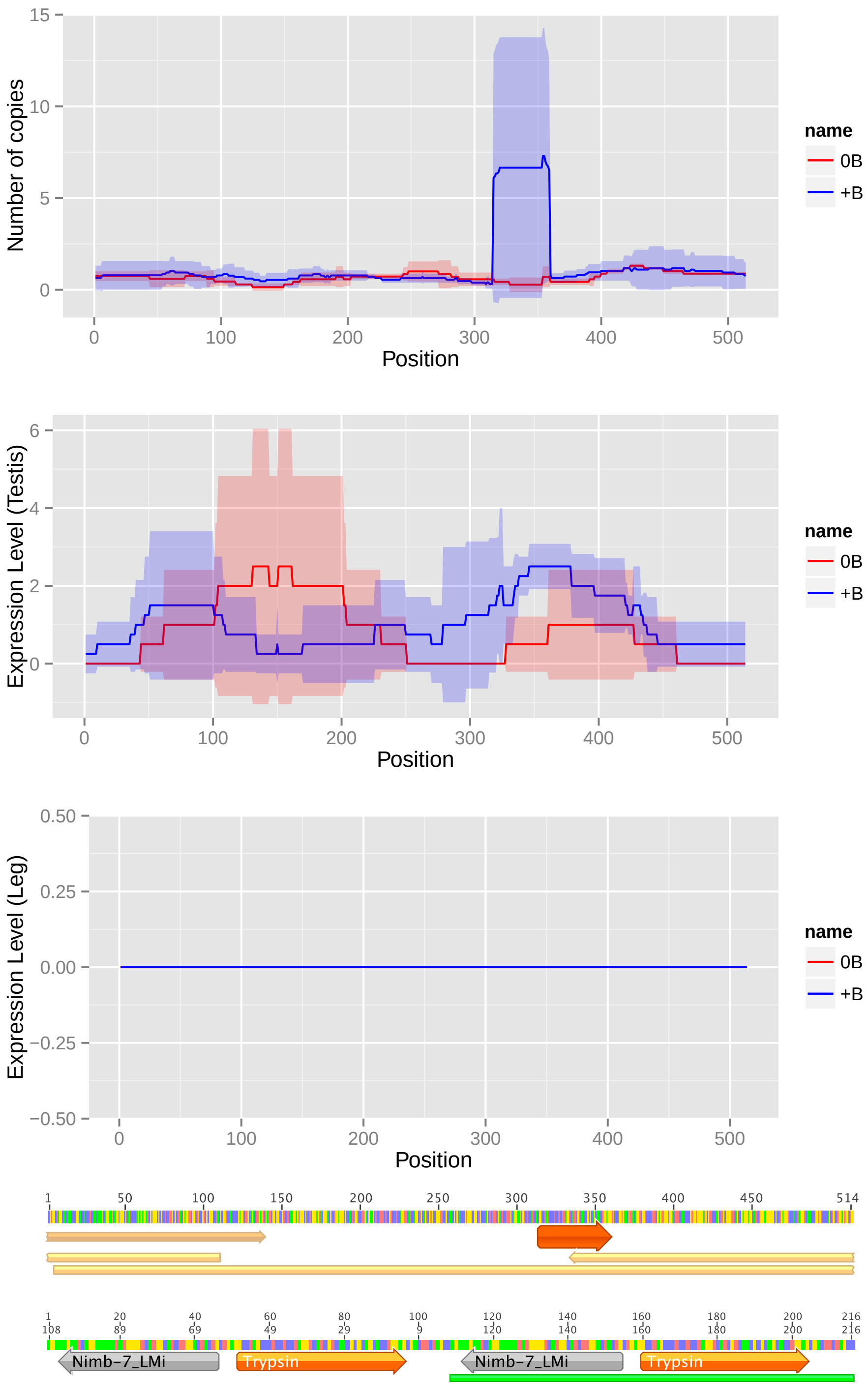


Figure 5.S24: Coverage for the *TRY1* gene in the gDNA RNA from testis and RNA from leg Illumina libraries from Cádiz individuals. The satDNA is mix between limbic and a transposon.

Supplementary Tables

Table 5.S1: Copy numbers for the B chromosome genes in each individual. Only high coverage regions in the B chromosome are considered.

Gene	0B_cadiz01	0B_cadiz04	0B_hp	+B_cadiz02	+B_cadiz03	+B_cadiz05	+B_cadiz06	+B_hp
PELI	1.37	1.18	1.08	**5.76**	**5.22**	**6.31**	**10.89**	**19.70**
VAR1	0.85	1.18	0.60	**2.52**	**2.57**	**2.59**	**2.22**	**6.05**
CC151	0.91	0.88	1.03	**3.03**	**3.06**	**3.00**	**4.41**	**5.98**
MDM2_A	0.41	0.59	1.06	**1.71**	**0.42**	**0.99**	**2.07**	**4.86**
FMO	0.43	0.80	0.40	**1.48**	**1.36**	**1.53**	**2.78**	**3.56**
ZSC22	1.71	1.81	1.22	**3.63**	**3.41**	**4.08**	**4.63**	**3.79**
SUOX_A	1.05	1.05	1.00	**1.96**	**2.73**	**1.91**	**2.65**	**4.40**
APC1	0.98	0.97	0.78	**2.06**	**2.70**	**2.02**	**2.74**	**4.48**
NR2CA	0.56	0.98	1.33	**1.74**	**2.00**	**1.30**	**2.08**	**6.32**
SUOX_B	0.92	0.98	0.95	**2.00**	**3.21**	**1.88**	**2.49**	**4.46**
MDM2_B	0.77	1.07	0.64	**1.95**	**1.41**	**1.88**	**2.65**	**2.64**
RBBP6	1.19	0.56	1.17	**1.61**	**1.76**	**1.55**	**1.57**	**2.45**
FAS2	1.16	1.16	0.70	**11.17**	**13.22**	**12.17**	**16.72**	**38.97**
RT28	0.77	0.78	1.21	**5.54**	**5.00**	**5.36**	**7.32**	**14.40**
VASA	1.99	1.92	1.86	**2.88**	**3.52**	**2.58**	**2.47**	**11.57**
TRET1	0.71	0.73	0.98	**2.19**	**1.21**	**1.55**	**1.57**	**3.74**
SLNL1	0.95	0.94	1.00	**1.51**	**1.24**	**1.32**	**1.19**	**2.08**
HEM1	1.63	1.99	0.36	**204.02**	**183.51**	**209.37**	**330.47**	**434.70**
BMBL	1.29	1.28	0.95	**11.41**	**8.71**	**11.77**	**19.04**	**22.17**
OR92	0.36	0.60	0.00	**23.28**	**51.82**	**28.22**	0.01	1.67
SV2	2.03	2.95	0.00	**44.03**	3.69	**43.87**	**66.69**	0.00
CL065	2.25	1.63	1.18	**117.35**	1.71	**96.81**	2.73	2.22
LSAMP	2.62	2.52	2.67	1.00	3.42	1.21	**29.81**	1.39
TRY1	0.13	0.64	0.04	0.68	**14.24**	0.70	**11.00**	0.00

Table 5.S2: KOG annotation for the genes found in the B chromosome. Genes *CC151*, *VAR1*, *NR2CA*, *MDM2_B*, *RBBP6*, *VASA*, *TRET1*, *SLNL1*, *HEM1*, *BMBL*, *OR92* and *SV2* were unannotated.

Gene	Hit	class	class description
PELI	KOG3842	T	Signal transduction mechanisms
MDM2_A	KOG4172	O	Posttranslational modification, protein turnover, chaperones
FMO	KOG1399	Q	Secondary metabolites biosynthesis, transport and catabolism
ZSC22	KOG2462	K	Transcription
SUOX_A	KOG0535	C	Energy production and conversion
APC1	KOG1858	DO	Cell cycle control, cell division, chromosome partitioning Posttranslational modification, protein turnover, chaperones
SUOX_B	KOG0535	C	Energy production and conversion
FAS2	KOG1445	Z	Cytoskeleton
RT28	KOG4078	J	Translation, ribosomal structure and biogenesis
CL065	KOG2726	J	Translation, ribosomal structure and biogenesis
LSAMP	KOG0162	Z	Cytoskeleton
TRY1	KOG3627	E	Amino acid transport and metabolism

Table 5.S3: Haplotypes defined by SNPs being at distances lower than one Illumina read (i.e. 100 nt) found in the B chromosome genes.

This table can be downloaded in https://dx.doi.org/10.6084/m9.figshare.3255556.

Conclusiones

1. Los microsatélites constituyen sólo el 0,5 % de los genomas de saltamontes, y no explican el tamaño gigante de los genomas de estos organismos. Se localizan tanto en regiones eucromáticas no codificantes, la mayoría de ellas próximas a TEs, como en los espaciadores de los genes para histonas y en el espaciador intergénico del ADNr 45S (IGS).

2. Los cromosomas B de *L. migratoria* y *E. plorans* son pobres en microsatélites. En el primer caso, los microsatélites son más frecuentes en el tercio eucromático proximal, coincidiendo con la localización de los genes para histonas. En los cromosomas B de *E. plorans*, los microsatélites están restringidos a la región distal y pequeñas regiones intersticiales.

3. Hemos desarrollado satMiner, un conjunto de herramientas para el análisis bioinformático del contenido en ADN satélite que, tras varias rondas de clusterización con RepeatExplorer y filtrado con DeconSeq, incrementa mucho la probabilidad de detectar familias de ADN satélite, incluso las que son poco abundantes en el genoma.

4. Proponemos el termino "satelitoma" para la colección completa de las distintas familias de ADN satélite en un genoma. El estudio en profundidad del satelitoma de *L. migratoria* mediante satMiner y FISH nos ha llevado a sugerir una hipótesis para la evolución del ADN satélite. Según nuestro modelo, todos los ADNs satélites se diseminan intragenómicamente y sólo algunos se amplifican localmente dando lugar a clústers visibles mediante FISH. Este patrón es válido tanto para ADNs satélites cortos como largos y es aplicable también a los ADNs satélites de bacterias.

5. El análisis del satelitoma en el saltamontes *E. monticola* ha demostrado que el autosoma megamérico (S8) lleva la mayor cantidad de familias de ADN satélite diferentes al igual que el autosoma megamérico (S9) en *L. migratoria*. Esto sugiere la posibilidad de que el ADN satélite juegue un papel en la heterocromatinización facultativa de este cromosoma durante la meiosis.

6. Seis de las 13 familias de ADN satélite presentes en el autosoma S8 de *E. monticola* estaban también presentes en el cromosoma B. Sin embargo, el cromosoma B carece de genes de histonas y de otros tres ADNs satélites, todos presentes intersticialmente en el cromosoma S8. Esto

permite delimitar el origen del cromosoma B en el tercio proximal del autosoma S8, desde el centrómero a una región localizada entre el ADN satélite EmoSat11-122 y el cistrón de histonas.

7. El análisis mediante FISH mostró que el cromosoma B de *E. monticola* contiene dos ADN satélites específicos que no dieron señal por FISH en los cromosomas A. El análisis bioinformático mostró que ambos ADNs satélites surgieron *de novo* en el cromosoma B, uno de ellos a partir de pocas repeticiones en tándem presentes en el genoma 0B y el otro a partir de un retrotransposón RTE. Esto indica que encontrar ADN satélites específicos de cromosomas B no soporta necesariamente su origen interespecífico.

8. Describimos por primera vez la secuencia completa del cistrón para los genes de histonas en ortópteros. Éste muestra una estructura y una ordenación similar al de dípteros. En base a la variación estructural encontrada en el espaciador 5, hemos desarrollado un marcador molecular para la presencia de cromosomas B en esta especie, que puede ser muy útil en investigaciones futuras.

9. El análisis de la abundancia de elementos repetidos en el genoma de *L. migratoria* revela que alrededor del 55 % del genoma está compuesto por ADN repetitivo. En comparación con los cromosomas A, el cromosoma B de *L. migratoria* está enriquecido en ADN satélite y genes de histonas, pero empobrecido en elementos transponibles. Más del 70 % del ADN repetitivo del cromosoma B corresponde a una única familia de ADN satélite (LmiSat02-176). Además, el cromosoma B contiene otros seis ADNs satélites. El único cromosoma A que lleva los 7 ADNs satélites observados mediante FISH en el cromosoma B es el megamérico (S9), por lo que este cromosoma podría haber jugado un papel en el origen del B, junto con el autosoma S8 (el portador de genes de histonas), tal vez mediante una translocación S8-S9.

10. Una región intersticial del cromosoma B acumula varios tipos diferentes de elementos transponibles. El análisis bioinformático permitió inferir la secuencia de ADN de una quimera que incluye 29 elementos transponibles de 18 familias distintas, la mayoría de los cuales estaban incompletos. Esto sugiere que el cromosoma B es un sumidero evolutivo para algunos elementos transponibles.

11. Hemos desarrollado un método bioinformático para encontrar genes codificadores de proteínas en los cromosoma B, que consiste en el mapeo de lecturas Illumina obtenidas a partir de ADN genómico con y sin cromosomas B, sobre las CDSs de un transcriptoma *de novo*

construido para la misma especie. Este método es útil para especies sin un buen genoma de referencia.

12. Hemos localizado 24 genes codificadores de proteínas en el cromosoma B de *L. migratoria*, la mitad de los cuales están completos y los restantes son pseudogénicos. La divergencia en secuencia entre los genes de los cromosomas A y de los B sugiere una edad para el cromosoma B entre 1 y 4 millones de años.

13. La mayoría de los genes del B de *L. migratoria* estaban transcripcionalmente activos, demostrando que los cromosomas B no son genéticamente inertes. Algunos genes completos en el B podrían tener un papel muy relevante para su transmisión, especialmente APC1 y dos genes MDM2, que son ubiquitin ligasas E3 involucradas en la regulación de la división celular. Por tanto, los cromosomas B tienen el potencial para comportarse como verdaderos parásitos al manipular la expresión génica en el genoma hospedador y así protagonizar una carrera de armamentos transcripcional con los cromosomas A.

Conclusions

1. Microsatellites constitute only 0.5% of grasshopper genomes, and do not explain the gigantic size of grasshopper genomes. They are located in non-coding euchromatic regions, most of them next to transposable elements, and also in histone gene spacers and the 45S rDNA intergenic spacer (IGS).

2. The B chromosomes in *L. migratoria* and *E. plorans* are impoverished in microsatellites. In the first case, microsatellites are more frequent in the proximal euchromatic third of the B chromosome, coinciding with histone gene location. In *E. plorans* B chromosomes, microsatellites are restricted to the distal region and small interstitial regions.

3. We have developed satMiner, a toolkit for the bioinformatic analysis of satellite DNA content which, after several rounds of clustering with RepeatExplorer and filtering with DeconSeq, increases very much the likelihood of detecting satellite DNA families being very scarce in the genome.

4. We suggest the term "satellitome" for the full set of different satellite DNA families in a genome. The high throughput analysis of the satellitome in *L. migratoria*, by means of satMiner and FISH, has led us to suggest a hypothesis on the evolution of satellite DNA. According to our model, all satellite DNAs disseminate intragenomically and only some of them are locally amplified to yield clusters visible by FISH. This pattern is valid for both short and long satellite DNAs, and is also applicable to bacterian satellite DNAs.

5. Satellitome analysis in the grasshopper *Eumigus monticola* has shown that the megameric (S8) autosome carries the highest number of different satellite DNA families, likewise the megameric (S9) autosome in *L. migratoria*. This suggests the possibility that satellite DNA might play a role in facultative heterochromatinization of this chromosome during meiosis.

6. Six out of the 13 satellite DNA families found in the *E. monticola* S8 autosome were also present in the B chromosome. However, the B chromosome lacks histone genes and three other satellite DNAs, all being interstitially located on the S8 chromosome. This allows delimiting B chromosome origin to the proximal third of the S8 autosome, i.e. from the centromere to a site located between the EmoSat11-122 satellite DNA and the histone gene cluster.

7. FISH analysis showed that the B chromosome in *E. monticola* contains two specific satellite DNAs which failed to yield FISH signals on A chromosomes. Bioinformatic analysis showed that both satellite DNAs arose *de novo* in the B chromosome, one of them through amplification of a few tandem repeats existing in the 0B genome, and the other from an RTE retrotransposon. This indicates that finding B-specific satellite DNAs does not necessarily support the interspecific origin of B chromosomes.

8. We describe here, for the first time, the complete sequence of the histone gene cluster in Orthoptera. It shows a structure and gene arrangement similar to those in Diptera. On the basis of structural variation found in spacer 5, we have developed a molecular marker for B chromosome presence in this species, which may be very useful in future research.

9. The analysis of abundance of repetitive elements in the *L. migratoria* genome has revealed that about 55% of it is composed of repetitive DNA. In comparison with A chromosomes, the B chromosome in *L. migratoria* is enriched in satellite DNA and histone genes, but is impoverished in transposable elements. More than 70% of the repetitive DNA found in the B chromosome belong to a single satellite DNA family (LmiSat02-176). In addition, the B chromosome contains six other satellite DNAs. The only A chromosome carrying the 7 satellite DNAs visualized by FISH on the B chromosome is the megameric (S9) autosome, for which reason this chromosome might have played a role in B chromosome origin, along with S8 (the histone genes carrier), perhaps through an S8-S9 translocation.

10. An interstitial region in the B chromosome of *L. migratoria* accumulates several types of transposable elements. Bioinformatic analysis allowed inferring the DNA sequence of a chimera including 29 transposable elements from 18 different families, most of which were incomplete. This suggests that the B chromosome is a sink for transposable elements.

11. We have developed a bioinformatic protocol to search for protein-coding genes in B chromosomes, consisting in mapping Illumina reads obtained from genomic DNA with and without B chromosomes, on the CDSs of a *de novo* transcriptome built for the same species. This method is thus useful for any species lacking a reference genome.

12. We have found 24 protein-coding genes in the B chromosome of *L. migratoria*, half of which were complete and the remaining were pseu-

dogenic. Sequence divergence between A and B chromosome genes suggested a B chromosome age between 1 and 4 million years.

13. Most B chromosome genes in *L. migratoria* were transcriptionally active, showing that B chromosomes are not genetically inert. Some of the B chromosome complete genes might play a very relevant role in B chromosome transmission, especially APC1 and two MDM2 genes, which are E3 ubiquitin ligases involved in cell division regulation. Therefore, B chromosomes have the potential to behave as true parasites by manipulating gene expression in the host genome and thus carrying out a transcriptional arms race with A chromosomes.

Perspectivas

Durante el desarrollo de esta tesis doctoral hemos aprendido no solo acerca del origen y la naturaleza de los cromosoma B de *Locusta migratoria* y *Eumigus montícola*, sino también sobre la elementos repetidos a nivel más general. No obstante, quedan abiertas nuevas incógnitas para desvelar en el futuro, tales como las que describimos a continuación.

La relación de los microsatélites con elementos transponibles, observada en el capítulo 1, sugiere que es necesario entender mejor el papel de los elementos transponibles en la abundancia y distribución de microsatélites a lo largo del genoma.

El estudio en profundidad del satelitoma, mediante la metodología propuesta en el capítulo 2, abre un campo nuevo de estudio para este tipo de secuencias. Esperamos que la aplicación de nuestro protocolo a otros genomas permita descubrir nuevas generalidades sobre el ADN satélite. Algunos aspectos interesantes que quedan por estudiar son, por ejemplo, cómo evoluciona el satellitoma al nivel intrapoblacional, y cómo lo hace entre especies emparentadas. También la transcripción de este tipo de secuencias, y su posible papel en la regulación génica, deberían ser estudiados con más detalle.

Para el cromosoma B de *E. monticola,* que es analizado en el capítulo 3, sería interesante realizar un análisis genómico y transcriptómico encaminado a determinar el contenido del cromosoma B para otros elementos repetitivos, además del satelitoma analizado aquí, así como para genes codificadores de proteínas. Una posibilidad interesante sería encontrar, como en *L. migratoria*, genes relacionados con la división celular. Finalmente, la presencia de dos ADNs satélites específicos del cromosoma B (al nivel de FISH) puede servir como marcador para estudiar el comportamiento meiótico de éste y su transmisión al nivel citológico.

En el capítulo 4, estudiamos el contenido en ADN repetitivo del cromosoma B de *L. migratoria*. Será interesante, en próximas investigaciones, probar la existencia física de la quimera de transposones inferida mediante análisis bioinformático, comprobando el orden de los transposones mediante PCR y visualizando su localización en el cromosoma B mediante FISH. Alternativamente, podemos realizar esta comprobación mediante las nuevas plataformas de secuenciación que generan fragmentos de varias kilobases. Esta quimera, además, puede ser diana para generar nuevos marcadores moleculares para la presencia del cromosoma B.

Con respecto al origen del cromosoma B, probablemente a partir de alguno(s) de los cromosomas A, las secuencias repetidas no han dado una información muy clara. No obstante, como observamos en el capítulo 5, tal vez la respuesta final puede venir de la mano de los genes codificadores de proteínas localizados en el cromosoma B. Un análisis mediante FISH, utilizando varias sondas que abarquen varias kilobases del gen *APC1*, por ejemplo, o BACs que lo contengan, permitiría identificar el cromosoma A del que se originó el cromosoma B. Además, sería interesante estudiar la presencia de estos genes en los cromosomas B de otras poblaciones naturales, localizadas a lo largo de la distribución de la especie, para comprobar cuáles de ellos están en todas ellas y, por tanto, podrían ser cruciales para su transmisión y mantenimiento.

Otras publicaciones

Los trabajos publicados durante mi periodo pre-doctoral incluyen el primer capítulo así como las siguientes publicaciones:

1. Silva, D. M. Z. A., Daniel, S. N., Camacho, J. P. M., Utsunomia, R., Ruiz-Ruano, F. J., Penitente, M., Pansonato-Alves, J. C., Hashimoto, D. T., Oliveira, C., Porto-Foresti, F., Foresti, F. Origin of B chromosomes in the genus *Astyanax* (Characiformes, Characidae) and the limits of chromosome painting. Molecular Genetics and Genomics, in advance online publication 16 March 2016.

2. Utsunomia, R., Silva, D. M. Z. A., Ruiz-Ruano, F. J., Araya-Jayme, C., Pansonato-Alves, J. C., Scacchetti, P. C., Hashimoto, D. T., Oliveira, C., Trifonov. V. A., Porto-Foresti, F., Camacho J. P. M., Foresti, F. Uncovering the ancestry of B chromosomes in Moenkhausia sanctaefilomenae (Teleostei, Characidae). PLoS One 11(3): e0150573.

3. Doña, J.*, Ruiz-Ruano, F. J.*, Jovani, R. DNA barcoding of Iberian Peninsula and North Africa Tawny Owls Strix aluco suggests the Strait of Gibraltar as an important barrier for phylogeography. Mitochondrial DNA, in advance online publication 14 October 2015.

4. Camacho, J. P. M., Shaw, M. W., Cabrero J., Bakkali, M., Ruíz-Estévez, M., Ruiz-Ruano, F. J., Martín-Blázquez, R., López-León, M. D. (2015). Transient microgeographic clines during B chromosome invasion. The American Naturalist 186 (5): 675-681.

5. Montiel, E. E.*, Ruiz-Ruano, F. J.*, Cabrero, J., Marchal, J. A., Sánchez, A., Perfectti, F., López-León, M. D., Camacho, J. P. M. (2015). Intragenomic distribution of RTE retroelements suggests intrachromosomal movement. Chromosome Research 23(2): 211-223.

6. Ruiz-Estévez, M.*, Ruiz-Ruano, F. J.*, Cabrero, J., Bakkali, M., Perfectti, F., López-León, M. D., Camacho, J. P. M. (2015). Non-random expression of rDNA units in a grasshopper showing high intragenomic variation for the ITS2 region. Insect Molecular Biology 24(3): 319-330.

7. Guillén, Y., Rius, N., Delprat, A., Williford, A., Muyas, F., Puig, M., Casillas, S., Ràmia M., Egea, R., Negre, B., Mir, G., Camps, J., Moncunill, V., Ruiz-Ruano, F. J., Cabrero, J., de Lima, L. G., Dias, G. B., Ruiz, J.C., Kapusta, A., Garcia-Mas, J., Gut, M., Gut, I. G., Torrents, D., Camacho, J. P. M., Kuhn, G. C. S., Feschotte, C., Clark, A. G., Betrán, E.,

Barbadilla, A., Ruiz, A. (2015). Genomics of ecological adaptation in cactophilic Drosophila. Genome Bioloy and Evolution, 7(1): 349-366.

8. Camacho, J. P. M., Ruiz-Ruano, F. J., Martín-Blázquez, R., López-León, M.D., Cabrero, J., Lorite, P., Cabral-de-Mello, D. C., Bakkali, M. (2015). A step to the gigantic genome of the desert locust: chromosome sizes and repeated DNAs. Chromosoma 124(2): 263-275.

9. Camacho, J.P.M.; Cabrero, J.; López-León, M.D.; Cabral-de-Mello, D.C.; Ruiz-Ruano, F.J. (2014). Grasshoppers (Orthoptera). In: Protocols for Cytogenetic Mapping of Arthropod Genomes (Ed. Sharakhov, I.V.) pages 381-438.

10. Anjos, A., Ruiz-Ruano, F. J., Camacho, J. P. M., Loreto, V., Cabrero, J., de Souza, M. J. & Cabral-de-Mello, D. C. (2015). U1 snDNA clusters in grasshoppers: chromosomal dynamics and genomic organization. Heredity 114: 207-219.

11. Teruel, M., Ruiz-Ruano, F. J., Marchal, J. A., Sánchez, A., Cabrero, J., Camacho, J. P. M., & Perfectti, F. (2014). Disparate molecular evolution of two types of repetitive DNAs in the genome of the grasshopper Eyprepocnemis plorans. Heredity 112: 531–542.

12. Silva, D. M. Z. A., Pansonato-Alves, J. C., Utsunomia, R., Araya-Jaime, C., Ruiz-Ruano, F. J., Daniel, S. N., Hashimoto, D. T., Oliveira, C., Camacho, J. P. M., Porto-Foresti, F., Foresti, F. Delimiting the origin of a B chromosome by FISH mapping, chromosome painting and DNA sequence analysis in Astyanax paranae (Teleostei, Characiformes). PLoS ONE 9(4): e94896.

13. Ruiz-Ruano, F. J., Camacho J. P. M., Cabrero, J., López-Rodríguez, M., Tierno de Figueroa, J. M. (2014). Peripatric origin of the only cave-restricted stonefly species known (Insecta: Plecoptera). Arthropod Systematics & Phylogeny 72(1): 3-10.

14. Bessa, J., Luengo, M., Rivero-Gil, S., Ariza-Cosano, A., Maia, A. H., Ruiz-Ruano, F. J., Caballero, P., Naranjo, S., Carvajal, J. J, & Gómez-Skarmeta, J. L. (2014). A mobile insulator system to detect and disrupt cis-regulatory landscapes in vertebrates. Genome Research 24(3): 487-495.

15. Cabrero, J., Bakkali, M., Navarro-Domínguez, B., Ruiz-Ruano, F. J., Martín-Blázquez, R., López-León, M. D., & Camacho, J. P. M. (2013). The Ku70 DNA-repair protein is involved in centromere function in a grasshopper species. Chromosome Research 21(4): 393-406.

16. Ruiz-Ruano, F. J., Ruiz-Estévez, M., Rodríguez-Pérez, J., López-Pino, J. L., Cabrero, J., & Camacho, J. P. M. (2011). DNA amount of X and B chromosomes in the grasshoppers Eyprepocnemis plorans and Locusta migratoria. Cytogenetic and genome research 134(2): 120-126.

Agradecimientos

Ya sea por lo que han hecho bien o por lo que ha hecho mal, de todas estas personas he aprendido cómo quiero ser y cómo no quiero ser.

En primer lugar, quiero agradecer a mis padres, los que más han sufrido por mis ausencias durante todos los años que ha durando la tesis. Tal vez lo menos importante que tengo de ellos es su herencia genética y epigenética. Han sido los únicos que han estado a mi lado en todo momento, sobre todo cuando más he necesitao un consejo, unas palabras de ánimo o un táper de lentejas. Gracias por enseñarme lo que cuesta ganar un chusco de pan y que con esfuerzo se puede llegar lejos. También agradezco a mis abuelos por su cara de orgullo cuando he aparecido por la puerta y por los buenos alimentos del terrero que siempre me han estado suministrando, ojalá pudieran estar todos en este momento. También quiero agradecer a mi hermana, con quien he convivido media vida. Y a mis otros familiares cercanos y vecinos, por tanto que me han enseñado.

También a mis amigos de cuando era mozuelo, en especial a, Jose, Itano, Migueles, Nhinho Spoa, L'avalo y Zizou, por tan buenos momentos que hemos pasado juntos.

Suena el Quatuor de Pierre-Max Dubious, la Obertura 1812 de Tchaikovsky, La Isla de las Perlas, la banda sonara de Rocky, Hosanna in excelsis, Manolete, un popurrí de Mario Bros y el Danzón nº 2 de Arturo Márquez. Dos tercios de mi vida siendo músico y cuando suena mi clarinete añoro aquellos momentos en los que podía dedicar horas y horas a la banda de música o el quarteto de clarinetes. Con ellos empecé a hacer cosas de mayores sin mis padres y es una de las experiencias más enriquecedoras de mi vida. Mención especial a Jose 'el Zocato' quien, a pesar de nuestra diferencia de edad, gracias a la música nos hicimos muy buenos amigos. Gracias por esos ratos de matemáticas y astronomía, por ser mi primera referencia como universitario, por ser mi padrino de Linux y por enseñarme tus "Cagadillas Cósmicas" como muestra de lo poderosa que es la programación. Con Jose y Miguel pasamos muy buenos ratos dedicados a nuestro Orange's Clarinet Quartet de tres que pasó a ser de cinco, seis, hasta ni se sabe... También mención a Nazaret, Rafa "Palote" y Olga.

A mis amigos de la carrera, sobre todo a Marcos (alias "Carlitos") y Alfonso. Por todos esos buenos momentos frikis de los que sería indecoroso hablar aquí, pero que me han servido desconexión durante los interminables días de prácticas o las horas de estudio en época de exámenes. A Alfonso por esas buenas conversaciones que hemos tenido de ciencia con unas jarras de por medio.

Un plato de lasaña, un bacalao al horno con alioli y un cuchillo que no corta. Agradezco a los Comedores Universitarios, el mejor servicio que tiene esta universidad, gracias por la buena organización y dedicación de su personal, por tan buenos momentos que he pasado allí. Gracias a ellos he estado bien nutrido cuando la comida de mi madre escaseaba. Se hacen largos los veranos sin comedores.

Mención especial a la comunidad de software libre porque sin una comunidad tan generosa nunca hubiera podido aprender tanto sobre Linux y programación por mi cuenta. En especial agradecer a mis amigos de la Oficina de Software Libre, tan bien dirigida por JJ, que confiaron en mí para dar mis primeras charlas de LyX y LaTeX, y con quienes pasé buenos momentos de frikismo informático.

’Uy un espejo! A Ángel Martín gracias por sus clases de Genética porque además de las leyes de Mendel despertó mi curiosidad por aprender más sobre Linux y LaTeX. A Paco Perfectti por esas clases de más allá de la Genética y por sus buenas referencias sobre mí, que fueron importantes para entrar al Departamento de Genética como alumno interno. Gracias por tus valiosos comentarios, aunque no siempre hemos estado de acuerdo en todo.

Han pasado casi nueve años desde que nos conocemos y cada día aprendo algo nuevo de él. Gracias a mi director Juan Pedro, a quien además considero mi padre científico. Gracias por plantearme continuamente retos que me han tenido horas y horas pensando, y que me han hecho sentirme realizado cuando los hemos resuelto. Hemos formado un magnífico tándem y hemos llegado mucho más lejos de lo que esperábamos. Gracias a la meticulosidad con que has revisado mi trabajo, he tenido que dar mi máximo. Algún día tendrás que contarme de dónde sacas el tiempo para trabajar tanto.

Recuerdo con emoción mi primera tarde tomando fotos en el microscopio con Pepi. Gracias por transmitirme tu pasión por la citogenética y por las palizas de FISH que semana tras semana te has pegado. Agradezco a Lola, por sus demoledoras dosis de realismo y por darse otros buenos tutes de FISH. Sin vosotras dos, en esta tesis la bioinformática no hubiera recibido tanto *feedback* de la citogenética.

Gracias al resto de profesores del departamento porque de todos vosotros he aprendido algo valioso. Especialmente a Bakkali por esos primeros pasos que dimos juntos con la secuenciación masiva, aunque para bien o para mal nuestros caminos se separaron. A Pepe Oliver, por dejarme acceso a sus supermáquinas, especialmente cuando para mí un ordenador de 8 Gb de RAM era un monstruo. A Esther Viseras por toda la pasión por la docencia que me ha sabido transmitir. También a Roberto y Rafa Jiménez quienes, junto a Paco Perfectti y otros estudiantes del departamento, hemos organizado saraos divulgativos durante los últimos años.

Gracias a todos mis compañeros de poyata. Juntos hemos sobrevivido no solo a horas y horas de laboratorio húmedo y laboratorio *in silico*, sino a los esperados días de muestreo, a interminables mudanzas, maliciosos ataques con pintura azul en nuestra habitación de los bichos, descongelaciones de los frigoríficos a -20 y -80ºC durante días festivos e inundaciones del sótano, incluso sobrevivimos al saltamontes de la feria de Sanlúcar. Gracias a María Teruel, con quien no pasé mucho tiempo, pero quien dejó muy buenas preguntas en el laboratorio. Gracias a Tati, quien fue una guía durante mis primeros pasos en el laboratorio, gracias por enseñarme algo más que ha hacer PCRs o clonaciones. A Eli, nuestra mamá en el laboratorio, de quien aprendí a ser ordenado y a utilizar la inteligencia social. Mercy, gracias por ser tan tenaz en tus últimos meses de tesis, a veces me he visto reflejado en ti y en los momentos de cansancio siempre he arrimado el hombre un poquito más. Gracias por los *hashs* y fiestas varias con el equipo de *frisbee* al que al final terminaste enganchándome. A María Lucena, juntos empezamos nuestros primeros pasos como alumnos internos y fue gratificante ver a otra persona tan perdida por el laboratorio. Gracias a Eva, con quien compartí momentos curiosos en el laboratorio y quien me ayudó a encontrar alojamiento en mi estancia en Uppsala. Gracias a Jesús, de quién además de aprender de redes y filogenias, aprendí lo bueno que es ser meticuloso. A mi maestro *jedi* Moha, de quien me considero su *padawan,* por recordarme continuamente cómo no se acababan las tesis, por estar ahí los fines de semana en el laboratorio, por desear de mayor ser como yo, por irse antes de hora porque le cerraban el Mercadona, por hacer *spoilers* de Juego de Tronos y por tus útiles consejos. A Bea, por ser con quien he pasado más buenos y malos momentos en el laboratorio. Todos estos años, gracias por enseñarme, por intentar entenderme, por alegrarte de mis progresos y por tantas conversaciones enriquecedoras. Me alegra saber que en pocos días también estarás escribiendo los agradecimientos de tu tesis. A Rubén, con quien podría formar un dúo cómico hablando de movidas y paranoias. Gracias por esos momentos donde nos hemos puesto los rulos y me has hecho caso al tirar por la sombra. A Carmen y Xiomara, por mostrar tanto interés esos primeros momentos enseñándoles en el laboratorio. A Mode, por estar siempre cubriéndome las espaldas. A Pedro Lanzas, que siempre me sorprende con un conocimiento nuevo. A Carolina, el *yin*, por siempre tener una sonrisa y una conversación absurda, sobre todo cuando hablamos nuestro viaje a Marruecos. A María Martín, el *yang*, el toque norteño del laboratorio siempre dispuesta a todo tanto dentro como fuera del laboratorio. Con ella el estudio de los Bs tiene el futuro garantizado. A Merce, por interesarse siempre por cómo me va. Un recuerdo también para mis otros compañeros de Departamento Ester Quesada, Fany, Darío, Pedro Sola, Laura, Alicia, Cristina Arquellada, Dilam, Cristina Gómez y Ricardo, y todos aquellos que he olvidado.

A mis vecinos becarios del Departamento de Ecología, sobre todo a Javi. Siento no poder haberme echado tantos cigarros como cafés me has ofrecido, pero con el tiempo hemos pasado buenos momentos tanto antes como durante y después del trabajo. A Jorge, por sus enriquecedoras conversaciones sobre ciencia que han llegado a terminar en publicación. Un placer colaborar contigo.

A mis amigos de Brasil Diogo, Roberto, Tatiana, Maresa, Duílio, Zeca, Ricardo y Érica, por los interesantes problemas que me han planteado. Sin sus preguntas, esta tesis tendría menos respuestas.

A José Luis Gómez-Skarmeta por abrirme las puertas de su laboratorio en el CABD del CSIC y a Zé por ser mi mentor en el trabajo con el pez cebra. Gracias a ambos por enseñarme las bases de la ciencia, aquello que no está en los libros, o si está debe ser en libros muy aburridos. Gracias también a Ana Ariza.

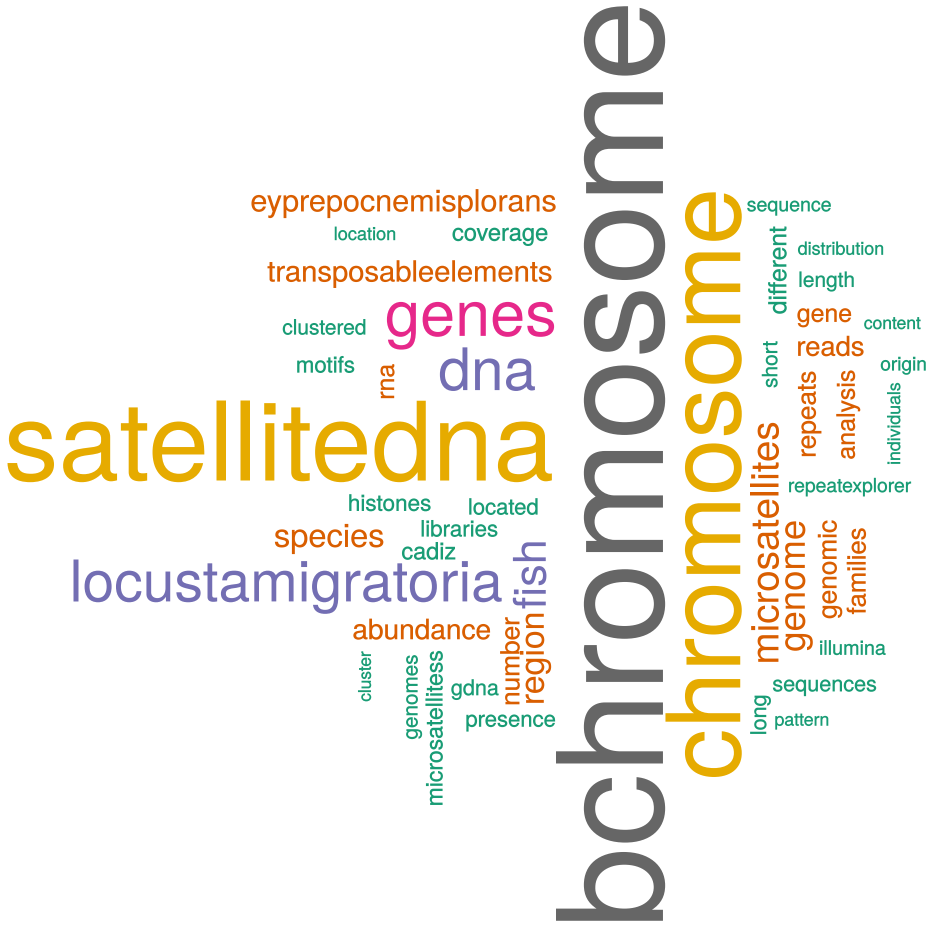
A Jochen Wolf por permitirme hacer una fantástica estancia en el EBC de la Universidad de Upsala. Ahí encontré nuevas ideas y motivación para los dos últimos años de mi tesis. Gracias en especial a Nagarjun, Jelmer, Alex Suh, Taki, Severin y Linnea.

La mitad de esta tesis no hubiera sido posible sin la colaboración de Francisco Jiménez-Cazalla, gracias a él conseguí el valiosísimo material de Cádiz.

A mis amigos del equipo de ultimate frisbee, por esos buenos ratos que nos hemos echado forzando *flick*, forzando *back* y forzando bar.

Gracias a Indira por toda su pasión en mejorar mi vida 1.0.

Gracias a la Universidad, la Consejería y el Ministerio, sin su saber hacer no hubiera valorado tanto lo que le debo a esta gente.


bchromosome
chromosome
satellitedna
locustamigratoria
genes
dna
eyprepocnemisplorans
location
coverage
transposableelements
clustered
motifs
rna
histones
located
species
libraries
cadiz
fish
abundance
number
region
cluster
genomes
microsatelitess
gdna
presence
sequence
distribution
different
length
gene
content
reads
short
origin
repeats
analysis
individuals
microsatellites
repeatexplorer
genome
genomic
families
illumina
sequences
long
pattern