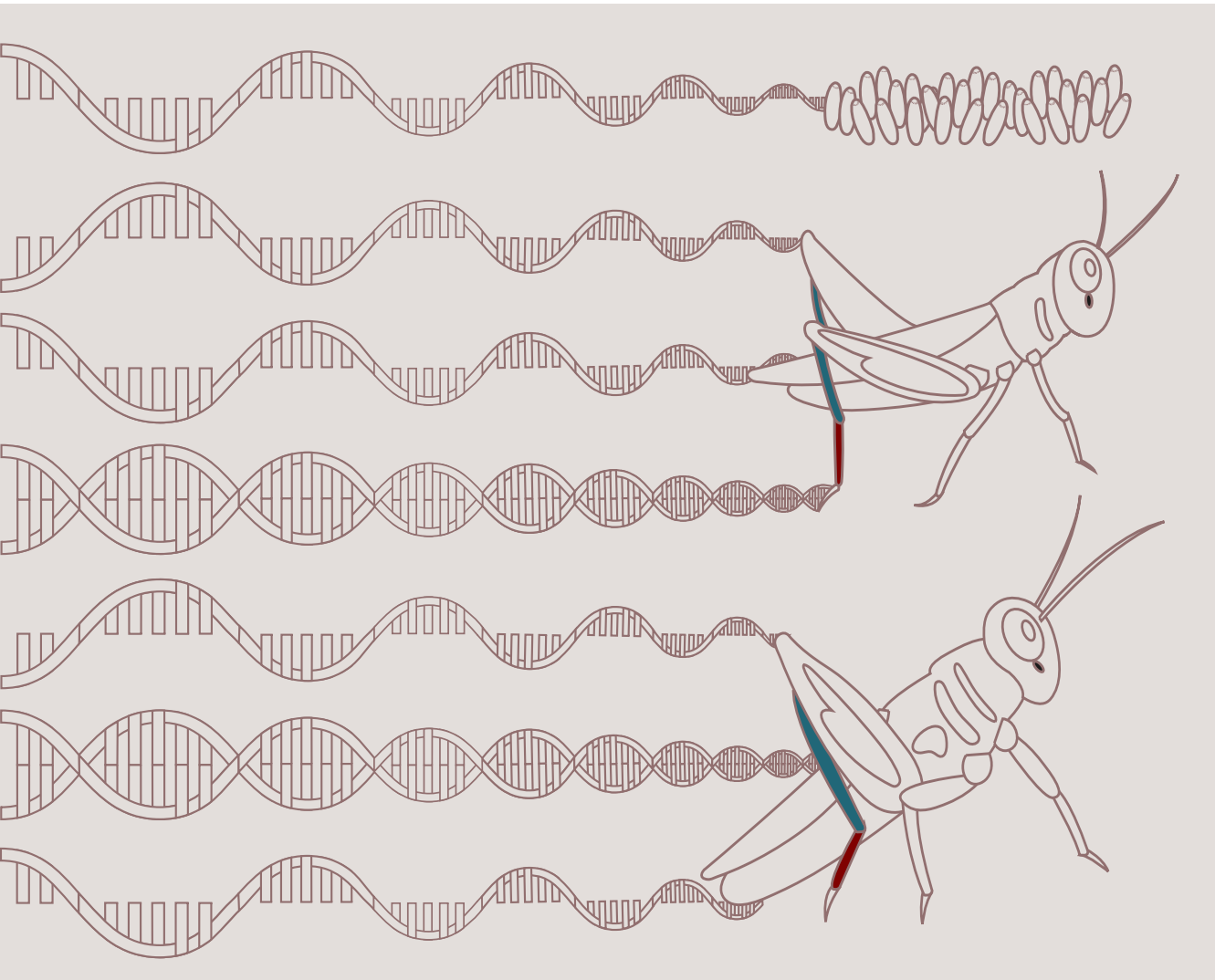
DOCTORAL THESIS

GENOMICS AND TRANSCRIPTOMICS OF THE B CHROMOSOMES OF THE GRASSHOPPER *Eyprepocnemis plorans*

María Martín Peciña

University of Granada
Department of Genetics
PhD program in Fundamental and Systems Biology



GENÓMICA Y TRANSCRIPTÓMICA DE LOS CROMOSOMAS B DEL SALTAMONTES *Eyprepocnemis plorans*

GENOMICS AND TRANSCRIPTOMICS OF THE B CHROMOSOMES OF THE GRASSHOPPER *Eyprepocnemis plorans*

UNIVERSIDAD DE GRANADA

María Martín Peciña

PhD Thesis

Fundamental and Systems Biology

2020

Editor: Universidad de Granada. Tesis Doctorales
Autor: María Martín Peciña
ISBN: 978-84-1306-798-8
URI: http://hdl.handle.net/10481/67821

La doctoranda María Martín Peciña y el director de tesis Juan Pedro Martínez Camacho: Garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por la doctoranda bajo la dirección del director de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Lugar y fecha:

Granada, a 16 de diciembre de 2020

Director de la Tesis:	Doctoranda:
Juan Pedro Martínez Camacho	María Martín Peciña

El presente trabajo se ha realizado en el Grupo de Genética Evolutiva del Departamento de Genética de la Universidad de Granada.

La investigación ha sido financiada por el Ministerio de Economía y Competitividad a través del proyecto CGL2015-70750-P.

Durante la realización de este proyecto de Tesis Doctoral he disfrutado de una beca predoctoral de Formación del Personal Universitario (FPU), del Ministerio de Educación, Cultura y Deporte, con la siguiente referencia: FPU13/01553.

La estancia realizada durante esta Tesis en el Laboratory of Computational Biology (Center for Human Genetics) VIB en la KULeuven ha sido financiada por el programa de Ayudas para la Movilidad Internacional de Estudiantes de Doctorado Universidad de Granada y CI-BIOTIC (2015/2016).

Table of Contents

List of Figures

List of Tables

See each specific chapter for supplementary tables.

List of Common Acronyms

aFC: transcriptional fold change of Alt variants (B-specific) between +B samples

AL: array length

Alt: sequence variant specific from B-libraries

arFC: transcriptional fold change between Alt and Ref variant in the same +B sample

Bs: B chromosomes

B: banded FISH pattern

BAM: binary Alignment Map

BED: browser Extensible Display

bp: base pair

CDS: coding sequence

D: dotted FISH pattern

DAPI: 4', 6 diamidino-2-phenylindole

DB: dotted-banded FISH pattern

DEG: differential expressed gene

DIF: differences between FIT measured applying different maxRU thresholds

FDR: false discovery rate

FISH: fluorescence *in situ* hybridization

FIT: congruence between FISH pattern and maxAL found

gFC: log2 of genomic fold change measured usually in +B individuals with respect to 0B ones

IC: irregular coverage B-gene

ID: interarray distance in Chapter 1

IGS: intergenic spacer

ITS: internal transcribed spacer

Kdiv: Kimura divergence

LINE: long interspersed nuclear element

LTR: long terminal repeat

maxAL: maximum array length

maxID: maximum interarray distance

maxRU: maximum number of repeat units

min: minutes

minTS: minimum target size in nucleotides for FISH visualization

mtDNA: mitochondrial DNA

NGS: next generation sequencing

NS: no signal after FISH

nt: nucleotides

PERU: proportion of external repeat units

PI: polymerization index

rDNA: ribosomal DNA

Ref: sequence variant from A chromosomes

rFC: transcriptional fold change of Ref variants between samples

RU: number of repeat units

RUL: repeat unit length

satDNA: satellite DNA

sec: seconds

SINE: short interspersed nuclear element

SNP: single nucleotide polymorphism

snRNA: small nuclear RNA

TBE: tris-borato EDTA

TE: transposable element

tFC: log2 of transcriptional fold change measured usually in 1B individuals with respect to 0B ones

TI: transcriptional intensity

TR: tandem repeat

tRNA: transfer RNA

TS: target size in nucleotides for FISH visualization

TSI: tandem structure index

UC: uniform coverage B-gene

UTR: untranslated region

Summary

The eukaryotic genome, apart from harboring canonical genomic elements, is vulnerable to the presence of other components that can be accessory and even harmful for the host. B chromosomes are one of these elements and frequently behave like parasitic elements triggering genomic conflicts with the standard (A) chromosomes, which constitute the host genome. Although dispensable elements, these chromosomes exhibit a selfish behavior that is based on their differential transmission (drive) by subverting Mendelian laws of inheritance, across generations, and on their deleterious effects decreasing carrier fitness. Despite the high variety of species in which B chromosomes have been reported, in few of them authors have reached a level of knowledge comparable to that of the B chromosome system in the grasshopper *Eyprepocnemis plorans*. In fact, this system was the reference on which the near-neutral model of B chromosome evolutionary dynamics was advocated. This model explains the maintenance of Bs in natural populations through drive mechanisms followed by their neutralization by the host genome. This would allocate the B chromosome to extinction due to genetic drift and natural selection against individuals with high number of Bs, unless the B polymorphism is regenerated by the appearance of a new variant being able to drive again. The motor of this evolutionary dynamics is thus the intragenomic conflict between the B chromosomes and the host genome, which is the core of this doctoral thesis.

Traditionally, the study of B chromosomes has been a subject of cytogenetics. Although this discipline has provided unparalleled contributions to the understanding of these elements, in recent years the synergy between cytogenetics, molecular techniques and bioinformatics is glimpsed as the most effective methodology to decipher the B chromosomes enigma. Thanks to the great cytogenetic characterization of B chromosomes in *E. plorans* during the past years, this thesis is hold up by a solid base on which building new knowledge about their genomic nature and delineating some molecular details of the conflict with the host genome in which they are embedded. In

fact, the B24 variant (Torrox, Spain) is our focus of study, since it shows accentuated parasitic features in comparison with other B chromosome variants in *E. plorans*. Genomics and transcriptomics are the cornerstones of this doctoral thesis as tools to understand and go deeper in the effects that a B chromosomes may have on the host genome, both derived from their own activity and from that of the A chromosomes as a response to their presence.

The first challenge to overcome consisted in revealing the molecular content of the B chromosomes in *E. plorans*, a species with a genome size of 11 Gb. Until now, only the presence of a satellite DNA (sat180pb), rDNA, 4 transposable elements and 10 protein-coding genes had been identified in the B chromosomes of *E. plorans*. With this in mind, in Chapter 1 of this thesis we characterized the satellitome of *E. plorans* using bioinformatic, molecular and cytogenetic methods. Then, applying the satMiner protocol, we uncovered 112 families of tandem repeats (TR) in the genome of *E. plorans* that were physically mapped using FISH. Among these families, we found a TR located in the centromere of all chromosomes and two families producing a specific FISH signal on B chromosomes, the two latter being useful as molecular markers for these elements. Furthermore, we performed a low-coverage MinION sequencing in *E. plorans* to unravel the correspondence between cytogenetic and molecular features of tandem repeats and satellite DNA. The results of these experiments suggest that the different families of tandem repeats in *E. plorans* differ in their molecular and cytogenetic properties. Many of them were characterized by emitting evident FISH signals in the chromosomes, presenting a high state of polymerization and homogenization of arrays as well as relatively long repeating units; which is reminiscent of the characteristics traditionally assumed for satellite DNA. However, we also found other families that present the alternative characteristics and could act as seeds from which new satellites could emerge and grow.

The characterization of the *E. plorans* satellitome is part of a *de novo* repetitive DNA database for this species (Chapter 2). For construction of this database we used the RepeatExplorer, DNApipeTE, MITObim and tRNAscan softwares. In this way, we were able to assemble hundreds of transposons, ribosomal DNA, mitochondrial DNA, several snRNAs families and thousands of tRNAs. This set of elements served as a reference to map two genomic libraries of *E. plorans* males, one with four B chromosomes and the other without them, in order to quantify the abundance of each repetitive element in the

Bs of this species. The repetitive DNA was very abundant in the B chromosomes of *E. plorans* (86.3%), in particular satellite DNA and tandem repeats made up 65% of the B chromosome. Most of the B-located repetitive sequences were also found in the A chromosomes, so we deduced an intragenomic origin of B chromosomes likely from the chromosome 9, as pointed out by the high number of elements shared between both chromosomes. In addition, the abundance of the B-located repetitive elements in individuals from Torrox (Spain) was positively correlated with that of the B chromosomes from distant populations of *E. plorans* (Tanzania, Egypt and Armenia), which supports the hypothesis of a common origin of Bs in this species. On the other hand, in the Appendix of the thesis we tested, in hominids, a method for phylogenetic reconstruction based on the abundance of repetitive DNA. This analysis established the methodology we followed to build a phylogeny with repeats abundance from individuals belonging to different populations of *E. plorans*, which reinforced the idea about the common origin of their B chromosomes.

However, this high proportion of repetitive DNA in the B chromosomes of *E. plorans* still allows place for other types of sequences, as protein-coding genes, that could play a role in the evolutionary dynamics of Bs. In this thesis, we increase up to 42 the number of protein-coding genes located on the B chromosomes of *E. plorans*, several of them being involved in functions related to the cell cycle. We found specific SNPs in the B-carrying gDNA libraries for 17 out of these 42 genes. Among them, the *ndl* gene was the one showing the highest number of SNPs, which suggests that it could be ancestral on the B chromosomes. Furthermore, we identified the *cdc16* gene in the Bs of *E. plorans* which codes for a subunit of the APC/C complex as does the *apc1* gene, located in the B chromosomes of the grasshopper *Locusta migratoria*. Most of these genes were also found in the B chromosomes of African and Asian populations of *E. plorans*, giving additional proof for a common origin of the B chromosomes in this species.

The molecular characterization of B chromosomes led to the discovery of some molecular markers, such as EplTR112-11, which allowed us to visualize the B chromosomes in *E. plorans* testis by FISH, as explained in Chapter 4 of this thesis. In this way, we show that B chromosomes are expelled during the maturation of spermatids that occurs after male meiosis, which would be a response from the host genome to get rid of these parasitic chromosomes. We believe that this is a clear sign of the intragenomic conflict caused by the B chromosomes that we also address here from a transcriptomic

point of view.

The heterochromatic nature of B chromosomes has supported assertions about the complete genetic silencing of these elements and their total inactivity. However, in the last few years, this idea has been gradually discarded in the light of transcriptomic analyzes. In the case of *E. plorans*, only the minute expression of rRNA and *cap-g* B-copies were known. In this thesis we explored the transcriptional activity of repetitive DNA sequences and protein-coding genes that we identified in the B chromosomes of *E. plorans*. For this purpose, we performed RNA-seq of embryos, legs and gonads in both sexes of 1B-carrying and B-lacking individuals. The analysis of B-specific SNPs revealed that not only the A chromosomes express copies of B-genes but also the own B-located copies are actively expressed. As we discussed in Chapter 3, this transcriptional activity occurred especially in genes involved in cell cycle functions (*cap-g*, *cip2a*, *mycb2* and *ndl*) and in gonads. In fact, we found a more pronounced expression of B-paralog copies of the gene *ndl* in ovary than that observed for the copies on the A chromosomes, which could suggest a role for these copies for B-drive during female meiosis in *E. plorans*.

Regarding the transcription of repetitive DNA in individuals carrying B chromosomes, we observed that B-located elements were, in general, more highly expressed in B-carrying individuals than in B-lacking ones, especially in gonads (Chapter 2). However, the expression of B-copies of repetitive elements, showing specific SNPs, was much lower than in the case of protein-coding genes. In fact, all repetitive elements in which we find B-specific SNPs showed rather less expression of the B-copies than of those derived from the host genome. But this does not prevent that the transcription intensity (TI) landscape of the B-located repetitive DNA changed with respect to that shown by the As. In this way, while in individuals without B chromosomes the mitochondrial DNA was full well the element that presented a higher TI (68.5%), the TI of B-specific variants in 1B-carrying individuals did not follow this trend. In this case, we observed an increase in activity derived from transposons (52.8%), tRNAs (20.5%)) or histones (12%) among other repetitive elements.

Finally, in Chapter 5, with the aim of unveiling how the host genome reacts to the presence of B chromosomes, we carried out a differential expression analysis including embryos of both sexes, with and without one B chromosome, and from two different crosses: in one of them the B chromosome was transmitted through the maternal progenitor (P1) and in the other through the paternal one (P2). In addition, we performed

the same analysis on legs and gonads from adult individuals of both sexes and, again, on individuals with and without a B chromosome. This experiment allowed us to track gene expression changes associated with the presence of B chromosomes between sexes, tissues and different ontogenetic stages. In this way, we showed that the B chromosome causes more changes in gene expression in embryos, regardless of pod origin, than those associated with sexes, probably because embryos analyzed here were still poorly sexually developed. On the contrary, in adults we observed more differentially expressed genes (DEGs) associated with sex than with the presence of B chromosomes. Moreover, the proportion of common DEGs between sexes related to the presence of one B chromosome was high in the case of embryos (P1 = 62.6%, P2 = 41.6%) but it decreased dramatically in adults (leg = 14.7%, gonad = 3%), suggesting different responses depending on sex to the intragenomic conflict caused by the Bs in adults of *E. plorans*.

Comparison of B chromosome effects between the different samples analyzed clearly revealed that the gonad, in particular the ovary, is the tissue in which the presence of Bs is associated with the highest number of DEGs, followed by embryos and then legs. Among the DEGs associated with the presence of Bs, we found transposons, protein-coding genes and some of the B-located genes, the latter showing up-regulation in the immense majority of 1B samples. In addition, while in embryos the presence of one B chromosome was associated with significant up-regulation of transposons (P1 = 40% of DEGs, P2 = 48%) which could reflect a possible response to the stress caused by B chromosomes, in adults, most DEGs corresponded to protein-coding genes (leg = 24% of DEGs, gonad = 51%). In fact, in ovary, we found down-regulation of many of these genes (230 DEGs), which was missing in the remaining samples, and could represent a host response to deal with B drive which takes place precisely in this organ.

Throughout this doctoral thesis we have deciphered much of the molecular content of B chromosomes in *E. plorans*, using bioinformatic, molecular and cytogenetic tools, which have allowed us to address a challenge that remained unattainable. In addition, using these tools we have also contributed to clarify the differences between satellite DNA and tandem repeats after the characterization of the *E. plorans* satellitome, which also constitutes part of a *de novo* repetitive DNA database in this species. Based on this genomic approach, we have revealed that B chromosomes contain dozens of protein-coding genes that show active transcription, especially in ovary. However, in B-carrying individuals, not only the B-located genes are expressed, as there is also a forceful

response from the host genome, presumably addressed to manage the intragenomic conflict caused by these selfish elements. In fact, the results included here draw a scene comprising B chromosome elimination in males but accumulation though females of *E. plorans*, which could have its molecular mirror on the high expression of B-genes and the under-expression of some genes in the host genome that we observe in ovary. These results constitute the first window opened to unveil the molecular details of this intragenomic conflict.

Resumen

El genoma eucariota, además de albergar elementos genómicos canónicos, es vulnerable a la presencia de otros componentes que pueden ser accesorios e incluso perjudiciales para el hospedador. Los cromosomas B son uno de estos elementos y frecuentemente se comportan como parásitos, provocando un conflicto genómico con los cromosomas estándar (A) que constituyen el genoma hospedador. Aunque dispensables, estos cromosomas exhiben un carácter egoísta que se fundamenta en su transmisión diferencial (impulso o *drive*), subvirtiendo las leyes Mendelianas de la herencia generación tras generación, y en sus efectos deletéreos reduciendo la *fitness* de los individuos portadores. A pesar de la gran variedad de especies en las que se han descrito cromosomas B, en pocas de ellas se ha alcanzado un grado de conocimiento equiparable al del sistema de cromosomas B del saltamontes *Eyprepocnemis plorans*. De hecho, este sistema sentó las bases del modelo evolutivo casi-neutro, que explica el mantenimiento de los cromosomas B en poblaciones naturales mediante mecanismos de *drive* que tienden a ser neutralizados por el genoma hospedador. Esto condenaría al cromosoma B a la extinción por deriva genética y selección contra la presencia de números elevados de cromosomas B, por sus elevados efectos deletéreos, a menos que se produjera la regeneración del polimorfismo para el B mediante la aparición de una nueva variante que recuperase el *drive*. El motor de esta dinámica evolutiva es el conflicto intragenómico generado entre los cromosomas B y el genoma hospedador, que constituye el eje de esta tesis doctoral.

Tradicionalmente, el estudio de los cromosomas B ha sido materia de la citogenética. Aunque esta disciplina ha contribuido sin parangón a la comprensión de estos elementos, en los últimos años las sinergia entre la citogenética, las técnicas moleculares y la bioinformática se vislumbra como la metodología más eficaz para descifrar el enigma de los cromosomas B. En este sentido, los cromosomas B de *E. plorans* han sido caracterizados citogenéticamente con bastante detalle por lo que partimos de una base sólida sobre la que construir nuevo conocimiento acerca de su

naturaleza genómica y la del conflicto con el genoma hospedador en el que están inmersos. De hecho, la variante B24 (Torrox, España) es nuestro foco de estudio aquí, ya que presenta características parasíticas acentuadas en comparación con otras variantes en *E. plorans*. Por otro lado, la genómica y la transcriptómica son las piedras angulares de esta tesis doctoral con la que ampliamos y profundizamos sobre los posibles efectos que pueden tener los cromosomas B en el genoma hospedador, tanto derivados de su propia actividad génica como de la que inducen en el genoma hospedador como respuesta a su presencia.

El primer reto a superar comprendió conocer el contenido molecular de los cromosomas B de *E. plorans*, una especie con 11 Gb de genoma, puesto que hasta el momento solo se había descrito en ellos la presencia de un ADN satélite (sat180pb), ADNr 45S, 4 elementos transponibles y 10 genes codificantes de proteínas. Con este objetivo en mente, en el Capítulo 1 de esta tesis caracterizamos el satelitoma de *E. plorans* utilizando herramientas bioinformáticas, moleculares y citogenéticas. Así, aplicando el protocolo satMiner identificamos 112 familias de repeticiones en tándem (TRs) en el genoma de *E. plorans* que además mapeamos físicamente mediante FISH. Entre estas familias se encuentra un TR localizado en el centrómero de todos los cromosomas, así como dos familias que producen señal de FISH solo en los cromosomas B, por lo que estas últimas pueden utilizarse como marcadores moleculares de estos elementos. Además, realizamos una secuenciación MinION de baja cobertura en *E. plorans* para desentrañar la correspondencia entre las características citogenéticas y moleculares de las repeticiones en tándem y el ADN satélite. Los resultados que arrojaron estos experimentos apuntan a que las distintas familias de repeticiones en tándem en *E. plorans* difieren en sus propiedades moleculares y citogenéticas. Muchas de ellas se caracterizaron por emitir señales de FISH evidentes en los cromosomas, presentar un alto estado de polimerización y homogenización de arrays así como unidades de repetición relativamente largas; lo que recuerda a las características asumidas tradicionalmente para el ADN satélite. Sin embargo, también encontramos otras familias que presentan las características alternativas y podrían actuar como semillas para la formación y crecimiento de nuevos satélites.

La caracterización del satelitoma de *E. plorans* se enmarca dentro de la construcción *de novo* de una base de datos de ADN repetitivo para esta especie (Capítulo 2). Para la elaboración de esta base de datos utilizamos los programas RepeatExplorer, DNApipeTE,

MITObim y tRNAscan. De esta forma pudimos ensamblar cientos de transposones, el ADN ribosómico, el mitocondrial, varias familias de snRNAs y miles de tRNAs. Esta referencia nos sirvió para mapear dos librerías genómicas de machos de *E. plorans*, uno con cuatro cromosomas B y otro sin ellos, con el fin de cuantificar la abundancia de cada elemento repetitivo en sus cromosomas B. El ADN repetitivo resultó ser muy abundante en los cromosomas B de *E. plorans* (86,3%), en particular el ADN satélite y las repeticiones en tándem constituyeron el 65% del cromosoma B. La mayor parte del ADN repetitivo localizado el estos cromosomas se encontró también en los As, por lo que deducimos un origen intragenómico de los cromosomas B posiblemente a partir del cromosoma 9, como apunta el alto número de elementos compartidos entre ambos cromosomas. Por otro lado, la abundancia de los distintos elementos de ADN repetitivo en el cromosoma B de Torrox (España) se correlacionó positivamente con la de cromosomas B provenientes de poblaciones distantes de *E. plorans* (Tanzania, Egipto y Armenia) lo que apoya la hipótesis de un origen común de los Bs en esta especie. Además, en el Apéndice de la tesis testamos en homínidos un método de reconstrucción filogenética basado en la abundancia de repetitivo. A partir de este análisis establecimos la metodología a seguir para construir una filogenia con abundancia de repetitivo en individuos pertenecientes a diferentes poblaciones de *E. plorans*, lo que reforzó la idea del origen común de los cromosomas B.

Esta elevada proporción de ADN repetitivo en los cromosomas B de *E. plorans* aún deja margen de espacio para que existan otro tipo de secuencias localizadas en ellos, como genes codificantes de proteínas, que puedan desempeñar alguna función en la dinámica evolutiva de los Bs. En esta tesis aumentamos hasta 42 el número de genes codificantes de proteínas localizados en los cromosomas B de *E. plorans*, estando varios de ellos involucrados en funciones relacionadas con el ciclo celular. Además, hemos encontrado SNPs específicos de las librerías con cromosomas B para 17 de estos 42 genes, siendo *ndl* el gen que presentó el número más alto de SNPs lo que sugiere que puede ser ancestral en los cromosomas B. Además, otro de los genes que identificamos en los cromosomas B fue *cdc16* que codifica para una subunidad del complejo APC/C, al igual que hace el gen *apc1* localizado en los cromosomas B del saltamontes *Locusta migratoria*. La mayor parte de estos genes se encontraron también en los cromosomas B de poblaciones africanas y asiáticas de *E. plorans* lo que apunta de nuevo a un origen común de los cromosomas B en esta especie.

La caracterización molecular del cromosomas B incluyó la identificación de marcadores, como EplTR112-11, que nos permitieron visualizar los cromosomas B en testículos de *E. plorans* mediante FISH, tal como explicamos en el Capítulo 4 de esta tesis. De esta forma demostramos que los cromosomas B son expulsados de la célula durante la maduración de las espermátidas, que se produce tras la meiosis masculina, lo que podría suponer una respuesta del genoma hospedador para deshacerse de estos cromosomas parásitos. Con estos resultados, nos adentramos progresivamente en el conflicto intragenómico causado por los cromosoma B que abordamos también mediante la transcriptómica.

La naturaleza heterocromática de los cromosomas B ha sustentado aseveraciones sobre el completo silenciamiento de estos elementos y su total inactividad. Sin embargo, esta idea se ha descartado recientemente a la luz de los análisis transcriptómicos. En el caso de *E. plorans* solo se conocía la expresión ínfima de copias de rRNA y del gen *cap-g* localizadas en los cromosomas B. En esta tesis doctoral analizamos la actividad transcripcional de las secuencias de ADN repetitivo y genes codificantes de proteínas localizadas en los cromosomas B de *E. plorans*, y para ello llevamos a cabo RNA-seq de embriones, patas y gónadas de ambos sexos, en individuos con un cromosoma B y sin él. Mediante el uso de SNPs específicos de los cromosomas B hemos analizado la actividad transcripcional de éstos que ha revelado que muchos de los genes que contienen los Bs se expresan activamente. Como exponemos en el Capítulo 3, esta actividad transcripcional se produce especialmente en genes involucrados en funciones relativas al ciclo celular (*cap-g*, *cip2a*, *mycb2* y *ndl*) y en gónadas. De hecho, para el gen *ndl* encontramos, en ovario de hembras con B, una expresión mucho más acentuada de las copias parálogas de los cromosomas B que de las copias estándar de los cromosomas A, lo que sugeriría que alguno de estos transcritos del B podría tener un papel relevante en el *drive* del propio B durante la meiosis femenina de *E. plorans*.

Respecto a la expresión de ADN repetitivo en individuos con cromosomas B, hemos observado que los elementos localizados en estos cromosomas también se expresaron más en individuos con cromosomas B que en los que no los tenían, especialmente en gónada (Capítulo 2). Sin embargo, la expresión de copias de repetitivo procedentes de los cromosomas B (que mostraron SNPs específicos) fue mucho menor que para los genes codificantes de proteínas. De hecho, todos lo elementos en los que encontramos SNPs específicos del cromosoma B presentaron una expresión mucho menor de las

copias específicas del B que de las derivadas del genoma hospedador. Pero esto no impide que el paisaje de intensidad transcripcional (TI) del ADN repetitivo localizado en el cromosoma B presente un cambio respecto al que muestran los As. De esta manera, mientras que en los individuos sin cromosomas B el ADN mitocondrial fue, con notable diferencia, el elemento que presentó una TI mayor (68.5%), la TI de las copias específicas del B en los individuos portadores no siguió esta tendencia. En esta caso observamos un incremento en la actividad derivada de transposones (52,8%), tRNAs (20.5%) o histonas (12%) entre otros elementos repetitivos.

Finalmente, en el Capítulo 5, con el ánimo de profundizar sobre los efectos de los cromosomas B en el genoma de *E. plorans* llevamos a cabo un análisis de la expresión diferencial incluyendo embriones de ambos sexos, con y sin un cromosoma B, provenientes de dos cruces diferentes: en uno de ellos el cromosoma B se transmitió por vía materna (P1) y en el otro por vía paterna (P2). Además, realizamos el mismo análisis en patas y gónadas de adultos para ambos sexos, y de nuevo, en individuos con un cromosoma B o sin él. Este experimento nos ha permitido rastrear los cambios de expresión génica asociados a la presencia de cromosomas B entre sexos, tejidos y distintas etapas ontogénicas. De esta forma, demostramos que el cromosoma B provoca en embriones de ambas puestas más cambios de expresión génica que los asociados al sexo, probablemente debido a que los embriones estudiados estaban aun poco avanzados en la diferenciación sexual. En adultos, al contrario, observamos más genes diferencialmente expresados (DEGs) asociados al sexo que a la presencia de cromosomas B. Además, la proporción de DEGs comunes entre sexos asociados a la presencia de un cromosoma B fue elevada en el caso de embriones (P1= 62,6%, P2= 41,6%) pero disminuyó drásticamente para adultos (pata= 14,7%, gónada= 3%), lo que sugiere, en este último caso, respuestas distintas en función del sexo frente al conflicto intragenómico causado por el B en adultos de *E. plorans*.

La comparación del efecto del cromosoma B entre distintas muestras reveló con notoriedad que es la gónada, y en particular el ovario, el tejido en el que la presencia de los Bs se asocia a un mayor número de DEGs, seguido de embriones y patas. Entre los genes diferencialmente expresados en presencia del B encontramos transposones, genes codificantes de proteínas y los propios genes localizados en ellos, presentándose estos últimos sobreexpresados en la mayoría de las muestras con un cromosoma B. Además, mientras que en embriones la presencia de los cromosomas B tiene como respuesta una

sobreexpresión importante de transposones (P1= 40% de DEGs, P2= 48%), lo que reflejaría una posible respuesta al estrés causado por los Bs, en adultos la mayor parte de los genes diferencialmente expresados son codificantes de proteína (pata= 24% de DEGs, gónada= 51%). De hecho, en ovario encontramos una infraexpresión de muchos de estos genes (230 DEGs), que no tiene lugar en otras muestras, y que podría suponer una respuesta del genoma hospedador frente a los mecanismos de acumulación que presenta el cromosoma B de *E. plorans* precisamente en este tejido.

A lo largo de esta tesis doctoral desciframos gran parte del contenido molecular de los cromosomas B de *E. plorans* valiéndonos de herramientas bioinformáticas, moleculares y citogenéticas que nos han permitido abordar un reto que permanecía inalcanzable. Además, con estas herramientas también contribuimos a clarificar las diferencias entre el ADN satélite y las repeticiones en tándem a partir del análisis del satelitoma de *E. plorans,* que forma parte de una nueva base de datos de ADN repetitivo para esta especie. Apoyados en esta base genómica, desvelamos que los cromosomas B contienen decenas de genes codificantes de proteínas que además presentan una expresión activa, especialmente en ovario. Sin embargo, en los individuos con cromosomas B no solo se expresan los genes que éstos contienen, también se produce una respuesta contundente por parte del genoma hospedador, probablemente encaminada a gestionar el conflicto intragenómico provocado por estos elementos egoístas. De hecho, los resultados que presentamos en esta tesis dibujan un escenario en el que los cromosomas B se eliminan en los machos de *E. plorans* pero se acumulan en las hembras, lo que podría tener su espejo molecular en la alta expresión de copias de genes del B y la infraexpresión de algunos genes de los As que observamos sobre todo en ovario, retratando las bases moleculares de un verdadero conflicto intragenómico.

Introduction

B chromosomes as enigmatic elements of genomes

Conflict is not alien to us, it exists in nature in a broad variety of ways and meanings. The eukaryotic genomes, apart from the set of genes found in standard (A) chromosomes, also harbor a huge amount of selfish genetic elements which main function is to ensure their own transmission in spite of reducing, in some cases, the carrier's fitness. These selfish genetic elements usually get transmission advantages by disobeying Mendelian laws of inheritance, with transmission rates higher than those of the standard chromosomes (the expected 0.5 value). Transposable elements, segregation distorters, several cytoplasmic factors and B chromosomes are some of the best-known selfish genomic elements (Camacho et al., 2000).

Since their discovery more than a century ago (Wilson, 1907), the definition of B chromosomes has not been out of controversy, but currently there is a general consensus following the proposal that Camacho and Parker stated during the First B-Chromosome Conference (1993): “They are dispensable chromosomes present in some individuals from some populations of some species, that have probably arisen from the A chromosomes but not recombining with them, thus following their own evolutionary pathway" (see Beukeboom, 1994). Furthermore, B chromosomes are also characterized by presenting mechanisms of accumulation (Östergren, 1945), mitotic and/or meiotic, which allow them to segregate in a non Mendelian way and transmit with higher rates than that of A chromosomes (Jones, 1985).

B chromosomes have been described in many species of eukaryotes, being found in a total of 2,087 species of plants, 736 species of animals and 14 species of fungi (D'Ambrosio et al., 2017; last accessed October 2020). There are species that tolerate a high number of B chromosomes, such as the mouse *Apodemus peninsulae* in which Volobujev and Timina (1980) characterized individuals harboring up to 24 B

chromosomes. However, in natural populations, it is not that frequent to find individuals with more than three or four B chromosomes (Camacho et al., 2000).

Drive mechanisms

A defining feature of B chromosomes is that they are not usually transmitted in a Mendelian way because they have mechanisms that allow for their accumulation though the offspring. In general, this accumulation process can occur before, during or after meiosis.

Premeiotic mechanisms act during the embryonic development, due to mitotic nondisjunction and preferential destination of cells with the highest number of B chromosomes passing towards the germ line. This type of behavior has been described in grasshoppers such as *Calliptamus palaestinensis* (Nur, 1963), *Camula pellucid*a (Nur, 1969) and *Locusta migratoria* (Viseras et al., 1990) and for the B chromosomes of the plant *Crepis capillaris* (Rutishauser and Röthlisberger, 1966).

Meiotic accumulation usually takes place during female meiosis. It is based on asymmetry of the meiotic products, so that the B chromosomes preferentially segregate to the oocyte instead of to the polar corpuscle. This accumulation mechanism has been observed in some orthopteran species such as *Myrmeleotettix maculatus* (Hewitt, 1976) or in our species under study, *Eyprepocnemis plorans* (Zurita et al., 1998; Bakkali et al., 2002), and also in plants such as *Lillium callosum* (Kayano, 1957). Recently, experimental observations by Akera et al. (2017) have put into evidence that selfish elements take advantage of the female spindle asymmetry to bias their transmission. They found that CDC42 signaling, depending on cell polarization guided by chromosomes, play a fundamental role in asymmetry and drive. Congruent with these findings, the drive of B chromosomes in the plant *Aegilops speltoides* is directed by nondisjunction and unequal spindle organization as empirically confirmed by Wu et al. (2019).

Postmeiotic accumulation occurs mainly in plants, during the maturation of the male gametophyte. In this way, the B chromosomes undergo non-disjunction and the two chromatids preferentially migrate to the generative nucleus. This phenomenon can occur during the first mitosis of the pollen grain, as in *Fetusca pratensis* (Bosemark, 1954), or in the second mitosis, as it happens with corn (Roman, 1974). This mechanism was first proposed in rye (Hasegawa, 1934).

There are other types of mechanisms of B chromosome accumulation that are

derived from its own effects, as it occurs for the B chromosome of the wasp *Nasonia vitripennis.* This particular B chromosome is called PSR (Paternal Sex Ratio) because its presence affects the sex ratio by increasing the proportion of males. Like other Hymenoptera, this wasp is haplodiploid, with haploid males and diploid females. The PSR chromosome causes condensation and inactivation of the paternal chromosome set transforming the fertilized eggs, which would have developed into females in the absence of Bs, in haploid males carrying B chromosome. Therefore, the transmission rate of this chromosome is close to 100% (Werren, 1991). Very recently, Dalla Benetta et al. (2020) identified the presence of the *haploidize*r gene in the PSR which is active in testis and its expression causes the paternal genome elimination, thus the female-to-male conversion.

We could also consider the GRC (germline restricted chromosomes) and the L chromosomes found in several species as special cases of B chromosomes. The L chromosomes (germline limited) are eliminated from the soma and in *Sciara coprophila* are though to be indispensable despite their lost in other related species (Singh and Belyakin, 2018). The GRC is widespread among songbirds and appears only in the female germline (expelled from the nucleus during spermatogenesis) since it is eliminated from somatic cells during the embryonic development (Torgasheva et al., 2019). Therefore, these chromosomes would show a highly-efficient premeiotic accumulation, ensuring its own transmission while reducing the cost of their presence in somatic tissues. Interestingly, the GRC of the zebra finch contains active crucial genes for sexual development that are consequently eliminated from somatic cells where the GRC is absent (Kinsella et al., 2019). These recent discoveries open the door to an outlook where the gene content of B chromosome would play a critical role for its own maintenance and accumulation.

Evolutionary dynamics of B chromosomes

The drive exhibited by the B chromosomes increases their rate of transmission through the offspring, which is determinant for their evolutionary success. The interplay between drive (i.e. transmission ratio) and the effects of B chromosomes in the carrier individuals is the basis for most of the models explaining the evolution of B chromosomes in natural populations (for review see Camacho et al., 2000). In fact, in absent of drive, the spread of B chromosomes could only be explained by putative beneficial effects to the carriers, otherwise they will disappear through mutation and drift.

There are few cases of B chromosomes causing benefits to the host. One example is the case of the plant *Allium schoenoprasu* in which B-carrying individuals show no B-drive and better germination rates under stressful conditions than those specimens lacking B chromosomes (Plowman and Bougourd, 1994). This case represents an heterotic model for the evolution of B chromosomes (White, 1973) which is based in a balance between the positive effects of B chromosomes (which do not drive) when they appear in low numbers and their detrimental effects when they are abundant. Also, the well-known B chromosome of *Nectria haematococca,* lending carrier individuals resistance to pisatin (Miao et al., 1991), could be an interesting candidate to behave as the heterotic B chromosomes. However, it still shows drive, a non-expected feature for a heterotic B chromosome, otherwise the frequency of Bs will hugely increase becoming a burden for the host.

However, most of the B chromosomes described so far show parasitic behavior by means of drive and negative effects for carrier individuals. Therefore, the parasitic model for the evolution of Bs (Östergren, 1945; Jones, 1985; Shaw and Hewitt, 1990) hypothesizes that the B chromosome dynamics is set up in the balance between B accumulation through drive and B decrease due to their detrimental effect in carriers. This assumption was empirically confirmed in several B chromosome systems as for Bs in *Myrmeleotettix maculatus* (Shaw and Hewitt, 1985) and *Pseudococcus affinis* (also *P. obscurus;* Nur and Brett, 1985). Interestingly, in many cases this equilibrium between accumulation and negative effects tips the scales in favor of the host by buffering negative effects or suppressing drive. This fact could even be positive for B chromosomes maintenance as it is claimed for other vertically transmitted parasites reducing their virulence (Lipsitch et al., 1995). Therefore, the B chromosome systems evolve towards the generation of less harmful B variants or drive suppression (Nur and Brett, 1988; Jimenez et al., 1995; Perfectti et al, 2004), usually the latter leading to the former and both resulting in a neutralized B chromosome by the host genome. Therefore, the accumulative behavior of B chromosome that makes them appear as genomic parasites is not found always and continuously in natural populations, thus somehow hiding their selfish side.

Along years, authors have described several B chromosome systems apparently representing different evolutionary stages. These findings serve as a clue for the proposal of a near-neutral model for B chromosome dynamics at a long term view where

all these evolutionary stages could be interlinked. The existence of long periods in which the B chromosomes show a stable frequency in many populations is the result of a balance between the increase in frequency due to the accumulation of Bs and the decrease produced by the selection against the carrier individuals (Bakkali et al., 2002; Perfectti et al., 2004; Voltolin et al., 2010; Lanzas et al., 2018). This balance can be altered by the evolution, in the host genome, of suppressor genes of the accumulation of Bs (hypothesized for *E. plorans* by Herrera et al., 1996), approaching its transmission rate to the Mendelian 0.5 value. This causes the B chromosomes to become quasi-neutral elements that can only evolve by drift and selection against carriers with many Bs, leading to the extinction of B chromosomes or their transformation to new variants. With this in mind, Camacho et al. (1997) proposed a biological cycle for the long-term evolution of parasitic chromosome using the B-system of *E. plorans* as case of study.

This model could be divided in three stages. First of all, the invasion stage occurs in a few generations thanks to the accumulation of B chromosomes. In the second stage, the host genome reacts by suppressing the accumulation of B chromosomes. This quasi-neutral stage is the longest one, lasting tens or hundreds of generations, and during which the B chromosomes evolve by drift and selection against individuals with many Bs. As the latter acts on very rare individuals in populations, the disappearance of Bs is a very slow process. During the third stage, mutations can occur on the B chromosomes that confer again the parasitic character. Therefore, in that moment, the B chromosomes will be able to accumulate and start a new cycle after this regeneration phase (Zurita et al., 1998). This cycle reflects an arm race between the B chromosomes and the host genome where probably a Red Queen dynamic (Van Valen, 1973) is taking place at intragenomic, individual and/or population levels.

Molecular structure and composition of B chromosomes

B chromosomes are usually the same size as those of the standard complement (As), but there are examples of B chromosomes larger than the A chromosomes, as in the case of those in the plant *Plantago lagopus* (Dhar et al., 2002). In the opposite case, there are also B chromosomes that are the smallest of the genome or mini B chromosomes such as those found in the dichromosomatic Brachycome plant (Houben et al., 1999) or the fish *Poecilia formosa* (Schartl et al., 1995). Schmid et al. (2006) have described, in the *Alburnus alburnus* fish one of the largest B chromosome found in a vertebrate, accounting for almost 10% of the genome size, although the GRC of the zebra finches

surpasses this proportion (Smith et al., 2018).

First cytogenetic approaches to unveil the molecular content of B chromosomes revealed a heterochromatic element. Later on, molecular analysis confirmed that B chromosomes harbor a diversity of repetitive DNA sequences, mainly ribosomal DNA (rDNA), mainly satellite DNA (satDNA) and transposable elements (TEs) (Camacho et al., 2000; see Table 4.2 of Camacho, 2005). In some cases, these repetitive elements may be present also in A chromosomes, as in *Crepis capillaris* (Jamilena et al., 1994) or the repetitive unit pSsP216 in the fly *Drosophila subsilvestris* (Gutknecht et al., 1995). Alternatively, that repetitive DNA can be specific to the B chromosome, as it is the case of a satellite DNA found in the PSR chromosome (Paternal Sex Ratio) from *Nasonia vitripennis* (Eickbush et al., 1992) or various satDNAs described in rye (Klemme et al., 2013).

Due to the lack of recombination with the rest of the genome, the B chromosome could be an ideal place for accumulation of mobile elements (Beukeboom, 1994; Camacho et al., 2000). The presence of transposons has been detected in several B chromosomes (see Table 4.2 of Camacho, 2005). The most striking case is that of the PSR chromosome of *N. vitripenis*, which is practically invaded by the retrotransposon NATE ("Nasonia Transposable Element") (McAllister, 1995; McAllister and Werren, 1997). Therefore, transposable elements may be involved in the origin and evolution of B chromosomes. Lamb et al. (2007) reported one of the first cases about the formation of a specific element in the B chromosomes of maize, named StarkB. The StarkB element is located in the heterochromatic region of B chromosomes and it is transcriptionally active. Interestingly, it is composed of sequences derived from a LTR-type (i.e. Long Terminal Repeat) mobile element present in the A chromosomes and also by specific sequences from the B chromosome. The insertion of mobile elements on the B chromosomes could truncate, interrupt or alter in certain ways the order of other sequences, such as protein-coding genes, located in the B chromosome, thus affecting their potential activity (Camacho et al., 2000). Very recently, Shams and Raskina (2020) have found that the content of some TEs and a TR widely varies between populations of the plant *Aegilops speltoides* and also between B-lacking and B-carrying specimens of the same population, usually being more abundant in the latter case. Transposable elements may also play a role in ectopic recombination (Montgomery et al., 1991), favoring the transfer of sequences between different chromosomes, and may even be

involved in the movement of DNA between cytoplasmic organelles and the B chromosome. For example, in *Brachycome dichromosomatica*, the presence of chloroplast DNA on the B chromosome has been explained by the activity of the Bd49 retrotransposon (Franks et al., 1996).

The arising of high-throughput sequencing technologies allowed researchers to depict a more detailed analysis about some repetitive DNA elements of B chromosomes (Cheng and Lin, 2003; Bugrov et al., 2007; Marques et al., 2018). Unfortunately, there is still a gap regarding the quantitative analysis of the complete repetitive DNA landscape of B chromosomes, likely because of the stickiness found in assembling a reference for the repetitive fraction of genomes. However, some authors have recently made great efforts to address these difficulties as Ruiz-Ruano et al. (2018) rendering the repetitive landscape of B chromosomes in *L. migratoria* or Ebrahimzadegan et al. (2019) for that in the plant *Festuca pratensis*.

In contrast to the accepted repetitive composition of Bs, it has been stated for a long time that B chromosomes do not contain genes (Camacho et al., 2000; Jones and Houben, 2003; Burt and Trivers, 2006). This is probably due to the difficulties of detecting them efficiently and the high enrichment of repetitive elements that can hide any signal of single copy genes or sequences found in low copy number (Navarro-Domínguez, 2016b). In addition, due to their dispensable nature, which frees them from selective pressure, and the absence of recombination with A chromosomes, DNA sequences in B chromosomes are under a context of the Muller's ratchet (Muller, 1964). Therefore, it is plausible to anticipate that putative gene sequences located in Bs, most of them being residues from the ancestral A chromosome from which B chromosome arose, will show a high number of mutations with respect to their homologous A chromosome paralogs, and will be highly fragmented by insertions of repetitive DNA (Green, 1990).

However, a clear case of the existence of single copy genes came out around thirty years ago in the B chromosome of the fungus *Nectria haematococca*. The activity of that B-gene confers, to the carrier individuals, resistance against pisatin, a cytoalexin produced by the host plant, thus promoting the fungal pathogenicity in carrier individuals (Miao et al., 1991). Things changed from 2005 onwards, authors found complete and fragmented protein-coding genes in the B chromosomes of several species (Graphodatsky et al., 2005; Teruel et al., 2010; Yoshida et al., 2011; Martis et al., 2012; Trifonov et al., 2013; Banaei-Moghaddam et al., 2013; Valente et al., 2014; Huang et al.,

2016; Carmello et al., 2017; Ma et al., 2017; Navarro-Domínguez et al., 2017a,b; Ruiz-Ruano et al., 2019). This year, Ahmad et al. (2020) have published a paper in which they applied a bioinformatic protocol to extensively identify B-located genes in several species. Interestingly, they found that the function of the B-located genes could be related to the evolutionary process of B chromosomes in most of the species included in the study. However, they did not address any question about the transcriptional activity of these B-genes.

Due to their heterochromatic nature, B chromosomes were thought to be transcriptionally inactive. There are, however, some cases where activity has been observed in these chromosomes, such as the B chromosome present in the frog *Leiopelma hochstetteri* when they are in lampbrush state, the polygenic B chromosome of the mosquito *Simulium juxtacrenobium* (Brockhouse et al., 1989) and a neoB in the wasp *Nasonia vitripennis* (Perfectti and Werren, 2001). Up until now, evidences about transcription activity from different B chromosome systems have not stopped growing (Leach et al., 2005; Van Vugt et al., 2005; Carchilan et al., 2007; Ruiz-Estévez et al., 2012; Zhou et al., 2012; Trifonov et al., 2013; Banaei-Moghaddam et al., 2013; Valente et al., 2014; Huang et al., 2016; Ma et al., 2017; Navarro Domínguez et al., 2017b; Ruiz-Ruano et al., 2019; Hong et al., 2020). Therefore, the general thought that parasitic B chromosomes are genetically inert elements (Camacho et al., 2000) is outdated in light of the extensive findings of transcription from B chromosome sequences. Now, the main dilemma states in ascertaining the extent to which there is expression coming from the B chromosomes that could affect their own evolutionary dynamic or/and that of the host, or in the contrary, expression of B-located sequence is simply fluff (Dalla Benetta et al., 2019).

Effects of B chromosomes

The effect that a B chromosome could have on carrier individuals is one of the most important and controversial issues around these genomic elements. First of all, it should be noted that the presence of the B chromosomes implies an increase in the amount of genomic DNA that will inevitably use the nuclear machinery for its replication. Therefore, an increase in cell and nuclear size can be expected, as well as a longer duration of the cell cycle in carrier individuals. These effects have been described in several species (Jones and Rees, 1982). On the other hand, B chromosomes rarely produce visible effects on carrier individuals. Despite this consideration, there are some exceptions such as the *Haplopappus gracilis* plant where the presence of the B chromosome changes the color

of the achenes (Jackson and Newmark, 1960) or the maize where individuals carrying B chromosomes show streaked leaves (Staub, 1987). The effect of B chromosomes on carrier individuals may also be due to the activity of genes they harbor, such as the cases of the *N. haematococca* fungus, with its antipisatin gene (Miao et al., 1991), and *N. vitripennis,* with its neo-B containing the *123+* gene that affects the color that is expressed in the wasp's eyes (Perfectti and Werren, 2001). B chromosomes can also have an effect on endophenotypic characters, such as amount of proteins, RNA, the expression of NORs (Nucleolar Organizing Regions) or the chiasma frequency. For example, Kirk and Jones (1970) observed that the amount of RNA and nucleolar proteins decreased when the number of B chromosomes increases. Also, in the grasshopper *E. plorans*, the expression of the NORs from the A chromosomes is affected by the presence of B chromosomes, so that the number of them that are active is greater in 1B individuals than in 0B individuals (Cabrero et al., 1987). At this molecular level, the presence of B chromosomes has been shown to alter the expression of certain genes located in the standard set as well as in the own B chromosomes (Bergerard et al., 1972; Ruiz-Rejón et al., 1980; Oliver et al., 1982; Plowman and Bougourd, 1994; Akbari et al., 2013; Huang et al., 2016; Navarro-Domínguez et al., 2019; Hong et al, 2020; see previous section for more references).

However, the most far-reaching effects of B chromosomes are those that affect characters related to biological efficacy such as vigor and fertility. In several species, authors reported a decrease in those features in carrier individuals, so that the B chromosomes can be considered genomic parasites (Camacho et al., 2000).

Very recently, Ruban et al. (2020) described in the plant *A. speltoides* the programmed elimination of B chromosomes from roots as a process to avoid the possible harmful effects derived from the expression of the B-content in that tissue. If we delve into this question, the recent findings of active genes in the B chromosomes operating during their drive in the host (Kinsella et al., 2019; Ruiz-Ruano et al., 2019; Dalla Benetta et al., 2020) lead to considered the accumulation of Bs as an unavoidable effect of the B-genes expression.

Origin of B chromosomes

B chromosomes presumably originate from the A chromosomes of the same species where they are found, i.e. intraspecific origin, or from a related species after

hybridization between them, i.e. interspecific origin (see Camacho, 2005). These accessory chromosomes could ultimately be, for instance, a by-product of karyotypic evolution, originating from processes such as the polysomy of an A chromosome, from fragments containing centromeres resulting from Robertsonian translocations (Hewitt, 1974) or from amplification of the paracentromeric region of an A chromosome (Key & Hägele, 1971).

The intraspecific hypothesis is supported by the existence on the B chromosomes of DNA sequences that also appear in the A chromosomes. For example, this hypothesis can be applied to explain the origin of the B chromosomes in *Crepis capillaris* since the microdissection of the B chromosome showed that B-located sequences were also found in other standard chromosomes (Jamilena et al., 1994, 1995). The intraspecific origin of Bs is widespread in scientific literature, for example, the case of Bs in *Drosophila melanogaster* (Hanlon et al., 2018), the migratory locust (Teruel et al., 2009; Ruiz-Ruano et al., 2018) or the plant *A. speltoides* (Ruban et al., 2020) where the molecular content of the B chromosomes has served as a tool to identify the most likely ancestral chromosome of the standard set from which the B derived. Furthermore, the own B chromosome could be involved in the arising of sex chromosomes, as proposed by Conte et al. (2020) for the origin of the sex megachromosome in Oreochromini fishes, or viceversa, such as the B chromosomes of *Characidium gomesi* derived from the W sex chromosome (Pansonato-Alves et al., 2014).

B chromosomes can also originate as a result of hybridization processes between related species or subspecies (Battaglia, 1964). The clearest evidence in favor of this theory is the existence of B-specific sequence that are not found in the host genome but that have certain homology with sequences from the genome of the related species. The most documented example is that of the PSR chromosome in the *Nasonia* wasp. In this case, a phylogenetic analysis of a retroelement showed that sequences located on the PSR chromosome, were more similar to the copies present in related species of the genus *Trichomalopsis* than to existing copies in *Nasonia*'s own genome (McAllister and Werren, 1997). More recently, an interspecific origin of B chromosomes in the frog *Hypsiboas albopunctatus* was also discussed based on the exclusive hybridization of the B chromosomes when using probes obtained from their own microdissection (Gruber et al., 2014). Furthermore, this theory has been empirically demonstrated by observing the *de novo* formation of a B chromosome through controlled crosses between related

species (Sapre and Deshpande, 1987; Shartl et al., 1995; Perfectti and Werren, 2001).

Another point to consider regarding the origin of B chromosomes is whether they arose in a common origin of related species/subspecies/populations or, in the contrary, their origin was independent. There are several cases of common origin of Bs, for instance, the B chromosomes of rye (Martis et al., 2012; Marques et al., 2012). Likewise, the B chromosomes in the fish genus *Astyanax* would also have arisen from a common origin in light of the results found by Silva et al., (2016) after cytological and molecular analysis of repetitive sequences. In addition, the B chromosomes of *A. latifasciata* and Lake Victoria cichlids may have a common origin as suggested by Valente et al. (2014). Rajičić et al. (2017) described a common origin for the B chromosomes in geographically distant populations of the mouse *Apodemus flavicollis.* This idea was also true for the origin of the canid B chromosome, harboring all of them the proto-oncogene cKIT (Graphodatsky et al., 2005; Becker et al., 2011). In contrast to this idea for the origin of Bs, an independent origin was found by Makunin et al. (2016) for two B-chromosome-carrying members of the Cetartiodactyla, the Siberian roe deer (*Capreolus pygargus*) and the grey brocket (*Mazama gouazoubira*). Other cases of independent origin of B chromosomes have been described regarding those of some *Characidium* species (Serrano et al., 2017) or in some grasshopper species of the genus *Podisma* (Bugrov et al., 2007). Furthermore, a multiple origin of different B chromosome variants within the same species was described in the harvest mouse *Reithrodontomys megalotis* (Peppers et al., 1997) or in different cytotypes of the plant *Prospero autumnale* (Jang et al., 2016).

Finally, the analysis of sequences located in Bs allowed to relatively date the origin of these elements. B chromosomes arisen from ancestral events, thus being quite old, have been reported in the fish *Prochilodus lineatus* (Artoni et al., 2006) or in the grasshopper species *L. migratoria* (Teruel et al., 2010) or *Rhammatocerus brasiliensis* (Melo et al., 2020). On the other hand, the origin of certain B chromosomes could be much more recent, as in *D. albomicans* (Zhou et al., 2012) or in the plant *Plantago lagopus* (Dhar et al., 2002). Most of these results came from molecular and cytological analysis of repetitive elements but recent progress in the identification of protein-coding genes in B chromosomes has made possible to address this question from a different approach, as the early origin of B chromosomes in *Astatotilapia latifasciata* proposed by Valente et al. (2014).

Eyprepocnemis plorans as a model for the study of B chromosomes

The species *E. plorans*

The grasshopper *Eyprepocnemis plorans* (Orthoptera, Acrididae), the object of study in the present thesis, was described for the first time by Charpentier in 1825. This species includes four subspecies: *E. plorans plorans* (Charpentier, 1825), *E. plorans ornatipes* (Walter, 1870), *E. plorans ibandana* (Giglio-Tos, 1907) and *E. plorans meridionalis* (Uvarov, 1921). Dirsh (1958) highlighted that *E. plorans* is a very variable species but it gathers four subspecies frequently found in overlapping locations. *E. plorans ibandana* is found in Central and West Africa, the location of *E. plorans ornatipes* (larger, with relatively longer tegmina) covers a wide area of Africa, including also Ethiopia and part of Somalia, as it is the case of *E. plorans meridionalis* (with tegmina usually not reaching the hind knees) but being more abundant in South and East Africa (see Fig. 1 in John and Lewis, 1965). In this thesis we will focus on the subspecies *E. plorans plorans*, which is distributed along the entire Mediterranean coast, the Caucasus, Turkey, Turkmenistan, Iran and the south-Western Arabia (Dirsh, 1958). Within the Iberian Peninsula, it is found throughout the Mediterranean region, from Tarragona to Huelva. For simplicity, I will refer to this subspecies only with the specific name from now on.

This grasshopper presents a single annual generation, from July to March, exhibiting the maximum population density in the month of October, especially for males (Hernández and Presa, 1984). It is a polyphagous species, with gregarious capacity and a high dispersal power. The size of the *E. plorans* genome is ~11 Gb (Ruiz-Ruano et al., 2011) and its chromosomal complement is the typical of the Acrididae family. Therefore, it comprises in males a telocentric X chromosome and 22 autosomes which are classified into three groups according to their size: long (L1-L2), medium (M3-M8) and small (S9-S11) chromosomes, with X being of intermediate size between chromosomes L2 and M3. The chromosomal sex determination in this species is X0/XX, being males X0 and females XX. In most populations, authors have found the presence of stable B chromosomes, thus appearing in the same number in all the cells of the organisms. This point has made this species an ideal model for the study of B chromosomes, helped also by their straightforward collection from the field and easy handling in the lab.

B chromosomes are widespread in the populations of *E. plorans*

The B chromosome system of *E. plorans* is one of the most deeply studied. These B chromosomes show stable mitotic behavior, so, as stated above, every cell in an individual has the same number of B chromosomes. In addition, Bs in *E. plorans* are very polymorphic, in fact authors have reported more than 50 variants of them (Henrique-Gil et al., 1984; López-León et al., 1993; Bakkali et al., 1999; López-León et al., 2008).

The presence of B chromosomes has been observed in most of the Spanish populations (Camacho et al., 1980; Henriques-Gil et al., 1984; Cabrero et al., 1997; Riera et al., 2004) being B1, B2, B5 and B24 the more abundant variants. The frequency of B chromosomes varies widely, both between populations as between individuals, there are B-lacking individuals but also specimens carrying six of them within the same population (Camacho et al., 2003). However, the presence of B chromosomes has not been observed in populations located at the head of the Segura river, in the province of Albacete (Cabrero et al., 1997).

The B1 chromosome is the one most broadly distributed in the Iberian Peninsula, being considered the ancestral variant (Henriques-Gil et al., 1990). Later studies have shown that this is only true for the western Mediterranean region, from Sicily and Tunisia to the Iberian Peninsula and Morocco (Cabrero et al., 2013). The existence of other prevalent variants in different Spanish populations indicates that there have been, at least, three processes of substitution of one variant for another: i) that of the B1 variant by the B2 one in the coastal populations of Granada and eastern Málaga, ii) that of the B1 variant by the B5 one in populations near Fuengirola (Málaga) and iii) the replacement of the B2 variant by the chromosome B24 in the Torrox population in the province of Malaga (Zurita et al., 1998). In the latter case, there is recent evidence that the B24 is extending its geographical distribution to the Torrox neighboring populations, such as Algarrobo (to the west) and Nerja (to the east) (Manrique-Poyato et al., 2013). Interestingly, females of *E. plorans* could be the responsible of this interpopulation spreading of Bs as the role of males as dispersive subjects is unlikely (Manrique-Poyato et al., 2020).

The B chromosome variants B1, B2 and B5 are considered as polymorphisms neutralized by the host genome. They do not have accumulation mechanisms nor effects on traits related to the biological efficacy of individuals when carrying a low number of chromosomes B, although fertility in females carrying a high number of B2 chromosomes is reduced (López-León et al., 1992; Camacho et al., 1997; Martín-Alganza et al., 1997).

Interestingly, through controlled crosses, Herrera et al. (1996) demonstrated the existence of genes that suppress the accumulation of B chromosomes in *E. plorans*. These authors observed that the transmission of Bs showed Mendelian inheritance in females with 1B when crossed with B-lacking males from the same population. However, when these females copulated with males from an Albacete population, where there are no Bs, they found the accumulation of B chromosomes. The authors concluded that in the genome of *E. plorans* belonging to populations with B chromosomes there are genes that suppress the accumulation of these elements whereas in populations without Bs these genes are absent so drive is restored.

The B chromosome variant that exhibits the higher degree of parasitism in *E. plorans* is the B24, which showed a transmission rate (0.696) higher than the expected by the Mendelian laws (0.5), while decreasing fertility in carrier females (Zurita et al., 1998). This selfish behavior is the main reason to choose the B24 as the B chromosome variant in which this thesis focuses. The accumulation presented by the B24 variant seems to have been critical for displacing, in the population of Torrox, the B2 variant previously neutralized by the host genome (Zurita et al., 1998). Subsequently, it was observed that the B24 variant was rapidly neutralized (in only six years) by the host genome, having lost the ability to drive (Perfectti et al., 2004) and being less virulent on fertility (Manrique-Poyato et al., 2006). This rapid suppression of B24 accumulation could be explained by the existence, on chromosomes A, of a single locus (or a few loci) with great neutralizing effect (Perfectti et al., 2004).

The findings of several variants for B chromosomes in *E. plorans* showing different parasitism and neutralization levels made this grasshopper the outstanding species supporting the near-neutral model for B chromosome dynamics in natural populations described by Camacho et al. (1997) as explained above. This model served also to support the evolution of other B chromosomes systems as that in frogs of the genus *Oreobates* (Ferro et al., 2016) or in the grasshopper *Rhammatocerus brasiliensis* (Melo et al., 2020).

The B chromosome drive in *E. plorans* takes place during female meiosis, and is based on their preferential segregation towards the anaphase pole that will form the secondary oocyte (Zurita et al., 1998; Bakkali et al., 2002). As mentioned above, almost all natural populations of this grasshopper carry B chromosomes, and the high success of these supernumerary chromosomes in *E. plorans* results from their drive during

female meiosis. Thus, the decisive step where B chromosomes in general, and those of *E. plorans* in particular, play their evolutionary destiny is the cell cycle, specifically during meiosis.

The DNA composition of their B chromosomes is poorly understood

The B chromosomes of *Eyprepocnemis plorans* are heterochromatic and enriched in two classes of repetitive sequences that are also located in the A chromosomes. These two sequences are the 180 bp satellite DNA, which is found as part of the three heterochromatin bands of the B24 variant, and the 45S ribosomal DNA, located at the distal end of the B chromosome. These two elements are arranged in the X chromosome of *E. plorans* in a similar order with respect to the centromere than that observed in the B2 variant. This fact supported the hypothesis of a B2 variant derived from the X chromosome (López-León et al., 1994). Notwithstanding, same authors put in question the role of the X chromosome as the main ancestor of Bs after outcomes from chromosome painting when using X and B chromosomes probes that hybridized with most of the autosomes. However, this study still supports the intragenomic origin of the B chromosome in *E. plorans* (Teruel et al., 2009). Later experiments, involving the ITS1 and ITS2 regions from the B, the X and the S11 chromosomes, revealed that the B-located sequences resemble those of the littlest autosome pair (Teruel et al., 2014).

The relative order of the sat180 and the rDNA is conserved in most of the B chromosomes found in the Western Mediterranean region (Bakkali et al., 1999; Cabrero et al., 1999), suggesting a common origin for all variants (Cabrero et al., 2014). In contrast, the relative proportion of rDNA and 180pb satDNA located in the B chromosomes is different in Eastern Mediterranean populations compared to the Western ones (Abdelaziz et al., 2007; López-León et al., 2008), which lack the 5S ribosomal DNA that is present in the formers (Cabrero et al., 2003). This led to consider a multiregional origin of the B chromosomes, however, this possibility was later on ruled out when observing the high similarity between very distant populations for a specific sequence of the B chromosome (Muñoz-Pajares et al., 2011).

In addition to both repetitive elements mentioned above, it is also known the presence of some transposable elements located in the euchromatin region of Bs in *E. plorans*, such as EploGypI, EploMar20 and EploRTE5 (belonging to the Gypsy, Mariner and RTE families respectively). However, their abundance in the B chromosomes is quite

low (Montiel et al., 2012). As expected from the presence of rDNA in the B chromosomes of *E. plorans*, the accumulation of the R2 retrotransposon have taken place in that precise location (Montiel et al., 2014) since this element shows preferential insertion in a well-defined target into the 28S gene (Browne et al., 1984; Jakubczak et al., 1991). Also the 5S rRNA gene have been found in the B chromosomes of *E. plorans* but only on those from individuals belonging to a Caucasus population (Cabrero et al., 2003). Remarkably, it was not until three years ago that the first protein-coding genes were found in B chromosomes of *E. plorans* (Navarro-Domínguez et al., 2017a). Over four decades, our group has studied many aspects of B chromosomes of *E. plorans*, however, there is still a lack of information regarding other DNA sequences located in Bs.

During years, the B chromosomes of *E. plorans* were thought to be transcriptionally silenced due to their heterochromatic nature and hypoacetylated stage during male meiosis (Cabrero et al., 2007). On the other hand, no phenotype and few endophenotype effects have been claim for the B chromosomes of *E. plorans*. In addition to the increase in chiasma frequency for *Eyprepocnemis plorans* (Camacho et al., 2002), there have been detected some alterations associated with the presence of B chromosomes in other stress markers, such as nucleolar size (Teruel et al., 2007) and a decrease in the levels of the Hsp70 protein (Teruel et al., 2011). In this latter case, the lower levels of Hsp70 in B-carrying individuals compared to 0B ones is likely the consequence of post-transcriptional regulation and not involving a decrease in the transcription activity of that gene (Navarro-Domínguez et al., 2016a). First discoveries concerning transcriptional effects of B chromosomes involved the presence of ribosomal DNA transcripts coming specifically from B chromosomes (Ruiz-Estévez et al., 2012). However, this rDNA transcription was described only in few males belonging to several populations (Ruiz-Estévez et al., 2013). Furthermore, the relative rRNA input of the B chromosome was negligible compared with that of the standard set (Ruiz-Estévez et al., 2014), suggesting that B chromosomes are quite silenced.

Nevertheless, the recent discovery of ten protein-coding genes located in the B chromosome of *E. plorans*, five of which were actively transcribed (Navarro-Domínguez et al., 2017a), means that B chromosomes are not so silenced as previously thought. Therefore, it is feasible that the key for B chromosome success may lie on its own gene content. This statement is supported by the active transcription of the complete CDSs of genes such as *cip2a* and *kif20a*, or some fragments of *ckap2*, *cap-g* and *mycb2*, all being

genes involved in functions related to cell division (Navarro-Domínguez et al., 2017a). Last year, same authors identified 46 differentially expressed genes (30 of them up-regulated) associated to the presence of a B chromosome in the grasshopper *E. plorans*. In fact, most of these genes are involved in functions related to host-parasite adaptation processes such as responses to stress, protein modifications, ovary function or regulation of gene expression, explaining some well-known effects of the B chromosome presence (Navarro-Domínguez et al., 2019). These gene expression changes associated to the presence of B chromosomes in *E. plorans* suggest the possibility of a transcriptional crosstalk taking place between A and B chromosomes under an intragenomic arms race scene.

One of the most enigmatic questions that arises in the study of supernumerary chromosomes is how they are able to perform the drive mechanisms allowing them to invade and establish in natural populations. As in *E. plorans* the drive of B chromosomes occurs during female meiosis, a possible answer to this issue could lie on meiosis manipulation through the expression of genes contained in the B itself. This fact places the spotlight in the search of B-genes with functions related to cell cycle control, which is one the main goals of this thesis.

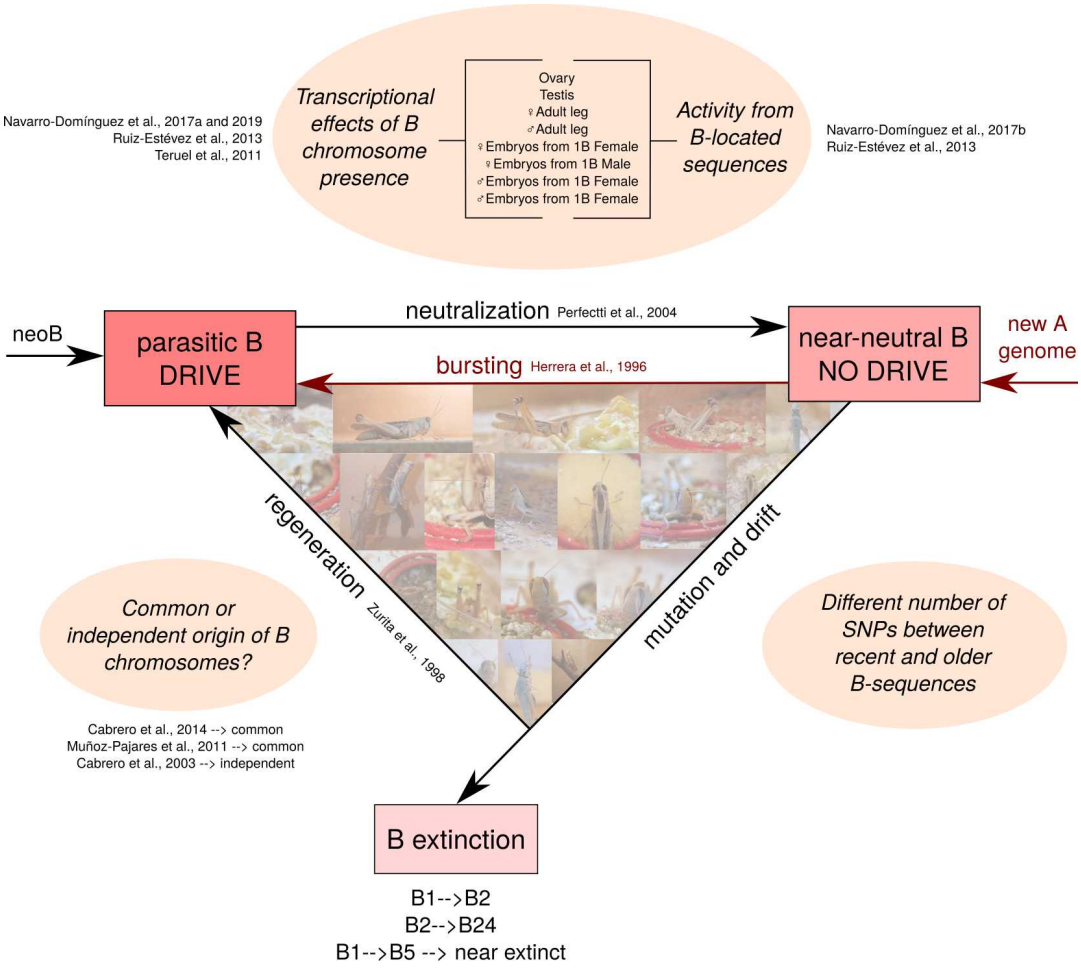


Figure I.1. Conceptual framework for the B chromosome system of *Eyprepocnemis plorans* relying on the model proposed in the same species for long term evolution of Bs. Find the three evolutionary stages of B chromosomes reported by Camacho et al. (1997) and empirically supported by reference cited in the diagram: parasitic, near-neutral and extinction or regeneration of B chromosomes variant. Note that the concept "bursting" is added here to represent the relighting of drive for the same B chromosome variant when the genomic background of the B-lacking individual changes (i.e. crosses between 1B females and 0B males from a B-lacking population described in Herrera et al., 1996). Circles in soft orange show some approached and ideas through which this PhD thesis could help to deeply understand the nature and dynamics of B chromosomes in *E. plorans*.

Objectives of this PhD Thesis

General objective and initial hypothesis

The general objective of this thesis is to unveil the molecular content of the B chromosomes of the grasshopper *Eyprepocnemis plorans* and the transcriptional changes driven by their presence in the genome. The initial hypothesis is that *"the presence of B chromosomes triggers some changes in gene expression of carrier individuals, probably coming from the own B chromosomes contributing to their maintenance and transmission but also from the host genome leading to silence B chromosome, thus counteracting the expression of the genes contained in it"*. This hypothesis is suggested by our previous results about B chromosomes in this species.

The first evidence for changes in the activity of genes located in the A chromosomes was obtained for the activity of nucleolus organizing regions (NORs) (López-León et al., 1995) and the heat shock protein Hsp70 (Teruel et al., 2011). Our recent findings of ten protein genes that reside on the B chromosome of *E. plorans*, five of which are actively expressed in B-carrying individuals (Navarro-Domínguez et al., 2017a) indicate that B chromosomes could be not as silenced as previously thought. This also suggests the existence of a transcriptional conflict between A and B chromosomes, as evidenced by our first results revealing transcription changes associated with the presence of the B chromosomes in this species (Navarro-Domínguez et al., 2019). However, it is still unknown the extent to which those transcripts come directly from B chromosomes or if they are an response from the host genome to the presence of these selfish elements. The present thesis aims to decipher the DNA content of the parasitic (B) chromosomes of *Eyprepocnemis plorans* and unveil deeper details on the interaction between the standard genome and these enigmatic genomic elements.

Specific objectives

1. Characterization and building of a complete repetitive DNA database of *E. plorans* including *de novo* assembling of tandem repeats, transposable elements and other repetitive sequences (histones, rDNA, mtDNA, snRNA and tRNAs) using Illumina reads.

2. Unveiling the repetitive DNA composition of the B chromosomes of *E. plorans* and their transcriptional activity through a quantitative approach to confirm whether B chromosomes are rich in these kind of sequence as suggested by previous studies. We will also determine the chromosomal location of putatively B-located repetitive sequences by FISH.

3. Addressing the possible origin of B chromosomes in *E. plorans* focused in the analysis of their content in repetitive DNA, FISH results of B-located sequences and analysis of MinION long reads. To accomplish this goal, we will also analyze the repetitive content of individuals of *E. plorans* coming from different and distant populations (Tanzania, Egypt and Armenia).

4. Identification of protein-coding genes located in the B chromosomes of *E. plorans* using several biological replicates harboring different number of B chromosomes and a comprehensive reference transcriptome assembled from RNA libraries belonging to different sexes and developmental stages of *E. plorans*.

5. Searching for B-specific transcripts showing SNPs signatures found exclusively in carrier individuals and analyzing then their expression activity in RNA libraries respect to the transcripts coming from the standard set of chromosomes. This will put into evidence the specific transcription activity from B chromosomes.

6. Tracking the possible elimination of B chromosomes in testis of *E. plorans* using a marker developed to detect the presence of these elements. We will also explore this phenomenon in *Eumigus monticola* in which the B chromosomes are, in contrast to *E. plorans*, mitotically unstable.

7. Analysis of gene expression changes in embryos belonging to different pods and sexes of *E. plorans* associated with the B chromosome presence. We will study two different pods, in one of them the female parental will be the carrier individual and in the other one that will be the male parental. This experimental design will allow us to explore differences in gene expression depending on the sex from which the B chromosome is inherited.

8. Analysis of gene expression changes in female and male adults of *E. plorans* associated with the presence of a parasitic chromosome. We will check those differences in legs and gonads, the latter being the place in where expression of B chromosomes is expected to play an important role regarding their maintenance and accumulation through the offspring.

References

Abdelaziz M, Teruel M, Chobanov D, Camacho JPM, Cabrero J. (2007). Physical mapping of rDNA and satDNA in A and B chromosomes of the grasshopper *Eyprepocnemis plorans* from a Greek population. *Cytogenetic and Genome Research*, 119(1–2), 143–146.

Ahmad SF, Jehangir M, Cardoso AL, Wolf IR, Margarido VP, Cabral-de-Mello DC, et al. (2020). B chromosomes of multiple species have intense evolutionary dynamics and accumulated genes related to important biological processes. *BMC Genomics*, 21, 656.

Akbari OS, Antoshechkin I, Hay BA, Ferree PM. (2013). Transcriptome profiling of *Nasonia vitripennis* testis reveals novel transcripts expressed from the selfish B chromosome, paternal sex ratio. *G3 (Bethesda)*, 3(9), 1597–1605.

Akera T, Chmátal L, Trimm E, Yang K, Aonbangkhen C, Chenoweth DM, et al. (2017). Spindle asymmetry drives non-Mendelian chromosome segregation. *Science*, 358(6363), 668–672.

Artoni RF, Vicari MR, Endler AL, Cavallaro ZI, de Jesus CM, de Almeida MC, et al. (2006). Banding pattern of A and B chromosomes of *Prochilodus lineatus* (Characiformes, Prochilodontidae), with comments on B chromosomes evolution. *Genetica*, 127(1–3), 277–84.

Bakkali M, Cabrero J, López-León MD, Perfectti F, Camacho JPM. (1999). The B chromosome polymorphism of the grasshopper *Eyprepocnemis plorans* in North Africa: I. B variants and frequency. *Heredity*, 83(4), 428–434.

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(1), 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(7), 2536–2544.

Becker SE, Thomas R, Trifonov VA, Wayne RK, Graphodatsky AS, Breen M. (2011). Anchoring the dog to its relatives reveals new evolutionary breakpoints across 11 species of the Canidae and provides new clues for the role of B chromosomes. *Chromosome Research*, 19(6), 685–708.

Bergerard J, Carton Y, Lespinasse R. (1972). Analyse eléctrophorétique en gel de polyacrylamidedes protéines de l'hémolymphe de *Locusta migratoria* L. Influence des chromosomes surnuméraires dans la sous-espèce *migratorioides* Reiche et Fairmaire. *C. R. Acad. SC. (Paris)* 275, 7833786.

Battaglia E. (1964). Cytogenetics of B chromosomes. *Caryologia,* 17, 245–299.

Beukeboom LW. (1994). Bewildering Bs an impression of the 1st B-chromosome conference. *Heredity*, 73(3), 328–335.

Bosemark NO. (1954). On accessory chromosomes in *Festuca pratensis*. II. Inheritance of the standard type of accessory chromosomes. *Hereditas*, 40, 425–437.

Brockhouse C, Bas JAB, Fereday RM, Strauss NA. (1989). Supernumerary chromosomes evolution in the *Simulium vernum* group (Diptera: Simulidae). *Genome* 32, 516–521.

Browne MJ, Read CA, Roiha H, Glover DM. (1984). Site specific insertion of a type I rDNA element into a unique sequence in the *Drosophila melanogaster* genome. *Nucleic Acids Research*, 12(23), 9111–9122.

Bugrov AG, Karamysheva TV, Perepelov EA, Elisaphenko EA, Rubtsov DN, Warchałowska-Sliwa E, et al. (2007). DNA content of the B chromosomes in grasshopper *Podisma kanoi* Storozh. (Orthoptera, Acrididae). *Chromosome Research*, 15(3), 315–325.

Burt A, Trivers R. (2006). Genes in Conflict: The Biology of Selfish Genetic Elements *Harvard University Press*.

Cabrero J, Alché JD, Camacho JPM. (1987). Effects of B chromosome of the grasshopper *Eyprepocnemis plorans* on nucleolar organiser regions activity. Activation of a latent NOR on a B chromosome fused to an autosome. *Genome*, 29, 116–121.

Cabrero J, López-León MD, Gómez R, Castro AJ, Martín-Alganza A, Camacho JPM. (1997). Geographical distribution of B chromosomes in the grasshopper *Eyprepocnemis plorans*, along a river basin, is mainly shaped by nonselective historical events. *Chromosome Research* 5, 194–198.

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

Cabrero J, Bakkali M, Bugrov A, Warchalowska-Sliwa E, López-León MD, Perfectti F, et al. (2003). Multiregional origin of B chromosomes in the grasshopper *Eyprepocnemis plorans*. *Chromosoma* 112, 207–211.

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(1–2), 135–142.

Cabrero J, López-León MD, Ruiz-Estévez M, Gómez R, Petitpierre E, Rufas JS, et al. (2014). B1 was the ancestor B chromosome variant in the western Mediterranean area in the grasshopper *Eyprepocnemis plorans*. *Cytogenetic and Genome Research*, 142(1), 54–58.

Camacho JPM, Carballo AR, Cabrero J. (1980). The B-chromosome system of the grasshopper *Eyprepocnemis plorans* subsp. *plorans* (Charpentier). *Chromosoma* 80, 163–176.

Camacho JPM, Shaw MW, López-León MD, Pardo MC, Cabrero J. (1997). Population dynamics of a selfish B chromosome neutralized by the standard genome in the grasshopper *Eyprepocnemis plorans*. *The American Naturalist*, 149(6), 1030–1050.

Camacho JPM, Sharbel TF, Beukeboom LW. (2000). B-chromosome evolution. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 355(1394), 163–178.

Camacho JPM, Bakkali M, Corral JM, Cabrero J, López-León MD, Aranda I, et al. (2002). Host recombination is dependent on the degree of parasitism. *Proceedings of the Royal Society of London B: Biological Sciences*, 269(1505), 2173–2177.

Camacho JPM, Cabrero J, López-León MD, Bakkali M, Perfectti F. (2003). The B chromosomes of the grasshopper *Eyprepocnemis plorans* and the intragenomic conflict. *Genetica*, 117(1), 77–84.

Camacho JPM. (2005). B chromosomes. The evolution of the genome (Gregory TR, ed.), 223–286.

Carchilan M, Delgado M, Ribeiro T, Costa-Nunes P, Caperta A, Morais-Cecílio L, et al. (2007). Transcriptionally active heterochromatin in rye B chromosomes. *Plant Cell* 19, 1738–1749.

Carmello BO, Coan RL, Cardoso AL, Ramos E, Fantinatti BE, Marques DF, et al. (2017). The hnRNP Q–like gene is retroinserted into the B chromosomes of the cichlid fish

Astatotilapia latifasciata. *Chromosome Research*, 25(3–4), 277–290.

Charpentier. (1825). De Orthopteris Europaeis. Horae entomologicae, adjectis tabulis novem coloratis, apud A. *Gosohorsky, Wratislaviae* 61–181.

Cheng YM, Lin BY. (2003). Cloning and characterization of maize B chromosome sequences derived from microdissection. *Genetics*, 164(1), 299–310.

D'Ambrosio U, Alonso-Lifante MP, Barros K, Kovařík A, Mas de Xaxars G, Garcia S. (2017). B–chrom: a database on B–chromosomes of plants, animals and fungi. *New Phytologist*, 216(3), 635–642.

Dalla Benetta E, Akbari OS, Ferree PM. (2019). Sequence Expression of Supernumerary B Chromosomes: Function or Fluff? *Genes (Basel)*, 10(2), 123.

Dalla Benetta E, Antoshechkin I, Yang T, Nguyen HQM, Ferree PM, Akbari OS. (2020). Genome elimination mediated by gene expression from a selfish chromosome. *Science Advances*, 6(14), eaaz9808.

Dhar MK, Friebe B, Koul AK, Gill BS. (2002). Origin of an apparent B chromosome by mutation, chromosome fragmentation and specific DNA sequence amplification. *Chromosoma,* 111(5), 332–340.

Dirsh VM. (1958). Revision of the genus *Eyprepocnemis* Fieber, 1853 (Orthoptera: Acridoidea). *Proceedings of the Royal Entomological Society of London. Series B*, 27, 33–45.

Ebrahimzadegan R, Houben A, Mirzaghaderi G. (2019). Repetitive DNA landscape in essential A and supernumerary B chromosomes of *Festuca pratensis* Huds. *Scientific Reports*, 9(1), 19989.

Eickbush DG, Eickbush TH, Werren JH. (1992). Molecular characterization of repetitive DNA sequences from a B chromosome. *Chromosoma*, 101(9), 575–583.

Ferro JM, Taffarel A, Cardozo D, Grosso J, Puig MP, Suárez P, et al. (2016). Cytogenetic characterization and B chromosome diversity in direct–developing frogs of the genus *Oreobates* (Brachycephaloidea, Craugastoridae). *Comparative Cytogenetics*, 10(1), 141–156.

Franks TK, Houben A, Leach CR, Timmis JN. (1996). The molecular organization of a B chromosome tandem repeat sequence from *Brachycome dichromosomatica*. *Chromosoma*, 105, 223–230.

Giglio-Tos. (1907). Spedizione al Ruwenzori di S.A.R. Luigi Amedeo di Savoia Duca degli Abruzzi. XVI: Ortotteri nuovi (diagnose preventive). *Bollettino dei Musei di Zoologia ed Anatomia Comparata della R. Università di Torino (Boll. Musei Zool. Anat. Comp. R. Univ. Torino)*, 22(547), 1–3.

Graphodatsky AS, Kukekova AV, Yudkin DV, Trifonov VA, Vorobieva NV, Beklemisheva VR, et al. (2005). The proto-oncogene C-KIT maps to canid B-chromosomes. *Chromosome Research*, 13(2), 113–122.

Green DM. (1990). Muller's ratchet and the evolution of supernumerary chromosomes. *Genome*, 33(6), 818–824.

Gruber SL, Diniz D, Sobrinho-Scudeler PE, Fausto Foresti, Haddad CF, 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(3), 185–197.

Gutknecht J, Sperlich D, Bachmann L. (1995). A species specific satellite DNA family of *Drosophila subsilvestris* appearing predominantly in B chromosomes. *Chromosoma*,

103, 539–544.

Hasegawa N. (1934). A cytological study on 8–chromosome rye. *Cytologia,* 6, 68–77.

Henrique-Gil N, Santos JL, Arana P. (1984). Evolution of a complex polymorphism in the grasshopper *Eyprepocnemis plorans*. *Chromosoma* 89, 290–293.

Henriques-Gil N, Arana P. (1990). Origin and substitution of B chromosomes in the grasshopper *Eyprepocnemis plorans*. *Evolution* 44, 747–753.

Hernández F, Presa JJ. (1984). Sobre la biología de *Eyprepocnemis plorans* (charpentier, 1825) (Orthoptera: Acrididae), en la huerta de Murcia (S.E. España). *Boletín del Servicio de Plagas*, 10, 245–249.

Herrera JA, López-León MD, Cabrero J, Shaw MW, Camacho J. (1996). Evidence for B chromosome drive suppression in the grasshopper *Eyprepocnemis plorans*. *Heredity*, 76(6), 633.

Hewitt G. (1976). Meiotic drive for B-chromosomes in the primary oocytes of *Myrmeleotettix maculatus* (Orthopera: Acrididae). *Chromosoma*, 56(4), 381.

Hewitt GM. (1974). The integration of supernumerary chromosomes into the orthopteran genome. *Cold Spring Harbor Symposium on Quantitative Biology*, 38, 183–194.

Hong ZJ, Xiao JX, Peng SF, Lin YP, Cheng YM. (2020). Novel B–chromosome–specific transcriptionally active sequences are present throughout the maize B chromosome. *Molecular Genetics and Genomics*, 295(2), 313–325.

Houben A, Thompson N, Ahne R, Leach CR, Verlin D, Timmis JN. (1999). A monophyletic origin of the B chromosomes of *Brachycome dichromosomatica* (Asteraceae). *Plant Systematics and Evolution*, 219(1–2), 127–135.

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), 88.

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

Jakubczak JL, Burke WD, Eickbush TH. (1991). Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. *Proceedings of the National Academy of Sciences USA*, 88(8), 3295–3299.

Jang TS, Parker JS, Weiss–Schneeweiss H. (2016). Structural polymorphisms and distinct genomic composition suggest recurrent origin and ongoing evolution of B chromosomes in the *Prospero autumnale* complex (Hyacinthaceae). *New Phytologist*, 210(2), 669–679.

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

Jamilena M, Garrido-Ramos M, Ruiz-Rejón M, Ruiz-Rejón C. (1995). Characterisation of repeated sequences from microdissected B chromosome of *Crepis capillaris*. *Chromosoma* 104, 113–120.

Jimenez MM, Romera F, Gallego A, Puertas MJ. (1995). Genetic control of the rate of transmission of rye B chromosomes. II. 0b times 2b crosses. *Heredity*, 74, 518–523.

John B, Lewis KR. (1965). Genetic speciation in the grasshopper *Eyprepocnemis plorans*. *Chromosoma*, 16, 308–344.

Jones N, Houben A. (2003). B chromosomes in plants: escapees from the A chromosome genome? *Trends in Plant Science*, 8(9), 417–423.

Jones RN, González-Sánchez M, González-García M, Vega JM, Puertas MJ. (2008). Chromosomes with a life of their own. *Cytogenetic and Genome Research*, 120, 265–280.

Jones RN, Rees H. (1982). B chromosomes. Academic press.

Jones RN. (1985). Are B chromosomes selfish?. The Evolution of Genome Size, 30, 397–425.

Kayano H. (1957). Cytogenetic studies in *Lilium callosum*. III. Proceedings of the Japanese Academy; 553–558. Preferential segregation of a supernumerary chromosome in EMCs.

Keyl HG, Hägele K. (1971). B chromosomen bei chironomus. *Chromosoma*, 35, 403–417.

Kinsella CM, Ruiz-Ruano FJ, Dion-Côté AM, Charles AJ, Gossmann TI, Cabrero J, et al. (2019). Programmed DNA elimination of germline development genes in songbirds. *Nature Communications*, 10(1), 5468.

Kirk D, Jones RN. (1970). Nuclear genetic activity in B-chromosome rye, in terms of the quantitative interrelationships between nuclear proteins, nuclear RNA and histone. *Chromosoma*, 31, 241–254.

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(2), 550–8.

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.

Lanzas P, Perfectti F, Garrido-Ramos MA, Ruíz-Rejón C, González–Sánchez M, Puertas M, et al. (2018). Long-term monitoring of B-chromosome invasion and neutralization in a population of *Prospero autumnale* (Asparagaceae). *Evolution*, 72(6), 1216–1224.

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(1), 269–278.

Lipsitch M, Nowak MA, Ebert D, May RM. (1995). The population dynamics of vertically and horizontally transmitted parasites. *Proceedings of the Royal Society of London B: Biological Sciences*, 260, 321–327.

López-León MD, Cabrero J, Camacho JPM, Cano MI, Santos JL. (1992). A widespread B chromosome polymorphism maintained without apparent drive. *Evolution*, 46, 529–539.

López-León MD, Pardo MC, Cabrero J, Viseras E, Camacho JPM, Santos JL. (1993). Generating high variability of B chromosomes in the grasshopper *Eyprepocnemis plorans*. *Heredity*, 71, 352–362.

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

López-León MD, Cabrero J, Dzyubenko VV, Bugrov AG, Karamysheva TV, Rubtsov NB, Camacho JPM. (2008). Differences in ribosomal DNA distribution on A and B chromosomes between eastern and western populations of the grasshopper *Eyprepocnemis plorans plorans*. *Cytogenetic and Genome Research*, 121, 260–265.

Ma W, Gabriel TS, Martis MM, Gursinsky T, Schubert V, Vrána J, et al. (2017). Rye B chromosomes encode a functional argonaute-like protein with in vitro slicer activities similar to its A chromosome paralog. *New Phytologist*, 213(2), 916–928.

Makunin AI, Kichigin IG, Larkin DM, O'Brien PC, Ferguson-Smith MA, Yang F, et al. (2016). Contrasting origin of B chromosomes in two cervids (Siberian roe deer and grey brocket deer) unravelled by chromosome-specific DNA sequencing. *BMC Genomics*, 17(1), 618.

Manrique-Poyato MI, Muñoz-Pajares AJ, Loreto V, López-León MD, Cabrero J, Camacho JPM. (2006). Causes of B chromosome variant substitution in the grasshopper *Eyprepocnemis plorans*. *Chromosome Research*, 14, 693–700.

Manrique-Poyato MI, López-León MD, Cabrero J, Perfectti F, Camacho JPM. (2013). Spread of a new parasitic B chromosome variant is facilitated by high gene flow. *PLoS ONE*, 8(12), e83712.

Manrique-Poyato MI, Cabrero J, López-León MD, Perfectti F, Gómez R, Camacho JPM. (2020). Interpopulation spread of a parasitic B chromosome is unlikely through males in the grasshopper *Eyprepocnemis plorans*. *Heredity (Edinb)*, 124(1), 197–206.

Marques A, Klemme S, Houben A. (2018). Evolution of Plant B Chromosome Enriched Sequences [published correction appears in Genes (Basel). (2019 Jan 26;10(2),]. *Genes (Basel)*, 9(10), 515.

Marques A, Klemme S, Guerra M, Houben A. (2012). Cytomolecular characterization of *de novo* formed rye B chromosome variants. *Molecular Cytogenetics*, 5(1), 34.

Martín-Alganza A, Cabrero J, López-León MD, Perfectti F, Camacho JPM. (1997). Supernumerary heterochromatin does not affect several morphological and physiological traits in the grasshopper *Eyprepocnemis plorans*. *Hereditas*, 126, 187–189.

Martis MM, Klemme S, Banaei-Moghaddam AM, Blattner FR, Macas J, Schmutzer T, 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(33), 13343–13346.

McAllister BF. (1995). Isolation and characterization of a retro–element from B chromosome (PSR) in the parasitic wasp *Nasonia vitripennis*. *Insect Molecular Biology*, 4, 253–262.

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

Melo AS, Cruz GAS, Félix AP, Rocha MF, Loreto V, Moura RC. (2020). Wide dispersion of B chromosomes in *Rhammatocerus brasiliensis* (Orthoptera, Acrididae). *Genetics and Molecular Biology*, 43(3), e20190077.

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

Montgomery EA, Huang SM, Langley CH, Judd BH. (1991). Chromosome rearrangement by ectopic recombination in *Drosophila melanogaster:* genome structure and evolution. *Genetics*, 129, 1085–1098.

Montiel EE, Cabrero J, Camacho JP, López-León MD. (2012). Gypsy, RTE and Mariner transposable elements populate *Eyprepocnemis plorans* genome. *Genetica*, 140(7–9), 365–374.

Montiel EE, Cabrero J, Ruiz-Estévez M, Burke WD, Eickbush TH, Camacho JP, et al. (2014). Preferential occupancy of R2 retroelements on the B chromosomes of the grasshopper *Eyprepocnemis plorans*. *PLoS One*, 9(3), e91820.

Muller HJ. (1964). The relation of recombination to mutational advance. *Mutation Research*, 106, 2–9.

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(3), 853–863.

Navarro-Domínguez B, Cabrero J, Camacho JP, López-León MD. (2016a). B-chromosome effects on Hsp70 gene expression does not occur at transcriptional level in the grasshopper *Eyprepocnemis plorans*. *Molecular Genetics and Genomics*, 291(5), 1909–1917.

Navarro-Domínguez BM. (2016b). Análisis de los cambios de expresión génica asociados a la presencia de cromosomas B en el saltamontes *Eyprepocnemis plorans*. Universidad de Granada. Retrieved from http://hdl.handle.net/10481/44096.

Navarro-Domínguez B, Ruiz-Ruano FJ, Cabrero J, Corral JM, López-León MD, Sharbel TF, et al. (2017a). Protein-coding genes in B chromosomes of the grasshopper *Eyprepocnemis plorans*. *Scientific Reports*, 7, 45200.

Navarro-Domínguez, B, Ruiz-Ruano, FJ, Camacho JPM, Cabrero J, López-León MD. (2017b). Transcription of a B chromosome CAP-G pseudogene does not influence normal condensin complex genes in a grasshopper. *Scientific Reports*, 7(1), 17650.

Navarro-Domínguez B, Martín-Peciña M, Ruiz-Ruano FJ, Cabrero J, Corral JM, López-León MD, et al. (2019). Gene expression changes elicited by a parasitic B chromosome in the grasshopper *Eyprepocnemis plorans* are consistent with its phenotypic effects. *Chromosoma*, 128(1), 53–67.

Nur U. (1963). A mitotically unstable supernumerary chromosome with an accumulation mechanism in a grasshopper. *Chromosoma*, 14(4), 407–422.

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

Nur U, Brett BL. (1985). Genotypes suppressing meiotic drive of a B chromosome in the mealybug, *Pseudococcus obscurus*. *Genetics*, 110(1), 73–92.

Nur U, Brett BLH. (1988). Genotypes affecting the condensation and transmission of heterochromtic B chromosomes in the mealybug *Pseudococcus affinis*. *Chromosoma*, 96(3), 205–212.

Oliver JL, Posse F, Martínez-Zapater JM, Enriquez AM, Ruiz-Rejón M. (1982). B chromosomes and E1 isoenzyme activity in mosaic bulbs of *Scilla autumnalis*. *Chromosoma*, 85, 399–403.

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

Pansonato-Alves JC, Serrano É, Utsunomia R, Camacho JP, da Costa Silva GJ, Vicari MR, et al. (2014). Single origin of sex chromosomes and multiple origins of B chromosomes in fish genus *Characidium*. *PLoS One*, 9(9), e107169.

Peppers JA, Wiggins LE, Baker RJ. (1997). Nature of B chromosomes in the harvest mouse *Reithrodontomys megalotis* by fluorescence in situ hybridization (FISH). *Chromosome Research*, 5(7), 475–479.

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

Perfectti F, Corral JM, Mesa JA, Cabrero J, Bakkali M, López-León MD, et al. (2004). Rapid suppression of drive for a parasitic B chromosome. *Cytogenetic and Genome Research*, 106(2–4), 338–43.

Plowman AB, Bougourd SM. (1994). Selectively advantageous effects of B chromosomes on germination behaviour in *Allium schoenoprasum* L. *Heredity*, 72, 587–593.

Rajičić M, Romanenko SA, Karamysheva TV, Blagojević J, Adnađević T, Budinski I, et al. (2017). The origin of B chromosomes in yellow–necked mice (*Apodemus flavicollis*) – Break rules but keep playing the game. *PLoS One*, 12(3), e0172704.

Riera L, Petitpierre E, Juan C, Cabrero J, Camacho JPM. (2004). Evolutionary dynamics of a B chromosome invasion in island populations of the grasshopper *Eyprepocnemis plorans*. *Journal of Evolutionary Biology*, 17, 716–719.

Roman H. (1947). Mitotic nondisjunction in the case of interchanges involving the B–type chromosome in maize. *Genetics*, 32(4), 391.

Ruban A, Schmutzer T, Wu DD, Fuchs J, Boudichevskaia A, Rubtsova M, et al. (2020). Supernumerary B chromosomes of *Aegilops speltoides* undergo precise elimination in roots early in embryo development. *Nature Communications*, 11(1), 2764.

Ruiz-Estévez M, Cabrero J, Camacho JPM. (2012). B–chromosome ribosomal DNA is functional in the grasshopper *Eyprepocnemis plorans*. *PLoS ONE*, 7(5), e36600.

Ruiz-Estévez M, López-León MD, Cabrero J, Camacho JP. (2013). Ribosomal DNA is active in different B chromosome variants of the grasshopper *Eyprepocnemis plorans*. *Genetica*, 141(7–9), 337–45.

Ruiz-Estévez M, Badisco L, Broeck J, Perfectti F, López-León M, Cabrero J, 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(6), 1209–1216.

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.

Ruiz-Ruano FJ, Ruiz-Estévez M, Rodríguez-Pérez J, López-Pino JL, Cabrero J, Camacho JP. (2011). DNA amount of X and B chromosomes in the grasshoppers *Eyprepocnemis plorans* and *Locusta migratoria*. *Cytogenetic and Genome Research*, 134(2), 120–126.

Ruiz-Ruano FJ, Cabrero J, López-León MD, Sánchez A, Camacho JPM. (2018). Quantitative sequence characterization for repetitive DNA content in the supernumerary chromosome of the migratory locust. *Chromosoma*, 127(1), 45–57.

Ruiz-Ruano FJ, Navarro-Domínguez B, López-León MD, Cabrero J, Camacho JPM. (2019). Evolutionary success of a parasitic B chromosome rests on gene content. *BioRxiv*, 683417.

Rutishauser A, Rothlisberger E. (1966). Boosting mechanism of B-chromosomes in *Crepis capillaris*. In *Chromosomes today*, 1, 28–30.

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

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

Schmid M, Ziegler CG, Steinlein C, Nanda I, Schartl M. (2006). Cytogenentics of the bleak (*Alburnus alburnus*), with special emphasis on the B chromosomes. *Chromosome Research*, 14, 231–242.

Serrano ÉA, Utsunomia R, Scudeller PS, Oliveira C, Foresti F. (2017). Origin of B chromosomes in *Characidium alipioi* (Characiformes, Crenuchidae) and its relationship with

supernumerary chromosomes in other *Characidium* species. *Comparative Cytogenetics*, 11(1), 81–95.

Shams I, Raskina O. (2020). Supernumerary B Chromosomes and Plant Genome Changes: A Snapshot of Wild Populations of *Aegilops speltoides* Tausch (*Poaceae*, Triticeae). *International Journal of Molecular Sciences*, 21(11), 3768.

Shaw MW, Hewitt GM. (1990). B chromosomes, selfish DNA and theoretical models: where next?. Oxford surveys in evolutionary biology, 7 (ed. D Futuyma and J Antonovics), 197–223. *Oxford University Press*.

Silva DMZA, Daniel SN, Camacho JP, Utsunomia R, Ruiz-Ruano FJ, Penitente M, et al. (2016). Origin of B chromosomes in the genus *Astyanax* (Characiformes, Characidae) and the limits of chromosome painting. *Molecular Genetics and Genomics*, 291(3), 1407–1418.

Singh PB, Belyakin SN. (2018). L Chromosome Behaviour and Chromosomal Imprinting in *Sciara Coprophila*. *Genes (Basel)*, 9(9), 440.

Smith JJ. (2018). Programmed DNA Elimination: Keeping Germline Genes in Their Place. *Current Biology*, 28(10), R601–R603.

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

Teruel M, Cabrero J, Perfectti F, Camacho JPM. (2007). Nucleolus size variation during meiosis and NOR activity of a B chromosome in the grasshopper *Eyprepocnemis plorans*. *Chromosome Research*, 15(6), 755–765.

Teruel M, Cabrero J, Montiel EE, Acosta MJ, Sánchez A, Camacho JP. (2009). Microdissection and chromosome painting of X and B chromosomes in *Locusta migratoria*. *Chromosome Research*, 17(1), 11–18.

Teruel M, Cabrero J, Perfectti F, Camacho JPM. (2010). B chromosome ancestry revealed by histone genes in the migratory locust. *Chromosoma*, 119(2), 217–225.

Teruel M, Sørensen JG, Loeschcke V, Cabrero J, Perfectti F, Camacho JPM. (2011). Level of heat shock proteins decreases in individuals carrying B–chromosomes in the grasshopper *Eyprepocnemis plorans*. *Cytogenetic and Genome Research*, 132(1–2), 94–99.

Teruel M, Ruiz-Ruano FJ, Marchal JA, Sánchez A, Cabrero J, Camacho JP, et al. (2014). Disparate molecular evolution of two types of repetitive DNAs in the genome of the grasshopper *Eyprepocnemis plorans*. *Heredity (Edinb)*, 112(5), 531–42.

Torgasheva AA, Malinovskaya LP, Zadesenets KS, Karamysheva TV, Kizilova EA, Akberdina EA, et al. (2019). Germline-restricted chromosome (GRC) is widespread among songbirds. *Proceedings of the National Academy of Sciences USA*, 116(24), 11845–11850.

Trifonov VA, Dementyeva PV, Larkin DM, O'Brien PC, Perelman PL, Yang F, et al. (2013). Transcription of a protein–coding gene on B chromosomes of the Siberian roe deer (*Capreolus pygargus*). *BMC Biology*, 11, 90.

Uvarov. (1921). Notes on the Orthoptera in the British Museum. I. the group Euprepocnemini. *Transactions of the Entomological Society of London*, 7, 106–144.

Van Valen L. (1973). A new evolutionary law. *Evolutionary Theory*, 1, 1–30.

Van Vugt JJ, de Nooijer S, Stouthamer R, de Jong H. (2005). NOR activity and repeat sequences of the paternal sex ratio chromosome of the parasitoid wasp *Trichogramma kaykai*. *Chromosoma*, 114(6), 410–419.

Valente GT, Conte MA, Fantinatti BE, Cabral-de-Mello DC, Carvalho RF, Vicari MR, et al. (2014). Origin and evolution of B chromosomes in the cichlid fish *Astatotilapia latifasciata*

based on integrated genomic analyses. *Molecular Biology and Evolution*, 31(8), 2061–2072.

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(1), 23–29.

Volobujev VT, Timina NY. (1980). Unusually high number of B–chromosomes and mosaicism by them in *Apodemus peninsulae* (Rodentia, Muridae). *Tsitologia Genetika*, 14, 43–45.

Voltolin TA, Senhorini JA, Oliveira C, Foresti F, Bortolozzi J, Porto-Foresti F. (2010). B–chromosome frequency stability in *Prochilodus lineatus* (Characiformes, Prochilodontidae). *Genetica*, 138(3), 281–284.

Walker F. (1870). Catalogue of the Specimens of Dermaptera Saltatoria in the Collection of the British Museum, London 3, 425–604.

Werren JH. (1991). The paternal-sex-ratio chromosome of *Nasonia*. *The American Naturalist*, 137(3), 392–402.

White MJD. (1973). Animal cytology and evolution, 3rd edn. London: Cambridge University Press.

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

Wu D, Ruban A, Fuchs J, Macas J, Novák P, Vaio M, et al. (2019). Nondisjunction and unequal spindle organization accompany the drive of *Aegilops speltoides* B chromosomes. *New Phytologist*, 223(3), 1340–1352.

Yoshida K, Terai Y, Mizoiri S, Aibara M, Nishihara H, Watanabe M, et al. (2011). B chromosomes have a functional effect on female sex determination in lake victoria cichlid fishes. *PLoS Genetic*s, 7(8), e1002203.

Zhou Q, Zhu H, Huang Q, Zhao L, Zhang G, Roy SW, et al. (2012). Deciphering neo–sex and B chromosome evolution by the draft genome of *Drosophila albomicans*. *BMC Genomics*, 13(1), 109.

Zurita S, Cabrero J, López-León MD, Camacho JPM. (1998). Polymorphism regeneration for a neutralized selfish B chromosome. *Evolution*, 52, 274–277.

Materials and methods

Biological materials and sampling

The biological materials used in this thesis belong, in most cases, to the grasshopper *Eyprepocnemis plorans*. We collected *E. plorans* specimens in a natural population at Torrox (Málaga, Spain) (36.737558N, -3.953546W) (Table M.1). They are included in the subspecies *E. plorans plorans* and many of them carried the B24 variant, which is the most parasitic B chromosome in this species.

Some individuals were prepared for cytogenetic and molecular analyses while others were maintained alive in the laboratory, and some of them were used to perform controlled crosses to obtain ten-day-old embryos for chromosome and transcriptome analyses (Table M.1).

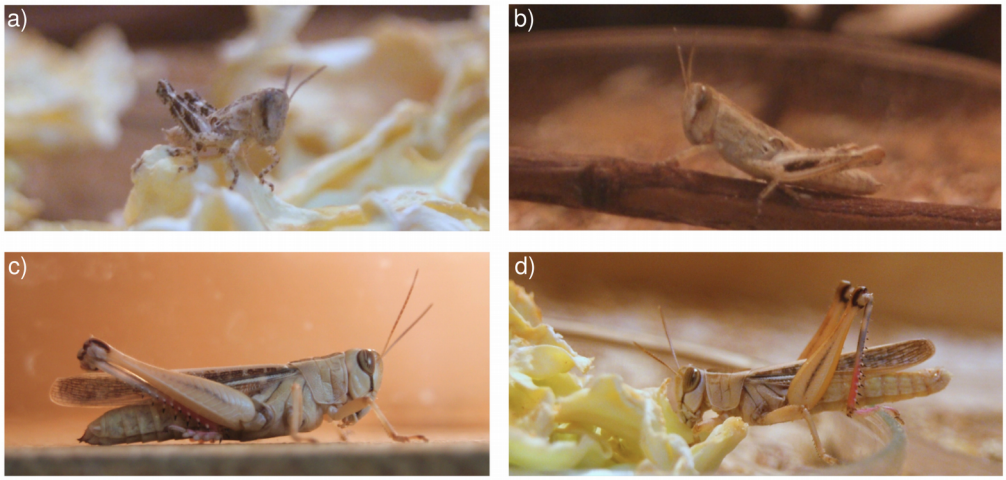


Figure M.1. Photographs of *E. plorans* individuals captured in Torrox and grown in the lab. a) First stage nymph, b) second stage nymph, c) female adult and d) male adult of *E. plorans*.

In addition, we analyzed *E. plorans* specimens from Alhama de Murcia (Murcia, Spain), Otívar and Salobreña (Granada, Spain), Tanzania, Egypt and Armenia, as well as other grasshopper species such as *Locusta migratoria* and *Eumigus monticola*. Finally,

we used Illumina libraries deposited in SRA corresponding to whole genome DNA from several hominid species: *Homo sapiens*, *Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla*, *Pongo pygmaeus* and *Macaca mulatta* (see Appendix).

Table M.1. Summary of the biological material used in this PhD thesis. Dev stage: developmental stage.

Species	Population	Year	Dev stage	Sex	Analysis	Chapter
Eyprepocnemis plorans	Torrox	2012	Embryo	Both	FISH	1, 2
			Adult	Male	gDNA Illumina HiSeq	1, 2
					gDNA MinION seq	1, 2
					PCR	1, 2
		2014	Embryo	Both	FISH	1, 2
					RNA Illumina HiSeq	2, 3, 5
			Adult	Female	gDNA Illumina HiSeq	3
				Male	gDNA Illumina HiSeq	3
					qPCR	3
					FISH	4
		2016	Adult	Female	gDNA Illumina HiSeq	3
					RNA Illumina HiSeq	2, 3, 5
				Male	gDNA Illumina HiSeq	3
					RNA Illumina HiSeq	2, 3, 5
					qPCR	3
	Alhama	2016	Adult	Male	FISH	4
	Otívar	2016	Adult	Male	FISH	4
	Salobreña	2016	Adult	Male	FISH	4
	Armenia	2016	Adult	Male	gDNA Illumina HiSeq	3
	Egypt	2016	Adult	Male	gDNA Illumina HiSeq	3
	Tanzania	2016	Adult	Male	gDNA Illumina HiSeq	3
Locusta migratoria	Padul	2012	Adult	Male	gDNA Illumina HiSeq	1
Eumigus monticola	Hoya de la Mora	2016	Adult	Male	FISH	4
Homo sapiens	Iberian populations in Spain	2013	Adult	Both	SRA libraries	Appendix
Pan troglodytes	Gabon	2013	Adult	Both	SRA libraries	Appendix
Pan paniscus	Democratic Republic of Congo	2013	Adult	Both	SRA libraries	Appendix
Gorilla gorilla	Western lowland	2013	Adult	Both	SRA libraries	Appendix
Pongo pygmaeus	Borneo	2013	Adult	Both	SRA libraries	Appendix
Macaca mulatta	India	2014	Adult	Male	SRA libraries	Appendix

Cytogenetic methods

Materials preparation

Some adult individuals collected in the field were immediately processed to get the tissue of interest whereas the others were maintained in the laboratory at 28ºC under light/darkness 12:12 hours photoperiod, some of which were used for controlled crosses to obtain embryos.

Adult males were anaesthetized prior to dissection to extract all testis follicles, some of which were fixed in 3:1 ethanol:acetic acid and stored at 4ºC for subsequent analysis, and the remaining follicles, and also body remains, were frozen in liquid nitrogen and stored in a freezer at -80ºC. Females were anaesthetized prior to dissection and ovarioles extraction, some of which were immersed in 2% colchicine (in saline solution) for 2 hours and then fixed in 3:1 ethanol:acetic acid and stored at 4ºC. The remaining ovarioles and body remains were frozen in liquid nitrogen and stored at -80ºC. We also processed alive adults of both sexes grown in our lab from controlled crossed to ascertain the number of B chromosomes of each individual while getting material for gDNA and RNA extraction for molecular and bioinformatics studies.

Egg pods from females that came gravid from the field or from our controlled crosses, were incubated at 28ºC for ten days to obtain embryos or until first instar nymph raising. Ten-day-old embryos were dissected from the eggs and immersed in 1ml of 0.05% colchicine, in saline solution, for 2 hours. Subsequently, they were subjected to an osmotic shock, adding 1 ml of distilled water, and then fixed for cytogenetic analysis as described in Camacho et al. (1991). Alternatively, sibling embryos coming from the same controlled cross were individually disaggregated in saline solution to obtain a cell suspension which was divided into three parts for cytogenetic analysis and gDNA and RNA extraction (see details below).

Determining the number of B chromosomes in alive specimens

Prior to performing controlled crosses, we needed to know the number of B chromosomes carried by the individuals collected in the field to establish appropriate mating couples. In males, the number of B chromosomes was determined following the procedure described in López-León (1992), consisting in extracting some testis follicles through a small cut performed between the second and third abdominal segments, and making squash preparations (using two of these testis follicles) stained with 2%

lactopropionic orcein. In females, however, we performed C-banding of hemolymh nuclei obtained from the abdomen with a needle, as reported in Cabrero et al. (2006).

Combined protocol to determine chromosome number and extract gDNA and RNA for genomic studies in a same embryo or adult individual

For Chapter 2, 3 and 5 of this PhD thesis, we needed to characterize each single embryo or adult of *E. plorans* in terms of sex and B chromosome number, and also obtaining enough material for gDNA and RNA extraction for sequencing. For this purpose, we adapted previous protocols of our research group to a new pipeline described below.

Embryos

Sibling embryos from controlled crosses (between adults collected in 2014) were dissected from the eggs after ten days of incubation at 28ºC, a developmental stage which shows abundant mitotic divisions allowing chromosome analysis with only a few sampled cells. The main steps of this protocol are:

1) Egg dissection in insect saline solution.

2) Each embryo is set in a 1.5 mL tube with 25 µL of insect saline solution and is broken up using a micropipette to get a homogenized cell suspension, which is divided into three parts: one for RNA extraction, one for DNA extraction and the latter for cytogenetic analysis to determine the sex and B chromosome number of embryos.

3) 15 µL of the homogenate is mixed up with 50 µL of Qiazol from the RNA extraction Kit "RNeasy Lipid Tissue Mini Kit (Qiagen)" and the tube is vigorously shaken in a Vortex for 15 seconds for proper homogenization. The tubes with individual embryo cell suspensions are kept at -20ºC, and then are stored at -80°C until RNA extraction with the afore mentioned kit as explained the Molecular methods section.

4) Keep 1 µL of the homogenate in a 0.2 mL tube for DNA extraction.

5) The remaining 9 µL are used for cytogenetic analysis. For this purpose, add 9 µL of 0.1% colchicine in insect saline solution and incubate for 45 min at room temperature.

6) During this time, we can perform gDNA extraction by cellular exposure at 95 ºC during 5 minutes (heat shock).

7) After 45 minutes of incubation in colchicine (see step 5), we perform an osmotic shock by adding 10 µL of distilled water and wait for 6 minutes at room temperature.

8) Centrifuge at 8.000 g for 1 minute to collect embryo remains at the bottom of the

tube. Discard supernatant, add 10 µL of 70% acetic acid and resuspend it. The material is then ready to make slides, as described in Camacho et al. (2015), for visualizing chromosomes in embryos.

This protocol allowed us to characterize each embryo for sex and number of B chromosomes. As embryos came from 1B x 0B crosses, they were classified as 0B males, 1B males, 0B females or 1B females.

Adults

With some adult individuals collected in 2016, we performed controlled crosses in the culture room of the research group. Egg pods were incubated at 28 ºC until hatching, at which moment the 1st instar nymphs were transferred to wooden boxes (one per cross) to raise them to adults.

After 7 days as adults, presence of B chromosomes in each individual was analyzed by hemolymph C-banding (described in Cabrero et al., 2006) and, at the same time, we extracted and stored their gonads, hind legs and body remains in liquid nitrogen at -80ºC. Legs were used for DNA and RNA extraction whereas gonads were used only for RNA extraction, and were separately stored to avoid cross contamination.

After chromosome analysis, we selected 12 sibling individuals coming from the same cross including three 0B males, three 0B females, three 1B males and three 1B females, to perform Illumina sequencing of hind leg and gonad RNA and also of leg gDNA.

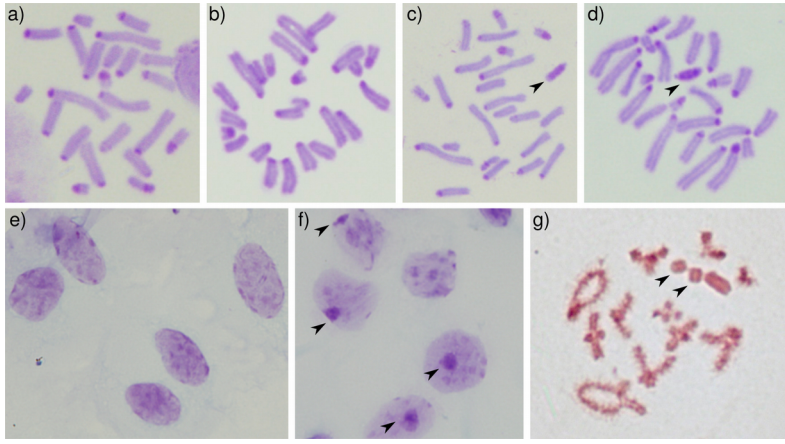


Figure M.2. C-banding of *E. plorans* embryos a) 0B male, b) 0B female, c) 1B male and d) 1B female and hemolymph nuclei in e) 0B and f) 1B adult females. Orcein staining of meiotic chromosomes from testis follicles g) 2B male. B chromosomes are marked with arrowheads.

Fluorescence *in situ* hybridization (FISH)

PCR products were labelled through "nick translation" using 2.5 units of DNA polymerase I/DNAasa I (Invitrogen) and about 250 ng of DNA probe was used in each FISH experiment following the technique described in Cabrero et al. (2003). The fluorochromes used were tetrametilrodamina-5-dUTP and/or fluoresceina-11-dUTP (Roche), which yield red and green fluorescence, respectively. To track probe location on chromosomes we counterstained with DAPI, and the slides were mounted in Vectashield (Vector, USA). For chromosome visualization, we used a BX41 epifluorescence Olympus microscope equipped with a DP70 cooled digital camera for photography.

To map repetitive DNA on chromosomes, we preferentially used embryo preparations as they show abundant mitotic metaphases (Chapters 1 and 2). Embryo preparations were performed by Meredith's technique, as described in Camacho et al. (1991). The slides were dehydrated in a series of 70%, 90%, and absolute ethanol, and then incubated in an oven at 60°C overnight.

FISH on testis preparations (in Chapter 4) was performed on squash preparations following the protocol described in Camacho et al. (2015). The preparations were previously permeabilized to facilitate probe entry in the cell and get better hybridization performance. The permeabilization consists in the removal of most of the cytoplasm from the cells, by treatment with 50 µg/ml pepsin in 0.01 NHCl in a wet chamber with 2x SSC at 37 ºC during a controlled time (one or more 1 min rounds) to avoid complete digestion of the material by the pepsin.

Molecular methods

gDNA and RNA extraction

We extracted genomic DNA (gDNA) of embryo cell suspensions by 5 minutes heat shock (95 ºC) in a thermocycler while gDNA from hind legs was extracted using the kit GenElute Mammalian Genomic DNA Miniprep (Sigma) after grinding the tissue in liquid nitrogen using a porcelain mortar. DNA quality and concentration was assessed in a 1% agarose gel and also using Tecan's Infinite 200 NanoQuant and Agilent 2100 spectrophotometers. Quantification of gDNA aliquots for qPCR was performed with a Qubit 4 Fluorometer (Invitrogen) using the Qubit 1X dsDNA HS Assay kit (ThermoFisher), which allowed measuring low DNA concentrations with high accuracy.

For RNA isolation from embryo cell suspensions (immersed in Quiazol) and from adult gonads, we used the RNeasy® Lipid Tissue Mini Kit (Qiagen, Hilden, Germany), whereas for RNA extraction from hind legs we used the Real Total RNA Spin Plus kit (Durviz). In all cases, we subjected the samples to Amplification Grade DNase I (Sigma) on the column of the extraction kit (20 units of DNAse) during 30 minutes of incubation. Concentration and integrity per biological replicate was checked in MOPS agarose gel (1.5%), in the NanoQuant Infinite 200 PRO Tecan and the Agilent 2100.

Primers design

Primer pairs for PCR and qPCR experiments were designed using the Primer3 software (Untergasser et al., 2012).

Following Ruiz-Ruano et al. (2016), satellite DNA (satDNA) was amplified using divergent primers (i.e. anchored in opposite orientations) with an optimal melting temperature of 60ºC. However, for monomers shorter than 50 bp, we manually designed primers with a similar Tm and with the less stable extensive dimers predicted by the PerlPrimer software (Marshall, 2004).

Primers for PCR amplification of transposable elements (TE) were designed to amplify one fragment of 300-400 bp including, when possible, some characteristic TE regions, such as transposase, retrotransposase or integrases domains. To study the possible link between satDNA and TEs (Chapter 1), we designed mixed primers, i.e. one primer of the pair being complementary to the satDNA and the other to the TE putatively linked to the former.

For qPCR amplification of protein-coding genes, we considered, in general, primer size between 18 and 22 nucleotides, product size ranging between 80 and 200 nucleotides, primer melting temperature (Tm) between 58 and 65ºC allowing a maximum Tm of 2ºC between each primer of the pair, a maximum self complimentary of 4 and a maximum 3' self complimentary of 1.

Polymerase Chain Reaction (PCR)

PCR was used in this thesis to test the reliability of *in silico* assemblies (e.g. satDNA-transposon junctions in Chapter 1, or prior to amplicon Sanger sequencing in Chapters 1 and 2). In addition, PCR was used to generate probes for FISH (Chapters 1, 2 and 4) and, in other cases, to test the presence of B chromosomes in *E. plorans* individuals (Chapters 2, 3 and 5).

We performed the PCR reactions using the kit Horse-PowerTaq DNA polymerase (Canvax) in a total volume of 25 µL in a master mix containing 1x PCR buffer, 2 mM MgCl 2, 200 µM of dNTPs, 0.4 µM of each primer, 10 ng of DNA and 1 unit of Taq polymerase (MBL002). In general, the PCR programs used for amplification of linear genomic elements started with an initial denaturation step at 95ºC during 5 min, followed by 30 cycles at 94ºC for 30 sec, 55-60-65ºC as hybridization temperatures and 30 sec of extension at 72ºC, finishing with an extension step at 72ºC for 7 min. For tandem repeat sequences with monomer greater than 50 bp we performed 35 cycles with the same temperature, but with a denaturation step of 20 sec, 40 sec of hybridization and 20 sec of extension, reducing the hybridization time to 10 seconds in the case of monomers shorter than 50 bp. We usually repeated PCRs adjusting melting temperatures and conditions depending on the performance of the above test.

PCR results were checked in an agarose gel (Seakem LE Agarose, BioWhittaker Molecular Applications) at 1.5% or 2% in TBE using SYBR safe® (Invitrogen) as DNA gel stain. Finally, they were visualized in a Gel DocTM XR+ System (Bio-Rad) with Image Lab TM Software version 6.0.0 (Bio-Rad). Amplicon sizes were determined using the HyperLadderTM 50bp (Bioline) molecular-weight marker. When necessary, we reamplified the product of PCR cutting the gel band, squeezing it into a square of parafilm and then we repeated the PCR with 0.5 µL of the resulting solution. In case of FISH probe generation and Sanger sequencing, the PCR product was cleaned with the GenElute® PCR Clean-Up (Sigma) kit.

Quantitative Polymerase Chain Reaction (qPCR)

We used qPCR to test for the presence of some genes in the B chromosomes of *E. plorans* by estimating their number of copies in gDNA samples of *E. plorans* males harbouring different number of B chromosomes (Chapter 3). In general, we followed the protocol described by Navarro-Domínguez et al. (2017) with some modifications.

Quantitative PCR (qPCR) was performed in a Chromo4 real-time PCR thermocycler (BioRad), using SensiMix TM SYBR No-ROX Mix (Bioline). Reaction mixture contained 5µl of 3 ng/µl gDNA, 5 µl of SensiMix SYBR No-ROX Mix and 2.5 µl of each 2.5 µM primer in a total 15 µl volume. Electronic pipettes (Eppendorf Research® Pro) were used in order to minimize pipetting errors. In each experiment, a negative control (blank) was included for each pair of primers and also a reference sample, called calibrator, consisting in a

mixture of gDNA from several individuals which was used to normalize variation between plates. Finally, we performed every reaction in duplicate to reduce technical variation.

The qPCR program consisted of an initial denaturation stage (95ºC, 10 minumtes), followed by 40 cycles of 94ºC during 15 minutes, 15 minutes at the optimum primer annealing temperature and 72ºC during 15 minutes, reading plate fluorescence in each cycle. At the end of the 40th cycle, a dissociation curve from 72 to 95ºC was included to assess reaction specificity for each primer pair (melting curve). Fluorescence was measured and processed using Opticon Monitor 3.1.32 (Bio-Rad Laboratories, Inc).

Primer pair optimization

The first step was to determine the optimum annealing temperature (Ta) for each primer pair by testing identical reactions across a range of annealing temperatures. We used a gradient program in the thermocycler with temperatures ranging normally from 58 to 62°C but adjusted depending on primer annealing temperature. Each primer pair was tested for three temperatures, approx. between 2°C less and 2°C more than its Ta. All these reactions were made with an *E. plorans* gDNA mixture (calibrator) from six individuals with different number of B chromosomes. Then, qPCR products were run on 1.5% TBE-agarose gels to determine the optimum annealing temperature corresponding to the reaction producing the expected size band (amplicon size) and primer specificity was also checked in melting curves.

Efficiency calculation

After primer optimization, the amplification efficiency (fold increase per cycle) for each primer pair was calculated by a standard curve of log quantity (x-axis) vs. C(t) cycle (y-axis), performed on five 1:10 serial dilutions of the 3 ng/µl calibrator gDNA. Outliers were removed from the standard curve when efficiency was compromised and when the coefficient of variation between technical replicates was too high. The variance explained (r^2) was always greater than 0.93 (in Chapter 3). The efficiency (E) was calculated according to the equation:

$$E = 10^{(-1/slope)}; E_{\%} = (E-1)*100$$

Relative quantification of genomic abundance

After efficiency calculation, qPCR experiments were carried out with 3 ng/µl gDNA of the samples. As for efficiency calculation, negative controls were included and each reaction was performed in duplicate. Relative quantities (RQs) were calculated using each gene

efficiency referred to the calibrator sample, based on Pfaffl (2001), and using the following equations:

a) $\Delta Ct = CtC - CtS$ with CtC = Ct value of the calibrator and Cts = Ct value of each sample.

b) $Q = E^{\Delta Ct}$ with E = primer pair efficiency

c) RQ = Q/NF*

*Normalization factor: in order to increase data accuracy, hypothetical copy number of the amplification target was estimated using the URI Genomics & Sequencing Center's calculator, available at http://cels.uri.edu/gsc/cndna.html (created by Staroscik, 2004) and divided by genome size:

NF = ((initial amount in ng*6.022*10^{23})/(amplicon length in bp*1*10^{9}*650))/genome size

Statistical analysis

To ascertain whether B chromosomes in *E. plorans* harbor paralog copies of the protein-coding genes assayed, we tested whether gene RQs increased with the number of B chromosomes which, following Navarro-Domínguez et al. (2017), was our null hypothesis (H_0). For this purpose, we performed a linear regression analysis for each gene, with the RQ values as dependent variable and the number of B chromosomes as the independent variable. Prior to the regression analysis, we tested whether the RQ values for each gene fitted a normal distribution, using the Kolmogorov-Smirnov (K-S) and Shapiro-Wilks (S-W) tests. For a better fit to a normal distribution, we transformed all RQ values to natural logarithms (ln). In addition, we estimated statistical significance using DABEST (Ho et al., 2019) which performs bootstrap-coupled estimation to test and display standardized effect sizes of B chromosome presence on RQ values.

Sanger sequencing

Sanger sequencing has been performed in Macrogen (Macrogen Europe, Amsterdam, Holland).

Illumina HiSeq gDNA and RNA sequencing

Illumina sequencing of gDNA included 21 *E. plorans* libraries from the population of Torrox (Málaga, Spain), see details of sequencing in Table M.2. In addition, we sequenced the gDNA of *E. plorans* from Tanzania, Egypt and Armenia, in each case one male harbouring B chromosomes and one more male lacking them. Finally, we used gDNA libraries of *E. plorans* males deposited in SRA under the accessions SRR2970625

(gDNA_0B) and SRR2970627 (gDNA_4B) in Chapters 1, 2 and 3.

For the analysis of primate repetitive DNA sequences, we used gDNA libraries found in SRA for several species (see Appendix for details about these libraries).

We also sequenced several RNA libraries of *E. plorans* for transcriptome analysis through Illumina HiSeq of male and female embryos and adults, with or without B chromosomes (Table M.2).

Table M.2. *E. plorans* Illumina sequencing performed in this PhD thesis. Pop: population, B chr: B chromosome content, Dev stage: developmental stage, N lib: number of libraries.

Pop	Nucleic acid	B chr	Dev stage	Tissue	Sex	N lib	Platform*	Gb per lib
Torrox	gDNA	0B	Adult	Leg	Female	3	HiSeq X	~7
						3	HiSeq X Ten	~10
			Adult	Leg	Male	3	HiSeq X	~7
						3	HiSeq X Ten	~10
		1B	Adult	Leg	Female	3	HiSeq X Ten	~10
					Male	3	HiSeq X Ten	~10
		4B	Adult	Leg	Male	3	HiSeq X	~7
	RNA	0B	Adult	Leg	Female	3	HiSeq 4000	~7
					Male	3	HiSeq 4000	~6
				Gonad	Female	3	HiSeq 4000	~7
					Male	3	HiSeq 4000	~6
			Embryo (F1BxM0B)	Cell whole embryo	Female	3	HiSeq 2000	~6
					Male	3	HiSeq 2000	~5
			Embryo (F0BxM1B)	Cell whole embryo	Female	3	HiSeq 4000	~8
					Male	3	HiSeq 4000	~8
		1B	Adult	Leg	Female	3	HiSeq 4000	~7
					Male	3	HiSeq 4000	~6
				Gonad	Female	3	HiSeq 4000	~7
					Male	3	HiSeq 4000	~6
			Embryo (F1BxM0B)	Cell whole embryo	Female	3	HiSeq 2000	~6
					Male	3	HiSeq 2000	~6
			Embryo (F0BxM1B)	Cell whole embryo	Female	3	HiSeq 4000	~6
					Male	3	HiSeq 4000	~7
Tanzania	gDNA	0B	Adult	Leg	Male	1	HiSeq 2500	~8
		1B	Adult	Leg	Male	1	HiSeq 2500	~8
Egypt	gDNA	0B	Adult	Leg	Male	1	HiSeq X	~7
		+B	Adult	Leg	Male	1	HiSeq X	~7
Armenia	gDNA	0B	Adult	Leg	Male	1	HiSeq X	~8
		+B	Adult	Leg	Male	1	HiSeq X	~8

*Companies in which we performed sequencing were:

HiSeq X: Beijing Novogene Bioinformatics Technology Co., Ltd (Headquarters).
HiSeq 2500: Beijing Novogene Bioinformatics Technology Co., Ltd (Headquarters).
HiSeq X Ten: Macrogen, Inc. (Seúl, Corea del Sur).
HiSeq 2000: BGI (Beijing Genomics Institute, China).
HiSeq 4000: BGI (Beijing Genomics Institute, China).

MinION nanopore sequencing

We carried out a long-read sequencing of a 4B male in an Oxford Nanopore MinION device with flow cell version R9. We performed library preparation by means of the Nanopore Genomic Kit (SQK-LSK108) and using magnetic beads from CleanNA (CleanNGS). The starting DNA quantity for sequencing was 5 µg, without previous fragmentation, getting ~ 645 Mb in 127,000 reads (0.058x).

Bioinformatic methods

Most of the scripts implemented in this thesis can be found in the GitHub repositories https://github.com/mmarpe/ and https://github.com/fjruizruano/ngs-protocols. Below a follows a brief explanation of the main bioinformatics methods and resources used here. For further detail, see each specific chapter.

Quality control of NGS

Quality control and statistics of Illumina sequencing was performed with FastQC (Andrews, 2010), trimming was carried out using Trimmomatic (Bolger et al., 2014) and the filtering of possible contaminant or undesirable reads was done with DeconSeq v0.4.3 (Schmieder and Edwards, 2011).

Sequences viewing and editing

We used Geneious 4.8.5 of Biomatters Ltd., Auckland, New Zealand (Drummond et al., 2009) for sequence editing and visualization and, in particular cases, we used this software for sequence alignment. In addition, for viewing and exploration of BAM and BED files, we used IGV 2.4.16 (Robinson et al., 2011).

Repetitive DNA analysis

Assembly and clustering

To get a comprehensive database of repetitive DNA, we followed specific approaches for each different kind of repetitive DNA found in the genome of *E. plorans*. For transposable

elements, we performed an initial assembly using the RepeatExplorer (RE) software (Novák et al., 2013) which groups together reads with at least 80% of identity into a same cluster. Then, it represents, in a graph, the connections between reads, which can give us an idea about the structure of that element in the genome. We further assembled repetitive DNA using dnaPipeTE vb0.31 (Goubert et al.,2015). Finally, we reduced redundancy of this two assemblies using CD-HIT-EST (Fu et al., 2012) and selected transposable elements before annotation. In the case of satellite DNA (satDNA), we used the satMiner toolkit (Ruiz-Ruano et al., 2016) to assemble the satDNA sequences found in the genome of *E. plorans*. We assembled U4, U5 and U6 snDNAs and mitochondrial DNA using MITObim (Hahn et al., 2013) with known seed sequences of homologous genes from several related species. The rDNA and the histone cistron were assembled starting at RE clusters containing these elements and performing several rounds of manual assembling using reference sequences. For tRNAs identification, we used the tRNAscan-SE software (Lowe and Eddy, 1997).

Annotation

Repetitive sequences were annotated on the basis of their similarity to known repeats in other species, using RepeatMasker v4.0.5 (http://www.repeatmasker/org, Smit et al., 2013) and searching against GenBank and RepBase databases (http://www.ncbi.nlm.nih.gov/genbank, http://www.girinst.org/repbase/). They were also investigated for the presence of ORFs and specific structural features like tandem subrepeats, potential long terminal repeats (LTRs), terminal inverted repeats (TIRs), transposase and retrotransposase domains, using ORF Finder, BLAST (Altschul et al., 1997), HMMER (Finn et al., 2011) and dotmatcher (Rice et al., 2000). *E. plorans* mitochondrial DNA was annotated and characterized with MITOS (Bernt et al., 2013).

Abundance and divergence estimation

We aligned the Illumina reads to our repetitive DNA database with the RepeatMasker program using cross-match as search engine. We run RepeatMasker with option "-a" to produce a *.align file that contains information about the number of nucleotides aligned for each reference sequence as well as divergence regarding the reference. Furthermore, we generated a repetitive landscape with the script included in RepeatMasker called “calcDivergenceFromAlign.pl”. It consists of a graphical representation of the abundance for each range of divergence for every repeated sequences so we can get a general idea of abundance and diversity for certain elements. We also used RepeatMasker to align and

annotate the MinION long-read library of *E. plorans* with our custom database of repetitive DNA as reference.

Protein-coding genes analysis

Assembly and clustering

We first generated a *de novo* transcriptome which was used as reference to map genomic reads from B-carrying and B-lacking individuals, in order to identify genes putatively located on B chromosomes of *E. plorans*, because they showed higher coverage associated to B chromosome presence. RNA-seq libraries were assembled using Trinity v2.5.1 (Haas et al., 2013) with default options, including *in silico* normalization with 50x maximum coverage. Since we were only interested in knowing the presence of protein-coding genes, we extracted the CDSs using the TransDecoder software (http://transdecoder.github.io) and then reduced redundancy with CD-HIT-EST with local alignment and the greedy algorithm. For differential expression analysis we maintained the first assembly without CDS extraction and redundancy reduction.

Annotation

We annotated contigs and CDSs with the Trinotate pipeline (https://trinotate.github.io) using the SWISS-PROT database (Boeckmann et al., 2003) and functionally annotated sequences with Gene Ontology (GO; Ashburner et al., 2000). In addition, we performed functional annotation using Eukaryotic Orthologous Groups of proteins (KOG; Tatusov et al., 2003) looking for the protein sequence on the WebMGA server (http://weizhong-lab.ucsd.edu/metagenomic-analysis/server/kog). Finally, we further annotated transcripts and CDSs using Blast2GO (Götz et al., 2018).

Abundance estimation

For identification of B chromosome genes, we mapped several gDNA libraries (0B, 1B and 4B) of *E. plorans* with SSAHA2 (Ning et al., 2001) on the *de novo* set of CDSs. Abundance data was normalized as copy number per genome coverage (as described in Chapter 3) . To select CDSs potentially located in the B chromosomes of *E. plorans* we calculated the gFC for each CDS in the 4B libraries of *E. plorans*. The gFC was obtained as log2 (copies in 4B libraries/copies in 0B libraries). To set a gFC threshold to identify B-located CDSs, we assumed 1 gene copy in the 0B libraries and 3 in the 4B ones, 1 copy from As and 2 from Bs in an haploid genome. Therefore, the expected gFC threshold to consider a CDSs as located in Bs was 1.585. Then we selected the most complete transcript from the selected

CDSs and performed additional mapping of all gDNA and RNA libraries with SSAHA2, to get estimates of coverage per site along the transcript. After this, we calculated the proportion of nucleotides showing higher abundance in 4B libraries than in 0B ones (prop_nt4B>0B), we considered as uniform coverage genes (UC, probably complete in B chromosomes) those with a prop_nt4B>0B equal or higher than 0.90 and genes with irregular coverage (IC, incomplete in B chromosomes) those genes with a prop_nt4B>0B value below 0.9. We then averaged these estimates per contig and calculated gFC in high and low coverage regions. Finally, we performed SNP calling with a custom script after SSAHA2 mapping to get specific variation for genes found on B chromosomes.

Differential gene expression analysis

To estimate abundance of each transcript in each RNA sample, the raw reads (prior to *in silico* normalization) were mapped against the *de novo* reference transcriptome using Bowtie (Langmead et al., 2009), in order to obtain the gap-free alignment required by RSEM (Li and Dewey, 2011) to estimate read abundance per gene. Bowtie and RSEM were used as implemented in the Trinity software. Differential expression analysis was performed with the edgeR package in R 3.6.1 (*v3.26.8*; Robinson et al., 2010). Read counts were normalized according to the TMM method described in Robinson et al. (2010), with the "calcNormFactors" function in edgeR. Then, we identified DE features using biological replicates to fit a negative binomial generalized log-linear model with "glmFIT" and performing a likelihood ratio test to select DE features with "glmLRT" function of edgeR. We confidently considered as DEGs only those genes whose FDR remained below 0.05 and having a log2-fold-change higher than 1 between the target RNA samples.

Phylogenetic analysis

The Tree Analysis Using New Technology - TNT software (Goloboff et al., 2008) for Linux 64 (no taxon limit), updated version of 11/Dec/13, was chosen for phylogenetic analyses performed in the appendix section of this thesis. Additional phylogenetic methods in appendix included Consense v3.695 of the PHYLIP package (Felsestein, 2005) to obtain consensus trees. FigTree 1.4.3 (http://tree.bio.ed.ac.uk/) was used for graphical view and representation of phylogenetic trees.

Graphical representation of data analysis

We used R 3.6.1 (R Core Team, 2019) to graphically represent data and results, in particular we used the packages ggplot2 (v3.2.1; Wickham, 2016), ggrepel (v0.8.1;

Slowikowski, 2019), gplots (v3.0.1.1; Warner et al., 2020), grid (v3.6.1; Murrell, 2005), gridExtra (v2.3.; Auguie, 2017), moonBook (v0.2.3; Moon, 2018) and RColorBrewer (v1.1.2; Neuwirth, 2014) and webr (v0.1.5; Moon, 2020). Inkscape v0.91 (http://www.inkscape.org/) was used for further edition of figures when required.

References

Alboukadel K. (2020). ggpubr, 'ggplot2' Based Publication Ready Plots. R package version 0.2.5. https://CRAN.R-project.org/package=ggpubr.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. (1997). Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389–3402.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1), 25–29.

Auguie B. (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra.

Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, et al. (2013). MITOS: improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, 69(2), 313–9.

Bolger AM, Lohse M, Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. (2003). The SWISS–PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31(1), 365–370.

Cabrero J, Bakkali M, Bugrov A, Warchalowska-Sliwa E, López-León MD, Perfectti F, et al. (2003). Multiregional origin of B chromosomes in the grasshopper *Eyprepocnemis plorans*. *Chromosoma*, 112, 207–211.

Cabrero J, Manrique-Poyato MI, Camacho JPM. (2006). Detection of B chromosomes in interphase hemolymph nuclei from living specimens of the grasshopper *Eyprepocnemis plorans*. *Cytogenetic and Genome Research*, 114(1), 66–69.

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, Cabrero J, López-León MD, Cabral-de-Mello DC, Ruiz-Ruano FJ. (2015). Grasshoppers (Orthoptera). In: Sharakhov IV (ed) Protocols for cytogenetic mapping of arthropod genomes. *CRC Press*, 381–438.

Drummond AJ, Ashton B, Cheung M, Heled J, Kearse M. (2009). Geneious v.4.8.5. Biomatters ltd. Aukland, New Zealand.

Felsenstein J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Seattle: Department of Genome Sciences, University of Washington.

Finn RD, Clements J, Eddy SR. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(Web Server issue), W29–W37.

Fu L, Niu B, Zhu Z, Wu S, Li W. (2012). CD–HIT: accelerated for clustering the next–generation sequencing data. *Bioinformatics*, 28(23), 3150–3152.

Goloboff PA, Farris JS, Nixon KC. (2008). TNT, a free program for phylogenetic analysis. *Cladistics*, 24, 774–786.

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. (2008). High–throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, 36(10), 3420–3435.

Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. (2015). De novo assembly and annotation of the Asian tiger mosquito (*Aedes albopictus*) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (*Aedes aegypti*). *Genome Biology and Evolution*, 7(4), 1192–1205.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(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(13), e129.

Ho J, Tumkaya T, Aryal S, Choi H, Claridge-Chang A. (2019). Moving beyond P values: data analysis with estimation graphics. *Nature Methods*, 16(7), 565–566.

Langmead B, Trapnell C, Pop M, Salzberg SL. (2009). Ultrafast and memory–efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3), R25.

Li B, Dewey CN. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323.

López-León MD. (1992). Significado biológico de la hecterocromatina supernumeraria de *Eyprepocnemis plorans* (Tesis doctoral). Universidad de Granada. Retrieved from http://hdl.handle.net/10481/14156.

Lowe TM, Eddy SR. (1997). tRNAscan–SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5), 955–964.

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

Moon KW. (2020). webr: Data and Functions for Web–Based Analysis. R package version 0.1.5. https://github.com/cardiomoon/webr.

Moon KW. (2015). R statistics and graphs for medical papers. Hannaare, Seoul.

Navarro-Domínguez B, Ruiz-Ruano FJ, Camacho JPM, Cabrero J, López-León MD. (2017). Transcription of a B chromosome CAP-G pseudogene does not influence normal condensin complex genes in a grasshopper. *Scientific Reports*, 7(1), 17650.

Neuwirth E. (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1–2. https://CRAN.R-project.org/package=RColorBrewer.

Ning Z, Cox AJ, Mullikin JC. (2001). SSAHA: a fast search method for large DNA databases. *Genome Research*, 11(10), 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.

Pfaffl MW. (2001). A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Research*, 29(9), e45–e45.

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rice P, Longden I, Bleasby A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6), 276–277.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1), 24–26.

Robinson MD, McCarthy DJ, Smyth GK. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–140.

Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM. (2016). High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific Reports*, 6, 28333.

Schmieder R, Edwards R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS One* , 6(3), e17288.

Slowikowski K. (2019). ggrepel: Automatically Position Non–Overlapping Text Labels with 'ggplot2'. R package version 0.8.1. https://CRAN.R-project.org/package=ggrepel.

Smit AFA, Hubley R, Green P. (2013). RepeatMasker Open–4.0. http://www.repeatmasker.org.

Staroscik A. (2004). URI Genomics and Sequencing Center. Calculator for determining the number of copies of a template. Available at https://web.uri.edu/gsc/dsdna-calculator/.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4, 41.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Research*, 40, e115–e115.

Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, et al. (2019). gplots: Various R Programming Tools for Plotting Data. R package version 3.0.1.1. https://CRAN.R-project.org/package=gplots.

Wickham H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer–Verlag New York.

1. A step forward to decipher the correspondence between the molecular and cytological nature of satDNA

Abstract: The term satellite DNA (satDNA) encompasses a variety of tandemly arrayed and repeated genomic elements, traditionally known to constitute the building blocks of heterochromatin. Even though satDNA is accepted to play an important role in genomic structure and function, its definition is still ambiguous, as it is usually based on arbitrary conditions depending on repeat unit length or fluorescence *in situ* hybridization (FISH) pattern. Here we perform a thorough search for tandem repeats (TRs) in the grasshopper *Eyprepocnemis plorans*, by applying the satMiner toolkit on Illumina reads, which uncovered 112 different TR families. We then analyzed their physical structure at array level on long MinION reads, and also at chromosome level by FISH. The combination of the Illumina, MinION and FISH data allowed us to explore several molecular and cytological properties of these 112 TR families.

We tested the consistency between the degree of polymerization and classical cytological properties of satDNA, i.e. conspicuous FISH signal on chromosomes, in the set TR families of *E. plorans*. Molecular and cytological data showed great congruence that reached the maximum for those TR families showing maxRU>=7 repeat units in an array. In addition, we found that several TR families were related to some transposable elements of *E. plorans*, especially those families yielding no signal after FISH. The combination of the Illumina, MinION and FISH data allowed us to explore several molecular and cytological properties of TR families in a discriminant analysis.

This analysis showed two markedly different groups of TRs in *E. plorans* in terms of those properties, one group comprised families with homogenized arrays, longer RULs, higher number of repetitions in arrays, no association to TEs and showing banded FISH signals while the other group contains families with the alternatives to these properties and could be view as simple tandem repeats. These tandem repeats could behave as seeds for arising of new satDNAs, therefore both are dynamic elements interacting together with TEs to reconfigure the structure of genomes, their evolution and even their function.

Keywords: *FISH*, *Illumina, long reads*, *satellite DNA, transposable elements*

Introduction

The structural organization of genomes has been extensively studied due to its evolutionary, functional and regulatory implications for organisms. Satellite DNA (satDNA) is a repetitive component of eukaryotic genomes widely known to form part of heterochromatin regions, centromeres and telomeres (López-Flores and Garrido-Ramos, 2012) but it also takes a key role in genome rearrangements, speciation (King, 1987; Bachmann et al., 1989; Raskina et al., 2008; Robles-Rodríguez et al., 2017) or even gene regulation and diseases (Usdin, 2008; Gemayel et al., 2010; Bilgin Sonay et al., 2015; Gymrek et al., 2016; Hannan, 2018). SatDNAs form arrays of head-to-tail tandemly repeated units (monomers), they are usually AT rich and their monomer length can highly vary from a few base pairs to more than 1 kb. Satellite DNAs can comprise over 50% of some eukaryotic genomes (Wei et al., 2014) and they are known to change rapidly in sequence and genomic location, which can cause genetic incompatibilities between closely related species (Henikoff et al., 2001; Ferree and Barbash, 2009; Melters et al., 2013). SatDNA families were historically classified according to their repeat unit size in microsatellites (2-5 bp), minisatellites (5-15 bp) and satellites with longer monomers and larger and more stable array sizes than mini- or microsatellites (Charlesworth et al., 1994).

In spite of the above supported statements about satDNA an increasing controversy about its true nature and organization in genomes is arising at the same time that new research is done using different technical approaches. The lack of suitable genomic and molecular approaches to study tandem repeats has stymied progress towards understanding the nature and dynamics of satDNA. In fact, it is not clear which are the features differentiating satDNA (apart of its larger arrays nature) from simply tandem repeats (TRs) which often leads to the use of both terms arbitrarily in scientific literature to describe the same phenomenon (Silva et al., 2019; Paço et al., 2019; Easterling et al., 2020).

Traditionally, satDNA organization as single and isolated continuous blocks of repeats has been supported by molecular and cytological techniques such as fluorescence *in situ* hybridization (FISH) and gradient centrifugation (Peacock et al., 1974; Lohe and Brutlag, 1986). More complex characterization of satDNA locus has been grounded by restriction enzymes analysis (Sun et al., 1997). Recently, next generation

sequencing has represented a turning point in the analysis of satDNA (or, simply, tandem repeats) in an increasing number of species, however it is still particularly challenging to assemble satDNA, thus representing the mayor gaps genome assemblies (Miga, 2015; Tørresen et al., 2019). For example, among several tandem repeats in the human genome some are derived from transposable elements (TEs) (Ahmed and Liang, 2012), some classical satDNAs are found mainly in HORs (higher order repeats) as the alpha satellite (Aldrup-MacDonald et al., 2016) and others can be found in several locations around the genome (Warburton et al., 2008). In addition, in recent studies there has been described almost 800,000 tandem repeats in the human genome including mononucleotide repeats, microsatellites and minisatellites, more than 6,000 located within coding regions (Duitama et al., 2014). In *Drosophila*, the study of the complex structure of satellites *Responder* (a satDNA that initially exists as a dimer of two repeats) and *1.688* was recently addressed by Khost et al. (2017) finding that both satDNA families have uninterrupted blocks of homogeneous repeats alternating with "islands" of complex DNA enriched in TE insertions. The TCAST satellite in *Tribolium castaneum* can be found in the form of short arrays (up to tetramers) or embedded within a complex unit similar to DNA transposon (Feliciello et al., 2011; Brajković et al., 2012) mainly in centromeric regions whereas Cast1-Cast9 tandem repeats families in *T. castaneum* are located in euchromatic regions of the genome (Pavlek et al., 2015). On the other hand, the link between TEs and satDNA has been pointed out in the study of adjacent sequences to satDNA in mollusks species (Satović et al., 2016), the BIV160 satellite among mollusk resembling MITEs (Plohl et al., 2010; Satović and Plohl, 2013), common TE dimers as sources of tandem insertions and tandem repeats expansion (i.e. SINEs – Batistoni et al., 1995; Helitron transposon in *Drosophila* – Dias et al., 2015, 2016; Hobo transposons – McGurk and Barbash, 2018). The first genome wide study of satDNA families in grasshoppers was done in *Locusta migratoria* (Ruiz-Ruano et al., 2016) and reported 62 families of satDNA showing clustered, mixed and even non-signal FISH pattern of chromosomal location. Up to now, only one satDNA family has been characterized in the genome of *E. plorans*, the so called sat180pb being highly abundant in the B chromosomes of this species (López-León et al., 1994, 1995).

This increasing knowledge of satDNA is revealing the diverse range of traits and features that could build up its definition. This leads to the need of a consensus analysis of all key aspects of satDNA to really get to know this kind of repetitive elements. For this

purpose, we have *de novo* identified a 112 collection of TR families in the grasshopper *E. plorans* using Illumina short reads libraries at low sequencing coverage through the satMiner protocol (Ruiz-Ruano et al., 2016), analyzed their spatial arrangement from MinION long reads and study their chromosome location through FISH. We have unveiled a wide range of monomer lengths (from 4 to 455nt), arrays sizes (from 1 to 311mer) and location patterns (banded – B, dotted and banded – DB, dotted – D and yielding no FISH signal – NS) for TR families in *E. plorans*. In addition, we were able to state some molecular and cytological properties (length of the repeat unit, content in A+T, array homogenization, maximum number of repeats, association to TEs and FISH pattern) to unravel the nature and dynamics of tandem repeats in the genome.

Materials and methods

Biological material, sequencing and data acquisition

E. plorans males were collected in Torrox (Málaga), a population where individuals carrying B chromosomes are quite common (in particular the variant B24). Males were anesthetized before dissecting out some testis follicles, a few of which were fixed in 3:1 ethanol-acetic acid for cytological analysis, and the remaining testis and other body parts were frozen into liquid nitrogen for nucleic acid extraction. Two or three fixed follicles where squashed on a slide in a drop of 2% acetic orcein to determine the number of B chromosomes in each male. We extracted gDNA from hind legs of two males, one 0B and the other carrying 4B chromosomes, using the GenElute Mammalian Genomic DNA Miniprep kit (Sigma). Quality was checked by 1% TBE-agarose gel electrophoresis and concentration was measured by the Infinite M200 Pro NanoQuant (Tecan).

The gDNA from the 0B and 4B individuals was sequenced on an Illumina HiSeq 2000 platform, each yielding about 5 Gb of paired-end reads (2×101 nucleotides). These Illumina sequences are available in NCBI SRA database under SRR2970625 (gDNA 0B) and SRR2970627 (gDNA 4B) accession numbers. In addition, the gDNA from the same 4B male was sequenced in an Oxford Nanopore MinION device with flow cell version R9. We performed library preparation by means of the Nanopore Genomic Kit (SQK-LSK108) and using magnetic beads from CleanNA (CleanNGS). The starting DNA quantity for sequencing was 5 µg, without previous fragmentation, getting ~ 645 Mb in 127,000 reads.

We used also a MinION library of a *L. migratoria* male to study arrangement of TR families in the genome of this species. Details about sequencing of this library can be found in Ruiz-Ruano et al. (2018a).

For analysis of polymerization of TR families in humans we downloaded PacBio SMRT libraries from the NCBI repository ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_15kb/. This downloaded data came from one shotgun library that was prepared from a DNA sample extracted from a large homogenized growth of B-lymphoblastoid cell lines. As explained in the repository, size selection was performed targeting a narrow size band ~15kb using the sageELF DNA size-selection system from SAGE Science. We downloaded ~34 Gb in 2,675,338 circular consensus sequences (ccs) meaning ~10x genome coverage, these ccs were calculated using default parameters and a predicted read accuracy filter of Q20 (99%) using SMRT Link 6.0.

Characterization of TR families in *E. plorans* using short reads

We performed quality trimming of both Illumina libraries (0B and 4B) with Trimmomatic (Bolger et al., 2014), by removing adapters and keeping complete read pairs with Q> 20. We then followed the satMiner protocol (Ruiz-Ruano et al., 2016; details in https://github.com/fjruizruano/satminer/), which consists in several consecutive runs of RepeatExplorer (Novák et al., 2013) each followed by filtering out previously identified TRs using the DesconSeq software (Schmieder and Edwards, 2011).

We selected 2x250,000 reads in a random manner with SeqTK (https://github.com/lh3/seqtk) from the 0B library and run RepeatExplorer (RE) with default options but adding a custom database of repetitive sequences (Chapter 2 of this thesis). We selected clusters potentially containing TRs by their spherical or ring shaped graph. Then we confirmed their tandemly repeated structure dottploting the contigs showing the highest coverage in each cluster with Geneious v4.8 (Drummond et al., 2009). The reads matching these detected TRs were filtered out from the 0B library by DesconSeq, and then we prepared a new batch of mismatched reads as an input dataset for a new RE run and TR identification, followed by a subsequent filtering. In total, we performed 7 runs of RE + DesconSeq until no new TR families were found, indicating that we had found a great representation of TR families in the 0B library. We then applied this same protocol for three consecutive rounds to the 4B library, but filtering out all the TR

families that had previously been found in the 0B library and in the first two rounds of the 4B one. Through this strategy we aimed to obtain a collection of TR families, as complete as possible, in the *E. plorans* genome.

To search for possible homologies between some of the TR families identified in *E. plorans*, we performed RepeatMasker v4.0.5 (Smit et al., 2013) using the cross-match search engine between all-to-all comparisons of TR families. This allowed us to group TR families into superfamilies if they matched each other in the RepeatMasker analysis and showed identity lower than 80% when their consensus sequences were aligned. All sequences showing identity higher than 80% were considered as variants within a same family and were not included in subsequent analysis due to difficulties to clearly distinguish or differentiate them in long reads showing sequencing error rates ~15%. We built a minimum spanning tree for DNA sequences in each superfamily with Arlequin v3.5 (Excoffier and Lischer, 2010) which was graphically reproduced using R 3.6.1 (R Core Team, 2019).

Physical mapping of TR location in *E. plorans*

We tested the reliability of all TR families trying to amplify them by PCR, a strategy which also served to generate DNA probes for fluorescent *in situ* hybridization (FISH) analysis. To design appropriate primers, we aligned a collection of TR monomers extracted from the RE clusters to get a consensus sequence and choose the most conserved region to design primers in opposite orientation, ensuring a minimal distance between them or, in unavoidable cases, overlapping them up to 3 bp at the 5′ end. For this purpose, we used the Primer3 software (Untergasser et al., 2012) with an optimal melting temperature of 60 °C (primer details in Table S1.8). For families with repeat unit (RU) longer than 50 bp, we performed PCR amplification with a starting renaturation step of 95 °C for 5 min, 35 cycles with 94 °C for 20 s, with 55–65 °C as annealing temperature for 40 sec and 72 °C for 20 sec and a final 7 min extension step. A ladder pattern of PCR products was checked by electrophoresis in a 2% agarose gel. We trimmed the monomer band and extracted DNA by squeezing it in a parafilm square to reamplified 0.5 µL of the resulting solution. For TRs shorter than 50 bp, we reduced annealing time to 10 sec in order to get longer amplicons. We purified all PCR products using the GenElute PCR Clean Up kit (Sigma) and confirmed their sequence by Sanger sequencing of amplicons in Macrogen Inc (Corea) prior to probe labeling.

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 (FISH). The probes were labeled with tetramethylrhodamine-5-dUTP (red signal) or fluorescein-12-dUTP (green signal) from Roche. FISH was performed on 10-day-old embryos of *E. plorans* that were immersed during two hours in 1 ml of 0.1% colchicine prior to an osmotic shock of 20-30 min with 1 ml of distilled water. Finally, embryos were fixed in 3:1 ethanol:acetic acid and kept at 4ºC. Embryo preparation for FISH was performed following the Meredith's technique (Meredith, 1969) with some modifications (Camacho et al., 1991) and FISH was performed using the protocol described in Cabrero et al. (2003).

Abundance and divergence estimation of TR families in short reads of *E. plorans*

We used RepeatMasker (Smit et al., 2013) with "-a" option to estimate abundance and divergence for each TR family in the Illumina gDNA paired libraries trimmed as described above. We estimated the average divergence within each TR family considering distances from the sequences applying the Kimura 2-parameter model with the script calcDivergenceFromAlign.pl within the RepeatMasker suite. We normalized abundance data per number of selected nucleotides and thus represented abundance as percentage of the library, i.e., genome proportion. In order to test for possible bias in TR abundance estimations, due to possible homology with other repetitive elements, we performed two runs of RepeatMasker using two different databases as reference for mapping, one including only the TR families found in *E. plorans* and other, more comprehensive, adding all kinds of repetitive elements found in this species (Chapter 2 of this thesis). We numbered TR families in order of decreasing abundance in the 0B library when using as reference the most complete database of repetitive DNA.

Identification of TR arrays in long reads of *E. plorans*

We searched for arrays of TR families in long reads of *E. plorans* by mapping the MinION library, with RepeatMasker, against a reference database including the TR families previously identified on short reads and, in a second run, using as reference the complete database (including TEs, satDNA and other repetitive sequences found in *E. plorans*) (Chapter 2 of this thesis), in order to avoid misidentification of TRs due to cross-homology with other kind of repetitive DNA. Additionally, we applied a -div 35 to accommodate sequence similarity to distinguish between variants and families of TRs, i.e

80% (Ruiz-Ruano et al., 2016) plus the error sequencing rate of MinION reads, i.e. 5-15% (Rang et al., 2018). Finally, when multiple matches coincided on the same region of a read, we selected the high score alignment from the output file.

To evaluate array junctions and the effect of inter-array distances on the distribution of array sizes, we processed the output of RepeatMasker by transforming it to a BED format file and merging TR matches found at difference distances using BEDTools (Quinlan and Hall, 2010) implemented in a custom script found in (https://github.com/mmarpe/satION/). We considered three maximum inter-array distances (maxIDs): 3 nt, 1 RU (length of one repeat unit) and the whole read. On this basis, two consecutive arrays for a same TR family were fused together if they were separated at distances lower than the maxID established.

Molecular description of each TR family of *E. plorans*

For each TR family, we determined repeat unit length (RUL) and A+T content (%) of the consensus sequence. To analyze possible classification of TRs based on RUL and/or A+T, we grouped them into quartiles ($Q_{RUL}1$-$Q_{RUL}4$, and $Q_{A+T}1$-$Q_{A+T}4$) to compare other TR properties described below, e.g. polymerization degree and FISH pattern.

In the MinION long reads, we determined array length (AL) from the RepeatMasker output file, using a custom script (https://github.com/mmarpe/satION/), and scored the number of repeat units (NRU) as AL/RUL. We then estimated the number of monomers (NM) per TR family by scoring the number of arrays showing NRU between 0.5 and 1.5 and defined a polymerization index (PI) expressing the proportion of nucleotides, for a given TR family, arranged into 2 RUs or higher polymerization degrees, which was calculated as one minus the number of nucleotides arranged into arrays with NRU<=1.5, divided by all the nucleotides found for such TR family. We performed all these estimations considering the three maxIDs (“3 nt”, “1 RU” and “read”) described above. To test if PI was a good measure for polymerization degree, we compared it with the TSI (Tandem Structure Index, 1-PERU) values estimated for the same TRs from Illumina reads. In that respect, PERU (Proportion of External Repeats Units) was calculated after RepeatMasker mapping of Illumina reads against consensus sequence of each TR family and counting how many of them were “external/mixed” and “internal/pure” (i.e. at least ~89% of the read length) including also non-concordant reads pairs.

Long reads pose an obvious limitation to the maximum detectable array length (maxAL), determined by read length, for which reason we tested if read length could actually bias the distribution of AL and NRU. For this purpose, we estimated AL and NRU distributions by discarding short arrays located at less than 50 nt from any of read ends and, as a second approach, in a subset of MinION reads longer that 5,000 nt where detection of short arrays would theoretically be less dependent on read length, and compared the obtained AL and NRU distributions with those observed on the full long read collection.

Finally, we calculated the Kimura divergence (Kdiv) for each array and analyzed its relationship with AL, NRU and PI. Although the high sequencing error rate of MinION reads (15%) presumably introduces high experimental noise in Kdiv calculations, we considered that this noise is random and affects equally to all arrays and TR families, specially as we used this parameter only at within-TR level. Therefore, it is likely that these divergence estimations are not biased and can be useful to the purposes of quantitative comparison between TR families.

Congruence between molecular and cytological properties of TRs

As satDNA has traditionally been defined on the basis of its tandem structure and FISH pattern, we wondered whether our collection of 112 TR families could serve to define it with higher accuracy. Previous literature described satDNA by showing conspicuous signals on chromosomes after FISH analysis, and this implies that AL should surpass a minimum target size (minTS) for FISH, which Schwartzacher and Heslop-Harrison (2000) established at about 1 kb. With this in mind, we generated a spreadsheet including all molecular and cytological parameters, mentioned above, to test if every TR family fitted the maximum array length (maxAL) and FISH pattern expected for satDNA definition. As a test for this proposal, we generated a binary variable, named FIT, being equal to 1 if the FISH pattern (signal or no signal) was consistent with the maxAL observed for the same TR (i.e. maxAL>=minTS and FISH=1, or maxAL<minTS and FISH=0), or 0 if both parameters were inconsistent (i.e. maxAL>=minTS and FISH=0, or maxAL<minTS and FISH=1).

The method to obtain information from the data collection in the Dataset 1.5 spreadsheets included sorting it in order of decreasing values for a desired parameter (e.g. maxNRU to investigate polymerization degree), and then dividing the set of TRs into two groups (e.g. maxNRU=3 implies considering all TRs showing maxNRU>3 as the upper

group and all those showing maxNRU<=3 as the lower group). We then calculated the average FIT for each group and the difference in FIT between the two groups (DIF). The next step consisted in adding the 4 RU class to the 3 RU one, thus changing the composition of the upper and lower TR groups and calculating the new values for the average FIT per group and DIF between groups. We thus used the DIF parameter in resemblance to the Otsu's algorithm for automatic image thresholding to separate foreground and background (Otsu, 1979). In our case, we searched for variable values that maximized DIF between the two groups of TRs (separated at 3 RU, 4 RU, …, until 30 RU thresholds) while maximizing FIT in the upper group. This allowed finding the minimum degree of polymerization that maximized DIF and FIT, thus revealing which were the best molecular properties (e.g. minTS or maxNRU) yielding the best FIT according to the observed FISH pattern.

Relationship between TRs and other repetitive elements

Using a custom script (https://github.com/mmarpe/satION/), we scored the number of MinION reads that each TR array shared with transposable elements and other repetitive elements mapped with RepeatMasker against the complete repetitive database. The same script also extracted the distance between each TR array and every TE or repeat element, as well as the length of each match. At this respect, we considered that about constant distance between most arrays for a given TR family and a particular TE would indicate linkage between the TR and the TE. We also searched for homology between each TR family and the repetitive sequences included in RepBase (Bao et al., 2015), using the Censor software (Kohany et al., 2006). We also aligned TRs and TEs to detect any homology between them, using Geneious v.4.8.5 and, if found, we analyzed whether TR and TE were clustered together in RepeatExplorer when using Illumina reads showing homology with the TR, previously selected with BLAT (Kent, 2002). In case of TR-TE clustering, we performed an additional test by PCR amplification anchoring one primer on the TR and the other on the TE. PCR primers were designed using the Primer3 software (Untergasser et al., 2012) and PCR conditions were similar to those described above for FISH probe generation, adjusting annealing temperature for each primer pair.

Statistical analysis

The statistical analyses were performed using the Statistica 6.0 software and R 3.6.1. The analyses performed included non-parametric Spearman rank correlation, Wilcoxon

matched-pairs test, Wilcoxon one-sample test, Mann-Whitney test, Kruskal-Wallis test and MCA (multiple correspondence analysis), PCA (principal component analysis) and DA (discriminant analysis).

Results

Searching for TRs using Illumina short reads

After seven rounds with the 0B library and three with the 4B one, the satMiner protocol yielded 112 TR families in the genome of *E. plorans* (see Table 1.1), with 38 of them being grouped into 13 superfamilies each showing 50%-80% identity between the included TR families (see Fig. S1.2 for MSTs graphs). Eight TR families were identified in the 4B-carrying library (EplTR008-426, EplTR041-399, EplTR058-30, EplTR059-98, EplTR064-226, EplTR077-47, EplTR079-166 and EplTR112-11), although only the latter was practically absent in the 0B genome suggesting that it is B chromosome specific. This TR family was used in Cabrero et al. (2017; Chapter 4 of this thesis) as a marker of B chromosome presence in spermatids to test B elimination in *E. plorans* males.

The consensus sequences for the 112 TR families showed repeat unit length (RUL) ranging from 4 (EplTR096-4) to 455 nt (EplTR020-455) (mean= 155.6 nt, SE= 11.2), and AT content (%A+T) ranging between 23.3% (EplTR084-30) and 84.2% (EplTR028-19) (mean= 59.9%, SE= 0.9). RepeatMasker (RM) mapping against the consensus TR sequences showed 10.01% of TRs in the 0B genome, and a conspicuously higher amount (16.23%) in the 4B genome, in consistency with the heterochromatic nature of these B chromosomes (Henriques-Gil et al., 1984; Camacho et al., 1991). However, these abundance estimates were lower when we used the complete database of repetitive DNA as reference for the RM mapping (see Table S1.1), specifically 8.38% for the 0B genome and 14.79% for the 4B one, suggesting that some TR families can be structurally linked to other repetitive DNA (e.g. TEs). Therefore, we considered that these latter estimates were more accurate. In the 0B genome, the five TR families experiencing the highest decrease in abundance when using the complete database were, in order of decreasing change, EplTR072-98 (98.01% decrease), EplTR111-34 (97.94%), EplTR103-242 (97%), EplTR092-24 (92.77%) and EplTR032-439 (91.96%). On the other hand, in the 4B genome the five TR families showing the highest change in abundance depending on the reference database used were EplTR072-98 (99.6%), EplTR108-30 (99.24%), EplTR111-34 (98.41%), EplTR103-242

(96.60%) and EplTR041-399 (93.26%).

We also scored the abundance of the 112 TR families, described above, in two ways: i) using only them as reference, and ii) using the most complete database including other repetitive elements (described in Chapter 2), and their results were highly positively correlated (0B: r_S=0.89, p<0.00001; 4B: r_S=0.88, p<0.00001)

A comparison of abundances between the 4B and 0B genomes pointed out to several TR families as good candidates to be abundant in the B chromosomes. We thus found 80 TR families (see Table 1.1) showing gFC>0 (gFC = log2 (nt_4B_norm/nt_0B_norm), thus being overrepresented in the 4B library and potentially located on B chromosomes (see Chapter 2 for details).

Heterogeneous FISH patterns of TR location in *E. plorans*

We determined the chromosomal location of the 112 TR families found through satMiner by means of FISH. We classified the FISH patterns found for the 112 TRs into the same four types described in Ruiz-Ruano et al. (2018): banded (B), dotted-banded (DB), dotted (D) and no-signal (NS) patterns (see Fig. 1.1 for examples). 47 out of the 112 TR families showed a B pattern, characterized by the presence of conspicuous FISH bands on one or more chromosome pairs, i.e. the classical pattern for satDNA. D and DB patterns were observed for 28 TR families, and were characterized by the presence of small dots of fluorescence signals (D, 15) or even the presence of additional small bands (DB, 13). Finally, 37 TR families failed to show bands or dots, and the FISH signal was absent or, in some cases, it was barely a weak halo on chromosomes, for which reason we classified them as NS type (see Table 1.1 and Dataset 1.1 for details about FISH patterns for all TR families). Interestingly, we found significant differences in abundance between the four FISH patterns of TRs in both 0B and 4B (not shown here) libraries (Kruskal-Wallis test: H= 12.82, df= 3, p= 0.00503). The Mann-Whitney tests showed that TR families yielding any kind of FISH signal (B, DB and D) showed higher abundance than those failing to show FISH signal (NS) (z= 3.26, p= 0.001126), whereas the KW test showed not significant differences between B, DB and D patterns (H= 2.07, df= 2, p= 0.3545).

The telomeric repeat was identified as the EplTR022-5 family, and its sequence was identical to that described for other grasshopper species (Frydrychová et al., 2004). FISH analysis showed the typical location of this family on the telomeres of all chromosomes. We also found that the EplTR005-49 family, quite abundant, was the only TR being

located on the centromeric region of all chromosomes, on which basis we consider that it is the best candidate to participate in centromeric function in *E. plorans*. Finally, we found that the EplTR112-11 and EplTR106-323 families showed FISH signals only on B chromosomes, thus suggesting that they are B-specific (although the latter showed some counts in the 0B library). This makes them putatively useful as markers to identify B chromosomes even in interphase cells (Cabrero et al., 2017; see Chapter 4).

Some TR families showed colocalization on the same chromosomes, especially those families showing sequence homology thus belonging to a same superfamily (see Supplementary Information). To test for possible cross-hybridization, in these cases we performed FISH by means of labeled oligos designed for specific short regions differing in sequence between the different TR families (see sequences in Table S1.9). These analyses reinforced the conclusion that these TR families actually share a common chromosomal location, which was also supported by the finding of several arrays for the colocated TR families being interspersed in a same MinION read (see Table S1.1 and Fig. S1.3).

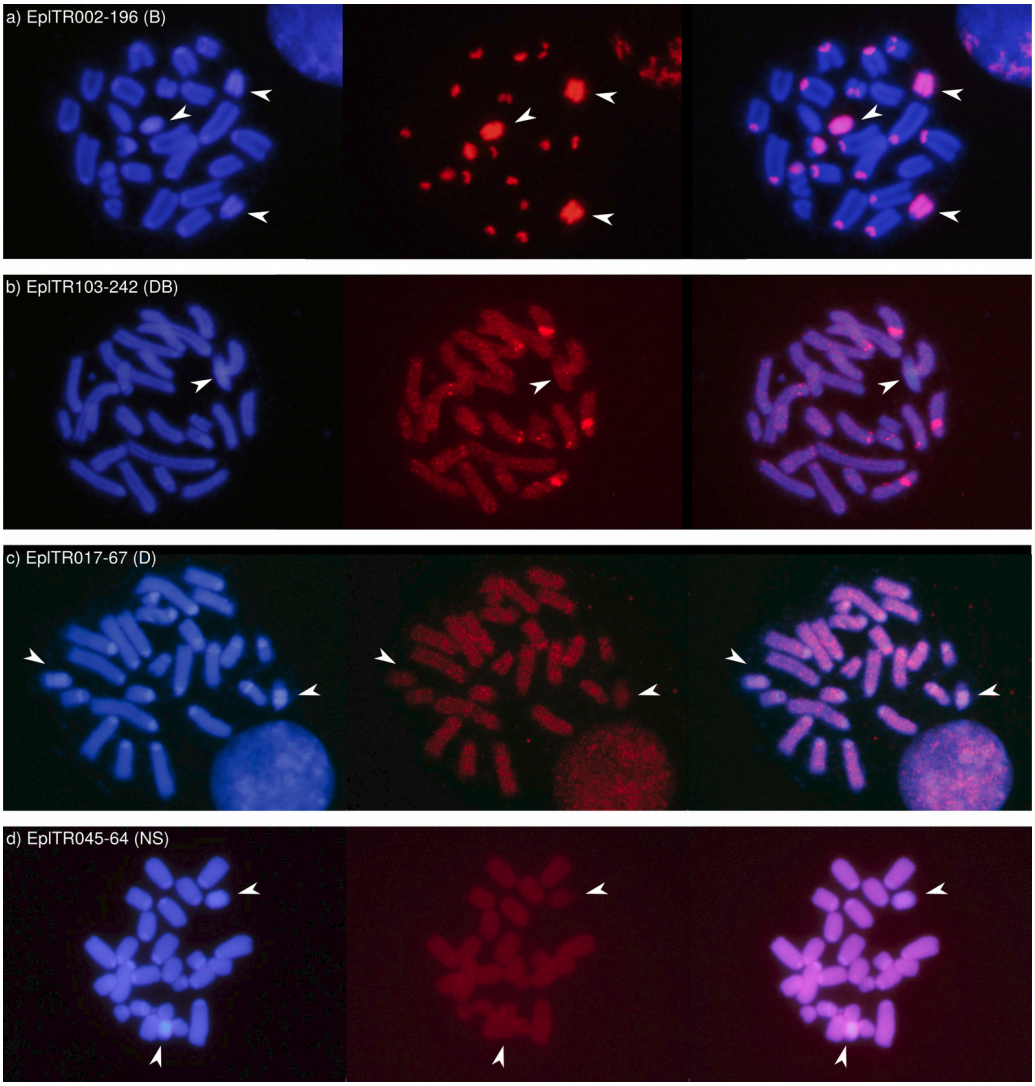


Figure 1.1. Examples of the four FISH patterns found in the 112 TR families within the *E. plorans* genome. a) EplTR002-196 representing the conventional banded pattern (B) described for satDNA. b) Dotted-banded (DB) pattern characterized by very small FISH signals dispersed across whole chromosome length and some incipient bands. c) Dotted (D) pattern being similar to the latter but without showing any band. d) Absence of FISH signal characterizing the NS pattern. B chromosomes are pointed out by arrowheads.

Identification of TR arrays in MinION long reads of *E. plorans*

TR abundance in MinION reads was positively correlated with that found in Illumina reads regardless of the reference database used for estimation, the TR-exclusive one (r_S=0.5166, p< 0.00001, N=112) or the full repeat database (r_S=0.6956, p<0.00001, N= 112). For subsequent analysis, we considered TR abundances obtained with the most complete database of repetitive DNA (see Chapter 2) as it yielded more accurate figures by discounting possible TR homologies with elements as abundant as, for instance, TEs.

Four TR families were not detected in the MinION reads: EplTR087-344 (DB), EplTR105-39 (NS), EplTR108-30 (NS) and EplTR111-34 (NS), the former being DB with long RUL and the three latter being NS with shorter RUL. The remaining TR families were found in the MinION reads, but 13 of them failed to show polymerization stages higher than dimers (i.e. >1.5 RU): EplTR013-320 (D), EplTR020-455 (D), EplTR023-301 (D), EplTR024-391 (D), EplTR032-439 (DB), EplTR048-249 (D), EplTR053-401 (B), EplTR078-144 (D), EplTR080-159 (NS), EplTR091-65 (B), EplTR93-50 (B), EplTR103-242 (DB) and EplTR104-37 (DB). Six out of these TRs showed D pattern for FISH, three were DB, three B and one NS, indicating high predominance of dotted patterns, i.e. 9 out of the 13 TRs that showed D or DB patterns. These 17 TR families were discarded for subsequent analysis as they gave no useful information about polymerization degree (Table 1.1).

Surprisingly, the telomeric repeat EplTR022-5 was found in a very low number of MinION reads (55 reads), contrary to the expected for a repetitive sequence found on all chromosomes (see FISH pattern in Dataset 1.1). For this reason, we decided to discard this TR family for subsequent analysis, as well as EplTR096-4, because their extremely short repeat units might imply some difficulty in sequencing which might bias abundance estimations. However, the remaining 93 TR families were useful for analysis of array size distribution, inter-array distances and polymerization index in long MinION reads.

RUL quartile classification included 22 of these TRs in $Q_{RUL}1$ (11-47 nt), 24 in $Q_{RUL}2$ (49-124 nt), 24 in $Q_{RUL}3$ (131-233 nt) and 23 in $Q_{RUL}4$ (238-426 nt). Likewise, AT content, 25 were included in $Q_{A+T}1$ (23.3-54.5%), 22 in $Q_{A+T}2$ (54.6-60.5%), 24 in $Q_{A+T}3$ (60.6-64.1%) and 22 in $Q_{A+T}4$ (64.5-84.2%).

Abundance estimations for these 93 TR families in the 4B individual were positively correlated with RUL whether they had been estimated on Illumina (r_S= 0.469, p< 0.00001

0.000127) or MinION (r_S=0.512, p< 0.00001) reads, and negatively with A+T% in the case of MinION (r_S= -0.31642, p= 0.002) but not Illumina (r_S= -0.14618, p= 0.16207) reads, suggesting that TRs with longer RUL and lower AT content showed higher abundance in the 4B individual. In fact, abundance differed significantly between RUL quartiles for both Illumina (0B: H= 20.107, p= 0.0001613; 4B: H= 25.406, p= 1.27e-05) and MinION (H= 29.976, p= 1.396e-06) reads. Post-hoc comparisons by the Mann-Whitney test revealed that the main differences were found between $Q_{RUL}1$ and the three other quartiles (Illumina 0B: $Q_{RUL}1$-$Q_{RUL}2$: p= 0.0177; $Q_{RUL}1$-$Q_{RUL}3$: p= 0.0003; $Q_{RUL}1$-$Q_{RUL}4$: p= 0.0003; Illumina 4B: $Q_{RUL}1$-$Q_{RUL}2$: p= 0.0093; $Q_{RUL}1$-$Q_{RUL}3$: p= 3.2e-05; $Q_{RUL}1$-$Q_{RUL}4$: p= 3.2e-05; MinION 4B: $Q_{RUL}1$-$Q_{RUL}2$: p= 0.0442; $Q_{RUL}1$-$Q_{RUL}3$: p= 2.5e-05; $Q_{RUL}1$-$Q_{RUL}4$: p= 1.9e-05; $Q_{RUL}2$-$Q_{RUL}3$: 0.0035; $Q_{RUL}2$-$Q_{RUL}4$: 0.0022), with no significant differences between $Q_{RUL}2$, $Q_{RUL}3$ and $Q_{RUL}4$ (or at low significance in the case of MinION abundance), indicating that the shortest TRs tend to be less abundant.

However, the four A+T% quartiles showed low significant differences when abundance was estimated on MinION reads (4B: H= 10.612, p= 0.01402), but not on Illumina reads (0B: H= 3.4482, p= 0.3275; 4B: H= 4.2281, p= 0.2379), suggesting some difference in sequencing bias between the two techniques in respect to AT content.

Kimura divergence (Kdiv) was accurately measured on Illumina reads, due to the low error rate of this technology. We found significant differences between RUL quartiles for 0B Illumina reads (0B: H= 11.985, p= 0.007435; 4B: H= 11.91, p= 0.007697). The Mann-Whitney test showed significant differences between $Q_{RUL}1$ and $Q_{RUL}3$ in 0B (p= 0.026), $Q_{RUL}1$ and $Q_{RUL}4$ (0B: p= 0.026; 4B: p= 0.011), and also between $Q_{RUL}2$ and $Q_{RUL}4$ (0B: p= 0.026; 4B: p= 0.034) and $Q_{RUL}2$ and $Q_{RUL}3$ in 0B (p= 0.026), with the longest TRs (higher quartiles) showing lower divergence. However, no differences in K2P divergence were observed between the four A+T quartiles (0B: H= 4.9115, p= 0.1784; 4B: H= 5.7379, p= 0.1251). Taken together, these results revealed that $Q_{RUL}1$ TRs (with RUL up to 47 nt long) display abundance and divergence figures being remarkably different to those observed in the remaining TRs.

Table 1.1. Basic features of TR families found after satMiner in 0B and 4B males of *E. plorans* (SF – superfamily, RUL – repeat unit length, AT% – content of AT in consensus sequence). See also normalized abundance (Prop.), divergence (Div.) and gFC as log2(Prop4B/Prop0B). FISH pattern included banded (B), dotted-banded (DB), dotted (D) and no-signal (NS) TR families.

ID satMiner	SF	RUL	AT%	Illumina HiSeq 2000				gFC 4B	4B male MinION Prop.	FISH
				0B male		4B male				
				Prop.	Div.	Prop.	Div.			
EplTR001-180	1	180	58.9	0.02149047	5.93	0.04360279	5.67	1.20	0.03161887	B
EplTR002-196	1	196	53.9	0.01715752	13.56	0.05441618	13.3	1.84	0.04666456	B
EplTR003-159		159	63.3	0.00878571	6.59	0.00861395	6.61	0.15	0.00609522	B
EplTR004-196	2	196	54.6	0.00763427	4.71	0.00616012	4.74	-0.13	0.00017620	D
EplTR005-49		49	49	0.00718695	14.87	0.00961228	15.3	0.59	0.01117925	B
EplTR006-147	3	147	57.8	0.00285965	6.05	0.00792312	6.03	1.65	0.00655739	B
EplTR007-246	4	246	57.3	0.00252802	5.64	0.00236046	5.62	0.08	0.00192743	B
EplTR008-426	5	426	51.3	0.00147199	14.30	0.00133080	14.35	0.03	0.00078055	B
EplTR009-300		300	48.3	0.00132331	8.45	0.00119319	8.53	0.03	0.00037810	DB
EplTR010-202	1	202	54	0.00125661	18.28	0.00104703	18.27	-0.09	0.00086955	B
EplTR011-174		174	63	0.00084656	21.19	0.00078860	21.28	0.07	0.00054190	NS
EplTR012-131		131	52.7	0.00081676	8.78	0.00077009	8.78	0.09	0.00081441	NS
EplTR013-320		320	63.1	0.00059389	17.97	0.00053637	18	0.03	--	D
EplTR014-285	6	285	40.4	0.00056636	4.75	0.00053662	4.87	0.10	0.00091871	B
EplTR015-351	5	351	56.1	0.00050611	11.88	0.00054301	10.15	0.28	0.00028075	DB
EplTR016-31		31	64.5	0.00049771	10.75	0.00048001	10.14	0.12	0.00024316	NS
EplTR017-67	7	67	62.7	0.00049199	4.38	0.00046299	4.35	0.09	0.00022340	D
EplTR018-75		75	61.3	0.00044634	10.98	0.00040399	10.96	0.03	0.00030936	D
EplTR019-75	3	75	56	0.00043342	9.41	0.00042179	7.84	0.14	0.00002393	B
EplTR020-455		455	59.3	0.00037433	5.50	0.00034682	5.41	0.07	--	D
EplTR021-165		165	64.6	0.00029330	6.90	0.00026141	7.01	0.01	0.00008220	DB
EplTR022-5		5	60	0.00026483	0.84	0.00023395	3.12	0.00	0.00001222	B
EplTR023-301		301	68.8	0.00024679	12.90	0.00022630	13.04	0.05	--	D
EplTR024-391		391	63.7	0.00022792	9.82	0.00021596	9.77	0.10	--	D
EplTR025-238	8	238	59.7	0.00020548	14.50	0.00026294	13.41	0.53	0.00022936	B
EplTR026-176		176	65.9	0.00020085	9.66	0.00015099	9.58	-0.24	0.00012773	B
EplTR027-176		176	56.2	0.00018810	6.33	0.00015989	6.49	-0.06	0.00020434	DB

EplTR028-19	10	19	84.2	0.00018620	4.30	0.00016737	4.22	0.02	0.00005605	NS
EplTR029-411		411	56.7	0.00018379	1.02	0.00015689	1.06	-0.05	0.00012500	B
EplTR030-267	4	267	61.4	0.00016719	15.73	0.00014918	15.07	0.01	0.00017309	B
EplTR031-324		324	54.9	0.00016263	8.19	0.00009182	8.62	-0.65	0.00003944	B
EplTR032-439		439	54	0.00015161	8.08	0.00014303	7.85	0.09	--	DB
EplTR033-158		158	69	0.00015011	12.27	0.00014066	12.28	0.08	0.00002369	NS
EplTR034-123		123	63.6	0.00014967	6.12	0.00013326	5.97	0.01	0.00005493	D
EplTR035-239	8	239	58.2	0.00014550	10.70	0.00013024	10.79	0.02	0.00015224	B
EplTR036-240	9	240	61.6	0.00013829	2.79	0.00012724	2.75	0.06	0.00006677	B
EplTR037-239	8	239	61.1	0.00013392	3.73	0.00011717	4.15	-0.02	0.00009552	B
EplTR038-17		17	52.9	0.00012601	10.10	0.00012056	10.22	0.11	0.00004472	D
EplTR039-118		118	64.6	0.00012024	19.36	0.00010427	19.41	-0.03	0.00002361	DB
EplTR040-113		113	60.2	0.00011984	2.10	0.00011663	2.06	0.14	0.00000506	D
EplTR041-399	5	399	50.1	0.00011273	9.05	0.00009513	9.41	-0.07	0.00010623	DB
EplTR042-225	4	225	61.3	0.00011061	1.00	0.00008220	1.13	-0.25	0.00005440	B
EplTR043-266	6	266	41.5	0.00010695	10.17	0.00009414	9.93	-0.01	0.00005814	B
EplTR044-210		210	61	0.00010686	14.41	0.00012188	15.37	0.37	0.00015447	B
EplTR045-64		64	45.3	0.00010634	17.66	0.00009813	17.6	0.06	0.00009953	NS
EplTR046-136		136	48.5	0.00010309	2.31	0.00007893	2.33	-0.21	0.00008869	B
EplTR047-95		95	70.5	0.00010306	8.99	0.00009144	9.01	0.00	0.00000051	NS
EplTR048-249		249	67.5	0.00009777	9.60	0.00008736	9.45	0.01	--	D
EplTR049-367		367	61.6	0.00009717	4.30	0.00018647	3.72	1.12	0.00015139	B
EplTR050-124		124	64.5	0.00009689	11.59	0.00008889	11.46	0.05	0.00003456	DB
EplTR051-111		111	75.7	0.00009412	14.16	0.00008790	14.42	0.08	0.00007775	NS
EplTR052-54		54	61.1	0.00008981	12.46	0.00008355	12.37	0.07	0.00003777	NS
EplTR053-401		401	65.3	0.00008894	8.32	0.00008337	8.22	0.08	--	B
EplTR054-269		269	60.2	0.00008494	4.31	0.00008185	4.64	0.12	0.00004196	B
EplTR055-64		64	64.1	0.00008157	16.58	0.00007595	16.98	0.07	0.00002830	NS
EplTR056-209		209	63.6	0.00007651	16.03	0.00009403	14.74	0.47	0.00009152	B
EplTR057-231		231	62.8	0.00007165	8.32	0.00007599	7.96	0.26	0.00001826	NS
EplTR058-30	11	30	50	0.00006762	10.12	0.00005965	10.05	-0.01	0.00006655	NS
EplTR059-98		98	41.8	0.00006395	15.47	0.00005720	15.45	0.01	0.00007834	NS
EplTR060-33	7	33	54.5	0.00006292	18.92	0.00005586	18.95	0.00	0.00021078	NS

EplTR061-99	12	99	55.5	0.00005604	18.11	0.00005259	18.37	0.08	0.00000589	NS
EplTR062-21		21	71.4	0.00005595	13.12	0.00004922	12.9	-0.01	0.00001293	NS
EplTR063-239	9	239	60.7	0.00005328	7.31	0.00005475	7.72	0.21	0.00006383	B
EplTR064-226		226	49.6	0.00005093	3.89	0.00004800	4.26	0.09	0.00007510	B
EplTR065-233		233	58.5	0.00004918	18.88	0.00007514	16.71	0.79	0.00004030	B
EplTR066-252	4	252	61.1	0.00004778	14.10	0.00005277	11.71	0.32	0.00001025	B
EplTR067-27		27	74.1	0.00004641	5.26	0.00004227	5.23	0.04	0.00001580	NS
EplTR068-69		69	40.6	0.00004589	8.95	0.00003962	9.09	-0.04	0.00008283	D
EplTR069-87		87	70.1	0.00004472	6.10	0.00004501	6.25	0.18	0.00002164	NS
EplTR070-21	10	21	76.2	0.00004404	10.19	0.00004346	10.38	0.16	0.00002738	NS
EplTR071-180		180	60.6	0.00004138	4.31	0.00004047	4.28	0.14	0.00003293	B
EplTR072-98	2	98	50.5	0.00004000	13.25	0.00003485	13.35	-0.02	0.00000956	D
EplTR073-297		297	57.6	0.00003944	6.70	0.00006876	7.07	0.98	0.00018557	B
EplTR074-241	9	241	61	0.00003212	7.54	0.00003396	7.6	0.26	0.00003382	B
EplTR075-136		136	66.9	0.00003106	8.47	0.00003634	7.33	0.40	0.00002698	NS
EplTR076-139		139	56.2	0.00002862	9.39	0.00002130	10.35	-0.25	0.00002421	B
EplTR077-47		47	53.2	0.00002841	11.89	0.00002555	11.51	0.02	0.00003228	B
EplTR078-144		144	69.6	0.00002604	11.03	0.00002168	10.89	-0.09	--	D
EplTR079-166		166	34.3	0.00002537	6.01	0.00004416	5.25	0.98	0.00006312	B
EplTR080-159		159	67.3	0.00002371	8.14	0.00002664	8.42	0.34	--	NS
EplTR081-65		65	66.2	0.00002319	6.43	0.00002341	6.15	0.19	0.00003994	NS
EplTR082-91		91	76.9	0.00002302	15.80	0.00002114	16.16	0.05	0.00000687	NS
EplTR083-238	9	238	60.1	0.00002149	10.53	0.00001919	10.02	0.01	0.00003259	B
EplTR084-30		30	23.3	0.00002141	12.90	0.00001846	13.03	-0.04	0.00002339	NS
EplTR085-81	13	81	63	0.00002087	5.26	0.00002381	5.35	0.37	0.00001959	NS
EplTR086-82		82	64.6	0.00001978	23.50	0.00001482	23.79	-0.24	0.00001253	B
EplTR087-344		344	66	0.00001930	8.93	0.00001910	8.45	0.16	--	DB
EplTR088-238	9	238	60.5	0.00001898	8.80	0.00001933	9.1	0.20	0.00002743	B
EplTR089-226	9	226	58.7	0.00001798	12.50	0.00001625	12.26	0.03	0.00000662	B
EplTR090-40	11	40	60	0.00001754	15.81	0.00001683	15.94	0.12	0.00000865	B
EplTR091-65		65	50.8	0.00001659	10.40	0.00001319	11.72	-0.16	--	B
EplTR092-24		24	70.8	0.00001490	12.29	0.00001535	12.33	0.22	0.00000176	NS
EplTR093-50		50	60	0.00001485	8.60	0.00001253	9.33	-0.07	--	B

EplTR094-36		36	52.8	0.00001433	22.05	0.00001257	22.43	-0.01	0.00000340	NS
EplTR095-36		36	61.1	0.00001279	4.55	0.00001131	4.27	0.00	0.00000046	D
EplTR096-4		4	50	0.00001030	25.30	0.00000613	23.29	-0.57	0.00000444	B
EplTR097-51	12	51	60.8	0.00000972	10.80	0.00000844	9.99	-0.03	0.00000090	NS
EplTR098-33		33	72.7	0.00000946	19.71	0.00000890	19.42	0.09	0.00000010	NS
EplTR099-33		33	63.6	0.00000877	23.47	0.00000857	23.18	0.14	0.00000273	NS
EplTR100-32		32	75	0.00000849	17.61	0.00000779	17.74	0.05	0.00000192	NS
EplTR101-18		18	55.6	0.00000767	14.87	0.00000682	15.08	0.00	0.00000172	NS
EplTR102-110	13	110	65.5	0.00000754	11.51	0.00000875	11.2	0.39	0.00000174	NS
EplTR103-242		242	62	0.00000722	9.16	0.00000769	6.74	0.27	--	DB
EplTR104-37		37	75.7	0.00000664	2.77	0.00000574	2.64	-0.03	--	DB
EplTR105-39		39	69.2	0.00000544	9.05	0.00000453	8.99	-0.09	--	NS
EplTR106-323		323	63.8	0.00000243	10.19	0.00011914	0.69	5.79	0.00007643	B
EplTR107-33		33	66.7	0.00000214	26.23	0.00000173	25.44	-0.13	0.00000094	DB
EplTR108-30		30	70	0.00000110	9.81	0.00000090	9.84	-0.10	--	NS
EplTR109-15		15	53.3	0.00000086	5.94	0.00000080	5.24	0.06	0.00000036	DB
EplTR110-11		11	54.5	0.00000064	12.41	0.00000042	12.19	-0.45	0.00000025	NS
EplTR111-34		34	67.6	0.00000015	15.97	0.00000036	16.72	1.47	--	NS
EplTR112-11		11	54.5	0.00000002	30.88	0.00003492	9.1	11.09	0.00000205	B

Degree of polymerization of TR families in the genome of *E. plorans*

We measured all array lengths (AL) for each TR family on MinION reads and calculated the number of repeat units (RU) within each array as the AL/RUL quotient. For each TR family, we defined two new variables: maximum array length (maxAL) and maximum number of repeat units (maxRU= maxAL/RUL), both being determined by the longest array found in any MinION read showing sequences for a given TR.

Graphical representation of AL values for each TR family revealed four distribution patterns (see Fig. 1.2, Dataset 1.2 and Table S1.1): right-skewed (RS: 57 TR families), unimodal (UM: 15), bimodal (BM: 6) and uniform (U: 15) distributions. Interestingly, the most abundant class (RS) (e.g. EplTR001 and EplTR002) is characterized by a high presence of short TR arrays (1 RU, 2 RUs or 3 RUs) which is reminiscent of the satDNA

dissemination hypothesis by Ruiz-Ruano et al. (2016). To investigate at which extent MinION read length (see Fig. S1.1) influences the frequency of these extremely short arrays, we built the same histograms but discarding those arrays found at less than 50 nt from each end of the read and selecting reads longer than 5,000 nt. If so, we would expect a decrease in the number of short arrays that actually derived from longer arrays that were truncated during MinION sequencing. However, the histograms looked very similar (see Fig. 1.2 and Dataset 1.3) and we concluded that the presence of a high amount of extremely short arrays is not a byproduct of sequencing. We also built the histograms of array sizes after scoring them at three maximum inter-array distances (maxID), specifically "3 nt", "1 RU" and "read", but they changed inappreciably (see Fig. 1.2 and Supplementary Dataset 1.4).

We defined a polymerization index (PI) for each TR family (see Table 1.2) by subtracting the abundance (in nt) of arrays showing NRU<=1.5 (i.e. lower than dimers) to the total abundance of each TR family. PI was positively correlated with the maximum number of repeats found for each TR family regardless the maxID considered (3 nt: r_S= 0.48, p= 0.00001; 1 RU: r_S= 0.39, p= 0.00092; read: r_S= 0.45, p= 0.00006). To ascertain whether PI can be considered an accurate index reflecting the polymerization stage of a TR, we analyzed its correlation with TSI (which measures the proportion of RUs structured in tandem using Illumina reads, see Materials and methods). When the 93 TR families were included, no clear correlation was apparent between PI and TSI at the three maxIDs scored ("3 nt": r_S= 0.1, p= 0.36; "1 RU": r_S= 0.16, p= 0.12; "read": r_S= 0.21, p= 0.045). However, when this analysis was separately performed for each RUL quartile it became apparent that TSI and PI showed no correlation at the $Q_{RUL}1$ scored ("3 nt": r_S= -0.01, p= 0.958; "1 RU": r_S= -0.01, p= 0.958; "read": r_S= 0.14, p= 0.534) whereas they were positively correlated at the three maxIDs for $Q_{RUL}2$ ("3 nt": r_S= 0.58, p= 0.00283; "1 RU": r_S= 0.60, p= 0.00215; "read": r_S= 0.57, p= 0.00376) and $Q_{RUL}3$ ("3 nt": r_S= 0.81, p= 0.000001; "1 RU": r_S= 0.68, p= 0.000246; "read": r_S= 0.59, p= 0.002215) and only at "3 nt" for $Q_{RUL}4$ ("3 nt": r_S= 0.51, p= 0.0136; "1 RU": r_S= 0.29, p= 0.1865; "read": r_S= 0.37, p= 0.07915). This result provides new evidence that $Q_{RUL}1$ TRs, i.e. those being shorter than 48nt, behave differently to longer TRs and that the others show a good correspondence between the tandem structure (TSI) and polymerization (PI) indexes.

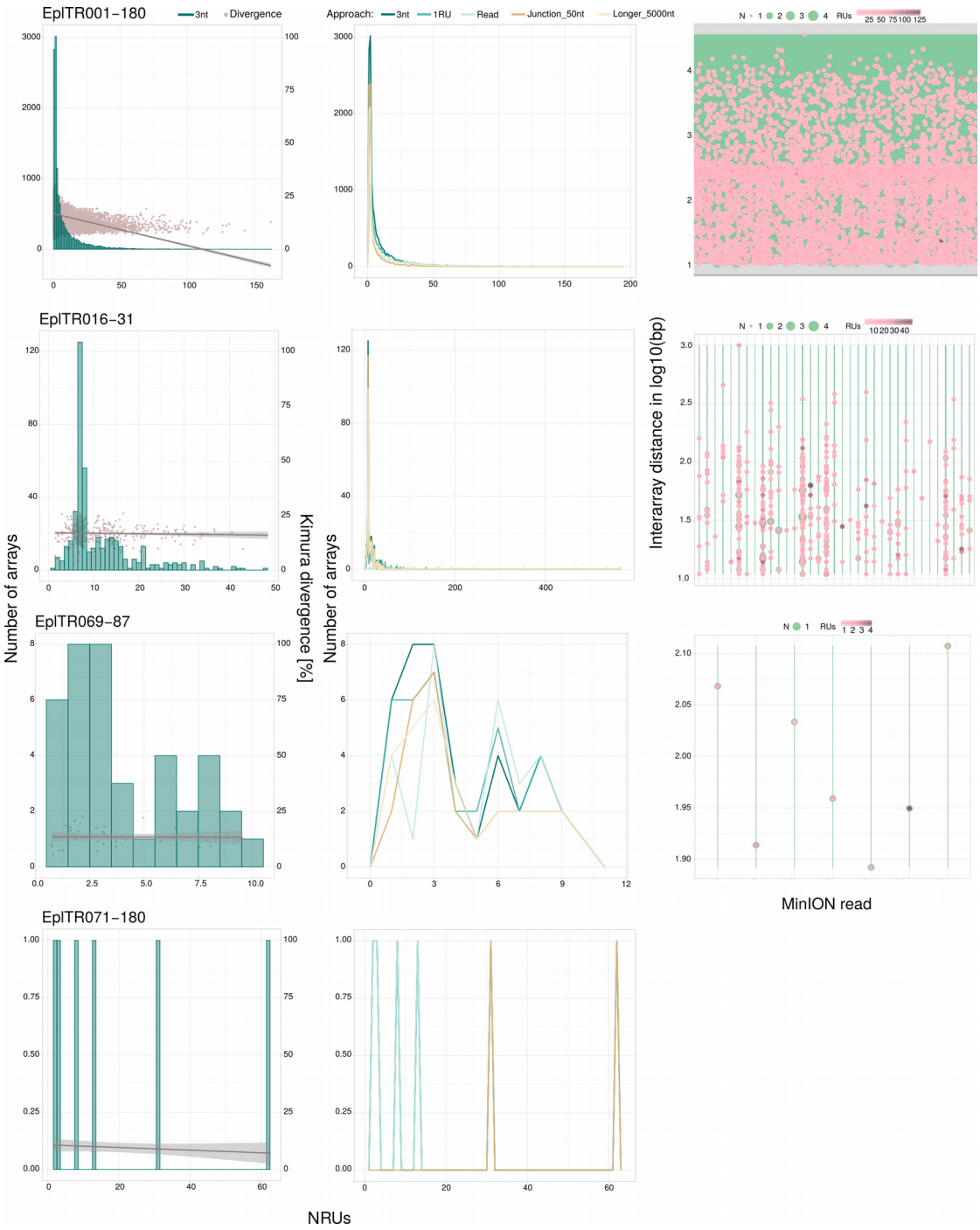


Figure 1.2. Patterns of NRU abundance distribution. Right-skewed (RS) histograms are represented here by the TR family EplTR001-180 with high number of short arrays in the genome. The family EplTR016-31 represents a unimodal (UM) distribution with the highest peak corresponding around

7RU arrays. In contrast, the family EplTR069-87 shows a bimodal (BM) distribution of RU with two peaks, around 2 RUs and 8RUs. RUs in arrays of the family EplTR071-180 are distributed uniformly (U) in the genome. See that these distributions do not change considering different maxIDs nor selecting longest reads (>5000 nt) or excluding terminal arrays (junctions < 50 nt). Finally, we show examples of raw inter-array distance within reads in the four TR families displayed in the graphs, except for EplTR071-180 were all reads harbored a single TR array.

Table 1.2. Polymerization properties (PI – polymerization index, maxAL – maximum array length, maxRU – maximum number of repeats unit in one array) of TR families found in MinION reads (93 families). Parameters calculations were performed considering different maximum inter-array distance (maxIDs) to collapse arrays: 3 nt, 1 RU or the read length.

ID satMiner	FISH	PI			maxAL			maxRU		
		3 nt	1 RU	Read	3 nt	1 RU	Read	3 nt	1 RU	Read
EplTR001-180	B	0.974	0.985	0.995	29195	29195	34711	162.19	162.19	192.84
EplTR002-196	B	0.973	0.983	0.994	32222	32222	42001	164.40	164.40	214.29
EplTR003-159	B	0.994	0.997	1.000	28174	28174	31258	177.20	177.20	196.59
EplTR004-196	D	0.003	0.005	0.011	384	424	740	1.96	2.16	3.78
EplTR005-49	B	0.999	0.999	1.000	15266	15266	27458	311.55	311.55	560.37
EplTR006-147	B	0.969	0.976	0.999	20288	20288	24496	138.01	138.01	166.64
EplTR007-246	B	0.654	0.945	0.992	3986	24823	24940	16.20	100.91	101.38
EplTR008-426	B	0.018	0.029	0.049	2604	3500	3624	6.11	8.22	8.51
EplTR009-300	DB	0.095	0.189	0.229	6857	7783	9196	22.86	25.94	30.65
EplTR010-202	B	0.035	0.896	0.996	510	10076	16680	2.52	49.88	82.57
EplTR011-174	NS	0.706	0.728	0.777	732	732	1185	4.21	4.21	6.81
EplTR012-131	NS	0.807	0.813	0.832	756	756	966	5.77	5.77	7.37
EplTR014-285	B	0.903	0.990	0.994	11625	22040	22040	40.79	77.33	77.33
EplTR015-351	DB	0.320	0.345	0.428	2247	2318	3535	6.40	6.60	10.07
EplTR016-31	NS	1.000	1.000	1.000	1503	4205	17564	48.48	135.65	566.58
EplTR017-67	D	0.921	0.922	0.952	2165	2165	4522	32.31	32.31	67.49
EplTR018-75	D	0.931	0.931	0.941	433	433	649	5.77	5.77	8.65
EplTR019-75	B	0.972	0.972	0.978	439	439	886	5.85	5.85	11.81
EplTR021-165	DB	0.688	0.699	0.709	4573	4573	4573	27.72	27.72	27.72
EplTR025-238	B	0.865	0.913	0.980	17820	17820	17820	74.87	74.87	74.87
EplTR026-176	B	0.996	0.996	0.996	12596	12596	16190	71.57	71.57	91.99
EplTR027-176	DB	0.995	0.997	0.997	17171	17336	27239	97.56	98.50	154.77
EplTR028-19	NS	0.995	0.995	0.995	124	124	297	6.53	6.53	15.63

EplTR029-411	B	0.888	0.956	0.960	5696	5696	11878	13.86	13.86	28.90
EplTR030-267	B	0.604	0.837	0.854	2251	14917	18933	8.43	55.87	70.91
EplTR031-324	B	0.499	0.914	0.914	1280	4879	4879	3.95	15.06	15.06
EplTR033-158	NS	0.043	0.068	0.068	336	373	373	2.13	2.36	2.36
EplTR034-123	D	0.028	0.076	0.222	226	364	530	1.84	2.96	4.31
EplTR035-239	B	0.762	0.948	0.984	6982	12441	20863	29.21	52.05	87.29
EplTR036-240	B	0.597	0.815	0.950	1367	4190	11740	5.70	17.46	48.92
EplTR037-239	B	0.821	0.973	0.996	5634	8750	14369	23.57	36.61	60.12
EplTR038-17	D	1.000	1.000	1.000	593	902	2156	34.88	53.06	126.82
EplTR039-118	DB	0.431	0.431	0.431	235	235	235	1.99	1.99	1.99
EplTR040-113	D	0.755	0.800	0.800	682	682	731	6.04	6.04	6.47
EplTR041-399	DB	0.559	0.679	0.701	3694	5626	5626	9.26	14.10	14.10
EplTR042-225	B	0.535	0.761	0.911	1201	3197	8842	5.34	14.21	39.30
EplTR043-266	B	0.827	0.977	0.977	2012	8551	14022	7.56	32.15	52.71
EplTR044-210	B	0.934	0.994	0.994	3566	30799	30799	16.98	146.66	146.66
EplTR045-64	NS	0.960	0.960	0.960	765	765	765	11.95	11.95	11.95
EplTR046-136	B	0.998	1.000	1.000	13634	13634	13634	100.25	100.25	100.25
EplTR047-95	NS	0.747	0.747	0.747	245	245	245	2.58	2.58	2.58
EplTR049-367	B	0.973	0.973	0.973	17483	17483	17483	47.64	47.64	47.64
EplTR050-124	DB	0.283	0.384	0.414	251	251	334	2.02	2.02	2.69
EplTR051-111	NS	0.966	0.980	0.983	879	987	1043	7.92	8.89	9.40
EplTR052-54	NS	0.899	0.913	0.935	352	354	564	6.52	6.56	10.44
EplTR054-269	B	0.915	1.000	1.000	3016	6953	6953	11.21	25.85	25.85
EplTR055-64	NS	0.395	0.395	0.404	193	193	193	3.02	3.02	3.02
EplTR056-209	B	0.900	0.994	1.000	3325	19814	19814	15.91	94.80	94.80
EplTR057-231	NS	0.240	0.273	0.303	549	549	549	2.38	2.38	2.38
EplTR058-30	NS	1.000	1.000	1.000	1625	2573	7843	54.17	85.77	261.43
EplTR059-98	NS	0.747	0.829	0.837	641	641	641	6.54	6.54	6.54
EplTR060-33	NS	0.994	0.994	0.997	903	998	1635	27.36	30.24	49.55
EplTR061-99	NS	0.910	0.944	0.944	507	848	1585	5.12	8.57	16.01
EplTR062-21	NS	1.000	1.000	1.000	427	427	589	20.33	20.33	28.05
EplTR063-239	B	0.455	0.767	0.983	1566	5068	8871	6.55	21.21	37.12
EplTR064-226	B	0.879	0.975	0.980	5417	22323	22323	23.97	98.77	98.77

EplTR065-233	B	0.744	0.995	1.000	3224	10790	10790	13.84	46.31	46.31
EplTR066-252	B	0.248	0.305	0.534	748	1171	1769	2.97	4.65	7.02
EplTR067-27	NS	0.996	0.996	0.996	112	112	112	4.15	4.15	4.15
EplTR068-69	D	0.994	0.994	0.994	994	994	994	14.41	14.41	14.41
EplTR069-87	NS	0.962	0.962	0.972	829	829	829	9.53	9.53	9.53
EplTR070-21	NS	1.000	1.000	1.000	93	93	138	4.43	4.43	6.57
EplTR071-180	B	1.000	1.000	1.000	11244	11244	11244	62.47	62.47	62.47
EplTR072-98	D	0.305	0.305	0.305	503	503	503	5.13	5.13	5.13
EplTR073-297	B	0.938	0.975	0.988	13889	18833	18833	46.76	63.41	63.41
EplTR074-241	B	0.501	0.802	0.959	1759	3746	5827	7.30	15.54	24.18
EplTR075-136	NS	0.751	0.760	0.760	5988	5988	5988	44.03	44.03	44.03
EplTR076-139	B	0.629	0.629	0.629	6152	6152	9125	44.26	44.26	65.65
EplTR077-47	B	1.000	1.000	1.000	7900	12685	12685	168.09	269.89	269.89
EplTR079-166	B	0.991	1.000	1.000	8558	8558	8558	51.55	51.55	51.55
EplTR081-65	NS	0.852	0.863	0.970	810	810	810	12.46	12.46	12.46
EplTR082-91	NS	0.263	0.263	0.431	258	258	258	2.84	2.84	2.84
EplTR083-238	B	0.451	0.806	0.978	1829	5658	11973	7.68	23.77	50.31
EplTR084-30	NS	0.996	0.996	1.000	267	403	789	8.90	13.43	26.30
EplTR085-81	NS	0.967	0.967	0.967	609	609	609	7.52	7.52	7.52
EplTR086-82	B	1.000	1.000	1.000	1431	1431	1431	17.45	17.45	17.45
EplTR088-238	B	0.221	0.907	0.958	1337	4950	14427	5.62	20.80	60.62
EplTR089-226	B	0.312	0.655	0.884	519	1919	1919	2.30	8.49	8.49
EplTR090-40	B	0.933	0.933	0.981	519	590	1059	12.98	14.75	26.48
EplTR092-24	NS	0.509	0.509	0.583	239	239	239	9.96	9.96	9.96
EplTR094-36	NS	0.923	0.923	0.949	238	238	309	6.61	6.61	8.58
EplTR095-36	D	1.000	1.000	1.000	111	111	111	3.08	3.08	3.08
EplTR097-51	NS	1.000	1.000	1.000	224	224	224	4.39	4.39	4.39
EplTR098-33	NS	1.000	1.000	1.000	66	66	66	2.00	2.00	2.00
EplTR099-33	NS	1.000	1.000	1.000	264	264	264	8.00	8.00	8.00
EplTR100-32	NS	0.961	0.961	0.961	140	140	140	4.38	4.38	4.38
EplTR101-18	NS	1.000	1.000	1.000	504	504	504	28.00	28.00	28.00
EplTR102-110	NS	0.648	0.648	0.648	286	286	286	2.60	2.60	2.60
EplTR106-323	B	0.830	0.926	0.941	3143	14704	14704	9.73	45.52	45.52

EplTR107-33	DB	1.000	1.000	1.000	177	177	177	5.36	5.36	5.36
EplTR109-15	DB	1.000	1.000	1.000	99	99	99	6.60	6.60	6.60
EplTR110-11	NS	1.000	1.000	1.000	83	83	83	7.55	7.55	7.55
EplTR112-11	B	1.000	1.000	1.000	405	405	794	36.82	36.82	72.18

Kimura divergence on MinION reads

Even though MinION sequencing shows a high error rate, our intention was to evaluate whether Kimura divergence (Kdiv) changes with array length. For this purpose, in each TR family, we calculated the mean Kdiv in arrays showing 1 RU, 2 RU, 3 RU, etc., polymerization degree in the MinION reads. These Kdiv estimates on minion reads (mean= 15.7%, SD= 3.9) were about 50% higher than those previously obtained in Illumina reads (mean= 10.7%, SD=5.5), indicating that sequencing errors logically inflate them (Wilcoxon matched pairs test: $p< 0.000001$). However, Kdiv estimates by both methods were highly correlated ($r_S= 0.71$, $p< 0.000001$) suggesting that sequencing errors impacted randomly on the 93 TR families. Anyway, our purpose was not comparing between TR families but analyzing whether, within each TR, higher degrees of polymerization showed lower Kdiv values as a consequence of the amplification process leading to higher homogenized states, expected for well-constituted satDNAs. We analyzed this issue by means of Spearman correlation analysis, at TR family level, between the Kdiv observed for each NRU value. Interestingly, 54 families showed negative correlation between NRU per array and Kdiv (significant for 31 families) and 39 showed positive correlation between them (significant in 19 families) (see Table S1.4). As a whole, the 54 TR families where the correlation was negative displayed higher maxRU (mean= 37 RUs, SD= 59) and maxAL values (mean= 5,541 nt, SD= 7,508) than the 39 where the correlation was positive (mean= 16 RUs, SD= 25, and mean= 1,937 nt, SD= 4,635, respectively) (Mann-Whitney test: z= 2.62, p= 0.0087, and z= 4.84, p= 0.000001, respectively). This separates the 93 TR families into two classes on the basis of maxRU and maxAL, one reaching higher values for these two parameters and showing negative correlation with Kdiv, and the other showing the reverse pattern, with only the first class fitting the expectation of TR homogenization after amplification. This is reinforced by the fact that the first group (N= 54) also showed longer repeat units (RUL mean: 189 nt versus 91 nt) (z= 4.64, p= 0.000003), higher tandem structure (TSI mean: 0.72 versus 0.43) (z= 4.77, p= 0.000002) and higher frequency of FISH signal (80% versus 36%) (z= 3.59, p=

0.00034). These differences indicate that well polymerized satDNAs in *E. plorans* are included in the 54 TRs that showed negative correlation between the NRU in an array and the Kdiv observed between arrays.

Congruence between cytogenetic and molecular properties of TRs

SatDNA has classically been defined on the basis of its array structure and conspicuous bands on chromosomes after FISH analysis (López-Flores and Garrido-Ramos, 2012; Garrido-Ramos, 2017). This molecular and cytological definition was valid until the high-throughput search for tandem repeats facilitated by the application of Next Generation Sequencing technologies, for instance, through the RepeatExplorer (RE) software developed by Novak et al. (2013). Intensive application of RE rounds with sequential filtering, led us to develop the satMiner toolkit which allowed finding TRs showing extremely low abundance in the genome of the migratory locust and failing to show FISH bands (Ruiz-Ruano et al., 2016). Here we check the molecular nature of these no-signal TRs using the 93 TR families selected above among the 112 families found in the grasshopper *E. plorans*.

For this purpose, we designed a spreadsheet to search for those maxRU values, found in MinION reads for every TR family, which maximize the congruence (FIT) between the presence (B, D and DB patterns) or absence (NS pattern) of FISH signal and whether maxAL surpassed or not, respectively, the minimum target size (minTS) for visualizing a FISH signal (Dataset 1.5). We thus generated a collection of spreadsheets including calculations considering measurement of the former parameters under different minID values ("3 nt", "1 RU" and "read"). This is justified by the fact that TRs are prone to replication slippage leading to short insertions or deletions, and that a given FISH band may contain up to 3 Mb of lineal DNA (Schwartzacher and Heslop-Harrison 2000), which is higher than all MinION reads lengths obtained here (Fig. S1.1).

In each spreadsheet, we first sorted the 93 TR families in decreasing order of maxRU values, and then we sequentially split them into two groups in a stepwise manner. In the first step, we set the threshold between 2 RU/3 RU to compare FIT between the two groups established (2 RUs= lower group; 3 RU-upward= higher group) and calculated their difference (DIF) in FIT values. In subsequent steps, we set the 3 RU/4 RU, 4 RU/5 RU, …, (N-1) RU/N RU thresholds (up to N RU= 30) and performed the same calculations.

Using this spreadsheet collection, we first tried to find which minTS values satisfied

the expectations above, covering minTS values between 500 and 3,000 bp at 100 bp intervals. For each minTS value, we recorded the maximum values for DIF (among all NRU classes, from 3 RU to 30 RU) at "3 nt", "1 RU" and "read" levels, and then compared between them to search for the minTS value which maximized DIF. This revealed that the "3 nt" level of analysis was the most discriminant by showing a peak at 1,700 nt which slightly decreased from there onwards (see Table S1.5 and Fig. S1.4). We interpret these results as an indication that the minimum target size for FISH visualization, in this collection of TRs, was 1,700 bp, which was close to the 1,000 bp anticipated by Schwarzacher and Heslop-Harrison (2000). On this basis, all subsequent calculations were performed at 1,700 nt as minTS.

To ascertain which is the minimum polymerization degree that TRs from which we find higher levels of consistency between their molecular and cytological properties, we searched for a polymerization degree (NRU= 3 RU-20 RU at 1 RU intervals, and 20 RU-30 RU bp at 5 RU intervals) from which onwards DIF and FIT reached their maximum values. This revealed that TRs that reached the 7 RU state showed the maximum values for FIT at any maxID considered and maximum DIF was reached at 6 RUs at "3 nt" and "1 RU" levels and 7 RUs with the read level. As DIF values were quite similar between 6 and 7 RUs states we considered the threshold of 7 RUs for subsequent analysis being more conservative (see Fig 1.3a). Interestingly, the set of TRs in *E. plorans* showed statistical difference in terms of PI (W = 599, p = 0.0004), maxAL (W = 260.5, p = 6.991e-10) and FISH pattern (KWa chi-squared = 8.0861, df = 1, p = 0.004461, "maxRU>=7" = 32 banded out of 54 TRs, "maxRU<7" = 10 banded out of 40 TRs) considering the threshold of 7 RUs.

To investigate whether the former conclusion is valid for all kinds of TRs, we divided them into two groups showing values above or below the median value for RUL (long and short groups) or AT content (AT_{poor} and AT_{rich} groups). This showed that the 6-7 RUs threshold is valid for all four groups of TRs established (Fig 1.3c). The same approach showed similar results considering $RUL_{shorter}$ and RUL_{longer} TR families (Fig 1.3b).

This 7 RUs threshold for high congruences between FISH and molecular features of TRs could be applied also to the *L. migratoria* satellites found by Ruiz-Ruano et al. (2016) (see Table S1.12). We found significance differences in terms of PI (W= 504, p= 0.0005878, mean "maxRU>=7"= 0.938, mean "maxRU<7"= 0.458), maxAL (W= 548, p= 3.688e-05, mean "maxRU>=7"= 3955.025, mean "maxRU<7"= 651.813) and FISH pattern (KW chi-squared= 7.1587, df= 2, p= 0.02789, "maxRU>=7"= 5 nc/na pattern out of 40 families, "maxRU<7"= 7

nc/na pattern out of 16 families) between “maxRU>=7” and “maxRU<7” TR families of this species. We also explored the effect of this threshold in the case of the human satellites found in UCSC (see Table S1.13) and we distinguished a group of 17 TRs with “maxRU>=7” and 3 elements that that showed maximum number of repeats in arrays below 7. These two groups showed significant differences in respect to PI (W= 54, p= 0.007657, mean “maxRU>=7”= 0.948, mean “maxRU<7”= 0.071) but not to maxAL (W= 35, p= 0.4707, mean “maxRU>=7”= 7,087.778, mean “maxRU<7”= 3,324.667). Furthermore, the 3 TR families (HSAT1, REP522 and SST1; UCSC names) are known to have a long RUL (from 568 to 1,800 nt) what could explain the lack of differences in terms of maxAL although they are moderately repetitive (Epstein et al., 1987); also several authors have identified homology between these TR families and transposons (Fromer et al., 1984; Röschenthaler et al., 1992; Brown et al., 1990).

Congruence between FISH pattern and maxAL considering the minTS of 1,700 nt in the set of TR families included here reach a success rate of 73.1% at 3 nt maxID, 78.5% at 1 RU and 80.6% considering the whole read. This increasing in the success rate when increasing maxID is the result of B/DB/D TR families that do not reach the maxAL expected for yielding FISH signal (i.e. 1,700 nt) but they do so when collapsing close array at higher maxIDs thus making congruence between maxAL and FISH pattern (e.g. EplTR010-202, EplTR066-252 or EplTR089-226) and confirming that FISH resolution is below the maxIDs considering here as suggested by Schwarzacher and Heslop-Harrison (2000). We found some TRs that yielded FISH signal even though all arrays found were shorter than 1,700 nt considering the read as maxID. However, most of their maxAL values were near the minTS established so the low coverage we have in MinION library could hide longer arrays (e.g. EplTR019-75, EplTR086-82 or EplTR090-40). On the other hand, some NS families did not display FISH bands despite showing arrays longer the minTS threshold (e.g. EplTR016-31, EplTR058-30 or EplTR075-136). In this case, the only explanation is that the FISH technique failed or there were differences in condensation between chromosome regions reducing accessibility of target DNA increasing the 1,700 nt threshold for detection of FISH signal (Speel et al., 2006; Wang et al., 2006).

Taken together, these results show high congruence between cytological and molecular data in the satellitome of *E. plorans*, specially for TRs with high PI, maxAL and maxRUs.

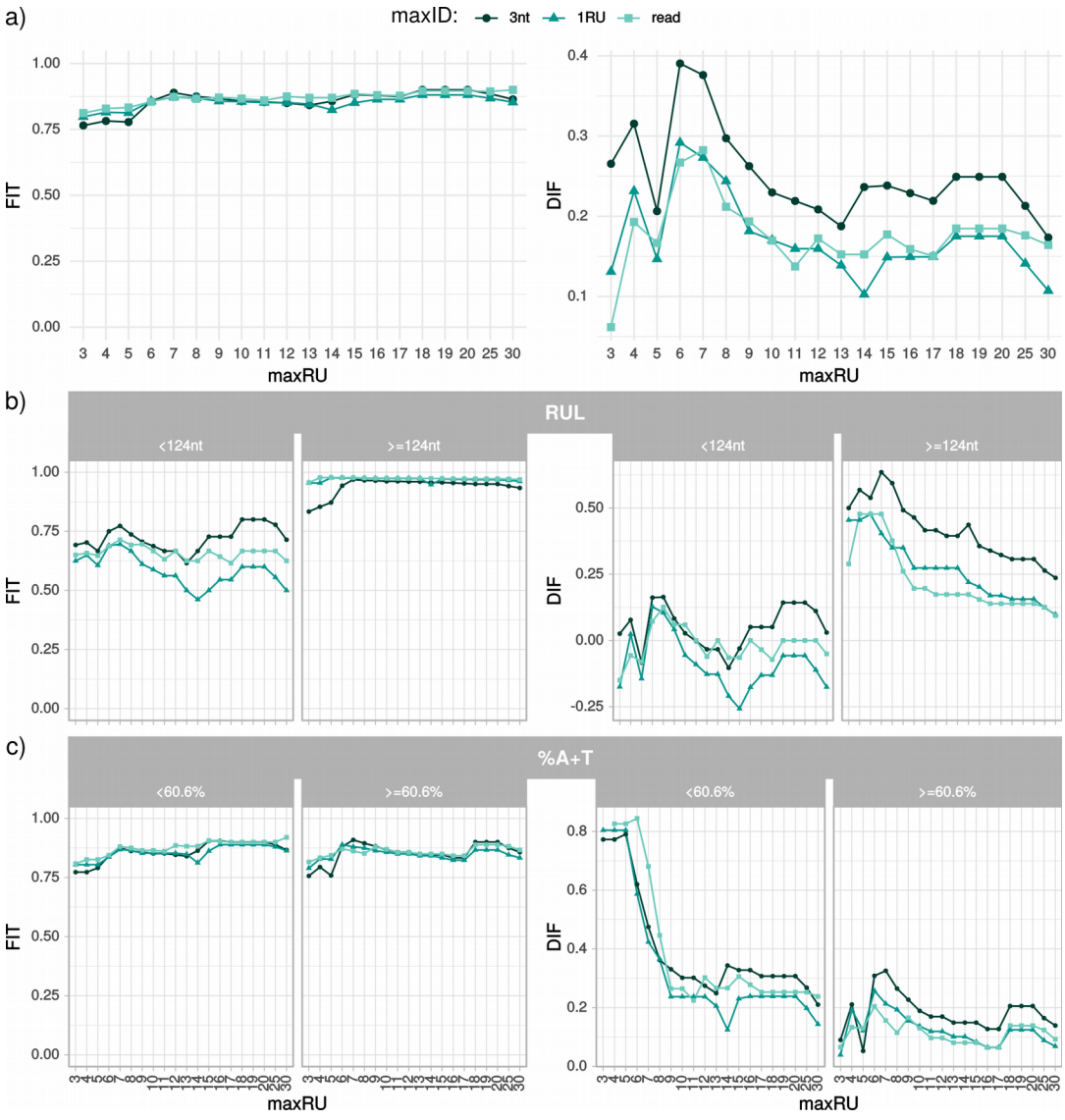


Figure 1.3. Results of FIT and DIF with maxAL and FISH pattern (considering 1,700 nt as the minimum array length detected by FISH). a) Results including the 93 TR families found in MinION reads, FIT (left) grows up from 6-7 maxRU especially at 3 nt of maxID, the same occurs for DIF (right). b) FIT and DIF splitting TR families by the median RUL, TRs with RUL>=124nt performed better in FIT (left) and DIF (right). c) FIT and DIF according to median A+T% value, TRs families showed similar FIT (left) independently of A+T% and the same is true for DIF (right) with the exceptions of some unexpected high values at low maxRU thresholds (i.e. 3, 4 and 5) when retaining TR families with AT content below 60.6%. These high DIF values are due to the existence of very few TR families with AT content <60.6% and low maxRU, most of them failing to reach FISH expectations according to maxRU.

Several TRs found in the *E. plorans* genome are linked (or belong to) TEs

To investigate the existence of possible relationships between each TR family and other repetitive elements, we annotated the MinION reads against the most complete repetitive database available in *E. plorans* (Chapter 2) and scored i) the number of reads shared by each TR family and other repetitive elements, ii) the number of TEs sharing more than 50% of TR reads, and iii) the total number of repetitive elements sharing reads with each TR family (Tables S1.1 and S1.6). We interpreted that the second parameter indicated physical association between a TR and a given repetitive element, as more than 50% of TR reads were shared with it. Interestingly, we found many TR families sharing reads with TEs. Moreover, the number of TEs sharing >50% of their reads with a given TR family was negatively correlated with its maxRU (r_S= -0.40, df= 91, p= 0.00007) and TSI (r_S= -0.51, df= 91, p< 0.000001) values, suggesting that TRs showing lower polymerization degree and tandem structure are more likely to be associated with TEs. By contrast, we found no significant correlation between parameter iii and maxRU (r_S= -0.03, df= 91, p= 0.75) or TSI (r_S= -0.14, df= 91, p= 0.19), indicating that sharing more than 50% of a TE reads with a TR is indicative of TR-TE linkage.

To try to understand the kind of relationship between TRs and TEs, in the case of TR families sharing more than 50% of its MinION reads with one or more TEs, we scored the physical distance (in nt) between TR and TE arrays (Table 1.3 and S1.6). Following Ahmed and Liang (2012), we considered that a TR derived from a TE if sequence similarity was >55% or else their positions on MinION reads overlapped or were contiguous (see Fig. S1.5a). In addition, we also considered TR-TE association when they tended to show roughly the same distance on many different positions (Fig. 1.5b) or homology to the sequence of a given TE (Fig. 1.5c). Exceptionally, EplTR082-91 showed constant distance in respect to one TE (Gypsy) and variable distance in respect to another (hAT) (Fig. S1.5d). In total, we found that 37 TRs were physically associated with 38 TEs, as EplTR075-136 was homologous to two different families of Polinton TEs (see Fig. S1.5e). The TEs most frequently containing TRs were Polinton (13 TRs) and Gypsy (9), followed by SINEs (6), CR1 (3), Helitron (3), Penelope (2), Jockey (1) and Kolobok (1). Therefore, 17 were DNA transposons and 21 were RNA ones. Remarkably, TRs yielding no FISH signal (NS pattern) were the most represented, with 22 families among the 37 showing TE association, followed by dotted TRs (9 D families) and dotted-banded families (3 DB) or banded (3 B). Interestingly, we found that the total of D TR families showed relationships with TEs.

As additional evidences and validations for TR-TE association, when we found homology between TRs and associated TEs, we performed TR clustering with RE software using short reads homologous to the TE, and PCR with primers pairs anchored to amplify the putative junction of both kind of elements (primers in Table S1.10). We considered this as an additional proof of relation between the TR and the TE. 28 TR families out of the 37 showing physical association with TEs (criteria from Ahmed and Liang, 2012) were validated by RE clustering or TE-TR amplification by PCR (Table 1.3). Again, we found a high number of NS or D TR families (12 and 8 respectively) showing validated TE association whereas DB or B families were less abundant (3 and 2 families respectively).

The association between TRs and TEs in the genome of *E. plorans* put in evidence a possible role of TEs as sources of TRs and satDNAs. This hypothesis in *E. plorans* is supported by the existence of 16 TR families, out of the 37 linked to TEs, that reached their maximum degree of polymerization (maxAL and maxRU) in reads not-containing the TE. Then, these TRs would have got genomic independence outside the TE as a first step to behave as true satDNAs (e.g. EplTR017-67 and EplTR030-267). On the other hand, some of the 21 remaining TR families, with maximum polymerization found in reads shared with the corresponding linked TE, showed also long arrays with high number of NRU (e.g. EplTR021-165 or EplTR090-40) when associated to TEs, thus suggesting that TR polymerization could begin at some locations inside or near the TE. Furthermore, we found that, in average, maxAL and maxRU of TRs when sharing reads with TE is quite short (196 nt and 2.7 RUs respectively). This put forward the function of some TE regions as seeds for TR formation in the genome of *E. plorans* by means of amplification of short repetitive (i.e. dimers or trimers) unit located on their sequence or in the surrounding regions.

Table 1.3. TR families sharing >50% of their reads with transposable elements (TEs). When a TR shares >50% of reads with a TE, distance between both is conserved (Mode), they showed homology (TR-TE Hom) and we were able to cluster the TE using TR-homologous reads (RE clust. of TE) and/or amplified a tandem repeat PCR pattern (PCR) anchoring primers to the TE, that TR is considerer as related to TE. We included in this table only pairs of TE-TRs showing homology (for full table see Table S1.6) and for non-homologous pairs we included all TEs sharing at least a 50% of TR reads.

ID satMiner	FISH	TE sharing reads			TR in shared reads				Distance between TR-TE			TR-TE Hom.	Overl. >RU	RE clust. of TE	PCR	TE assoc.	TE assoc. val.
		N	N >50% reads	TE	N reads	% reads	AL av.	RU av.	Mode	Mode Freq.	Mode Prop.						
EplTR004-196	D	380	1	SINE_02	689	0.68	111.43	0.57	0	301	0.486	Yes	Yes	Yes	Ladder	Yes	Yes
EplTR011-174	NS	368	1	SINE_02	561	0.50	278.00	1.60	3	3	0.006	Yes	Yes	No	Single	Yes	--
EplTR017-67	D	250	3	Polinton_04	262	0.53	195.08	2.91	0	11	0.043	Yes	Yes	Yes	Single	Yes	Yes
EplTR018-75	D	374	1	Gypsy_038	609	0.58	176.85	2.36	1	109	0.179	Yes	--	Yes	Single	Yes	Yes
EplTR021-165	D	208	1	Helitron_03	124	0.67	295.19	1.79	0	13	0.105	Yes	Yes	Yes	Ladder	Yes	Yes
EplTR030-267	B	143	1	CR1_2	57	0.56	273.68	1.03	20	6	0.107	Yes	--	Yes	Dimer	Yes	Yes
EplTR031-324	B	23	1	SINE_04	12	0.63	365.93	1.13	0	11	0.917	No	Yes	--	--	Yes	--
EplTR033-158	NS	163	2	Helitron_09	56	0.57	148.75	0.94	0	5	0.091	Yes	Yes	Yes	Ladder	Yes	Yes
EplTR034-123	D	240	1	Penelope_01	226	0.80	109.22	0.89	0	45	0.218	Yes	Yes	Yes	Ladder	Yes	Yes
EplTR038-17	D	141	4	Polinton_07	63	0.54	214.64	12.63	38	2	0.032	Yes	Yes	Yes	Dimer	Yes	Yes
EplTR039-118	B	172	1	Helitron_03	87	0.84	148.53	1.26	-77	8	0.092	Yes	Yes	Yes	Dimer	Yes	Yes
EplTR040-113	D	22	2	Jockey_1	9	0.69	188.30	1.67	0	4	0.444	Yes	Yes	Yes	Ladder	Yes	Yes
EplTR041-399	DB	136	1	SINE_06	83	0.94	547.30	1.37	0	58	0.795	Yes	Yes	Yes	Ladder	Yes	Yes
EplTR045-64	NS	212	2	Polinton_07	234	0.85	226.77	3.54	29	8	0.035	Yes	Yes	Yes	Ladder	Yes	Yes
EplTR047-95	NS	5	5	RTE_01	2	1.00	164.00	1.73	995.5	--	--	No	--	--	--	--	--

				hAT_01	1	0.50	245.00	2.58	--	--	--	No	--	--	--	--	--
				Helitron_02	1	0.50	245.00	2.58	--	--	--	No	--	--	--	--	--
				Penelope_01	1	0.50	245.00	2.58	--	--	--	No	--	--	--	--	--
				Helitron_01	1	0.50	245.00	2.58	--	--	--	No	--	--	--	--	--
EplTR051-111	NS	166	1	Gypsy_003	99	0.66	308.80	2.78	1540	3	0.031	Yes	--	No	Single	Yes	--
EplTR052-54	NS	177	1	Gypsy_003	76	0.52	120.98	2.24	950	2	0.026	Yes	--	No	Single	Yes	--
EplTR055-64	NS	243	1	Penelope_07	115	0.53	81.23	1.27	11	5	0.043	Yes	Yes	Yes	Dimer	Yes	Yes
EplTR057-231	NS	95	1	CR1_1	38	0.86	241.41	3.01	353	3	0.079	Yes	--	No	Dimer	Yes	Yes
EplTR059-98	NS	196	1	Polinton_12	126	0.58	222.24	2.27	81	5	0.040	Yes	--	No	Ladder	Yes	Yes
EplTR060-33	NS	272	1	Polinton_04	231	0.50	291.28	8.83	-11	11	0.048	Yes	Yes	Yes	Single	Yes	Yes
EplTR061-99	NS	27	1	Gypsy_001	9	1.00	253.27	2.56	0	3	0.333	Yes	Yes	Yes	Dimer	Yes	Yes
EplTR062-21	NS	98	2	Polinton_03	37	0.73	132.47	6.31	18	2	0.067	Yes	Yes	Yes	Single	Yes	Yes
EplTR066-252	B	48	1	CR1_2	11	0.58	200.36	0.80	--	--	--	No	--	--	--	--	--
EplTR067-27	NS	186	2	Polinton_01	176	0.98	56.99	2.11	53	20	0.114	Yes	--	Yes	Ladder	Yes	Yes
EplTR068-69	D	171	1	Gypsy_003	93	0.57	321.06	4.65	2107	2	0.022	Yes	--	No	Single	Yes	--
EplTR069-87	NS	56	3	Polinton_05	19	0.59	344.13	3.96	35	2	0.105	Yes	Yes	Yes	Dimer	Yes	Yes
EplTR072-98	D	143	1	SINE_02	39	0.61	86.33	0.88	0	6	0.176	Yes	Yes	No	Ladder	Yes	Yes
EplTR075-136	NS	72	2	Polinton_08	27	0.69	153.61	1.13	1527	2	0.074	Yes	--	No	Single	Yes	--
				Polinton_01	24	0.62	140.63	1.03	855	2	0.083	Yes	--	No	Single	Yes	--
EplTR076-139	B	118	2	hAT_02	28	0.56	117.82	0.85	--	--	--	No	--	--	--	--	--
				SINE_02	27	0.54	117.78	0.85	--	--	--	No	--	--	--	--	--
EplTR081-65	NS	128	1	Polinton_12	80	0.81	156.22	2.40	140	3	0.04	Yes	--	Yes	Single	Yes	Yes

EplTR082-91	NS	112	4	Gypsy_059	34	0.87	102.16	1.12	392	2	0.06	Yes	--	No	--	Yes	--
EplTR084-30	NS	138	1	Jockey_1	41	0.51	165.50	5.52	--	--	--	No	Yes	--	--	Yes	--
EplTR085-81	NS	88	3	Polinton_05	26	0.54	262.85	3.25	0	5	0.19	Yes	Yes	Yes	Ladder	Yes	Yes
EplTR090-40	B	34	2	Kolobok_01	12	0.75	191.00	4.78	0	5	0.45	Yes	Yes	Yes	Ladder	Yes	Yes
EplTR092-24	NS	52	5	CR1_4	22	0.88	35.46	1.48	0	13	0.59	Yes	Yes	Yes	Single	Yes	Yes
EplTR094-36	NS	53	1	SINE_01	9	0.53	126.70	3.52	--	--	--	No	--	--	--	--	--
EplTR095-36	D	13	4	Gypsy_012	3	1.00	98.67	2.74	0	2	0.67	Yes	Yes	Yes	Single	Yes	Yes
EplTR097-51	NS	1	1	Gypsy_001	3	1.00	193.33	3.79	0	3	1.00	Yes	Yes	Yes	Dimer	Yes	Yes
EplTR098-33	NS	5	5	SINE_02	1	1.00	66.00	2.00	--	--	--	No	Yes	--	--	Yes	--
				hAT_01	1	1.00	66.00	2.00	--	--	--	No	--	--	--	--	--
				Nimb_3	1	1.00	66.00	2.00	--	--	--	No	--	--	--	--	--
				RTE_02	1	1.00	66.00	2.00	--	--	--	No	--	--	--	--	--
				Gypsy_008	1	1.00	66.00	2.00	--	--	--	No	--	--	--	--	--
EplTR099-33	NS	25	1	Gypsy_043	9	1.00	195.89	5.94	--	--	--	Yes	--	No	Single	Yes	--
EplTR100-32	NS	63	3	hAT_02	9	0.56	81.11	2.53	--	--	--	No	--	--	--	--	--
				RTE_01	8	0.50	73.75	2.30	--	--	--	No	--	--	--	--	--
				SINE_02	8	0.50	77.38	2.42	--	--	--	No	--	--	--	--	--
EplTR102-110	NS	25	7	Polinton_05	4	0.67	207.5	1.89	--	--	--	Yes	Yes	Yes	Ladder	Yes	Yes
EplTR107-33	NS	23	6	Jockey_1	2	0.50	154.50	4.68	--	--	--	No	--	--	--	--	--
				Gypsy_085	4	1	151.25	4.58	--	--	--	No	--	--	--	--	--
				hAT_01	3	0.75	142.67	4.32	--	--	--	No	--	--	--	--	--
				RTE_01	3	0.75	142.67	4.32	--	--	--	No	--	--	--	--	--

				Gypsy_065	3	0.75	142.67	4.32	--	--	--	No	--	--	--	--	--
				SINE_02	3	0.75	146.00	4.42	--	--	--	No	--	--	--	--	--
EplTR110-11	NS	27	8	hAT_02	3	1.00	53.67	4.88	--	--	--	No	Yes	--	--	Yes	--
				hAT_01	3	1.00	53.67	4.88	--	--	--	No	--	--	--	--	--
				CR1_1	2	0.67	58.50	5.32	--	--	--	No	--	--	--	--	--
				RTE_02	2	0.67	58.50	5.32	--	--	--	No	--	--	--	--	--
				RTE_01	2	0.67	58.50	5.32	--	--	--	No	--	--	--	--	--
				Gypsy_012	2	0.67	39.00	3.55	--	--	--	No	--	--	--	--	--
				SINE_01	2	0.67	39.00	3.55	--	--	--	No	--	--	--	--	--
				SINE_02	3	1.00	53.67	4.88	--	--	--	No	--	--	--	--	--

Molecular and cytological differences between TR families

SatDNA has been classically defined by the formation of long arrays at molecular level and the visualization of conspicuous bands after the FISH technique. Our present results allowed exploring the molecular-cytological interface through a comparative analysis between several parameters measured here. For this purpose, we generated seven variables from the cytological and molecular data: i) RUL (length of the repeat unit), ii) A+T% (content of AT in the repeat unit), iii) Kimura divergence (Kdiv) suggesting homogenization by negative correlation with array length (yes= negative correlation, no= positive correlation), iv) TE-association (yes= association, no= no association detected), v) TE-association validated through RE clustering or PCR (yes= validated, no= no validated nor detected), vi) median maxRU of 8 RUs as threshold (yes >= 8RUs, no < 8RUs) and vii) the FISH pattern of each TR (B, DB, D and NS). As a first approach, we performed a multiple correspondence analysis (MCA) to study the contribution of the molecular and cytological variables studied here to the differences found between TR families. This analysis revealed two dimensions (Dim1 and Dim2) explaining 37.2% and 13.9% of the inertia, respectively (Fig 1.4a). The categories of the variables including high RULs (above the median, i.e. 124 nt), no TE association, B or DB FISH patterns, maxRU>=8, array homogenization (Kdiv correlated negatively with array length) and low A+T% were grouped together in Dim1 and apart from the alternative categories of variables (D/NS pattern, TE association, short RULs, maxRU<8, high AT content and no array homogenization). Dim2 grouped the D pattern with association to TEs, this relation was expected since we found TE association for all D TRs (9) in the genome of *E. plorans* (Table 1.3) which supports the conclusion that TRs displaying the D pattern of FISH are most likely TEs containing TRs.

These five molecular properties (counting TE association and TE association validated as one) clearly define two different groups of TRs, one including the B and DB patterns and showing long RULs, low A+T%, rare association with TEs, high tandem structure (maxRU>=8) and high sequence homogenization. These properties have been claimed for true satDNA, with the exception of the AT content since satDNA tend to be enriched in AT. In this case, we believe that the anomalously low AT content is due to the presence of 11 TRs with AT content higher than 70% but none in all those TRs showing FISH signal.

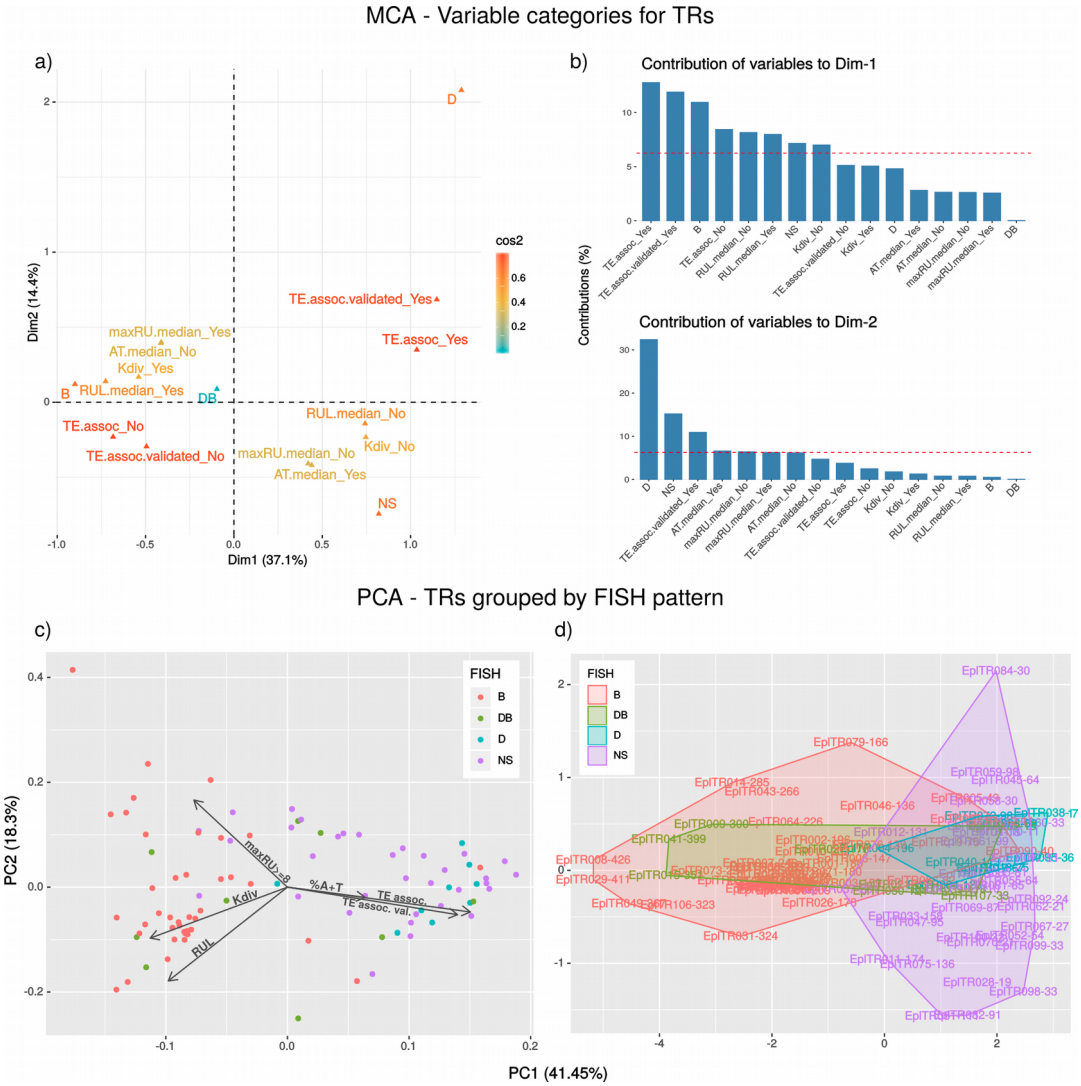


Figure 1.4. a) MCA of seven molecular and cytological properties of TRs studied here. The states of these features split into two groups in Dim1, one of them comprising the conventional features of satDNA: array homogenization, high RUL and maxRU (above 7RUs here), genomic independence (no association to TEs here) and well-defined FISH signal (B and DB patterns). b) Contribution of the variables to the difference between TR families, FISH data and TE association appear as the most contributing properties for both MCA dimensions. c) PCA analysis and d) DA analysis for the six molecular features of TRs using FISH pattern as grouping factor. Note that B and DB TRs group together thus resembling in their molecular features and it was so also for D and NS TR families.

Altogether, these properties roughly delineate satellite DNA definition, in particular the FISH pattern was, apart from TE association, the variable explaining great part of the inertia in Dim1 and Dim2 (Fig. 1.4 b) that is why it was chosen as grouping factor for PCA (principal component analysis) and DA (discriminant analysis) subsequent approaches. We explored the 93 TR families found in the MinION library of *E. plorans* in terms of the five molecular variables indicated above by a principal component analysis (PCA) and a linear discriminant analysis (DA). This analysis split again the set of 93 TR families into two differentiated groups (Fig. 1.4 c). B and DB TRs shared similar molecular features: long RULs, uncommon TE association, high polymerization degree (>8RUs), array homogenization and low A+T%. In contrast, D and NS TRs showed the alternatives categories for these variables.

The graphical representation of the DA, with these six molecular parameters as independent variables and the FISH pattern as grouping variable, reinforces the existence of these two groups of TRs based on FISH pattern as DB TRs families resembles the group of B families regarding their molecular features and so is true for D TRs with respect to NS families (Fig. 1.4 d). The exploratory discriminant analysis (DA) showed significant assignations to the four FISH groups (Wilks Lambda: 0.27318, F=7.70, p=9.743e-16, better performance than when using other variables, see Table S1.14), with TE association (Wilks Lambda= 0.54, p= 8.869e-12), RUL (Wilks Lambda= 0.35, p= 9.695e-09) and maxRU>=8 (Wilks Lambda= 0.31, p= 3.713e-02) as the only parameters showing significant contribution. On the basis of the classification of TRs according to FISH pattern, DA correctly classified 75% of cases, with high success for B and NS patterns (95% and 79% respectively) but lower for D and DB ones (44% and 0%, respectively). Remarkably, all nine DB TRs were misassigned to B (5) or NS (4) patterns and five D TRs were classified as NS pattern while other four NS TRs were predicted as being D families, thus reinforcing the conclusion that D-NS and B-DB TRs probably represent TR of different nature, the latter showing satDNA properties although some DB (predicted as NS) could behave as incipient satDNAs retaining properties of simple tandem repeats. Taken together all present results, and bearing in mind the logical limitations of our present data in respect to the material and techniques performed, we believe that the most reasonable way to define satDNA is through FISH banding.

Interestingly, some of the misassigned TRs in terms of “FISH” could represent intermediate steps in a possible evolutionary dynamics of satDNAs. For example, the

EplTR021-165 (DB pattern) or the EplTR090-40 (B pattern) could represent a satDNAs emerging from a tandem repeat located in a TE but lacking the homogenization claimed for mature satDNAs (Table S1.7). Therefore, although there are differences between families in terms of molecular and cytological properties, the discussion about they classification as satDNAs or TRs could be empty as they are both part of the same evolutionary dynamics.

These results confirm the high resemblance between B-DB TR families and D-NS the former one comprising the molecular and cytological properties of satDNA such as longer repeat units, high degree of tandem repetitions, array homogenization and genomic independence (no association with other elements, i.e TEs). Therefore, it is possible to effectively distinguish well-defined satDNA from simply tandem repeats in the pool of families identified in the genome of *E. plorans* and, moreover, some of these simply tandem repeats could be identified as seeds representing intermediate steps in the evolution of satDNAs.

Discussion

Satellite DNA has been widely studied from the perspective and methodological approaches of several genetic disciplines. This wide range of views gave rise to ambiguous definitions for these repetitive genomic elements. SatDNA sequences can vary in their location and cytological appearance, they have been found making up the heterochromatin part of genomes as historically declared (Plohl et al., 2012; Garrido-Ramos et at., 2017) but they are also located in the euchromatin (Brajković et al., 2012; Pavlek et al., 2015; Pita et al., 2017; Sproul et al., 2020). The only consensus about the essential requirement for a satDNA is its tandemly repeat structure but it would be nameless if every tandem repeat in the genome, even dimers, were considered as satDNA. This need for a way of differentiation between satDNA and TRs was also claimed by Silva et al. (2019) in *Drosophila* where authors distinguish between satDNA and dispersed tandem repeats regarding abundance and clustering structure in RepeatExplorer.

As a result of its tandemly nature, satDNA is usually detected by FISH as an intense band in chromosomes so it has been useful as molecular marker of chromosomes (Mestriner et al., 2000; Baumgartner et al., 2006; Cabrero et al., 2017, Kroupin et al., 2019)

although also other FISH patterns have been described (Palacios-Gimenez et al., 2017; Utsunomia et al., 2019). Trying to fill the gap between the molecular and cytological levels, in respect to how the molecular properties (RUL, %AT, homogenization, degree of polymerization and TE association) of TRs translate into cytological properties (their visibility by FISH), we characterized the satellitome of the grasshopper *E. plorans* through Illumina sequencing, long MinION reads and FISH.

Following the satMiner protocol (Ruiz-Ruano et al., 2016) in two Illumina libraries of *E. plorans*, one individual 0B and one 4B, we identified 112 TR families in the genome. Then, we performed FISH for the 112 TR families that were classified according to their signal as suggested in Ruiz-Ruano et al. (2018b). Finally, we searched for these TR families in the MinION long reads belonging to the same 4B individual used in the Illumina sequencing and analyzed their degree of polymerization. Then, we discarded TRs missed in long reads and all those showing maxRUs equal or lower than 1.5, thus we retained 93 TR families. We then considered different thresholds of inter-array distance (3 nt, 1 RU and the read) to include two arrays as part of the same one to explore polymerization.

The abundance of the 112 TR families in the genome of *E. plorans* (at least 8.38% of 0B male genomes) and the remaining 93 TR families validated in MinION reads (8.17% in 0B males) was higher than that found in *L. migratoria* (2.39% of the Southern genome and 2.74% of the Northern one; Ruiz-Ruano et al., 2016) and thus consistent with the heterochromatin characterization of both species (Camacho et al., 1991) showing that the genome of *E. plorans* reacts more intensely to C banding technique than the one of *L. migratoria*. In addition, abundance estimations of each TR family in Illumina reads and MinION long reads was verified by a strong positive correlation between them.

Interestingly, this abundance was positively correlated with FISH pattern. TR families showing well-defined FISH bands were more abundant and families being scarce in the genome failed to show FISH signal. However, some TR families showed a high abundance but they did not yield FISH signal (EplTR011, EplTR024 and EplTR026 as examples) and, the contrary was also true, some TR families yielding an intense FISH signal were poorly represented in the genome of *E. plorans* (e.g. EplTR050, EplTR096 and EplTR101). To unveil the molecular structure behind the FISH performance of these TR families, we analyzed their polymerization degree in long MinION reads. We scored the arrays of each family by the number of repetition they carry and we realized that many of the TR

families show several monomer and dimers that were located distantly enough (more than one long read) to be considered as possible dissemination seeds for arising of new TR arrays as proposed by Ruiz-Ruano et al. (2016) in *L. migratoria;* Rodrigues et al. (2019) in some Characidae fishes and according to the library hypothesis of Salser et al. (1976). In addition, FISH patterns showed significant differences regarding the maximum array length for each TR family, as expected, TR families arranged in longer arrays give banded FISH patterns although there are also exceptions that could mark differences between the nature of TR families.

To better understand these liaisons between the molecular nature and cytogenetics of tandem repeats we designed a set of spreadsheets to find the number of repeats from which TRs show the maximum congruence between the maximum array length and the FISH signal so that the set of TRs in *E. plorans* behave as expected in terms of molecular and cytological properties.

The data of the 93 TR families in *E. plorans* behave consistently with respect to the FISH signal and the maxAL found in MinION reads, except for a few exceptions (see Results). In addition, the greatest difference between the congruence of cytogenetic and molecular variables occurred around maxRU=7. This threshold divides the set of TR families into two groups, the upper one (maxRU>=7) showing higher PIs, obvious FISH signals and longer maxALs (as expected for real satDNA). This maxRU threshold was close to the median maxRU (i.e. 8 RU) but for DA analysis we chose the latter value to avoid dependence between maxRU and the FISH pattern/molecular variables, and being also more astringent to find differences between TR groups.

Although most of the "maxRU<7" families showed a NS FISH pattern (18 out of 33 TR families) and TRs showing maxRU>=7 yield a B FISH signal (32 out of 42 TR families), there were some satDNA or TR families that should not be detected as banded (B) through FISH considering 1,700 nt as the minimum array size that can be visualized by FISH (Fig. S1.4). However that could be possible as the signal of nearby arrays could be added when considering higher maxIDs such as the whole read. In fact, several authors have indicated that the resolution of FISH in mitotic chromosomes could be at distances in the range of Mb (Schwartzacher and Heslop-Harrison 2000; Cheng et al., 2002). This explains, for instance, the FISH pattern of EplTR031-324, EplTR042-225 or EplTR063-239 while the lack of coverage could explain that of EplTR107-33, EplTR109-15 or EplTR112-11. On the opposite case, there are three TR families showing a maximum array size

higher than 1,700 nt at read maxID that could not be detected by FISH (NS pattern) as EplTR016-31, EplTR058-30 and EplTR075-136 (the latter one surpassed the maxAL of 1,700 nt even considering 3 nt maxID). This can be explained by the differences in condensation between chromosome regions reducing accessibility of target DNA increasing the 1 kb threshold for detection of FISH signal (Speel et al., 2006; Wang et al., 2006) or simply by problems with the FISH technique.

We found that 37 TR families in *E. plorans* were related to transposable elements (TEs), most of them failing to yield FISH signal. In particular, we identified connections to DNA transposons (Helitron, Polinton and Kolobok), retrotransposons (Gypsy, CR1, Penelope and Jockey) and SINEs. Interestingly, TR families related to TEs were found in arrays with low repeat number in average (1, 2 or 3 RUs) when sharing reads with TEs. This fact could imply that several satDNA in *E. plorans* could have arisen from TEs. In fact, some of these TEs have been pointed out as sources of satellites as the case of Helitrons in *D. melanogaster* (Dias et al., 2015) or other DNA transposons (Belyayev et al., 2020), retrotransposons in plants (Sharma et al., 2013; Vondrak et al., 2020) or SINEs (Bois et al., 2001; Tu et al., 2004). However, some TR families reached their maxRUs in reads shared with the linked TEs suggesting a lack of genomic independence claimed for true satDNAs (Silva et al., 2019).

Our analysis announced differences in the molecular and cytological nature of the total set of 93 TR families in *E. plorans*. To check if this assumption was true we performed a discriminant analysis including several parameters: RUL, AT%, Kdiv, maxRU>=medianRU (8RUs), TE association and FISH pattern. As a result, TRs were split into two groups, one of them retaining the well-known properties of satDNAs (Garrido-Ramos, 2017): long RULs, lower Kdiv in longer arrays (homogenization), maxRU>=8 (high degree of polymerization), genomic independence (no association to TEs) and well-defined banded in FISH (B and DB patterns). These differences between TR families characterized using satMiner is not a byproduct of that approach as in a single run of TAREAN (Novák et al., 2017) since the EplTR081-65 showing a NS pattern was identified as a well-defined satDNA (see Table S1.11) although we found it associated with a Polinton TE and lacking homogenization (see Table S1.7). Therefore, NS satDNAs could not be an artifact of our approach to identified satellites using satMiner. In addition, satDNAs showing a NS pattern have been described also for the *Astyanax* genus of fishes (Silva et al., 2017), the grasshopper *Pyrgomorpha conica* (Ruiz-Ruano et al., 2018b) or in the moss

Physcomitrella patens (Kirov et al., 2018). Therefore, the exploration of these difference between properties of satDNA is recommended when characterizing a new satellitome.

However, in the light of our analysis NS TRs are probably simple tandem repeats, often linked to TEs, that behave as seeds that could develop to well-constituted satDNA depending on the evolutionary pathways they follow. Extraordinarily, some NS TRs could retain other properties of satDNA (e.g. EplTR016-31 or EplTR058-30, see Table S1.7) in spite of their lack of FISH signal probably due to short RULs that needs more polymerization to be detected through FISH. Therefore, we could ultimately consider satDNA and TRs as part of the same evolutionary dynamics, thus, discussions about which term is better would be more focused on form than content.

The existence of several FISH patterns explained by different molecular properties seeds evidences of a cyclic model for the intragenomic evolution of satDNA. The case of *E. plorans* supports the presence of simple tandem repeats that are usually undetectable by FISH (NS pattern, e.g. EplTR028-19 or EplTR033-158) some of them linked to TEs (all D TRs are associated to TEs). Some of these tandem repeats could start a polymerization processed (e.g. EplTR045-64 or EplTR049-367) helped or not by TEs (where some of they could grow up above the limit for FISH detection thus becoming B or DB patterned) and reaching then a stage in which they begin the homogenization of arrays (e.g, EplTR036-240 or EplTR040-113), thus developing as well-defined satDNAs. In this way, a new conventional satDNA would birth under the synergy of high degree of polymerization and array homogenization (e.g. EplTR003-159 or EplTR030-267), usually being long enough to yield FISH signal and gaining genomic independence (see Table S1.7 for more examples of satDNA possible dynamics).

Interestingly, a conventional satDNA could disseminate short seeds around the genome serving as new roots for array polymerization (Ruiz-Ruano et al., 2016). This process would take place easily helping with the mobile nature of TEs in the case of TR-TE association (Paço et al., 2019). Those short blocks of satDNA could also be involved in the arrangement of complex repetitive regions as the centromeric islands described in *D. melanogaste*r (Le et al., 1995 and Chang et al., 2019) or the complex TE-satDNA networks described by Satović et al. (2016) in mollusc species. On the contrary, we have found that several TRs behaving as true satDNA families show arrays in form of monomers and dimers dispersed in the genome (Fig. 1.2 and Dataset 1.2 and 1.3) and this amount of short arrays is not correlated with the number of reads shared with other

repetitive elements so they could be involved in dissemination processes acting as seed for the birth and/or growing of new arrays of satDNAs as suggested by Mestrović et al. (1998) or Ruiz-Ruano et al. (2016) after following some of the mechanisms proposed by Lower et al. (2018). In fact, we have found an unexpected high amount of families with RS distribution of arrays RU in well-polymerized satDNAs (e.g. EplTR001-180 or EplTR002-196) than in simple TRs (chi = 6.587, df = 2, p= 0.03712). This finding suggests that the cyclic model for satDNA evolution would be closed with the dissemination of a short array from a satDNA (or from the degradation of old arrays that could be interrupted by other sequences) that starts over again the process of polymerization and homogenization. Eventually, the disseminated arrays could be very different from the original sequence (mutational processes, e.g. Belyayev et al., 2019) and a new satDNA family could emerge from there.

Altogether, our results indicate that satDNAs are ubiquitous repetitive elements in the genome that appear in complex structures, from dimers to large arrays harboring high number of repeats and differ from simply tandem repeats in their molecular and cytological features. This is the first exploration of TR families through different methodologies (Illumina, MinION and FISH) put forwards an empirical and plausible model for satDNA intragenomic evolution and development from tandem repeats. Finally, we found that dissemination of satDNA in short arrays around the genome is likely and in some cases these short arrays could be related to transposable elements, pointing out TEs as sources of satDNAs and TRs thus shaping the architecture of genomes.

References

Ahmed M, Liang P. (2012). Transposable elements are a significant contributor to tandem repeats in the human genome. *Comparative and Functional Genomics*, 2012, 947089.

Aldrup-MacDonald ME, Kuo ME, Sullivan LL, Chew K, Sullivan BA. (2016). Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome Research*, 26(10), 1301–1311.

Bachmann L, Raab M, Sperlich D. (1989). Satellite DNA and speciation: a species specific satellite DNA of *Drosophila guanche*. *Journal of Zoological Systematics and Evolutionary Research*, 27(2), 84–93.

Bao W, Kojima KK, Kohany O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6, 11.

Batistoni R, Pesole G, Marracci S, Nardi I. (1995). A tandemly repeated DNA family originated from SINE-related elements in the European plethodontid salamanders (Amphibia, Urodela). *Journal of Molecular Evolution*, 40, 608–615.

Baumgartner A, Weier JF, Weier HU. (2006). Chromosome-specific DNA repeat probes. *Journal of Histochemistry and Cytochemistry*, 54(12), 1363–1370.

Belyayev A, Josefiová J, Jandová M, Mahelka V, Krak K, Mandák B. (2020). Transposons and satellite DNA: on the origin of the major satellite DNA family in the *Chenopodium* genome. *Mobile DNA*, 11, 20.

Belyayev A, Josefiová J, Jandová M, Kalendar R, Krak K, Mandák B. (2019). Natural History of a Satellite DNA Family: From the Ancestral Genome Component to Species-Specific Sequences, Concerted and Non-Concerted Evolution. *International Journal of Molecular Sciences*, 20(5), 1201.

Bilgin Sonay T, Carvalho T, Robinson MD, Greminger MP, Krützen M, Comas D, et al. (2015). Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. *Genome Research*, 25(11), 1591–9.

Bolger AM, Lohse M, Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120.

Bois PR, Southgate L, Jeffreys AJ. (2001). Length of uninterrupted repeats determines instability at the unstable mouse expanded simple tandem repeat family MMS10 derived from independent SINE B1 elements. *Mammalian Genome*, 12(2), 104–111.

Brajković J, Feliciello I, Bruvo-Mađarić B, Ugarković D. (2012). Satellite DNA–like elements associated with genes within euchromatin of the beetle *Tribolium castaneum*. *G3 (Bethesda)*, 2(8), 931–941.

Brown WR, MacKinnon PJ, Villasanté A, Spurr N, Buckle VJ, Dobson MJ. (1990). Structure and polymorphism of human telomere-associated DNA. *Cell*, 63(1), 119–132.

Cabrero J, Bakkali M, Bugrov A, Warchalowska-Sliwa E, López-León MD, Perfectti F, Camacho JP. (2003). Multiregional origin of B chromosomes in the grasshopper *Eyprepocnemis plorans*. *Chromosoma*, 112(4), 207–11.

Cabrero J, Martín-Peciña M, Ruiz-Ruano FJ, Gómez R, Camacho JPM. (2017). Post-meiotic B chromosome expulsion, during spermiogenesis, in two grasshopper species. *Chromosoma*, 126(5), 633–644.

Camacho JPM, Cabrero J, Viseras E, López-León MD, Navas-Castillo J, Alché JD. (1991). G–banding in two species of grasshoppers 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. *Nature*, 371(6494), 215–220.

Cheng Z, Buell CR, Wing RA, Jiang J. (2002). Resolution of fluorescence in-situ hybridization mapping on rice mitotic prometaphase chromosomes, meiotic pachytene chromosomes and extended DNA fibers. *Chromosome Research*, 10(5), 379–387.

Dias GB, Heringer P, Svartman M, Kuhn GC. (2015). Helitrons shaping the genomic architecture of *Drosophila:* enrichment of DINE-TR1 in α- and β-heterochromatin, satellite DNA emergence, and piRNA expression. *Chromosome Research*, 23(3), 597–613.

Dias GB, Heringer P, Kuhn GC. (2016). Helitrons in *Drosophila:* Chromatin modulation and tandem insertions. *Mobile Genetic Elements*, 6(2), e1154638.

Drummond AJ, Ashton B, Cheung M, Heled J, Kearse M. (2009). Geneious v.4.8.5. Biomatters ltd. Aukland, New Zealand.

Duitama J, Zablotskaya A, Gemayel R, Jansen A, Belet S, Vermeesch JR, et al. (2014). Large-scale analysis of tandem repeat variability in the human genome. *Nucleic Acids Research*, 42(9), 5728–5741.

Easterling KA, Pitra NJ, Morcol TB, Aquino JR, Lopes LG, Bussey KC, et al. (2020). Identification of tandem repeat families from long-read sequences of *Humulus lupulus*. *PLoS One*, 15(6), e0233971.

Epstein ND, Karlsson S, O'Brien S, Modi W, Moulton A, Nienhuis AW. (1987). A new moderately repetitive DNA sequence family of novel organization. *Nucleic Acids Research*, 15(5), 2327–2341.

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(3), 564–567.

Feliciello I, Chinali G, Ugarković D. (2011). Structure and population dynamics of the major satellite DNA in the red flour beetle *Tribolium castaneum*. *Genetica*, 139(8), 999–1008.

Ferree PM, Barbash DA. (2009). Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*. *PLoS Biology*, 7(10), e1000234.

Frommer M, Prosser J, Vincent PC. (1984). Human satellite I sequences include a male specific 2.47 kb tandemly repeated unit containing one Alu family member per repeat. *Nucleic Acids Research*, 12(6), 2887–2900.

Frydrychová R, Grossmann P, Trubac P, Vítková M, Marec F. (2004). Phylogenetic distribution of TTAGG telomeric repeats in insects. *Genome*, 47(1), 163–178.

Garrido-Ramos MA. (2017. Satellite DNA: An Evolving Topic. *Genes (Basel)*, 8(9), 230.

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual Review of Genetics*, 44, 445–477.

Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, et al. (2016). Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature Genetics*, 48(1), 22–9.

Hannan AJ. (2018). Tandem Repeats and Repeatomes: Delving Deeper into the 'Dark Matter' of Genomes. *EBioMedicine*, 31, 3–4.

Henikoff S, Ahmad K, Malik HS. (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*, 293(5532), 1098–1102.

Henriques-Gil N, Santos JL, Arana P. (1984). Evolution of a complex polymorphism in the grasshopper *Eyprepocnemis plorans*. *Chromosoma* , 89, 290–293.

Kent WJ. (2002). BLAT--the BLAST-like alignment tool. *Genome Research*, 12(4), 656–664.

Khost DE, Eickbush DG, Larracuente AM. (2017). Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. *Genome Research*, 27(5), 709–721.

King M. (1987). Chromosomal rearrangements, speciation and the theoretical approach. *Heredity*, 56, 1–6.

Kirov I, Gilyok M, Knyazev A, Fesenko I. (2018). Pilot satellitome analysis of the model plant, *Physcomitrellapatens*, revealed a transcribed and high-copy IGS related tandem repeat. *Comparative Cytogenetics*, 12(4), 493–513.

Kohany O, Gentles AJ, Hankus L, Jurka J. (2006). Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, 7, 474.

Kroupin P, Kuznetsova V, Romanov D, Kocheshkova A, Karlov G, Dang TX, et al. (2019). Pipeline for the Rapid Development of Cytogenetic Markers Using Genomic Data of Related Species. *Genes (Basel)*, 10(2), 113.

Lohe AR, Brutlag DL. (1986). Multiplicity of satellite DNA sequences in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences USA*, 83, 696–700.

López-Flores I, Garrido-Ramos MA. (2012). The repetitive DNA content of eukaryotic genomes. *Genome Dynamics*, 7, 1–28.

López-León MD, Neves N, Schwarzacher T, Heslop-Harrison JS, 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.

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.

Lower SS, McGurk MP, Clark AG, Barbash DA. (2018). Satellite DNA evolution: old ideas, new approaches. *Current Opinion in Genetics & Development*, 49, 70–78.

McGurk MP, Barbash DA. (2018). Double insertion of transposable elements provides a substrate for the evolution of satellite DNA. *Genome Research*, 28(5), 714–725.

Miga KH. (2015). Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome Research*, 23(3), 421–426.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, et al. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology*, 14(1), R10.

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(8), 1062–1068.

Meredith R. (1969). A simple method for preparing meiotic chromosomes from mammalian testis. *Chromosoma*, 26(3), 254–8.

Mestriner CA, Galetti PM Jr, Valentini SR, Ruiz IR, Abel LD, Moreira-Filho O, et al. (2000). Structural and functional evidence that a B chromosome in the characid fish *Astyanax scabripinnis* is an isochromosome. *Heredity (Edinb)*, 85 (Pt 1), 1–9.

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(6), 792–793.

Novák P, Ávila Robledillo L, Koblížková A, Vrbová I, Neumann P, Macas J. (2017). TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Research,* 45(12), e111.

Otsu N. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62–66.

Paço A, Freitas R, Vieira-da-Silva A. (2019). Conversion of DNA Sequences: From a Transposable Element to a Tandem Repeat or to a Gene. *Genes (Basel)*, 10(12), 1014.

Palacios-Gimenez OM, Dias GB, de Lima LG, Kuhn GCES, Ramos É, Martins C, et al. (2017). High-throughput analysis of the satellitome revealed enormous diversity of satellite DNAs in the neo-Y chromosome of the cricket *Eneoptera surinamensis*. *Scientific Reports*, 7(1), 6422.

Pavlek M, Gelfand Y, Plohl M, Meštrović N. (2015). Genome-wide analysis of tandem repeats in *Tribolium castaneum* genome reveals abundant and highly dynamic tandem repeat families with satellite DNA features in euchromatic chromosomal arms. *DNA Research*, 22(6), 387–401.

Peacock WJ, Brutlag D, Goldring E, Appels R, Hinton CW, Lindsley DL. (1974). The organization of highly repeated DNA sequences in *Drosophila melanogaster* chromosomes. *Cold Spring Harbor Symposia on Quantitative Biology*, 38, 405–416.

Pita S, Panzera F, Mora P, Vela J, Cuadrado Á, Sánchez A, et al. (2017). Comparative repeatome analysis on *Triatoma infestans* Andean and Non-Andean lineages, main vector of Chagas disease. *PLoS One*, 12(7), e0181635.

Plohl M, Petrović V, Luchetti A, Ricci A, Satović E, Passamonti M, et al. (2010). Long–term conservation vs high sequence divergence: the case of an extraordinarily old satellite DNA in bivalve mollusks. *Heredity (Edinb)*, 104(6), 543–551.

Plohl M, Meštrović N, Mravinac B. (2012). Satellite DNA evolution. *Genome Dynamics*, 7, 126–152.

Quinlan AR, Hall IM. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841 - 842.

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R–project.org/.

Rang FJ, Kloosterman WP, de Ridder J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, 19(1), 90.

Raskina O, Barber JC, Nevo E, Belyayev A. (2008). Repetitive DNA and chromosomal

rearrangements: speciation-related events in plant genomes. *Cytogenetic and Genome Research*, 120, 351–357.

Rodríguez FR, de la Herrán R, Navajas-Pérez R, Cano-Roldán B, Sola-Campoy PJ, García-Zea JA, et al. (2017). Centromeric Satellite DNA in Flatfish (Order Pleuronectiformes) and Its Relation to Speciation Processes. *Journal of Heredity*, 108(2), 217–222.

Rodrigues PHM, Dos Santos RZ, Silva DMZA, Goes CAG, Oliveira C, Foresti F, et al. (2019). Chromosomal and Genomic Dynamics of Satellite DNAs in Characidae (Characiformes, Teleostei) Species. *Zebrafish*, 16(4), 408–414.

Röschenthaler F, Schäble KF, Thiebe R, Zachau HG. (1992). Of orphons and UHOs. Delimitation of the germline repertoire of human immunoglobulin kappa genes. *Biological Chemistry Hoppe-Seyler*, 373(4), 177–186.

Ruiz-Ruano FJ, López-León M, Cabrero J, Camacho JPM. (2016). High–throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific Reports*, 6, 28333.

Ruiz-Ruano FJ, Cabrero J, López-León MD, Sánchez A, Camacho JPM. (2018a). Quantitative sequence characterization for repetitive DNA content in the supernumerary chromosome of the migratory locust. *Chromosoma*, 127(1), 45–57.

Ruiz-Ruano FJ, Castillo-Martínez J, Cabrero J, Gómez R, Camacho JPM, López-León MD. (2018b). High-throughput analysis of satellite DNA in the grasshopper *Pyrgomorpha conica* reveals abundance of homologous and heterologous higher-order repeats. *Chromosoma*, 127(3), 323 - 340.

Salser W, Bowen S, Browne D, el-Adli F, Fedoroff N, Fry K, et al. (1976) . Investigation of the organization of mammalian chromosomes at the DNA sequence level. *Federation Proceedings*, 35(1), 23–35.

Satović E, Plohl M. (2013). Tandem repeat-containing MITEs in the clam *Donax trunculus*. *Genome Biology and Evolution*, 5(12), 2549–2559.

Satović E, Vojvoda Zeljko T, Luchetti A, Mantovani B, Plohl M. (2016). Adjacent sequences disclose potential for intra-genomic dispersal of satellite DNA repeats and suggest a complex network with transposable elements. *BMC Genomics*, 17(1), 997.

Schmieder R, Edwards R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6), 863–864.

Sharma A, Wolfgruber TK, Presting GG. (2013). Tandem repeats derived from centromeric retrotransposons. *BMC Genomics*, 14, 142.

Silva BSML, Heringer P, Dias GB, Svartman M, Kuhn GCS. (2019). De novo identification of satellite DNAs in the sequenced genomes of *Drosophila virilis* and *D. americana* using the RepeatExplorer and TAREAN pipelines. *PLoS One*, 14(12), e0223466.

Silva DMZA, Utsunomia R, Ruiz-Ruano FJ, Daniel SN, Porto-Foresti F, Hashimoto DT, et al. (2017). High-throughput analysis unveils a highly shared satellite DNA library among three species of fish genus *Astyanax*. *Sci Reports*, 7(1), 12726. Erratum in: Sci Rep. (2020 Jan 8; 10(1), 190.

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

Smit AFA, Hubley R, Green P. (2013). RepeatMasker Open–4.0. http://www.repeatmasker.org.

Speel EJ, Hopman AH, Komminoth P. (2006). Tyramide signal amplification for DNA and mRNA

in situ hybridization. *Methods in Molecular Biology*, 326, 33–60.

Sproul JS, Khost DE, Eickbush DG, Negm S, Wei X, Wong I, et al. (2020). Dynamic Evolution of Euchromatic Satellites on the X Chromosome in *Drosophila melanogaster* and the simulans Clade. *Molecular Biology and Evolution*, 37(8), 2241–2256.

Sun X, Wahlstrom J, Karpen G. (1997). Molecular structure of a functional *Drosophila* centromere. *Cell*, 91, 1007–1019.

Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, et al. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, 47(21), 10994–11006.

Tu Z, Li S, Mao C. (2004). The changing tails of a novel short interspersed element in *Aedes aegypti*: genomic evidence for slippage retrotransposition and the relationship between 3' tandem repeats and the poly(dA) tail. *Genetics*, 168(4), 2037–2047.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Research*, 40, e115–e115.

Usdin K. (2008). The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. *Genome Research*, 18(7), 1011–1019.

Utsunomia R, Silva DMZA, Ruiz-Ruano FJ, Goes CAG, Melo S, Ramos LP, et al. (2019). Satellitome landscape analysis of *Megaleporinus macrocephalus* (Teleostei, Anostomidae) reveals intense accumulation of satellite sequences on the heteromorphic sex chromosome. *Scientific Reports*, 9(1), 5856.

Vondrak T, Ávila Robledillo L, Novák P, Koblížková A, Neumann P, Macas J. (2020). Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats. *Plant Journal*, 101(2), 484–500.

Wang CJ, Harper L, Cande WZ. (2006). High-resolution single-copy gene fluorescence in situ hybridization and its use in the construction of a cytogenetic map of maize chromosome 9. *Plant Cell*, 18(3), 529–544.

Warburton PE, Hasson D, Guillem F, Lescale C, Jin X, Abrusan G. (2008). Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics*, 9, 533.

Wei KH, Grenier JK, Barbash DA, Clark AG. (2014). Correlated variation and population differentiation in satellite DNA abundance among lines of *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences USA*, 111(52), 18793–18798.

Supplementary Information for Chapter 1

FISH using oligos. Some TR families coincided in FISH location and show certain homology in consensus sequences as EplTR001-180, EplTR002-196, EplTR006-147 EplTR019-75 or EplTR071-180. To double check chromosomal location of these families we designed oligos in specific regions of the consensus sequence of each TR that were previously aligned using Geneious v4.8 (Drummond et al., 2009) and performed FISH with that oligos (see Table S1.9). Minimum length required for oligos was 20 nt and they were designed in a region showing as maximum 50% of identity with the coinciding TR. Oligos were labeling through terminal transferase (TdT) and FISH conditions were the same as described in Materials and Methods. Results showed the same FISH pattern as the one yielded by PCR amplified probes and some of the TR families failed in oligo FISH so we considered first FISH results for subsequent analysis.

Identification of TR arrays in long reads of *L. migratoria* and *H. sapiens* to check the 7mer satDNA congruence. We searched for TR arrays in the MinION library of *L. migratoria* included in Ruiz-Ruano et al. (2018) (0.02x genomic coverage) using RepeatMasker v4.0.5 (Smit et al., 2013), with the same options as for *E. plorans,* against the *L. migratoria* database of repetitive DNA described in Ruiz-Ruano 2018. However, for TR we included exclusively the most abundant variant of each family due to the lack of resolution of variants in long reads having high error rates. We calculated abundance, PI, maxAL and maxRU in MinION reads (Table S1.12). 56 out of 62 TR families were found in MinION reads however only 49 showed a maxRU higher than 2 probably due to the low genomic coverage of the library. We considered the 56 TR families found in the library to compare properties between families with maxRU>=7 and maxRU<7 and we found significant difference between both groups in respect to PI, maxAL and FISH pattern (FISH details in Ruiz-Ruano et al., 2016).

To test the 7 RUs threshold in human we downloaded 34 Gb of CCS PacBio reads from the NCBI repository indicated above (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_15kb/) and mapped them using RepeatMasker to the database of repetitive DNA of *H. sapiens* that can be found in RepBase (Bao et al., 2015; last accessed may 2019). However we analyzed only those TR families that were described in UCSC (Karolchik et al., 2004) as "Satellite" repClass summing up 21 TR families (Table S1.13). We found all these families

in the long reads although two of them (REP522 and HSAT1) were not tandemly repeated (maxRU<2 RUs). We found differences between families of showing maxRU>=7 and those with maxRU<7 in terms of PI but not of maxAL.

References

Bao W, Kojima KK, Kohany O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6, 11.

Drummond AJ, Ashton B, Cheung M, Heled J, Kearse M. (2009). Geneious v.4.8.5. Biomatters ltd. Aukland, New Zealand.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, 32(Database issue), D493–6.

Ruiz-Ruano FJ, López-León M, Cabrero J, Camacho JPM. (2016). High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific Reports*, 6, 28333.

Ruiz-Ruano FJ, Cabrero J, López-León MD, Sánchez A, Camacho JPM. (2018). Quantitative sequence characterization for repetitive DNA content in the supernumerary chromosome of the migratory locust. *Chromosoma*, 127(1), 45–57.

Smit AFA, Hubley R, Green P. (2013). RepeatMasker Open-4.0. http://www.repeatmasker.org.

Supplementary Figures for Chapter 1

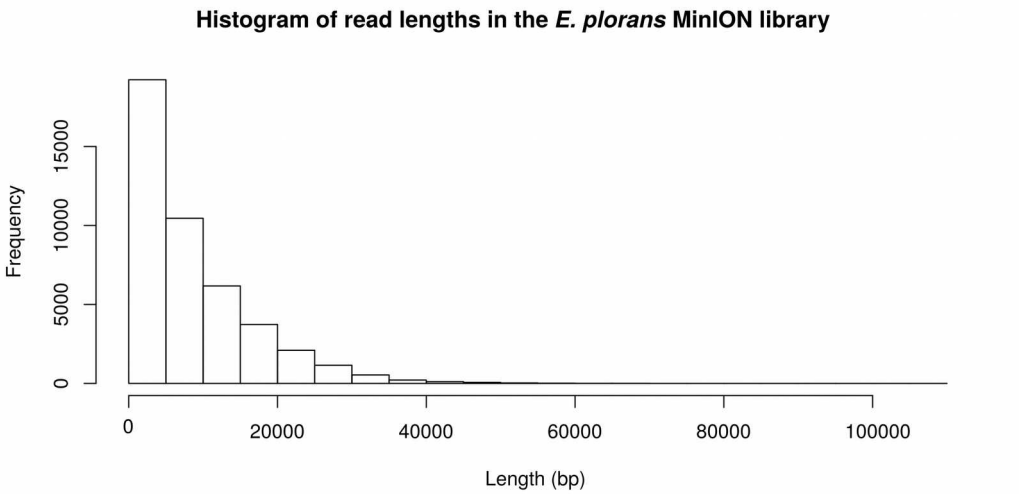


Figure S1.1. Read length distribution of the Oxford Nanopore MinION library of *Eyprepocnemis plorans.*

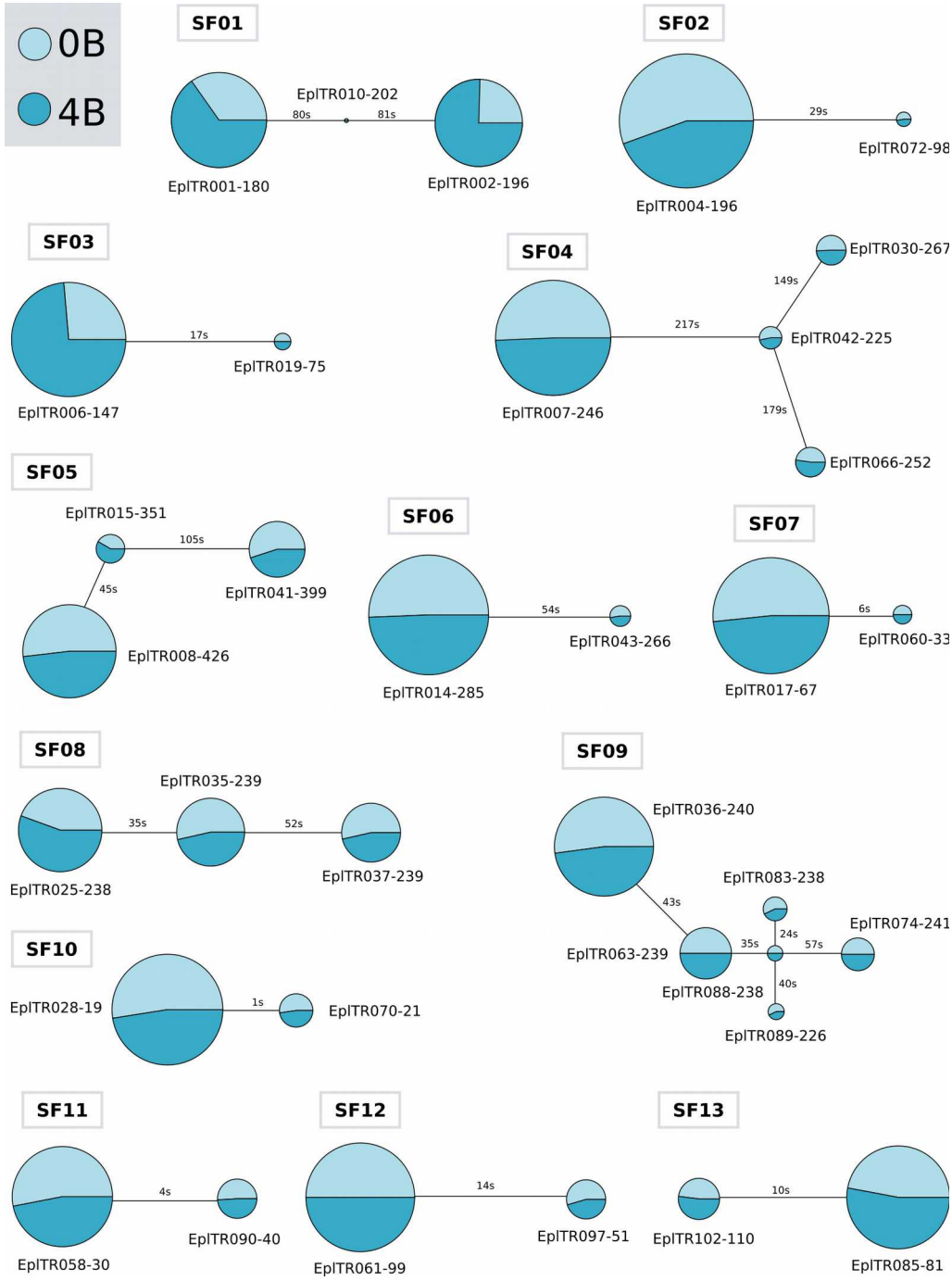


Figure S1.2. Minimum spanning trees for each TR superfamily (SF) of *E. plorans*. Within each SF, link between families and circle sizes are proportional to the number of substitutions and abundance respectively. See that TR abundance in 0B (light blue) and 4B (dark blue) libraries is shown as proportion as sectors inside circles representing TR families.

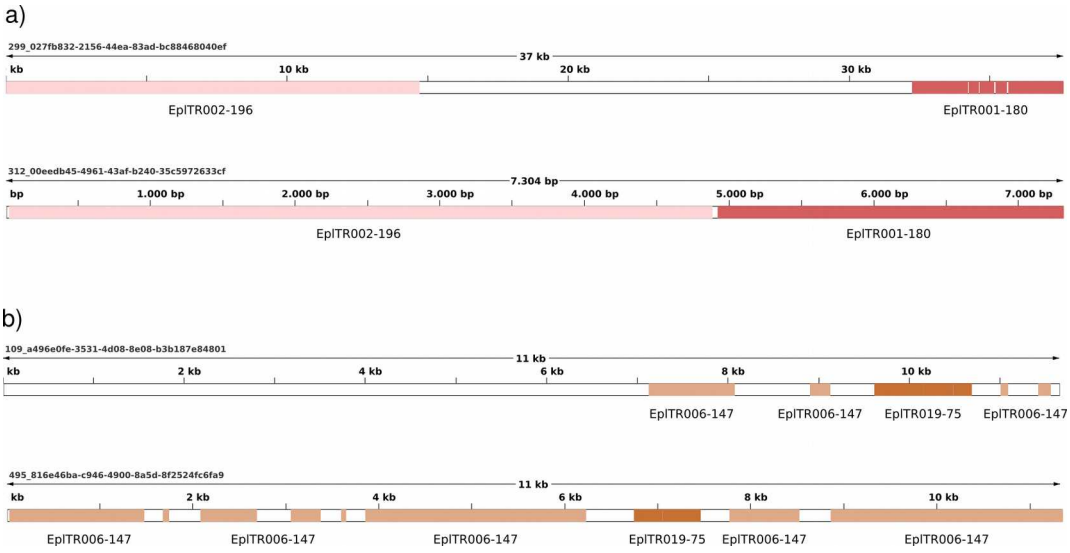


Figure S1.3. Examples of TR families with similar FISH location and co-existence in the same read. a) Two reads shared by EplTR001-180 and EplTR002 TRs belonging to the same SF01 superfamily and yielding similar FISH signals. b) Examples of two reads containing both families of the SF03 superfamily, EplTR006-147 and EplTR0019-75 with resemblance in FISH results.

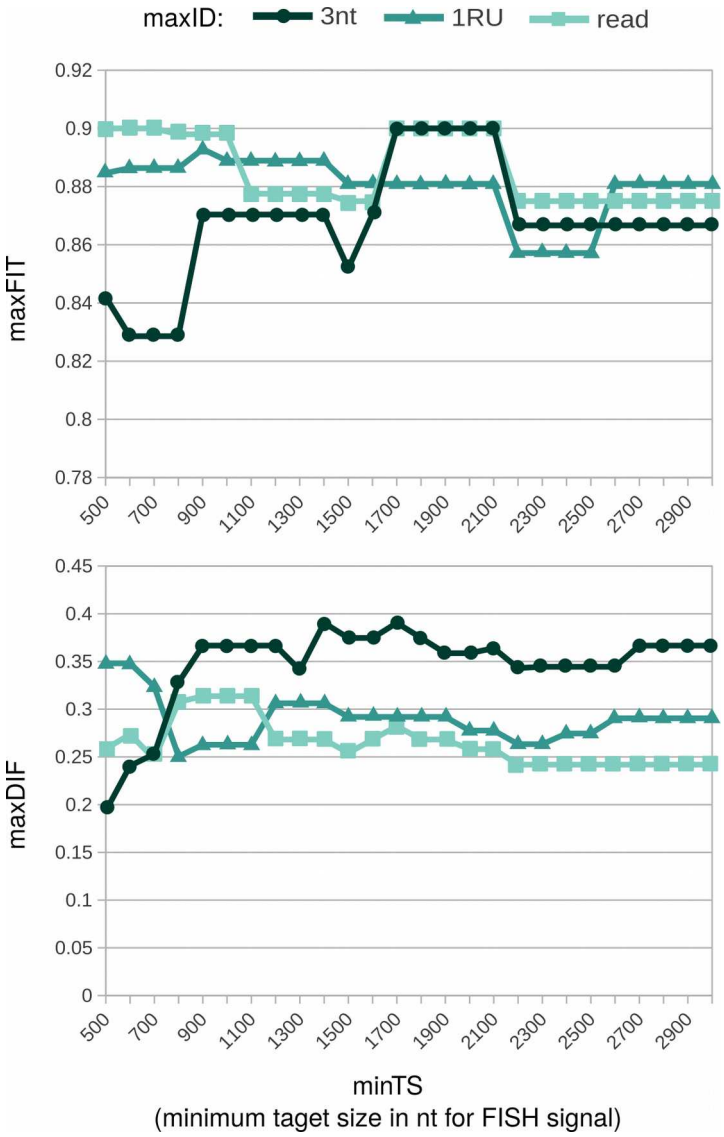


Figure S1.4. Graphical representation of FIT (up) and DIF (down) values at 3 nt, 1 RU and the read length as maxID (maximum inter-array distances) and considering different minTS (minimum target size for FISH visualization). The best DIF and FIT values are found 1,700 nt, specially at 3 nt as maxID.

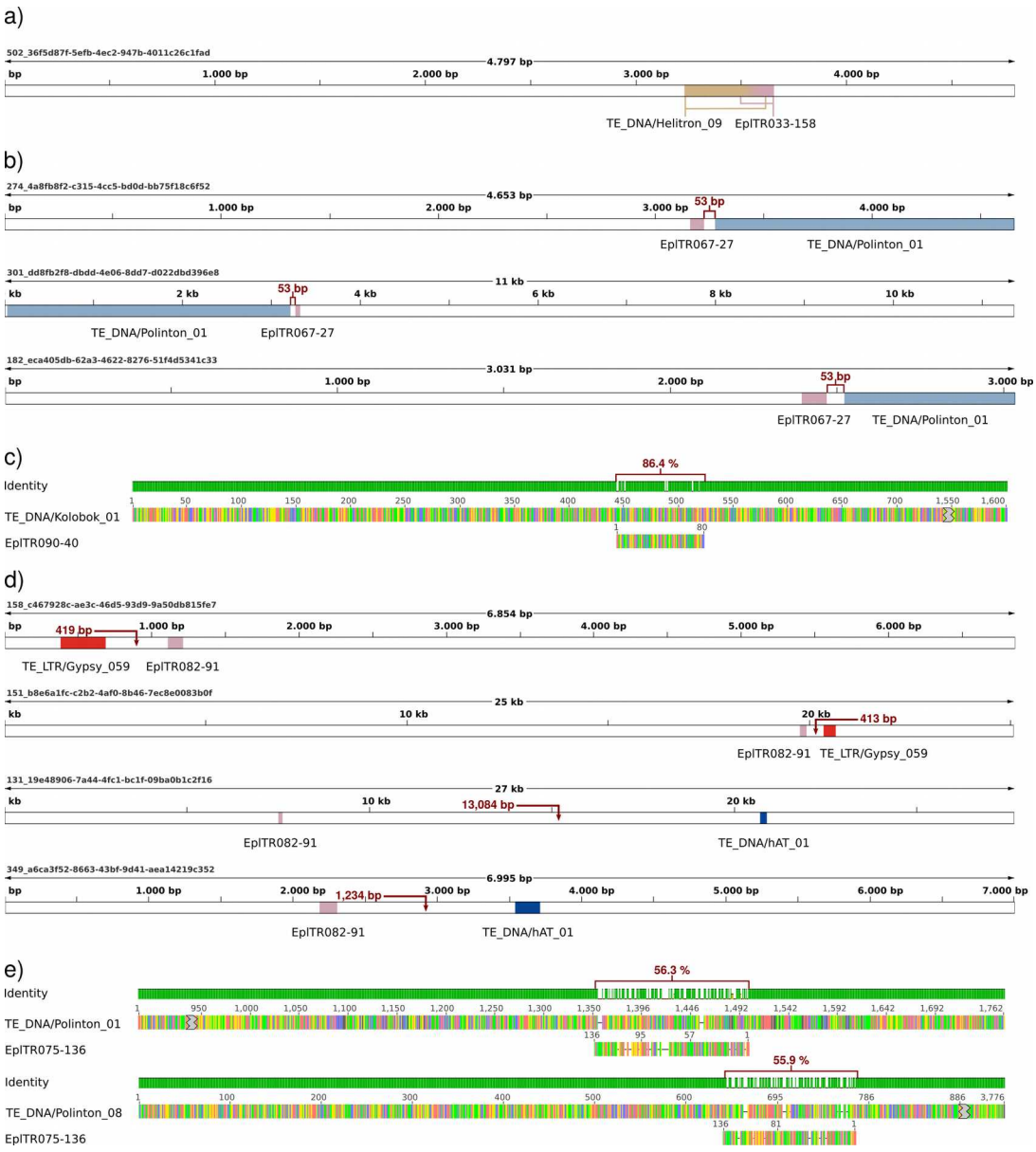


Figure S1.5. Examples of TR families accomplishing the criteria for TR-TE association. a) Overlapping matches in the same MinION read. b) Conserved distance between the TR and the TE. c) TR-TE homology. d) EplTR082-91 located at similar distances in different reads from Gypsy_059 (up) but at wide ranging ones in respect to hAT_01 (down). e) EplTR075-136 showing homology in sequence with to different Polinton families.

Supplementary Datasets and Tables for Chapter 1

Dataset 1.1. FISH results for the 112 TR families identified in the genome of *E. plorans* using satMiner.
Dataset available in: https://figshare.com/s/a035ec047946ef1de483

Dataset 1.2. Histograms of RU (repeats unit) in arrays for the 93 TR families of *E. plorans* found in MinION reads considering 3 nt as maxID (maximum inter-array distance). We included Kimura divergence (%) of each TR array represented by a scatter plot referred to a secondary y-axis (right) showing also logarithmic regression of data.
Dataset available in: https://figshare.com/s/cdd54f9245b8632a1cb9

Dataset 1.3. Spaghetti plot of distribution of RU in arrays for the 93 TR families of *E. plorans* in MinION reads. Each line graph represents a RUs distribution considering different max IDs (dark turquoise – 3 nt, middle turquoise – 1 RU – , light turquoise – read) and methodological approaches (middle brown – excluding terminal arrays/junctions<50 nt, light brown – removing data from reads shorter 5,000 nt).
Dataset available in: https://figshare.com/s/f916aafb6f8ea4ad637b

Dataset 1.4. Graphical representation of inter-array distance for the 93 TR families of *E. plorans* found in MinION reads. Log10 of the raw distances (nt) between each pair of contiguous arrays is indicated with a pink circle in an vertical line representing each read. If the same distance value is found between different arrays in the same read it is represented by a surrounding green circle indicating frequency. The colour of the circle turns from light to dark pink according to the increase in RUs in the array. Note that TR families counting on one array per read can not be included in these graphics.
Dataset available in: https://figshare.com/s/9b0ba3693b1a71753a28

Dataset 1.5. Congruence between molecular and cytological data applied to the 93 TR families of *E. plorans* found in MinION reads. Information regarding maxIDs (maximum inter-array distances) considered (3 nt, 1 RU and the read) are included in different sheets. The sheets "summary" contains values of minTS (minimum target site detected by FISH), RUL (repeat unit length) and A+T% (AT content) intervals that can be modified to explore data performance automatically.
Dataset available in: https://figshare.com/s/bb2d5eda5db0ec083c7d

Supplementary Tables can be downloaded from:
https://figshare.com/s/de57d278fcf2cb951d97
Table S1.1. Features of the 93 TR families of *E. plorans* found in MinION read. We included sequences properties, abundances in Illumina and MinION libraries, polymerization characteristics, FISH results and amount of TE (transposable elements) sharing long reads with each TR family. TR families are named according to the species name "Epl", kind of element ("TR"), number based on abundance in Illumina reads including the repetitive DNA database as reference (column O) and the RUL of each TR family.
Table S1.2. Polymerization properties of the 93 TR families found in MinION reads of *E.*

plorans at 3 nt, 1 RU and the read maxIDs (maximum inter-array distance). We show values of PI (polymerization index), maxAL (maximum array length), maxRU (maximum repeat unit per array) and the difference between measurements at maxIDs of 1 RU and the read in respect to 3 nt. See also data of TSI (tandem structure index) in Illumina reads for each TR family.

Table S1.3. Results of Spearman correlation between abundances (in Illumina and MinION 4B libraries) and polymerization features (PI, maxAL, maxRU) at different maxIDs of the 93 TR families of *E. plorans* found in long reads.

Table S1.4. Results of Spearman correlation between Kimura divergence of each TR array and the number of repeats it harbours for each of the 93 TR families of *E. plorans* found in MinION reads. High error rate in MinION reads affects divergence estimations, we used divergence in MinION reads to performed comparisons between families assuming that the error affects all equally and the signal is maintained but absolute divergence values are not reliable.

Table S1.5. maxFIT and maxDIF values at 3 nt, 1 RU and the read length as maxID (maximum inter-array distances) and considering different minTS (minimum target size for FISH visualization). The best DIF and FIT values are found 1,700 nt, specially at 3 nt as maxID.

Table S1.6. TR families sharing >50% of their reads with transposable elements (TEs). When a TR shares >50% of reads with a TE, distance between both is conserved (Mode), they showed homology (TR-TE Hom) or overlapped annotation (Overlapped annot.) TR is considered as related to TE (TE assoc.). We validated this association whether we were able to cluster the TE using TR-homologous reads (RE clust. of TE) and/or amplified a tandem repeat PCR pattern (PCR) anchoring primers to the TE (TE assoc. validated).

Table S1.7. Discriminant analysis of the 93 TR families of *E. plorans* comprising the molecular parameters: RUL (repeat unit length), A+T% (AT content), maxRU>=7 (maximum repeat unit found higher or equal to 7 RUs), TE assoc. (association to TEs), TE assoc. validated (association to TEs validated by RE reclustering or PCR amplification). The FISH pattern was used as a grouping factor. See highlighted in red misassignation of prediction. In "Notes" we indicated some remarks about possible stages of satDNA evolution in function of their homogenization and polymerization degree, we also indicate if it was related to TEs.

Table S1.8. Primers used for amplification and probe labeling of the 112 TR families identified in *E. plorans* with satMiner.

Table S1.9. Sequence of oligos labeled for FISH of TR families of *E. plorans* that co-localized when FISH was performed with PCR-probes.

Table S1.10. Primers used for amplification TR families homologous to transposable elements (TEs) of *E. plorans*.

Table S1.11. Results to one run of TAREAN in a random selection of 200,000 Illumina paired-end reads of *E. plorans*. Note that NS families (EplTR016-31 and EplTR081-65) were also identified with TAREAN.

Table S1.12. Polymerization features of 56 out of 62 TR families found in MinION reads of *L. migratoria*. FISH pattern was retrieved from Ruiz-Ruano et al. (2016).

Table S1.13. Polymerization features of the 21 TR families of *H. sapiens* described as "Satellites" in UCSC.

Table S1.14. Wilk's Lambda from the discriminant analysis (DA) showing signification considering different grouping variables. The FISH pattern is the parameter that shows higher signification when assigning groups to TRs based on the remaining variables.

2. Repetitive DNA content in the B chromosomes of the grasshopper *Eyprepocnemis plorans*

Abstract: B chromosomes, also known as supernumerary, are dispensable elements to the standard set of A chromosomes reported in almost all eukaryotic groups, including the grasshopper *Eyprepocnemis plorans*. The current view is that B chromosomes are parasitic elements similar to selfish DNA but there is still a gap of knowledge about their origin and molecular make up in several species. Due to the fact that they do not recombine with the standard complement (As), B chromosomes follow their own evolutionary pathway, thus escaping from selective pressures which leads to the accumulation of repetitive elements.

The present work is based on the analysis of the repetitive DNA contained in two genomic libraries of *E. plorans* from Torrox (Spain), one of them lacking B chromosomes (0B) and the other one carrying four of them (4B), obtained through Illumina sequencing. We built a *de novo* database of repetitive DNA in this species that we used as a reference to quantitatively identify the repetitive elements being over-represented in the 4B library, hence likely located in the B chromosomes of *E. plorans*. Some of these elements were analyzed using fluorescent *in situ* hybridization to determine their chromosomal location. This approach, along with the search of some B-located repetitive elements in long MinION reads, allowed us to propose the origin of B chromosomes from the chromosomes 9 of the A complement. Furthermore, the origin of Bs would have likely taken place in a common event as inferred after quantification of repetitive DNA in Bs of *E. plorans* from African and Asian populations. Finally, we found B specific transcripts of repetitive elements that were more intensely expressed in gonads that in other biological samples as legs or embryos. However, this transcriptional activity from the B chromosomes is lower than the one from the As suggesting a possible silencing of repetitive DNA in B chromosomes of *E. plorans*.

Keywords: *B chromosomes, FISH*, *Illumina*, *repetitive DNA*, *satDNA, TEs*

Introduction

Grasshoppers are generally characterized by large genome sizes. In particular, the genome of *Eyprepocnemis plorans* is around 11 Gb in size (Ruiz-Ruano et al., 2011), which is about three times the size of the human genome. This genome is organized into 22 autosomes plus a sexual X0 or XX pair, for males and females respectively, all of them being acrocentric chromosomes. Typically, one or more supernumerary chromosomes are added to this standard complement. These chromosomes, commonly known as B chromosomes, exist in thousands of eukaryote species although they are dispensable for carrier individuals. They also have a parasitic character that is based on their detrimental effects on the host when they are in high numbers inside the cell and their accumulation across generations through different mechanisms (Jones, 1991).

Despite recent and progressive advances in data analysis and sequencing technologies, the molecular content of the B chromosomes of *E. plorans* remains a mystery. The B chromosomes, by not recombining with the As, follow their own evolutionary pathway, without being subjected to selective pressures. For this reason, they constituted and ideal place in which repetitive DNA can disseminate and amplify. In fact, most of the B chromosomes described so far possess a molecular content rich in repetitive DNA, especially satDNA, in concordance with their heterochromatic nature (see Camacho et al., 2005 for review). For example, the B chromosomes of maize contain sequences along their length that, in contrast, are restricted to centromeric regions on the A chromosomes (Lamb et al., 2005). In addition, some retrotransposons have been found on B chromosomes, this is the case of the Stark B retroelement from maize B chromosomes that is composed of fragments from the A genome as well as B-specific sequences (Lamb et al., 2007) or the retrotransposon NATE located in the PSR chromosome (B chromosome) of *Nasonia vitripennis* (McAllister, 1995). Tandem repeats are also common components of B chromosomes, Bs of *Creppis capillaris* (Jamilena et al., 1994) share some satellites DNA with A chromosomes but there are also specific satellites DNAs of B chromosomes as several satellites described in rye (Kemmel et al., 2013). All of these aside, there are few studies that extensively address the analysis of repetitive DNA on B chromosomes apart from those of Ruiz-Ruano et al. (2018) in the migratory locust, Coan and Martins (2018) for TEs in the Bs *Astatotilapia latifasciata* and Ebrahimzadegan et al. (2019) in *Festuca pratensis.* These kind of analysis requires a

previous characterization of the genomic repetitive DNA which is often an arduous task due to the high amount and diversity of repetitive DNA classes including satDNA, TEs, tRNAa, histones, rDNA, mtDNA, snRNA genes and also some pseudogenic families.

For a long time, it has been claimed that B chromosomes do not contain genes (Camacho et al., 2000; Jones and Houben, 2003; Burt and Trivers, 2006). This is probably due to the difficulty of detecting them efficiently and the high enrichment of repetitive elements that can hide any gene signal of single copy genes or in low number of copies (Navarro-Domínguez, 2016). However, about 15 years ago, first reports appeared identifying single copy genes or pseudogenes on the B chromosomes in a wide range of species and the list keeps growing (Graphodatsky et al., 2005; Yoshida et al., 2011; Martis et al., 2012; Trifonov et al., 2013; Banaei-Moghaddam et al., 2013; Valente et al., 2014; Huang et al., 2016; Carmello et al., 2017; Ma et al., 2017; Navarro-Domínguez et al., 2017a, b, 2019, Ruiz-Ruano et al., 2019; Dalla Benetta et al., 2020). Furthermore, many of these investigations demonstrate the transcriptional activity of the sequences contained in the B chromosomes, so the widespread assertion about the inertness of the B chromosomes has been discarded in light of these evidences (Camacho et al., 2000). In the case of *E. plorans*, Navarro-Domínguez et al. (2017a) found ten protein-coding genes residing in their B chromosome, five of which were actively transcribed. In addition, they also detected several changes in gene expression associated with the presence of B chromosomes in *E. plorans,* suggesting the possibility of a transcriptional crosstalk taking place between A and B chromosomes in B-carrying individuals of the species (Navarro-Domínguez et al., 2019).

The repetitive DNA represents the outstanding fraction of B chromosomes, in fact, those of *E. plorans* are known to contain mainly 180 bp satellite DNA, located in their heterochromatin region, and ribosomal DNA distantly in this chromosomes (López-León et al., 1994). Also a few sequences of the R2 mobile element have been described to be located in those extra chromosomes (Montiel et al., 2014). The presence of ribosomal RNA transcripts specifically coming from B chromosomes was confirmed by Ruiz-Estévez et al. (2012), although it was detected only in a few males (Ruiz-Estévez et al., 2013) and the relative rRNA contribution of the B chromosome compared to that from the host genome was negligible (Ruiz-Estévez et al., 2014).

In this study we go deep into the repetitive nature of the B chromosomes of *E. plorans.* With this in mind, we followed a quantitative approach similar to the one

indicate in Ruiz-Ruano et al. (2018) but starting from the construction of a *de novo* repetitive DNA database for this grasshopper. The characterization of the genomic repetitive DNA of *E. plorans* allowed us to identify several repetitive elements located in the B chromosomes apart from the sat180bp, rDNA and R2 previously described. In that respect, the characterization of B-located repetitive DNA may help to complete their molecular composition, shedding light on the origin and evolution of B chromosomes of *E. plorans*. In fact, the repetitive DNA make up of B chromosomes has helped to elucidate the B chromosome origin of several species as those in *Eumigus monticola* (Ruiz-Ruano et al., 2017), *Moenkhausia sanctaefilomenae* (Utsunomia et al., 2016) or *Drosophila melanogaster* (Hanlon et al., 2018).

Current bioinformatic approaches allow the *de novo* assembly of repetitive DNA from massive sequencing genomic data without the need for a reference genome. This is especially relevant when studying non-model species as *E. plorans.* As there is no reference genome for this grasshopper and the genome size of this species is too big to inexpensively get a high-coverage through sequencing, we bet on low coverage sequencing to explore the repetitive DNA content of its genome. This method assures enough representation of transposable elements (TEs) and other repetitive DNA sequences in genomic libraries, thus enabling their assembly (Goubert et al., 2015). Employing NGS, our approach is focused on the comparative analysis of repetitive DNA in 0B versus 4B gDNA libraries of the grasshopper *E. plorans*, which allowed us to get the over-represented sequences in the 4B genome that are candidates to be located on the B chromosomes. Some of these B-located sequences have been verified by fluorescent *in situ* hybridization. In addition, we explored their transcriptional activity searching for B-specific SNPs in gDNA and RNA libraries, the latter belonging to embryos (from two different pods) and adults (leg and gonads) from both sexes of *E. plorans*.

Our new findings highlight the repetitive nature of B chromosomes of the grasshopper *E. plorans* mainly composed of satDNA. Furthermore, we show transcription directly from B-located sequences in embryos, legs and gonads of *E. plorans*. Interestingly, the identification of B-specific elements allowed us to hypothesize the origin of the B chromosomes from the chromosome 9 of the A complement. Finally, the high resemblance between the DNA content, including the presence of B-specific element, of the B chromosomes from Spanish, African and Asian populations of *E. plorans* supports the common origin of Bs in this species.

Materials and methods

Biological material, DNA isolation and sequencing

E. plorans individuals were collected in Torrox (Málaga), a population where the prevalent B chromosome variant is the B24. Some gravid females were bred in the laboratory in order to obtain embryos for FISH experiments. Then, egg pods were incubated at 28 °C for 10 days, after which embryos were fixed in 3:1 ethanol-acetic acid and used for FISH studies. Males were anesthetized before dissecting out some testis follicles, one of which was fixed in 3:1 ethanol-acetic acid for cytological analysis as described in Camacho et al. (1991), while the remaining testis and body were frozen in liquid nitrogen for nucleic acid extraction. Two or three fixed follicles where squashed in slide with a drop of 2% acetic orcein to determine the number of B chromosomes of each male.

We extracted gDNA from hind legs of two males, one 0B and the other carrying 4 B chromosomes, using the GenElute Mammalian Genomic DNA Miniprep kit (Sigma). Quality was checked by 1.5% TBE-agarose gel electrophoresis and concentration measured with Infinite M200 Pro NanoQuant (Tecan) which also allowed us to verify gDNA quality. The two libraries of gDNA from these two males of *E. plorans* were sequenced on a single lane and generated 101 bp paired-end sequences using Illumina HiSeq 2000 platform in Macrogen (Seoul, Korea). A total of 80,269,402 and 69,181,368 reads were obtained from the 0B and 4B sample (~0.7x genome coverage), respectively. The quality trimming of reads was performed using Trimmomatic (Bolger et al., 2014), removing adapters and retaining complete reads pairs with Q>20 (~0.5x genome coverage). Raw reads of *E. plorans* used in this study are available in NCBI SRA database under accession numbers SRR2970625 (0B gDNA) and SRR2970627 (4B gDNA).

We also sequenced the genomic DNA from 2 *E. plorans* males from the same population located in Armenia (one 0B, another +B) and another 2 from Egypt (one 0B, another +B) that were sent to us by external research groups collaborators. The individuals were cytogenetically studied to characterize their content on B chromosomes. In addition, the presence or absence of B chromosomes was confirmed by PCR of the SCAR marker (Muñoz-Pajares et al., 2011). Genomic DNA for sequencing was extracted from one leg of each individual (previously fixed in liquid nitrogen and stored at -80 °C) using the same protocol indicated above. That gDNA was sent to Novogene

Bioinformatics Technology Co., Ltd (Headquarters) to sequence four libraries (each one belonging to one of the four males analyzed). We obtained about 8 Gb for Armenian males and about 7 Gb for the Egyptian ones. In both cases they were sequenced in an Illumina HiSeq X platform yielding of 150 bp reads, representing a genomic coverage of around 0.74x and 0.67x per individual respectively. Two *E. plorans* males (one 0B and one +B), received from colleges in Tanzania, were also processed as described above and its gDNA was sequenced using a Illumina HiSeq 2500 platform in the same company as for Egypt and Armenia samples, yielding about 8 Gb of paired-end 125 bp reads (~0.8x).

For transcription analysis we sequenced female and male embryos of *E. plorans* that were grown in the lab from adults individuals (around 40 individuals) collected also in the population of Torrox (Málaga) (36.737558N, -3.953546W). We processed captured individuals *in vivo* as follows: we extracted several testis follicles from males through a small cut in the abdomen and cytologically analyzed primary spermatocytes at diplotene or metaphase I to score the number of B chromosome of each individual. For the same purpose, in the case of females we performed C-banding on hemolymph nuclei as described in Cabrero et al. (2006). Then, we set up several controlled crosses between 0B and 1B individuals of *E. plorans* to get 0B and 1B embryos. Sibling embryos from these controlled crosses were processed to obtain material for RNA extraction as well as for cytogenetic analysis to determine their sex and B chromosome content. Doing so, we were able to get three sibling embryos from each of the following groups: females 0B, males 0B, females 1B and males 1B. We got this set of embryos from two different controlled crosses, the female parental was the B-carrying progenitor in one of them (F1BxM0B, embryos P1 from now on) and the male parental was the one who harbored the B chromosome in the other (F0BxM1B, embryos P2), counting with 24 embryos in total. We extracted RNA from embryos using the RNeasy Lipid Tissue kit (Qiagen), following manufacturer's recommendations. Then we sequenced the 24 RNA libraries in Beijing Genomics Institute, BGI (China), but in different moments. Sequencing of the 12 embryo libraries coming from the P1 cross was performed in two lanes of Illumina HiSeq 2000 which yielded about 3 Gb of 2x100 nt reads per library (we discarded one library of a 0B male for subsequent analysis due to sequencing failure). Remaining libraries from the P2 cross were sequenced in an Illumina HiSeq 4000 platform producing 6 Gb of 2x150 nt read per library approximately.

In 2016 we collected about 14 adults of *E. plorans* from Torrox (36.737558N,

-3.953546W) and they were all processed *in vivo* to set up controlled crosses in order to get sibling adults for RNA-seq experiments. Every egg pod from controlled crosses in the lab was maintained in an incubator chamber at 28 ºC until hatching, moment in which they were transferred to wooden boxes (one for each cross) until their development to adults. After seven days since the last shedding to adults, each individual was studied cytogenetically by hemolymph C-banding (described in Cabrero et al., 2006) and at the same time we extracted and fixed the gonads, hind legs and body in liquid nitrogen and preserved at -80 ºC. When characterized, we selected individuals coming from the same cross counting with at least 3 males 0B, 3 females 0B, 3 males 1B and 3 males 1B for each of which we performed two extractions of RNA per individual, one from the leg (Total RNA Spin Plus, Durviz) and one from the gonads (RNeasy Lipid Tissue kit, Qiagen) followed by sequencing. We sequenced a total of 24 libraries from the different samples of *E. plorans* individuals that we processed: 2 sexes x 2 tissues (leg/gonad) x 2 categories (0B/1B) x 3 biological replicates. These libraries were also sequenced in BGI but in an Illumina HiSeq 4000 platform yielding about 6 Gb of 150 nt paired-end reads per library. In sum, we used for transcription studies 47 RNA libraries from *E. plorans* adults and embryos of both sexes and having/lacking B chromosomes.

Construction of a repetitive DNA database for *E. plorans*

Transposable elements (TEs)

We used RepeatExplorer (RE) (Novák et al., 2013) for *de novo* identification of repetitive element in the genome of *E. plorans*. This software is able to automate the classification of TEs based on homology, structure, and target-site duplication.

A total of 400,000 trimmed Illumina reads (half from 0B genome and half from the 4B genome) were used to perform comparative analysis resulting on cluster abundance. We also conducted two more runs of reclustering in order to merge some connected clusters from the previous RE results, being able to identify new repetitive elements from Illumina reads.

In addition, we used dnaPipeTE (Goubert et al., 2015) for further annotation and assembly of TEs. This software produces precise estimates of repetitive DNA content and TE consensus sequences performing well on very low coverage sequencing. We performed two runs of this pipeline using as input a subset of 3,000,000 Illumina 0B paired-end reads and the same number of reads for the 4B Illumina library respectively.

Finally, we grouped together all the reference sequences of TEs acquired by means of RepeatExplorer and dnaPipeTE and reduce the redundancy of the database using CD-HIT-EST (Fu et al., 2012) with local alignment and greedy algorithm, and grouped those sequences showing 80% or higher similarity in at least 80% of length to get the final set of transposons sequences from the genome of *E. plorans*.

In order to deeply classify TEs, we performed similarity searches against several databases. In particular, these sequences databases comprised known repeats included RepeatMasker v4.0.5 (http://www.repeatmasker/org, Smit et al., 2013) and the set of sequences deposited in GenBank and RepBase (http://www.ncbi.nlm.nih.gov/genbank, http://www.girinst.org/repbase/). They were also analyzed for the occurrence of ORFs and particular structural features as tandem subrepeats, terminal inverted repeats (TIRs), potential long terminal repeats (LTRs), transposase and retrotransposase domains, using ORF Finder, BLAST (Altschul et al., 1997), HMMER (Finn et al., 2011) and Dotmatcher (Madeira et al., 2019).

Satellite DNA (satDNA)

For satDNA mining in the genome of *E. plorans* we made use of the satMiner protocol (Ruiz-Ruano et al., 2016; details in https://github.com/fjruizruano/satminer/) as explained in the Chapter 1 of this thesis. We selected 2x250,000 quality trimmed reads in a random manner with SeqTK (https://github.com/lh3/seqtk) from the 0B library and run RepeatExplorer with default options but adding a custom database of repetitive sequences (transposons sequences indicated above and other grasshoppers repetitive sequences, unpublished data). We perform 7 runs of RepeatExplorer until no more satDNA was identified in the 0B library, then we filtered out previously detected satDNAs from the 4B library and applied the same protocol 3 more times for the 4B library to get the most complete collection of satDNAs for *E. plorans*.

Mitochondrial DNA (mtDNA), ribosomal DNA (rDNA), histones, small nuclear RNA (snRNA) and transfer RNA (tRNAs)

We acquired the reference sequence of U1 (KJ606066) and U2 (KT963542) of *E. plorans* snRNAs deposited in GenBank. Then, we assembled U4, U5 and U6 snRNAs and mitochondrial DNA using MITObim (Hahn et al., 2013) with known seed sequences of each particular sequence coming from several species: *Peripolus nepalensis* (NC_029135) for mitochondial DNA and *Drosophila melanogaster* for U4 (NR_001670), U5

(NR_001933), and U6 (NR_002081) snRNA genes. Mitochondrial DNA of *E. plorans* was further annotated and characterized with MITOS (Bernt et al., 2013). The rDNA and the histone cistron were assembled starting at clusters of RE containing these elements and performing several rounds of manual assembly with *E. plorans* Illumina reads.

For tRNAs identification, we used the tRNAscan-SE program (Lowe and Eddy, 1997) with the total of overlapping reads from the 0B and 4B paired-end Illumina libraries. Then we removed redundancies with CD-HIT-EST applying the options -M 0 -aS 0.8 -c 0.8 -G 0 -g 1.

In silico identification of over-represented repeats in the 4B library

Newly characterized repeats by the above-explained approach were used as a reference for abundance comparison between the 0B and 4B libraies of *E. plorans*. For this purpose, all quality-filtered pairs of Illumina reads were aligned from each gDNA library (~0.5x and ~0.4x of genome coverage in the 0B and 4B libraries respectively) to the reference sequences by means of RepeatMasker. As the read files were too big to perform a single RepeatMasker, read files were split using a python custom script to run RepeatMasker separately for every subset of reads and achieving full results at the end of the process. Abundance of each repetitive element (measured in nucleotides) was normalized by nucleotides used for each analysis. Then, the log2 of the quotient between the abundance in the 4B and 0B libraries (gFC4B) was the tool to detect sequences located on the B chromosomes, a gFC>0 point out a sequence potentially located in the B chromosomes.

RepeatMasker also estimates divergence that, together with abundance data, make it possible to construct a repetitive landscape that is highly informative about performance of both parameters. Furthermore, using these data we constructed a subtractive landscape by comparing abundances of each repetitive sequence at certain divergence values in 4B and 0B libraries of *E. plorans*. This was a initial step to detect over-represented elements in the 4B library with respect to the 0B one, probably because they are contained in the B chromosomes.

As described in Ruiz-Ruano et al. (2018) we estimated the abundance of each repetitive element in a single B chromosome. For this calculation we used normalized abundance in 0B and 4B libraries indicated above. We assumed that the abundance of an element in the 4B genome is the abundance in the 0B genome plus the abundance in the

B chromosomes everything weighted by genome size.

We performed the same analysis of abundance in gDNA libraries of *E. plorans* from Tanzania, Egypt and Armenia including one individual harboring B chromosomes and another lacking them for each population.

We also searched for shared reads between different repetitive element in a long reads MinION library from the same 4B male of *E. plorans* sequenced previously by Illumina. The MinION library was annotated with RepeatMasker using as a reference the *de novo* database of *E. plorans* repetitive DNA as explained in the Chapter 1 of this thesis.

Finally, every repetitive element was named based on the class and family of repeats it belongs to, followed by a number indicating its order in decreasing abundance in the 0B library in respect to the rest of the elements in that repetitive family.

B-specific variation of repetitive DNA in *E. plorans*

We performed analysis of coverage variation along sequences of repetitive DNA potentially located in the B chromosomes of *E. plorans* after mapping of 0B and 4B gDNA libraries using SSAHA2 (Ning et al., 2001). The SSAHA2 output was processed to identify SNPs specific of 4B libraries thus belonging to repetitive sequence located in B chromosomes, as explained in Chapter 3 for protein-coding genes.

We also mapped 47 libraries of RNA to explore differences in repetitive DNA expression between lacking and carrier individuals belonging to both sexes, different developmental stages and tissues (embryos and adult leg and gonads) of *E. plorans*. Additionally, we quantified differences in expression of Alt (B specific variants) and Ref (sequence variants of As) to test whether B chromosomes are transcriptionally active concerning repetitive DNA.

Fluorescence *in situ* hybridization (FISH)

Probes for FISH were generated by PCR with MBL-Taq DNA polymerase (MBL002) from template DNA isolated from adult males of *E. plorans* originating from Torrox population and containing B chromosomes. The primers used are listed for tandem repeats in Table S1.6 (Chapter 1) and for the remaining repetitive classes in Table S2.9, they were designed from NGS sequences using Primer3 v0.4.0 (Untergasser et al., 2012). PCR conditions were: initial denaturation 5 min at 94 °C, 30 cycles of 30 sec at 94 °C, 30 sec at annealing temperature of primer pairs (usually between 57-62 ºC), and 2 min at 72 °C, and a final elongation of 7 min at 72 °C. PCR products were visualized in a 1.5% agarose

gel and then were cleaned with GenElute® kit PCR Clean-Up (Sigma). These probes were verified by Sanger sequencing by Macrogen (Seoul, Korea).

Preparations of mitotic chromosomes were performed from embryos according to Camacho et al. (1991) methodology that is focused on Meredith's technique (1969). Double FISH was performed as described in Cabrero et al. (2003b). We also carried out FISH in testis of individuals from Torrox, Fuengirola, Salobreña, Mundo, Valentín, Cameroon and Turkey that were fixed in 1:3 acetic:ethanol and stored at 4 °C. Testicular follicles preparations were obtained as described in Cabrero et al. (1999). Slides were dehydrated in a series of 70%, 90% and absolute ethanol and then incubated in an oven at 60 °C overnight. We used amplicons from the PCR experiments as probes.

As indicated in Montiel et al. (2012), PCR products were labelled through “nick translation” using 2.5 units of DNA polymerase I/DNAasa I (Invitrogen) and about 250 ng of DNA probe was used in each FISH experiment following the technique described in Cabrero et al. (2003). The fluorochromes used were tetrametilrodamina-5-dUTP and/or fluoresceina-11-dUTP (Roche), which yield red and green fluorescence, respectively. To track probe location on chromosomes we counterstained them with DAPI, and the slides were mounted in Vectashield (Vector, USA). For chromosome visualization, we used a BX41 epifluorescence Olympus microscope equipped with a DP70 cooled digital camera for photography.

Results

The repetitive DNA database of *E. plorans*

The graph based clustering of Illumina reads through RepeatExplorer followed by satMiner, dnaPipeTE, MITObim and tRNAscan-SE protocols allowed us to construct a comprehensive database of repetitive DNA from *E. plorans* genome composed of 1,726 sequences.

The repetitive database of *E. plorans* comprises the mitochondrial DNA (split in 39 entries belonging to each gene, see Figure S2.1a), the histone cluster (as 11 sequences from each gene and spacers that we were able to assemble, Figure S2.1c), the ribosomal DNA (7 sequences in Figure S2.1b plus the 5S rDNA), 5 sequences of snRNAs (U1, U2, U4, U5 and U6), 1,038 tRNAs sequences, 112 families of tandem repeats and 513 sequences of transposable elements (TEs).

We further classified TEs according to Wiker et al. (2007) in class I – retrotansposons (325 families), class II – subclass I – DNA transposons (136 families) and class II – subclass II – DNA transposons (30 families). Regarding retrotransposons, we found 158 families of LTR retrotransposons (av. length= 1,648 nt, 56 autonomous families) belonging to Bel (6), Copia (21) and Gypsy (131) superfamilies. On the other hand we assembled 155 sequences of non-LTR retroransposons (av. length= 2,103 nt, 50 autonomous families) of the superfamilies CR1 (8), Daphne (8), I (5), Jockey (7), Kiri (5), Nimb (9), Penelope (37), R1 (10), R2 (1), R4 (1), RTE (26), Tx1 (31) and Vingi (7); and 12 SINEs (av. length= 872 nt, all of them non-autonomous by definition). DNA transposons comprised 166 sequences (av. length= 1,211 nt, 46 autonomous families) included in the superfamilies Academ (7), EnSpm (1), Harbinger (2), hAT (38), ISL2EU (2), Kolobok (10), Mariner (52), MuDR (1), P (1), PiggiyBac (14), Sola (7), Transib (1), Helitron (14) and Polinton (16) (those two last families belonging to subclass II while the rest were subclass I of DNA transposons) (see Table S1 for details). We also found 22 TE sequences unclassified that were annotated as unknown transposons (av. length= 3,379 nt, all non-autonomous). For details about TEs classification see Table S2.1.

The 112 tandem repeat families (TRs) of *E. plorans* included RULs (repeats unit lengths) from 4 (EplTR096-4) to 455 nt (EplTR020-455) (av. length~156 nt) and the content in AT of consensus sequences ranged from 23.3% (EplTR084-30) to 84.2% (EplTR028-19) (av~60%). The 112 TR families were grouped in 13 superfamilies in terms of sequence identity higher than the 50% and lower than 80%, while sequences with higher than 80% of identity were considered as variants and were not included in the database.

The mitochondrial DNA of *E. plorans* was assembled completely and annotated as indicated in Figure S2.1a. However, the histones clusters and ribosomal DNA were incomplete in their IGS regions (Fig S2.1).

All newly identified repetitive sequences in the genome of *E. plorans* were combined in a reference database that was used to estimate the amount of repetitive elements in each genome (0B and 4B) and consequently in B chromosomes.

Abundance of repetitive elements in B chromosomes of *E. plorans*

The estimated proportion of repetitive genome with RE in the B-lacking individual of *E. plorans* was 64.9% whereas it was 65.6% in the 4B individual suggesting a high content of repetitive DNA in the B chromosomes of this species. However, the *de novo* repetitive

DNA database represented a 48.3% and 52.4% after per-sequence RepeatMasker analysis of the 0B and 4B gDNA of *E. plorans* respectively. This decline in repetitive DNA proportion results as a consequence of missing repetitive sequences in our database that we were not able to characterize, so the actual repetitive proportion of the *E. plorans* genome would be probably similar to that calculated by RE.

After RepeatMasker analysis of reference sequences in 0B and 4B libraries of *E. plorans* we determined which repetitive elements were over-represented in the 4B library (gF4B>0, gFC=log2(4B abundance/0B abundance)) after normalization by library and genome size. Additionally, we estimated the repetitive DNA proportion in a single B chromosome considering the genome sizes of the Bs and the A standard set of chromosomes (see Materials and methods). See Table S2.2 for details in gFCs, abundances values and proportion of each repetitive sequence in the Bs of *E. plorans*.

As expected after repetitive proportion estimated by RE in *E. plorans* individuals, the B chromosome of this species is highly enriched in repetitive DNA that represents a 86.3% of the B genomic content (see Figure 2.1c). In particular, satDNA and TRs make up a 65% of the B chromosomes whereas transposable elements represents the 15.7% of the Bs being retrotransposons the most abundant TEs in B chromosomes of *E. plorans*. The 5.4% of the Bs comprises other repetitive sequences (histones, mtDNA, rDNA, snRNAs and tRNAs) but the ribosomal DNA was markedly the most abundant of them (5%). The remaining proportion of the B chromosomes (13.7%) was reserved to single copy genes or missing repetitive sequence in our *de novo* database (Figure 2.1c). We could assume that Bs are depleted in TEs and enriched in satDNA and TRs compared to the standard set of chromosomes where the proportion of these elements are 37.6% and 8.5% respectively, even so, LINEs are between the most abundant TEs in the As and also in B chromosomes (see Fig 2.1c). However, the repetitive DNA composition of B chromosomes correlates positively with the one of the As (r_S= 0.5862, p < 2.2e-16) which could suggest an intragenomic origin of B chromosomes.

This subtractive approach yielded 1,088 B-located repetitive sequences (gFC4B>0), 80 of which were satDNAs/TRs and 454 sequences were TEs (118 families of DNA transposons, 126 LINEs, 150 LTR retrotransposons, 10 SINEs, 16 Polintons, 13 Helitron transposons and 21 unknown TEs). We found also 500 tRNAs, 3 snRNAs and the sequences corresponding to different parts of histones (11 sequences), mtDNA (32 sequences) and ribosomal DNA (8 sequences). Interestingly, a high number of repetitive

DNA sequences included in the reference database (~ 63% of all sequences) were present in the B chromosomes of *E. plorans*, being tRNAs practically the main class of repetitive DNA lacked in B chromosomes (~ 84% of sequences absent in Bs).

Despite the high number of TR and satDNA families found in B chromosomes of *E. plorans*, only two of them, EplTR001-180 and EplTR002-196, are enough to take up more than a half B chromosome (21.5% and 34.3% of the B chromosome content respectively). EplTR005-49 and EplTR006-147 were also abundant in the B chromosomes of *E. plorans* (2.8% and 4.7% respectively). The family EplTR005-49 was located in the centromeric regions of all chromosomes, thus being the best candidate to perform as the centromeric satDNA in this species (see Chapter 1). Interestingly, this centromeric satDNA is located also interstitially in the B chromosomes of *E. plorans*. Despite their moderate abundance in the B chromosomes of *E. plorans*, we found two satDNAs families, EplTR106-323 and EplTR112-11, that showed high gFCs values in 4B male of *E. plorans* and the latter family was almost inexistent in 0B individuals. FISH analysis of EplTR106-323 and EplTR112-11 yielded specific signals on B chromosomes (see Chapter 1 and Fig. 2.2), so they could be used as B chromosome markers as it was done by Cabrero et al. (2017) using the family EplTR112-11 to track the elimination of B chromosomes during the spermiogenesis of *E. plorans*.

Regarding TEs located in the B chromosomes, the most abundant superfamily of LINEs was RTE transposons (2.3% of the B chromosome), Gypsy was the most represented superfamily of LTR transposons in B chromosomes (4.7%) whereas hATs were the DNA transposons making up the highest proportion in Bs (1.1%). Polinton and unknown transposons were also present in B chromosomes of *E. plorans* surpassing the 1% of their content (1.2% and 1.7% respectively). Among the specific TE families showing a high gFC in the 4B male of *E. plorans,* we found LINE/Daphne_8, LINE/R2_1, LINE/RTE_26, LINE/R1_10, SINE_08 and piggyBac_04. These elements were selected for subsequent FISH analysis together with some others.

Regarding the remaining repetitive classes (histones, mtDNA, snRNAs, rDNA and tRNAs), only the ribosomal DNA represented more than the 1% of the Bs, in particular the 5%, as stated above. In particular, the 28S gene and the IGS were the most abundant rDNA regions in B chromosomes of *E. plorans* (1.2% and 2.8% respectively).

Interestingly, we found a negative correlation between Kimura divergence of each reference sequence in the 4B library and its gFC in the 4B individual of *E. plorans* (r_S=

-0.142009, p= 4.442e-09). This correlation was also negative compared gFC and Kimura divergence in the 0B library (r_S= -0.1014057, p= 2.929e-05) although it was lower than the one considering divergence in the 4B library. In addition, Kimura divergence was lower in average in the 4B library than in the 0B one for sequences located in the B chromosomes (Kdiv0B_av.= 9.1, Kdiv4B_av.= 8.8; excluding tRNAs) and these differences were significant considering sequences with gFC>1.58 (W = 42, p-value = 0.02622, Kdiv0B_av.= 16.2, Kdiv4B_av.= 7.5). These results suggest the amplification of particular variants of repetitive sequence in the B chromosomes of *E. plorans*.

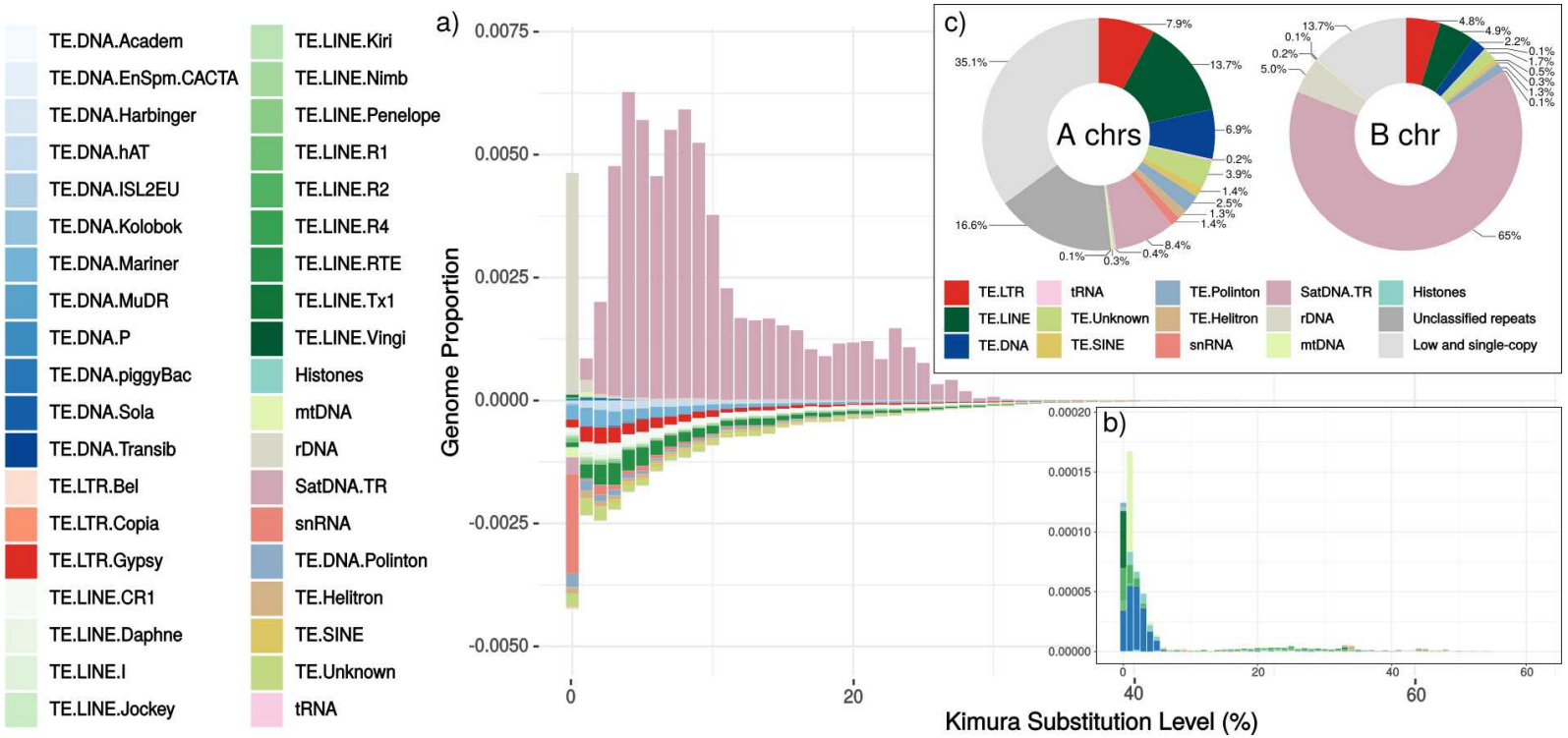


Figure 2.1. a) Subtractive landscape of repetitive DNA between the 4B and 0B genomes of *E. plorans*, positive values show repetitive classes and superfamilies over-represented in the 4B library, thus potentially located in the B chromosomes (i.e. satDNA/TRs or rDNA. b) Positive values of the same subtractive landscape but excluding satDNA/TRs and the ribosomal DNA which hide the presence of TEs such as piggyBac DNA TEs or R1, R2 and Tx1 LINEs. c) Proportion of the different repetitive DNA classes in each B chromosome of *E. plorans* (right) and in the standard complement (As). Note the high content of satDNA/TRs (65%) in the B chromosomes followed by the ribosomal DNA (5%) and LINE TEs (4.9%).

FISH of repetitive DNA potentially located in B chromosomes

SatDNA and TRs

As indicated in the Chapter 1 of this thesis we performed FISH for the 112 TR families found in *E. plorans* (see Dataset 2.1) although the B chromosomes of this species only contains 80 out of the 112 TR families. FISH showed signal for 51 out of the 80 B-located TRs but only 12 of them displayed evident signals in B chromosomes (well-defined bands or dots) whereas the remaining 39 TR were scattered in Bs. We found significant differences in terms of gFC (KW chi-squared= 17.06, df = 2, p= 0.0001975) and B proportion (KW chi-squared= 12.072, df= 2, p= 0.002391) between TRs yielding well-defined signals in B chromosomes (12 TR families, gFC4B av.= 1.917, Bprop av.= 0.053), dispersed patterns in Bs (15 families, gFC4B av.= 0.056, Bprop av.= 9.725e-7) and failing to produce FISH signal (53 families, gFC4B av.= 0.234, Bprop av.= 2.391e-4). Those TRs showing FISH signal, especially producing bands or dots, showed higher gFCs and B proportion than those scattered or non-signal TRs, thus FISH results are in concordance with the abundance of TRs found in B chromosomes through bioinformatic methods.

TRs yielding clear FISH signals showed the characteristic structure of tandem repeats, appearing concentrated in small regions or bands on chromosomes. 7 TR families out of the 12 producing well-defined FISH signals (EplTR001-180, EplTR002-196, EplTR006-147, EplTR019-75, EplTR071-180, EplTR087-15 and EplTR109-344) were located in the three heterochromatin bands described in the B chromosome variant, B24, of *E. plorans* found in Torrox. This FISH pattern is similar to the one showed by the well-known sat180bp initially found in B chromosomes of *E. plorans* (López-León et al, 1994; Cabrero et al., 2003b), in fact, the family EplTR001-180 represents this satDNA. The EplTR015-351 appeared as a band in the centromeric region of B chromosomes and their short arm, but also it yielded little dots located interstitially in this chromosome. The family EplTR007-246 was completely interstitial in B chromosomes. As expected, we found the centromeric satellite EplTR005-59 in the centromere region of Bs but, surprisingly, it was also located in the whole short arm and interstitially in the B chromosomes (Table 2.1).

In order to test if the centromeric satellite EplTR005-49 was also present in other variants of B chromosomes of *E. plorans*, we performed FISH on testicular follicles of *E. plorans* from different populations: Torrox (B24), Fuengirola (B5), Salobreña (B2), Mundo (B1), Valentín (0B), Cameroon (0B) and Turkey (0B). As expected, EplTR005-49 was

located in the centromeric and paracentromeric zones of all chromosomes (Fig. S2.2) and also in the small arm of B chromosomes (B1, B2, B5).

Finally, the families EplTR106-323 and EplTR112-11 showed specific signal on B chromosomes so they could behave as B chromosome markers. Both TRs were interstitial in B chromosomes but yielding different patterns and showed a band in the centromeric and pericentromeric regions. However, the EplTR112-11 showed a great band in the distal part of B chromosomes (Table 2.1).

Transposable elements (TEs)

We carried out FISH experiments for mobile elements belonging to each TE superfamily that in light of the bioinformatic analysis were over-represented in the B chromosomes of *E. plorans,* 32 TE probes in total (see Table S2.3). Most of these TEs, 18 families, failed to yield FISH signals or it was scattered and scarce in all chromosomes. In addition, we found that 6 TE families were dotted in all A chromosomes and, whereas three of them were rare in Bs (LTR/Bel_2, LINE/R1_02 and LINE/R1_10), the remaining three TE families, all belonging to the piggyBac superfamily, were concentrated in B chromosomes (DNA/piggyBac_01, 04 and 08). Both DNA/piggyBac_08 and LINE/R1_10 accumulated interstitially and in the distal part of the largest pair of chromosomes respectively (pair 1 sorted by size) although LINE/R1_10 appeared only in one chromosome of the pair (Table 2.1).

The B chromosomes of *E. plorans* contained well-defined interstitial regions of DNA/Kolobok_02, LINE/Daphne_8, LINE/RTE_26, LINE/Tx1_01 and LINE/Tx1_05. Interestingly, FISH signals for LINE/Daphne_8, LINE/Tx1_01 and LINE/Tx1_05 were apparently B-specific however abundance estimates in the 0B library showed that they were also present in A chromosomes (Table 2.1).

Furthermore, three TE families (DNA/hAT_11, LINE/Tx1_22 and SINE_08) form clusters in Bs and A chromosomes. The retrotransposon SINE_08 appeared dispersed in most chromosomes but concentrated on some rDNA-containing chromosomal regions, such as the distal end of B chromosome and the paracentromeric band of chromosome 9. This sequence is also located in the distal region of chromosome X but rDNA is near the centromere of this chromosome (López-León et al., 1994), hence these two elements were not in the same region of the Xs. The location of hAT_11 in B chromosome was similar to that of SINE_08 however it was located in chromosomes 10 and X, in all cases

in rDNA regions. Finally, LINE/Tx1_22 appeared in centromeric regions of all chromosomes (Bs included) and in the short arms, pericentromeric region and interstitially in the B chromosomes (see Table 2.1 for details).

Other repetitive elements

We also performed FISH for two spacer of the histone cluster (spacer 1 and 3), the snRNA/U2, four tRNAs and 28S gene of the rDNA. All these element, but histone spacers and rDNA, did not produce any FISH signa. Histones spacer showed a band in the distal part of the second pair of chromosomes but failed to yield signals in Bs. The 28S gene was expectedly found in the rDNA regions of *E. plorans* as previously described by Cabrero et al. (2003a) including the distal part of B chromosomes (Table 2.1 and S2.3).

Possible origin of B chromosomes from chromosome 9

As most of the repetitive elements located in B chromosomes of *E. plorans* were also found in A ones, we explore the possibility that there were a good candidate among the A chromosomes from which the B chromosomes of *E. plorans* could have arisen. For that purpose, we identified and counted the number of A chromosomes yielding FISH signal for B-located repetitive sequences (identified bioinformatically by their abundance in 0B and 4B libraries). The chromosome 9 was the one containing more of these B-located sequences, 24 repetitive elements in total, followed by the chromosome 8 sharing 22 repetitive elements with the B chromosome of *E. plorans*. Then, we restricted this counting exclusively to repetitive sequence yielding FISH signals in Bs. Likewise, the chromosome 9 shared the highest number of repetitive sequences with the Bs, 16 elements, followed by the X chromosomes with 13 elements in common with the B chromosomes. In addition, three TR families (EplTR009-300, EplTR015-351 and EplTR032-439) were found specifically in 9 and B chromosomes and there were no other chromosomes sharing three or more specific sequence with the Bs (Table 2.1). These results put a spotlight on the chromosome 9 as the main source for B chromosome origin.

Table 2.1. Genomic proportion and Kimura divergence (%) of some B-located repetitive elements in 0B and 4B *E. plorans* males from Torrox (Spain) sorted decreasingly by their proportion in B chromosomes. We include those B-located sequences (gFC4B>0) that yielded FISH signal, thus excluding those showing NS pattern (for extended version see Table S2.3). See details about FISH signal such as pattern (Patt.; B= banded, D= dotted and DB= dotted-banded), number of A pairs showing signal (NA), chromosome location (1, 2, 3,…, B; p= pericentomeric, i= interstitial, d= distal, *= retrieved from Montiel et al., 2014, i3= reference to the three heterochromatic bands described in the B24 chromosomes of *E. plorans*). Find indicated in the last column the ID of the B-located repetitive element that shared at least one MinION reads with certain chromosomes-specific element.

ID	gDNA Abundances						FISH														
	Male 0B		Male 4B		gFC 4B	B prop	Patt.	N A	1	2	X	3	4	5	6	7	8	9	10	11	B	Shared MinION B?
	Prop.	Div	Prop	Div																	
EplTR002-196	1.72E-02	13.56	5.44E-02	13.3	1.840	3.43E-01	B	9	p	p	p	p	p	p			p	p		p	p,i,i,i3	
EplTR001-180	2.15E-02	5.93	4.36E-02	5.67	1.196	2.15E-01	B	9	p	p	p	p	p	p	p			p		p	p,i,i,i3	
EplTR006-147	2.86E-03	6.05	7.92E-03	6.03	1.646	4.71E-02	B	10	p		p	p	p	p		p	p	p	p	p	p,i,i,i3	
EplTR005-49	7.19E-03	14.87	9.61E-03	15.3	0.595	2.84E-02	B	12	p	p	p	p	p	p	p	p	p	p	p	p	p,i,i	
Eplo_ribo/28S	1.09E-03	0.76	2.39E-03	0.68	1.306	1.24E-02	B	7			p				i	i	i	p	p	p	p	
EplTR003-159	8.79E-03	6.59	8.61E-03	6.61	0.147	7.29E-03	B	10			p	p	p	p	p	p	p	i	p	p		
TE_SINE_08	2.18E-04	7.47	3.48E-04	5.31	0.853	1.36E-03	B	2			d							p			p	
TE_DNA/piggyBac_01	3.28E-04	11.21	4.26E-04	8.7	0.552	1.18E-03	D	12													all dotted	
EplTR007-246	2.53E-03	5.64	2.36E-03	5.62	0.076	1.06E-03	B	8					p	p	p	p	p	p	p,i	p	i,i	
EplTR106-323	2.43E-06	10.19	1.19E-04	0.69	5.788	1.02E-03	B	0													p,i,i,d	
TE_DNA/piggyBac_04	1.59E-04	3.73	2.53E-04	3.63	0.847	9.82E-04	D	12													all dotted	
TE_DNA/hAT_11	4.05E-04	10.15	4.63E-04	9.01	0.371	9.18E-04	B	2			p								p		p	
TE_DNA/Kolobok_02	3.73E-04	1.83	4.33E-04	1.63	0.390	8.96E-04	DB	12													i,i,i	
EplTR049-367	9.72E-05	4.3	1.86E-04	3.72	1.116	8.77E-04	B	1									i					SINE_08

EplTR015-351	5.06E-04	11.88	5.43E-04	10.15	0.277	8.28E-04	DB	12										p				p,i,i		
TE_LINE/R2_1	3.83E-05	14.79	1.20E-04	12.43	1.826	7.55E-04	D	0														i *		
EplTR025-238	2.05E-04	14.5	2.63E-04	13.41	0.531	7.07E-04	B	5	p	p,i		d		p,i		i								
TE_DNA/piggyBac_08	9.35E-05	6.33	1.34E-04	4.77	0.692	4.45E-04	DB	1	i													all dotted	hAT_11	
EplTR019-75	4.33E-04	9.41	4.22E-04	7.84	0.136	3.32E-04	B	11	p		p	p	p	p	p	p	p	p	p	p	p	p,i,i,i3		
EplTR014-285	5.66E-04	4.75	5.37E-04	4.87	0.098	3.07E-04	B	7.5					p	p	p	p	p	p	p	i				
EplTR112-11	1.81E-08	30.88	3.49E-05	9.1	11.090	3.05E-04	B	0														sa,p,i,i,d		
EplTR073-297	3.94E-05	6.7	6.88E-05	7.07	0.977	2.96E-04	B	1									p						No	
EplTR065-233	4.92E-05	18.88	7.51E-05	16.71	0.787	2.76E-04	B	1						i									No	
EplTR017-67	4.92E-04	4.38	4.63E-04	4.35	0.088	2.39E-04	D	12														scarce		
EplTR008-426	1.47E-03	14.3	1.33E-03	14.35	0.030	2.39E-04	B	1									i					scarce	Tx1_01	
EplTR044-210	1.07E-04	14.41	1.22E-04	15.37	0.365	2.38E-04	B	1												d			No	
EplTR056-209	7.65E-05	16.03	9.40E-05	14.74	0.473	2.30E-04	B	0.5										i					No	
TE_LINE/Tx1_01	8.95E-05	19.55	1.05E-04	15.22	0.404	2.24E-04	D	12														i	--	
TE_LINE/Tx1_05	5.24E-05	18.21	6.96E-05	12.63	0.583	2.02E-04	D	12														i,i	--	
Histones/spacer3	1.26E-04	2.03	1.35E-04	2.09	0.270	2.01E-04	B	1		p													No	
EplTR079-166	2.54E-05	6.01	4.42E-05	5.25	0.975	1.90E-04	B	2			i							i					No	
EplTR009-300	1.32E-03	8.45	1.19E-03	8.53	0.026	1.87E-04	DB	1										i,i				scarce	Tx1_01, Tx1_05, SINE_08, Hat_11, Kolobok_02	
Histones/spacer1	1.30E-04	1.9	1.33E-04	1.89	0.211	1.58E-04	B	1		p													No	

EplTR020-455	3.74E-04	5.5	3.47E-04	5.41	0.065	1.34E-04	D	12														all dotted	
TE_LINE/RTE_26	1.16E-05	13.27	2.44E-05	6.25	1.247	1.24E-04	D	12														i,i	No
EplTR024-391	2.28E-04	9.82	2.16E-04	9.77	0.098	1.23E-04	D	12														scarce	
EplTR040-113	1.20E-04	2.1	1.17E-04	2.06	0.136	9.18E-05	D	12														abundant	
EplTR013-320	5.94E-04	17.97	5.36E-04	18	0.028	9.14E-05	D	12														scarce	
EplTR066-252	4.78E-05	14.1	5.28E-05	11.71	0.319	9.14E-05	B	9	p		p			p	p	p	p	p	p	p			
EplTR038-17	1.26E-04	10.1	1.21E-04	10.22	0.112	7.84E-05	D	12														abundant	
EplTR032-439	1.52E-04	8.08	1.43E-04	7.85	0.091	7.66E-05	DB	0.5										i				scarce	No
EplTR018-75	4.46E-04	10.98	4.04E-04	10.96	0.032	7.64E-05	D	12														scarce	
TE_LINE/Daphne_8	2.66E-06	27.37	1.05E-05	5.97	2.157	7.12E-05	DB	12														i	No
EplTR023-301	2.47E-04	12.9	2.26E-04	13.04	0.050	6.79E-05	D	12														scarce	
EplTR063-239	5.33E-05	7.31	5.47E-05	7.72	0.215	6.61E-05	B	1					d										No
EplTR054-269	8.49E-05	4.31	8.19E-05	4.64	0.122	5.79E-05	B	2										p		p			
TE_LINE/R1_06	1.94E-05	24.49	2.35E-05	20.68	0.456	5.57E-05	D	0														scarce	
TE_LINE/R1_10	7.21E-06	12.63	1.22E-05	8.13	0.938	5.10E-05	D	1	i													scarce	hAT_11
EplTR074-241	3.21E-05	7.54	3.40E-05	7.6	0.256	4.82E-05	B	1									d						No
TE_LINE/Tx1_22	1.14E-05	26.85	1.55E-05	19.33	0.618	4.72E-05	B	12	p	p	p	p	p	p	p	p	p	p	p	p	p,i,i	
EplTR036-240	1.38E-04	2.79	1.27E-04	2.75	0.055	4.17E-05	B	1							p								No
EplTR053-401	8.89E-05	8.32	8.34E-05	8.22	0.082	4.03E-05	B	1									i						No
TE_LTR/BEL_2	1.46E-05	8.13	1.72E-05	7.82	0.410	3.73E-05	D	0														scarce	
EplTR071-180	4.14E-05	4.31	4.05E-05	4.28	0.144	3.35E-05	B	8	p	p	p	p	p	p	p			i				p,i,i,i3	
EplTR050-124	9.69E-05	11.59	8.89E-05	11.46	0.051	2.69E-05	DB	12	p	p	p	p	p	p	p	p	p	p,i	p	p		all	

EplTR064-226	5.09E-05	3.89	4.80E-05	4.26	0.090	2.53E-05	B	1								p								No
EplTR088-238	1.90E-05	8.8	1.93E-05	9.1	0.202	2.20E-05	B	1			i													No
EplTR087-344	1.93E-05	8.93	1.91E-05	8.45	0.160	1.76E-05	DB	12	p	p	p	p	p	p	p	p	p	p,i	p	p	p,i,i,i3?		
EplTR021-165	2.93E-04	6.9	2.61E-04	7.01	0.009	1.47E-05	DB	0.5						p							scarce	Tx1_05	
EplTR035-239	1.45E-04	10.7	1.30E-04	10.79	0.016	1.22E-05	B	2	p,i	i,i	d	i	p	p										
EplTR103-242	7.22E-06	9.16	7.69E-06	6.74	0.267	1.14E-05	DB	1.5							i		i							
EplTR090-40	1.75E-05	15.81	1.68E-05	15.94	0.116	1.13E-05	B	6							p	p	p	p	p	p				
EplTR030-267	1.67E-04	15.73	1.49E-04	15.07	0.011	9.86E-06	B	5						p		i	p	i		p				
EplTR048-249	9.78E-05	9.6	8.74E-05	9.45	0.013	6.76E-06	D	12													all dotted			
EplTR034-123	1.50E-04	6.12	1.33E-04	5.97	0.008	6.33E-06	D	12													scarce			
EplTR077-47	2.84E-05	11.89	2.55E-05	11.51	0.022	3.41E-06	B	1						p										No
EplTR089-226	1.80E-05	12.5	1.63E-05	12.26	0.030	2.91E-06	B	2				i					i							
EplTR083-238	2.15E-05	10.53	1.92E-05	10.02	0.012	1.34E-06	B	1.5				i				i								
EplTR109-15	8.62E-07	5.94	7.97E-07	5.24	0.063	3.00E-07	DB	12	p	p	p	p	p	p	p	p	p	p	p	p	p,i,i,i3?			

To further characterize the relationships between A and B chromosome to elucidate the possible origin of Bs in terms of repetitive DNA, we analyzed which repetitive elements yielding specific FISH signals in one chromosome (Bs included) were found in the same long MinION read of *E. plorans*. This would be assumed as signs that at least, in that case, both elements are located in the same chromosome. Interestingly, we found that one TR family specific of chromosome 9, EplTR009, shared several long reads with some B located transposable element validated by FISH (DNA/hAT_11, DNA/Kolobok_02, LINE/Tx1_01, LINE/Tx1_05 and SINE_08), for details see Fig 2.2 and Table S2.3. In some of these cases, the size of the TR array was higher than 1,000 nt, which is the minimum target for FISH proposed by Schwartzacher and Heslop-Harrison (2000), this implies that these particular reads are located in chromosome 9, the chromosome where we found exclusively the FISH signal. We performed the same analysis considering repetitive elements yielding specific FISH signals in other A pairs as in chromosome 1 (for DNA/piggyBac_1, LINE/R1_10), chromosome 2 (histones), chromosome 4 (EplTR063-239), chromosome 5 (EplTR021-165, EplTR065-233 and EplTR077-47), chromosome 6 (EplTR036-240), chromosome 7 (EplTR064-226), chromosome 8 (EplTR008-426, EplTR049-367, EplTR053-401, EplTR073-297 and EplTR074-241) and chromosome 11 (EplTR044-210). However, in most of these cases we did not find reads shared with B-specific elements except for some elements specific to chromosomes 1, 2, 5 and 8 (Table S2.3) and still some of these reads were also shared by EplTR009 (specific of chr. 9). This was so especially in the case of hAT_11-containing reads, so we could not discard that these reads came also from the chromosome 9.

All these results point to the chromosome 9 as the likely principal contributor to B chromosome origin in *E. plorans*, although other chromosomes would also be involved in the building of the B chromosome.

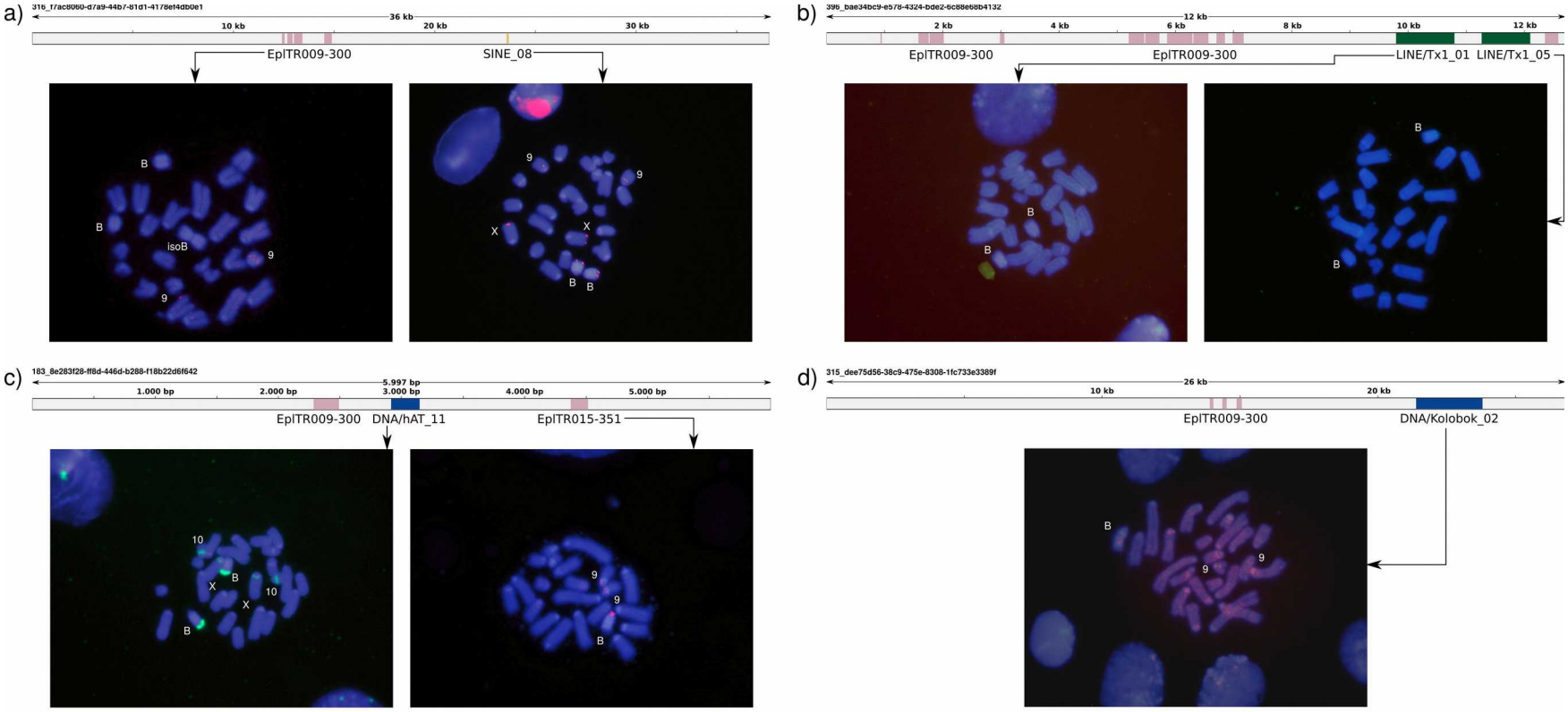


Figure 2.2. Examples of *E. plorans* MinION reads containing the chromosome 9 specific EplTR009-300 family and different TEs yielding FISH signals in the B chromosomes, the chromosome location of each repetitive element is shown. a) MinION long reads including the EplTR009 and TE_SINE_08 including their chromosome location. b) Example of read containing the EplTR009-300 and the LINEs Tx1_01 and Tx1_05, FISH results are shown for both TEs. c) MinION read shared by EplTR009-300, TE_DNA/hAT_11 and EplTR015-351 (located in Bs and in chr.9). d) Read including the chr. 9 specific TR and the TE_DNA/Kolobok_02.

Repetitive DNA in B chromosomes of *E. plorans* from African and Asian populations

We sequenced the gDNA of two males of *E. plorans*, one male carrying B chromosomes and the other lacking them, belonging to three populations located in Tanzania, Egypt and Armenia (six individuals in total). Then, we mapped these gDNA libraries against the *de novo* database of *E. plorans* to ascertain the repetitive DNA composition of the B chromosomes from individuals in these populations.

The B-specific TRs, EplTR106-323 and EplTR112-11, found in B chromosomes of *E. plorans* from Torrox (Spain) were also present in the Bs of this species from the three African and Asian populations. In fact, the EplTR112-11 family was completely absent in the As of the male from Tanzania and Armenia but in the latter population its abundance was low also in the +B genome (see Table S2.4). We found that the proportion of satDNA and TRs in the B chromosomes of males from Africa and Asia was quite poorer than that in Bs from Torrox, especially in Armenia, which is in concordance with the low amount of the sat180bp found by Cabrero et al. (2003b) in B chromosomes from this population. In contrast, in the B chromosomes of *E. plorans* from Tanzania there are plenty of transposable elements and TRs but they are depleted in rDNA whereas in Egypt the amount of rDNA in Bs increases and there is a reduction in the proportion dedicated to TE/LINES and TRs (see pie charts in Figure S2.3). The B chromosomes of the male from Armenia present a markedly different composition from that of Bs in Torrox. We were able to detect a very few proportion of total repetitive DNA in Bs from Armenia which implies that our *de novo* repetitive database, built from Torrox reads, does not include a great part of repetitive DNA from Armenian individuals or that we are missing repetitive elements that are abundant in the Bs from Armenia.

Interestingly, there was a significant positive correlation between gFC of repetitive elements found in B chromosomes from Torrox (excluding tRNAs) and the gFCs of those elements in males of Tanzania (r_S= 0.116, p= 0.00492) and Egypt (r_S= 0.270, p= 3.471e-11). However, this correlation was not significant between gFC in +B males from Torrox and Armenia (r_S= 0.047, p= 0.2562). The positive correlation was maintained between the gFC of individuals from Tanzania and Egypt (r_S= 0.417, p< 2.2e-16), Tanzania and Armenia (r_S= 0.507, p< 2.2e-16) and Egypt and Armenia (r_S= 0.646, p< 2.2e-16). These results indicate that the B chromosome of *E. plorans* from African populations, in particular from Egypt, resembles better the Bs from Torrox in Spain that those of males from Armenia (see gFC

scatter plots in Fig. SF3) and they still point to a common origin of B chromosomes in the species.

We also performed SNP calling of B-located sequences in these populations and the amount of B-specific SNPs found in Tanzania was slightly higher than in Egypt and Armenia. In particular, we found 2,483 SNPs in B-located repetitive elements in the male from Tanzania whereas 1,975 and 2,144 were found in B-sequences of males from Egypt and Armenia respectively (see Table S2.9). This could suggest an older B chromosomes in Tanzania with respect other two populations which is in concordance with results of phylogenetic signal of repetitive elements (see below and Figure 2.3)

Finally, we performed a phylogenetic analysis of males from each population in terms of abundance of each repetitive DNA element found in Bs of all populations (Figure S2.3), thus focusing in the history of B chromosomes and avoiding random homoplasy. For that purpose, we mapped libraries against our repetitive DNA database using RepeatMasker and normalized data by genome and library sizes. We used the species *Heteracris adspersa* as an outgroup. The gDNA of *H. adspersa* was extracted from the hind leg as indicated in the Materials and methods section and sequenced in Novogene Bioinformatics Technology Co., Ltd (Headquarters) through a Illumina HiSeq X platform as for Armenian and Egyptian libraries.

The resulting tree showed that *E. plorans* from Torrox appeared more recently than that of the African and Asian populations. Moreover, the population of Tanzania seems to be ancestral with the 1B male in the base of the tree. This result, together with the high correlation between gFCs in Torrox and Tanzania population, could highlight a possible origin of Bs in African ancestor from Tanzania. However, deeper analysis to confirm this hypothesis should be done as here we count on few *E. plorans* individuals and for some of them we do not know exactly the number of Bs they harbored (i.e. Egypt and Armenia). In addition, as recommended by Martín-Peciña et al. (2019), it would be better to perform phylogenetic analysis of repetitive DNA abundance after characterizing the repetitive DNA content of each species/populations.

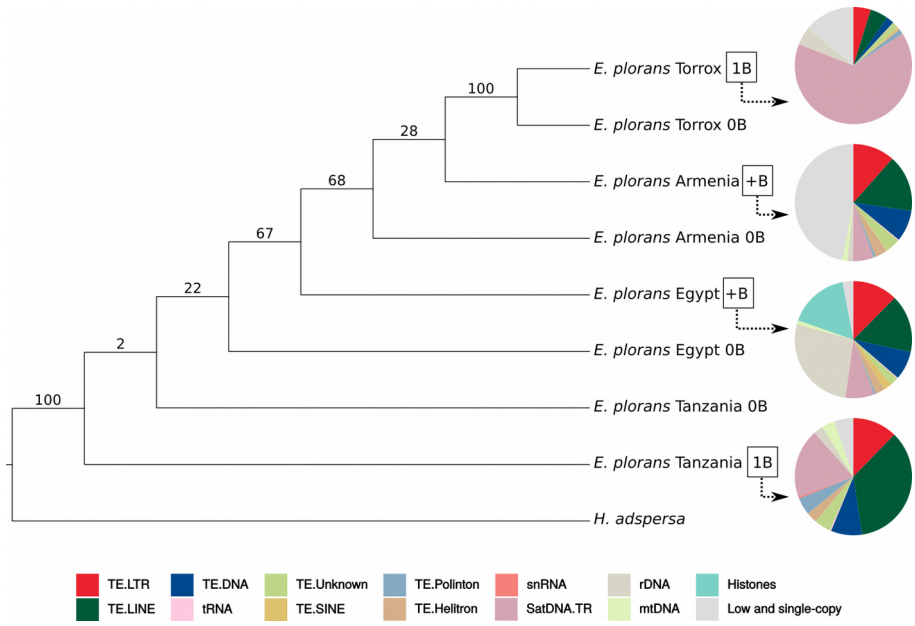


Figure 2.3. Phylogenetic tree of 0B and +B males of *E. plorans* from different populations (Spain, Armenia, Egypt and Tanzania) using abundances of B-located repetitive elements. Bootstrap support of each node is specified on the tree. Note that the 1B male from Tanzania appeared basal in the tree one which could suggest a first origin of Bs in Africa. See also the pie charts showing the proportion of repetitive classes in one B chromosome from the four populations of *E. plorans* analyzed here.

Transcriptional activity of some repetitive elements from the B chromosomes

As a first approach to identify which repetitive elements were transcriptionally active in 1B samples we mapped the different RNA libraries we sequenced (0B and 1B female and male embryos from two different pods, and leg and gonads of adults individuals lacking Bs and carrying 1B belonging to both sexes) against the *de novo* repetitive DNA database of *E. plorans* and then we calculated the tFC as log2(1B/0B) for each repetitive element in each of the biological samples we got. Considering counts of all RNA libraries in average, several repetitive elements located in B chromosomes of *E. plorans* also showed a positive tFC so they were transcriptionally active in 1B libraries. Furthermore, it was especially in adult gonads where we found a slightly higher tFC of B-located repetitive elements than in the rest of the samples (see Fig. SF2.4).

Then, we searched for SNPs showing variants being specific to B-carrying individuals (not found in 0B genomes or transcriptomes) to address whether the high expression of

repetitive elements in 1B samples came directly from the B chromosomes. We named Ref the reference variant present in all individuals and Alt the B-specific alternative variant, however we could not discard that some Ref variants could be also located in B chromosomes. We calculated the quotient between Alt and Ref variants from 1B and 0B RNA samples respectively for each repetitive element and divided that value by the quotient between of genomic abundance of both variants to obtain the transcription intensity (TI) of that repetitive element in the B chromosomes.

We found 15,162 B-specific SNPs in a total of 543 repetitive elements of *E. plorans* (Table S2.5), however only 303 B-located repetitive elements showed transcription of Alt variants in some of the RNA samples and 94 out of them did it with a TI higher than 1 (Fig. 2.4b). Taking into account the total repetitive element of each class in average, there was a huge shift in transcription activity in 0B samples (from TI of Ref variants) compared to the 1B-carrying ones (from TI of Alt variants). Expression of mtDNA was stood out in 0B samples with the 68.5% of the total TI for repetitive elements, followed by the rRNA contributing with a 5.6% of TI (see Fig. 2.4a left). In contrast, we found transcription of a mixture of repetitive elements from the B chromosomes in the 1B transcriptomes (Fig. 2.4a right) with a predominance of tRNAs (20.5%) and Polinton transposable elements (19.2%). However, this averaged TI of Alt variants was lower than 1 for all repetitive DNA classes, thus quite weak compared to the averaged expression of Ref variants by class (Table S2.6). The TI of each repetitive elements and RNA sample separately showed that gonads were the tissue in which there were more transcriptional activity of B-repetitive element compared to embryos of both pods and adult legs. Several families of transposable elements from the classes DNA/hAT, DNA/Polinton, LINE/Penelope, LINE/Tx1, and LTR/Gypsy surpassed the TI of 1 in several RNA samples and also some elements of other repetitive DNA classes did it (for details in particular elements see Fig. 2.4b and Table S2.7).

In addition, we found TI in Alt variants higher than 1 for 27 and 33 repetitive elements in 1B samples of ovary and testis respectively, being this number lower for legs (13 and 17 elements in female and male legs respectively) and embryos from pod 1 (13 and 14 elements in female and male embryos respectively) and pod 2 (with 10 elements in females and 17 in males). This result indicates that it is in gonads of *E. plorans*, especially in testis, where the repetitive DNA from B chromosomes is being transcribed more intensely.

Figure 2.4. a) Proportion of transcriptional intensity (TI) from each repetitive DNA class for Ref variants in 0B samples (left) and Alt variant in 1B samples (right), note the change in transcription activity in 1B samples. b) Transcriptional intensity of different repetitive DNA families in 1B RNA samples of *E. plorans*, only those families surpassing the TI of 1 are included in the graphic (i.e. 93 families).

Discussion

The results obtained here, in the frame of massive sequencing, bioinformatics and molecular cytogenetics, shed new light on an issue still without revealing: the molecular content of B chromosomes of the grasshopper *Eyprepocnemis plorans*. Until now, the only repetitive DNA elements known to be located in these B chromosomes were the ribosomal DNA, located distally, and the tandem repeat sat180bp whose location matches the heterochromatin bands of Bs (López-León et al., 1994). The vast majority of B chromosomes in this species, analyzed in many circum-Mediterranean populations have shown these two DNAs (Cabrero et al., 1999; López-León et al., 2008). In the present work we extended the number of repetitive sequences located in Bs and depicted the repetitive landscape of the B chromosomes of *E. plorans* in a quantitative way.

We confirm in *E. plorans* the repetitive nature proposed for B chromosomes (Camacho et al., 2005) whose Bs are made up with a 86.3% of repetitive DNA. Moreover, we reinforce also the heterochromatic view of B chromosomes in this species (Henriques-Gil et al., 1984) with TR families representing a 65% of their molecular composition. As expected, the sat180bp (here EplTR001-180) occupies a high proportion of B chromosomes in *E. plorans* (21.5%) although the TR family that has successfully prospered in Bs was the EplTR002-196 (34.3%) whose sequence is similar to the one of the former TR (see Fig. S2.5), both families showing the same FISH pattern in Bs. We also verified quantitatively the presence of rDNA in Bs of *E. plorans* (5%) previously described by (López-León et al., 1994). In addition, we uncovered the presence of 454 transposable elements in B chromosomes representing about a 15.7% of their content being lower than that found in the standard set of A chromosomes (37.6%). Hence, we could state that the B chromosomes of *E. plorans* are enriched in TR families but depleted in TEs. However, we still found several TE families accumulated in B chromosomes of *E. plorans* (e.g. TE_LINE/Tx1_01, TE_LINE_Tx1_05 or TE_LINE/RTE_26, see Dataset 2.1) and yielding a banded FISH pattern (e.g. TE_DNA/hAT_11 or TE_SINE_8, see Dataset 2.1), this contrasts with the low amount of some TEs in Bs of this species estimated by Montiel et al. (2012). Several transposable elements found in B chromosomes of *E. plorans* (e.g. TE_LINE/Daphne_8, TE_DNA/Kolobok_02, TE_DNA/piggyBac_04, TE_DNA/piggyBac_08, TE_DNA/Polinton_01) are scattered in A chromosomes of the species, but appeared concentrated in certain regions of Bs, indicating that these chromosomes are an ideal

place for the accumulation of mobile elements, especially DNA transposons. This may be due to silencing of B chromosomes and the resulting lack of selective pressure, so these elements would behave as shelters where transposons could hide and get out of their selective removal. Alternatively, B chromosomes may be an evolutionary sink for some mobile elements, such as retrotransposon R2 in *E. plorans* (Montiel et al., 2014). This transposon is specifically inserted into the 28S rDNA and their transposition depends on the expression of rDNA. As the B chromosome rDNA is strongly muted, R2 elements falling into chromosomes B are mostly trapped in this location.

Moreover, it seems that there is a tendency towards the amplification of some repetitive variants in B chromosomes of *E. plorans* revealed by a decrease in the Kimura divergence of several repetitive families in the 4B individuals compared to the 0B one (Kdiv0B_av.= 9.1, Kdiv4B_av.= 8.8; if gFC>1.58 then Kdiv0B_av.= 16.2, Kdiv4B_av.= 7.5). This mechanism has been claimed by several authors as a way to stabilize the raise of new B chromosomes (Reed et al., 1994; Leach et al., 1995) and as been described after the formation of new variants of Bs in rye (Marques et al., 2012).

Interestingly, we discovered for the first time repetitive elements detected specifically in B chromosomes of *E. plorans* by FISH, as EplTR106-323 and EplTR112-11, yielding well-defined bands in Bs thus they can be used as markers of their presence in cells as performed by Cabrero et al. (2017). The repetitive families TE_LINE/Tx1_01, TE_LINE/Tx1_05, TE_LINE/Daphne_8 and TE_LINE/RTE_26 showed also weak specific FISH signal in B chromosomes but due to their molecular nature they might be not suitable for detection of B chromosomes through FISH.

However, these B-FISH specific sequences allowed us to propose a intragenomic origin of B chromosomes in this species from the chromosome 9 of the standard set. This hypothesis is supported by similar repetitive contents in A chromosomes and Bs of *E. plorans* (see Results section) and the presence of B-specific sequences in long MinION reads including sequences specific from the chromosome 9 (see Fig. 2.2) or the high number of repetitive elements showing FISH signal in the chromosome 9 and also in the Bs (see Table 2.1 and S2.3).

Special mention should be made for the case of TE_SINE_08. This sequence co-localizes with the ribosomal DNA on B and 9 chromosomes but it is also found in the X chromosomes but without sharing location with the rDNA (TE_SINE_08 is located distally in the X chromosome while rDNA is near the centromere). Given this fact, and the

presence of all repetitive sequences analyzed in both B and A chromosomes, B chromosomes of *E. plorans* probably have and intragenomic origin. The X chromosome was initially proposed as a possible origin of the B chromosome in *E. plorans* in Western populations (i.e. Spain and Morocco; López-León et al., 1994; Cabrero et al., 2003b). However, same authors put in doubt the role of the X chromosome as source of Bs after results from chromosome painting using probes derived from the X and B chromosomes that hybridized with several autosomes. Nevertheless, they still confirmed the intragenomic origin of the B chromosome in *E. plorans* (Teruel et al., 2009). On the other hand, recent evidence from the sequences of the ITS1 and ITS2 regions of rDNA suggested that the Bs sequences look like those of the smallest autosome (Teruel et al., 2014) and the same was also proposed for the origin of B chromosomes in Eastern populations (i.e. Caucasus) in a context of a multiregional origin of Bs in *E. plorans* (Cabrero et al., 2003b). In all previous studies about this issue, the chromosome 9 was ruled out as a candidate for B chromosome origin since the relative order of the sat180bp and the rDNA was the opposite in this chromosome than in the Bs (Cabrero et al., 2003b), but we should considerer that the location of sat180bp is highly polymorphic in the chromosome 9 of this species and the same order of sat180bp and rDNA in both chromosomes have been described in some specimens of *E. plorans* (Cabrero et al., 2003a; Abdelaziz et al., 2007). In addition, some inversions could have taken place in the moment of B chromosomes formation as it have been proposed for the arisen of neo-sex chromosomes (Lahn et al., 1999; Palacios-Gimenez et al., 2018; Natri et al., 2019).

Here, neither the chromosome 11 nor the X are robust candidates for Bs origin concerning the repetitive DNA landscape, being the chromosome 9 the one gathering all requisites. Furthermore, in populations where the chromosomes 9 of *E. plorans* contains poor amount of sat180bp, or it is absent in that chromosome, the B chromosomes variants described there are also scarce in sat180bp compared to other populations in which the presence of sat180bp on chromosomes 9 was detected by FISH (Cabrero et al., 1997; López-León et al., 2008).

Regarding the A chromosomes, the only TR family that occupies the centromeric region of all of them was the EplTR005-49, so this TR is the best candidate to be related to the centromeric function in this species. This possibility should be studied in more detail because the comprehension of the structure and regulation of centromeres of chromosomes A and B is a prerequisite for a better understanding of the segregation of B

chromosomes (Houben et al., 2011). B chromosomes of different species carry tandem repeats located at the centromere of both B and A chromosomes, such as maize (Lamb et al., 2005). The centromere of B chromosome of the maize share sequences with other chromosomes of the standard set, in addition, these sequences are found elsewhere in the maize B chromosome (as in the case of Bs of *E. plorans* for EplTR005-49). These centromeric sequences of maize B chromosomes follow their own evolutionary pathway, so they have differed markedly from A chromosomes showing lower similarity with A sequences than with the homologous sequences found in other locations of B chromosomes (Peng and Cheng, 2011). Additionally, comparison between rye A and B centromeres also revealed differences in the centromere repeat composition (Banaei-Moghaddam et al., 2012). In this regard, it would be interesting, in the future, to analyze EplTR005-49 sequence divergence between A and B chromosomes of *E. plorans*. The centromeric location of EplTR005-49 have been verified by FISH in other populations of *E. plorans* carrying different B chromosomes variants (Fig. S2.2) which reinforces the possible role of this TR as centromeric satDNA in distant populations of *E. plorans*.

We also found the presence of several other repetitive DNA sequences in the B chromosomes of Tanzanian, Egyptian and Armenian populations of *E. plorans* (Table S2.4). In some cases, the presence of mobile elements in B chromosomes has been studied to obtain valuable information about the origin of Bs. For example, the accumulation of transposons in rye B chromosomes, which are also found in the As, suggests the possible intraspecific B chromosome origin of this species (Klemme et al., 2013). However, the greatest similarity between some TEs located on B chromosomes of the wasp *Nasonia vitripennis* and those found in the wasps genus *Trichomalopsis* supports the interspecific source B chromosomes of *N. vitripennis* (McAllister and Werren, 1997). In the case of *E. plorans*, the similar and correlated repetitive DNA composition of B chromosomes from individuals belonging to distant populations analyzed here (Tanzania, Egypt, Armenia and Spain) pointed to a common origin of B chromosomes in the species which is in line with previous conclusions obtained by Muñoz-Pajares et al. (2011) and Cabrero et al. (1999). This idea is also supported by the phylogenetic signal yielded by the abundances of B-located repetitive sequences from all the populations above mentioned. In view of these results, the B chromosome of *E. plorans* would have arisen in middle Africa supported by the basal position of the 1B individual from the Tanzanian population in the phylogenetic tree (see Fig. 2.3). This

region in Africa represented a hotspot scenario in the speciation process of *E. plorans* (John and Lewis, 1965), therefore, it would be interesting to compare the repetitive DNA content present on B chromosomes between different subspecies of *E. plorans*, to test the possible origin of B chromosomes by hybridization between subspecies.

Despite the high number of repetitive elements located in the B chromosomes of *E. plorans*, only some of them were transcriptionally active and just a limited fraction showed more expression coming from B chromosomes than from As (TI>0, see Fig. S2.4b and Tables S2.6 and S2.7) specially in adults gonads. However, it is true that, considering classes of repetitive DNA, we found a turnover of repetitive DNA expression in B chromosomes compared to that from the standard set (Fig. S2.4b) with an increase in transcription activity of TEs together with a reduction in mtDNA expression. This pattern of few and not highly transcribed B-located repetitive elements was also found by Ruiz-Ruano et al. (2018) in the grasshopper *L. migratoria* and by Coan and Martins (2018) in the B chromosomes of the FISH *A. latifasciata*.

Therefore, putative silencing mechanisms may act on the B chromosome to avoid detrimental consequences of repeat transcription in the genome while other kind of sequence such as protein-coding genes could be active and perhaps functional involving some key aspect for B chromosome maintenance and evolution as argue for genes found in B chromosomes of *E. plorans* by Navarro-Domínguez et al. (2017a and 2019). Although there is still much to know about the molecular content of B chromosomes in *E. plorans*, especially regarding to possible sequences represented by fewer copies, this work contributes to the understanding of the molecular make up and origin of B chromosomes in *E. plorans* and the role of repetitive DNA in B chromosome evolutionary dynamics.

References

Abdelaziz M, Teruel M, Chobanov D, Camacho JP, Cabrero J. (2007). Physical mapping of rDNA and satDNA in A and B chromosomes of the grasshopper *Eyprepocnemis plorans* from a Greek population. *Cytogenetic and Genome Research*, 119(1–2), 143–146.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–10.

Banaei-Moghaddam AM, Schubert V, Kumke K, Weiβ O, Klemme S, Nagaki K, et al. (2012). Nondisjunction in favor of a chromosome: the mechanism of rye B chromosome drive during pollen mitosis. *Plant Cell*, 24, 4124–4134.

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(7), 2536–2544.

Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, et al. (2013). MITOS: improved *de novo* metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, 69(2), 313–9.

Bolger AM, Lohse M, Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120.

Burt A, Trivers R. (2006). Genes in conflict: The biology of selfish genetic elements.

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, Gómez R, Castro AJ, Martín-Alganza A, Camacho JP. (1997). Geographical distribution of B chromosomes in the grasshopper *Eyprepocnemis plorans,* along a river basin, is mainly shaped by non-selective historical events. *Chromosome Research*, 5(3), 194–198.

Cabrero J, Perfectti F, Gómez R, Camacho JPM, López-León MD. (2003a). Population variation in the A chromosome distribution of satellite DNA and ribosomal DNA in the grasshopper *Eyprepocnemis plorans*. *Chromosome Research*, 11, 375–381.

Cabrero J, Bakkali M, Bugrov A, Warchalowska-Sliwa E, López-León MD, Perfectti F, et al. (2003b). Multiregional origin of B chromosomes in the grasshopper *Eyprepocnemis plorans*. *Chromosoma*, 112, 207–211.

Cabrero J, Manrique-Poyato MI, Camacho JPM. (2006). Detection of B chromosomes in interphase hemolymph nuclei from living specimens of the grasshopper *Eyprepocnemis plorans*. *Cytogenetic and Genome Research*, 114(1), 66–69.

Cabrero J, Martín-Peciña M, Ruiz-Ruano FJ, Gómez R, Camacho JPM. (2017). Post-meiotic B chromosome expulsion, during spermiogenesis, in two grasshopper species. *Chromosoma*, 126(5), 633–644.

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

Camacho JPM, Sharbel TF, Beukeboom LW. (2000). B-chromosome evolution. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 355(1394), 163–178.

Camacho JPM. (2005). B chromosomes. The evolution of the genome (Gregory TR, ed.), 223–

286.

Carmello BO, Coan RL, Cardoso AL, Ramos E, Fantinatti BE, Marques DF, et al. (2017). The hnRNP Q-like gene is retroinserted into the B chromosomes of the cichlid fish *Astatotilapia latifasciata*. *Chromosome Research*, 25(3–4), 277–290.

Coan RLB, Martins C. (2018). Landscape of Transposable Elements Focusing on the B Chromosome of the Cichlid Fish *Astatotilapia latifasciata*. *Genes (Basel)*, 9(6), 269.

Dalla Benetta E, Antoshechkin I, Yang T, Nguyen HQM, Ferree PM, Akbari OS. (2020). Genome elimination mediated by gene expression from a selfish chromosome. *Science Advances*, 6(14), eaaz9808.

Ebrahimzadegan R, Houben A, Mirzaghaderi G. (2019). Repetitive DNA landscape in essential A and supernumerary B chromosomes of *Festuca pratensis* Huds. *Scientific Reports*, 9(1), 19989.

Finn RD, Clements J, Eddy SR. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(Web Server issue), W29–W37.

Fu L, Niu B, Zhu Z, Wu S, Li W. (2012). CD-HIT: accelerated for clustering the next–generation sequencing data. *Bioinformatics*, 28(23), 3150–3152.

Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. (2015). *De novo* assembly and annotation of the Asian tiger mosquito (*Aedes albopictus*) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (*Aedes aegypti*). *Genome Biology and Evolution*, 7(4), 1192–1205.

Graphodatsky AS, Kukekova AV, Yudkin DV, Trifonov VA, Vorobieva NV, Beklemisheva R, et al. (2005). The proto-oncogene C-KIT maps to canid B-chromosomes. *Chromosome Research*, 13(2), 113–122.

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(13), e129.

Hanlon SL, Miller DE, Eche S, Hawley RS. (2018). Origin, Composition, and Structure of the Supernumerary B Chromosome of *Drosophila melanogaster*. *Genetics*, 210(4), 1197–1212.

Henriques-Gil N, Santos JL, Arana P. (1984). Evolution of a complex polymorphism in the grasshopper *Eyprepocnemis plorans*. *Chromosoma*, 89, 290–293.

Houben A, Kumke K, Nagaki K, Hause G. (2011). CENH3 distribution and differential chromatin modifications during pollen development in rye (*Secale cereale* L.). *Chromosome Research*, 19, 471– 480.

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), 88.

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

John B, Lewis KR. (1965). Genetic speciation in the grasshopper *Eyprepocnemis plorans*. *Chromosoma*, 16, 308–344.

Jones N, Houben A. (2003). B chromosomes in plants: Escapees from the A chromosome genome? *Trends in Plant Science*, 8(9), 417–423.

Jones RN. (1991). B-chromosome drive. *The American Naturalist*, 137(3), 430–442.

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(2), 550–8.

Lamb JC, Kato A, Birchler JA. (2005). Sequences associated with A chromosome centromeres are present throughout the maize B chromosome. *Chromosoma*, 113, 337-349

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.

Lahn BT, Page DC. (1999). Four evolutionary strata on the human X chromosome. *Science*, 286(5441), 964–967.

Leach CR, Donald TM, Franks TK, Spiniello SS, Hanrahan CF, Timmis JN. (1995). Organisation and origin of a B chromosome centromeric sequence from *Brachycome dichromosomatica*. *Chromosoma*, 103, 708–714.

López-León MD, Neves N, Schwarzacher T, Heslop-Harrison TS, 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.

López-León MD, Cabrero J, Dzyubenko VV, Bugrov AG, Karamysheva TV, Rubtsov NB, et al. (2008). Differences in ribosomal DNA distribution on A and B chromosomes between eastern and western population of the grasshopper *Eyprepocnemis plorans* plorans. *Cytogenetic and Genome Research*, 121, 260–265.

Lowe TM, Eddy SR. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5), 955–964.

Ma W, Gabriel TS, Martis MM, Gursinsky T, Schubert V, Vrána J, et al. (2017). Rye B chromosomes encode a functional argonaute-like protein with in vitro slicer activities similar to its A chromosome paralog. *New Phytologist*, 213(2), 916–928.

Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, 47(W1), W636–W641.

Marques A, Klemme S, Guerra M, Houben A. (2012). Cytomolecular characterization of *de novo* formed rye B chromosome variants. *Molecular Cytogenetics*, 5(1), 34.

Martín-Peciña M, Ruiz-Ruano FJ, Camacho JPM, Dodsworth S. (2019). Phylogenetic signal of genomic repeat abundances can be distorted by random homoplasy: a case study from hominid primates. *Zoological Journal of the Linnean Society*, 185(3), 543–554.

Martis MM, Klemme S, Banaei-Moghaddam AM, Blattner FR, Macas J, Schmutzer T, 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 USA*, 109(33), 13343–13346.

McAllister BF. (1995). Isolation and characterisation of a retro-element from B chromosome (PSR) in the parasitic wasp *Nasonia vitripennis. Insect Molecular Biology*, 4, 253–262.

McAllister BF, Werren JH. (1997). Hybrid origin of a B chromosome (PSR) in the parasitic wasp *Nasonia vitripennis*. *Chromosoma*, 106(4), 243–253.

Meredith R. (1969). A simple method for preparing meiotic chromosomes from mammalian test. *Chromosoma*, 26, 254–258.

Montiel EE, Cabrero J, Camacho JP, López-León MD. (2012). Gypsy, RTE and Mariner transposable elements populate *Eyprepocnemis plorans* genome. *Genetica*, 140(7–9), 365–374.

Montiel EE, Cabrero J, Ruiz-Estévez M, Burke WD, Eickbush TH, Camacho JPM, et al. (2014).

Preferential occupancy of R2 retroelements on the B chromosomes of the grasshopper *Eyprepocnemis plorans*. *PLoS One*, 9(3), e91820.

Muñoz-Pajares AJ, Martínez-Rodríguez L, Teruel M, Cabrero J, Camacho JP, Perfectti F. (2011). A single, recent origin of the accessory B chromosome of the grasshopper *Eyprepocnemis plorans*. *Genetics*, 187(3), 853–863.

Natri HM, Merilä J, Shikano T. (2019). The evolution of sex determination associated with a chromosomal inversion. *Nature Communications*, 10(1), 145.

Navarro-Domínguez B. (2016). Análisis de los cambios de expresión génica asociados a la presencia de cromosomas B en el saltamontes *Eyprepocnemis plorans*. Universidad de Granada. Retrieved from http://hdl.handle.net/10481/44096.

Navarro-Domínguez B, Ruiz-Ruano FJ, Cabrero J, Corral JM, López-León MD, Sharbel TF, et al. (2017a). Protein–coding genes in B chromosomes of the grasshopper *Eyprepocnemis plorans*. *Scientific Reports*, 7, 45200.

Navarro-Domínguez B, Ruiz-Ruano FJ, Camacho JPM, Cabrero J, López-León MD. (2017b). Transcription of a B chromosome CAP–G pseudogene does not influence normal condensin complex genes in a grasshopper. *Scientific Reports*, 7(1), 17650.

Navarro-Domínguez B, Martín-Peciña M, Ruiz-Ruano FJ, Cabrero J, Corral JM, López-León MD, et al. (2019). Gene expression changes elicited by a parasitic B chromosome in the grasshopper *Eyprepocnemis plorans* are consistent with its phenotypic effects. *Chromosoma*, 128(1), 53–67.

Ning Z, Cox AJ, Mullikin JC. (2001). SSAHA: a fast search method for large DNA databases. *Genome Research*, 11(10), 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.

Palacios-Gimenez OM, Milani D, Lemos B, Castillo ER, Martí DA, Ramos E, et al. (2018). Uncovering the evolutionary history of neo–XY sex chromosomes in the grasshopper *Ronderosia bergii* (Orthoptera, Melanoplinae) through satellite DNA analysis. *BMC Evolutionary Biology*, 18(1), 2.

Peng SF, Cheng YM. (2011). Characterization of satellite CentC repeats from heterochromatic regions on the long arm of maize B-chromosome. *Chromosome Research*, 19(2), 183–191.

Reed KM, Beukeboom LW, Eickbush DG, Werren JH. (1994). Junctions between repetitive DNAs on the PSR chromosome of *Nasonia vitripennis:* association of palindromes with recombination. *Journal of Molecular Evolution*, 38, 352–362.

Ruiz-Estévez M, Cabrero, J, Camacho JPM. (2012). B-chromosome ribosomal DNA is functional in the grasshopper *Eyprepocnemis plorans*. *PLoS ONE*, 7(5), e36600.

Ruiz-Estévez M, López-León MD, Cabrero J, Camacho JPM. (2013). Ribosomal DNA is active in different B chromosome variants of the grasshopper *Eyprepocnemis plorans*. *Genetica*, 141(7–9), 337–345.

Ruiz-Estévez M, Badisco L, Broeck J, Perfectti F, López-León MD, Cabrero J, 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(6), 1209–1216.

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(2), 120–126.

Ruiz-Ruano FJ, López-León M, Cabrero J, Camacho JPM. (2016). High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific Reports*, 6, 28333.

Ruiz-Ruano FJ, Cabrero J, López-León MD, Camacho JPM. (2017). Satellite DNA content illuminates the ancestry of a supernumerary (B) chromosome. *Chromosoma*, 126(4), 487–500.

Ruiz-Ruano FJ, Cabrero J, López-León MD, Sánchez A, Camacho JPM. (2018). Quantitative sequence characterization for repetitive DNA content in the supernumerary chromosome of the migratory locust. *Chromosoma*, 127(1), 45–57.

Ruiz-Ruano FJ, Navarro-Domínguez B, López-León MD, Cabrero J, Camacho JPM. (2019). Evolutionary success of a parasitic B chromosome rests on gene content. *BioRxiv*, 683417.

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

Smit AFA, Hubley R, Green P. (2013). RepeatMasker Open–4.0. http://www.repeatmasker.org.

Teruel M, Cabrero J, Perfectti F, Acosta MJ, Sánchez A, Camacho JPM. (2009). Microdissection and chromosome painting of X and B chromosomes in the grasshopper *Eyprepocnemis plorans*. *Cytogenetic and Genome Research*, 125(4), 286–291.

Teruel M, Ruiz–Ruano FJ, Marchal JA, Sánchez A, Cabrero J, Camacho JPM, et al. (2014). Disparate molecular evolution of two types of repetitive DNAs in the genome of the grasshopper *Eyprepocnemis plorans*. *Heredity (Edinb)*, 112(5), 531–42.

Trifonov VA, Dementyeva PV, Larkin DM, O'Brien PC, Perelman PL, Yang F, et al. (2013). Transcription of a protein-coding gene on B chromosomes of the siberian roe deer (*Capreolus pygargus*). *BMC Biology*, 11(1), 90.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Research*, 40, e115–e115.

Utsunomia R, Silva DM, Ruiz-Ruano FJ, Araya-Jaime C, Pansonato-Alves JC, Scacchetti PC, et al. (2016). Uncovering the Ancestry of B Chromosomes in *Moenkhausia sanctaefilomenae* (Teleostei, Characidae). *PLoS One*, 11(3), e0150573.

Valente GT, Conte MA, Fantinatti BE, Cabral-de-Mello DC, Carvalho RF, Vicari MR, et al. (2014). Origin and evolution of B chromosomes in the cichlid fish *Astatotilapia latifasciata* based on integrated genomic analyses. *Molecular Biology and Evolution*, 31(8), 2061–2072.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12), 973–82.

Yoshida K, Terai Y, Mizoiri S, Aibara M, Nishihara H, Watanabe M, et al. (2011). B chromosomes have a functional effect on female sex determination in lake victoria cichlid fishes. *PLoS Genetics*, 7(8), e1002203.

Supplementary Figures for Chapter 2

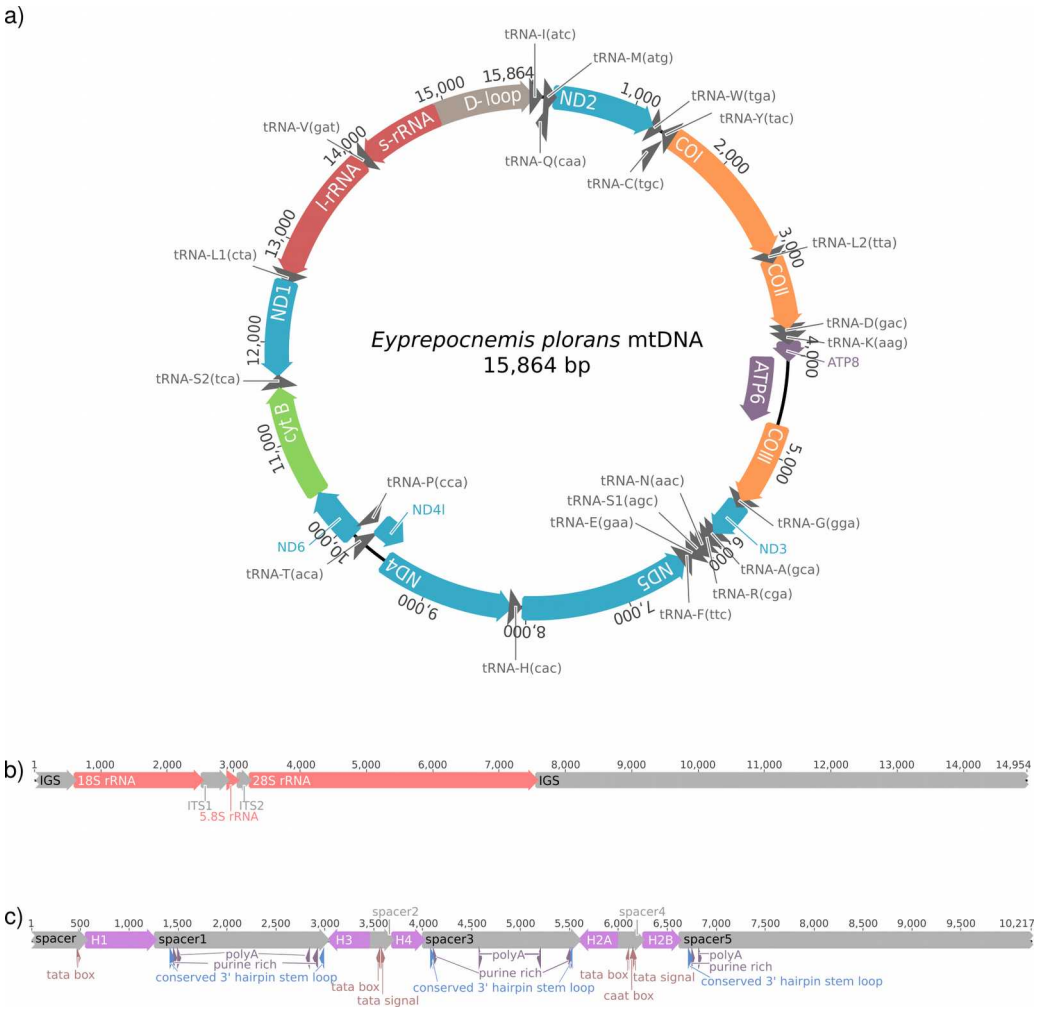


Figure S2.1. Annotation of the mitochondrial DNA (a), ribosomal DNA (B) and histone cluster (c) of *E. plorans*. See that, while the mtDNA was assembled completely, the rDNA and histones are truncated by the IGS and spacer 5 regions respectively.

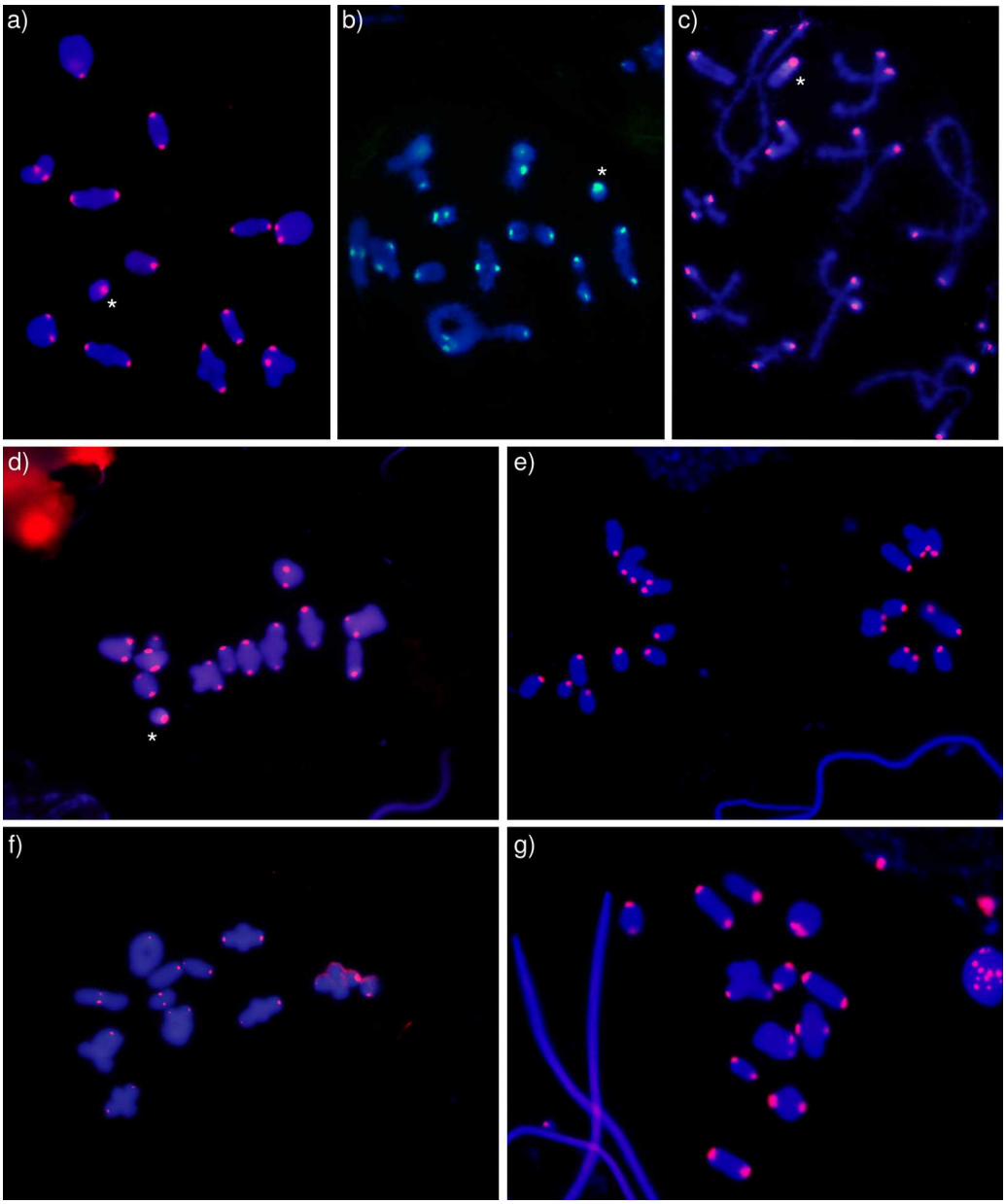


Figure S2.2. FISH of the centromeric satDNA EplTR005-49 in testis of *E. plorans* from Torrox-Spain (a), Salobreña-Spain (b), Mundo-Spain (c), Fuengirola-Spain (d), Valentín-Spain (e), Turkey (f) and Cameroon (g). This satDNA is located in the centromeric region of all chromosomes in individuals from all populations analyzed. The B chromosome is indicated with and asterisk in Torrox-Spain (B24), Salobreña-Spain (B2), Mundo-Spain (B1), Fuengirola-Spain (B5) showing FISH signal not only in the centromeric region but also in the short arm.

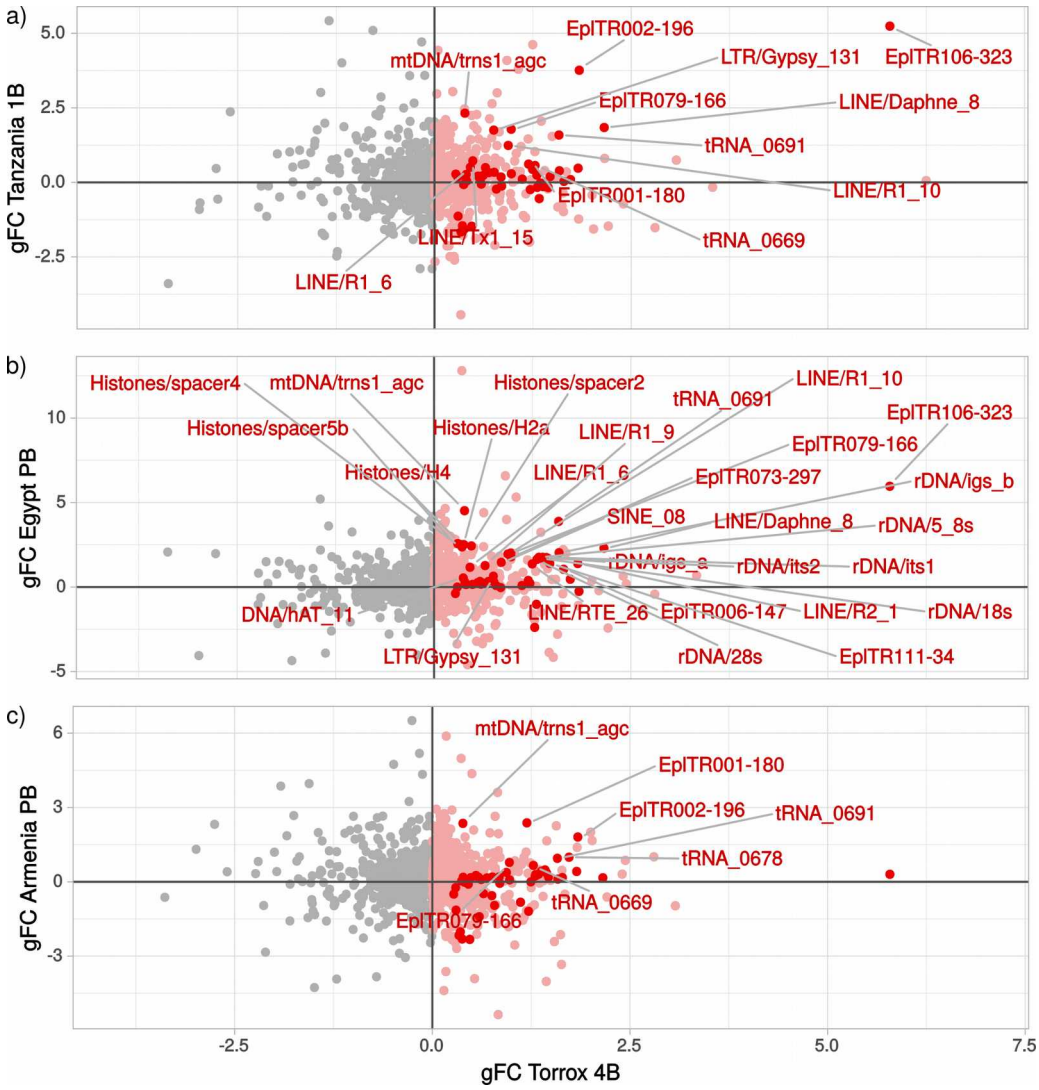


Figure S2.3. Comparison of gFC values for repetitive DNA elements of *E. plorans* between the population of Torrox and that of Tanzania (a), Egypt (b) and Armenia (c). Light red dots represents repetitive elements with gFC>0 in Torrox and highlighted in heavy red are those sequences showing the highest gFC values in each repetitive DNA class and superfamily. We show names for repetitive elements in strong red and with gFC>0.5 in the corresponding African or Asian population.

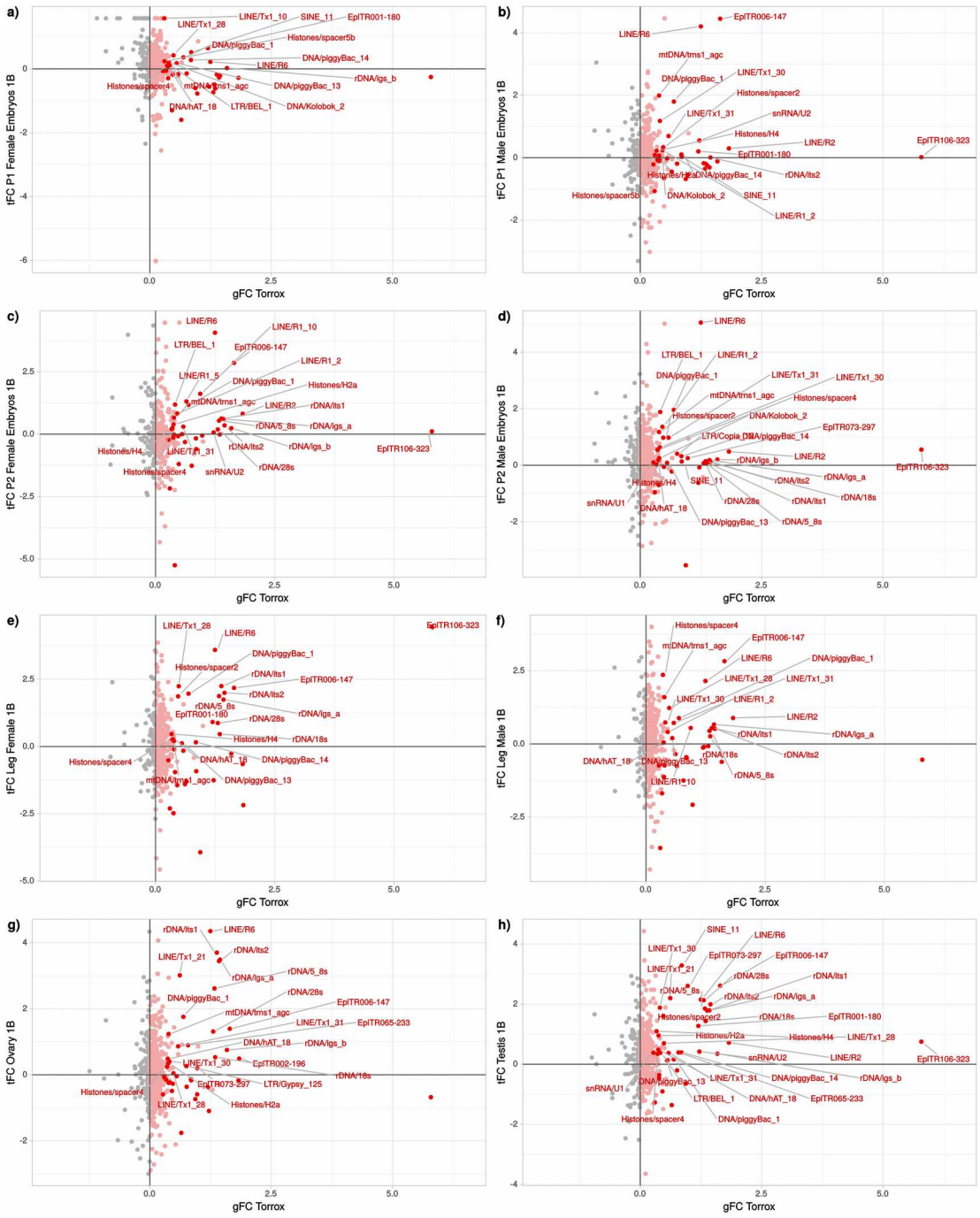


Figure S2.4. Transcriptional fold change [tFC= log2(1B/0B)] observed in 1B RNA libraries (Y-axis) in respect to gFC in female (a) and male (b) embryos from pod 1, female (c) and male (d) embryos from pod 2, adult legs from females (e) and males (f), ovary (g) and testis (h). Color codes are as in Figure S2.3.

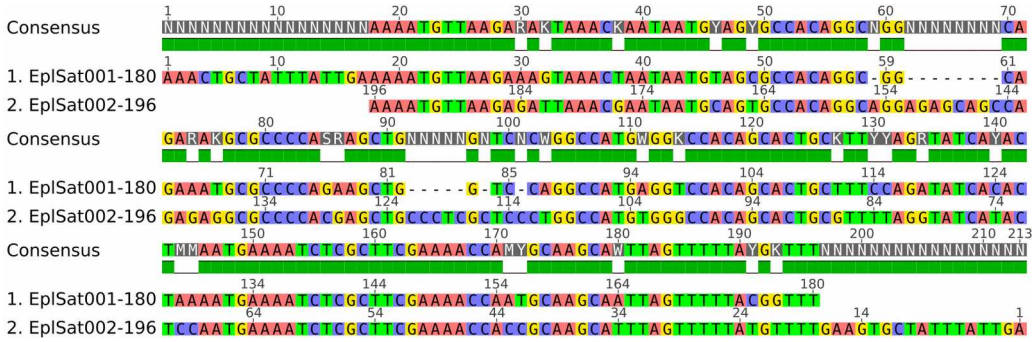


Figure S2.5. Alignment of reference monomers from EplTR001-180 and EplTR002-196 families showing a pairwise identity of 77.7%. The alignment was performed with Geneious 4.8.5.

Supplementary Datasets and Tables for Chapter 2

Dataset 2.1. FISH results for the 33 B-located repetitive sequences identified in the genome of *E. plorans*. For details about location of TR families see Dataset 1.1.
Dataset available in: https://figshare.com/s/c52b820785c9e7103a63

Supplementary Tables can be downloaded from:
https://figshare.com/s/26dd8e805a40f6d27b64

Table S2.1. Classification of the TE sequences comprised in the *de novo* database of the repetitive DNA from *E. plorans.*

Table S2.2. Genomic abundance and Kimura divergence for each repetitive DNA elements in 0B and 4B males of *E. plorans* from Torrox (Spain). The gFC4B was calculated as the log2 of the quotient between the genome proportion of the 4B individual and the 0B one, both values normalized by genome size. The proportion of a certain repetitive element in the one B chromosome was calculated as B_prop=((4Bgenome proportion*4B_C-value)-(0Bgenome proportion*0B_C-value))/(NºBs_haploid_genome*B size).

Table S2.3. Genomic proportion and Kimura divergence (%) of some B-located repetitive elements in 0B and 4B *E. plorans* males from Torrox (Spain) sorted decreasingly by their proportion in B chromosomes. See details about FISH signal such as pattern (Patt.; B= banded, D= dotted and DB= dotted-banded), number of A pairs showing signal (NA), chromosome location (1, 2, 3,…, B; p= pericentomeric, i= interstitial, d= distal). Find indicated in the last column the ID of the B-located repetitive element that shared at least one MinION reads with certain chromosome-specific elements.

Table S2.4. Genomic abundance and Kimura divergence for each repetitive DNA elements in 0B and +B males of *E. plorans* from Tanzania, Egypt and Armenia. The gFC was calculated as the log2 of the quotient between the genome proportion of the +B individual and the 0B one, both values normalized by genome size. The proportion of a certain repetitive element in the one B chromosome was calculated as B_prop=((+Bgenome proportion*+B_C-value)-(0Bgenome proportion*0B_C-value))/(NºBs_haploid_genome*B size). We are unaware of the number of B chromosomes of the +B individuals of *E. plorans* from Egypt and Armenia (the one from Tanzania was 1B) so we considerer both individuals as 2B males for data normalization as they were selected for sequencing as showing the most intense amplicon after PCR of the B chromosome SCAR marker. We considered also genome and B sizes as that found in Torrox.

Table S2.5. Genomic and transcriptomic abundance, normalized by genome and library size, of the 15,161 SNPs found in the B-located repetitive sequences (gFC4B>0) of *E. plorans.*

Table S2.6. Average transcription intensity (TI) for the different B-located repetitive DNA classes found in the RNA libraries from different biological samples of *E. plorans* analyzed.

Table S2.7. Transcription intensity (TI) for the different B-located repetitive sequences found in the RNA libraries from different biological samples of *E. plorans* analyzed. Note that sequences showing TI=0 in all samples are excluded from the list.

Table S2.8. Primers used for generation of FISH probes. For details about satDNA/TR primers see Table S1.6.

Table S2.9. Number of specific SNPs in the +B libraries of *E. plorans* from Tanzania, Egypt and Armenia.

3. B chromosomes of *Eyprepocnemis plorans* contain active protein-coding genes involved in cell division

Abstract: Supernumerary (B) chromosomes are dispensable genomic elements found in most kinds of eukaryotic genomes. Many show drive mechanisms that give them an advantage in transmission, but how they achieve them remains a mystery. The recent finding of protein-coding genes in B chromosomes has opened the possibility that their evolutionary success is based on their genetic content. Using a protocol based on mapping genomic DNA Illumina reads from B-carrying and B-lacking individuals on the coding sequences of *de novo* transcriptomes, we identified 42 protein-coding genes in the B chromosome of *Eyprepocnemis plorans*. Sequence comparison of A and B chromosome gene paralogs showed that the latter harbors B-specific nucleotide changes. These nucleotide signatures allowed identifying B-derived transcripts in B-carrying transcriptomes, showing some of them higher expression levels than A-derived ones thus self-disclosured as subverted genes. We found *ndl* as a subverted genes in *E. plorans*, in particular, it is strongly subverted in the ovary, the place in which the drive of B chromosomes have been described for this species. *Ndl* is involved in asymmetric cell division, a mechanisms that would enhance B drive in cells. In addition, one of the active B-genes in *E. plorans*, *cdc16*, codes for a subunit of the Anaphase Promoting Complex or Cyclosome (APC/C), an E3 ubiquitin ligase involved in the metaphase-anaphase transition. Remarkably, the gene *apc1* of *Locusta migratoria* B chromosomes also codes for a different subunit of the APC/C complex. This coincidence opens the door to the idea that all successful B chromosomes harbor active genes for regulation of cell division. Therefore, here we present the first protein-coding genes uncovered in a B chromosome that might be responsible for its drive operating in cell cycle regulation.

Keywords: *B drive*, *cell division, protein-coding genes, SNP, transcription*

Introduction

After 113 years since they were uncovered (Wilson, 1907), B chromosomes continue being an enigmatic part of eukaryote genomes. They were first considered as merely genetically inert passengers of eukaryote genomes (Rhoades and McClintock, 1935), a view supported by some authors (Randolph, 1941) but criticized by those who argued that B chromosomes are beneficial (Darlington and Upcott, 1941) or parasitic (Östergren, 1945) elements. In fact, as in most cases of selfish genetic elements, phenotypic effects of B chromosomes are usually modest (Burt and Trivers, 2006) but they frequently decrease fertility (Jones and Rees, 2006; Camacho, 2005). An extreme case was found in the wasp *Nasonia vitripennis*, where a B chromosome decreases the fitness of the host genome to zero by the elimination of all paternal chromosomes accompanying it in the spermatozoon (Nur et al., 1988).

Thirteen years ago, two findings suggested that B chromosomes were not inert elements, namely the first molecular evidence of gene activity on B chromosomes (Leach et al., 2005) and the finding of protein-coding genes in them (Graphodatsky et al., 2005). Later on, Carchilan et al. (2012) showed the transcription of ribosomal DNA in B chromosomes and, in the same year, Ruiz-Estévez et al. (2012) found that rDNA transcripts from a B chromosome are functional. The first evidence for transcription of a protein-coding gene on B chromosomes was provided by Tryfonov 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, some of which are pseudogenical and actively transcribed (Banaei-Moghaddam et al., 2013). Likewise, Valente et al. (2014) found that a B chromosome in the fish *Astatotilapia latifasciata* 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. Likewise, Navarro-Domínguez et al. (2017) found evidence for the presence of ten 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. Very recently, Ahmad et al. (2020) have found the presence of several protein-coding genes in the B chromosomes of distant species although they have not investigated the expression of these genes. In spite of this, authors stated that the

function of those B-genes could explain some of the effects of these parasitic chromosomes in carrier individuals.

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. However, 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 harbors few transcriptionally active sequences and might influence transcription in A chromosomes. 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 up-regulated genes are actually present in these B chromosomes. Notably, it has been recently shown that rye B chromosomes carry active Argonaute-like genes and that B-encoded AGO4B protein variants show in vitro RNA slicer activity (Ma et al., 2017), thus going a step further in demonstrating the functionality of B chromosome gene paralogs. Recently, Dalla Benetta et al. (2020) have found the *haploidizer* gene, located in the B chromosomes (PSR-Paternal Sex Ratio) of *Nasonia vitripennis*, that induces the paternal genome elimination. Ruban et al. (2020) proved B chromosome elimination from root during embryo stages of *Aegilops speltoides* and they hypothesized a role of Bs elimination leading to the silencing of root related genes. In addition, it has been shown that B chromosome presence in *E. plorans* triggers transcriptional changes being consistent with some of the effects previously reported in this species (Navarro-Domínguez et al., 2019) but still it is not clear whether those changes came specifically from the B chromosomes.

Since decades, grasshoppers have been especially useful in the study of chromosomes as they are quite easy-handled for cytogenomics techniques (Bidau and Martí, 2010). Therefore, it is not at all a surprise the high number of grasshopper species in which B chromosomes have been described. In particular, the species *E. plorans* has raised up as an outstanding model for B chromosome studies as they are easily maintained in the lab and B chromosome number is the same in each cell (mitotically stable B chromosome). The geographic distribution of the subspecies *E. plorans plorans* includes the entire Mediterranean coast, the Caucasus, Turkey, Turkmenistan, Iran and the south-west of the Arabian Peninsula (Dirsh, 1958). The presence of B chromosomes in this subspecies has been described in almost all natural populations hitherto analyzed

(for review, see Camacho et al., 2003) and the Bs drive during oogenesis could explain this widespread presence (Zurita et al., 1998; Bakkali et al., 2002).

Here we use NGS approaches to search for protein-coding genes in B chromosomes with the potential to manipulate cell division, by means of genomic and transcription analyses. We found that the B chromosome of the *E. plorans* contains at least 42 protein-coding genes. Several of these genes are involved in function related to cell cycle regulation and meiosis, as *ndl*, being intensely transcribed in gonads of B-carrying individuals and specifically the copies coming from the B chromosomes. Interestingly, one of the B-genes, *cdc16,* codes for a particular component of the multiprotein complex APC/C, as it does the gene *apc1* found in the B chromosomes of *Locusta migratoria*. *Cdc16* and *ndl* are genes involved in the control of metaphase-anaphase progression and in asymmetric cell division respectively. Both genes are transcriptionally active in the B chromosomes with transcripts showing B-specific SNPs. In particular, the gene *ndl* is expressed in a higher rate in the ovary of carriers females than in B-lacking one, thus giving a reasonable explanation to the non-disjunction events and non-Mendelian inheritance characterizing B chromosome drive mechanisms in *E. plorans*.

Materials and methods

Find here a description of the procedures that were done. For full details on how they were performed, see Supplementary Materials and Methods.

Biological material, nucleic acid isolation and sequencing

We collected females and males specimens of *E. plorans* in Torrox (Málaga) in 2014 (~60 individuals) and 2016 (~14 individuals). Half of the males collected in 2014 were processed as described in the “Materials and methods” section of this thesis: testis were cytologically analyzed to determine B chromosome presence, and body remains were frozen for DNA (hind leg) and RNA (testis and hind leg) extraction. The other half were processed *in vivo* by extracting several testis tubules through a small cut in the abdomen and cytologically analyze primary spermatocytes at diplotene or metaphase I to score the number of B chromosome of each individual. We also processed half of females extracting ovarioles from the body and fixing both parts in liquid nitrogen that were stored at -80 ^{o}C until gDNA extraction (presence of B chromosomes in fixed females was tested by PCR of the SCAR marker). We processed the other half of females *in vivo*

performing C-banding on hemolymph nuclei as described in (Cabrero et al., 2006) to ascertain the number of B chromosomes they harbored. Then we set up several controlled crosses between 1B and 0B individuals to get sibling embryos (from the same pod) for transcription analysis. We got enough embryos from two different crosses: in one of them the female progenitor was the one carrying the B chromosome (we will refer to this batch of embryos as P1, pod 1, from now on) and the one lacking it in the other cross (embryos P2 from now on). We sequence the RNA of 3 biological replicates of each biological sample of interest (males, females, 0B and 1B) from the two different controlled crosses. A total of 24 embryo libraries were sequenced in an Illumina platform and we used 23 of them for transcriptome assembly (one of them did not produce satisfactory sequencing results) and for tFC calculation between samples. In addition, we sequence the gDNA (Illumina HiSeq) of some adults individuals collected from the field, processed and stored as described above: 3 males 0B, 3 males 4B (actually 2B+1isoB) and 3 females 0B. For CDSs abundance estimations we used only male libraries while for SNP calling we used reads from both sexes.

Adults collected in 2016 were studied *in vivo* as described above to characterize the number of B chromosomes they carried prior to set up several crosses between 1B and 0B individuals to get sibling adults for RNA-seq experiments. When characterized, we selected adults coming from the same cross counting with at least 3 males 0B, 3 females 0B, 3 males 1B and 3 males 1B for each of which we performed leg and gonad gDNA and RNA extraction as described above followed by sequencing. We sequenced a total of 24 RNA libraries that we used in transcriptome assembly and for tFC calculation between samples. On the other hand, a total of 12 gDNA libraries from these individuals were sequenced and used for CDS abundance estimation in B-lacking and 1B-carriyng individuals (only males) and for SNP calling (both sexes).

In addition, we used a 0B and a 4B libraries of *E. plorans* males (Torrox, 2012) previously deposited in SRA under accession numbers SRR2970625 and SRR2970627 respectively to increase genomic coverage as much as we could for SNP calling.

Finally, we also included two gDNA libraries of *E. plorans* males, one 0B and the other one carrying B chromosomes, from each of the following populations: Tanzania, Egypt and Armenia (see Supplementary Methods for details).

De novo transcriptome assembly, annotation, mapping and selection

Bioinformatic procedures to search for B chromosome genes were similar to those described in Navarro-Domínguez et al. (2017) and are graphically summarized in Fig. S3.3 We applied this protocol to the 12 gDNA male libraries and 47 RNA-seq libraries of *E. plorans* adults and embryos of both sexes (Torrox 2014 and 2016), by generating a *de novo* transcriptome with Trinity v2.5.1 (Haas et al., 2013) which was used as a reference to map the gDNA reads to detect transcriptome contigs showing overabundance in B-carrying gDNA libraries compared to B-lacking ones, estimating the number of gene copies per haploid genome. To find contigs being candidate to reside in the B chromosome, we first selected those showing less than four copies per genome (thus discarding highly repetitive elements) and more than 0.5 copies (thus discarding contigs putatively showing coverage problems). Among these contigs, we selected, as possible candidate to be present in the B chromosome, those contigs showing a genomic fold change [gFC= log2(4B/0B)] associated to B chromosome presence being higher than 1.585 in the three 4B males of *E. plorans*. As 4B males of *E. plorans* were actually 2B+1isoB we applied a gFC threshold higher than 1 for each 4B library separately requiring a gFC>0.585 in the average of 1B libraries to identify genes potentially located in short arms of Bs (see Dataset 3.2 for selection process). This threshold was set by assuming that every A or B chromosome carried one copy of the gene. Before the identification of B genes following the steps described above, we explored the performance of different methodological approaches in yielding false positives depending on the number of biological replicates and B chromosomes in +B individuals, and also considering or not the range of 0.5-4 copies in 0B samples.

Sequence analysis for selected contigs

We first checked if the CDS was complete in each of the selected contigs and retrieved the full transcript (CDS plus both UTRs if possible) from the *de novo* transcriptomes. Then, we performed an additional mapping of all genomic and transcriptomic libraries using all the selected transcripts as a reference to get estimates of coverage per nucleotide site along each gene. We then calculated the proportion of nucleotides showing higher abundance in the 4B gDNA libraries compared to the 0B ones (prop_nt4B>0B). This proportion was used as an index of completeness, considering values above 0.90 as indication of uniform coverage pattern (UC genes) and below that value as reflecting irregular coverage (IC genes). Functional annotation of the selected

contigs was performed Eukaryotic Orthologous Groups of proteins (KOG).

We also used the software SSAHA2 to map the gDNA libraries of *E. plorans* from Tanzania, Egypt and Armenia against the B-genes identified in Torrox (Spain) in order to ascertain which B-genes were found in B-carrying individuals from these populations.

Transcription analysis of B chromosome genes

We calculated the transcription fold change (tFC) due to B chromosome presence as log2 of the quotient between 1B and 0B coverage in the RNA-seq libraries, which provided a first indication of B chromosome gene activity. To reliably identify transcripts coming from B-gene activity, we performed SNP analysis for all selected genes to search for B-specific sequence changes. At each variable nucleotide position, we considered as reference (Ref) the nucleotide being fixed in the 0B libraries, and as alternative (Alt) that being present only in +B gDNA or RNA. To increase the reliability of the nucleotidic variations observed, we applied an extra filter (before normalizing data by library and genome size) selecting those variants showing a quotient of Alt/Ref_0B copies equal or higher than the expected depending on the B chromosome number of +B libraries and assuming 1 Alt copies per B and 2 Ref copies from As, i.e. >1.1 (the average B chromosome number of 1B and 4B libraries of *E. plorans* mapped was 2.2, for genes located in short arms the expected quotient was 0.8). This procedure allow us to select variation linked directly to B chromosomes. Then, we used Geneious to translate Alt and Ref variant sequences and manually checked whether they showed synonym or non-synonym changes. We identified deleterious aminoacid substitutions using PROVEAN (Choi and Chan, 2015) and then we compared the expression of the neutral substitutions with the deleterious ones for each gene.

Additionally, we applied an approach to know the number of Alt and Ref copies located in B chromosomes. For this, we considered that the excess of Ref copies in +B genomes with respect to Ref copies in 0B genomes (represented by Ref+B/Ref0B) is due to the presence of gene copies in the B chromosomes being equal to those of the As. See Supplementary Materials and Methods for details about Alt and Ref copy number on B chromosomes calculation.

Comparing normalized reads counts for Alt and Ref gene variants in RNA samples we explored the differential expression of B chromosomes in biologically different samples (see Datasets 3.4). Additionally, comparing the expression of Alt with respect to Ref

variants in +B RNA samples we identify subverted genes expressing the Alt alleles in a higher amount than the Ref ones. Moreover, by means of read counts in gDNA and RNA of B-carrying individuals it is possible to estimate Alt variant frequency (i.e. FC=log2 Alt/Ref) in gDNA (gFC) and RNA (tFC) in the different samples testing expression of Alt copies compared to their abundance.

qPCR validation

Genomic overabundance associated with B chromosome presence was tested for 9 possible B-genes of *E. plorans* using gDNA from 6 males 0B and 13 males +B (6 1B, 3 2B, 2 3B and 2 4B). Primer pairs anchoring in the same exon were designed with Primer3 (Untergasser et al., 2012) preferentially on regions with low sequence variation and high read mapping coverage (Table S3.7). Quantitative PCR was carried out as described in "Materials and methods" section of this PhD thesis.

Results

Low number of biological replicates and B chromosomes lead to false positive genes in Bs

We performed exploratory analyses to evaluate different criteria to search for candidate contigs located in the B chromosomes. For this purpose, we compared the theoretical number of contigs corresponding to a B chromosome considering its size with the number of selected contigs meeting different criteria. If the number of contigs identified in B chromosomes was higher that the expected we considered that excess as false positives. We used for this purpose the coverage data of *E. plorans*, and performed additional contig selection to determine the effect of three factors: i) biological replicates (median gFC value vs gFC of 3 biological replicates), ii) high and low coverage filtering and (0.5-4 copies in 0B samples) and iii) number of of B chromosomes (1B vs 4B).

First, we found a reduction of false positives when we used 4B individuals instead of 1B ones in average and also separately (i.e. gFC applied considering samples as biological replicates). Secondly, this decrease was more pronounced applying the gFC threshold to 4B individuals separately (Dataset 3.1). Considering 1B individuals instead of 4B ones also avoids false positives when including biological replicates but their number is still higher compared to the use of 4B samples. The effect of the 0.5-4 copies threshold in 0B individuals did not affect to the number of false positives although the rate of

transposable elements markedly increased from 4 to 24% when retrieving contigs with >4 copies in 0B samples.

For all the above reasons, we decided to consider separately the gFC for every individual, include exclusively contigs with 0B copies between 0.5-4 and use individuals with higher number of Bs (4B instead of 1B). In this way, the gene fraction in B chromosome of *E. plorans* would be around 0.2% and Bs would be depleted in genes around a 92% with respect to the rest of chromosomes if contigs were equally distributed in the genome. This approach prevents false positives and allowed us to potentially identify B-genes with relevant roles for B chromosomes that is the main aim of this study.

Searching for protein-coding genes in the B chromosome

In order to obtain the most complete reference transcriptome of *E. plorans* we sequenced several RNA libraries from individuals at different developmental stages (embryos and adults), belonging to both sexes and having/lacking 1B chromosome, getting in total 47 libraries used for assembling a *de novo* reference transcriptome. We performed one transcriptome assembly separately per each sample category: embryos 0B from a F1BxM0B cross (embryos P1 0B), embryos 0B from a F0BxM1B cross (embryos P2 0B), adult legs 0B, adult gonads 0B, embryos 1B from a F1BxM0B cross (embryos P1 1B), embryos 1B from a F0BxM1B cross (embryos P2 1B), adult legs 1B, adult gonads 1B. Then we extracted CDSs and clustered each assembly with CDHit-EST. Finally, we joined together assembled CDSs from 0B samples and reduce redundancy in a last run of clustering to get a 0B reference set of CDSs and we did the same for 1B CDSs assemblies (see Fig. S3.2). This workflow of independent assembling and subsequent clusterings requires lower computational power than a whole transcriptome assembly. In addition, this procedure allows us to identify CDSs derived from 1B samples and directly getting the most complete transcripts (usually coming from 0B assemblies as transcripts from Bs are often truncated).

De novo transcriptome assembly of the 23 0B RNA libraries of *E. plorans* (female/male embryos, leg and gonads of females and males adults) yielded 66,646 CDSs longer than 300nt (N50= 1,191 nt) whereas from the 1B reference transcriptome (made up with 24 RNA libraries) we got 64,019 CDSs (N50= 122,1 nt). We compared these statistics with parameters of a global assembly using all libraries as a whole and values improved when assembling different biological samples independently. For example, the N50

increases from 504 nt in non-redundant CDSs for the global assembly to 1,191 and 1,221 nt in the set of CDSs (using 0B and 1B samples respectively) after independent assembling and clustering. After gDNA mappings to both reference transcriptomes (0B and 1B) we got counts for 64,776 and 62,373 CDS contigs respectively. We selected for subsequent analysis CDS contigs showing 0.5-4 copies in 0B gDNA libraries thus retaining 31,926 and 30,885 contigs from 0B and 1B reference transcriptomes respectively (see Fig. S3.2).

After gDNA SSAHA2 mappings and genomic coverage fold change calculations in 4B and 1B gDNA libraries[gFC= log2(4Bor1B/0B)], we selected 360 contigs showing gFC>1.585 in all three 4B males analyzed or a gFC>1 in all 4B libraries (and a gFC>0.585 in the average of 1B libraries in the last case) thus being potentially located in the short arms of Bs, as expected since 4B males were actually 2B+1isoB (Dataset 3.2). Finally, 142 of these contigs were annotated as different protein-coding genes (Table S3.1).

The transcript sequences of the 142 genes selected were used as reference to separately map the reads from each of all 0B and +B gDNA and RNA libraries (we included extra 0B and +B libraries to increase gDNA coverage in SNP calling, see Methods). After analysis of per nucleotide sequence coverage graphics and gFCs values we discarded 6 sequences for being assembling artifacts and 97 that showed gFC<1.585 at least one 4B individual or copy number higher than 4 in 0B libraries, see Table S3.2. We also retained for posterior analysis those genes with a gFC>1 in 4B individuals and a gFC>0.585 in 1B individuals; *kif20a* was kept since they were validated by qPCR in Navarro-Domínguez et al. (2017) although here the number of copies in 0B individuals was above 4.

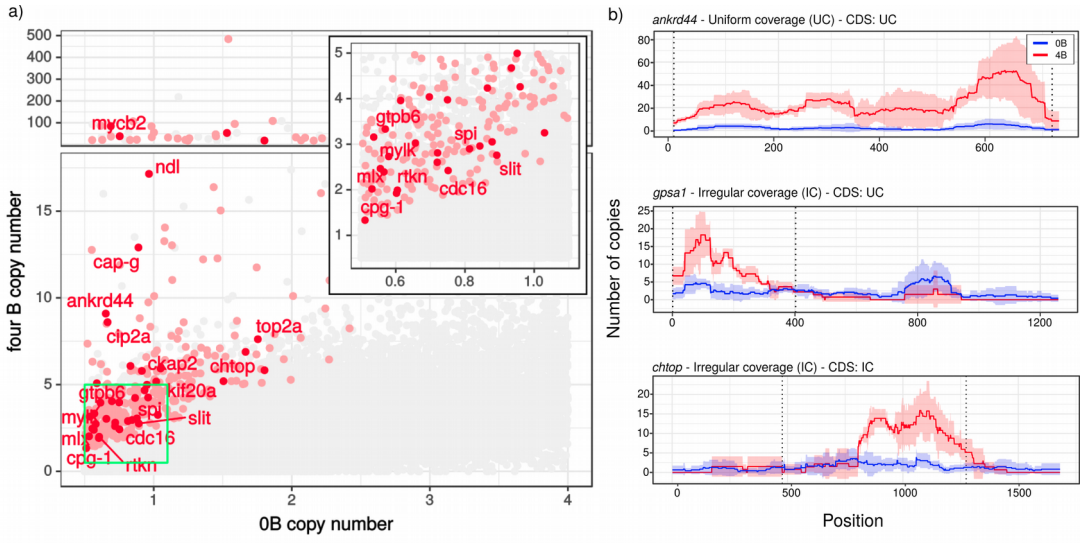


Figure 3.1. Detection of protein-coding genes residing in the B chromosome of *Eyprepocnemis plorans* by mapping gDNA Illumina reads obtained from three 4B-carrying and six B-lacking males on a *de novo* transcriptome built with Illumina reads coming from female and male embryos, legs and gonads RNA libraries for *E. plorans*. (a) Scatterplot showing the mean copy number for 62,811 CDSs in 0B and 4B males of *E. plorans*, with enlargement of the region included in the green box (inset). The CDSs that in presence of B chromosomes showed a genomic fold change [gFC= log2(4B/0B)] higher than 1.585 or 1 (requiring gFC>0.585 in the mean of 1B males) in the three 4B males of *E. plorans* (142 contigs, see Dataset 3.1) are shown in light red color, and those failing to reach this thresholds are noted in gray color. In addition, the 42 protein-coding genes which were validated by full transcript analysis are noted in dark red color for the contig showing the highest gFC value. (b) Examples of the coverage patterns found in B-genes of *E. plorans*: UC= uniform coverage, IC= irregular coverage, for both full transcript and CDS.

Therefore, the final list of B-located genes in *E. plorans* included 42 genes that regarding coverage variation along gene sequence can be classified in 14 UC genes (none of them showed a IC CDS) and 28 IC genes (5 of them with a UC CDS) (Dataset 3.3, Fig. 3.1). Interestingly, four of these genes (*ank2*, *gar-2*, *krba2* and *smf*) came exclusively from the 1B reference transcriptome so due to our assembly workflow we were able to retrieve them.

Table 3.1. Basic structural features of the 42 B-genes of *E. plorans*. Average genomic fold change (Av_gFC) was calculated as log2 the 4B/0B quotient between the mean nucleotidic coverage along transcript length. The distribution pattern of coverage along transcript length was determined by the proportion of nucleotides more abundant in 4B libraries than in 0B ones. Proportion higher than 0.90 indicated uniform coverage suggesting that the CDS is complete in the B chromosome (UC pattern). Values lower than 0.90 indicated irregular patterns (IC).

Gene	CDS assembling	Av_ gFC	Copies in B chr.	Coverage pattern		gDNA (copy numbers)									
						HC region					LC region				
				Gene	CDS	0B		4B		gFC	0B		4B		gFC
						Mean	SD	Mean	SD		Mean	SD	Mean	SD	
ank2	Lacks 5' end	4.96	29.94	UC	UC	1.98	0.75	61.87	0.61	4.96					
mycb2	Lacks 5' end	4.81	17.32	UC	UC	1.28	0.36	35.93	8.34	4.81					
mtg1	Complete	4.62	15.38	UC	UC	1.30	0.66	32.07	7.60	4.62					
ndl	Lacks 3' end	4.11	12.96	IC	IC	1.24	0.28	27.17	6.02	4.45	0.84	0.41	0.85	0.34	0.01
cap-g	Complete	3.70	8.08	IC	UC	0.92	0.55	17.08	5.79	4.22	0.83	0.67	1.30	0.68	0.66
ckap2	Lacks 3' end	2.71	3.96	IC	IC	0.80	0.48	8.71	3.98	3.45	1.11	0.27	2.79	1.03	1.33
ankrd44	Lacks 5' end	3.33	10.54	UC	UC	2.33	1.72	23.40	10.54	3.33					
qvr	Complete	0.10	1.27	IC	IC	0.30	0.67	2.83	1.18	3.24	1.05	0.64	0.82	0.38	-0.36
cip2a	Complete	3.21	4.32	UC	UC	1.04	0.32	9.67	3.10	3.21					
cyp6k1	Lacks 5' and 3' ends	0.95	3.84	IC	IC	0.96	1.15	8.64	2.54	3.16	3.48	0.66	3.42	1.79	-0.03
slit	Complete	1.69	3.72	IC	IC	1.04	0.54	8.49	0.33	3.02	1.12	0.28	1.52	0.41	0.44
mef1	Lacks 5' and 3' ends	1.78	1.81	IC	IC	0.54	0.67	4.15	1.33	2.95	1.10	1.21	0.08	0.04	-3.82
c0j52_04360	Lacks 5' end	1.91	2.68	IC	UC	0.83	1.04	6.19	2.62	2.90	0.60	0.36	0.62	0.62	0.03
gtpb6	Complete	2.89	3.70	UC	UC	1.16	0.48	8.55	3.17	2.89					
gf14c	Lacks 5' end	2.88	2.31	UC	UC	0.73	0.97	5.35	2.34	2.88					

slv	Complete	2.85	5.09	UC	UC	1.64	1.48	11.83	6.85	2.85					
mylk	Lacks 3' end	1.44	1.48	IC	IC	0.52	0.53	3.48	0.97	2.74	1.06	0.41	1.03	0.39	-0.04
cdc16	Complete	2.12	2.43	IC	IC	0.90	0.33	5.76	2.20	2.68	0.63	0.51	0.73	0.56	0.20
sgp1	Complete	0.53	2.54	IC	IC	1.09	1.00	6.18	0.78	2.50	0.92	0.36	0.61	0.22	-0.59
chtop	Complete	1.50	4.30	IC	IC	1.85	0.86	10.45	2.35	2.50	1.18	0.66	0.91	1.18	-0.38
gar-2	Lacks 5' and 3' ends	2.33	2.13	UC	UC	1.06	0.69	5.33	1.43	2.33					
gas8	Complete	1.03	2.46	IC	UC	1.30	1.07	6.23	1.06	2.26	1.34	0.61	1.03	0.70	-0.38
kpyc	Lacks 5' end	2.26	1.74	UC	UC	0.92	1.25	4.40	1.48	2.26					
or85c	Lacks 5' end	1.39	1.82	IC	IC	0.98	0.32	4.62	0.44	2.24	0.69	0.29	0.62	0.50	-0.14
top2a	Complete	2.13	3.26	IC	UC	1.77	0.91	8.28	2.77	2.23	0.86	0.67	1.16	0.38	0.44
or9	Lacks 3' end	1.83	1.76	IC	IC	0.96	1.05	4.48	0.51	2.22	0.46	0.67	0.42	0.20	-0.14
smf	Lacks 5' end	2.17	2.00	IC	IC	1.15	2.14	5.15	1.44	2.16	0.00	0.00	0.03	0.06	--
wdr49	Lacks 5' and 3' ends	2.15	2.42	UC	UC	1.41	1.84	6.26	1.53	2.15					
mlx	Lacks 3' end	1.59	1.77	IC	IC	1.09	0.75	4.63	0.90	2.09	0.58	0.57	0.14	0.03	-2.02
stum	Lacks 5' end	1.21	2.06	IC	IC	1.28	0.89	5.40	0.65	2.08	1.34	0.96	1.73	0.60	0.37
ormdl2	Lacks 5' and 3' ends	1.38	1.28	IC	IC	0.79	1.55	3.35	0.24	2.08	4.81	3.31	4.70	2.53	-0.03
25kd	Lacks 3' end	1.17	1.67	IC	IC	1.04	0.88	4.39	1.71	2.08	1.29	0.77	1.14	1.03	-0.18
cg32809	Lacks 5' and 3' ends	2.02	1.68	UC	UC	1.10	0.96	4.46	1.84	2.02					
cpg-1	Lacks 3' end	1.56	1.64	IC	IC	1.09	0.60	4.38	0.72	2.00	0.61	0.18	1.40	0.68	1.19
gpsa1	Lacks 5' end	0.51	4.01	IC	UC	2.67	0.65	10.68	1.88	2.00	2.16	0.93	1.01	0.64	-1.09
spi	Lacks 5' end	-0.85	1.31	IC	IC	0.96	0.36	3.57	0.38	1.90	0.93	0.62	0.34	0.36	-1.47
znf423	Lacks 3' end	0.79	1.66	IC	IC	1.23	0.76	4.55	0.74	1.89	1.47	1.06	1.46	0.64	-0.01

idgf4	Lacks 5' end	1.45	2.34	IC	IC	1.75	0.91	6.44	1.28	1.88	1.37	0.48	2.44	0.47	0.83
rtkn	Lacks 3' end	1.28	1.97	IC	IC	1.53	1.49	5.47	1.24	1.84	1.39	1.48	1.81	1.65	0.38
krba2	Lacks 5' end	1.80	0.95	UC	UC	0.77	0.51	2.68	0.47	1.80					
nphs1	Lacks 5' end	0.38	3.36	IC	IC	2.77	1.81	9.48	1.06	1.78	2.75	1.99	2.72	0.79	-0.01
kif20a	Lacks 5' end	1.29	2.95	UC	UC	4.08	2.51	9.98	6.35	1.29					

Since total coverage was low in our experiments, we are likely missing genes located in B chromosomes but still some genes found here show important KOGs function that could be determinant for the Bs maintenance and evolution (see Table 3.2).

Table 3.2. KOG annotation for the genes found in the B chromosome of *E. plorans.*

Gene	Hit	Class	Class description
25kd	KOG4169	Q	Secondary metabolites biosynthesis, transport, and catabolism
ank2	KOG4177	M	Cell wall/membrane/envelope biogenesis
	KOG0510	J	Translation, ribosomal structure and biogenesis
	KOG4412	O	Post-translational modification, protein turnover, and chaperones
ankrd44	KOG0504	E	Amino acid transport and metabolism
	KOG4177	M	Cell wall/membrane/envelope biogenesis
cap-g	KOG2025	B	Chromatin structure and dynamics
		D	Cell cycle control, cell division, chromosome partitioning
cdc16	KOG1173	D	Cell cycle control, cell division, chromosome partitioning
cip2a	KOG0161	Z	Cytoskeleton
cyp6k1	KOG0158	Q	Secondary metabolites biosynthesis, transport, and catabolism
gar-2	KOG4220	T	Signal transduction mechanisms
gas8	KOG0161	Z	Cytoskeleton
gf14c	KOG0841	O	Post-translational modification, protein turnover, and chaperones
gpsa1	KOG2711	C	Energy production and conversion
gtbp6	KOG0410	G	Carbohydrate transport and metabolism
idgf4	KOG2806	G	Carbohydrate transport and metabolism
kif20a	KOG0247	Z	Cytoskeleton
	KOG0242	Z	Cytoskeleton
	KOG0161	Z	Cytoskeleton
kpyc	KOG2323	G	Carbohydrate transport and metabolism
krba2	KOG0017	C	Energy production and conversion
mef1	KOG0465	J	Translation, ribosomal structure and biogenesis
mlx	KOG1319	K	Transcription
mtg1	KOG2485	P	Inorganic ion transport and metabolism
mycb2	KOG1428	O	Post-translational modification, protein turnover, and chaperones
	KOG1869	A	RNA processing and modification
mylk	KOG0032	T	Signal transduction mechanisms

ndl	KOG3627	O	Post-translational modification, protein turnover, and chaperones
	KOG1215	T	Signal transduction mechanisms
nphs1	KOG3515	O	Post-translational modification, protein turnover, and chaperones
	KOG4221	T	Signal transduction mechanisms
ormdl2	KOG3319	P	Inorganic ion transport and metabolism
slit	KOG4237	T	Signal transduction mechanisms
	KOG4289	T	Signal transduction mechanisms
	KOG0472	L	Replication, recombination and repair
slv	KOG1623	K	Transcription
smf	KOG3482	A	RNA processing and modification
stum	KOG4717	T	Signal transduction mechanisms
top2a	KOG0355	B	Chromatin structure and dynamics
wdr49	KOG0266	K	Transcription
znf423	KOG3623	K	Transcription
	KOG2462	K	Transcription
	KOG1074	K	Transcription

B-specific sequence variation

To ascertain whether B-chromosome genes carry sequence signatures, we searched for DNA sequence variation being specific to B-carrying individuals and thus to the B chromosome gene paralogs. Specifically, we searched for nucleotide variations being present in B-carrying gDNA libraries as expected depending on B chromosome number, but absent in the gDNA and RNA libraries from 0B samples.

This yielded 180 single nucleotide polymorphisms (SNPs) representing B-specific variation (Alt variant) in 17 B-genes (Tables 3.3 and 3.4), 121 of which were found in CDS regions whereas the remaining 59 were in UTR regions. Remarkably, the genes showing IC pattern, as a whole, did not differ from those showing UC pattern for the number of synonymous (dS) and non-synonymous (dN) substitutions or divergence per nucleotide site (ps) in the CDS (two-sided permutation t-test: $p > 0.05$).

Table 3.3. Nucleotidic variation found in the B chromosome gene paralogs. To score the number of variants, we distinguished between substitutions (s), deletions (d) or insertions (i) in respect to the A chromosome paralogs. We considered deletions and insertions as a single mutation event whether they involved 1-15 nt. dS= number of synonymous substitutions; dN= number of non-synonymous substitutions; ps= proportion of variable sites.

Gene	CDS length	Number of variants				dS	dN	dN/dS	ps
		Total	In 5'-UTR	In 3'-UTR	In CDS				
ank2	333	3s	0	1s	2s	0	2		0.0060
ankrd44	717	2s	0	0	2s	1	1	1.00	0.0028
cap-g	3669	2s	0	0	2s	0	2		0.0005
cdc16	1758	6s+1d	2s	2s+1d	2s	0	2		0.0011
cip2a	2475	6s+1d	2s+1d	1s	3s	2	1	0.50	0.0012
ckap2	2085	1s	0	0	1s	0	1		0.0005
gf14c	536	1s	0	0	1s	1	0	0.00	0.0019
gtpb6	1557	27s+1d	2s	15s+1d	10s	4	6	1.50	0.0064
kif20a	1365	1s	0	0	1s	0	1		0.0007
kpyc	459	1s	0	0	1s	1	0	0.00	0.0022
mtg1	966	3s	1s	1s	1s	0	1		0.0010
mycb2	2673	15s+2d	0	5s+2d	10s	2	8	4.00	0.0037
ndl	6237	56s+6d+1i	5s	0	51s+6d+1i	25	33	1.32	0.0093
ormdl2	402	2s	0	0	2s	0	2		0.0050
slit	4404	4s	0	2s	2s	1	1	1.00	0.0005
slv	657	15s+2d	5s+1d	5s+1d	5s	5	0	0.00	0.0076
top2a	4767	19s+2d	2s	1d	17s+1d	4	14	3.50	0.0038
Total	--	180	21	38	121	46	75	1.63	--

In addition, considering count of Ref variants found in +B gDNA libraries with respect to Ref counts in the 0B ones, we calculated the number of Alt and Ref variants of a gene harbored in the B chromosomes. Interestingly, most of B-genes have at least 0.5 copies of Ref variants located in the B chromosome, the exception to this were *cip2a*, *gtpb6*, *kpyc*, *ormdl2* and *slv* (Table S3.5). This is important as Ref variants of genes located in B chromosomes are the best candidates to preserve their function, enhancing or weakening the global effect of that gene.

Sequence analysis of the 42 B-genes revealed high variation between them for the number of nucleotide differences in respect to the A chromosome paralogs. At one

extreme, there were 25 genes showing no differences at all, and 4 showing a single difference. On the other hand, 8 genes showed from 2 to 7 SNPs and 5 B-genes showed high number of nucleotide difference compared to A paralogs, from 17 to 64 (see Table 3.3 and Dataset 3.3). We interpret these results as reflecting differences in the time of gene arrival to the B chromosome, the oldest genes accumulating a higher number of substitutions in respect to the A chromosome paralogs, and the youngest genes being those showing few or no changes at all, in resemblance with the evolutionary strata found on sex chromosome differentiation (Lahn and Page, 1999) and the evolution of the germ-line restricted chromosome in songbirds (Kinsella et al., 2019). It is plausible to assume that the most important genes for the B chromosome should be those that arrived first, as one or more of them should be crucial for the drive mechanisms that guaranteed the initial invasion of this parasitic element (Camacho et al., 1997).

We validated the list of 42 B-located genes by qPCR amplification selecting key genes related to cell cycle and sex development that showed a gFC value higher than the threshold in different numbers of 4B males analyzed. Results from qPCR of *tsl* and *tssk6* genes (gFC>1.585 in only 2 males 4B after mapping to the CDSs) did not validate the presence of this genes in Bs. *Foxc2* and *drm* genes show a gFC>1.585 in all 4B individuals after CDS mappings but gFC<1.585 in some of the three 4B males after calculations using coverage of full transcripts and we could not validate their presence in B chromosomes through qPCR. In contrast, B presence validation of *cdc16*, *cpg-1*, *ndl* and *rtkn* was successful after qPCR (Table S3.2 and Fig. S3.1). Interestingly, the gene *wtap* was validated by qPCR although it surpassed the gFC threshold in only one 4B individual so we decided to excluded it from the list. Genes that were validated by qPCR showed a gFC>1.585 after transcript coverage analysis in all 4B libraries so we considered as validated the list of 42 genes indicated above. B-genes described in Navarro-Domínguez et al. (2017) are within the gFC range of validated genes except *kif20a* that was included in the list because it was qPCR validated by the authors and *hyi* that was discarded of the gene selection process since its gFC was below the threshold in all 4B individuals (see Table S3.3 for final list of 42 B-genes).

The analysis of qPCR results showed that this technique could be extremely variable between individuals causing false positives (i.e. *wtap*) or negatives. Although several B-genes included in the final list of both species were validated by qPCR, our filtering approach requiring the surpass of the gFC threshold in all +B individuals separately was

probably robust enough to validate the presence of genes in B chromosomes, moreover when they showed B specific variation.

B chromosomes from distant population of *E. plorans* share similar gene content

We mapped the sequences of the 42 genes identified in the B chromosomes of *E. plorans* from Torrox (Spain) to gDNA libraries from distant population of the same species in order to ascertain the number of genes present in the B chromosomes from that regions. Interestingly, we found 29 out of the 42 B-genes of Torrox in the B chromosomes of *E. plorans* from Tanzania and Egypt while 25 B-genes were identified in *E. plorans* from Armenia (see Table S3.8). Furthermore, 16 out of the 42 B-genes (*25kd*, *ankrd44*, *cap-g*, *cdc16*, *cip2a*, *ckap2*, *gas8*, *gpsa1*, *gtpb6*, *idgf4*, *kif20a*, *mtg1*, *mycb2*, *ndl*, *or85c* and *spi*) were present in the B chromosomes of *E. plorans* from the four populations here analyzed (i.e. Torrox, Tanzania, Egypt and Armenia). In addition, 90 out of the 180 SNPs identified for B-genes found in Torrox were also present in the 1B library of Tanzania, 96 were found in +B males from Egypt and only 7 of them appeared in the B-carrying male from Armenia (see Table S3.4).

Considering the same high coverage regions for B-genes in all *E. plorans* populations (i.e. those found in Torrox, Table S3.3), there was a significant positive correlation between the gFC of B-genes from Torrox with that of Tanzania (r_S= 0.61725, p= 0.00006) and Egypt (r_S= 0.35135, p= 0.03054). However, that correlation was not significant in the case of B chromosomes from Armenian population of *E. plorans* (r_S= 0.08867, p= 0.60176). In spite of this, the correlation between the gFCs of B-genes from Tanzania, Egypt and Armenia was significant (Tanzania and Egypt: r_S= 0.72605, p< 0.00001; Tanzania and Armenia: r_S= 0.40364, p= 0.01619; Egypt and Armenia: r_S= 0.36062, p= 0.03072). Therefore, the B chromosomes from distant populations of *E. plorans* harbor similar gene content which implies a common origin of B chromosomes in this species.

Finally, *de novo* SNP calling for B-genes of *E. plorans* from African and Asian populations yielded 132 SNPs in 24 B-genes of *E. plorans* from Tanzania (5.5 SNPs/gene, Table S3.9), 93 of them for 18 B-genes in Egypt (5.17 SNPs/gene, Table S3.10) and 56 SNPs for 17 B-genes in Armenia (3.29 SNPs/gene, Table S3.11). This result suggests that B chromosomes from Tanzania are older than those of *E. plorans* from Egypt or Armenia which is congruent with the findings about repetitive DNA content in the B chromosomes of *E. plorans* from these populations (Chapter 2 of this thesis).

Activity of B chromosome gene paralogs

A first indication of gene transcription in the B chromosome was given by the fold change observed in embryos, testis and hind leg transcriptomes [tFC= log2(1B/0B) RNA], showing that several contigs being over-represented in the B-carrying gDNA libraries were also over-represented in the transcriptomes of *E. plorans* (Fig. 3.2 and 3.3). In fact, 23 out of 42 B-genes in *E. plorans* showed tFC>1 in embryo, gonad or leg transcriptomes (Table S3.3). An overview of the tFCs (i.e. log2[RNA count 1B/RNA count 0B]) for the 42 B-genes found in the B chromosomes of *E. plorans* revealed a higher number of B-genes showing tFC>1 in gonads (16 genes) than in legs (9 genes). For embryos, there were 11 B-genes with a tFC>1 in those embryos coming from P2 and the number decreased to 6 B-genes in embryos from P1. This results suggest that B-genes are transcriptionally active specially in gonads of 1B individuals.

In addition, we clustered different biological samples in a heatmaps in terms of RNA counts for Alt and Ref variants of B-genes and they grouped together in first place depending on tissue (embryos, gonads or legs). Interestingly, gonads and embryos formed a group apart from legs (Fig. 3.4a), suggesting that the expression pattern of B-genes in the latter tissue would be different than that from gonads or embryos. Then, biological samples clustered based on whether they carried or not 1B chromosomes. Embryos split into to groups depending on B chromosome presence instead of pod origin (P1 or P2 embryos) although afterwards they were divided meeting that criteria, thus the pattern of B-genes expression in embryos is more affected by the own presence of B chromosome than by the different origin of embryos. In the case of gonads, we observed that 0B gonad from both sexes and 1B testis group together apart from 1B ovaries suggesting specific expression patterns in ovaries due to B chromosomes presence which would be in concordance with the determining function of this tissue for drive of B chromosomes in *E. plorans*.

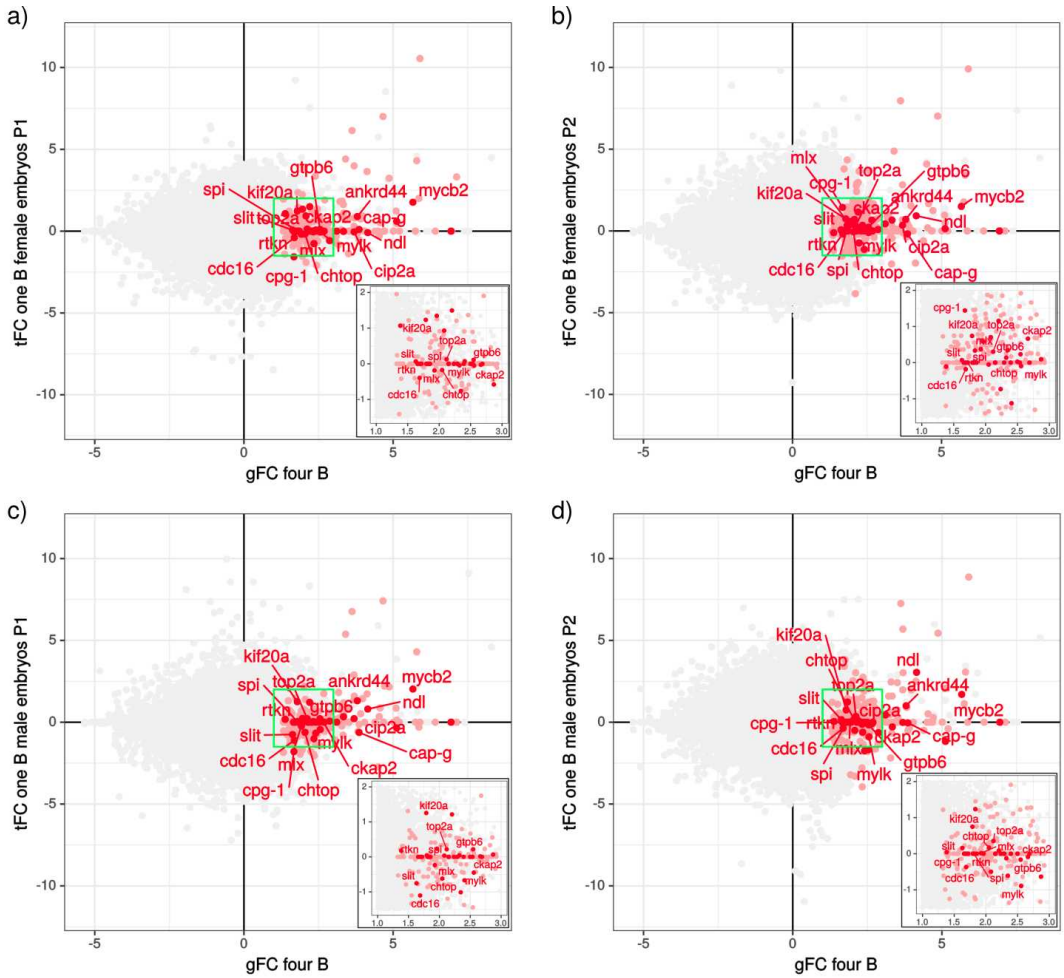


Figure 3.2. Transcription of B chromosome genes in female embryos from P1 (a) and P2 (b), and males embryos from P1 (c) and P2 (d) of *E. plorans*. Transcriptional fold change [tFC= log2(1B/0B)] observed in B-carrying RNA libraries (Y-axis) in respect to gFC in 4B males (X-axis). Color codes in A and B are as in Figure 3.1. The green box is enlarged in the inset. Note that many B-genes showed tFC≥1.

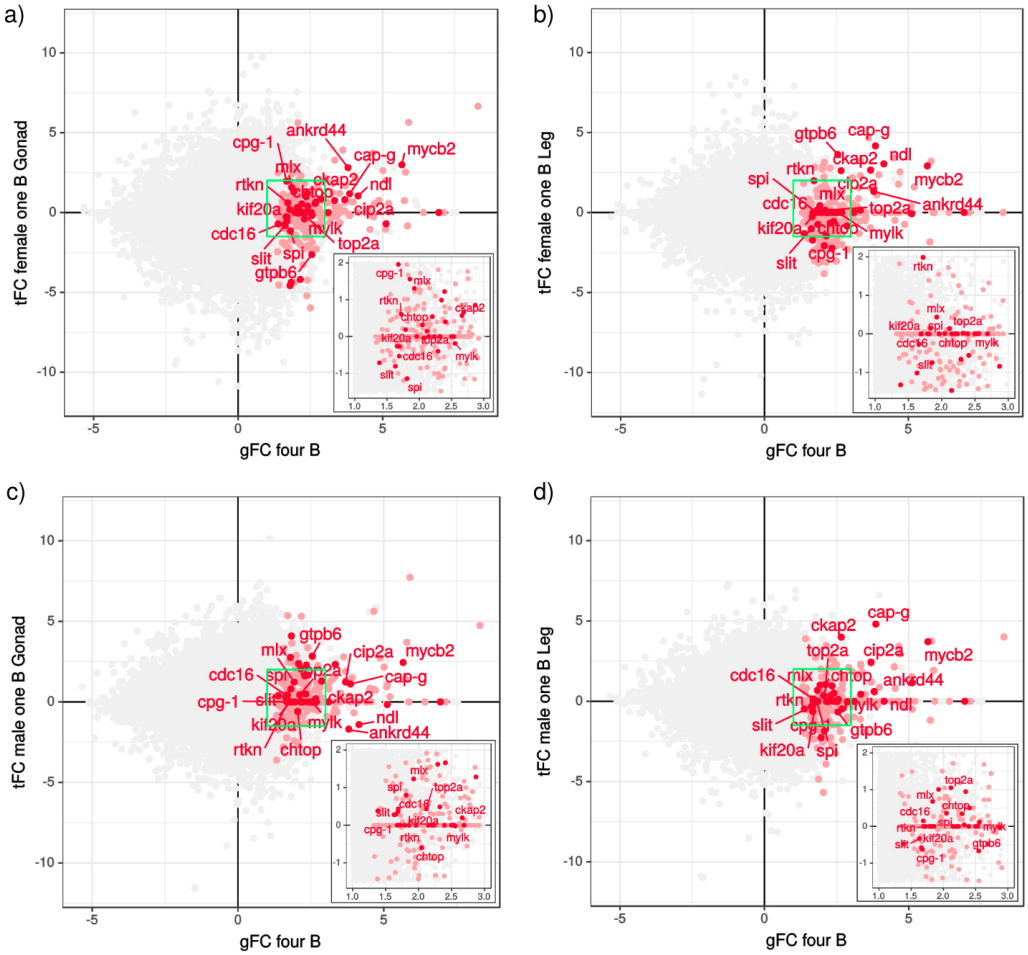


Figure 3.3. Transcription of B chromosome genes in ovary (a), females legs (b), testis (c) and male legs (d) of *E. plorans*. Transcriptional fold change [tFC= log2(1B/0B)] observed in B-carrying RNA libraries (Y-axis) in respect to gFC in 4B males (X-axis). Color codes in A and B are as in Figure 3.1. The green box is enlarged in the inset. Note that many B-genes showed tFC≥1.

We analyze the transcriptional activity of genes present in the B chromosomes by comparing counts in RNA-seq libraries from different tissues and sexes for all the SNPs using the fold change for the Alt/Alt ratio (aFC) and the Ref/Ref ratio (rFC) in 1B individuals of *E. plorans*. Additionally, we tested if B-paralogs showed a higher expression than A-paralogs calculating the fold change for the Alt/Ref ratio (arFC). Thus, we considered as subverted genes those with higher expression of Alt alleles than the Ref ones in 1B samples, i.e. an arFC>0.

Transcriptional activity of B-genes in embryos of E. plorans

In the case of embryos, 2 out of 17 genes (*ank2* and *ndl*) showed aFC>1 in male embryos from P2 compared with those from P1 whereas one gene (*ndl*) did it in the case of females embryos. Concerning the rFC in 1B embryos we found 3 genes (*ank2*, *ckap2* and *slv*) with rFC>1 in male and female embryos from P2 respect to those of P1. Considering each pod of embryos separately, none of the B-genes showed a rFC>1 in any of the group of embryos (neither P1 or P2 embryos) between females and males. However, we found 2 out of 17 genes (*ankrd44* and *ndl*) showing higher expression (aFC>1) of Alt alleles in 1B males than in 1B females in embryos from P1 and the same number of genes doing so in embryos from P2 (*ndl* and *slit*).

Regarding expression of Alt alleles with respect to Ref ones in 1B embryo libraries, *ank2* in females embryos from P1 was the only gene showing arFC>1 we found considering all comparisons in all 1B samples (females and males from P1 and P2 embryos). In contrast, the *cip2a* gene appeared as a subverted gene (i.e. arFC>0), thus showing higher expression of Alt alleles than Ref ones in all 1B embryos (see Fig. 3.4b).

As suggested by these results, expression of B-genes is quite similar between pods and sexes although we found that embryos from P2 showed higher expression of the *ndl* cell cycle gene that those from P1.

Transcriptional activity of B-genes in adults of E. plorans

When we looked at adults, we found an aFC>1 in testis respect to male leg when carrying 1B in 6 out of the 17 genes whereas 4 genes showed an aFC>1 in ovary with respect to females leg. The same comparison but for rFC in 1B libraries yielded 9 genes highly expressed between testis and male legs whereas 12 genes showed higher transcription of Ref alleles in ovary compared to leg in females. Remarkably, the Alt allele for 2 genes showed an aFC>1 in female legs compared to the male ones in contrast to 6 genes that

showed it in ovary compared to testis (Table S3.6). Expression of the Ref allele in 1B samples shows a rFC>1 for 9 out of 17 genes between testis and male leg and in 12 out 17 comparing ovary to female leg (Table S3.6). On the other hand, arFC was higher than 1 for 3 genes in males, *cap-g*, *kif20a* and *mycb2*, and 4 genes in females, same than in males and *top2a*, in 1B leg samples. Regarding gonads, we found 2 genes in 1B testis showing an arFC>1, *cip2a* and *cap-g,* and also 2 genes above this tFC value in 1B ovaries, *cap-g* and *ndl* (see Fig. S3.4 and Table S3.6). Considering genes with arFC>0 in 1B samples, i.e. subverted genes, we found six of these genes in *E. plorans* male legs (*ankrd44*, *cap-g*, *cip2a*, *kif20a*, *mycb2* and *top2a*) and 4 in female legs (*cap-g*, *kif20a mycb2* and *top2a*). Likewise, we found 4 subverted genes in *E. plorans* testis (*cap-g*, *cip2a*, *mycb2* and *ndl*) and 3 in ovary (*cap-g*, *mycb2* and *ndl*). From the total of genes showing Alt expression in *E. plorans*, *cap-g* was the one showing the higher arFC ratio in 1B libraries considering the total set of RNA samples. However, when looking exclusively at ovary, the gene *ndl* surpassed this value (see Fig. 3.3b and S3.5). This tendency could suggest an important role of *ndl* B-transcripts in the performance of B chromosomes in the ovary, the tissue in which their drive takes place.

Transcriptional activity of B-genes in E. plorans comparing between tissues

Almost all B-genes showed expression of Alt variants for several SNPs in all 1B tissues. However, the number of SNPs for which we found Alt expression reached the maximum in ovary with Alt expression for 130 SNPs while in testis we found expression of Alt variants in 64 SNPs. In the case of legs, 78 and 75 SNPs showed Alt expression in females and males respectively. In embryos, we found counts of Alt variants in 55 and 57 SNPs for female and male embryos from P1 respectively, and in 52 SNPs for female embryos and 59 for male P2 embryos. Again, these findings suggest that it is in the ovary of *E. plorans* where genes copies located in the B chromosomes are being transcribed in a greater extent compared to other tissues.

We also analyzed aFC, rFC and arFC values between tissues without considering different sexes, thus averaged expression of Alt and Ref in 1B embryos, adults legs and gonads. In that respect we found that 9 out of 17 B-genes showed higher expression of Alt alleles (i.e. aFC>1) in gonads compared to legs while the aFC was higher than 1 for 4 B-genes in gonads compared to embryos and 6 genes did it in embryos respect to adults legs. On the other hand, for 10 out of 17 B-genes the expression of Ref variants was higher (i.e. rFC>1) in gonads than in legs. Likewise, 5 B-genes showed a rFC>1 in gonads

compared to embryos whereas 10 out of 17 B-genes did it in embryos with respect to legs. When considering the arFC>1, we found that one gene exhibited higher expression of Alt alleles than Ref ones in gonads and embryos, the *ndl* and *ank2* genes respectively, whereas 3 B-genes did it in legs (*cap-g*, *kif20a* and *top2a*). In general, we could state that Alt alleles showed a slightly higher expression in gonads followed by embryos and then in legs.

Finally, when comparing the expression of Ref alleles in 1B samples to 0B ones we noticed that most of the genes showed higher expression in 1B individuals than in 0B ones (see genes in blue and green in Table S3.6) which implies that the presence of B chromosomes causes changes in the expression of genes regardless whether they came from the A chromosomes or from copies located in the Bs.

In total, we found expression of B specific copies (Alt variants) for 14 out of 17 genes showing SNPs in at least one condition. In addition, subverted genes in 1B individuals keep this turnover in expression when comparing Alt counts in 1B samples to Ref counts in 0B ones, pointing out a general over-regulation of these genes when the B chromosomes is present in the genome (see Figure S3.4). Taken together, these results indicate that the B chromosome is not completely silenced but, on the contrary, shows high levels of transcription, especially in gonads for both grasshopper species (Fig. 3.3. *cdc16* and *ndl* bar graphs of RNA counts, figures for each gene can be found in Datasets 3.4).

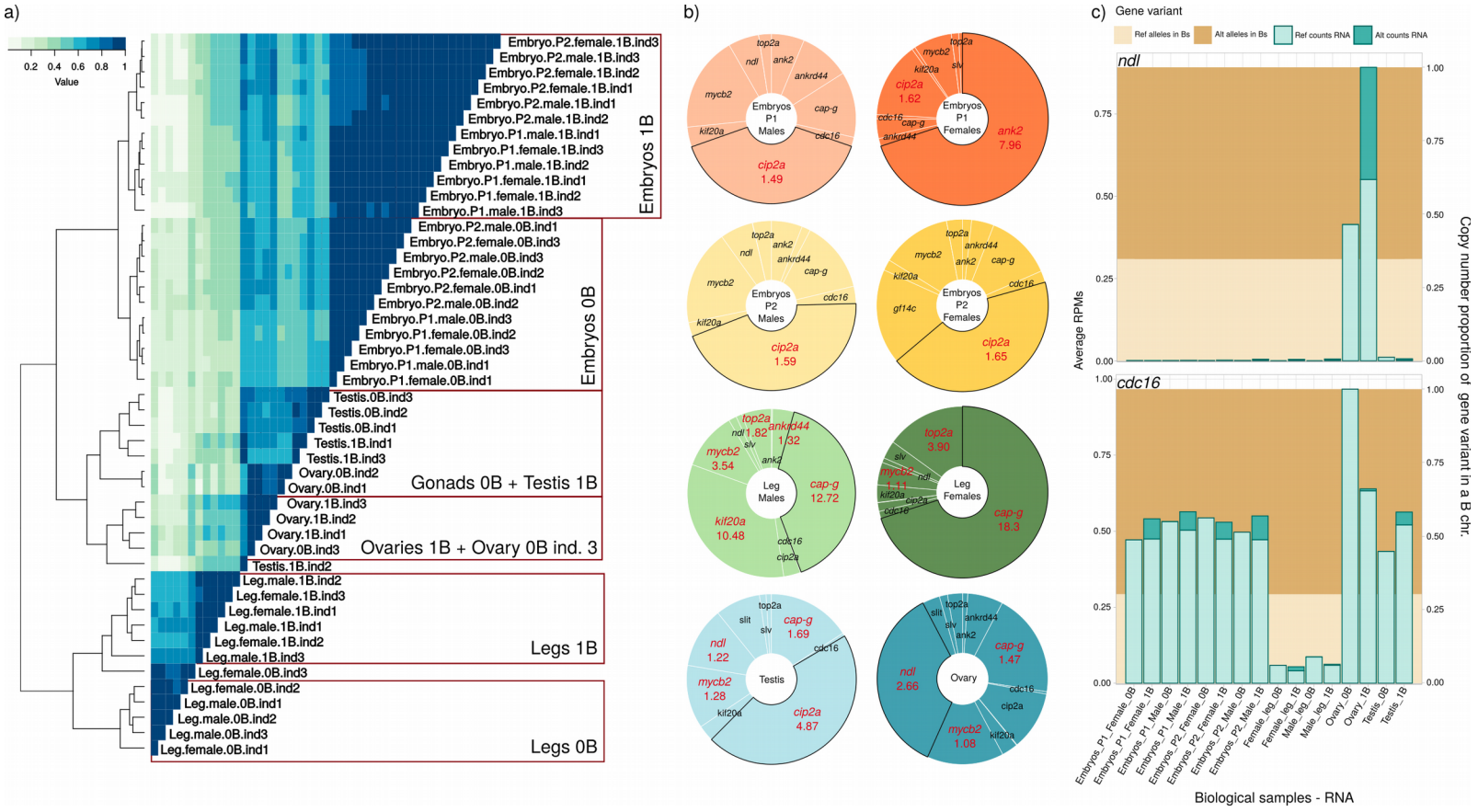


Figure 3.4. (a) Heatmaps of RNA samples using counts of Alt and Ref variants for B-located genes of *E. plorans*. Samples are divided based firstly on tissue and then in terms of B chromosome presence. (b) Pie graphs representing the proportion of Alt/Ref count in 1B libraries for each B-gene respect to the total of genes. Subverted genes (arFC>0) are highlighted in red and showing the value of the Alt/Ref ratio. For further detailed pie graphs see Figure S3.5. (c) See bar graphs of example cell cycle genes, *cdc16* and *ndl,* showing counts (reads per million) in all RNA samples analyzed, embryos, legs and gonads from both sexes of *E. plorans*. At the background of each graph we represent the proportion of Alt and Ref variants in a B chromosomes in respect to the total of copies of that gene in the B. Note that in gonads we found high expression of genes, in particular of the Alt variant.

Discussion

We uncover here the presence of 42 protein-coding genes in the B chromosomes of *E. plorans,* 9 of them already described in Navarro-Domínguez et al. (2017). Although these authors described an important set of B-genes in *E. plorans*, here we notably swell that list putting into evidence the importance of increasing genomic coverage in the purpose of B-genes searching. Navarro-Domínguez et al. included one 0B and one 4B gDNA libraries and were not able to perform SNP calling, here we include six 0B, three 1B and three 4B males for CDSs abundance estimation and thirteen 0B, six 1B and four 4B gDNA libraries for SNP calling what allowed us to find 180 SNPs in the B-genes of *E. plorans*. In addition, our comprehensive *de novo* transcriptome, built from 47 RNA libraries, also helped to increase the list of B-genes compared with the previous one (Navarro-Domínguez et al. used two females, one 0B and one 1B). Of course, this is by no means the complete set of genes located in B chromosomes, and future research with higher coverage will surely uncover some more B-genes in *E. plorans*. However, our analysis about false positives (Datasets 3.1) showed that this list is likely free of false positives protein-coding genes as we included for identification of B genes several biological replicates (3 males of *E. plorans*), high number of B chromosomes (individuals with 4Bs instead of 1B chromosome) and the initial set of contigs was delimited to the ones with 0.5-4 copies in 0B samples. This was not the case for the thousand of genes identified in the B chromosomes of the fish *Astatotilapia latifasciata* by Valente et al. (2014). To search for protein-coding genes in Bs of the latter species authors included only one individual 0B and other 2B without filtering contigs by coverage in 0B samples what could lead to false positives. In addition, most of the genes they found were very fragmented and transcriptionally inactive while in *E. plorans* almost all B-genes show expression in, at least, one of the RNA samples analyzed.

As summarized in Table 3.3, 17 B-genes of *E. plorans* showed B specific SNPs, becoming a subset of highly-reliable B-genes so these genes are the focus of this work. In addition, the finding of SNPs being exclusive of B-carrying individuals allowed defining genomic variation carrying B-specific signatures (Alt alleles), and all of them were detected at high proportion in the 1B Illumina RNA libraries, specially in ovary. Among the 17 B-genes of *E. plorans* showing B specific variation, the *cip2a* gene was subverted

(i.e. higher expression of Alt alleles than Ref ones in 1B samples) in male and female embryos from P1 and P2, and the *ank2* gene being subverted only in female embryos from P1. In the case of adults, we found 6 subverted genes in legs of males (*ankrd44*, *cap-g*, *cip2a*, *kif20a*, *mycb2* and *top2a*), 4 in females legs (*cap-g*, *kif20a mycb2* and *top2a*), 4 in testis (*cap-g*, *cip2a*, *mycb2* and *ndl*) and 3 in ovary (*cap-g*, *mycb2* and *ndl*). Remarkably, all these genes are involved in cell cycle function thus being excellent candidates to play crucial roles for B chromosome accumulation and maintenance.

Since B chromosomes in *E. plorans* show transmission advantage through female meiotic drive (Zurita et al., 1998), the most important genes for a B chromosome like these are those involved in cell cycle regulation. Among the cell cycle genes found in the B chromosomes of *E. plorans*, *ndl,* a B-gene present in the four populations of *E. plorans* analyzed (i.e. Torrox, Tanzania, Egypt and Armenia), is the one showing the higher number of nucleotide changes. Therefore, it could be quite old in the Bs with 56 substitutions (33 non-synonymous), 6 deletions and 1 insertion compared to the paralog located in A chromosomes. Notably, PROVEAN revealed only two deleterious substitutions at the beginning of the *ndl* CDS that were lower expressed than the neutral ones, specially in ovary (see Table S3.3). The cell cycle genes *ankrd44*, *cap-g*, *cdc16*, *cip2a, ckap2, mycb2, slit* and *top2a* also display aminoacidic substitutions in B-paralogs with respects to the A sequences being tolerated only for *ankrd44, cip2a* and *ckap2* (Table 3.3). Note that, for *cdc16* the deleterious prediction score (-2.505) indicated a low confident result. However, in at least one group of biological samples (embryos, leg, testis and ovary), these deleterious substitution showed less expression than the neutral ones in all genes except for *cap-g* (see Table S3.3).

As mentioned above, the B-drive mechanism in *E. plorans* takes place during female meiosis, by means of preferential migration of the B chromosome towards the secondary oocyte instead of to the first polar body (Zurita et al., 1998), a mechanism cytologically shown by Hewitt (1976) in the grasshopper *Myrmeleotettix maculatus*. Spindle asymmetry is the hallmark of oogenesis and constitutes the basis for non-Mendelian chromosome segregation (Akera et al., 2017) thus B chromosomes could take advantage of this mechanism to drive to the oocyte. We have found one gene directly involved in cell polarity and oocyte asymmetry in the B chromosomes of *E. plorans,* the gene *ndl*. *Ndl* seems to be one of the oldest genes in the B chromosomes of this grasshopper because of the high number of nucleotide changes it displays in these chromosomes (see

above) so it should be one of the crucial genes for Bs success. Although *ndl* shows a IC pattern in B chromosomes, its value of prop_nt4B>0B (0.83) is very closed to the 0.90 threshold we set for coverage pattern definition. This gene is highly transcribed in 1B gonad of *E. plorans*, especially in ovary where B specific copies (Alt alleles) are much more expressed than the Ref ones (Fig. 3.4, S3.4, S3.5 and Table S3.6). *Ndl* gene codes for the Nudel protease domain that is essential for embryonic dorsoventral patterning (LeMosy et al., 1998; Roth, 1998) and has a role prior to fertilization in the successful completion of oogenesis (Mineo et al., 2017). Therefore, this ovary-specific subversion of *ndl* expression in 1B samples could enhance oocyte asymmetry and cell polarity to assure the maternal transmission of B chromosomes in *E. plorans*.

Interestingly, the gene a*pc1,* that is located and transcriptionally active in the B chromosomes of *L. migratoria* (Ruiz-Ruano et al., 2019), codes for one of the subunit of the Anaphase Promoting Complex or Cyclosome (APC/C) as it does the gene *cdc16* that we found in the B chromosomes of *E. plorans* from Torrox (Spain), Tanzania, Egypt and Armenia. The APC/C complex is a cell cycle E3 ubiquitin ligase regulating the metaphase-anaphase transition during mitosis (Jörgensen et al., 2001; Peters 2002 and 2006), playing a role during meiosis (Harper et al., 2002; Barford, 2011) and increasing cell polarity and asymmetry during meiosis upon fertilization (Rappleye et al., 2002). The gene *cdc16* was first reported in *Saccharomyces cervisiae* (Lamb et al., 1994) being homologous of the *apc6* gene in humans (Barford, 2011) and it codes for a core TPR-type subunit of the APC/C complex. *Cdc16* is incomplete in B chromosome of *E. plorans* (IC pattern) and although its number of nucleotide substitution is not so high to be considered as an old and essential gene for B chromosomes it still could play an important role for B chromosomes. Although the IC pattern suggests the presence of incomplete paralogs in the B chromosome, our present approach cannot rule out a possible functional role through the RNA interference pathway (Banaei-Moghaddam et al., 2015). We found a high expression of *cdc16* in gonads of both sexes compared to leg, in 0B and 1B RNA libraries. However, the expression of B-located copies of the *cdc16* gene, i.e. showing Alt alleles, is very low (see Table S3.6). In addition, the *cdc16* gene showed a modest increased in transcription when comparing 1B and 0B testis but a decrease when comparing ovaries of carrier females with those from lacking ones (Fig. 3.4c and Table S3.6). This transcription could even come from B-copies of the gene carrying Ref alleles according to our estimations (see Table S3.5) but being

indistinguishable from that the standard set.

During cell division, APC/C activation is tightly controlled by the spindle assembly checkpoint (SAC) which, in presence of kinetochores being unattached to microtubules, generates the mitotic/meiosis 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. Therefore, the preferential migration of the B chromosome to the secondary oocyte in *E. plorans* actually occurs while APC/C is operating to promote the metaphase-anaphase transition. In grasshoppers, meiotic resumption of primary oocytes occurs upon fertilization during egg laying, so that recently laid eggs are at first meiotic metaphase (Henriques-Gil et al., 1986). In addition, Akera et al. (2019) show that flipping events to face selfish centromeres toward the egg pole in asymmetric female meiosis take time, and rapid progression through meiosis I prevents this drive of selfish elements (I.e B chromosomes). Interestingly, a delayed progression to anaphase I as been explained as a consequence of a reduced activity in the APC/C complex (McCarthy Campbell et al., 2009; McGuinness et al., 2009). It is thus tempting to speculate that the infra expressed *cdc16* gene in 1B ovaries of *E. plorans* could decrease the activity of the APC/C complex, extending the time to the anaphase onset and, therefore, giving B chromosomes enough time to perform meiotic drive.

The coincidence of the two genes, *apc1* and *cdc16,* coding for proteins belonging to the same protein complex (APC/C) in the B chromosomes of both grasshopper species, *L. migratoria* and *E. plorans* respectively, might simply be a chance event. However, the putative utility that the expression of these genes might have to B chromosome drive, in both cases, makes it worth to hypothesized that parasitic B chromosomes showing drive might all harbor protein-coding genes with functions related to chromosome segregation during mitotic and/or meiotic anaphase. Regarding the *apc1* gene located in the B chromosomes of *L. migratoria*, Ruiz-Ruano et al. (2019) found a higher expression of this gene in testis of B-carrying individuals than in those belonging to B-lacking ones. B chromosomes of *L. migratoria* accumulated through a premeiotic drive mechanism, in particular, during the first embrionic mitotic division. This results in a increased number of the B chromosome in germ line cells compared with somatic cells. Interestingly, the APC/C complex, in which the gene *apc1* is involved, promote the asymmetry to configure the anterior-posterior axis in embryos , specifically the APC/C is involved in the defining

of the P axis which will gave rise to the gonads in the future adult (Rappleye et al., 2002). Therefore, the high expression of the *apc1* gene in 1B testis of *L. migratoria* could ultimately increase this embryo asymmetry, after fertilization, causing B chromosomes to pass through the germ line.

Summarizing, through different mechanisms, the B-gene content might give B chromosomes the transmission advantage that they need to invade natural populations before the host genome evolves appropriate resistance. This fightback could come in the form of drive suppressor genes, as proposed by Herrera et al. (1996), or B chromosomes elimination (Cabrero et al., 2017; Chapter 4).

Taken together, our present results show the B chromosome as a subversive genomic element whose destiny may depend critically on the presence of active genes in it. Darlington and Upcott (1941) claimed that “new extra chromosomes appear from time to time in many species, but most of them come to nothing”. We can now complement this sentence by adding that "the remainder can become true B chromosomes if they are genetically well equipped".

References

Ahmad SF, Jehangir M, Cardoso AL, Wolf IR, Margarido VP, Cabral-de-Mello DC, et al. (2020). B chromosomes of multiple species have intense evolutionary dynamics and accumulated genes related to important biological processes. *BMC Genomics*, 21, 656.

Akera T, Chmátal L, Trimm E, Yang K, Aonbangkhen C, Chenoweth DM, et al. (2017). Spindle asymmetry drives non–Mendelian chromosome segregation. *Science*, 358(6363), 668–672.

Akera T, Trimm E, Lampson MA. (2019). Molecular Strategies of Meiotic Cheating by Selfish Centromeres. *Cell*, 178(5), 1132–1144.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment searchtool. *Journal of Molecular Biology*, 215, 403–410.

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(1), 14.

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.

Banaei-Moghaddam AM, Martis MM, Macas J, Gundlach H, Himmelbach A, Altschmied L, et al. (2015). Genes on B chromosomes: old questions revisited with new tools. *BBA Gene Regulatory Mechanisms*, 1849, 64–70.

Barford D. (2011). Structural insights into anaphase-promoting complex function and mechanism. *Philosophical Transactions* of the *Royal Society B*: *Biological Sciences*, 366, 3605–3624.

Bidau C, Martí DA (2010). 110 years of Orthopteran cytogenetics, the chromosomal evolutionary viewpoint, and Michael White's Signal contributions to the field. *Journal of Orthoptera Research*, 19(2), 165–182.

Burt A, Trivers R. (2016). Genes in Conflict: The Biology of Selfish Genetic Elements. Hardvard University Press.

Cabrero J, Manrique-Poyato MI, Camacho JPM. (2006). Detection of B chromosomes in interphase hemolymph nuclei from living specimens of the grasshopper *Eyprepocnemis plorans*. *Cytogenetic and Genome Research*, 114(1), 66–69.

Cabrero J, Martín-Peciña M, Ruiz-Ruano FJ, Gómez R, Camacho JPM. (2017). Post–meiotic B chromosome expulsion, during spermiogenesis, in two grasshopper species. *Chromosoma*, 126(5), 633–644.

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, Cabrero J, López-León MD, Bakkali M and Perfectti F. (2003). The B chromosomes of the grasshopper *Eyprepocnemis plorans* and the intragenomic conflict. *Genetica*, 117(1), 77–84.

Camacho JPM. (2005). B chromosomes. In Gregory TR, editor. The Evolution of the Genome. Elsevier. 223–286.

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.

Choi Y, Chan AP. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, 31(16), 2745–2747.

Dalla Benetta E, Antoshechkin I, Yang T, Nguyen HQM, Ferree PM, Akbari OS. (2019). Genome elimination mediated by gene expression from a selfish chromosome. *BioRxiv*, 793273.

Darlington CD, Upcott MB. (1941). The activity of inert chromosomes in *Zea mays*. *Journal of Genetics*, 41, 275–296.

Dirsh VM. (1958). Revision of the genus *Eyprepocnemis* Fieber, 1853 (Orthoptera: Acridoidea). *Proceedings of the Royal Entomological Society of London. Series B*, 27, 33–45.

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

Harper JW, Burton JL, Solomon MJ. (2002). The anaphase-promoting complex: it's not just for mitosis any more. *Genes & Development*, 16, 2179–2206.

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.

Herrera JA, López-León MD, Cabrero J, Shaw MW and Camacho J. (1996). Evidence for B chromosome drive suppression in the grasshopper *Eyprepocnemis plorans*. *Heredity*, 76(6), 633.

Hewitt GM. (1976). Meiotic drive for B-chromosomes in the primary oocytes of *Myrmeleotettix maculatus* (Orthoptera: Acrididae). *Chromosoma*, 56, 381–391.

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.

Jones RN, Rees H. (1982). B chromosomes. Academic Press.

Jörgensen PM, Gräslund S, Betz R, Ståhl 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 USA*, 113, 966–971

Kinsella CM, Ruiz-Ruano FJ, Dion-Côté AM, Charles AJ, Gossmann TI, Cabrero J, et al. (2019). Programmed DNA elimination of germline development genes in songbirds. *Nature Communications*, 10(1), 5468.

Lahn BT, Page DC. (1999). Four evolutionary strata on the human X chromosome. *Science*, 286, 964–967.

Lamb JR, Michaud WA, Sikorski RS, Hieter PA. (1994). Cdc16p, Cdc23p and Cdc27p form a complex essential for mitosis. *The EMBO Journal*, 13(18), 4321–4328.

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.

LeMosy EK, Kemler D, Hashimoto C. (1998). Role of Nudel protease activation in triggering dorsoventral polarization of the *Drosophila* embryo. *Development*, 125(20), 4045–53.

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.

Ma W, Gabriel TS, Martis MM, Gursinsky T, Schubert V, Vrána J, et al. (2017). Rye B chromosomes encode a functional Argonaute-like protein with in vitro slicer activities similar to its a chromosome paralog. *New Phytologist*, 213, 916–928.

Martis MM, Klemme S, Banaei-Moghaddam AM, Blattner FR, Macas J, Schmutzer T, 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 USA*, 109, 13343–13346.

McCarthy Campbell EK, Werts AD, Goldstein B. (2009). A cell cycle timer for asymmetric spindle positioning. *PLoS Biology*, 7(4), e1000088.

McGuinness BE, Anger M, Kouznetsova A, Gil-Bernabé AM, Helmhart W, Kudo NR, et al. (2009). Regulation of APC/C activity in oocytes by a Bub1–dependent spindle assembly checkpoint. *Current Biology*, 19(5), 369–80.

Mineo A, Furriols M, Casanova J. (2017). Transfer of dorsoventral and terminal information from the ovary to the embryo by a common group of eggshell proteins in *Drosophila*. *Genetics*, 205(4), 1529–1536.

Navarro-Domínguez B, Ruiz-Ruano FJ, Cabrero J, Corral JM, López-León MD, Sharbel TF, et al (2017). Protein-coding genes in B chromosomes of the grasshopper *Eyprepocnemis plorans*. *Scientific Reports*, 7, 45200.

Navarro-Domínguez B, Martín-Peciña M, Ruiz-Ruano FJ, Cabrero J, Corral JM, López-León MD, et al. (2019). Gene expression changes elicited by a parasitic B chromosome in the grasshopper *Eyprepocnemis plorans* are consistent with its phenotypic effects. *Chromosoma*, 128, 53–67

Nur U, Werren JH, Eickbush DG, Burke WD, Eickbush TH. (1988). A "selfish" B chromosome that enhances its transmission by eliminating the paternal genome. *Science*, 240, 512–514.

Östergren G. (1945). Parasitic nature of extra fragment chromosomes. Bot Not 2, 157–163.

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.

Rappleye CA, Tagawa A, Lyczak R, Bowerman B, Aroian RV. (2002). The anaphase-promoting complex and separin are required for embryonic anterior-posterior axis formation. *Developmental Cell*, 2(2), 195–206.

Randolph LF. (1941) Genetic characteristics of the B chromosomes in maize. *Genetics*, 26, 608–631.

Rhoades MM, McClintock B. (1935). The cytogenetics of maize. *Botanical Review*, 1, 292–325.

Roth S. (1998). *Drosophila* development: the secrets of delayed induction. *Current Biology*, 8(25), R906–10.

Ruban A, Schmutzer T, Wu DD, Fuchs J, Boudichevskaia A, Rubtsova M, et al. (2020). Supernumerary B chromosomes of *Aegilops speltoides* undergo precise elimination in roots early in embryo development. *Nature Communications*, 11(1), 2764.

Ruiz-Estévez M, López-León MD, Cabrero J, Camacho JPM. (2012). B-chromosome ribosomal DNA is functional in the grasshopper *Eyprepocnemis plorans*. *PLoS ONE*, 7, e36600.

Ruiz-Ruano FJ, Navarro-Domínguez B, López-León MD, Cabrero J, Camacho JPM. (2019). Evolutionary success of a parasitic B chromosome rests on gene content. *BioRxiv*, 683417.

Trifonov VA, Dementyeva PV, Larkin DM, O'Brien PC, Perelman PL, Yang F, et al. (2013). Transcription of a protein-coding gene on B chromosomes of the Siberian roe deer (*Capreolus pygargus*). *BMC Biology*, 11, 1.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Research*, 40, e115.

Valente GT, Conte MA, Fantinatti BE, Cabral-de-Mello DC, Carvalho RF, Vicari MR, et al. (2014). Origin and evolution of B chromosomes in the cichlid fish *Astatotilapia latifasciata* based on integrated genomic analyses. *Molecular Biology and Evolution*, 31, 2061–2072.

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.

Zurita S, Cabrero J, López-León MD, Camacho JPM. (1998). Polymorphism regeneration for a neutralized selfish B chromosome. *Evolution*, 52(1), 274–277.

Supplementary Materials and Methods for Chapter 3

Materials, nucleic acid isolation and sequencing. We collected adults individuals of *E. plorans* from the Torrox population (Málaga province, 36.737558N, -3.953546W) where the prevalent B variant is the most parasitic one, B24 (Camacho et al., 2003). In 2014 we collected about 30 males and 30 females individuals, half of them were processed and stored whereas the other half was analyzed *in vivo* to ascertain the number of B chromosomes they carried. We processed individuals as follows: from males we extracted and fixed some testis follicles in 3:1 ethanol-acetic acid for cytological analysis of primary spermatocytes at diplotene or metaphase I to score the number of B chromosome of each individual, and the other testis and body remains were separately frozen in liquid nitrogen and stored at -80 ºC for subsequent experiments. 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. In the case of females, we extracted the whole ovary from the body and we fixed both parts in liquid nitrogen and then they were stored at -80 ºC. The presence of B chromosomes in females was determined by PCR of the SCAR marker, Muñoz-Pajares et al. (2011). Among the processed and stored individuals we found 3 males 0B, 3 males 4B (2Bs+1isoB) and 3 females 0B. We extracted the genomic DNA from hind legs of these individuals using the GenElute Mammalian Genomic DNA Miniprep (Sigma-Aldrich) and sequenced 6 libraries in an Illumina HiSeq X platform by *Beijing. Novogene Co. Ltd* (*Headquarters*) producing about 6.5 Gb of 2x150nt reads per library. Males gDNA libraries were used for CDS abundance estimation and qPCR validations whereas libraries from both sexes were used for SNP calling.

For *in vivo* study of adults, we extracted several testis follicles from males through a small cut in the abdomen and cytologically analyze primary spermatocytes at diplotene or metaphase I to score the number of B chromosome of each individual. For the same purpose, in the case of females we performed C-banding on hemolymph nuclei as described in (Cabrero et al., 2006). Then, we set up several controlled crosses between 0B and 1B individuals of *E. plorans* to get 0B and 1B embryos. Sibling embryos from this controlled crosses were processed to obtain material for DNA and RNA extraction as well as for cytogenetic analysis to determine their sex and B chromosome content. Doing so, we were able to get three biological replicates of sibling embryos from each of the

following groups: females 0B, males 0B, females 1B and males 1B. As we could not able to get enough RNA for sequencing from one single embryo, we decided to perform one RNA extraction joining material of two embryos from the same group. Therefore, each of the three biological replicates is actually a mixture of two embryos thus reducing inter-individual variability.

We got this set of embryos from two different controlled crosses, the female parental was the B-carrying progenitor in one of them (F1BxM0B, P1 from now on) and in the other it was the male parental who harbored the B chromosome (F0BxM1B, P2), counting with 24 embryos in total. We extracted RNA from embryos using the RNeasy Lipid Tissue kit (Qiagen), following manufacturer's recommendations. Then we sequenced the 24 RNA libraries in *Beijing Genomics Institute, BGI* (China) but in different moments. Sequencing of the 12 embryo libraries coming from the P1 cross was performed in two lanes of Illumina HiSeq 2000 which yielded about 3 Gb of 2x100nt reads per library (we discarded one library of a 0B male for subsequent analysis due to sequencing failure). Remaining libraries from the P2 cross were sequenced in an Illumina HiSeq 4000 platform producing 6 Gb of 2x150 nt read per library approximately. These RNA libraries have been used in the present study for reference transcriptome assembly and expression analysis.

In 2016 we collected about 17 adults of *E. plorans* from Torrox (36.737558N, -3.953546W) and they were all processed *in vivo* to set up controlled crosses in order to get sibling adults for RNA-seq experiments. Every egg pod from controlled crosses in the lab was maintained in an incubator chamber at 28 ºC until hatching, moment in which they were transferred to wooden boxes (one for each cross) until their development to adults. After seven days since the last shedding to adults, each individual was studied cytogenetically by hemolymph C-banding (as described in Cabrero et al., 2006) and at the same time we extracted and fixed the gonads, hind legs and body in liquid nitrogen and preserved them at -80 ºC. Legs and gonads for DNA and RNA extraction were stores in separated packaged to avoid cross contamination. When characterized, we selected individuals coming from the same cross counting with at least 3 males 0B, 3 females 0B, 3 males 1B and 3 males 1B for each of which we performed one extraction of gDNA from the jumping leg of each individual (as described above) and two extraction of RNA per individual, one from the leg (Total RNA Spin Plus, Durviz) and one from the gonads (RNeasy Lipid Tissue kit, Qiagen) followed by sequencing.

The only cross producing enough adults of each biological group was a cross

between a 0B female and a 1B male. A total of 12 gDNA libraries were sequenced by *Macrogen, Inc. (Seul, South Corea)* in an Illumina HiSeq X Ten platform yielding about 10 Gb of 2x150nt read per library. This data was used for CDS abundance estimation in B-lacking and B-carrying individuals and for SNP calling. Regarding adults RNA, we sequenced a total of 24 libraries from the different samples of *E. plorans* individuals we processed: 2 sexes x 2 tissues (leg/gonad) x 2 categories (0B/1B) x 3 biological replicates. These libraries were also sequenced in *BGI* but in an Illumina HiSeq 4000 platform yielding about 6 Gb of 150 bp paired-end reads per library. These libraries were used for reference trancriptome assembly and for tFC calculation between samples. In sum, we used for reference transcriptome assembly 47 RNA libraries from *E. plorans* adults and embryos of both sexes and having/lacking B chromosomes.

In addition, we used a 0B and a 4B libraries of *E. plorans* males (Torrox, 2012) previously deposited in SRA under accession numbers SRR2970625 and SRR2970627 respectively to increase genomic coverage as much as we could for SNP calling. We also included two gDNA libraries of *E. plorans* males, one 0B and the other one carrying B chromosomes, from each of the following populations: Tanzania, Egypt and Armenia. We determined the number of B chromosomes of males from Tanzania after cytogenetic analysis of their fixed testis and the B-carrying male harbored 1B chromosome. However, the quality of samples from Egypt and Armenia was not good enough to perform cytogenetic analysis so we selected as +B males those showing the brightest band after PCR amplification of the SCAR marker and we considered 2B chromosomes for subsequent gFC calculation. Genomic DNA extraction of these males was performed as explained above for Torrox individuals and it was sequence using an Illumina HiSeq X platform at Beijing Novogene Bioinformatics Technology Co., Ltd (Headquarters) in the case of males from Egypt and Armenia yielding about 7 Gb and 8 Gb of 150 bp paired-end reads respectively. In contrast, gDNA of males from Tanzania was sequence in the same company but through a HiSeq 2500 Illumina platform, yielding about 8 Gb of 125 bp paired-end reads (~0.8x).

***De novo* transcriptome assembly, annotation, mapping and selection.** Custom scripts written to perform this analysis are deposited in a public repository with instructions to install and launch them (https://github.com/fjruizruano/whatGene).

We used a bioinformatic procedure similar to the protocol described by Navarro-Domínguez et al. (2017) to search for genes showing differential abundance between B-

carrying and B-lacking libraries (Fig. S3.3). We generated a *de novo* transcriptome which was used as a reference to map the genomic reads. In order to obtain the most complete reference transcriptome of *E. plorans* we assembly several RNA libraries from individuals at different developmental stages (embryos and adults), belonging to both sexes and having/lacking B chromosomes (47 libraries as indicated above). To avoid assembly artifacts and a redundant *de novo* transcriptome we assembly separately each kind of sample (embryos from P1, embryos from P2, adult legs and adults gonads) using Trinity (Haas et al., 2013) with *in silico* normalized libraries (50x maximum coverage) and default options. Then we extracted CDSs being longer than 100 aminoacids, using the Transdecoder software (Haas et al., 2013) and we reduced redundancy with CDHit-EST (Li et al., 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). We performed this assembly procedure separately considering 0B and 1B samples, getting two reference transcriptomes (after a final run of CDHit), 0B and 1B, so we were able to distinguish the +B specific contigs (from B chromosomes or expressed from the A genome in response to the B presence).

We annotated the clustered CDSs using RepeatMasker (Smit et al., 2013) with the RepBase (Bao et al., 2015) and a custom database of repetitive elements previously generated by us (Ruiz-Ruano et al., 2018 and Chapter 2 of this thesis) to annotate contigs for repetitive elements, and the Trinotate pipeline (https://trinotate.github.io) and the SWISS-PROT database (Boechmann et al., 2003) to annotate contigs for protein-coding genes.

We then mapped the gDNA and RNA reads against the reference transcriptome using SSAHA2 (Ning et al., 2001), we performed mappings to the 0B and 1B reference transcriptomes separately. 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, in contrast, present in the genomic reads. We accepted mappings with at least 40 nt with a 80% of minimum identity. We then scored the number of reads mapped per site along the CDS, using the Pysamstats utility (https://github.com/alimanfoo/pysamstats) integrated in the bam_coverage_join.py custom script. In addition, the coverage_graphics.py custom script expressed CDS coverage in each library as the number of copies per haploid genome using the following equation: copy number= (coverage x genome size)/library

size. Additionally, we estimated the expression level of the annotated CDSs in the 12 RNA-seq libraries as RPKM using the following equation: RPKM= (10^9*mapped reads)/ (total mapped reads*contig length).

To normalize genomic coverage, we calculated the genome size of 0B, 1B and 4B individuals from Torrox (Spain) as follows. According to Ruiz-Ruano et al. (2011), the *E. plorans* genome is 1.78 times larger than that of *Locusta migratoria*. On this basis these authors got estimates of DNA content of A and B chromosomes in *E. plorans*, using previous estimates of C value in *L. migratoria* (5.89 pg = 5.76 Gb). However, Wang et al. (2014) later showed after genome sequencing that this value was actually larger (6.3 Gb). We recalculated C, X and B sizes of *E. plorans* taking into account the 6.3 Gb genome of *L. migratoria* indicated by Wang et al. (2014) and that the *E. plorans* genome is 1.78 times bigger than the one of *Locusta*, the X chromosome is 12.3% the haploid genome and the B chromosome is a 49.61% the size of the X. Starting from the haploid genome value for *E. plorans* indicated by Ruiz-Ruano et al. (10.259 Gb) we obtain the new C-value (11.21 Gb), X (1.38 Gb) and B (0.68 Gb) chromosome sizes. Therefore, the haploid genome size of the different samples used here is:

$$G_{male0B} = (2C - X)/2 = (2*11.21 - 1.38)/2 = 10.52\ Gb$$

$$G_{female0B} = (2C)/2 = (2*11.21\ )/2 = 11.21\ Gb$$

$$G_{male1B} = (2C - X + B)/2 = (2*11.21 - 1.38 + 0.68)/2 = 10.86\ Gb$$

$$G_{female1B} = (2C + B)/2 = (2*11.21 + 0.68)/2 = 11.55\ Gb$$

$$G_{male4B} = (2C - X + 4B)/2 = (2*11.21 - 1.38 + 0.68)/2 = 11.88\ Gb$$

Using the copy number of contigs in each individual, we analyzed the effect of including biological replicates, increasing the number of B chromosomes in samples and considering contigs showing between 0.5 and 4 copies in 0B samples. For that purpose we counted the number of contig and their copy number above the gFC threshold, log2(copies +B/copies 0B), depending on the number of B chromosomes of the samples (gFC>0.585 for 1B samples and 1.585 for 4B ones, considering assuming that a single-copy gene would show two copies in a 0B genome, three in a genome carrying 1B and six copies when harboring 4Bs) and annotated them using Trinotate and SWISS-PROT database. We also estimate the number of contigs and copies from the total of contigs analyzed that would be locate in a B chromosome considering that they distributed uniformly around the genome and the size B chromosomes calculated above, note that this approach is quite conservative as B chromosome are expected to be depleted in

genes compared to the rest of chromosomes (Dataset 3.1). Therefore, if the number of contigs identified in B chromosomes was higher that the expected we considered that excess as false positives. We performed this analyzed using the mean coverage of contigs for all +B individuals and using individuals separately as biological replicates, including only 1B individuals or 4B ones and discarding or retaining contigs with >0.5 and <4 copies in 0B samples. The most reliable approach avoiding false positives was the one including biological replicates, higher number of B chromosomes and selecting contigs with number of copies on 0B samples between 0.5 and 4.

As a result of the above analysis, we separately estimated mean CDS coverage in both 0B and 1B/4B genomic libraries and then we filtered out CDSs according to coverage, excluding highly represented CDSs, because they were candidates to be repetitive 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 might have come from assembly errors. We then calculated the fold change in CDS coverage (gFC), due to B chromosome presence, as log2 of 4B/0B and 1B/0B.

To select contigs being candidate to reside in the B chromosome, we used several stepwise criteria. The first and least stringent criterion (applied to 4B and 0B averages) implied selecting all those contigs showing gFC$\geq$ 0.585 in 1B libraries, in 4B (2B+1isoB) libraries the gFC threshold was 1 for short arm genes or 1.585 for genes located in long arms of B chromosomes. For these calculations we assumed that a single-copy gene would show two copies in a 0B genome, three in a genome carrying 1B, four in a 2B+1isoB genome if the gene is located in the short arms of Bs and six copies in a 4B genome, also assuming the presence of a single copy in the B chromosome. This would yield, as example, an expected genomic 1B/0B ratio of 3/2= 1.5 and thus gFC= $\log_2(1.5)$= 0.585. As *E. plorans* 4B samples carried actually 2 B chromosome plus 1 isoB chromosome we were able to apply two different gFCs: 1.585 (considering 4 copies from B chromosomes) or 1 (counting with 2 copies from Bs, genes located in B short arm). Therefore in this filtering step we selected contigs with gFC>1.585 and gFC>1 in 4B libraries, when applying the gFC>1 we also required a gFC>0.585 in 1B libraries avoiding false positives. In subsequent filtering step we use exclusively 4B libraries as in 1B libraries coverage in Bs is very low thus limiting the power to identify B located genes. Then, we remove duplicates between contigs from 1B and 0B transcriptomes using

reciprocal BLASTN homology detection and excluding contigs matching repetitive elements of *E. plorans* using RepeatMasker. To decrease the probability of selecting contigs showing variation in coverage between different B-carrying individuals we applied a second filter selecting contigs showing gFC>1 in each of the three 4B libraries separately. Then we annotated all these contigs by Blast2GO (Conesa et al., 2005). Complete workflow of CDSs selection can be found in Datasets 3.2 whereas graphical representation of filtering steps is shown in Fig. S3.2.

Sequence analysis for selected contigs. The CDSs meeting the former criteria were submitted to additional analyses. We first checked if the CDS was complete in the contig. We performed functional gene annotation using Eukaryotic Orthologous Groups of proteins (KOG), by searching for the predicted protein sequence in the WebMGA server (http://weizhong-lab.ucsd.edu/webMGA/server/kog/).

Using the longest version obtained for all selected CDSs, we performed additional mapping of all genomic and transcriptomic libraries with SSAHA2, to get estimates of coverage per site along the transcript (CDS+UTRs when possible). We then averaged these estimates per contig and using the script coverage_graphics.py script, we graphically represented the coverage as mean copy number (gDNA) or RPM (RNA) ±1SD. Finally, we calculated the proportion of nucleotides in average showing higher abundance in +B libraries than in 0B ones (prop_nt4B>0B), we considered as uniform coverage genes (UC, probably complete in B chromosomes) those with a prop_nt4B>0B equal or higher than 0.90 and genes with irregular coverage (IC, incomplete in B chromosomes) those genes with a prop_nt4B>0B value below 0.9. We analyze this B-genes completeness for the full transcript an for the CDS.

We also used the software SSAHA2 to map the gDNA libraries of *E. plorans* from Tanzania, Egypt and Armenia against the B-genes identified in Torrox (Spain) in order to ascertain which B-genes were found in B-carrying individuals from these populations. As we could not determine the specific B variant harbored in these populations we used the chromosome size estimated in Torrox for a B24 carrier male as indicated by Ruiz-Ruano et al. (2011) applying corrections of Wang et al. (2014). Therefore we used the following genome sizes estimations:

$$G_{male0B\ (Tanzania,\ Egypt\ and\ Armenia)} = (2C - X)/2 = (2*11.21 - 1.38)/2 = 10.52\ Gb$$

$$G_{male1B\ (Tanzania)} = (2C - X + B)/2 = (2*11.21 - 1.38 + 0.68)/2 = 10.86\ Gb$$

$$G_{male2B(Egypt\ and\ Armenia)} = (2C - X + 2B)/2 = (2*11.21 - 1.38 + 2*0.68)/2 = 11.2\ Gb$$

Transcription analysis of B chromosome genes. As a first estimation of possible up-regulation of some genes due to the presence of some active copies in the B chromosome, we calculated a transcriptomic fold change (tFC) due to B chromosome presence as log2 of the quotient between 1B and 0B coverage in the RNA-seq libraries. For sequence-dependent inferences, however, we searched for B-specific sequence changes in the gDNA and RNA libraries. For this purpose, we performed SSAHA2 mappings of gDNA and RNA reads against the sequences of the selected genes, in order to perform a SNP analysis. We first merged the BAM files using SAMtools (Li et al., 2009) for the ten conditions of *E. plorans*, i.e., gDNA 0B, +B (1B plus 4B), and RNA from gonads, legs, embryos from P1 and P2, belonging to both sexes and having/lacking B chromosomes. We used the custom script SNP_calling_bchr.py to search for SNPs variants found at least 2 times in gDNA +B and with zero counts in the gDNA and RNA 0B libraries. At each position, we considered as reference (Ref) the nucleotide being present in the 0B, and as alternative (Alt) that being present only in +B gDNA or RNA. To increase the reliability of the nucleotidic variations observed, we applied an extra filter (before normalizing data by library and genome size) selecting those variants showing a quotient of Alt/Ref_0B copies equal or higher than the expected depending on the B chromosome number of +B libraries, i.e. >1.1 (the average B chromosome number of 1B and 4B libraries of *E. plorans* mapped was 2.2, for genes located in short arms the expected quotient was 0.8). This procedure allow us to select variation linked directly to B chromosomes (see selected SNPs in Table S3.4). We used this information to generate the A chromosome sequences with the Ref allele and the B chromosome sequences with the Alt allele, using the custom script sequence_ref_alt.py. We then used Geneious to translate them and manually check if they were synonym or non-synonym changes. We also used PROVEAN (Choi and Chan, 2015) to predict the functional effects of aminoacid substitutions in proteins translated from Alt variants and we calculated the ratio between expression of neutral versus deleterious SNPs in the different RNA samples we studied (see Table S3.3).

Additionally, we applied an approximation to know the number of Alt and Ref copies located in B chromosomes. For this, we consider that the excess of Ref copies in genomes +B with respect to Ref copies in genomes 0B (represented by Ref+B/Ref0B) is due to the presence in B chromosomes of genes copies equal to those located in the A chromosomes (in fact, these copies may be the most interesting ones to conserve their

function, although we will not be able to distinguish in RNA whether these Ref copies from Bs are expressed). To calculate Ref copy number in Bs we applied the formula:

Ref_{Bs}=(((Ref$_{+Bgenome}$/ Ref$_{0Bgenome}$)*copy number in 0B genome)-copy number in 0B)/number of B chromosomes in a haploid genome.

The number of Alt copies located in B chromosomes would be represented by:

Alt_{Bs}=((Alt/Ref$_{0Bgenome}$)*copy number in 0B genome)/number of B chromosomes in a haploid genome

As 'copy number in 0B' we have considered both normalized count of mappings to CDS and counts after the mappings to the complete transcripts (using selected SNPs). Copy number results for *cap-g* gene, 4.13 Alt and 5.34 Ref copies, were highly similar to those found in Navarro-Domínguez et al. (2017) where author estimated about 12 copies of *cap-g* in B24 chromosomes.

Comparing normalized reads counts for Alt and Ref gene variants in RNA samples we explored the differential expression of B-genes in biologically different samples (see Datasets S4 and examples in Fig. 3.4c) and comparing the expression of Alt with respect to Ref copies in 1B RNA samples we identify subverted genes expressing the Alt variant in a higher amount than the Ref one. In Figure S3.5 we show Alt/Ref ratios in 1B RNA libraries as percentage of the total of B genes and the ratio Alt/Ref for each gene. Moreover, by means of read counts in gDNA and RNA of B-carrying individuals it is possible to estimate Alt variant frequency, i.e. FC=log2 (Alt/Ref), in gDNA (gFC) and RNA (tFC) in the different samples testing expression of Alt copies compared to their abundance (see Fig. S3.4).

qPCR validation. The genomic overabundance associated with B chromosome presence was tested in *E. plorans* for of 9 genes on B chromosomes (5 genes with a gFC>1 in the three 4B individuals, 2 with gFC>1 in the three 4B individuals in CDS abundant estimations but not after transcript analysis and 2 genes with a gFC<1 in some of the 4B libraries). The presence of these genes was tested using gDNA from 6 males 0B and 13 males +B (6 1B, 3 2B, 2 3B and 2 4B) of *E. plorans*.

Primer pairs anchoring in the same exon were designed with Primer3 (Untergasser et al., 2005). We search for exon limits using the Exonerate software (Slater et al., 2005) and aligning the transcript sequence against the *L. migratoria* genome assembled by Wang et al. (2014). We preferably selected regions with low sequence variability and high read mapping coverage. Primer sequences and amplicon length are shown in Table S3.7.

Quantitative PCR was carried out as described in Navarro-Domínguez et al. (2017) with modifications indicated in the "Materials and methods" section of this thesis. qPCRs were performed on a Chromo 4 Real Time PCR thermocycler (Biorad). Each reaction mixture contained 25 ng of gDNA in 5 µl, 5 µl of SensiMix SYBR Kit (Bioline) and 2.5 µl of each 2.5 µM primer. Reactions were carried out in duplicate. We estimated the amplification efficiency (E) of each primer pair by means of a standard curve performed on a 10-fold dilution series of *E. plorans* gDNA mixture from different individuals. This gDNA pool was also used as an external calibrator for the qPCR reactions. The relative abundance of each gene in each sample was calculated according to $RQ = E^{CtC-CtS}$, where RQ = Relative quantity, E = Amplification efficiency (fold increase per cycle), CtC = Ct value of the calibrator sample and CtS = Ct value of each sample.

We perform a test of correlation between RQ values and number of B chromosomes of each sample used. In addition, for both species we perform data analysis with bootstrap-coupled estimation, using DABEST (Ho et al., 2019), to test and display standardized effect sizes of B chromosome number on RQ values (Fig. S3.1).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment searchtool. *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, 11

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E*,* et al. (2003). The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31, 365–370.

Cabrero J, Manrique-Poyato MI, Camacho JPM. (2006). Detection of B chromosomes in interphase hemolymph nuclei from living specimens of the grasshopper *Eyprepocnemis plorans*. *Cytogenetic and Genome Research*, 114(1), 66–69.

Camacho JPM, Cabrero J, López-León MD, Bakkali M, Perfectti F. (2003). The B chromosomes of the grasshopper *Eyprepocnemis plorans* and the intragenomic conflict. *Genetica*, 117(1), 77–84.

Choi Y, Chan AP. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, 31(16), 2745–2747.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21, 3674–3676.

Drummond AJ, Ashton B, Cheung M, Heled J, Kearse M. (2009). Geneious v.4.8.5. Biomatters ltd. Aukland, New Zealand.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J*,* 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.

Ho J, Tumkaya T, Aryal S, Choi H, Claridge-Chang A. (2019). Moving beyond P values: data analysis with estimation graphics. *Nature Methods*, 16(7), 565–566.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25, 2078–2079.

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

Muñoz-Pajares AJ, Martínez-Rodríguez L, Teruel M, Cabrero J, Camacho JPM, and Perfectti F. (2011). A single, recent origin of the accessory B chromosome of the grasshopper *Eyprepocnemis plorans*. *Genetics*, 187(3), 853–863.

Navarro-Domínguez B, Ruiz-Ruano FJ, Cabrero J, Corral JM, López-León MD, Sharbel TF, Camacho JPM. (2017). Protein-coding genes in B chromosomes of the grasshopper *Eyprepocnemis plorans*. *Scientific Reports*, 7, 45200.

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.

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.

Ruiz-Ruano FJ, Cabrero J, López-León MD, Sánchez A, Camacho JPM. (2018). Quantitative sequence characterization for repetitive DNA content in the supernumerary chromosome of the migratory locust. *Chromosoma*, 127, 45–57.

Slater GS. Birney E. (2005). Automated generation of heuristics for biological sequence comparison. BMC *Bioinformatics*, 6, 31.

Smit AFA, Hubley R, Green P. (2013). RepeatMasker Open–4.0. http://www.repeatmasker.org.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. (20 12). Primer3—new capabilities and interfaces. *Nucleic Acids Research*, 40, e115.

Wang X, Fang X, Yang P, Jiang X, Jiang F, Zhao D, et al. (2014). The locust genome provides insight into swarm formation and long-distance flight. *Nature Communications*, 5, 2957.

Supplementary Figures for Chapter 3

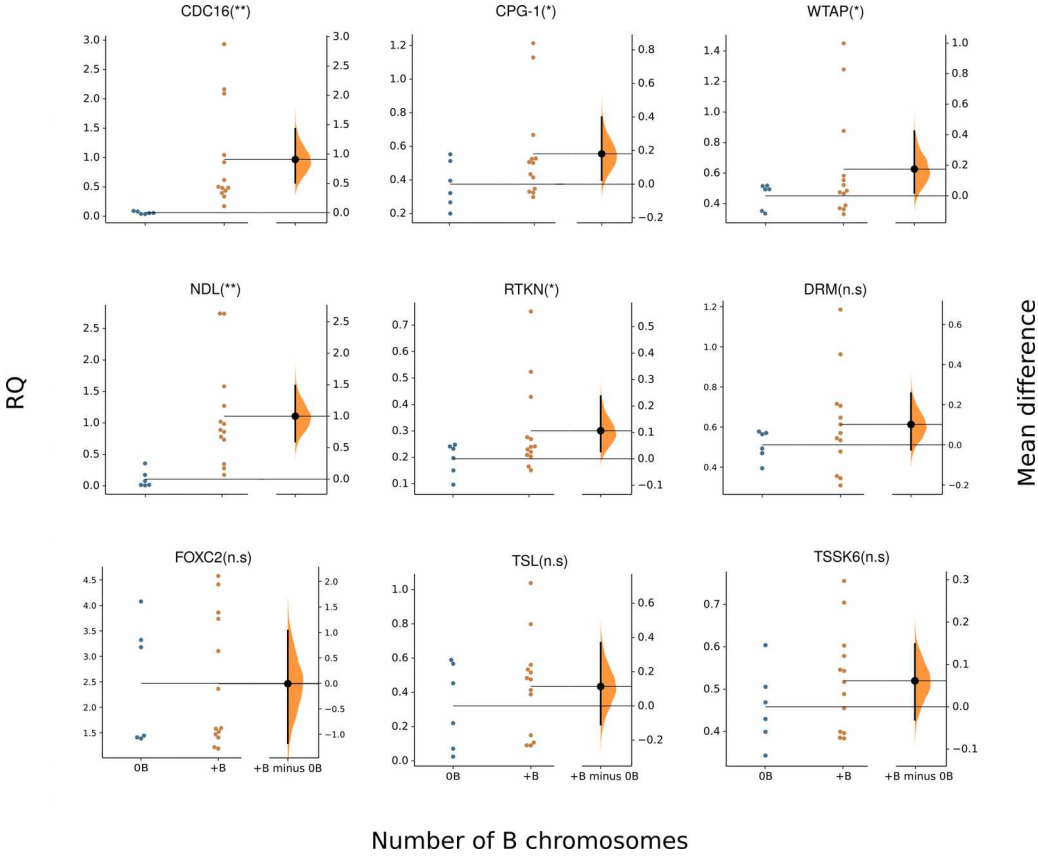


Figure S3.1. Results for all the qPCR analyses performed on genomic DNA for validation of B-located genes in *E. plorans*. **: p<0.05. *: validated by enough mean difference (non-overlapping confident intervals). n.s.: p>0.05.

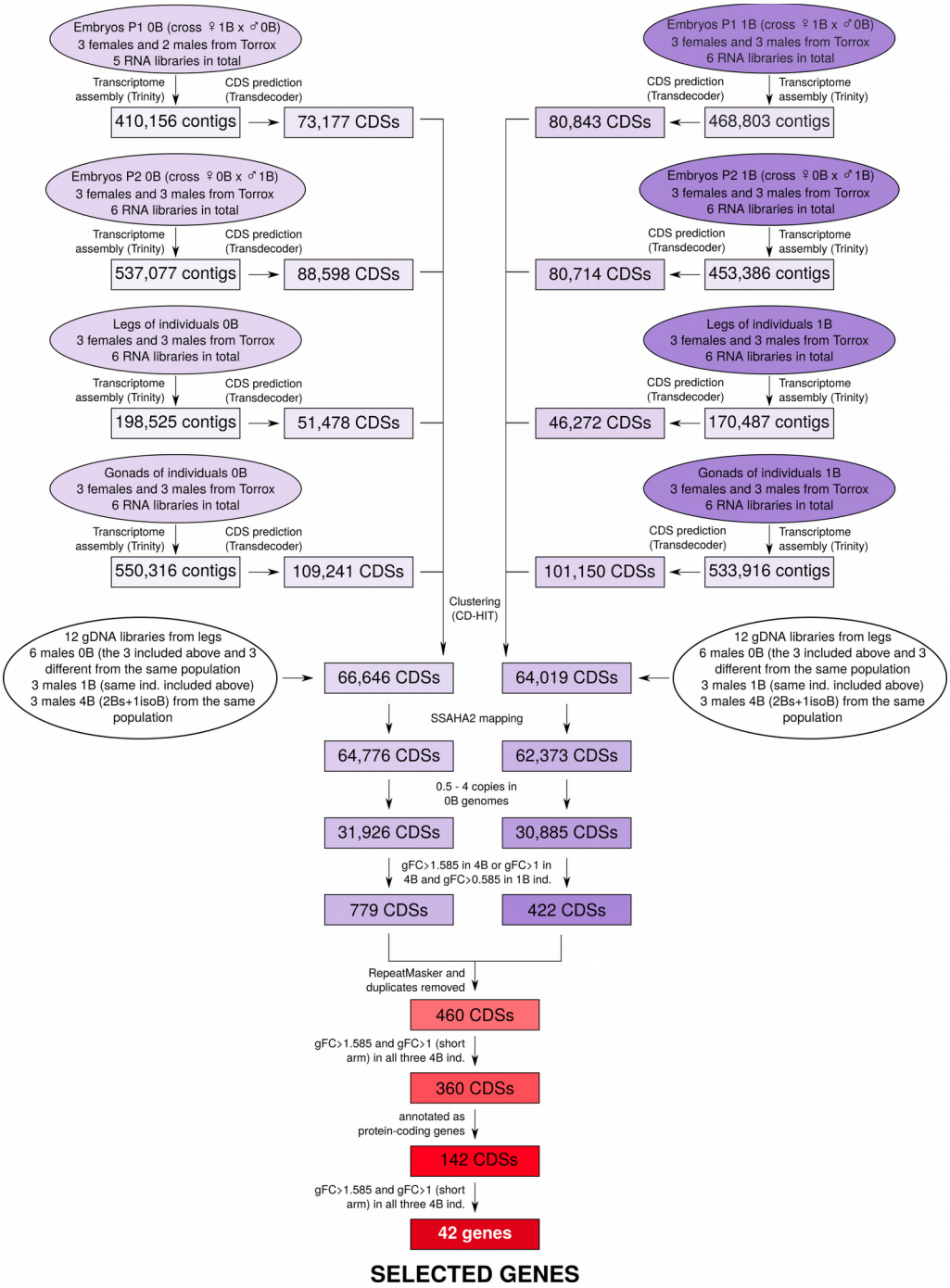


Figure S3.2. Pathway for gene search in the B chromosome of *E. plorans*.

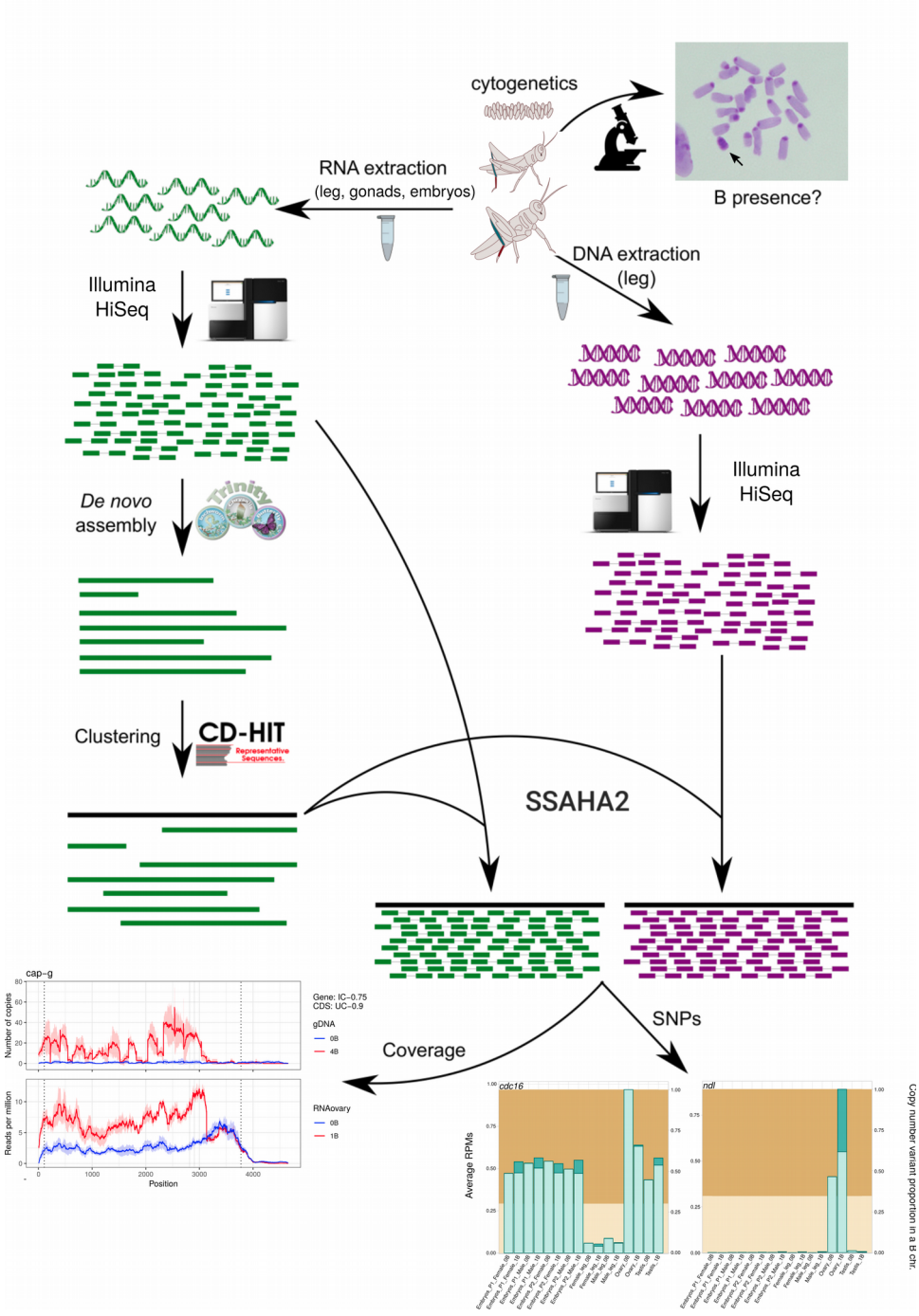


Figure S3.3. Bioinformatic protocol applied for gene selection, and coverage and SNP analyses.

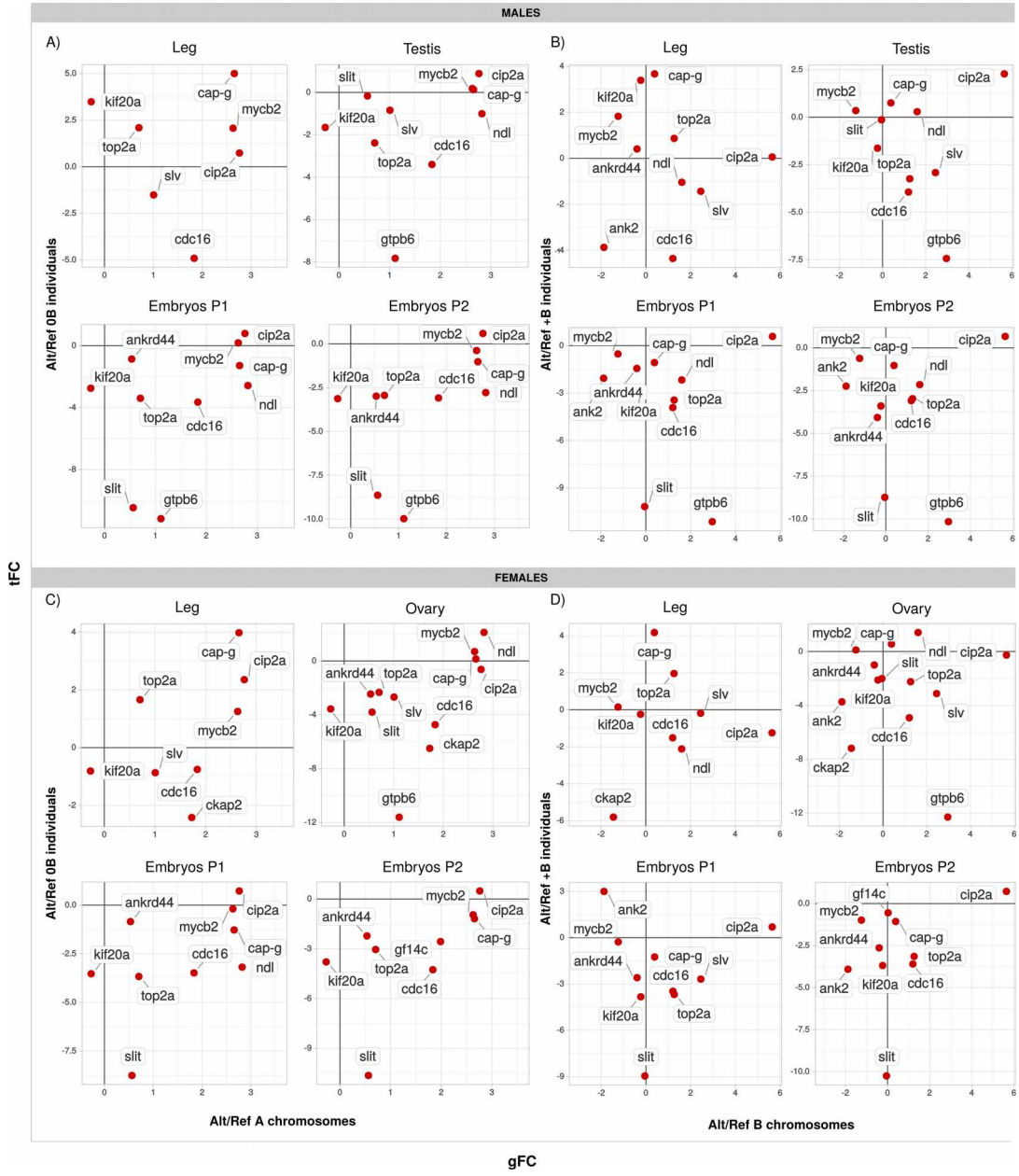


Figure S3.4. Scatterplots of B-genes showing nucleotide variation in B chromosome in respect to A paralogs for *E. plorans*. We show the tFC value of Alt/Ref (in 0B samples) versus the gFC value of Alt variant in respect to Ref (in 0B samples) in legs and gonads of males and females (A, C). Comparison between tFC and gFC of Alt variant in respect to the Ref ones in 1B libraries is displayed in B and D for legs and gonad of males (B) and females (D). Note that gFC Alt values are higher when comparing to Ref copies in A chromosomes than in Bs suggesting the presence of Ref variants in B chromosomes that can be also expressed.

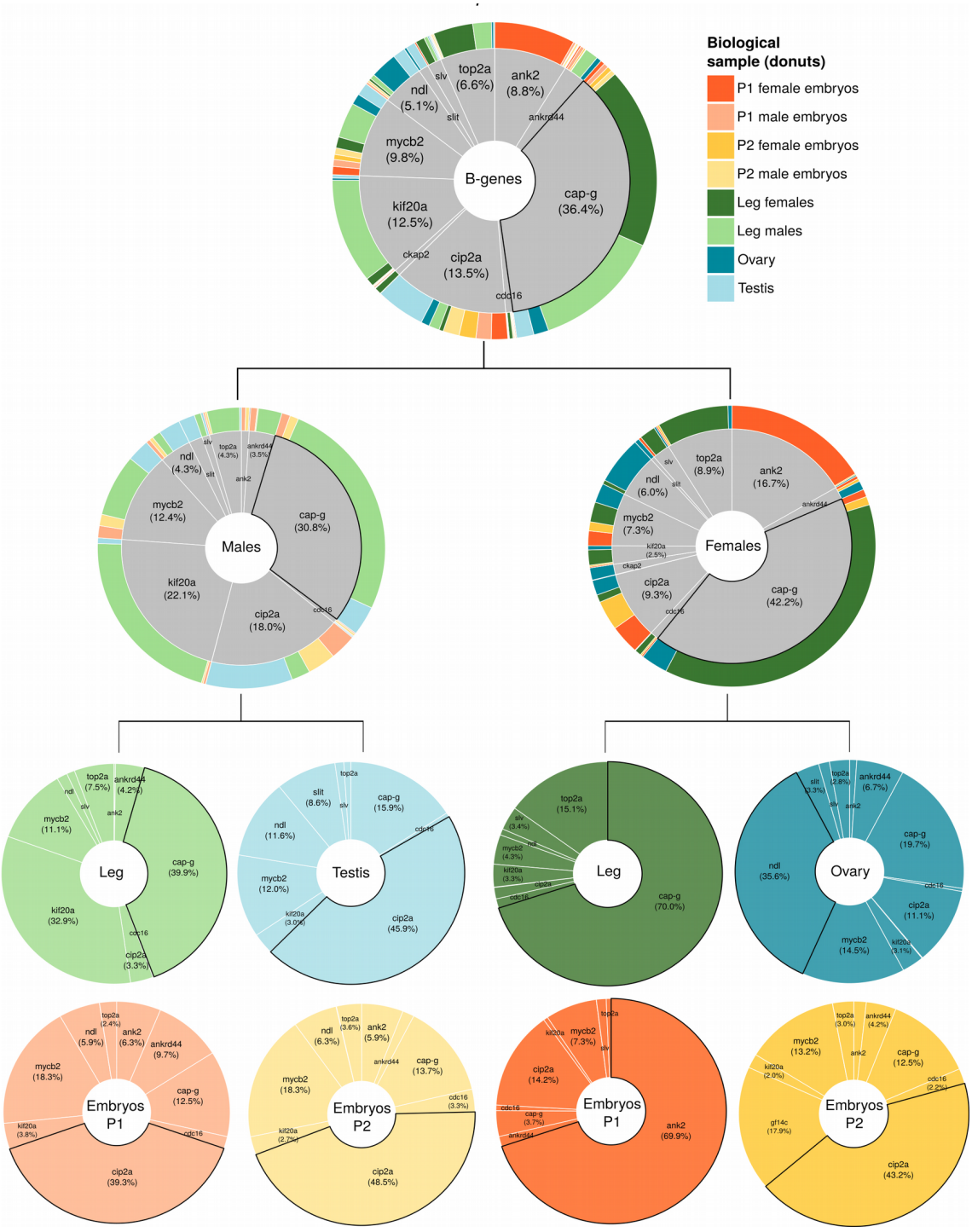


Figure S3.5. Alt/Ref ratios in different 1B RNA libraries shown in percentage (from the sum of all ratios) for B-located gene.

Supplementary Datasets and Tables for Chapter 3

Dataset 3.1. i) Data used for analysis of false positives rates after identification of B-genes in *E. plorans* following different methodological approaches: using 1B or 4B individuals, considering gFC threshold, log2(copies+B/copies0B), in biological replicates or in the average of +B individuals and including or excluding genes with copies in 0B samples <0.5 and >4. ii) Results of false positives rates in the different approaches followed, including biological replicates, increasing the number of B chromosomes (4Bs instead of 1B) and filtering by 0.5-4 copies in 0B samples avoid false positives and identification of repetitive elements.
Dataset available in: https://figshare.com/s/7c9e915a93f14f527d82

Dataset 3.2. Spreadsheet showing the procedure for CDS selection by applying five consecutive filters: i) Among the 64,776 CDSs showing mappings for *E. plorans* gDNA libraries using the transcriptome assembly of the 0B RNA libraries, we selected those with average number of copies in the 0B libraries between 0.5 and 4 (31,926 CDSs), we applied the same filter in the case of mapping to the reference transcriptome 1B selecting 30,885 out of 62,373 CDSs, ii) we then selected those CDSs showing genomic fold change due to B chromosome presence [gFC= log2(4B/0B)] higher than 1.58 considering 4 B chromosomes, contigs with a gFC lower than this value were (or lower than this value in filter iv) selected setting a gFC threshold of 1 for 4B samples (they were actually 2B+1isoB) and of 0.585 for 1B samples in order to identify CDSs potentially located in the short arm of the B chromosomes, iii) we removed duplicated sequences resulting from CDSs abundance estimation using a 0B and 1B reference transcriptome selecting the longer one after blasting sequences from both trancriptomes between them, in addition, we discarded sequences matching repetitive elements using RepeatMasker and a custom repetitive DNA database of grasshoppers as reference, a total of 460 out of 1,201 CDSs were selected for next filtering steps, iv) CDS showing gFC> 1.58 or a gFC>1 (and gFC>0.585 in average 1B libraries) in all three 4B individuals (360 CDSs) were selected v) 142 CDSs out of 361 were annotated for protein-coding genes.
Dataset available in: https://figshare.com/s/2ce4a3c96b3178f45182

Dataset 3.3. Coverage graphical display for the 42 protein-coding genes found in the *E. plorans* B chromosomes from the Spanish individuals collected in Torrox (Málaga). gDNA coverage is shown as number of copies, and RNA coverage is shown as reads per million of mapped reads. Additionally, we add a track showing the position of the B-specific SNPs found in the Torrox libraries. The coverage patterns inferred from the proportion of nucleotides showing higher abundances in 4B libraries than in the 0B ones. Note that a proportion higher than 0.90 indicates uniform coverage (UC) whereas below that value suggests irregular coverage (IC) patterns.
Dataset available in: https://figshare.com/s/69f0defb42e270351270

Dataset 3.4. Bar graphs of Alt and Ref variants counts on RNA of B-genes showing variation in respect to A paralogs of *E. plorans.* RNA counts are expressed in reads per million reads mapped in all RNA samples, leg and gonads of males and females. At the background of each graph we represent the proportion of Alt and Ref variants in a B chromosomes in respect to

the total of copies of that gene in the B.
Dataset available in: https://figshare.com/s/10ca81d587f9d81067c9

Supplementary Tables can be downloaded in:
https://figshare.com/s/30de98e12ee8c0cc4cac
Table S3.1. Contigs selected for thorough sequence analysis.
Table S3.2. List of genes sorted per decreasing gFC value in the high coverage region (hc_gFC), and qPCR results for genes validation.
Table S3.3. Genomic (gFC [log2(4B/0B)]) and transcriptomic (tFC [log2(1B/0B)]) fold change for B chromosome genes. SNPs analysis showing Presence of deleterious aminoacid changes (PDAC) and the ratio in RNA libraries between expression of neutral versus deleterious SNPs. REP1M= RNA-seq on male embryos from P1, REP2M= RNA-seq on male embryos from P2, REP1F= RNA-seq on female embryos from P1, REP2F= RNA-seq on female embryos from P2, RLM= RNA-seq on male leg. RT= RNA-seq on testis. RLF= RNA-seq on female leg. RO= RNA-seq on ovary.
Table S3.4. Counts of reference (Ref) and alternative (Alt= B-specific) variants in all gDNA and RNA libraries analyzed.
Table S3.5. Number of Ref and Alt copies located on a B chromosomes, gFCAlt1B (log2(NcopiesALT/NcopiesRef_Bchromosomes)) and gFCAlt0B (log2(NcopiesALT/NcopiesRef_Achromosomes)).
Table S3.6. Fold change in RNA libraries of *E. plorans* (tFC) from selected SNPs. We show several comparison between different RNA samples and Alt/Ref alleles. tFC > 1 are shaded in green, tFC< -1 in red and tFC>0<1 in blue.
Table S3.7. Primer pairs used for qPCR validation.
Table S3.8. Genomic fold change (gFC [log2(1Bor2B/0B)]) for B chromosome genes in *E. plorans* males from Tanzania, Egypt and Armenia.
Table S3.9. Counts of reference (Ref) and alternative (Alt= B-specific) alleles in all gDNA and RNA libraries analyzed in of *E. plorans* from Tanzania.
Table S3.10. Counts of reference (Ref) and alternative (Alt= B-specific) alleles in all gDNA and RNA libraries analyzed in of *E. plorans* from Egypt.
Table S3.11. Counts of reference (Ref) and alternative (Alt= B-specific) alleles in all gDNA and RNA libraries analyzed in of *E. plorans* from Armenia.

4. Post-meiotic B chromosome expulsion, during spermiogenesis, in two grasshopper species

Published as:
Cabrero J, Martín-Peciña M, Ruiz-Ruano FJ, Gómez R, Camacho JPM. (2017). Post-meiotic B chromosome expulsion, during spermiogenesis, in two grasshopper species. *Chromosoma*, 126(5), 633-644. doi:10.1007/s00412-017-0627-8.

Abstract: Most supernumerary (B) chromosomes are parasitic elements carrying out an evolutionary arms race with the standard (A) chromosomes. A variety of weapons for attack and defense have evolved in both contending elements, the most conspicuous being B chromosome drive and A chromosome drive-suppression. Here we show, for the first time, that most micronuclei and microspermatids formed during spermiogenesis in two grasshopper species contain expulsed B chromosomes. By using DNA probes for B-specific satellite DNAs in *Eumigus monticola* and *Eyprepocnemis plorans*, and also 18S rDNA in the latter species, we were able to count the number of B chromosomes in standard spermatids submitted to fluorescence *in situ* hybridization (FISH), as well as visualizing B chromosomes inside most microspermatids. In *E. plorans*, the presence of B-carrying microspermatids in 1B males was associated with a significant decrease in the proportion of B-carrying standard spermatids. The fact that this decrease was apparent in elongating spermatids but not in round ones demonstrates that meiosis yields 1:1 proportions of 0B and 1B spermatids and hence that B elimination takes place postmeiotically, i.e. during spermiogenesis, implying 5-25% decrease in B transmission rate. In *E. monticola*, the B chromosome is mitotically unstable and B number varies between cells within a same individual. A comparison of B frequency between round and elongating spermatids of a same individual revealed a significant 12.3% decrease. We conclude that B chromosome elimination during spermiogenesis is a defense weapon of the host genome to get rid of parasitic chromosomes.

Keywords: *FISH*, *micronuclei*, *microspermatids*, *parasitic*, *satellite DNA*

Introduction

Regular chromosome elimination from somatic cells has been reported in nematodes, insects, mites, finches, bandicoots and hagfish, and has been interpreted as a mechanism for gene silencing, dosage compensation, sex determination, or germline and soma differentiation (for review, see Wang and Davis, 2014). Of course, this variety of adaptations is evolutionarily viable provided that the somatically eliminated chromosomes have granted their presence in the germ line. For the same reason, chromosome elimination from germ cells is most likely the result of a genetic conflict where the standard genome tries to get rid of a disturbing harmful element such as parasitic chromosomes.

B chromosomes are considered genomic parasites which prosper in natural populations because they show an advantage in transmission (drive) counteracting their detrimental effects on host genome fitness (for review, see Camacho et al., 2000; Camacho, 2005; Burt and Trivers, 2006). The presence of B chromosomes evokes an evolutionary response in the host genome leading to suppress drive, and the two contending parts develop a true coevolutionary arms race (Camacho et al., 1997; Frank, 2000) which may elicit the emergence of new adaptations in the host genome. A suggestive example of these adaptations are germ-line restricted B chromosomes, such as those in the marsupial *Echymipera kalabu* (Hayman et al., 1969) and the ant *Leptothorax spinosior* (Imai, 1974), as this minimizes their harm to the somatic cells while still assuring their transmission to future generations.

The formation of aberrant meiotic products during spermatogenesis has been a recurrent subject in the literature on B chromosomes. The postmeiotic part of spermatogenesis is called spermiogenesis, during which the round-shaped spermatids resulting from meiosis undergo drastic morphological changes becoming them into spermatozoa. In grasshoppers, spermatozoa possess extremely elongated heads showing almost the same width as the tail. During spermatid elongation, DNA packaging changes to a highly condensed state facilitated by histone replacement with protamines. Electron microscopy studies have shown that grasshopper spermiogenesis can be divided into ten developmental stages (Szöllösi, 1975). Under optical microscopy, however, it is only possible to differentiate between the immature round spermatids, the mature spermatozoa (with fibrillar heads) and the intermediate stages with elongating

spermatids at several degrees of elongation.

In addition to the temporally differentiated round (immature) and elongating (maturing) spermatids, optical microscopy allows identifying other types of spermatids, on the basis of size. In addition to standard haploid spermatids, polyploid spermatids have frequently been reported in grasshoppers. For instance, Cabrero et al. (2013) showed the presence of 2C, 4C, 8C and 16C macrospermatids in males that had been RNAi knocked-down for the *Ku70* gene. In addition, a few tiny microspermatids can sometimes appear within cysts of standard spermatids. However, the frequency of macro- and microspermatids has shown to be significantly higher in B-carrying males of several species (see below).

Nur (1969) was the first in claiming that the production of macro- and microspermatids could be related with the presence of B chromosomes in the grasshopper *Camnula pellucida*. This author suggested that lagging B chromosomes could block cytokinesis in both meiotic divisions leading to the formation of restitution nuclei and thus 2C or 4C macrospermatids. Alternatively, lagging B chromosomes could be excluded from the standard meiotic products giving rise to microspermatids. Other authors have later found aberrant spermatid formation in other species carrying B chromosomes (for review, see Teruel et al., 2009). Partial support to Nur's claiming and slightly different explanations were later given by other authors. For instance, Bidau (1986) reported that unequal cytokinesis in *Metalaptea brevicornis* gave rise to macrospermatids and a small nuclear bud which sometimes could include the B chromosome. Likewise, Suja et al. (1989) "observed the presence of condensed Bs outside the nuclei in both recently formed secondary spermatocytes and early spermatids" thus supporting the hypothesis that lagging Bs can be eliminated from "standard nuclei". However, based on the fact that spermatocytes within a same cyst are connected by cytoplasmic bridges as a result of incomplete cytokinesis (Phillips, 1970), Suja et al. (1989) suggested that macrospermatids could also derive from B-provoked impairment of spermatid differentiation during early spermiogenesis, which would explain the lack of correspondence they observed between the number of centriolar adjuncts and ploidy level in spermatids of the grasshopper *Eyprepocnemis plorans*. In addition, Loray et al. (1991) found that the presence of B chromosomes in *Dichroplus elongatus* was associated with an increase in the frequency of macrospermatids even in testis tubules lacking this mitotically unstable B chromosome, and claimed for

physiological effects of B's affecting meiosis even in cells lacking them. This kind of systemic response could also be explained through some kind of gene expression change due to B presence (included gene expression in the B itself) whose effects would be exported to B-lacking testis tubules. This kind of effect would be compatible with the spermiogenesis impairment suggested by Suja et al. (1989), but not with the odd-even effect frequently reported for the frequency of aberrant spermatids in the case of mitotically unstable B chromosomes, as they are most abundant in testis tubules carrying odd numbers of B chromosomes (see Camacho et al., 2004; Teruel et al., 2009) whereas the physiological effect should erase this difference.

It is however unknown whether microspermatids actually contain B chromosomes, as no direct evidence has hitherto been provided in animals. In contrast, using DNA probes specific to A or B chromosomes, Chiavarino et al. (2000) showed that "micronuclei formed during male meiosis in maize can include both A and B chromosomes". On this basis, and given that B-carrying males showed higher frequency of microspermatids than 0B ones in *E. plorans* (0.73% and 0.22%, respectively), Teruel et al. (2009) suggested that most microspermatids in this species presumably include B chromosomes, with a consequent decrease in B transmission rate.

The finding of repetitive DNAs which B chromosomes are very enriched for (e.g. ribosomal DNA) or else being specific to them (e.g. some satellite DNAs) allows getting an easy estimation of their transmission rate by simply visualizing them in the meiotic products by FISH for DNA probes being highly specific to B chromosomes. For instance, Milani et al. (2017) found U2 repeats in a B chromosome in the grasshopper *Abracris flavolineata* being useful for B chromosome identification in interphase cells, and they can be also useful for B transmission studies.

Here we analyze the presence of B chromosomes in standard and aberrant spermatids in two grasshopper species harboring B chromosome systems differing in mitotic stability. In *E. plorans*, B chromosomes are mitotically stable, meaning that they show the same number of Bs in all cells from a same individual. In *Eumigus monticola*, however, B chromosomes are mitotically unstable, so that the number of B chromosomes differs between the cells from different testis tubules but not between cells within a same tubule (Ruiz-Ruano et al., 2017). In each species, we have used DNA probes for FISH analysis which allowed scoring the number of B chromosomes in standard spermatids and demonstrated the presence of B chromosomes in most

microspermatids observed in both species. Remarkably, the standard spermatids showed a significant decrease in the frequency of B chromosomes between their round and elongating stages, suggesting that B chromosomes are eliminated during spermiogenesis.

Materials and methods

Adult males of the grasshoppers *Eyprepocnemis plorans* and *Eumigus monticola* were collected in natural populations from Spain, the former species in Alhama de Murcia (Murcia province), Salobreña (Granada), Otívar (Granada) and Torrox (Málaga), and the latter in Hoya de la Mora (Sierra Nevada, Granada). For the present analysis, we chose *E. plorans* males carrying a single B chromosome belonging to several different variants: B1 (four males from Alhama de Murcia and one from Torrox), B2 (four males from Salobreña and six from Otívar) and B24 (four males from Torrox). A description of these B chromosomes can be found in Cabrero et al. (2014). In the case of *E. monticola*, we used here one male carrying a mitotically unstable B chromosome, thus showing different B number in different cells.

Males were anesthetized with ethyl acetate vapors before dissection. Testes were fixed in 3:1 ethanol:acetic acid and stored at 4 °C. The number of B chromosomes was analyzed in squash preparations of testis tubules stained with acetic orcein. Fluorescent *in situ* hybridization (FISH), including DNA probe preparation and FISH reaction, was performed following the protocols described in Camacho et al. (2015a, b) and Ruiz-Ruano et al. (2017). The DNA probes employed in *E. plorans* were 18S ribosomal DNA (rDNA), which shows the largest cluster on B1 and B2 variants, and a B-specific satellite DNA recently found by us (EplTR112-11 in Chapter 1, named initially as EplSat115-11 in the original paper by Cabrero et al., 2017) which shows FISH signals only on B chromosomes. In *E. monticola*, we used a B-specific satellite DNA (EmoSat26-41) previously reported by Ruiz-Ruano et al. (2017). The electron microscope Figure 4.2f was obtained by the methods reported in Teruel et al. (2009).

Statistical analysis of spermatid counts in *E. plorans* was performed by a goodness-of-fit chi square test with null hypothesis predicting that 1B males produce 0B and 1B standard spermatids at Mendelian 1:1 proportion. This test was separately applied to round and elongating spermatids and a heterogeneity chi square test was also employed

to analyze within-population heterogeneity before testing the 1:1 proportion at population level. In *E. monticola*, however, the mitotic instability of the B chromosome did not allow applying the same null hypothesis and we compared the number of spermatids with 0-3 B chromosomes, between round and elongating ones, by the RxC software (provided by G. Carmody, Ottawa, Ontario, Canada), which performs chi-square tests in contingency tables, with permutation, and calculates P values by Monte Carlo methods. 20 batches of 2500 replicates were performed.

Results

Mitotically stable B chromosomes in *Eyprepocnemis plorans*

The B1 and B2 variants carry the largest block of rDNA in B-carrying genomes (Fig. 4.1a), so that FISH with an rDNA probe allows easy identification of B-carrying and B-lacking round and elongating spermatids (Fig. 4.1b-d). B24, however, carries a smaller rDNA block (see Cabrero et al., 2014) and this marker does not discriminate properly between B24+ and B24- meiotic products. However, the B-specific satellite DNA (EplTR112-11) shows conspicuous clusters on both ends of the B24 chromosome (Fig. 4.1e) and is clearly apparent in spermatids as one or two small dots (Fig. 4.1f-h).

We analyzed the presence of B chromosomes in round (strictly circular) spermatid nuclei (Fig. 4.1b, f) and also in elongating ones (i.e. showing elliptic to spearhead shape) (Fig. 4.1c, d, g, h). The four males from the Alhama de Murcia population, carrying one B1 chromosome, showed about similar proportions of B-carrying and B-lacking spermatids at both round and elongating stages (Table 4.1), thus showing a Mendelian rate of B chromosome transmission (k_B). In the Salobreña and Torrox populations, which harbor the B2 and B24 variants, respectively, no significant difference was observed between B-carrying and B-lacking round spermatids (Table 4.1). However, two males in each population showed a significant deficit of B-carrying elongating spermatids, and chi-square tests applied to the totals in each population (availed by the heterogeneity chi-square test) yielded significant decreases in the transmission rate of these 1B males (k_B being 0.455 for B2 in Salobreña and 0.463 for B24 in Torrox) (Table 4.1).

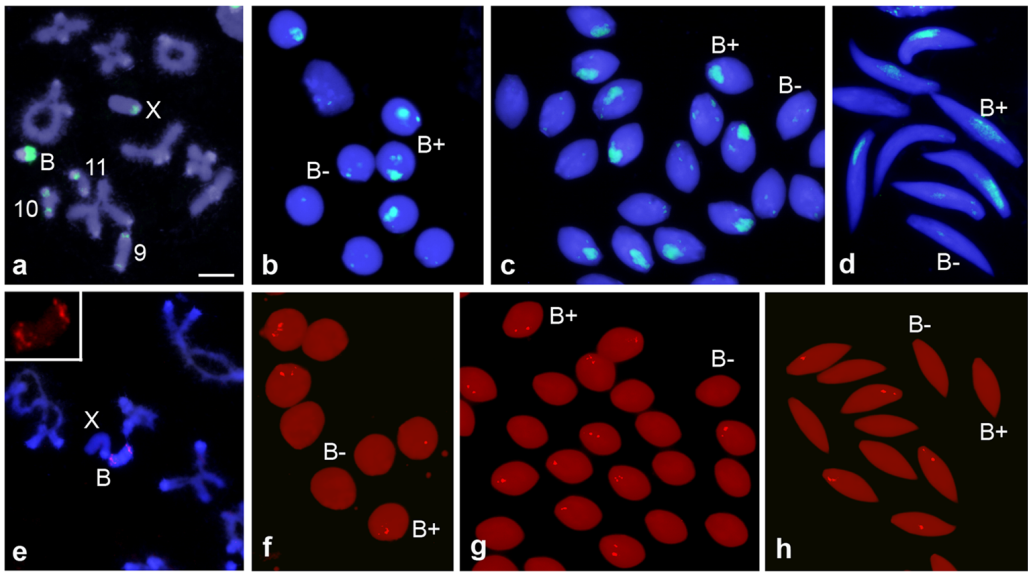


Figure 4.1. Detection of B chromosomes in primary spermatocytes at diplotene (a and e), round spermatids (b and f) and elongating spermatids (c, d, g and h) of the grasshopper *Eyprepocnemis plorans*, by means of FISH for 18S rDNA (a-d) and the B-specific EplTR112-11 satellite DNA (e-h), as DNA probes. Hybridization signals in a-e are merged with DAPI staining. Inset in e depicts the same B in the diplotene cell, at higher magnification, showing satellite location. In a, note that the B1 chromosome in the Alhama de Murcia population carries the largest cluster for 18S rDNA and that this allows identifying B-carrying spermatids in b-d. In f-h, note that B24-carrying spermatids are identified by the presence of the B-specific satellite (small dots) in Torrox males. Bar in a indicates 5 µm for a and e, and 10 µm for the remaining photographs.

In the Otívar population, which also harbors the B2 variant, we scored only elongating spermatids in six males and all of them showed k_B lower than 0.5, but the difference with the Mendelian one was not significant. However, as a whole, they showed a significant tendency to B elimination (Table 4.1).

In the Torrox population, we found a male carrying one B1 chromosome, a very unusual event in this population where B24 is the most frequent variant. Remarkably, this male showed about similar proportions of B-carrying and B-lacking round spermatid nuclei, but a significantly lower proportion of B-carrying elongating spermatids rendering a low B transmission rate (k_B= 0.375) (Table 4.1).

Table 4.1. Frequency of B-lacking (B-) and B-carrying (B+) spermatids in 19 males of the grasshopper *E. plorans* collected at four Spanish populations. B chromosome presence was identified by FISH for 18S rDNA and a B-specific satellite (EplTR112-11). B transmission rate (k_B) was calculated as the proportion of B-carrying spermatids. ST spermatid type, i.e., round (R) or elongated (E). Goodness-of-fit χ2 tested the null hypothesis that the B chromosome was transmitted at Mendelian rate (k_B = 0.5). Heterogeneity χ2 is calculated as the sum of all individual χ2 minus the χ2 of total spermatid numbers. All χ2 have one degree of freedom (df) except the heterogeneity one where they are calculated as the sum of individual df values minus 1 df of the χ2 for the sum of all spermatids. P values <0.05 are noted in bold-type letter.

Population	Id	B type	FISH marker	ST	B-	B+	k_B	chi	P
Alhama	m1	B1	rDNA	R	127	131	0.508	0.06	0.80334
				E	168	180	0.517	0.41	0.52005
	m2	B1	rDNA	R	76	60	0.441	1.88	0.17007
				E	107	85	0.443	2.52	0.11235
	m4	B1	rDNA	R	48	46	0.489	0.04	0.83657
				E	156	140	0.473	0.86	0.35238
	m12	B1	rDNA	R	87	89	0.506	0.02	0.88017
				E	141	142	0.502	0.00	0.95260
Total				R	338	326	0.491	0.22	0.64144
				E	572	547	0.489	0.56	0.45485
Heterogeneity				R				1.79	0.6165
				E				3.24	0.3554
Salobreña	m8	B2	rDNA	R	144	138	0.489	0.13	0.72087
				E	186	156	0.456	2.63	0.10476
	m10	B2	rDNA	R	212	232	0.523	0.90	0.34254
				E	245	188	0.434	7.50	**0.00616**
	m16	B2	rDNA	R	46	49	0.516	0.09	0.75824
				E	63	67	0.515	0.12	0.72572
	m25	B2	rDNA	R	125	92	0.424	5.02	**0.02508**
				E	115	97	0.458	1.53	0.21637
Total				R	527	511	0.492	0.25	0.61946
				E	609	508	0.455	9.13	**0.00251**
Heterogeneity				R				5.90	0.11683
				E				2.65	0.44811
Otívar	m11	B2	rDNA	E	237	203	0.461	2.63	0.10504
	m12	B2	rDNA	E	212	188	0.470	1.44	0.23014
	m14	B2	rDNA	E	271	258	0.488	0.32	0.57193
	m17	B2	rDNA	E	109	94	0.463	1.11	0.29244
	m18	B2	rDNA	E	177	163	0.479	0.58	0.44770
	m21	B2	rDNA	E	194	182	0.484	0.38	0.53601
Total				E	1200	1088	0.476	5.48	**0.01921**
Heterogeneity				E				6.45	0.26446

Torrox	m02	B24	Sat112-11	R	125	122	0.494	0.04	0.84862
				E	121	110	0.476	0.52	0.46922
	m18	B24	Sat112-11	R	99	120	0.548	2.01	0.15588
				E	155	153	0.497	0.01	0.90927
	m21	B24	Sat112-11	R	193	206	0.516	0.42	0.51517
				E	207	166	0.445	4.51	**0.03376**
	m27	B24	Sat112-11	R	207	187	0.475	1.02	0.31365
				E	198	158	0.444	4.49	**0.03401**
Total				R	624	635	0.504	0.10	0.75655
				E	681	587	0.463	6.97	**0.00830**
Heterogeneity				R				3.39	0.33493
				E				2.57	0.46287
Torrox	m26	B1	rDNA	R	156	155	0.498	0.00	0.95478
				E	172	103	0.375	17.31	**0.00003**
Grand total				R	1645	1627	0.497	0.10	0.75301
				E	3234	2833	0.467	26.50	**<0.00001**

The observed k_B in elongating spermatids implied only residual B loss in Alhama for the B1 variant (2.23%), but it was higher in Otívar (4.9%) and Salobreña (9.04%) for B2, as well as in Torrox for B24 (7.41%) and B1 (25.1%).

FISH analysis showed that these decreases in B transmission rate (k_B) were paralleled by the presence of B-carrying macro- and microspermatids (Fig. 4.2a, b), and we scored them in cysts containing round spermatids in 13 males (excepting those from Otívar) and in cysts of elongating ones in all 19 males analyzed (Table 4.2). Multiple regression analysis, with k_B as dependent variable and the proportion of B-carrying macro- and microspermatids as independent variables, showed that k_B was independent of the frequency of these two types of aberrant gametes in the cysts containing round spermatids (micronuclei: r= 0.08, N= 13, t= 0.24, df= 10, P= 0.82; macronuclei: r= -0.12, N= 13, t= 0.35, df= 10, P= 0.73). However, in the cysts of elongating spermatids, k_B was significantly negatively correlated with the frequency of B-carrying microspermatids (r= -0.54, N= 19, t= 2.47, df= 16, P= 0.025) but not with the frequency of B-carrying macrospermatids (r= -0.25, N= 19, t= 1.12, df= 16, P= 0.28). This suggests that microspermatid formation is related with a decrease in k_B whereas macrospermatids are not, confirming predictions by Teruel et al. (2009).

As Table 4.2 shows, not all microspermatids carried a B chromosome, the main exception being 29 round spermatids, all found in the m16 male from Salobreña, showing a nuclear bud containing a long chromosome carrying a small cluster of rDNA, which allowed identifying it as the X chromosome (Fig. 4.2c). All these 29 nuclei carried the B chromosome, and most of them showed the X chromosome still stuck to the nucleus, excepting one which was partially separated but still contacting by its end carrying rDNA, i.e. its centromeric region, and another nucleus showing the X chromosome completely separated from it (Fig. 4.2c). In elongating spermatids, we only observed four B-lacking microspermatids (Fig. 4.2c), one in m16 from Salobreña and three in m27 from Torrox, the 31 remaining microspermatids carrying the B chromosome.

It was highly remarkable that all B-carrying microspermatids observed by us were placed very close to a standard B-lacking spermatid (Fig. 4.2d, e), suggesting that the former derived from the same nuclei as the latter and that both share the same cytoplasm. This is also inferred from the fact that our preparations were made by squashing, so that the likelihood that the 47 B-carrying microspermatids were adjacent to a B-lacking standard spermatid would be negligible unless they share the same cytoplasm. In fact, some of the observed B-carrying microspermatids were physically in contact with an adjacent B-lacking nucleus whereas others did not contact with the nucleus and were found at different distance from the B-lacking nucleus (Fig. 4.2d,e), suggesting that microspermatids are finally expelled from the standard spermatids. Remarkably, a review of the photographs made by us in a previous analysis of spermatogenesis under electronic microscope (Teruel et al., 2009), revealed the presence of microspermatids sharing the same cytoplasm as standard spermatid nuclei, and also the presence of very similar dense bodies outside spermatids which could correspond to remains of microspermatids extruded from the cytoplasm (Fig. 4.2f).

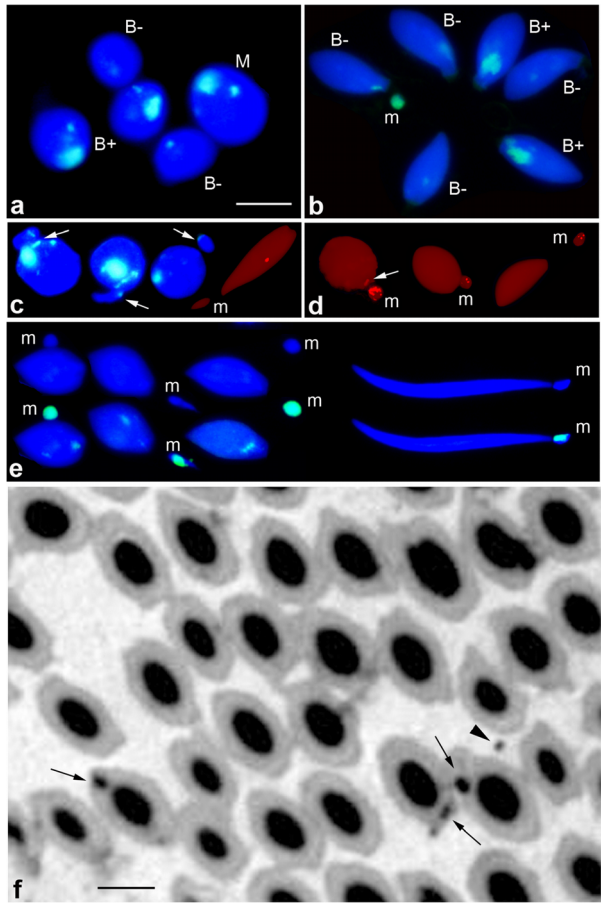


Figure 4.2. Presence of macro- and microspermatids in the grasshopper *E. plorans*. a) Two B-carrying (B+) and two B-lacking (B-) standard spermatids, and one macrospermatid (M). b) Six standard elongating spermatids, four of which lack B chromosomes (B-) and two carry the B chromosome (B+). Note the presence of a B-carrying microspermatid (m). c) Three round spermatids (on the left) showing a large chromosome being apparently extruded from the nucleus. Note that this chromosome carries a small rDNA cluster, which allows identifying it as the X chromosome (arrow). Note in the nucleus at the center that the centromere region, indicated by rDNA location, is still contacting the nucleus, whereas in the nucleus on the right the whole X chromosome has lost contact with the main nucleus. The elongating B-carrying standard spermatid, on the right, was exceptional by lying beside a B-lacking microspermatid (m). d) Examples of microspermatids (m) lying at different distances from a B-lacking standard spermatid. Note the presence of a small FISH signal in the main nucleus on the left (arrow). e) Additional examples of microspermatids (m) laying by a B-lacking standard spermatid, showing DAPI staining (upper row) and FISH+DAPI (lower row). f) Electron microscope photograph of cross sectioned standard spermatid nuclei (dense bodies), showing the presence of small dense bodies appearing to be microspermatids, some of which share the cytoplasm with the main nucleus (arrows) and one is outside (arrowhead). Bar in a indicates 10 µm for a-e, and that in f equals 1 µm.

Table 4.2. Frequency of B-carrying (B+) and B-lacking (B-) micro- and macrospermatids (M) in 19 males of the grasshopper *E. plorans*, collected at four populations, namely Alhama (A), Salobreña (S), Otívar (O) and Torrox (T). ST= Spermatid type, i.e. round (R) or elongated (E); kB= B transmission rate estimated in normal spermatids (see Table 4.1); frB+-= frequency of B-carrying or B-lacking microspermatids calculated as the proportion between the observed number and the total number of normal spermatids (see values in Table 1); frM= proportion of macrospermatids in respect to the total number of normal spermatids.

Pop	Id	B	ST	k_B	Microspermatids					Macrospermatids		
					B+	B-	Total	frB+	frB-	B+	B-	frM
A	m1	B1	R	0.51	0	0	0	0	0	0	0	0
			E	0.52	0	0	0	0	0	3	0	0.9%
	m2	B1	R	0.44	0	0	0	0	0	2	0	1.5%
			E	0.44	0	0	0	0	0	5	0	2.6%
	m4	B1	R	0.49	0	0	0	0	0	0	0	0
			E	0.47	1	0	1	0.3%	0	2	0	0.7%
	m12	B1	R	0.51	0	0	0	0	0	2	0	1.1%
			E	0.50	1	0	1	0.4%	0	8	0	2.8%
Total			R	0.49	0	0	0	0	0	4	0	0.6%
			E	0.49	2	0	2	0.2%	0	18	0	1.6%
S	m8	B2	R	0.49	3	0	3	1.1%	0	14	0	5.0%
			E	0.46	0	0	0	0	0	3	0	0.9%
	m10	B2	R	0.52	0	0	0	0	0	5	0	1.1%
			E	0.43	4	0	4	0.9%	0	1	0	0.2%
	m16	B2	R	0.52	0	29	29	0	30.5%	4	0	4.2%
			E	0.52	0	1	1	0	0.8%	0	0	0
	m25	B2	R	0.42	0	0	0	0	0	3	0	1.4%
			E	0.46	0	0	0	0	0	5	0	2.4%
Total			R	0.49	3	29	32	0.3%	2.8%	26	0	2.5%
			E	0.45	4	1	5	0.4%	0.1%	9	0	0.8%
O	m11	B2	E	0.46	0	0	0	0	0	0	0	0
	m12	B2	E	0.47	0	0	0	0	0	0	0	0
	m14	B2	E	0.49	1	0	1	0.2%	0	0	0	0
	m17	B2	E	0.46	1	0	1	0.5%	0	0	0	0
	m18	B2	E	0.48	1	0	1	0.3%	0	0	0	0
	m21	B2	E	0.48	2	0	2	0.5%	0	0	0	0
Total			E	0.48	5	0	5	0.2%	0	0	0	0
T	m2	B24	R	0.49	0	0	0	0	0	0	0	0
			E	0.48	0	0	0	0	0	0	0	0
	m18	B24	R	0.55	0	0	0	0	0	0	0	0
			E	0.50	0	0	0	0	0	0	0	0
	m21	B24	R	0.52	0	0	0	0	0	0	0	0
			E	0.45	0	0	0	0	0	0	0	0
	m27	B24	R	0.47	0	0	0	0	0	1	0	0.3%
			E	0.44	0	3	3	0	0.8%	13	0	3.7%
Total			R	0.50	0	0	0	0	0	1	0	0.1%
			E	0.46	0	3	3	0	0.2%	13	0	1.0%
T	m26	B1	R	0.50	13	0	13	4.2%	0	10	0	3.2%
			E	0.37	20	0	20	7.3%	0	15	0	5.5%
Grand total			R		16	29	45	0.5%	1.0%	31	0	1.0%
			E		31	4	35	0.9%	0.1%	40	0	1.1%

Assuming that every B-carrying microspermatid implied the conversion of a B-carrying standard spermatid into a B-lacking one due to B chromosome loss, we can calculate the expected frequency of B+ and B- standard spermatids and test whether this explain the observed k_B in round and elongating ones. In the case of round ones, the analysis of 3,272 standard spermatids indicated k_B= 0.497, and we found 16 B-carrying micronuclei (see Table 4.2). The expected frequencies of B+ and B- round spermatids is thus 3,272*0.5-16= 1,620 B+ and 3,272*0.5+16= 1,652 B- (k_B= 0.495), and a goodness-of-fit chi square test comparing these expected frequencies with the observed ones (1,627 and 1,645, respectively) indicated the absence of significant difference (χ^2= 0.06, df= 1, P= 0.8066). On the contrary, we observed 2,833 B+ and 3,234 B- elongating spermatids (k_B= 0.467) plus 31 B-carrying microspermatids, and the expected frequencies, namely 6,067*0.5-31= 3,002.5 B+ and 6,067*0.5+31= 3,064.5 B- (k_B= 0.495) differed significantly from the observed ones (χ^2= 18.94, df= 1, P= 0.00001). This indicates that the observed amount of microspermatids does not explain the decrease in k_B observed in elongating spermatids. A possible explanation is that a fraction of the microspermatids produced are finally degraded and lost, so that we are able to visualize only part of those actually formed. We calculated that the loss of B chromosomes in 200 microspermatids (instead of the 31 B-carrying ones observed) would have yielded the observed k_B= 0.467, implying that we detected only 16% of the B chromosomes lost as microspermatids.

Mitotically unstable B chromosomes in *Eumigus monticola*

The exclusive presence of EmoSat26-41 in the B chromosome of the grasshopper *E. monticola* (Fig. 4.3a, b) (see also Ruiz-Ruano et al., 2017) allows scoring the number of B chromosomes in spermatid nuclei submitted to FISH with a probe for this satellite DNA (Fig. 4.3c-h). B chromosomes in this species are mitotically unstable, implying that B number varies among cells within a same individual, but not within a same testis tubule. For this reason, we analyzed round and elongating spermatids in the same six testis tubules and compared B frequency between these two kinds of standard spermatids. In total, we analyzed 911 round spermatid nuclei (355 with 0B, 465 with 1B, 89 with 2B and 2 with 3B) and 442 elongating spermatids (210 with 0B, 193 with 1B, 34 with 2B and 5 with 3B), and found a significant decrease in the mean number of B chromosomes between round (0.71) and elongating (0.62) spermatid nuclei (RxC contingency test: P= 0.0004, SE= 0.0002). This suggests that B chromosomes in *E. monticola* undergo about 12.3% elimination during spermiogenesis [0.71-0.62)/0.71= 0.123], as was also evidenced by the

presence of 3% of B-carrying micronuclei associated with round spermatids (Fig. 4.3c, d) and 5% of B-carrying microspermatids associated with elongating spermatids (Fig. 4.3e, f). Likewise in *E. plorans*, the observed frequency of microspermatids is lower than the 12.3% decrease in B frequency, implying that we observed only about 42% (0.05/0.123) of B losses in form of microspermatids, presumably because many of them are finally degraded.

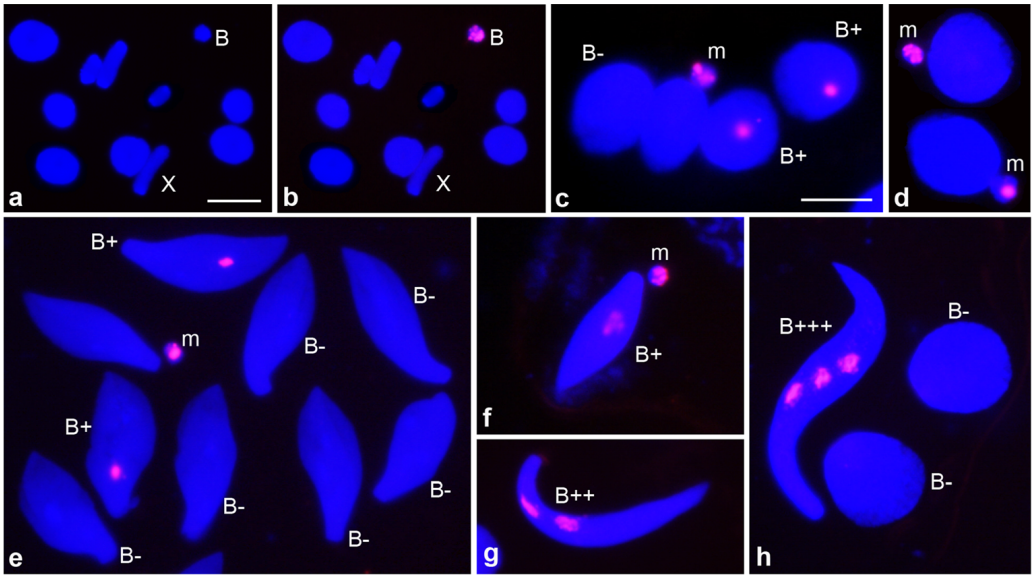


Figure 4.3. Detection of B chromosomes in primary spermatocytes at metaphase I (a-b) and spermatids (c-h) of the grasshopper *Eumigus monticola*, submitted to FISH for the EmoSat26-41 B-specific satellite DNA. Examples of B-carrying (B+) and B-lacking (B-) standard round and elongating spermatids are shown in c-h. In c-e, note the presence of microspermatids (m) beside a B-lacking standard spermatid. Mitotic instability of B chromosomes in this species explains the presence of B-carrying standard spermatids beside a B-carrying microspermatid (f), since standard spermatids in this species can carry two (g) or three (h) B chromosomes. The number of B chromosomes within a sperm nucleus is indicated by the number of plus signs. Bar in a indicates 5 µm for a and b, that in c indicates 10 µm for c, f and g, and that in e indicates 10 µm for d, e and h.

Discussion

Population invasion by a parasitic B chromosome needs some kind of drive (Camacho et al., 1997). In *E. plorans*, we observed that B chromosomes show drive in some populations (Zurita et al., 1998) but not in others (López-León et al., 1992), as a consequence of drive suppression (Herrera et al., 1996; Camacho et al., 1997). In *E. monticola*, however, nothing is known at this respect. However, our present results suggest that the loss during spermiogenesis would have impeded its birth as a B chromosome. It is thus likely that this B chromosome show drive at other stage of the reproductive cycle. Its mitotic instability suggests possible B accumulation based on mitotic non-disjunction during early cleavage divisions, with preferential destiny of mitotic products carrying more B chromosomes towards the germ line. This kind of accumulation has been reported for mitotically unstable B chromosomes of grasshopper species such as, for instance, *Calliptamus palaestinensis* (Nur, 1963), *Camnula pellucida* (Nur, 1969) and *Locusta migratoria* (Nur, 1969; Kayano, 1971; Viseras et al., 1990). In the latter species, premeiotic accumulation of B chromosomes represents about 30% increase in male B transmission, but it is counteracted by 20% decrease during subsequent stages of the reproductive cycle, including the formation of microspermatids, the net B transmission thus implying about 10% accumulation in males (Pardo et al., 1994). In addition, this B chromosome shows 62% accumulation during female transmission (Pardo et al., 1994), which explains the worldwide distribution of B chromosomes in *L. migratoria*.

Our results have shown a significant decrease in B transmission rate (k_B) during spermiogenesis in two species of grasshopper carrying B chromosomes. In the case of *E. plorans*, males carrying one mitotically stable B chromosome yielded meiotic products at the Mendelian rate, given that the half of round spermatids carried the B chromosome. Therefore, spermiogenesis in these males begins with 1:1 proportion of B-carrying and B-lacking round spermatids. In contrast, most males showed a tendency to a decreased proportion of B-carrying elongating spermatids, which was significant in five males and, as a whole, in Salobreña, Otívar and Torrox populations (see Table 4.1). Therefore, the k_B decrease takes place necessarily during spermiogenesis. We also demonstrate here that k_B was negatively correlated with the frequency of B-carrying microspermatids, suggesting that B chromosomes are lost during spermiogenesis in the form of

microspermatids. Our FISH visualization of B chromosomes within microspermatids constitutes the first direct demonstration of Nur's claiming that microspermatids are a way of B chromosome loss (Nur, 1969). Remarkably, we only observed B-carrying macrospermatids, even though some B-lacking ones might be expected if cytokinesis failures would take place in B-lacking secondary spermatocytes. This suggests a direct role of lagging B chromosomes in the formation of macrospermatids, as was also suggested by Nur (1969).

In addition, the fact that B-carrying microspermatids were always found beside a B-lacking standard spermatid, even in squash preparations, along with the presence of microspermatids sharing a common cytoplasma with standard spermatid nuclei at electronic microscope images, suggest a causal relationship between microspermatid formation, B chromosome loss and the decrease in k_B in standard spermatids. The conventional explanation for microspermatids is that they contain B chromosomes lagged during the precedent meiotic divisions, as was first suggested by Nur (1969) and later supported by other authors (Pearse and Ehrlich, 1979; Viseras and Camacho, 1985; Bidau 1987; Teruel et al., 2009; Abdel-Haleem et al., 2009). Our present results, however, challenge this hypothesis. Of course, we cannot rule out that some of the 16 B-carrying microspermatids found within cysts containing round standard spermatids could have derived from B chromosomes lagged during previous meiosis which failed to properly integrate into the main nucleus, and even some of the 31 B-carrying microspermatids observed in the cysts of elongating spermatids could actually have derived from them. But, if meiosis were the only source of microspermatids, we should observe similar values of k_B in round and elongating spermatids, and this was not the case in the two species analyzed here, thus clearly implying micronucleus formation during spermiogenesis.

The finding that micronuclei can be formed by nuclear budding in interphase cells could provide a mechanistic support to B chromosome elimination during spermiogenesis. Nuclear budding and micronucleus formation are common characteristics to many cell cultures frequently leading to chromosome elimination (Elston, 1963; Longwell and Yerganian, 1965). The classical mechanism of micronucleus formation claims that they incorporate lagging chromosomes during mitosis (Heddle and Carrano, 1977; Schubert and Oud, 1997; Fenech et al., 2011). However, recent findings have shown that nuclear budding and micronucleus formation can also occur in

interphase cells (Gernand et al., 2005, 2006; Utani et al., 2011; Ishii et al., 2016). Similarly, spermatids can form micronuclei during spermiogenesis without the involvement of any additional cell division. In fact, our results show remarkable similarities with some characteristics of interphase micronucleus formation described by the former authors.

For instance, Gernand et al. (2005, 2006) reported that the chromosomes destined to elimination occupied a peripheral location in interphase cells of interspecific hybrids. This appears to be a general tendency since other chromosomes being regularly eliminated also occupy peripheral locations, such as E chromosomes in Cecidomyiidae (Kloc and Zagrodzinska, 2001), the germ-line restricted chromosomes (GRC) in the zebra and the Bengalese finches (Schoenmakers et al., 2010; Del Priore and Pigozzi, 2014), and even acentric, autonomously replicating extrachromosomal structures called double-minute chromosomes (DMs) (Shimizu et al., 1998).

Interestingly, a tendency of B chromosomes to occupy peripheral locations in the nucleus during cell division was early noted by Avdulow (1933) in maize (see also Randolph, 1941; Darlington and Upcott, 1941; Carlton and Cande, 2002). Subsequent research has reached the same conclusion for B chromosomes in *Poa alpina* (Hakansson, 1948), *Dactylis* (Williams and Barclay, 1972) and rye (Jones, 1995; Morais-Cecílio et al., 1996; Langdon et al., 2000). In animals, the Paternal Sex Ratio (PSR) is an extremely parasitic B chromosome which localizes to the outer periphery of the paternal nucleus and at the tip of the sperm nucleus, but in this case the B chromosome escapes from elimination which is focused on the paternal standard set (Swin et al., 2012). These authors visualized PSR by FISH in spermatids and mature sperm and about 98% of them, in both cases, carried the B chromosome, so that we can infer that PSR is not eliminated at all during spermiogenesis.

Sex chromosomes in animals also occupy peripheral locations (see Turner, 2007; Finch et al., 2008; Calvente et al., 2013) and are inactivated during meiosis by means of epigenetic marks (Vaskova et al., 2010). Likewise, in *E. plorans*, X and B chromosomes are heterochromatic, they show frequent non-homologous association during first meiotic prophase (Camacho et al., 1980) and are hypoacetylated for H3K9 during entire meiosis (Cabrero et al., 2007). They also tend to occupy peripheral location in meiotic nuclei, which probably facilitates their elimination in the form of micronuclei. It is tempting to speculate that the high similarity between X and B chromosomes during meiosis may lead to eventual X chromosome elimination, presumably because some of the epigenetic

marks used for micronucleus formation are common to these two chromosomes.

Another resemblance of our present results with those in interspecific hybrids is that centromeric regions of pearl millet chromosomes are the last in being eliminated in wheat-pearl millet hybrids (Gernand et al., 2005). We observed this same fact in the case of the X chromosome elimination in m16 from Salobreña (see Fig. 4.2c). Interestingly, Gernand et al. (2005) suggested that micronucleus formation can eventually leave the centromeric region of the expulsed chromosome in the main nucleus thus opening the possibility to *de novo* formation of B chromosomes in interspecific crosses. Our Fig. 4.2d shows a micronucleus, beside a round spermatid, which harbor most of the B-specific satellite except a small FISH signal remaining in the main nucleus, indicating that micronucleus extrusion of the B chromosome can be incomplete, thus giving indirect support to Gernand et al. claiming.

In addition, Gernand et al. (2005) suggested that post-translational histone modification might play a role in chromosome elimination, as differential acetylation of histones H3 and H4 and methylation of histone H3 had been reported in chromosome elimination in sciarid flies (Goday and Ruiz, 2002) and in programmed DNA elimination in *Tetrahymena* (Taverna et al., 2002). In addition, the GRC chromosome in the zebra fich is silenced from early leptotene onwards, and is eliminated through micronucleus formation following metaphase I (Schoenmakers et al., 2010). It is thus presumable that the observed H3K9 hypoacetylation of X and B chromosomes in *E. plorans* (Cabrero et al., 2007) may serve as a signal for elimination through the evolutionary conserved mechanism suggested by Gernand et al. (2005, 2006).

Our present results suggest that even in organisms where chromosome elimination occurs only sporadically (e.g. B chromosome loss during spermiogenesis) interphase cells appear to show the ability to eliminate chromosomes through micronucleus formation. The parasitic nature of B chromosomes makes them an elimination target with high fitness reward for the host genome. Ideally, the best situation for a B chromosome would be to remain limited to the germ line by being eliminated from somatic cells thus minimizing harmful effects on the host. Examples of germ-line restricted B chromosomes have been found, for example, in the marsupial *Echymipera kalabu* (Hayman et al., 1969) and the ant *Leptothorax spinosior* (Imai, 1974). Even in this case B chromosome presence in the germ line is still a load for the host genome, as it has to replicate extra DNA without a reward, except in the case that the B chromosome

carries a gene whose activity result profitable for the host (e.g. see Miao et al., 1991). In most cases, however, it is expected that the host genome continues trying to get rid of the parasitic element. The existence of postmeiotic elimination mechanisms like that shown here might help in this task, but it does not always work. Suggestive examples are germ-line restricted chromosomes like those reported in diptera (Bauer and Beermann, 1952; Staiber, 1988; Herrick and Seger 1999; Goday and Esteban, 2001) or zebra finches (Pigozzi and Solari, 2005; Schoenmakers et al., 2010; Del Priore and Pigozzi, 2014), as they could actually be the last face of obstinately resistant parasitic B chromosomes.

Acknowledgments

We thank M. Teruel and J.D. Alché for their help to obtain the electron microscope photograph in Fig. 4.2f.

References

Abdel-Haleem AA, Sharaf HM, El-Kabbany AI. (2009). New record of B-chromosome through meiosis in the Egyptian locust *Anacridium aegyptium* (Acrididae) with indication to its origin. *Journal of King Saud University – Science*, 21, 163–166.

Avdulow NP. (1933). On the additional chromosomes in maize. *Bull Appl Bot Ser*, 2, 101–130.

Bauer H, Beermann W. (1952). Der chromosomencyclus der Orthocladiinen (Nematocera, Diptera). *Z Naturforsch*, 7, 557– 563.

Bidau CJ. (1986). Effects on cytokinesis and sperm formation of a B-isochromosome in *Metaleptea brevicornis adspersa* (Acridinae, Acrididae). *Caryologia*, 39, 165–177.

Bidau CJ. (1987). Influence of a rare unstable B-chromosome on chiasma frequency and nonhaploid sperm production in *Dichroplus pratensis* (Melanoplinae, Acrididae). *Genetica*, 73, 201–210.

Burt A, Trivers R. (2006). Genes in conflict: the biology of selfish genetic elements. Belknap Press of Harvard University Press, Cambridge.

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, 135–42.

Cabrero J, Bakkali M, Navarro-Domínguez B, Ruiz-Ruano FJ, Martín-Blázquez R, López-León MD, et al. (2013). The Ku70 DNA-repair protein is involved in centromere function in a grasshopper species. *Chromosome Research*, 21, 393–406.

Cabrero J, López-León MD, Ruiz-Estévez M, Gómez R, Petitpierre E, Rufas JS, et al. (2014). B1 was the ancestor B chromosome variant in the western Mediterranean area in the grasshopper *Eyprepocnemis plorans*. *Cytogenetic and Genome Research*, 142, 54–8.

Calvente A, Viera A, Parra MT, De La Fuente R, Suja JA, Page J, et al. (2013). Dynamics of cohesin subunits in grasshopper meiotic divisions. *Chromosoma*, 122, 77–91.

Camacho JPM. (2005). B chromosomes. In: Gregory TR, ed. The evolution of the genome. New Yor:K Academic Press. 223–286.

Camacho JPM, Carballo AR, Cabrero J. (1980). The B-chromosome system of the grasshopper *Eyprepocnemis plorans* subsp. *plorans* (Charpentier). *Chromosoma*, 80, 163–176.

Camacho JPM, Shaw MW, López-León MD, Pardo MC, Cabrero J. (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. Series B, Biological Sciences*, 355, 163–178.

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, 325–31.

Camacho JPM, Cabrero J, López-León MD, Cabral-de-Mello DC, Ruiz-Ruano FJ. (2015a). Grasshoppers (Orthoptera). In: Sharakhov IV (ed) Protocols for cytogenetic mapping of arthropod genomes. *CRC Press*, 381–438.

Camacho JPM, Ruiz-Ruano FJ, Martín-Blázquez R, López-León MD, Cabrero J, Lorite P, et al. (2015b). A step to the gigantic genome of the desert locust: chromosome sizes and

repeated DNAs. *Chromosoma*, 124, 263–275.

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

Chiavarino AM, Rosato M, Manzanero S, Jiménez G, González-Sánchez M, Puertas MJ. (2000). Chromosome nondisjunction and instabilities in tapetal cells are affected by B chromosomes in maize. *Genetics*, 155, 889–897.

Darlington CD, Upcott MB. (1941). The activity of inert chromosomes in *Zea mays*. *Journal of Genetics*, 41, 275–296.

Del Priore L, Pigozzi MI. (2014). Histone modifications related to chromosome silencing and elimination during male meiosis in Bengalese finch. *Chromosoma*, 123, 293–302.

Elston RN. (1963). Nuclear budding and micronuclei formation in human bone marrow, skin and fascia lata cells in vitro and in oral mucosa cells in vivo. *Acta pathologica et microbiologica Scandinavica. Section B*, 59, 195–199.

Fenech, M, Kirsch-Volders M, Natarajan AT, Surralles J, Crott JW, Parry J, et al. (2011). Molecular mechanisms of micronucleus, nucleoplasmic bridge and nuclear bud formation in mammalian and human cells. *Mutagenesis*, 26, 125–132.

Finch KA, Fonseka KGL, Abogrein A, Ioannou D, Handyside AH, Thornhill AR, et al. (2008). Nuclear organization in human sperm: Preliminary evidence for altered sex chromosome centromere position in infertile males. *Human Reproduction*, 23, 1263–1270.

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

Gernand D, Rutten T, Varshney A, Rubtsova M, Prodanovic S, Brüß, et al. (2005). Uniparental chromosome elimination at mitosis and interphase in wheat and pearl millet crosses involves micronucleus formation, progressive heterochromatinization, and DNA fragmentation. *Plant Cell*, 17, 2431–2438.

Gernand D, Rutten T, Pickering R, Houben A. (2006). Elimination of chromosomes in *Hordeum vulgare* x *H. bulbosum* crosses at mitosis and interphase involves micronucleus formation and progressive heterochromatinization. *Cytogenetic and Genome Research*, 114, 169–174.

Goday C, Esteban MR. (2001). Chromosome elimination in sciarid flies. *BioEssays*, 23, 242–250.

Goday C, Ruiz MF. (2002). Differential acetylation of histones H3 and H4 in paternal and maternal germline chromosomes during development of sciarid flies. *Journal of Cell Science*, 115, 4765–4775.

Hakansson A. (1948). Embryology of *Poa alpina* plants with accessory chromosomes. *Hereditas*, 34, 233–247.

Hayman DL, Martin PG, Waller PF. (1969). Parallel mosaicism of supernumerary chromosomes and sex chromosomes in *Echymipera kalabu* (Marsupialia). *Chromosoma*, 27, 371–380.

Heddle JA, Carrano AV. (1977). The DNA content of micronuclei induced in mouse bone marrow by gamma irradation: Evidence that micronuclei arise from acentric chromosomal fragments. *Mutation Research*, 44, 63–69.

Herrera J, López-León M, Cabrero J, Shaw M, Camacho JPM. (1996). Evidence for B chromosome drive suppression in the grasshopper *Eyprepocnemis plorans*. *Heredity*, 76, 633–639.

Herrick G, Seger J. (1999). Imprinting and paternal genome elimination in insects. In: *Results and Problems in Cell Differentiation*, 25, 41–71.

Imai HT. (1974). B–chromosomes in the Myrmicine ant, *Leptothorax spinosior*. *Chromosoma*, 45, 431–444.

Ishii T, Karimi-Ashtiyani R, Houben A. (2016). Haploidization via chromosome elimination: Means and mechanisms. *Annual Review* of *Plant Biology*, 67, 1–18.

Jones RN. (1995). B chromosomes in plants. Tansley Review No. 85. *New Phytologist*, 131, 411–434.

Kayano H. (1971. Accumulation of B chromosomes in the germ line of *Locusta migratoria*. *Heredity*, 27, 119–123.

Kloc M, Zagrodzinska B. (2001). Chromatin elimination – An oddity or a common mechanism in differentiation and development? *Differentiation*, 68, 84–91.

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

Longwell AC, Yerganian G. (1965). Some observations on nuclear budding and nuclear extrusions in a chinese hamster cell culture. *Journal of the National Cancer Institute*, 34, 53–69.

López-León M, Cabrero J, Camacho JPM, Cano M, Santos JL. (1992). A widespread B chromosome polymorphism maintained without apparent drive. *Evolution*, 46, 529–539.

Loray MA, Remis MI, Vilardi JC. (1991). Parallel polymorphisms for supernumerary heterochromatin in *Dichroplus elongatus* (Orthoptera): Effects on recombination and fertility. *Genetica*, 84, 155–163.

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

Milani D, Palacios-Gimenez OM, Cabral-de-Mello DC. (2017). The U2 snDNA is a useful marker for B chromosome detection and frequency estimation in the grasshopper *Abracris flavolineata*. *Cytogenetic and Genome Research*, 151(1), 36–40.

Morais-Cecílio L, Delgado M, Jones RN, Viegas W. (1996). Painting rye B chromosomes in wheat: Interphase chromatin organization, nuclear disposition and association in plants with two, three or four Bs. *Chromosome Research*, 4, 195–200.

Nur U. (1963). A mitotically unstable supernumerary chromosome with an accumulation mechanism in a grasshopper. *Chromosoma*, 14, 407–422.

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

Pardo MC, López-León MD, Cabrero J, Camacho JPM. (1994). Transmission analysis of mitotically unstable B chromosomes in *Locusta migratoria*. *Genome*, 37, 1027–34.

Pearse FK, Ehrlich PR. (1979). B chromosome variation in *Euphydryas colon* (Lepidoptera, Nymphalidae). *Chromosoma*, 73, 263–274.

Phillips DM. (1970). Insect sperm: their structure and morphogenesis. *Journal of Cell Biology,* 44, 243 –277.

Pigozzi MI, Solari AJ. (2005). The germ-line-restricted chromosome in the zebra finch: recombination in females and elimination in males. *Chromosoma*, 114, 403–409.

Randolph LF. (1941). Genetic characteristics of the B chromosomes in maize. *Genetics*, 26,

608–631.

Ruiz-Ruano FJ, Cabrero J, López-León MD, Camacho JPM. (2017). Satellite DNA content illuminates the ancestry of a supernumerary (B) chromosome. *Chromosoma,* 126(4), 487–500.

Schoenmakers S, Wassenaar E, Laven JSE, Grootegoed JA, Baarends WM. (2010). Meiotic silencing and fragmentation of the male germline restricted chromosome in zebra finch. *Chromosoma*, 119, 311–324.

Schubert I, Oud JL. (1997). There is an upper limit of chromosome size for normal development of an organism. *Cell*, 88, 515–520.

Shimizu N, Itoh N, Utiyama H, Wahl GM. (1998). Selective entrapment of extrachromosomally amplified DNA by nuclear budding and micronucleation during S phase. *Journal of Cell Biology*, 140, 1307–1320.

Staiber W. (1988). G-banding of germ line limited chromosomes in *Acricotopus lucidus* (Diptera, Chironomidae). *Chromosoma*, 97, 231–234.

Suja JA, García de la Vega C, Rufas JS. (1989). Mechanisms promoting the appearance of abnormal spermatids in B-carrier individuals of *Eyprepocnemis plorans* (Orthoptera). *Genome*, 32, 64–71.

Swim MM, Kaeding KE, Ferree PM. (2012). Impact of a selfish B chromosome on chromatin dynamics and nuclear organization in *Nasonia*. *Journal of Cell Science*, 125, 5241–5249.

Szöllösi A. (1975). Electron microscope study of spermiogenesis in *Locusta migratoria* (Insect Orthoptera). *Journal of Ultrastructure Research*, 50, 322–346.

Taverna SD, Coyne RS, Allis CD. (2002). Methylation of histone H3 at lysine 9 targets programmed DNA elimination in *Tetrahymena*. *Cell*, 110, 701–711.

Teruel M, Cabrero J, Perfectti F, Alché JD, Camacho JPM. (2009). Abnormal spermatid formation in the presence of the parasitic B24 chromosome in the grasshopper *Eyprepocnemis plorans*. *Sexual Development*, 3, 284–289.

Turner JMA. (2007). Meiotic sex chromosome inactivation. *Development*, 134, 1823–1831.

Utani K, Okamoto A, Shimizu N. (2011). Generation of micronuclei during interphase by coupling between cytoplasmic membrane blebbing and nuclear budding. *PLoS ONE*, 6(11), e27233.

Vaskova EA, Pavlova SV, Shevchenko AI, Zakian SM. (2010). Meiotic inactivation of sex chromosomes in mammals. *Russian Journal of Genetics*, 46, 385–393.

Viseras E, Camacho JPM. (1985). The B-chromosome system of *Omocestus bolivari:* changes in B-behaviour in M_4-polysomic B-males. *Heredity*, 54, 385–390.

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

Wang J, Davis RE. (2014). Programmed DNA elimination in multicellular organisms. *Current Opinion in Genetics & Development*, 27, 26–34.

Williams E, Barclay PC. (1972). Transmission of B-chromosomes in *Dactylis*. *New Zealand Journal of Botany*, 10, 573–584.

Zurita S, Cabrero J, López-León M, Camacho JPM. (1998). Polymorphism regeneration for a neutralized selfish B chromosome. *Evolution*, 52, 274–277.

5. Transcriptional changes between sexes, tissues and ontogenetic stages associated with the presence of a B chromosome in the grasshopper *Eyprepocnemis plorans*

Abstract: B chromosomes are supernumerary and dispensable elements that are found in the genome of many eukaryotes and often behave like genomic parasites. However, the true nature of the intragenomic conflict caused by them is still poorly known. In this chapter we address this problem by means of a comprehensive RNA-seq experiment, comparing transcriptomes of B-carrying and B-lacking male and female embryos and adults of the grasshopper *E. plorans*. The results showed that B chromosome presence is associated with gene expression changes for a number of A chromosome genes, in addition to the changes inherent to the protein-coding genes present in the B chromosome. The higher numbers of differentially expressed genes (DEGs) due to B presence were found in gonads (ovary= 643 DEGs, testis= 306 DEGs), followed by embryos (female= 95 DEGs, male= 130 DEGs) and finally in hind legs (female= 81 DEGs, male= 36 DEGs). Furthermore, while B-carrying embryos showed a pronounced up-regulation of transposable elements (~38% of all DEGs), gene expression changes in B-carrying adults mostly involved protein-coding genes (~50%, considering both up- and down-regulations), being most apparent in the ovary due to ~36% of protein-coding DEGs showing down-regulation, a figure being only ~15% in testis. Furthermore, we found a slight increase of DEGs in embryos that received a maternal B in respect to embryos which inherited a paternal B chromosome. These findings outline a scene in which the B chromosome in testis, and in embryos derived from a male with B chromosomes, triggers fewer changes in gene expression than in ovaries and embryos derived from a maternal B. The meaning of the higher transcriptional effects of B chromosome presence on the ovary is discussed in the light of the arms race between A and B chromosomes. As B chromosomes in *E. plorans* show their transmission advantage (i.e. drive) during oogenesis, this process is thus the best focus to neutralize. Our results suggest that resistance to parasitic B chromosomes could be mediated by gene expression changes.

Keywords: *DEGs*, *host genome, B chromosome, embryos, gonads*

Introduction

B chromosomes constitute a peculiar piece of the genome in many eukaryote organisms, being unnecessary and usually harmful to the carrier individuals. Despite this, they often achieve high population frequencies because of prominent systems of profitable transmission (drive). Up to 2005, the DNA known to be contained in B chromosomes was an extensive battery of repetitive elements, including satellite and microsatellite DNA, ribosomal DNA (rDNA) and transposable elements (see Camacho, 2005 for review). In that context, B chromosomes were thought to be mostly heterochromatic and essentially inert as they rarely trigger phenotypic effects on carrier individuals. Their genetic inertness was also supported by the scarcity of tritiated uridine observed on B chromosomes of the grasshopper *Myrmeleotettix maculatus* (Fox et al., 1974) and the mouse *Apodemus peninsulae* (Ishak et al., 1991).

The first indirect evidence of the presence of a single-copy gene in a B chromosome was PDA1, a gene that, in the pea-infesting fungus *Nectria haematococca*, encodes a particular cytochrome P450 protein that protects against pisatin, an anti-microbial delivered by the pea plant (Miao et al., 1991). In recent times, a variety of protein-coding genes have been found in B chromosomes from several species, such as the proto-oncogene C-unit in the red fox (*Vulpes vulpes*) and the raccoon canine (*Nyctereutes procyonoides*) (Graphodatsky et al., 2005), or the H3 and H4 histone genes in the B chromosomes of the grasshopper *Locusta migratoria* (Teruel et al., 2010). Furthermore, five distinctive protein-coding genes have been found in the B chromosomes of the cyclid fish *Lithocromis rubripinnis* (Yoshida et al., 2011), three protein-coding genes were also found in the Siberian roe deer (*Capreolus pygargus*) (Trifonov et al., 2013) and the vaccinia-related kinase gene (*vrk1*) located in the B chromosomes of two species of the genus *Apodemus* have been recently reported by Makunin et al. (2018).

The finding of protein-coding genes located on B chromosomes cast doubts on the inertness of these genetic elements. In that sense, much research was endeavored to evidence gene expression related to B chromosome presence. One of the first indirect proofs of B chromosomes transcriptomic activity was obtained in the frog *Leiopelma hochstetteri* (Green, 1988) and the mosquito *Simulium juxtacrenobium* (Brockhouse et al., 1989). Likewise, some B chromosomes have been appeared to contain ribosomal DNA (rDNA) having the capacity to sort out a nucleolus, in this way being functional (Bidau et

al., 2004; Teruel et al., 2007). However, rRNA transcripts particularly originating from B chromosomes have been just described in the plant *Crepis capillaris* (Leach et al., 2005), the parasitic wasp *Tricogramma kaykai* (van Vugt et al., 2003) and the grasshopper *Eyprepocnemis plorans* (Ruiz-Estévez et al., 2012). Furthermore, Carchilan et al. (2009) found, in rye, the transcription of a B-specific DNA sequence made up of high-copy families similar to mobile elements, and Zhou et al. (2012) found a B-linked scaffold partially transcribed. Trivonov et al. (2013) demonstrated the transcription of a B-specific protein-coding sequence in fibroblast culture of the Siberian roe deer (*Capreolus pygargus*) and Banaei-Moghaddam et al. (2013) reported, by NGS, that about 15% of the pseudogene-like sequences on B chromosomes are transcribed in rye. Recently, a functional B-derived Argonaute-like protein has been found also in rye (Ma et al., 2017).

These outcomes point out that certain fractions of the DNA contained in B chromosomes are likely transcriptionally active but the transcriptomic crosstalk between B chromosomes and the host genome has very rarely been examined. As a consequence, it is not clear the role of this B chromosome transcriptomic activity for the coevolutionary dynamic between the host genome and these selfish genetics elements. The new technologies of massive sequencing could be the way to meet these issues and unveil the underlying intragenomic conflict caused by B chromosomes presence. In that direction, Akbari et al. (2013) reported the differences in transcriptomic profiles of wild type and paternal sex ratio (PSR) chromosome carrying males. The PSR is a particular kind of B chromosome, found in the wasp *Nasonia vitripennis*, whose presence causes the conversion of diploid zygotes (destined to be females) to haploid males (Werren, 1991).

After 40 years studying the B chromosome system in the grasshopper *Eyprepocnemis plorans* at cytogenetic, population and molecular levels, we have obtained a high degree of knowledge about the evolutionary biology of this biological system, including the development of the near-neutral model (Camacho et al., 1997), which has made *E. plorans* a reference species in this field. B chromosomes of this species are mitotically stable and do not trigger apparent effects on the external phenotype (Camacho et al., 1980), but in high numbers they decrease the fertility of females (Zurita et al., 1998). With respect to gene expression of *E. plorans* B chromosomes, very little is still known. The B chromosome variant named B24 harbors 45S ribosomal RNA genes which are transcribed at low rates but are able to build nucleoli attached to the Bs, suggesting that the rRNA

genes of these B chromosome are actually functional (Ruiz-Estévez et al., 2013, 2014). Our latest genomic and transcriptomic experiments have given clear confirmation that *E. plorans* B chromosomes contain no less than ten protein-coding genes and that at least two of them are actively transcribed (Navarro-Domínguez et al., 2017).

Recent transcriptomic analyses in adult females (bodies and ovaries) have revealed the existence of changes in the expression of many genes related to previously known effects of B chromosomes, such as nucleotypic effects, stress response and detoxification, protein modification and turnover, ovarian function and gene expression regulation (Navarro-Domínguez et al., 2019). Those changes could be behind the meiotic drive, the fertility changes associated to B chromosomes (Zurita et al., 1998) or even the B chromosome expulsion during spermiogenesis (Cabrero et al., 2017). However, changes in gene expression associated with the presence of B chromosomes could be manifested, even to a greater extent, in the early ontogenic stages, where the host genome face the presence of this supernumerary DNA for the first time. The general objective of this study is, therefore, to reveal the molecular details of the adaptation of the genome of *E. plorans* to the presence of the B chromosomes at the level of the gene expression in different sexes, tissues and ontogeny stages. With this aim, we have analyzed transcriptomes of embryos and adults (legs and gonads) in both sexes, with and without B chromosomes. In particular, we shed light in mechanisms involved in the adaptation of the host genome to B chromosome presence in regard to the following questions: I) Is the presence of B chromosomes associated with changes in gene expression in *E. plorans* embryos and adults? II) Are these changes in gene expression similar between sexes? III) Is there a response from the A chromosomes to the presence of Bs in *E. plorans*?

Materials and methods

Sampling and RNA sequencing

Embryos

In order to investigate the net effect of B chromosome presence, we performed the current experiment on sibling embryos obtained from controlled crosses in the laboratory. This reduced interindividual variance due to genetic causes not related with B chromosome presence. For this purpose, we collected adult males and females of *E. plorans* at the Torrox population (Málaga, Spain) in 2014 and analyzed them *in vivo* to

ascertain the number of B chromosomes they harbored. This was done in males by extracting several testis follicles through a small cut in the abdomen and cytologically analyzing primary spermatocytes at diplotene or metaphase I to score the number of B chromosomes. In females, we performed it on hemocytes extracted from the abdomen, as described in Cabrero et al. (2006). We then started several controlled crosses between a 1B individual and a 0B one, and incubated the egg-pods laid by females to obtain ten-day-old embryos.

Embryos from these controlled crosses were analyzed through an approach devised by our group to get three samples from a same embryo, which might allow chromosome, genomic and transcriptomic analyses. Each embryo was thus dissected from the egg after ten days of incubation, a stage which allows analyzing them cytologically to ascertain sex and their number of B chromosomes, and molecularly to perform genomic and transcriptomic analyses on the DNA and RNA obtained from the two other embryo parts. The main steps of this protocol are indicated in the Materials and methods section of this thesis.

This protocol allowed us to classify the RNA obtained from each embryo within one of four categories: 0B males, 1B males, 0B females and 1B females, and performing an RNA-seq experiment with three biological replicates per category. In order to get enough RNA for Illumina sequencing, each biological replicate was obtained by mixing the cell suspensions from two embryos belonging to the same category, thus buffering also between-individual variability.

For RNA isolation from the cell suspensions of embryos, that were stored immersed in Quiazol, the RNeasy® Lipid Tissue Mini Kit (Qiagen, Hilden, Germany) was used, and we subjected the samples to Amplification Grade DNase I (Sigma) on the column of the extraction kit (20 units of DNAse, 30 minutes incubation). Concentration and quality per biological replicate was checked in MOPS agarose gel (1.5%), in the NanoQuant Infinite 200 PRO Tecan and the Agilent 2100.

We then sequenced the 12 RNA samples obtained (two B categories x two sexes x three replicates) from each of the two following egg-pods: one of them from a cross between a 1B female and a 0B male (P1 embryos from now on) and the other from the reverse cross, i.e. a 1B male with a 0B female (P2 embryos).

RNA from P1 embryos was sequenced in two lanes of Illumina HiSeq 2000 (Beijing

Genomics Institute, BGI, China) which, after removal of adapters and low-quality sequences with Trimmomatic (Bolger et al., 2014), yielded about ~65.5 Gb of 100 bp paired-end reads. The 12 remaining samples from P2 embryos were sequenced in one lane of the Illumina HiSeq 4000 platform also by BGI yielded about ~80 Gb of 150 bp paired-end reads after removal of adapters and low-quality sequences.

Adults

We captured 14 individuals of *E. plorans* in the population of Torrox (Málaga), 7 males and 7 females that were cytogenetically analyzed to characterize each one by the number of B chromosomes they carried. Males were characterized by studying a few testicular follicles (López-León, 1992a) while females were studied by C-banding of hemolymph (Cabrero et al., 2006), find details about the protocol used in the Materials and methods section of this thesis. Once the 14 individuals were characterized, all possible controlled crosses were established between 0B and 1B individuals in order to achieve sufficient 0B and 1B offspring of each of them. In this way we were able to prepare 3 crosses: C1 (F1BxM0B), C2 (F0BxM1B) and C3 (F0BxM1B). According to our previous inheritance analysis of this B chromosome system (López-León et al., 1992b; Zurita et al., 1998), only 0B and 1B offspring was expected from all three types of cross. Pods were incubated in an wet chamber at 28 ºC until hatching, at which point they were transferred to three wooden boxes (one for each cross) for their development into adults.

The only cross that produced enough individuals reaching adult stages that could be classified in three 0B males, three 0B females, three 1B males and three 1B males was the C3 cross. After 5 to 7 days after the last molting, each individual was studied cytogenetically by means of hemolymph banding and at the same time fixed in liquid nitrogen after extraction of their gonad (also fixed in liquid nitrogen) and kept at -80 ºC. This protocol for sample preparation included:

1) Dissection of the individual through a vertical cut along the abdomen followed by a hemolymph smear on a slide to characterize cytogenetically the presence or absence of B chromosomes in the adult by C-banding of interphase nuclei.

2) Gonad extraction and fixation in liquid nitrogen.

3) Fixation of the body in liquid nitrogen.

4) Conservation of body and gonad at -80 ºC.

This protocol should not last more than two minutes for each individual in order to

maintain RNA integrity and minimize possible changes in gene expression due to individuals manipulation.

As indicated above, once all the individuals from C3 had been cytologically analyzed, we chose three 0B males, three 0B females, three 1B males and three 1B males to perform hind leg and gonad RNA extraction and sequencing.

RNA was extracted from each fixed leg using the Real Total RNA Spin Plus kit (Durviz), while for the extraction of total RNA from the gonads, the RNeasy® Lipid Tissue Mini Kit (Quiagen) was used, in both cases we treated the samples with DNase I Amplification Grade (Sigma). The integrity and concentration of the extracted RNA was verified by 2% agarose gel electrophoresis and analysis on the Nanodrop.

RNA from the 24 gonad and leg samples was sent to BGI (Beijing Genomics Institute, China) for sequencing through Illumina HiSeq 4000 which yielded ~156 Gb (~6-7 Gb/library) of 150 bp paired-end reads after quality filtering using Trimmomatic (Bolger et al., 2014).

De novo transcriptome assembly and annotation

We generated a *de novo* transcriptome to be used as reference in the detection of differentially expressed genes (DEGs) in embryos of *E. plorans*. We included all sequenced embryo libraries involved in this study (0B and 1B males and females from P1 and P2) in order to get the most complete reference transcriptome as possible. To identify DEGs in adults and compare them with those found in embryos, we also assembled a *de novo* transcriptome using all libraries from adults and those from P2 embryos. We chose P2 embryos instead of P1 ones in order to avoid effects due to the parental B-content, as adults and P2 embryos both came from a cross between a 1B male and a 0B female, whereas P1 embryos derived from the reciprocal cross with the female parental being the B-carrier.

We used for the novo assembly of transcriptomes the Trinity software v2.5.1 (Haas et al., 2013). One of the main challenges when building a reference transcriptome is the assembling of low frequency reads. To alleviate this problem, we used the *in silico* normalization algorithm implemented in the Trinity suite, which selects reads on the basis of median kmer coverage and user-defined maximum coverage. We used a maximum limit of 50x coverage, as default option, which has been successfully applied in yeast and mouse (Haas et al., 2013) and reduces computing time. Trinity software was

run using default parameters for assembling of the above indicated libraries. It resulted in a FASTA file including the different assembled isoforms for every gene that was used as *de novo* reference transcriptome for subsequent analysis.

Both *de novo* transcriptomes were annotated following the Trinotate (release 2.0.2) annotation suite (https://trinotate.github.io), especially indicated for functional annotation of transcriptomes assembled using the Trinity software. Protein sequences were predicted from potential ORFs using the TransDecoder package (http://transdecoder.sourceforge.net/). Sequence homology search was performed with BLASTX of the transcripts and BLASTP of the predicted proteins against Swiss-Prot (Boeckmann et al., 2003) using default settings. In addition, protein domains were analyzed with HMMER and PFAM (Finn et al., 2011, 2016). Trinotate works with a boilerplate SQLite database to analyze the results of the searches described above. We further annotated DEGs using Blast2GO (Götz et al., 2018).

Differential expression analysis

To estimate read abundance for each contig in each sample (which is expected to be proportional to the expression level of that contig), the paired Illumina reads (prior to *in silico* normalization) were mapped against the *de novo* reference transcriptome using Bowtie (Langmead et al., 2009), in order to obtain the gap-free alignment required by RSEM (Li and Dewey, 2011) to estimate read abundance per gene or per isoform. The RSEM computation generates two files: one containing information on abundance (read counts) per Trinity transcript (isoforms-level) and the other containing the read counts for the sum of all the isoforms of a given gene (gene-level). We performed DE analysis at gene-level as it has been shown to be more accurate in terms of abundance estimates (Yi et al., 2018; Soneson et al., 2016) while isoform-level differences are still hard to interpret biologically (Dapas et al., 2017) specially for a non-model organisms such as *E. plorans* with no reference genome/annotation available and including a high complex set of isoforms that are expected to came from the B chromosomes of the species.

Differential expression analysis was performed with the edgeR package of the suite Bioconductor (Robinson et al., 2010) in R 3.6.1 (R Core Team, 2019). First, we filtered the RSEM count matrix to remove low-expressed transcripts that could bias subsequent analysis. For this purpose, we set a threshold of 15 counts in median per gene filtered with CPM (count per million), thus accounting for difference in library sizes as indicated

by Chen et al. (2016, 2020) (see cpm histograms and boxplots in Figure 5.1 and 5.2). In addition, we plotted correlation between samples in a heatmap in terms of gene counts. Then, we also estimated data dispersion and the biological coefficient of variation of each group of samples after filtering (P1 and P2 embryos, legs and gonads; see Figure S5.3). Prior to DE analysis we normalized the RSEM data of reads counts according to the TMM method described in Robinson and Oshlack (2010) for subsequent analysis. Then we applied a glm method with likelihood ratio test (glmFit and glmLRT functions) to identify DE genes between samples, as recommended when analyzing multiple factors (i.e. B-content, sex, embryo pod and tissues). To control for the rate of type I errors, the false discovery rate (FDR) method for multiple comparisons was applied to the p-values, and only those genes whose FDR remained below 0.05, and a fold-change (FD) above 1 between samples, were confidently considered as DEGs. Finally, we built MA and Volcano plots for all comparison here included using ggplot2 in the R 3.6.1 software.

We carried out this protocol for P1 and P2 embryos and also for P2 embryos plus adults separately. In this way, we performed two DE analysis: one of them to study embryos, thus including the *de novo* reference transcriptome assembled using all embryo libraries, and the other for P2 embryos and adults, including the corresponding *de novo* reference assembly.

Functional analysis using Gene Ontology (GO)

Trinotate suite provides a report where every sequence of the transcriptome is described per homology results and is associated with a Gene Ontology function (Ashburner et al., 2000). Gene ontology (GO) enrichment analysis was performed using the script run_Goseq.pl, included in the Trinity suite, which makes use of the R/Bioconductor package GOSEQ (Young et al., 2010). This analysis was performed using GO categories and gene length data extracted previously from the Trinotate and Blast2GO annotation, and the Trinity assembly.

In particular, we focused on common DEGs in embryos associated with a B chromosome and with those that were specific from ovaries and testis in the case of adults. The reason for this was the low number of DEGs shared between ovaries and leg in presence of a B chromosomes and the high number of those between embryos from both pods (P1 and P2) and sexes. We considered as background in GO enrichment analysis all genes detected in the different samples.

Then, ReviGO was used for visualization of GO terms assigned to DE transcripts (Supek et al., 2011). This software removes the redundant terms, calculates and summarizes the list of GO terms according to the enrichment in the cellular component, biological process, and molecular function and helps visualization of the remaining GO terms based on their semantic similarities in scatterplots. Finally, we plotted the p-values of GO terms using a custom script in R.3.6.1.

Statistical analysis

We performed a series of Chi-square calculations by Goodness of fit for within-experiments comparisons between number of DEGs. In the case of tests between tissues and developmental stages we carried out contingency analysis applying Yates correction when the number of DEGs in some of the observed cases was less than 5. Analysis was performed in R.3.6.1 and using spreadsheets.

Results

Transcriptional changes associated with the B chromosome presence in embryos

The *de novo* reference transcriptome for embryo analysis was built with all libraries from P1 and P2 embryos belonging to both sexes and having/lacking B chromosomes including three biological replicates for each group. We finally included 23 embryo libraries because one of the 0B-male libraries from P1 was discarded due to sequencing failure. The RNA concentration of that 0B male replicate was too low for normal library construction, for which reason BGI staff assayed a special kit on only 10 ng RNA. However, this yielded only 0.82 Gb, in high contrast to the ~6 Gb, on average, yielded by each of the remaining libraries. The *de novo* embryo reference transcriptome contained 1,127,894 transcripts belonging to 661,919 genes, the N50 length was 916 bp and the GC content was 40.46%.

Before the DE analysis of *E. plorans* embryos, we clustered in a heatmap all samples from P1 and P2 embryos separately in terms of CPM for each Trinity gene (Fig S5.4). At first glance, we noticed an incipient group of 1B samples in embryo samples from P1 that was not evident in the case of P2 embryos. This clustering pattern suggests a stronger effect of B chromosome presence in P1 compared to P2 embryos.

EdgeR analysis of differential expression for P1 and P2 embryo samples yielded correlation matrices grouping the replicates belonging to the four classes (0B females, 0B males, 1B females and 1B males) on the basis of their gene-expression similarity, with B-presence being a better classification factor than sex. This suggested that B-presence elicits more gene expression changes on ten-day old embryos than sex differences at this stage (Fig. 5.1a and 5.2a). This was more evident in P1 embryos which showed almost three times more DEGs associated to B-presence than P2 ones (Table 5.1, S5.4) and in which the division between 0B females and males was not so evident as for P2 ones.

We separately analyzed DEGs classified into four categories: i) protein-coding genes having shown to be B-linked (see Chapter 3), ii) protein-coding genes including remaining annotated genes (presumably located only in A chromosomes), iii) TEs or undetermined proteins and iv) non-annotated contigs.

When we looked at the number of DEGs (FDR<0.05 and logFC>|1|) in P1 embryos (Table 5.1), the presence of B chromosomes was associated with a high number of DEGs in both sexes, 254 in the F1B/F0B comparison and 229 in the M1B/M0B one. In high contrast, sex differences were extremely low in absence of B chromosomes (only 7) whereas in B-carrying embryos it was high (207 DEGs). Therefore, B presence triggers more than 200 DEGs in both sexes whereas sex implied only seven DEGs in absence of B chromosomes (see Fig. 5.1).

In P2 embryos, where the B chromosome was transmitted by the male parent, B presence also elicited most gene expression changes, but with higher intensity in male (106 DEGs) than female (67) embryos (see Fig. 5.2c). Likewise, sex differences were much higher in 1B embryos (87 DEGs) than in 0B ones (25 DEGs).

On the contrary, there were more DEGs between B-lacking females and males in embryos from P2 than in P1 although DEGs between sexes in presence of 1B chromosomes were again higher for P1 than P2, suggesting a stronger effects of the B chromosome on sex differences in the former pod. Again, all these difference were due to the effect of non-annotated and TEs DEGs.

Table 5.1. Comparison of the numbers of DEGs (obs.: observed, exp.: expected) found in P1 and P2 embryos in presence of a B chromosome and between sexes, performed also separately for each group of gene annotation (non-annotated, TEs, protein-coding genes from A chromosomes and B-genes). *First test that appears in each group is for totals in rows, second one for totals in columns.

Annotation	Sample comparisons	DEGs obs.			DEGs exp.		Goodness of fit*		Contingency	
		P1	P2	Total	P1	P2	chi (1:1)	P	chi	P
All DEGs	F1B/F0B	254	67	321	236.35	84.65	0.3	5.85E-01	9.8	**0.0018**
	M1B/M0B	229	106	335	246.65	88.35				
	Total	483	173	656			146.5	**1.01E-33**		
	F0B/M0B	7	25	32	21.01	10.99	210.6	**1.03E-47**	30.1	**4.02E-08**
	F1B/M1B	207	87	294	192.99	101.01				
	Total	214	112	326			31.9	**1.61E-08**		
Non-annotated	F1B/F0B	109	16	125	101.59	23.41	0.0	9.50E-01	5.7	**0.0165**
	M1B/M0B	95	31	126	102.41	23.59				
	Total	204	47	251			98.2	**3.78E-23**		
	F0B/M0B	1	7	8	5.56	2.44	120.7	**4.46E-28**	10.3	**1.34E-03**
	F1B/M1B	104	39	143	99.44	43.56				
	Total	105	46	151			23.1	**1.58E-06**		
TEs	F1B/F0B	99	32	131	91.89	39.11	0.9	3.37E-01	3.5	0.0619
	M1B/M0B	96	51	147	103.11	43.89				
	Total	195	83	278			45.1	**1.85E-11**		
	F0B/M0B	1	8	9	6.73	2.27	62.4	**2.79E-15**	16.9	**3.87E-05**
	F1B/M1B	70	16	86	64.27	21.73				
	Total	266	107	373			23.3	**1.42E-06**		
Protein-coding genes (As)	F1B/F0B	38	16	54	35.49	18.51	0.1	7.70E-01	1.1	0.3010
	M1B/M0B	31	20	51	33.51	17.49				
	Total	69	36	105			10.4	**1.28E-03**		
	F0B/M0B	4	10	14	6.55	7.45	31.2	**2.35E-08**	2.3	0.132
	F1B/M1B	32	31	63	29.45	33.55				
	Total	36	41	77			0.3	5.69E-01		
B-genes	F1B/F0B	8	3	11	7.50	3.50	0.0	1.00E+00	0.0	1
	M1B/M0B	7	4	11	7.50	3.50				
	Total	15	7	22			2.9	8.81E-02		
	F0B/M0B	0	0	0	0.00	0.00	–	–	–	–
	F1B/M1B	1	1	2	1.00	1.00				
	Total	1	1	2			–	–		

Contingency tests on intra-sexual comparisons (i.e. F1B/F0B and M1B/M0B) revealed that DEG numbers in males and females differ between pods. In particular, non-annotated DEGs associated with the presence of one B chromosomes were similar between sexes in P1 while in P2 they were higher in the case of males than females (Table 5.1). Likewise, inter-sexual comparisons (i.e. F0B/M0B and F1B/M1B) showed differences in the number of DEGs in B-lacking and B-carrying comparisons between sexes for P1 and P2, and they were significant only for non-annotated and TE DEGs (Table 5.1). We found more DEGs in P1 than P2 between sexes when the B chromosome was present, however, in P2 there were more DEGs than in P1 were it was absent, suggesting stronger effects of B chromosome presence for sexes in P1 than in P2 embryos.

A Venn diagram showed that 186 DEGs were associated with B-presence in male and female embryos whereas 68 and 43 were exclusive for females and males, respectively (Fig. 5.1b). In P2 embryos, a Venn diagram showed that 51 DEGs were associated with B-presence in male and female embryos whereas 16 and 55 were exclusive for females and males, respectively (Fig. 5.2b). A contingency chi-square test showed that P1 and P2 patterns were highly significantly different (chi= 1836, df= 1, $p< 0.000001$) due to a higher number of DEGs in P1 and also to the presence of more female-biased DEGs in P1 and more male-biased ones in P2.

These results revealed that B-presence is associated with higher numbers of DEGs than sex in ten-day-old embryos of *E. plorans*, but with some remarkable differences between P1 and P2, since the latter egg-pod showed lower total number of DEGs, with predominance of male-specific DEGs over female-specific ones (55m:16f), whereas P1 showed higher total number of DEGs with slight predominance of female-specific ones (43m:68f). These results could be interpreted into under two non-excluding explanations: i) that some kind of imprinting on the B chromosome during gametogenesis determines its degree of genetic silencing during the first embryonic stages, in which case we could conclude that B chromosomes are more silenced during spermatogenesis, or ii) that there was some difference in developmental age between P1 and P2 embryos, in which case we would conclude that those showing higher number of DEGs (i.e. P1) were slightly older than P2 ones, bearing also in mind that the starting point (the zygote) is expected to show no DEGs whereas they were extremely higher in the most sexually differentiated organs analyzed (i.e. ovary and testis).

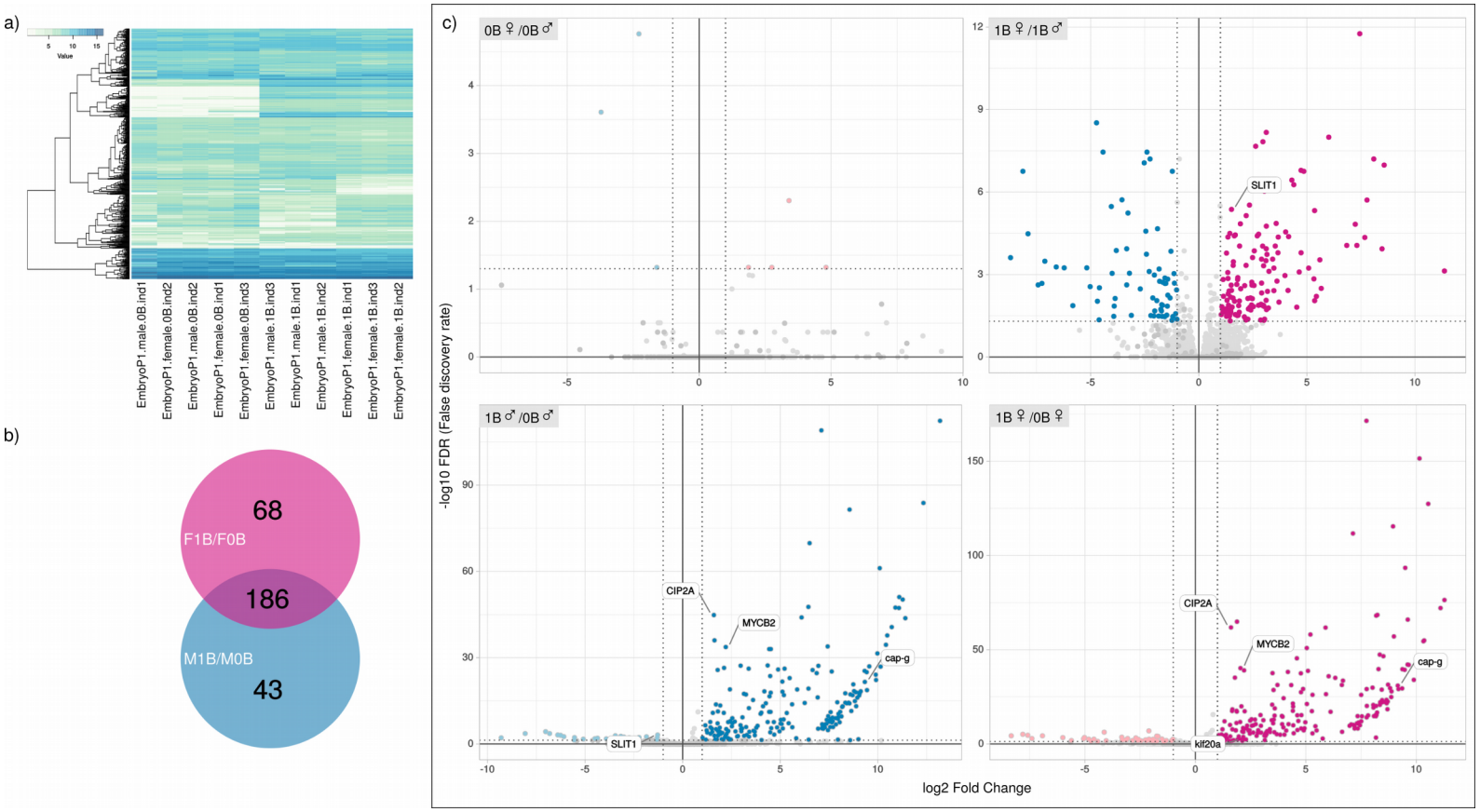


Figure 5.1. a) Heatmap of DEGs in P1 embryos of *E. plorans*, samples split in terms of B chromosome content. b) Venn diagram between DEGs associated with the presence of 1B chromosome in females and male P1 embryos, there is a high number of common DEGs between both sexes. c) Volcano plots for different comparisons in P1 embryos of *E. plorans*, note the impact of 1B chromosome in gene expression (down).

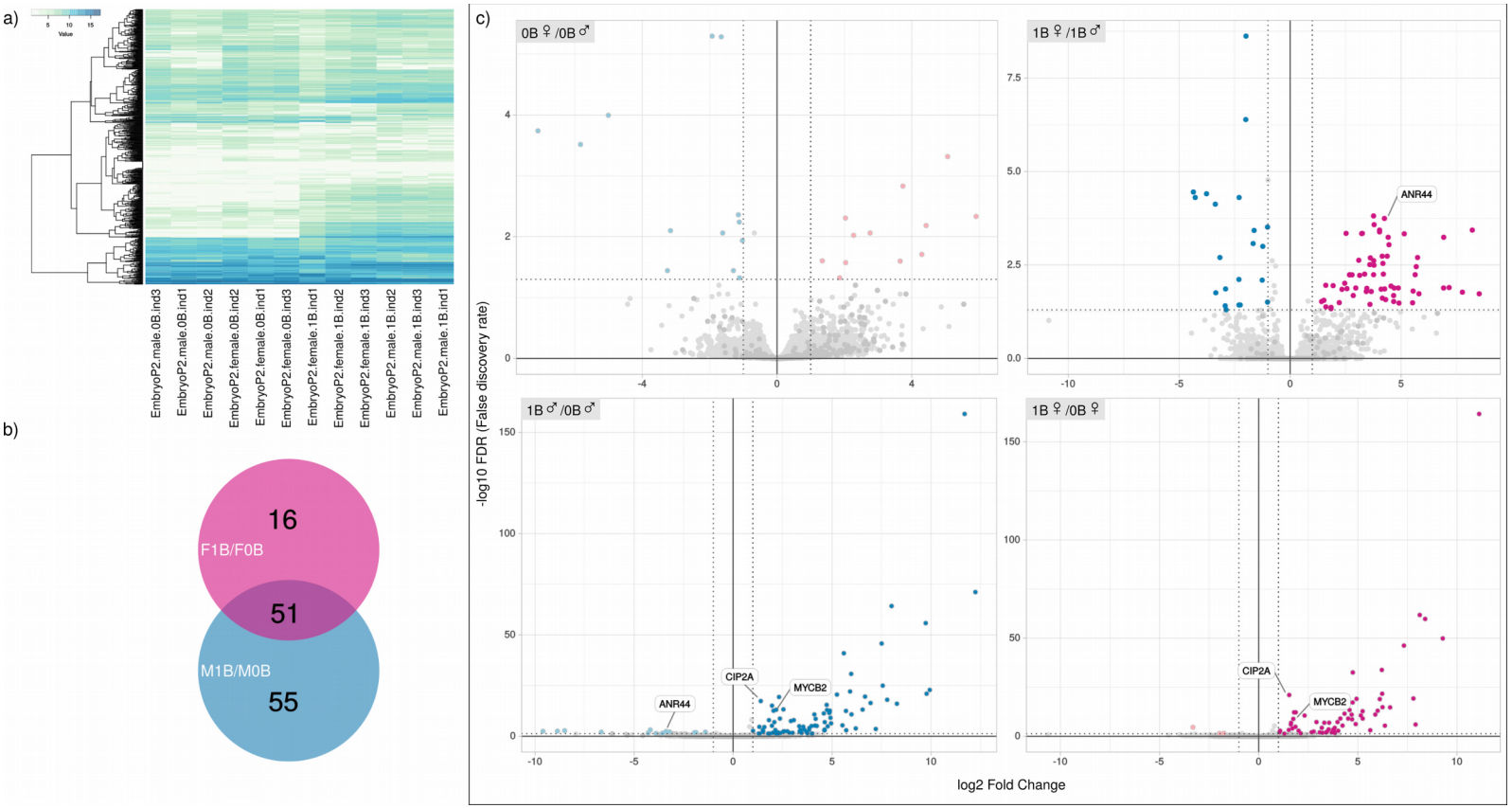


Figure 5.2. a) Heatmap of DEGs in P2 embryos of *E. plorans*, samples split in terms of B chromosome content, then by sexes. b) Venn diagram between DEGs associated with the presence of 1B chromosome in females and male P2 embryos, there is a high number of common DEGs between both sexes. c) Volcano plots for different comparisons in P2 embryos of *E. plorans*, note the impact of 1B chromosome in gene expression (down).

Transcriptional changes associated with the B chromosome presence in adults

For DE analysis in adults and their comparison with P2 embryos, we built a more inclusive *de novo* transcriptome with P2-embryo, leg and gonad libraries (including sex and B-presence as classifying factors), with three biological replicates per category. This transcriptome was thus built using a total of 48 libraries, and contained 1,390,496 transcripts associated with 846,154 trinity genes with an N50 of 823 bp and GC content of 41.16%.

The DE analysis of leg RNA libraries yielded 81 DEGs in 1B females compared to 0B ones whereas there were 36 DEGs for the same comparison in males (Fig. 5.3c). Therefore, in adults legs of *E. plorans* we detected a stronger effect of one B chromosome in females than in males (see Table 5.2). In addition, we found 198 DEGs between B-lacking females and males but only 7 DEGs when the B chromosome was present. Therefore, B chromosome presence reduces difference between sexes in hind legs. In gonads, however, the number of DEGs was much higher in both sexes, with 643 between 1B and 0B ovaries and 306 between 1B and 0B testis (Fig. 5.4c). In resemblance to the results in legs, there were more DEGs associated with B-presence in ovaries than in testis (Table 5.2). As expected for the high gonad complexity and the extreme physiological differences between ovary and testis, the number of DEGs associated to B-presence was notably lower than those found between ovaries and testis in both absence (14,258) or presence (13,793) of the B chromosome. This effect of sex in DEGs is also obvious in the heatmap of ovary and testis samples (Fig. 5.4a) in which libraries were first classified by sex and then by B-presence. In contrast, leg samples first clustered together in terms of B-content and then by sex (Fig. 5.3a).

Contingency tests confirmed that the number of intra-sex DEGs due to B-presence in female and males did not differ between legs and gonads since in both tissues we found more DEGs in female than in males associated with B chromosome presence (Table 5.2). On the contrary, the number of inter-sex DEGs was highly significantly different between leg and gonads for sexes comparison in presence/absence of B chromosomes, with about similar results in all types of DEGs (Table 5.2). These differences are mainly due to the low amount of DEGs between sexes in 1B individuals. This suggests that the presence a B chromosome balances gene expression between male and female legs, likely due to the response against B, which could be similar in both sexes.

Table 5.2. Comparison of the numbers of DEGs (obs.: observed, exp.: expected) found in legs and gonads in presence of a B chromosome and between sexes, performed also separately for each group of gene annotation (non-annotated, TEs, protein-coding genes from A chromosomes and B-genes). *First test that appears in each group is for totals in rows, second one for totals in columns.

Annotation	Sample comparisons	DEGs obs.			DEGs exp.		Goodness of fit*		Contingency	
		Leg	Gonad	Total	Leg	Gonad	chi (1:1)	P	chi	P
All DEGs	F1B/F0B	81	643	724	79.46	644.54	136.9	**1.27E-31**	0.1	7.47E-01
	M1B/M0B	36	306	342	37.54	304.46				
	Total	117	949	1066			649.4	**3.07E-143**		
	F0B/M0B	198	14258	14456	104.88	14351.12	15.2	**9.52E-05**	170.53	**5.68E-39**
	F1B/M1B	7	13793	13800	100.12	13699.88				
	Total	205	28051	28256			27441.9	**0.00E+00**		
Non-annotated	F1B/F0B	41	265	306	41.24	264.76	50.2	**1.37E-12**	0	9.44E-01
	M1B/M0B	21	133	154	20.76	133.24				
	Total	62	398	460			245.4	**2.58E-55**		
	F0B/M0B	69	6859	6928	37.9	6890.1	8.0	**4.67E-03**	52.6	**4.09E-13**
	F1B/M1B	5	6594	6599	36.1	6562.9				
	Total	74	13453	13527			13232.6	**0.00E+00**		
TEs	F1B/F0B	3	27	30	3.49	26.51	6.7	**9.53E-03**	0.00	9.90E-01
	M1B/M0B	2	11	13	1.51	11.49				
	Total	5	38	43			25.3	**4.84E-07**		
	F0B/M0B	8	440	448	4.11	443.89	0.6	4.36E-01	5.82	**1.59E-02**
	F1B/M1B	0	425	425	3.89	421.11				
	Total	8	865	873			841.3	**5.68E-185**		
Protein-coding genes (As)	F1B/F0B	31	336	367	27.8	339.2	80.4	**3.10E-19**	1.3	2.53E-01
	M1B/M0B	9	152	161	12.2	148.8				
	Total	40	488	528			380.1	**1.17E-84**		
	F0B/M0B	110	6906	7016	57.19	6958.81	6.2	**1.24E-02**	100.45	**1.21E-23**
	F1B/M1B	2	6721	6723	54.81	6668.19				
	Total	112	13627	13739			13294.7	**0.00E+00**		
B-genes	F1B/F0B	6	15	21	6	15	1.4	2.37E-01	0.15	7.03E-01
	M1B/M0B	4	10	14	4	10				
	Total	10	25	35			6.4	**1.12E-02**		
	F0B/M0B	11	53	64	6.02	57.98	1.0	3.09E-01	8.14	**4.33E-03**
	F1B/M1B	0	53	53	4.98	48.02				
	Total	11	106	117			77.1	**1.60E-18**		

Despite the high number of sexual differences in terms of DEGs between ovary and testis compared to that between female and male legs, changes in gene expression related to the presence of one B chromosome were higher in gonads than in legs (Table 5.2). This suggests an intense effect of B chromosome on gonads, especially on ovary, the tissue in which B chromosomes bet their maintenance through offspring. In fact, looking at DEGs annotation, we observe that the number of DEGs associated with the B chromosome presence was higher in gonads than in legs in all group of annotations except for B-genes. Number of B-genes differentially expressed show no significant differences between legs and gonads. This finding put into evidence that main differences between legs and gonads in presence of a B chromosome are more focused on the role of A chromosomes in the intragenomic battle than on the arms exhibited by the Bs.

Intra-sexual Venn diagrams in legs showed 15 DEGs common to males and females in presence of one B chromosome whereas 66 other were female-specific and 21 were male-specific (Fig. 5.3b). However, the number of common DEGs was less than twice higher in gonads (28) whereas female- and male-specific DEGs were one order of magnitude higher in gonads (615 and 278, respectively), as expected for the extensive physiological differences expected between ovary and testis Fig. 5.4b). These observed patterns in legs and gonads were significantly different ($chi= 32.68$, $df= 1$, $p< 0.000001$). When compared with the figures in the Venn diagram obtained in embryos, the observed patterns were also significantly different ($chi>400$ in P2-leg and P2-gonads comparisons, $p< 0.000001$ in both cases), indicating that gene expression changes associated to B presence change very much along the ontogeny.

Summing up, RNA-seq in adults showed that DEGs associated to B presence in gonads being about nine times more frequent than those found in legs. In addition, ovaries showed twice more DEGs associated with B presence than testes.

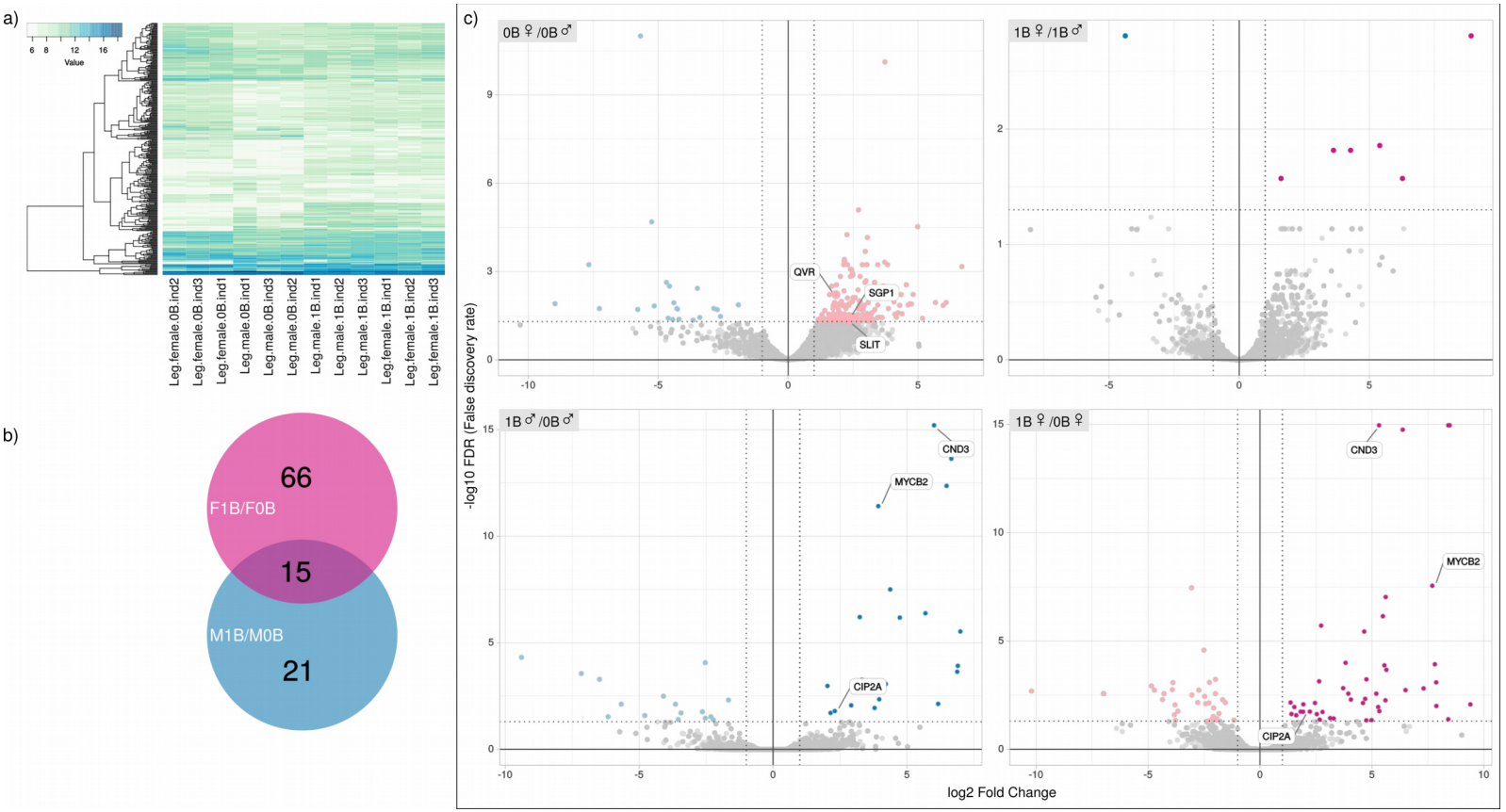


Figure 5.3. a) Heatmap of DEGs in adult legs of *E. plorans*, samples split in terms of B chromosome content, then by sexes. b) Venn diagram between DEGs associated with the presence of 1B chromosome in females and male legs, note the low number of common DEGs between both sexes. c) Volcano plots for different comparisons in legs of *E. plorans*, the effects of 1B chromosome in gene expression (down) are softer than in embryos and gonads.

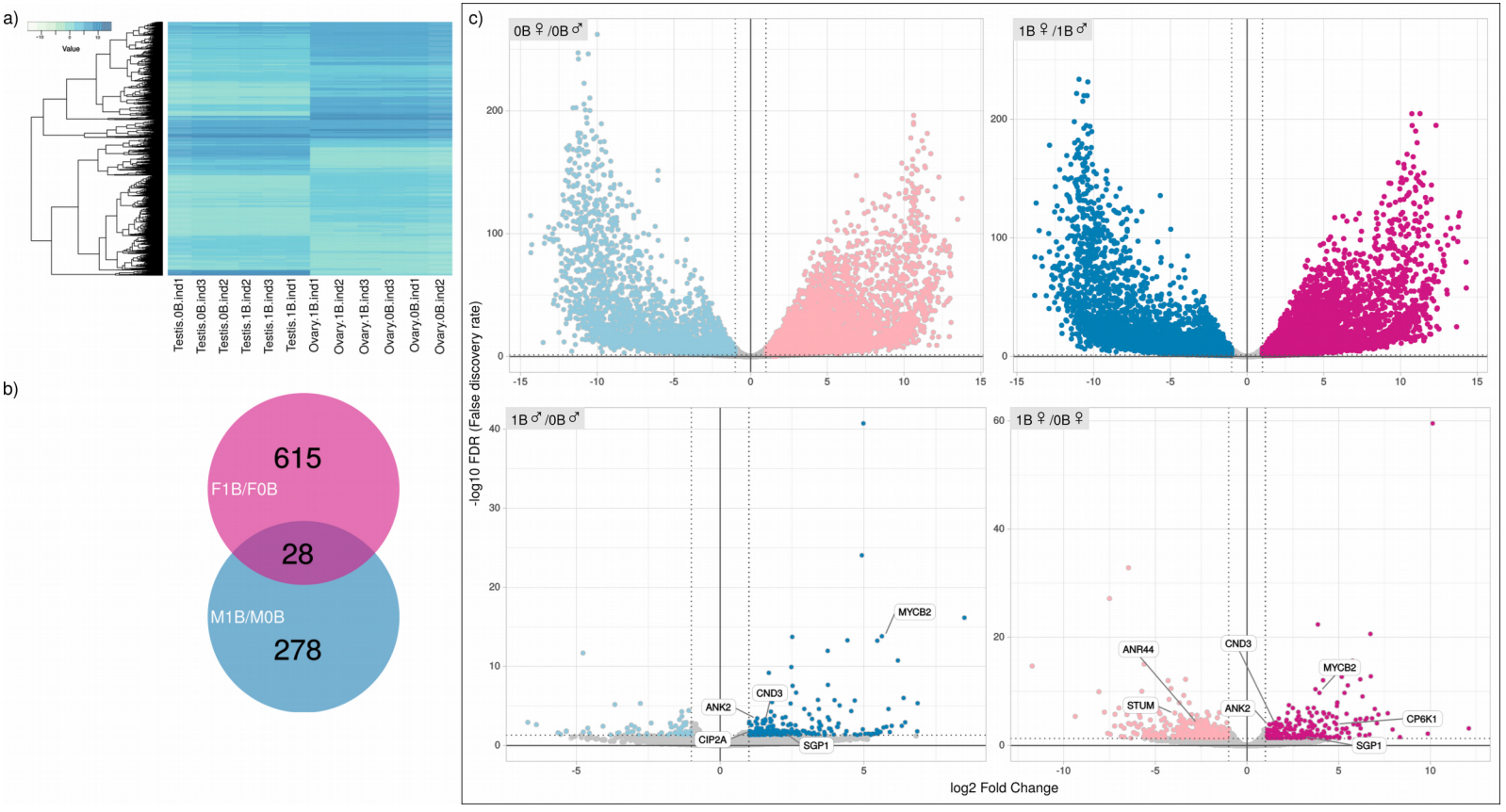


Figure 5.4. a) Heatmap of DEGs in gonads of *E. plorans*, samples split in terms of sexes first and then by B-content. b) Venn diagram between DEGs associated with the presence of 1B chromosome in ovary and testis, note the low number of common DEGs between both sexes. c) Volcano plots for different comparisons in gonads of *E. plorans*, the effects of 1B chromosome in gene expression (down) are stronger than in embryos and legs. See also huge difference between sexes (up) in absence of B chromosomes.

Transcriptional changes associated with the B chromosome presence during ontogeny

As explained above, we performed an additional DE analysis including P2 embryos and adults to explore differences between these two developmental stages. For this purpose, we plotted a heatmap including all samples and DEGs involved in this analysis. As seen in Figures 5.5a and b, P2 embryos and adult gonads grouped together and were well separated from adult legs which implies more resemblance in expression patterns between embryos and gonads than with legs. The next factor splitting samples was B-content for embryos and legs while it was sex in the case of gonads, thus resuming the same idea previously stated showing the high influence of sexes in the expression patterns found in gonads.

Concerning the number of DEGs yielded in this global analysis for P2 embryos, we found 95 DEGs between 1B and 0B females from P2 embryos. In the case of males, there were 130 DEGs between 1B-carrying and B-lacking males of *E. plorans* embryos. Therefore, we found again higher number of DEGs in males than in females. In addition, we found 38 and 90 DEGs between sexes in absence and presence of a B chromosome respectively (Table S5.3 and S5.6). These results were highly correlated (r_S= 0.8158, p< 0.00001) with those found in P2 embryos but considering the *de novo* transcriptome from embryos (for comparison between P1 and P2 embryos). Note that for comparisons of B chromosomes effects during ontogeny we include only P2 embryos that received the B chromosome from the male parent as it was in the case of adults.

Among all DEGs found in embryos, legs and gonads of males and females individuals harboring 1B chromosomes, we identified 9 genes differentially expressed in all embryos and legs of both sexes, 6 genes were so in legs and gonads of both sexes and there were 12 common DEGs in embryos and gonads. In addition, embryos and gonads shared 24 DEGs considering common genes at least between one sex, the same number of common DEGs were found between embryos and gonads and a lower amount (15 DEGs) in the case of embryos and legs (Fig 5.5c). Therefore, we found more similarities in terms of DEGs between embryos and gonads than with legs in the presence of one B chromosome. Remarkably, common DEGs genes between all samples were annotated as B-genes so it came across the activity of B-genes along all ontogenetic stages.

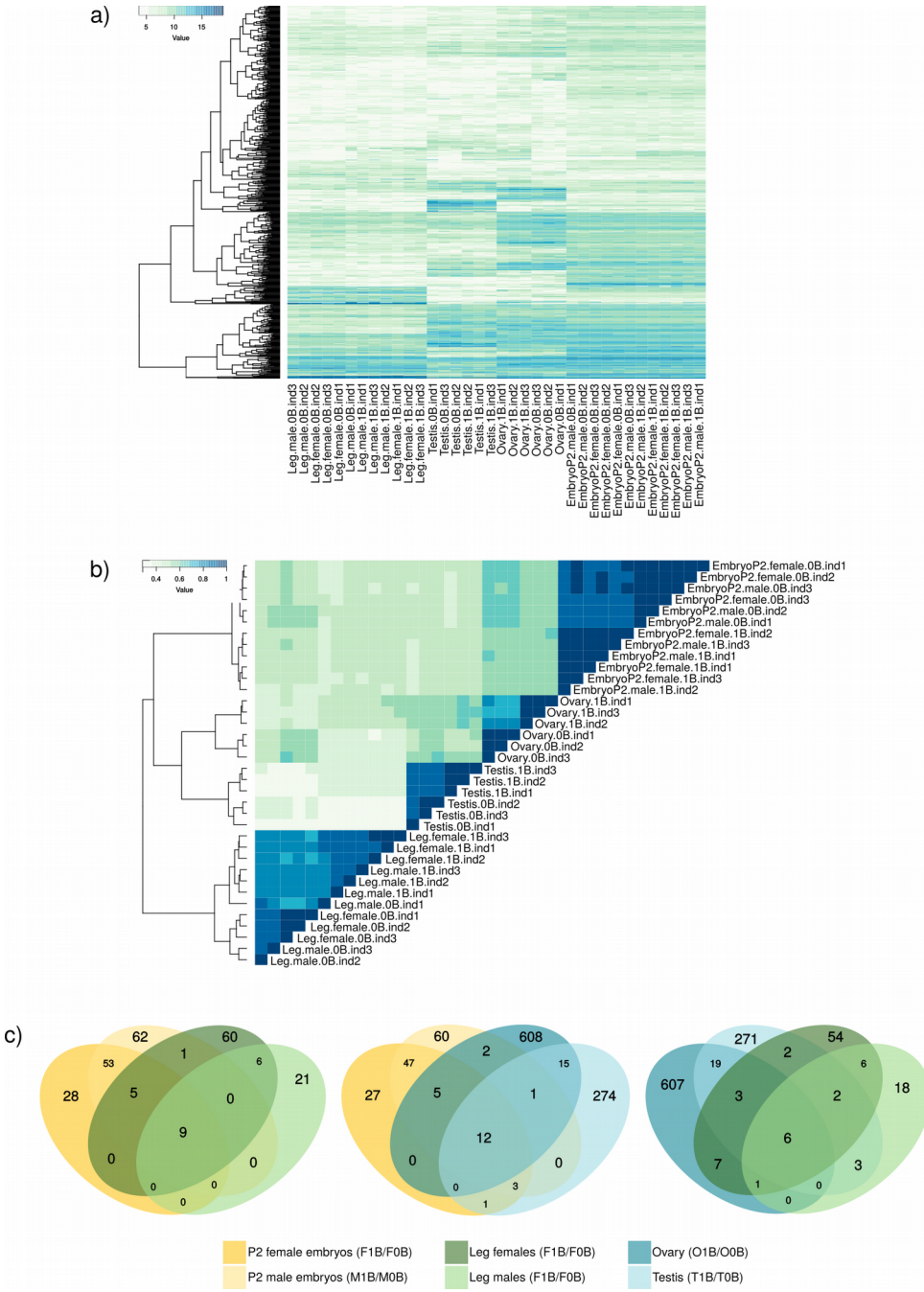


Figure 5.5. a) Heatmap of DEGs found in embryos, legs and gonads of *E. plorans* related to the occurrence of 1B chromosome. b) Clustering of samples considering DEGs counts. Note that embryos and gonads group together and apart from legs. c) Venn diagrams of DEGs in presence of 1B chromosome in embryos and legs (left), embryos and gonads (center), and gonads and legs (right).

We also performed a series of statistical analyses that revealed several interesting insights on B chromosome effects (Table 5.3, Table S.5.5):

P2 embryos-legs: The number of DEGs associated with the presence of one B chromosome (intra-sex comparisons) was much higher in embryos than in legs, in particular for males. This was the case for all kind of DEGs except for B-genes for which there were no difference between number of DEGs in legs and gonads. In case of inter-sex differences (Table S5.6) we found more DEGs between sexes in absence for legs but in was the opposite in embryos.

P2 embryos-gonads: Gonads showed higher number of DEGs than embryos, and the B chromosome evoked higher response in the ovary (intra-sex comparison). This was more evident for protein-coding DEGs and non-annotated DEGs, although in female there was also difference for TEs and B-genes. Inter-sex comparisons indicated much more DEGs between sexes in presence as well as in absence of B chromosomes in gonads compared to embryos (Table S5.6).

Interestingly, the lowest amount of difference between tissues associated with the B chromosome presence appeared for B-genes. Thus, A chromosomes may confront the occurrence of B chromosomes through different mechanisms in embryos, legs and gonads although B-genes are active in all developmental stages.

Moreover, contingency tests put into evidence different patterns of DEGs associated with the B chromosome between sexes in embryos compared to leg and gonad. In fact, there were more DEGs in case of males associated with B-presence in embryos but that number was higher in females than males for legs and gonads. This suggest that a presumably imprinting could be active in embryos but weaker in adults of *E. plorans*.

Expectedly, difference between sexes in absence of B chromosomes, were markedly lower in P2 embryos (38 DEGs) than in legs (198) or gonads (14,258). This could explain why B chromosomes cause more differences than sexes in embryos but not in adults of *E. plorans*. Remarkably, in the case of leg we found very few DEGs between sexes in absence of 1B chromosome compared to embryos and gonads. This could imply that the regulation of gene expression in embryos is focused towards the fight against genomic parasites (i.e. B chromosomes) since male-female cell differentiation is still emerging while, in adults, this fight is less predominant than sexual differences.

Table 5.3. Comparison of the numbers of DEGs (obs.: observed, exp.: expected) found in P2 embryos, legs and gonads in presence of a B chromosome, performed also separately for each group of gene annotation (no annotated, TEs, protein-coding genes from A chromosomes and B-genes). *First test that appears in each group is for totals in rows, second one for totals in columns.

Annotation	Sample comparisons	DEGs obs.			DEGs exp.		Goodness of fit*		Contingency	
		P2	Leg	Total	P2	Leg	chi (1:1)	P	chi	P
All DEGs	F1B/F0B	95	81	176	115.79	60.21	0.3	5.89E-01	22.5	**2.12E-06**
	M1B/M0B	130	36	166	109.21	56.79				
	Total	225	117	342			34.1	**5.22E-09**		
Non-annotated	F1B/F0B	59	41	100	69.76	30.24	0.1	7.27E-01	10.7	**1.07E-03**
	M1B/M0B	84	21	105	73.24	31.76				
	Total	143	62	205			32.0	**1.54E-08**		
TEs	F1B/F0B	13	3	16	13.84	2.16	0.7	4.11E-01	0.1	7.43E-01
	M1B/M0B	19	2	21	18.16	2.84				
	Total	32	5	37			19.7	**9.05E-06**		
Protein-coding genes (As)	F1B/F0B	18	31	49	23.87	25.13	5.1	**2.35E-02**	7.6	**5.92E-03**
	M1B/M0B	20	9	29	14.13	14.87				
	Total	38	40	78			0.1	8.21E-01		
B-genes	F1B/F0B	5	6	11	6	5	0.0	1.00E+00	0.7	3.92E-01
	M1B/M0B	7	4	11	6	5				
	Total	12	10	22			0.2	6.70E-01		
		P2	Gonad	Total	P2	Gonad	chi (1:1)	P	chi	P
All DEGs	F1B/F0B	95	643	738	141.44	596.56	77.7	**1.21E-18**	50.8	**1.03E-12**
	M1B/M0B	130	306	436	83.56	352.44				
	Total	225	949	1174			446.5	**4.19E-99**		
Non-annotated	F1B/F0B	59	265	324	85.64	238.36	21.2	**4.22E-06**	28.1	**1.16E-07**
	M1B/M0B	84	133	217	57.36	159.64				
	Total	143	398	541			120.2	**5.74E-28**		
TEs	F1B/F0B	13	27	40	18.29	21.71	1.4	2.32E-01	6.6	**1.04E-02**
	M1B/M0B	19	11	30	13.71	16.29				
	Total	32	38	70			0.5	4.73E-01		
Protein-coding genes (As)	F1B/F0B	18	336	354	25.57	328.43	63.0	**2.10E-15**	7.4	**6.54E-03**
	M1B/M0B	20	152	172	12.43	159.57				
	Total	38	488	526			385.0	**1.02E-85**		
B-genes	F1B/F0B	5	15	20	6.49	13.51	0.2	6.22E-01	1.1	2.95E-01
	M1B/M0B	7	10	17	5.51	11.49				
	Total	12	25	37			4.6	**3.26E-02**		

Are DEGs a response of the host genome to the presence of B chromosomes?

As already indicated, to further characterize the nature of DEGs found in *E. plorans* associated with the presence of one B chromosome, we annotated them using Trinotate (https://trinotate.github.io) and Blast2GO (Götz et al., 2018). Then, we polished the annotation by identification of B-genes, transposable elements (TEs), using RepeatMasker v4.0.5 (Smit et al., 2013) against the repetitive database of *E. plorans* (Chapter 2), or undetermined proteins (see Table S5.1 and S5.2). About 46% of DEGs found in embryos and adults failed to annotate and ~19% of them did it as TEs or undetermined proteins. Only a small fraction of DEGs was for genes located on the B-chromosomes (~3%) while about 32% of them were protein-coding genes presumably not located on B-chromosomes. Therefore, it is clear that some B-linked genes are transcribed in 1B individuals, but many A-linked genes do too, revealing an interesting transcriptional crosstalk between A and B chromosomes as a consequence of the arms race characterizing this B chromosome system (Camacho et al., 1997).

The analysis of total DEGs excluding B-genes, shown in Table S5.3, revealed that sex differences were characterized by more number of up-regulations, especially when the B chromosome was present. Up/down ratios ranged between 0.9 and 3.1 except in the case of hind legs which reached 12.4 and 6 ratios presumably due to the higher body size of the female and physiological differences between sexes probably derived from this fact.

On the other hand, the effect of the presence of the B chromosome in embryos was characterized by 20 up-regulations of B-linked genes (presumably due to the transcription of the B-paralogs) and only two down-regulations. Remarkably, the transcriptional changes observed in embryos for genes not linked to the B chromosome (thus indicating the response of the A genome to B presence) were also characterized by more up-regulations (between 4.3 and 20.3). In adults, B-linked genes showed only up-regulations except in ovary where 5 out of the 15 DEGs were down-regulations, suggesting that in this organ it is conceivable that some of the B-linked transcripts may function as repressors such as interference RNA.

The response of the A genome in embryos, indicated by the DEGs for gene not being linked to the B chromosome, was in the sense of up-regulated genes, notably in the case of TEs. This pattern was similar for females and males embryos from both pods (P1 and P2). In the case of adults, this response from A chromosomes was rather different in hind legs and gonads, as in the former there were similar numbers of up- and down-

regulations whereas in gonads there were contrasting patterns with more down-regulations in the ovary and more up-regulations in the testis. The biological meaning of these down-regulations in ovary is not so obvious but we would hypothesize that these changes in gene expression could be aimed to accommodate or prevent the drive of the B chromosome that takes places during female meiosis in *E. plorans* (Zurita et al., 1998). As such, these gene expression changes might represent the response of the A genome to counteract and/or resist to the parasitic B chromosome. Unveiling the ultimate details for these transcriptional interactions need further research.

When the same analysis was performed on each type of DEGs separately, it revealed that the immense majority of up-regulations were roughly similar for protein-coding genes, TEs and non-annotated DEGs. In contrast, the only bias towards down-regulation, which was observed in the ovary, was in fact exclusive to protein-coding regions.

To further compare up- and down-regulation patterns associated with the presence of a B chromosome between different developmental stages of *E. plorans* we performed several contingency analysis found in Table 5.4, S5.6 and S5.7.

As indicated in Table 5.4, there were no difference in the number of up- and down-regulated DEGs along ontogeny for B-linked DEGs. In the case of males, we found the same for protein-coding genes from A chromosomes and TEs although there were difference in the case of non-annotated genes with more down-regulated DEGs than expected. In contrast, the presence of a B chromosome in females was associated with different patterns of up- and down-regulated DEGs along ontogeny for all kind of annotated genes except for B-linked genes. In particular, we found an excess of up-regulated DEGs in embryos and of down-regulated ones in ovary.

This pattern was also apparent when we compared the number of DEGs associated with the presence of a B chromosome belonging to different annotation groups between different developmental stages (Table S5.7). We found differences between samples for the distribution of DEGs among annotation categories, except for P1 and P2 female embryos although we found it in the case of males. Males from P2 showed an excess of TEs and protein-coding DEGs compared to those from P1. From the comparison of P2 embryos with legs we found that embryos elicited more TEs differentially expressed than expected compared to legs and gonad in both sexes. On the other hand, in adults, both for legs and gonads, there were more protein-coding DEGs than expected compared to embryos, especially in the case of females.

Table 5.4. Comparison of the numbers of up- and down-regulated DEGs associated with one B chromosome presence along ontogeny, performed also separately in females and males and for each group of gene annotation (no annotated, TEs, protein-coding genes from A chromosomes and B-genes).

Comparison	Annotation	Sample	DEGs observed			DEGs expected		Contingency	
			Up	Down	Total	Up	Down	chi	P
F1B/F0B	All	P2	89	6	95	45.24	49.8	99.89	**2.09E-22**
		Gonad	255	388	643	306.19	336.8		
		Leg	46	35	81	38.57	42.4		
		Total	390	429	819				
	Non-annotated	P2	53	6	59	32.81	26.2	35.11	**2.38E-08**
		Gonad	126	139	265	147.38	117.6		
		Leg	24	17	41	22.8	18.2		
		Total	203	162	365				
	TEs	P2	13	0	13	8.47	4.5	8.53	**0.01**
		Gonad	13	14	27	17.58	9.4		
		Leg	2	1	3	1.95	1.0		
		Total	28	15	43				
	Protein-coding genes (As)	P2	18	0	18	6.45	11.5	36.08	**1.46E-08**
		Gonad	106	230	336	120.44	215.6		
		Leg	14	17	31	11.11	19.9		
		Total	138	247	385				
	B-genes	P2	5	0	5	4.04	1.0	1.85	0.39
		Gonad	10	5	15	12.12	2.9		
		Leg	6	0	6	4.85	1.2		
		Total	21	5	26				
M1B/M0B	All	P2	106	24	130	98.88	31.1	10.46	**0.01**
		Gonad	233	73	306	232.74	73.3		
		Leg	20	16	36	27.38	8.6		
		Total	359	113	472				
	Non-annotated	P2	66	18	84	63.88	20.1	18.32	**0.0001**
		Gonad	107	26	133	101.15	31.9		
		Leg	8	13	21	15.97	5.0		
		Total	181	57	238				
	TEs	P2	18	1	19	18.41	0.6	3.25	0.2
		Gonad	11	0	11	10.66	0.3		
		Leg	2	0	2	1.94	0.1		
		Total	31	1	32				
	Protein-coding genes (As)	P2	16	4	20	14.03	6.0	0.57	0.75
		Gonad	105	47	152	106.65	45.3		
		Leg	6	3	9	6.31	2.7		
		Total	127	54	181				
	B-genes	P2	6	1	7	6.67	0.3	0.62	0.73
		Gonad	10	0	10	9.52	0.5		
		Leg	4	0	4	3.81	0.2		
		Total	20	1	21				

These results highlight that the B chromosome presence is associated with a response of transposable elements in embryos while it takes place with protein-coding genes in the case of adults.

Finally, we explored whether there were difference for up- and down-regulation patterns between protein-coding DEGs coming from A and B chromosomes across different samples including in this study (Table S5.8). Interestingly, in any of the samples we found differences except in ovary and, in a lesser extent, in female legs. We observed a excess of B-genes up-regulations compared to A-genes in ovary where we found a high number of down-regulated protein-coding DEGs presumably from A chromosomes. Therefore, suggesting, apart from B transcription, a response from A chromosomes by means of down-regulated DEGs in ovary, the tissue in which the B-drive occurs.

Functional GO enrichment of DEGs

To explore the functional characterization of DEGs associated with the presence of a B chromosomes in the different samples of *E. plorans* studied here, we carried out an enrichment analysis in GO terms. This was done for common and specific DEGs associated with the presence of the B chromosome (0B vs 1B library comparisons) at different sexes and stages of development.

Among the few common DEGs between embryos and adults of both sexes (Fig. 5.5c) we find some non-annotated genes but also some B-genes, which reinforces the idea of the activity of these sequences in all the *E. plorans* samples analyzed here. The *mycb2* gene appeared differentially expressed in all ontogenetic stages while *cip2a* did so in embryos and legs.

The enrichment in GO terms produced significant results, considering an adjusted p-value as FDR (adjusted p-value <0.05), for common up-regulated DEGs in embryos, common up-regulated DEGs between gonads and testis, and also for specific down-regulated DEGs in ovary. For the rest of GO enrichment analysis, we set a p-value lower than 0.05 to consider a GO term as enriched in our comparisons.

In this way, the common DEGs in embryos were related to transposition and integration of DNA GO terms (Fig 5.6) such as *RNA-dependent DNA biosynthetic process* (GO:0006278) or *viral penetration into host nucleus* (GO:0075732). This is consistent with the high number of TEs differentially expressed in these samples (Table S5.7). However, the analysis of up-regulated DEGs specific for P1 embryos yielded several GO terms

related to the cell cycle and chromosomes as *chromatin silencing at centromere* (GO:0030702), *mitotic cell cycle process* (GO:1903047) or *chromatin organization* (GO:0006325) in females and, to a lesser extent, in males. On the other hand, down-regulated genes in P1 embryos, particularly in females, also showed an enrichment in some functions related to the cell cycle such us *anaphase-promoting complex-dependent catabolic process* (GO:0031145). These functions did not appear in P2 embryos in which we found that down-regulated DEGs were related to translation and development regulation (e.g. *developmental maturation* – GO:0044403, *transcriptional activator activity* – GO:0001077 or GO:0001228, *translation* – GO:0006412, *ribosome* – GO:0005840). This characterization of GO functions in embryos shows an association of maternally inherited B chromosome (P1) with cell cycle functions, whereas if the B was paternally inherited, translation and ribosomal activity decrease. This could reflect a higher effect of the B chromosome on the host genome when it is inherited through the maternal route, whereas it could be more silenced and attenuated when it is inherited through the male parent, which would be consistent with the elimination of B in spermiogenesis (Chapter 4).

In the case of hind legs, we found GO terms related to cytoskeleton and condensin (e.g. *regulation of cytoskeleton organization* – GO:0051493, *DNA packaging complex* – GO:0044815, *condensin complex* – GO:0000796) for common up-regulated DEGs associated with the presence of a B chromosome between both sexes (Fig. 5.6). On the other hand, down-regulated common DEGs were related to functions involved in metabolism and nervous system as *neuromuscular synaptic transmission* – GO:0007274, *positive regulation of skeletal muscle tissue development* – GO:0048643, *carbohydrate derivative metabolic process* – GO:1901135.

GO terms related to cytoskeleton, catabolism and nervous system were found for common up-regulated DEGs between ovary and testis such us regulation of cytoskeleton organization – GO:0051493, negative regulation of catabolic process – GO:0009895 or axon (GO:0030424). On the other hand, common down-regulated DEGs in male and female gonads were related to oxidative metabolism, for instance reactive oxygen species metabolic process – GO:0072593, peroxisome – GO:0005777 or oxidoreductase activity – GO:0016725 (Fig. 5.6). Apparently, these GO terms seem not to be directly related with cell cycle function or other roles involving B chromosome dynamics and maintenance, so we performed a GO enrichment analysis on specific DEGs in testis and ovary (Fig S5.5).

Figure 5.6. GO enrichment for common DEGs between males and females embryos from P1 and P2 (up), between female and male legs (left) and ovary and testis (right) in presence of one B chromosome. See in green and orange up- and down-regulated DEGs respectively. Note that there were not common down-regulated DEGs between embryos of *E. plorans*.

After this enrichment analysis, we found for up-regulated DEGs in ovary GO terms reflecting functions related to cytoskeleton and cell cycle as *cytoplasmic microtubule organization* – GO:0031122, *regulation of cell cycle checkpoint* – GO:1901976 or *regulation of centriole replication* – GO:0046599. On the other hand, specific down-regulated DEGs in ovary showed functions related to cardiac developments and apoptotic processes such as *cardiac ventricle formation* – GO:0003211, *negative regulation of extrinsic apoptotic signaling pathway in absence of ligand* – GO:2001240 or *negative regulation of anoikis* – GO:2000811. However, looking at unadjusted p-values there were several GO terms involved in cell cycle functions that were down-regulated in the presence of a B chromosome in ovary as *regulation of establishment of bipolar cell polarity* – GO:0061172, *regulation of stem cell proliferation* – GO:0072091 or *mitotic sister chromatid separation* – GO:0051306 . In contrast to common DEGs between testis and ovary, these GO terms resemble endophenotypic characteristics claim for B chromosomes (Navarro-Domínguez et al., 2019) and those DEGs could have a role in their evolutionary dynamics. In fact, Akera et al. (2019) stated that a delayed cell division promote drive of selfish elements (i.e. B chromosomes) and that could take place in *E. plorans* supported by the down-regulation of cell cycle DEGs that we report here in ovary, the organ in which the drive of the B chromosome takes place in this species.

In the case of specific DEGs in testis, the GO enrichment analysis showed that *male germ-line sex determination* (GO:0019100) and *negative regulation of centrosome duplication* (GO:0010826) were down-regulated in presence of a B chromosome (Fig. S5.5). Interestingly, in contrast to ovary, up-regulated genes in testis were directly involved in cell cycle, reproduction and silencing mechanisms as *cellular process involved in reproduction in multicellular organism* (GO:0022412), *maintenance of location in cell* (GO:0051651), *negative regulation of sequence-specific DNA binding transcription factor activity* (GO:0043433) or *negative regulation of DNA-templated transcription, termination* (GO:0060567). In fact, the gene *daxx,* coding for the Death Domain Associated Protein*,* was up-regulated at occurrence of a B chromosome in embryos, legs and testis but not in ovary. The protein DAXX acts as a corepressor down-regulating basal and activated transcription and also as an important role chromatin remodeling (Yang et al., 1997; Kwon et al., 2006; Tang et al., 2010). Therefore, we could speculate that the up-regulation of *daxx* in embryos, legs and testis might be involved in the silencing mechanisms for the B chromosome presence and perhaps also in their

elimination from testis (Chapter 4). On the contrary, *daxx* is not differentially expressed in ovary of *E. plorans*, the place in which the B chromosome is expected to exhibit its most effective resources to drive meiotically through offspring.

Discussion

The true nature of the B chromosome intragenomic conflict is still poorly understood. In this study we address this issue through the analysis of changes in gene expression between 1B and 0B embryos and adults of the grasshopper *E. plorans* belonging to both sexes. First attempt to uncover the transcription changes associated with the presence of B chromosomes in *E. plorans* by Navarro-Domínguez et al. (2019) verified that in females of this species the B chromosome triggers changes in gene expression consistent with their endophenotypic effects in the host. This experiment was carried out considering the whole body of only two females from Torrox (one with B chromosomes and the other without them). Therefore, despite their relevant findings, authors were unable to detect transcriptomic changes associated with sex or specific tissues in the presence of B chromosomes which is one of the focus in this study. Here we include adults of both sexes and consider gonads and legs separately. In addition, we also analyze the embryonic stage since it could be probably a key moment for an individual of *E. plorans* to adapt and deal with the B chromosome presence. However, it is technically a very hard task to extract enough biological material from a single embryo to properly perform a comprehensive RNA-seq experiment.

In this study we propose a molecular approach to extract enough biological material from a single 10-day old embryo of a grasshopper species to perform cytogenetic characterization, DNA analysis and RNA-seq experiment. This protocol could be of great importance to investigate transcriptomics changes between carriers and non-carriers of B chromosomes in other species or ontogeny stages where the biological material is scarce. This molecular approach could be very useful even to perform RNA-seq experiment between sexes if there are no other ways to distinguish between females and males apart from cytogenetic techniques.

During the past four decades, significant progress in understanding of B chromosome nature has been made through several points of view (Camacho, 2005; Houben, 2017). Currently, immersed in the time of NGS and massive data analysis, our

RNA-seq experiments on *E. plorans* have provided enormous amount of information on gene expression at the whole genome level related to B chromosome presence in both sexes, different tissues (i.e. legs and gonads) and developmental stages (adults and embryos).

In particular, we make clear that the B chromosome presence in *E. plorans* embryos from P1 as well as for P2 leads to more differences in gene expression that those occurring due to sex (i.e. between females and males lacking B chromosomes). It could be controversial to assert that a B chromosome effectively causes more gene expression changes than sex in *E. plorans* embryos without previous knowledge of the complete sex determination pathway of the species.

However, we found 25 transcripts annotated as several genes involved in sex determination in other well-known insects in the assembled transcriptome of *E. plorans* embryos such as DSX_DROME (doublesex), JANA_DROPS (Sex-regulated protein janus-A), TRA2_DROVI (transformer-2 sex-determinig protein) or SXL_DROME (protein sex-lethal). Therefore, some genes related to sex determination are being transcribed in our embryo samples although significant differences in gene expression were not found for those genes. In fact, sex determination in insects takes place at very early stage development (Bopp et al., 2014). For example, in *Bombix mori* sex determination occurs between 29 and 32 hours after oviposition (Sakai et al., 2014), transcription of *nix*, a male-determining factor of the mosquito *Aedes aegypti*, was first detected 3 to 4 hours after oviposition (Hall et al., 2015) and in the fruit fly *Bactrocera jarvisi* sex-specific expression of TRA transcripts occurred between 3 to 6 hours after egg laying, and the DSX isoform was established by 7 hours (Morrow et al., 2014). Regarding the determination of sex in grasshoppers, Nelsen (1931) described the presence of primordial germ cells, that in no case presented mitotic divisions, after the phase that the author called revolution and later Bentley et al. (1979) referred to it as katatrepsis. This stage begins at 45% of the embryonic development, while the pigmentation of the embryo's eyes occurs in the middle of development. In this experiment we processed the embryos at 10 days and always showing no eyes pigmentation, so we can assume that they were below the 50% of their embryonic development. Therefore, we cannot rule out that we have not captured the moment in which differences between the sexes begin to be observed. Accordingly, we did not found significant differences in gene expression between sexes in the absence of B chromosomes and 0B embryos from any of the pods did not clustered

in a sex-dependent manner when 1B samples were excluded from a correlation heatmap (see Figure S5.4c).

We compared DEGs between sexes in absence of B chromosomes, embryos present the lower number of DEGs compared with adult gonads and legs. Accordingly, we found more difference for gene expression between sexes in adults than those caused by the presence of 1B. Despite this, the number of DEGs was much higher in gonads, especially in ovary, than those in embryos and legs. In addition, we found more DEGs in embryos than in leg associated with the presence of one B chromosome.

It is worthy remarkable the influence of B chromosome on gene expression of ovary and female leg, stronger than in the case of males. This result is quite consistent bearing in mind that the number of chromosomes in a 1B female of *E. plorans* is 11 autosome pairs, one pair of sexual chromosomes (two XX) plus 1 B chromosome that will behave as an univalent chromosome that could cause troubles during cell division. On the contrary, a 1B male has 11 pairs of autosomes plus 1 univalent X chromosome and 1 B univalent chromosome, consequently these two univalent could recombine somehow and diminish the effects of a B chromosome during cell division. In fact, it was reported that X and B chromosomes display similar meiotic characteristics in male grasshoppers (Viera et al., 2004).

In P1 and P2 embryos of *E. plorans* we found an up-regulation of genes in the presence of one B chromosomes, most of them were annotated as transposable elements or undetermined proteins. Ge (2017) proved through big data analysis that transposable elements may play a role in establishing the expression landscape in early embryos, as well as Ansaloni et al. (2019) did in embryos of *Caenorhabditis elegans*. Moreover, this up-regulation of transposons in the presence of 1B chromosome could be a way to overcome stressful conditions as it was described for *Arabidopsis* (Le et al., 2014) or *Drosophila* (Rech et al., 2019). On the other hand, P1 embryos showed more DEGs, including B-genes, in the presence of 1B chromosomes than embryos from P2, thus a maternal B chromosome could be more active or have a stronger effect on the host genome than if it was paternally inherited. Indeed, in Chapter 4 of this thesis (Cabrero et al., 2017) we demonstrated the elimination of B chromosomes during spermiogenesis so it could exist some procedures aimed to weaken them that could impact also on gene expression in embryos of *E. plorans*. In this sense, the maternal imprinting was confirmed for B chromosomes of rye by Puertas et al. (1990) but more

research about this issue at molecular level could shed new light about this phenomenon.

In the case of adults, the gonads of *E. plorans* are the tissues in which the effect of B chromosomes is more striking. We found not only expression of B-genes among the DEGs but also other protein-coding genes that were much more abundant than the former.

These findings depict the battle of the intragenomic conflict between A and B chromosomes. Recently, Park et al. (2019) compared 1B and 0B transcriptomes from the plant *Lilium amabile* and they found the over-expression of several cell cycle genes. In addition, this year Boudichevskaia et al. (2020) found 341 up-regulated B-unique isoforms in meristematic cells of *Aegilops speltoides* embryos where the programmed elimination of B chromosomes occurs. Here we showed that in gonads of *E. plorans* there is a strong effect of B chromosome on gene expression, in particular in the case of ovary. This is consistent with the accumulation mechanism exhibit by B chromosomes of *E. plorans* though female meiosis, i.e. in gonads (Zurita et al., 1998). In fact, although most of the protein-coding DEGs found in testis associated with the presence of one B chromosome were up-regulated, it was not so in ovary where we found a high fraction of protein-coding genes not located in B chromosomes that were down-regulated. Therefore, that could suggest that in testis there is an up-regulation from A chromosomes aimed to fight against the B chromosomes (coherent with B chromosome elimination in this tissue) while in ovary the down-regulation of some protein-coding genes from A chromosomes could pave the way to the drive of B chromosomes and thus their maintenance in natural populations.

Our results indicate that the intragenomic conflict caused by the presence of a parasitic chromosome elicits more gene expression changes in gonads than in other tissues, thus putting into perspective the functional repercussion of the presence of these genetics elements.

References

Akbari OS, Antoshechkin I, Hay BA, Ferree PM. (2013). Transcriptome profiling of *Nasonia vitripennis* testis reveals novel transcripts expressed from the selfish B chromosome, paternal sex ratio. *G3 (Bethesda)*, 3(9), 1597–1605.

Akera T, Trimm E, Lampson MA. (2019). Molecular Strategies of Meiotic Cheating by Selfish Centromeres. *Cell*, 178(5), 1132–1144.

Ansaloni F, Scarpato M, Di Schiavi E, Gustincich S, Sanges R. (2019). Exploratory analysis of transposable elements expression in the *C. elegans* early embryo. *BMC Bioinformatics*, 20(Suppl 9), 484.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1), 25–9.

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(7), 2536–2544.

Bentley D, Keshishian H, Shankland M, Toroian-Raymond A. (1979). Quantitative staging of embryonic development of the grasshopper, *Schistocerca nitens*. *Journal of Embryology and Experimental Morphology*, 54(3), 47–74.

Bidau CJ, Rosato M, Martí DA. (2004). FISH detection of ribosomal cistrons and assortment-distortion for X and B chromosomes in *Dichroplus pratensis* (Acrididae). *Cytogenetic and Genome Research*, 106(2-4), 295–301.

Boeckmann B, Bairoch A, Apweiler R, et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31(1), 365–370.

Bolger AM, Lohse M, Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120.

Bopp D, Saccone G, Beye M. (2014). Sex determination in insects: variations on a common theme. *Sexual Development*, 8(1-3), 20–8.

Boudichevskaia A, Ruban A, Thiel J, Fiebig A, Houben A. (2020). Tissue-Specific Transcriptome Analysis Reveals Candidate Transcripts Associated with the Process of Programmed B Chromosome Elimination in *Aegilops speltoides*. *International Journal of Molecular Sciences*, 21(20), 7596.

Brockhouse C, Bas JAB, Fereday RM, Strauss NA. (1989). Supernumerary chromosomes evolution in the *Simulium vernum* group (Diptera: Simulidae). *Genome*, 32, 516–521.

Cabrero J, Manrique-Poyato MI, Camacho JPM. (2006). Detection of B chromosomes in interphase hemolymph nuclei from living specimens of the grasshopper *Eyprepocnemis plorans*. *Cytogenetic and Genome Research*, 114(1), 66–69.

Cabrero J, Martín-Peciña M, Ruiz-Ruano FJ, Gómez R, Camacho JPM. (2017). Post-meiotic B chromosome expulsion, during spermiogenesis, in two grasshopper species. *Chromosoma*, 126(5), 633–644.

Camacho JPM, Carballo AR, Cabrero J. (1980). The Bchromosome system of the grasshopper *Eyprepocnemis plorans* subsp. *plorans* (Charpentier). *Chromosoma*, 80, 163176.

Camacho JPM, Shaw MW, López-León MD, Pardo MC, Cabrero J. (1997). Population dynamics of a selfish B chromosome neutralized by the standard genome in the grasshopper *Eyprepocnemis plorans*. *The American Naturalist*, 149(6), 1030–1050.

Camacho JPM. (2005). B chromosomes. In: Gregory TR, ed., The evolution of the genome. San Diego: Elsevier, 223–286.

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.

Chen Y, Lun ATL and Smyth GK. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-

likelihood pipeline [version 2; peer review: 5 approved]. *F1000Research*, 5, 1438.

Chen Y, McCarthy D, Ritchie M, Robinson M, Smyth G. (2020). edgeR: differential analysis of sequence read count data. User's Guide. https://bioconductor.org/.

Dapas M, Kandpal M, Bi Y, Davuluri RV. (2017). Comparative evaluation of isoform-level gene expression estimation algorithms for RNA-seq and exon-array platforms. *Briefings in Bioinformatics*, 18(2), 260–269.

Finn RD, Clements J, Eddy SR. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(Web Server issue), W29–W37.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1), D279–85.

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.

Ge SX. (2017). Exploratory bioinformatics investigation reveals importance of "junk" DNA in early embryo development. *BMC Genomics*, 18(1), 200.

Graphodatsky AS, Kukekova AV, Yudkin DV, Trifonov VA, Vorobieva NV, Beklemisheva VR, et al. (2005). The proto-oncogene C-KIT maps to canid B-chromosomes. *Chromosome Research*, 13(2), 113–122.

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, 36(10), 3420–3435.

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

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8), 1494–512.

Hall AB, Basu S, Jiang X, Qi Y, Timoshevskiy VA, Biedler JK, et al. (2015). SEX DETERMINATION. A male-determining factor in the mosquito *Aedes aegypti*. *Science*, 348(6240), 1268–70.

Houben A. (2017). B Chromosomes - A Matter of Chromosome Drive. *Frontiers in Plant Science*, 8, 210.

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.

Kwon JE, La M, Oh KH, Oh YM, Kim GR, Seol JH, et al. (2006). BTB domain-containing speckle-type POZ protein (SPOP) serves as an adaptor of Daxx for ubiquitination by Cul3-based ubiquitin ligase. *Journal of Biological Chemistry*, 281(18), 12664-72.

Langmead B, Trapnell C, Pop M, Salzberg SL. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3), R25.

Le TN, Schumann U, Smith NA, Tiwari S, Au PC, Zhu QH, et al. (2014). DNA demethylases target promoter transposable elements to positively regulate stress responsive genes in *Arabidopsis*. *Genome Biology*, 15(9), 458.

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(1), 269–278.

Li B, Dewey CN. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323.

López-León MD. (1992a). Significado biológico de la hecterocromatina supernumeraria de *Eyprepocnemis plorans* (Tesis doctoral). Universidad de Granada. http://hdl.handle.net/10481/14156.

LópezLeón MD, Cabrero J, Camacho JPM, Cano MI, Santos JL. (1992b). A widespread B chromosome polymorphism maintained without apparent drive. *Evolution*, 46, 529539.

Ma W, Gabriel TS, Martis MM, Gursinsky T, Schubert V, Vrána J, et al. (2017). Rye B chromosomes encode a functional argonaute-like protein with in vitro slicer activities similar to its A chromosome paralog. *New Phytologist*, 213(2), 916–928.

Makunin AI, Rajičić M, Karamysheva TV, Romanenko SA, Druzhkova AS, Blagojević J, et al. (2018). Low-pass single-chromosome sequencing of human small supernumerary marker chromosomes (sSMCs) and *Apodemus* B chromosomes. *Chromosoma*, 127(3), 301–311.

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

Morrow JL, Riegler M, Frommer M, Shearman DC. (2014). Expression patterns of sex-determination genes in single male and female embryos of two *Bactrocera* fruit fly species during early development. *Insect Molecular Biology*, 23(6), 754–67.

Navarro-Domínguez B, Ruiz-Ruano FJ, Cabrero J, Corral JM, López-León MD, Sharbel TF, et al. (2017). Protein-coding genes in B chromosomes of the grasshopper *Eyprepocnemis plorans*. *Scientific Reports*, 7, 45200.

Navarro-Domínguez B, Martín-Peciña M, Ruiz-Ruano FJ, Cabrero J, Corral JM, López-León MD, et al. (2019). Gene expression changes elicited by a parasitic B chromosome in the grasshopper *Eyprepocnemis plorans* are consistent with its phenotypic effects. *Chromosoma*, 128, 53–67.

Nelsen OE. (1931). Life cycle, sex differentiation, and testis development in *Melanoplus differentialis* (Acrididae, Orthoptera). *Journal of Morphology*, 51(2), 467–525.

Park D, Kim JH, Kim NS. (2019). De novo transcriptome sequencing and gene expression profiling with/without B-chromosome plants of *Lilium amabile*. *Genomics & Informatics*, 17(3), e27.

Puertas MJ, Jiménez M, Romera F, Vega JM, Díez M. (1990). Maternal imprinting effect on B chromosome transmission in rye. *Heredity* , 64, 197–204.

Rech GE, Bogaerts-Márquez M, Barrón MG, Merenciano M, Villanueva-Cañas JL, Horváth V, et al. (2019). Stress response, behavior, and development are shaped by transposable element-induced mutations in *Drosophila*. *PLoS Genetics*, 15(2), e1007900.

Robinson MD, McCarthy DJ, Smyth GK. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26, 139–140.

Robinson MD, Oshlack A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11, R25.

Ruiz-Estévez M, Cabrero J, Camacho JPM. (2012). B-chromosome ribosomal DNA is functional in the grasshopper *Eyprepocnemis plorans*. *PLoS ONE*, 7(5).

Ruiz-Estévez M, López-León MD, Cabrero J, Camacho JPM. (2013). Ribosomal DNA is active in different B chromosome variants of the grasshopper *Eyprepocnemis plorans*. *Genetica*, 141(7-9), 337–45.

Ruiz-Estévez M, Badisco L, Broeck J, Perfectti F, López-León M, Cabrero J, 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(6), 1209–1216.

Sakai H, Aoki F, Suzuki MG. (2014). Identification of the key stages for sex determination in the silkworm, *Bombyx mori*. *Development Genes and Evolution*, 224(2), 119–123.

Smit AFA, Hubley R, Green P. (2013). RepeatMasker Open–4.0. http://www.repeatmasker.org.

Soneson C, Love MI, Robinson MD. (2016). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; peer review: 2 approved]. *F1000Research*, 4, 1521.

Supek F, Bošnjak M, Škunca N, Šmuc T. (2011). REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLoS ONE*, 6(7), e21800.

Tang J, Qu L, Pang M, Yang X. (2010). Daxx is reciprocally regulated by Mdm2 and Hausp. *Biochemical and Biophysical Research Communications*, 393(3), 542-545.

Teruel M, Cabrero J, Perfectti F, Camacho JPM. (2007). Nucleolus size variation during meiosis and NOR activity of a B chromosome in the grasshopper *Eyprepocnemis plorans*. *Chromosome Research*, 15(6), 755–765.

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, O'Brien PC, Perelman PL, Yang F, et al. (2013). Transcription of a protein-coding gene on B chromosomes of the Siberian roe deer (*Capreolus pygargus*). *BMC Biology*, 11, 90.

Van Vugt JJ, de Nooijer S, Stouthamer R, de Jong H. (2005). NOR activity and repeat sequences of the paternal sex ratio chromosome of the parasitoid wasp *Trichogramma kaykai*. *Chromosoma*, 114(6), 410–419.

Viera A, Calvente A, Page J, Parra MT, Gómez R, Suja JA, et al. (2004). X and B chromosomes display similar meiotic characteristics in male grasshoppers. *Cytogenetic and Genome Research*, 106(2-4), 302–8.

Werren JH. (1991). The paternal-sex-ratio chromosome of *Nasonia*. *The American Naturalist*, 137(3), 392–402.

Yang X, Khosravi-Far R, Chang HY, Baltimore D. (1997). Daxx, a novel Fas-binding protein that activates JNK and apoptosis. *Cell*, 89(7), 1067-76.

Yi L, Pimentel H, Bray NL, Pachter L. (2018). Gene-level differential analysis at transcript-level resolution. *Genome Biology*, 19(1), 53.

Yoshida K, Terai Y, Mizoiri S, Aibara M, Nishihara H, Watanabe M, et al. (2011). B chromosomes have a functional effect on female sex determination in lake victoria cichlid fishes. *PLoS Genetics*, 7(8), e1002203.

Young MD, Wakefield MJ, Smyth GK, Oshlack A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 11(2), R14.

Zhou, Q, Zhu H, Huang Q, Zhao L, Zhang G, Roy SW, et al. (2012). Deciphering neo-sex and B chromosome evolution by the draft genome of *Drosophila albomicans*. *BMC Genomics*, 13(1), 109.

Zurita S, Cabrero J, López-León MD, Camacho JPM. (1998). Polymorphism regeneration for a neutralized selfish B chromosome. *Evolution*, 52(1), 274–277.

Supplementary Figures for Chapter 5

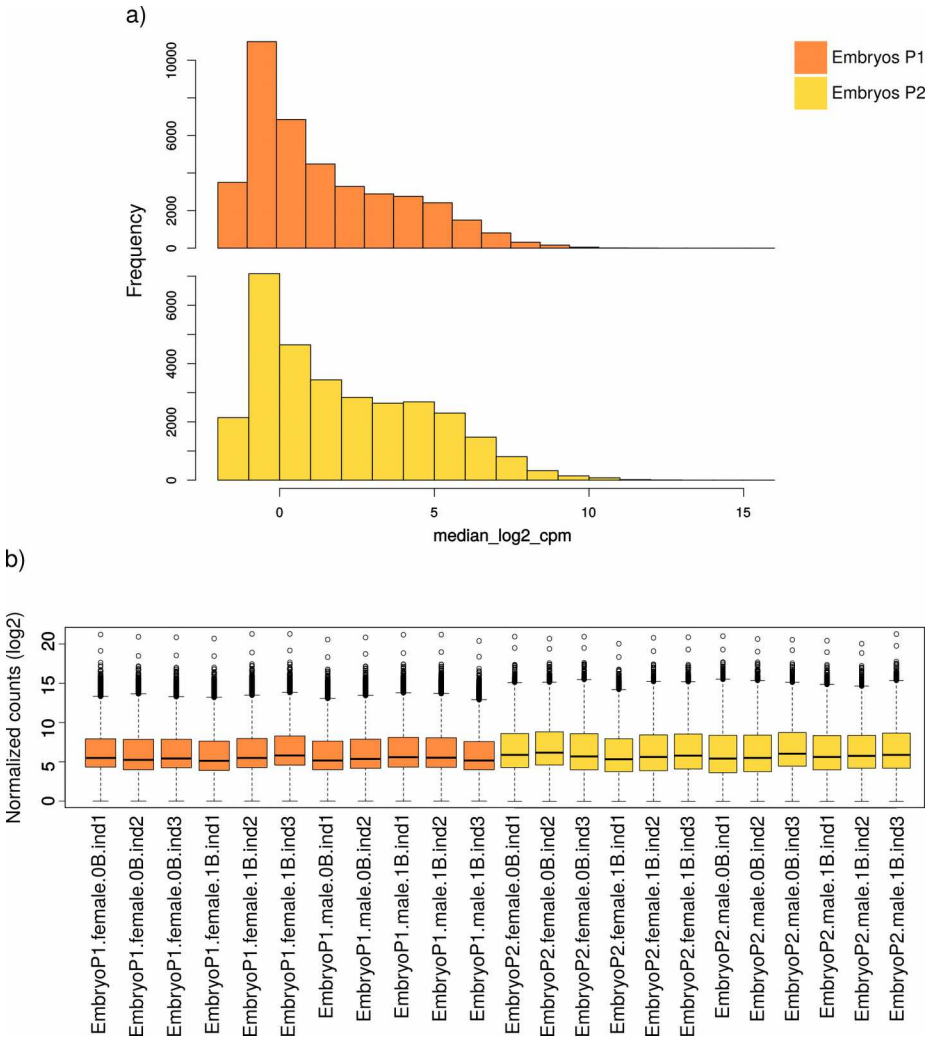


Figure S5.1. a) Histograms of filtered gene counts in cpm (counts per million) in P1 and P2 embryos of *E. plorans*. b) Boxplots of normalized counts for all P1 and P2 embryo samples.

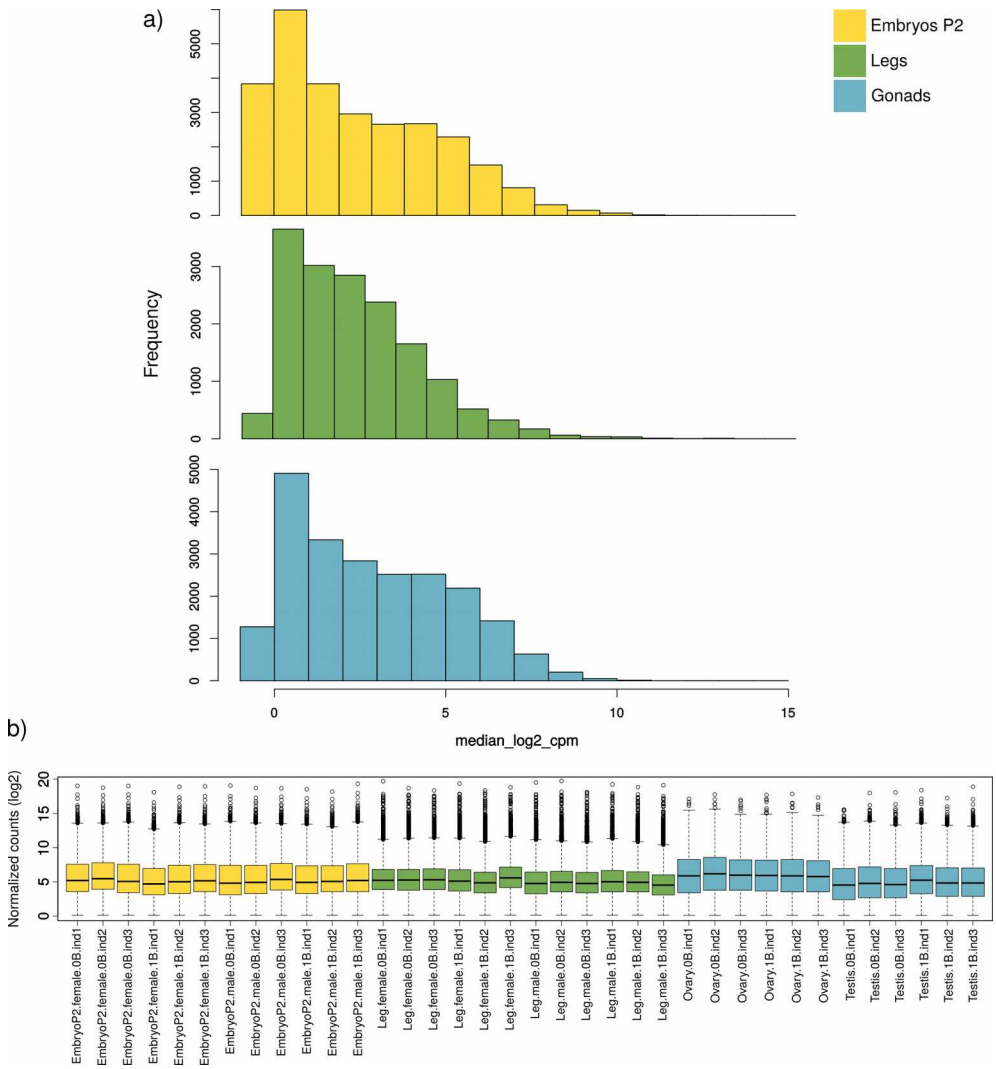


Figure S5.2. a) Histograms of filtered gene counts in cpm (counts per million) in P2 embryos, legs and gonads of *E. plorans*. b) Boxplots of normalized counts for all P2 embryo, leg and gonad samples.

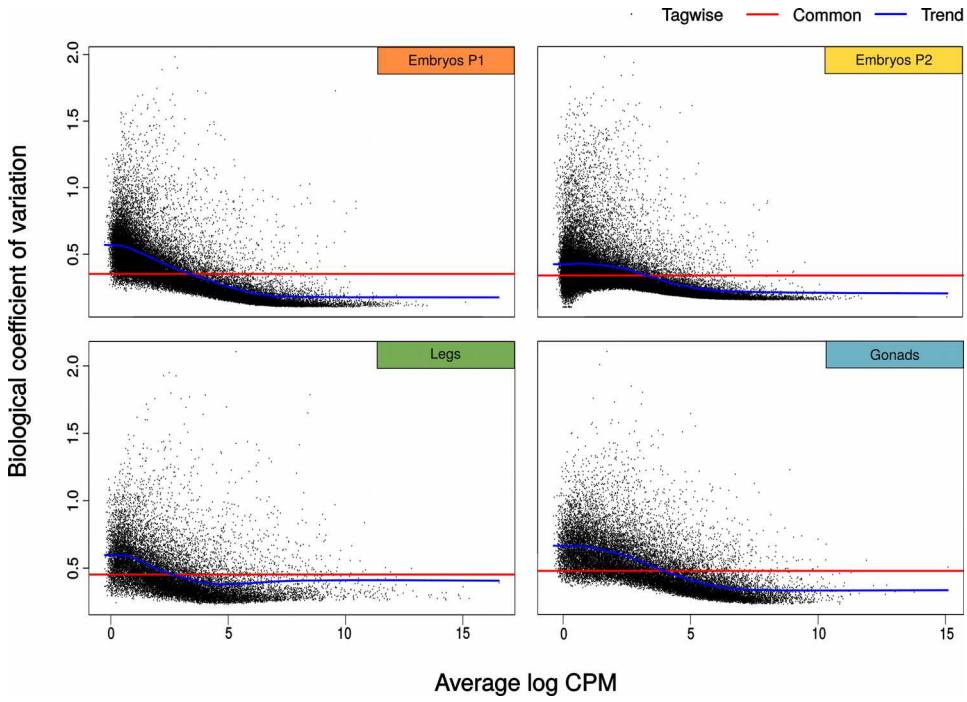


Figure S5.3. Biological variation plots representing data dispersion in embryos (P1 and P2), legs and gonads of *E. plorans* after counts filtering.

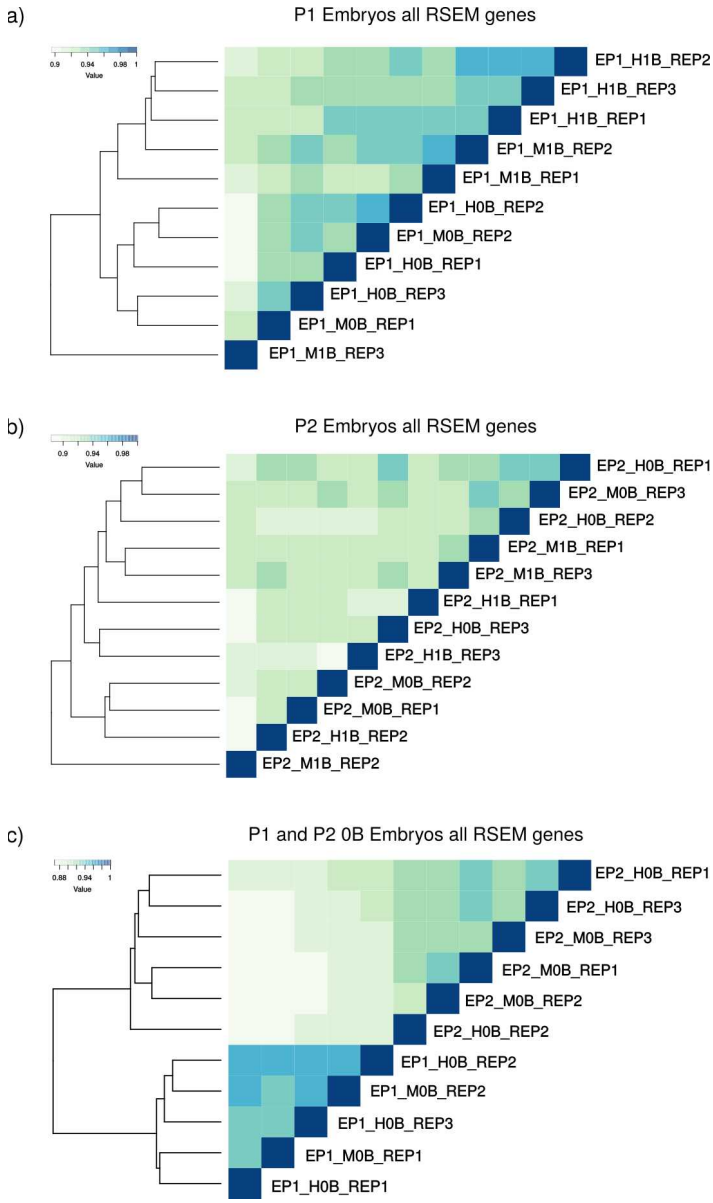


Figure S5.4. Sample correlation heatmaps of filtered count matrix in P1 embryos (a), P2 embryos (b) and both pods (c).

Figure S5.5. GO enrichment for specific DEGs in ovary (left) and testis (right) in presence of one B chromosome. See in green and orange up- and down-regulated DEGs respectively.

Supplementary Datasets and Tables for Chapter 5

Dataset 5.1. MA plots for all samples comparisons included in this study.
Dataset available in: https://figshare.com/s/0c24d76cd188bf5599a4

Dataset 5.2. Curated GO enrichment analysis using REVIGO for common and specific DEGs associated to the presence of one B chromosome for different sample comparisons.
Dataset available in: https://figshare.com/s/f834da50777f91d9ac3a

Supplementary Tables can be downloaded in:
https://figshare.com/s/869b9b94bfe907ca18ee

Table S5.1. Annotation, logFC (log2 fold-change), FDR (false discovery rate) for DEGs in four comparisons in P1 and P2 embryo samples of *E. plorans*: 1B females vs 0B females (F1B/F0B), 1B males vs 0B males (M1B/M0B), 0B females vs 0B males (F0B/M0B) and 1B females vs 1B males (F1B/M1B). We indicated the up- or down-regulation of the DEG depending on the logFC>1 or logFC<1 respectively.

Table S5.2. Annotation, logFC (log2 fold-change), FDR (false discovery rate) for DEGs in four comparisons in P2 embryos, leg and gonad samples of *E. plorans*: 1B females vs 0B females (F1B/F0B), 1B males vs 0B males (M1B/M0B), 0B females vs 0B males (F0B/M0B) and 1B females vs 1B males (F1B/M1B). Note that in case of gonads we used ovary and testis terms instead of female and male. We indicated the up- or down-regulation of the DEG depending on the logFC>1 or logFC<1 respectively.

Table S5.3. Number of DEGs between different samples of *E. plorans* (embryos, gonads and legs) annotated as B-genes, protein-coding genes, transposable elements or undetermined proteins (TE/Undetermined) or that failed to be annotated. We calculated the ratio between up- and down-regulated genes and also between the number of B-genes in respect to the total of protein-coding DEGs annotated.

Table S5.4. Statistical test for number of DEGs in embryos.

Table S5.5. Statistical test for number of DEGs in adults.

Table S5.6. Statistical test for number of DEGs at different developmental stages.

Table S5.7. Statistical test for different annotation of DEGs and developmental stages.

Table S5.8. Statistical test for up- and down-regulation patterns of A and B protein-coding genes.

Appendix.

Phylogenetic signal of genomic repeat abundances can be distorted by random homoplasy: a case study from hominid primates

Published as:
Martín-Peciña M, Ruiz-Ruano FJ, Camacho JPM, Dodsworth S. (2019). Phylogenetic signal of genomic repeat abundances can be distorted by random homoplasy: a case study from hominid primates. *Zoological Journal of the Linnean Society*, 185, 543–554. doi.org/10.1093/zoolinnean/zly077.

Abstract: The genomic abundance of different types of repetitive DNA elements contains a phylogenetic signal useful for inferring the evolutionary history of different groups of organisms. Here we test the reliability of this approach using the Hominidae family of primates whose consensus phylogeny is well accepted. We used the software RepeatExplorer to identify the different repetitive DNA clusters and quantify their abundances. With this data we performed phylogenetic analyses by maximum parsimony, including one, two or three individuals per species, technical replicates, and including or discarding two clusters of repetitive elements (i.e. a satellite DNA and an endogenous retrovirus) that generated random homoplasy, because they were abundant in *Pan* and *Gorilla* but almost absent in *Homo* and *Pongo*. The only phylogenetic tree congruent with the accepted topology for hominids, thus coinciding with that obtained from the mitogenomes of the same individuals, was the one built after filtering out the libraries for the two homoplasious clusters and using three individuals per species. Our results suggest some caution in the use of repeat abundance for phylogenetic studies, as some element abundances are homoplasious, which severely distorts the phylogenetic signal due to their differential amplification among evolutionary lineages.

Keywords: *Hominidae, homoplasy, inter-individual variation, phylogenetics, repetitive DNA*

Introduction

With the rise of high-throughput sequencing technologies, there has been an intersection between previously disparate fields of cytogenetics/genomics and phylogenetics. There are many approaches that seek to use genome-scale data for phylogenetic inference (often termed 'phylogenomics'), that usually aim to reduce the genome complexity to something manageable for phylogenetic purposes. Additionally, such data are very useful for characterizing repeats and other markers for efficiently producing cytogenetic probes. The simplest method, is one of 'genome skimming' sensu Straub et al. (2012), whereby whole-genome shotgun sequencing is performed but at a very low depth of coverage (less than 1X coverage and perhaps less than 0.1X). These datasets consist primarily of those sequences that are in high abundance, either in the genome itself or within the organism; this includes, predominantly, the high-copy organellar genome sequences (mitogenome, plastome – in plants) but also those sequences that are in high-copy in the nuclear genome. Amongst these high-copy nuclear sequences are mainly repetitive elements, an array of different types of repeat sequences, that include satellite (tandem) repeats, and transposable elements (TEs) such as retroelements (Class I TEs) and DNA transposons (Class II TEs). Often these data are discarded by researchers focusing on phylogenetics with such datasets, using instead only the reconstructed organellar genomes (e.g. Guschanski et al., 2013; Richter et al., 2015; Timmermans et al., 2016; Ren et al., 2017).

The importance of repetitive DNA abundance as a marker for the phylogenetic history of species has been increasingly explored (e.g. Sveinsson et al., 2013; Ricci et al., 2013; Cai et al., 2014). Several recent studies have shown that genomic repeat abundance, rather than the sequence itself, can be used as an informative character for phylogenetic inference (Novák et al., 2014; Dodsworth et al., 2015, 2016a, 2016b; Usai et al., 2017). Utilizing a recently developed pipeline for de novo repeat analysis from low-coverage sequence data, RepeatExplorer (Novák, Neumann & Macas, 2010; Novák et al., 2013), a high number of clusters are generated, each representing a putatively homologous repeat family/class. Within each cluster or element, the sequence divergence is low, and although this can be used for fine-scale classification of element types, particularly retroelements (e.g. Piednoël et al., 2014; Mascagni et al., 2015; Harkess et al., 2016; Tetreault & Ungerer, 2016), the sequence divergence is not sufficient to infer taxon

relationships. The abundance of homologous repeats does differ, however, and the abundance of elements is often indicative of evolutionary relatedness, i.e. phylogeny (e.g. in bananas – Novák et al., 2014; angiosperms and Drosophila – Dodsworth et al., 2015; in poplars – Usai et al., 2017). In some cases, however, this is not entirely clear-cut, due to the activity of some elements, particularly those in high abundance, that is more reflective of recent activity or, perhaps, differential processes of elimination from the genome (Pons et al., 2004; Ribeiro et al., 2017; Ustyantsev, Blinov & Smyshlyaev, 2017). This needs to be explored and tested in cases where the topology is "known", such that particular element histories can be teased apart and their impact on overall phylogenetic signal investigated.

Here we decided to test the abundance of repeats as adequate phylogenetic characters, particularly exploring the homoplasy of repeats, using the hominids as a case study. This group was selected due to the widely-accepted phylogenetic hypothesis based on much previous research and genome-scale data. Specifically, we set out to answer the following questions in this study:

1) Is the phylogenetic signal of genomic repeat abundance reliable in the case of the hominids?

2) Do certain clusters/repeats with homoplasious abundances adversely affect the phylogenetic signal?

3) Is one individual per taxon enough to build a reliable phylogenetic tree from genomic repeat abundances?

Materials and methods

Data acquisition

We downloaded high-throughput sequence data from 15 NCBI Short Read Archive (SRA) accessions, including Illumina reads (Illumina Inc., San Diego, CA) from three individuals belonging to five of the well-known species of the Hominidae family of primates (Table App.1): *Homo sapiens*, *Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla* and *Pongo pygmaeus*. We also downloaded Illumina read data from a *Macaca mulatta* individual to be used as an outgroup for phylogenetic analyses.

In order to avoid data biases based on different sequencing protocols all reads used in this study were chosen because they had been obtained under the same sequencing

platform (Illumina HiSeq 2000), thus yielding reads of 100 bp in length, except for the *M. mulatta* library in which Illumina read length was 101 bp. Chimpanzee, bonobo, gorilla and orangutan data were acquired from wild-born individuals sequenced within the same SRA BioProject (PRJNA189439; IBE CSIC-Universitat Pompeu Fabra; Prado-Martínez et al., 2013) whereas human short reads belong to the 1000 Genomes Project Phase 3 (PRJNA262923).

Table App.1. Taxon sampling of hominids from NCBI SRA.

Phylogeny ID	SRA RUN ID	BioSample ID	Geographic origin
HSAP1	ERR068394	SAMN00263022	Iberian populations in Spain
HSAP2	ERR050125	SAMN00014366	Iberian populations in Spain
HSAP3	ERR050124	SAMN00014365	Iberian populations in Spain
PTRO1	SRR748072	SAMN01920536	Gabon
PTRO2	SRR748062	SAMN01920534	Gabon
PTRO3	SRR748058	SAMN01920533	Gabon
PPAN1	SRR740802	SAMN01920509	Democratic Republic of Congo
PPAN2	SRR740794	SAMN01920508	Democratic Republic of Congo
PPAN3	SRR740768	SAMN01920506	Democratic Republic of Congo
GGOR1	SRR748092	SAMN01920490	Western lowland
GGOR2	SRR748096	SAMN01920491	Western lowland
GGOR3	SRR748097	SAMN01920492	Western lowland
PPYG1	SRR748020	SAMN01920551	Bornean
PPYG2	SRR748000	SAMN01920547	Bornean
PPYG3	SRR748004	SAMN01920548	Bornean
MMUL1	SRR1944168	SAMN03264679	Indian breed

Mitochondrial genome assembly, phylogenetic analysis and filtering

A total of 5,000,000 100/101-bp raw Illumina read pairs were randomly selected using the SeqTK software (https://github.com/lh3/seqtk) from each library downloaded from the SRA and were used for mitochondrial genome assembly with MITObim v1.8 (Hahn, Bachmann & Chevreux, 2013). The mitochondrial genomes used as reference for assembly are indicated in Table App.2 and were downloaded from NCBI GenBank reference sequences. Genome annotation was performed in GENEIOUS v4.8.5 (Biomatters

Ltd., Auckland, New Zealand) by aligning with the reference mitochondrial genome of each species. To verify its phylogenetic identity, a phylogenetic tree was built based on maximum parsimony analysis of a global alignment of the whole newly assembled mitochondrial genome of each individual included in this study. The Tree Analysis Using New Technology (TNT) software for Linux 64 (no taxon limit), updated version of 11/Dec/13 (Goloboff, Farris & Nixon, 2008), was used for phylogenetic reconstruction, using implicit enumeration. Prior to subsequent analyses of repetitive DNA abundance, all Illumina libraries were filtered out for mitochondrial DNA with the software DeconSeq v0.4.3 (Schmieder & Edwards, 2011), using as reference the mitochondrial genome for each species found in Table App.2.

Table App.2. Mitochondrial genome reference sequences

Taxa	NCBI reference sequence accession
Homo sapiens	NC_012920.1
Pan troglodytes	NC_001643.1
Pan paniscus	NC_001644.1
Gorilla gorilla	NC_001645.1
Pongo pygmaeus	NC_001646.1
Macaca mulatta	NC_005943.1

Preparation of read data for repeat analyses

SRA files were unpacked into FASTQ using the FASTQ-DUMP tool from the SRA Toolkit. Low-quality reads in FASTQ files were discarded using Trimmomatic (Bolger, Lohse & Usadel, 2014), by removing adapters and selecting read pairs with all their nucleotides with Q > 30, using the options "ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:30 MINLEN:[100/101]".

All samples were assumed to have a genome size of about 3.5 Gbp, based on data available in the Animal Genome Size Database, which showed only slight variation in genome size between the species used in this study (3.47–3.85 Gbp) (http://www.genomesize.com/ last accessed 12/11/2016), which is considered appropriate for this kind of study (Dodsworth et al., 2016a). Each accession was then sampled for 0.6% of the genome by randomly subsampling each Illumina dataset. This resulted in 200,000 reads per sample from all Hominidae accessions, randomly selected

with SeqTK and then converted into FASTA format.

Selected reads from each sample were labeled with a unique five-character prefix, making a total combined dataset of 1,200,000 reads for datasets of one individual per species, 2,200,000 reads for datasets of two individuals per species and 3,200,000 reads for the global dataset including all individual samples. Specifically, we prepared three different datasets of one individual (library or sample) per species plus *M. mulatta* as an outgroup (6 OTUs per dataset), three different datasets of two biological individuals per species plus *M. mulatta* as an outgroup (11 OTUs per dataset) and one dataset grouping together all libraries representing three biological individuals per species making a total of 16 OTUs for phylogenetic analysis, as shown in Table App.3.

Table App.3. Read sampling for repetitive DNA clustering and phylogenetic analyses

Dataset	Individuals per species	HSAP	PTRO	PPAN	GGOR	PPYG	MMUL	Total reads
1	1	HSAP1	PTRO1	PPAN1	GGOR1	PPYG1	MMUL1	1,200,000
2	1	HSAP2	PTRO2	PPAN2	GGOR2	PPYG2	MMUL1	1,200,000
3	1	HSAP3	PTRO3	PPAN3	GGOR3	PPYG3	MMUL1	1,200,000
4	2	HSAP1 HSAP2	PTRO1 PTRO2	PPAN1 PPAN2	GGOR1 GGOR2	PPYG1 PPYG2	MMUL1	2,200,000
5	2	HSAP1 HSAP3	PTRO1 PTRO3	PPAN1 PPAN3	GGOR1 GGOR3	PPYG1 PPYG3	MMUL1	2,200,000
6	2	HSAP2 HSAP3	PTRO2 PTRO3	PPAN2 PPAN3	GGOR2 GGOR3	PPYG2 PPYG3	MMUL1	2,200,000
7	3	HSAP1 HSAP2 HSAP3	PTRO1 PTRO2 PTRO3	PPAN1 PPAN2 PPAN3	GGOR1 GGOR2 GGOR3	PPYG1 PPYG2 PPYG3	MMUL1	3,200,000

RepeatExplorer (RE) clustering of samples

Clustering of Illumina reads was performed using the RE pipeline, implemented in a GALAXY server environment running locally in the University of Granada. RE clustering was used to identify genomic repeat clusters within each dataset, with default settings (minimum overlap = 55, cluster size threshold for detailed analysis = 0.01% and the "all reads are paired" option selected). For additional details about the clustering algorithm see Nóvak et al., 2010 and 2013. For further identification of repeat clusters, we used a custom repeat database of all primate repetitive DNA annotations included in RepBase (Bao, Kojima & Kohany, 2015) (http://www.girinst.org/repbase/ last accessed

20/11/2016). Following Dodsworth et al., 2016a, we used the 1,000 most abundant repeat clusters, as they represented enough of a proportion of the genome, for phylogenetic analyses. Read counts per cluster and sample information obtained from RE can be found in figshare under the accession https://figshare.com/s/c2ccda047dd502890dcb.

Phylogenetic analysis of clusters

The 1,000 most abundant clusters of each dataset were used to create the data matrices for phylogenetic inference. TNT software was chosen for phylogenetic analyses under the maximum parsimony principle (Goloboff & Mattoni, 2006; Goloboff et al., 2008). Cluster abundances were used as input (continuous characters). To make the cluster abundance values suitable as input for the TNT software, we divided all abundances by a factor calculated by dividing the abundance of the most abundant cluster by 65, so that all data would fall within the 0–65 range (with up to three decimals) as needed for continuous characters analysis with TNT. Further transformations (e.g. cubed root) were checked but provided no improvement on the factorial transformation. Implicit enumeration (branch and bound) tree searches were used for datasets in this study owing to the small number of taxa in each dataset. Resampling was performed using 10,000 replicates and symmetrical resampling was done by a modification of the standard bootstrap (Goloboff et al., 2003). FigTree v1.4.3 (http://tree.bio.ed.ac.uk/) was used for graphical view and representation of phylogenetic trees.

Filtering of disturbing clusters

After the first RE clustering, we found some clusters for satellite DNA and an endogenous retrovirus that were abundant in chimpanzee, bonobo and gorilla, but were absent in human and orangutan libraries. We identified these clusters by means of a Python script (https://github.com/mmarpe/phyl_rep_Hominidae/sel_clusters.py) that helped us to locate those clusters that had less than 25 reads in *Homo* and *Pongo* but that were abundant in the rest of hominid species. The identity of these clusters was confirmed by the RepeatExplorer annotation and further characterized by means of sequence homology search using BLASTn (Altschul et al., 1990) and CENSOR (Kohany et al., 2006) tools.

To test the effect of these clusters on the phylogenies built with the abundance of repeats, we performed two sets of phylogenetic analyses, one using unfiltered libraries and the other using libraries previously filtered out for these particular clusters. Filtering

was performed by DeconSeq software against the CL3 satellite consensus sequence (X74280.1 and X74281.1 GenBank accessions) (Royle, Baird & Jeffreys, 1994) and against the CERV1_INT, the internal sequence for the endogenous retrovirus (Skaletsky, Hughes & Page, 2004) included in RepBase.

Combinations of one or two individuals per species

Using a custom script, written in Python (https://github.com/mmarpe/phyl_rep_Hominidae/sample_mix.py), we phylogenetically analyzed all possible combinations of one or two individuals per taxon (243 phylogenetic trees each), with abundances obtained from a global RepeatExplorer run of all libraries involved in this study after the above filtering of clusters. The 1,000 most abundant clusters of each combination were phylogenetically analyzed by means of maximum parsimony implemented using TNT software as described previously. From the 1,000 top abundant cluster data obtained from the RE of all three individuals per species (all samples included in this paper) after filtering, this script constructs all possible cluster abundance datasets for all different abundances data combinations of two individuals per species or one single individual per species without samples repetitions, later it generates the trees derived from each dataset using the same parameters described above for the TNT software, and finally transforms the tree files from .nex format to .pdf format using FigTree to make their visualization more accessible.

The 243 trees produced from these combinations were grouped together in a file and, using Consense version 3.695 included in the PHYLIP package (Felsenstein 1989, 2005), we obtained the consensus tree for 2 individual/species cluster abundances combinations and for 1 individual/species combinations. This consensus tree consists of groups that occur as often as possible in the data through implementation of the Majority Rule (extended) method (Margush & McMorris, 1981).

Results

Mitogenome phylogenetic tree

In order to check the integrity and reliability of the libraries used, we assembled the full mitochondrial DNA sequence in each individual library, using MITObim v1.8, and built a mitochondrial phylogeny by means of maximum parsimony. This showed the absence of mis-tagging or sample confusion, since it coincided with the universally accepted Hominidae phylogeny (Roos & Zinner, 2017).

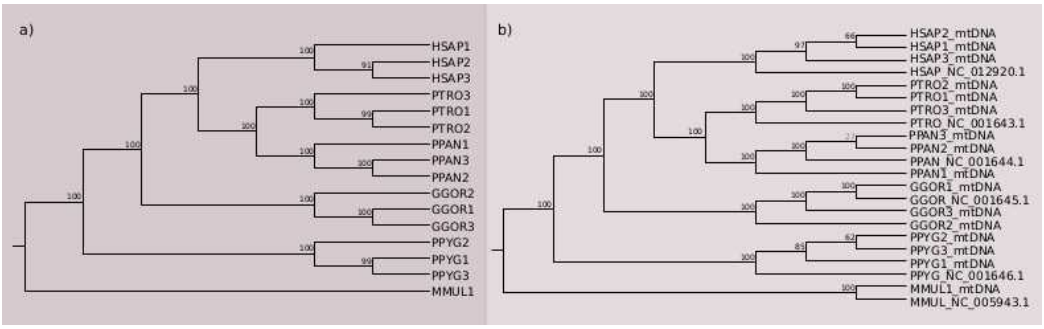


Figure App.1. Mitochondrial phylogeny of all samples (libraries) used in the present study (a) and with the reference mitogenome from each species used in the sample mtDNA assembly (b). In each case the trees represent the well-known (*Pongo* (*Gorilla* (*Pan* + *Homo*))) topology. Bootstrap support of each node is specified on the tree (values <50 in light gray indicate less robust nodes).

Phylogenetic analyses using unfiltered datasets

The first set of RE clustering and phylogenetic analyses was performed using the datasets indicated in Table App.3. None of the phylogenies obtained (Figure App.2) reflected the universally accepted phylogeny for the Hominidae family confirmed by the mitogenome phylogeny depicted above (Figure App.1). In all cases, *Homo sapiens* appeared in a basal position in the phylogeny and sometimes forming a clade with *Pongo pygmaeu*s (Figure App.2c-f). As we noticed that the topology of most trees shown in Figure App.2 supported the hypothesis of a *Pan*/*Gorilla* clade, we searched for clusters showing extremely high abundance similarity between humans and orangutans, which could be responsible for the observed phylogenetic distortion. For this purpose, we searched for clusters showing less than 25 reads in *Homo* and *Pongo* but showing higher abundance in *Pan* and *Gorilla*, using a custom script.

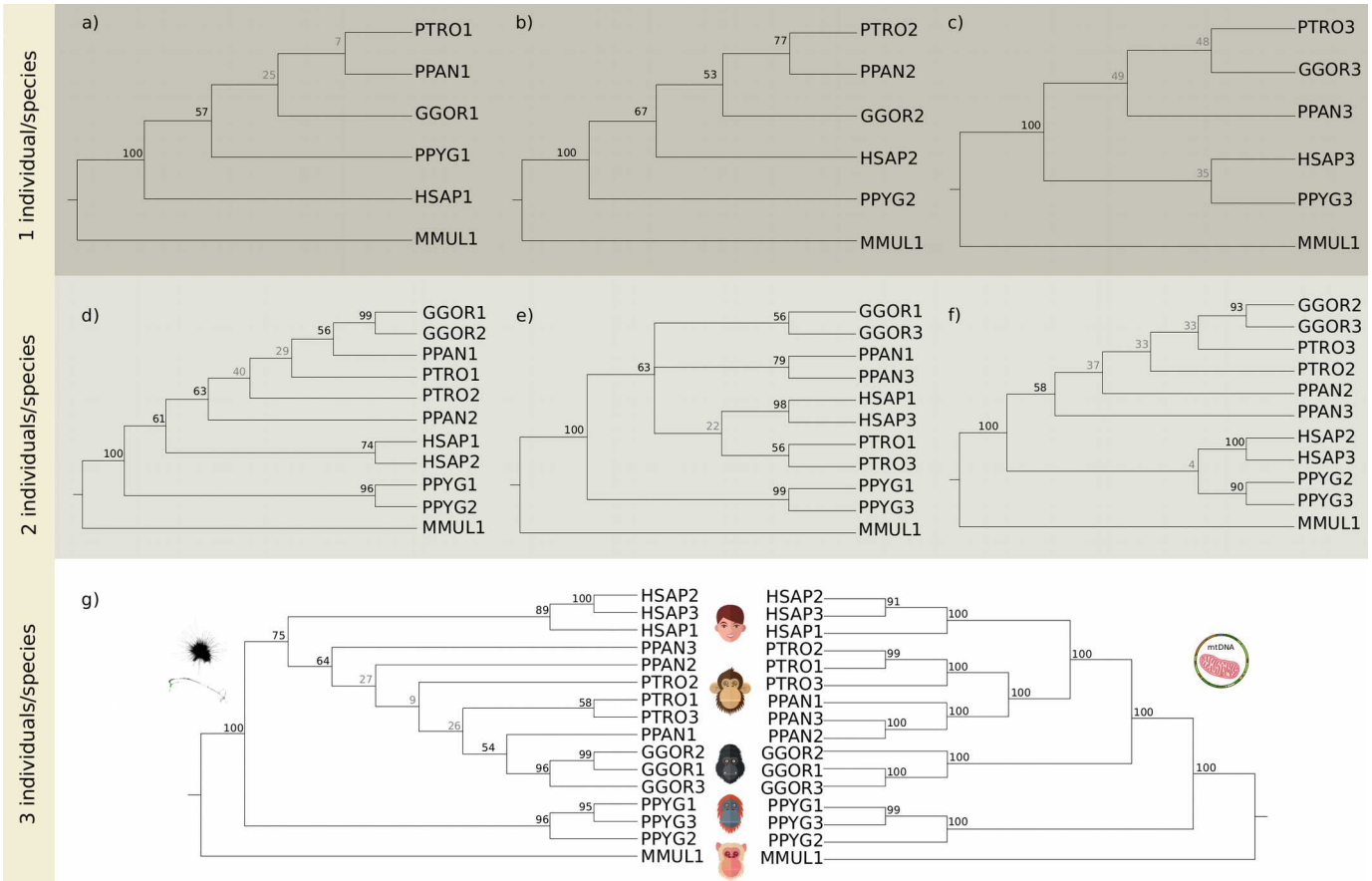


Figure App.2. Genomic repeat phylogenies of one (a, b, c), two (d, e, f) and all samples (g) after RE clustering of unfiltered libraries. Bootstrap support of each node is specified on the tree (values <50 in light gray indicate less robust nodes). Note that none of the trees match the topology of the mtDNA tree. Even with three individuals per species (g), the tree reconstructed using repetitive element abundances (on the left) placed *Homo* as the ancestor of *Pan* and *Gorilla*, in strong disagreement with the mtDNA tree (and current accepted placement).

Phylogenetic analyses using filtered datasets

We found two repetitive DNA elements, which were practically absent in *Homo sapiens* and *Pongo pygmaeus* (< 25 reads) but were abundant in *Pan* and *Gorilla* (Figure App.3a). These clusters were identified as a subterminal satellite repeat and an endogenous retrovirus (Figures App.3b, c). The repeat unit of the CL3 satellite is 32 bp long; it was isolated from the chimpanzee genome, found to be even more abundant in gorillas, but not detected in humans or orangutans (Royle et al., 1994). The endogenous retrovirus, CERV1/PTERV1, was found by means of the analysis of BAC chimpanzee genome sequences. It is integrated in the germline of African great ape and Old World monkey species but is absent from humans and Asian ape genomes (Yohn et al., 2005; Polavarapu, Bowen & McDonald, 2006). To evaluate the possible effect of these two repeats on the phylogenetic signal, we filtered these repeats out of all libraries and performed a new batch of phylogenetic analyses on the same datasets described in Table App.3, following the same protocol after filtering. As shown in Figure App.3c, the endogenous retrovirus was partially clustered in CL140 (cluster graphs of full ERVs should have a circular shape).

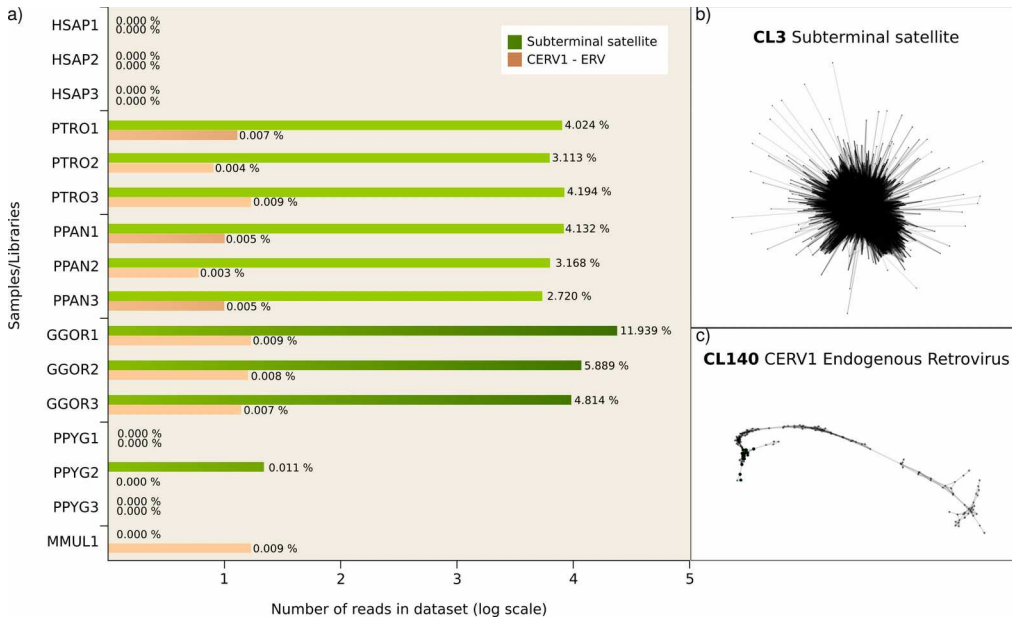


Figure App.3. a) Abundance of the CL3 subterminal satellite and the CERV1-ERV (CL140) retrovirus per individual. Number of reads as log-scaled bars and percentages shown next to bars indicate the proportion of each element per sample in the RE dataset. b-c) Graph-clusters of the two homoplasious repeats, CL3 satellite and CL140 CERV1 retrovirus.

We found some other clusters containing part of this ERV but they were less abundant and they were not discarded after the use of the script for filtering ERV reads, and we decided to include these small clusters in subsequent phylogenetic analyses as their presence did not influence the phylogenetic signal of the dataset as a whole. In addition, homoplasious clusters were filtered out of libraries using full reference sequences from RepBase, which means that the number of retained reads matching those repetitive elements is very low after filtering.

The phylogenies obtained (Figure App.4) failed to show the previous close relationship between *Homo* and *Pongo*, indicating that the discarded repeats were actually responsible for the distortion of the phylogenetic signal shown in the first set of analyses. In fact, the tree built with three individuals per species yielded a tree (Fig. App.4g) with essentially the same topology as the mitogenome tree, albeit with low node support in places.

This result demonstrates that some repeats can generate "random homoplasy" by differential amplification among different evolutionary lineages. In the present datasets, a satellite DNA and a retrovirus became highly abundant in the *Pan* and *Gorilla* lineages, whereas they did not prosper in the *Homo* and *Pongo* lineages, for which reason the two latter species showed a homoplasious rather than real phylogenetic relationship. This may present a serious problem for using the abundance of repeats for phylogenetic analysis in groups not as well-known as the hominids.

One or two individuals per species can yield poor phylogenetic trees

As shown in Figure App.4g, the phylogeny built with three individuals per species was very similar to that obtained with the mitogenomes, when the two homoplasy-generating repeats were filtered out from the libraries. However, trees built with one or two individuals per species were still better than those performed by the unfiltered libraries, because *Pongo* was ancestral with respect to *Gorilla*, *Pan* and *Homo*, but they did not resolve properly the phylogenetic relationships between the three latter taxa (see Fig. App.4a-f) as all these topologies show an unsolved *Homo*/*Pan*/*Gorilla* clade. According to the phylogenetic analysis of technical replicates (15 technical replicates, one per each biological sample used in this study, outgroup excluded), this issue of resolution may be due to inter-individual variation rather than sequencing bias (see supplementary material for technical replicates analysis).

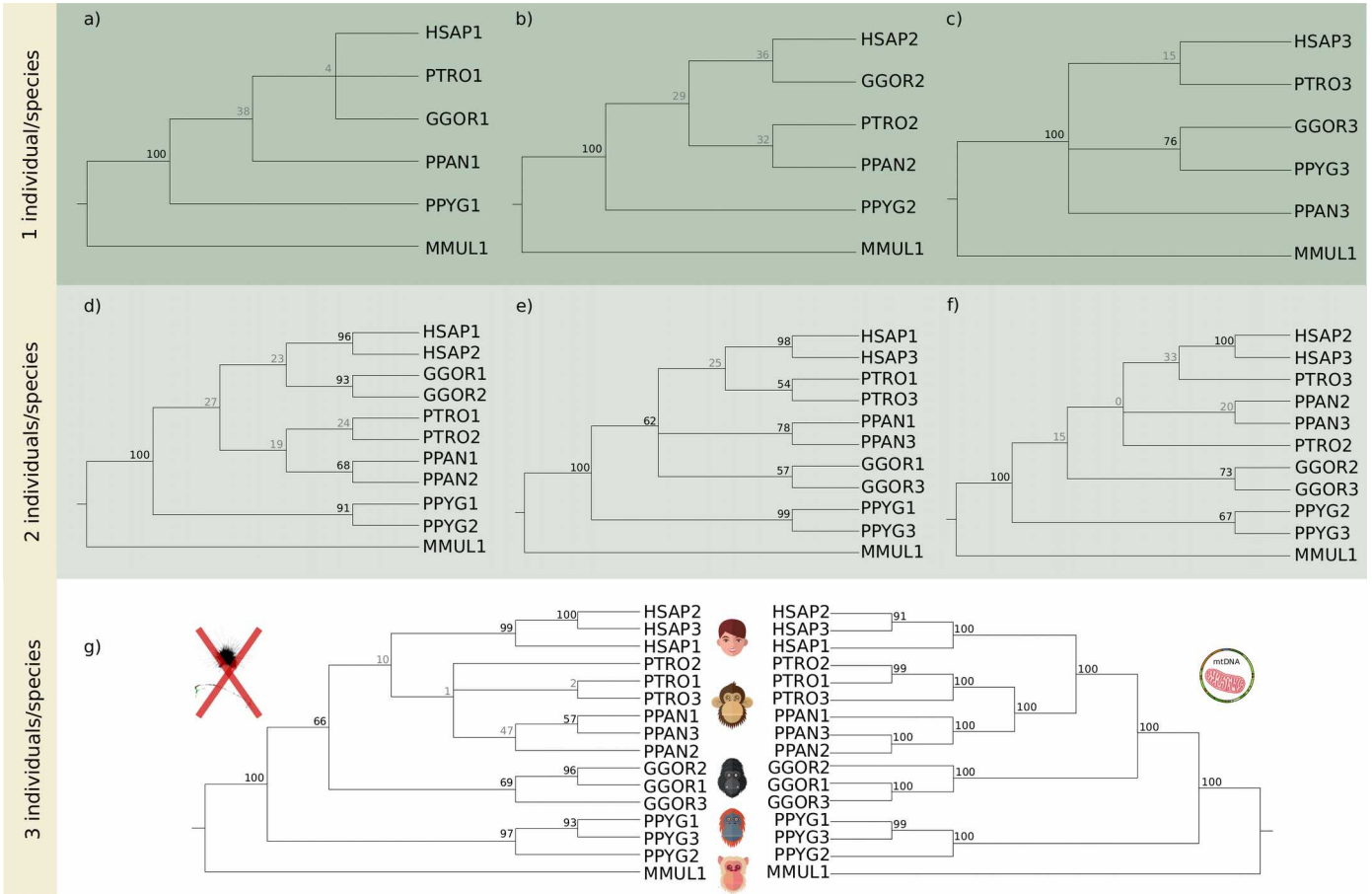


Figure App.4. Genomic repeat phylogenies of one (a, b, c), two (d, e, f) and all samples (g) after RE clustering of libraries previously filtered out for a satellite DNA (CL3) and an endogenous retrovirus (CL140 – CERV1) that have homoplasious abundance distributions. Bootstrap support for each node is specified on the tree (values <50 in light gray indicate less robust nodes).

To evaluate the effect of inter-individual (coincident with intraspecific in this case) variation in repeat abundance on phylogenetic reconstruction, we made all possible combinations of one or two individuals per species, chosen from the matrix of abundances obtained after RE clustering of the dataset including all three filtered libraries per species. We thus performed the phylogenetic inference for each combination, producing 243 trees for the combinations of one individual and 243 trees for the combinations of two individuals per species. The results showed that the consensus tree for the combinations of one individual did not reflect the phylogeny of the mitogenome (Fig. App.5a), even though 36 trees out of the set of 243 did. However, the consensus tree obtained from the combinations of two individuals clearly represented the phylogenetic relationships universally accepted for the Hominidae (Figure App.5b), even though only 16 trees out of the 243 showed the resolved and accepted topology.

We conclude that the phylogenetic inference obtained from genomic repeat abundance is highly dependent on inter-individual variation, and the use of only one or two individuals per taxon may potentially lead, with high probability [(243-36)/243= 0.85 with N=1 and (243-16)/243= 0.93 with N=2], to wrong phylogenetic inferences, at least in the case of the Hominidae family.

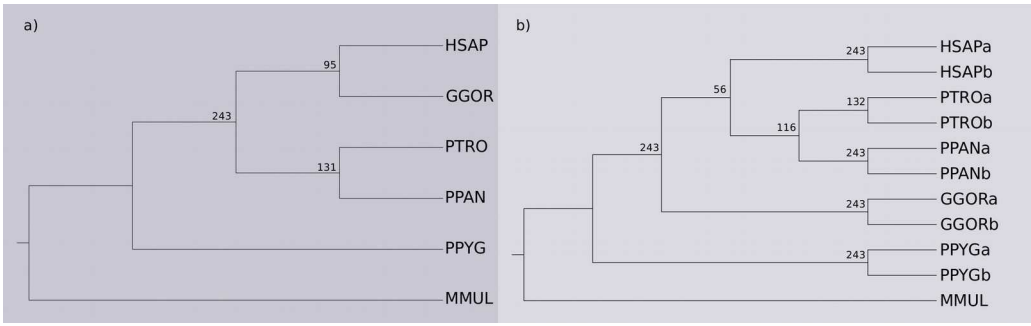


Figure App.5. Consensus phylogenetic trees obtained from all possible combinations of one (a) and two (b) individuals per species (after filtering of the two homoplasious repeats). Numbers beside nodes indicate the number of trees, out of 243, that support the split. Note that the consensus tree built with two samples per taxon (b) shows a similar topology to the mtDNA tree shown in Fig. App.1, albeit with low support for two nodes.

Discussion

Phylogeny of Hominidae using repeat abundance

The phylogenetic relationships of the Hominidae family have been the object of study and great interest for the scientific community for a long time and they have not been exempt of controversy (Holmquist, Miyamoto & Goodman, 1988; Dean & Delson, 1992; Grehan & Schwartz, 2009; Grehan & Schwartz, 2009). Currently, the (*Pongo* (*Gorilla* (*Pan* + *Homo*))) evolutionary reconstruction is universally accepted and well-established (Purvis, 1995; Arnason et al., 2000; Arnold et al., 2010; Perelman et al., 2011; Popadin et al., 2017), so we believe that it is an appropriate model to test the method of phylogenetic estimation from the abundance of genomic repeats. We compared our results with the reference tree built by mitogenomes (Fig. App.1), which agrees with the previously accepted topology for this group (chromosomal evidence – Seuánez, 1982; morphological data - Ciochon et al., 1983; identity of the alpha and beta hemoglobin sequences – Goodman et al., 1983; using DNA-DNA hybridization values – Sibley & Ahlquist, 1984; mitochondrial DNA analyses – Hayasaka, Gojobori & Horai, 1988; beta-globin gene clusters study – Koop et al., 1989). Our results show that there is phylogenetic signal present in repeat abundances, for the top 1,000 most abundant repetitive elements in the hominid nuclear genomes (Figs. App.2-4). Generally, we recovered phylogenetic hypotheses close to the accepted tree topology indicated above. However, this was only after adding more than one individual per species, and after filtering out two particular repeats that had high abundance but not in closely related taxa, therefore distorting the phylogenetic inference (Fig. App.4). The most acceptable phylogeny was inferred when making a consensus of all possible combinations of two-taxon datasets (Fig. App.5b) after RE clustering of three individuals per species and filtering out the two clusters causing homoplasy. Even then, some nodes are not well-supported according to bootstrapping, which underlies a lack of phylogenetic signal of repeat abundances for some parts of the tree.

Inter-individual variation affects phylogenetic inference

The abundance of repetitive elements appears to show high variation between individuals, so that ideally two or more individuals per species should be used for phylogenetic analysis based on repeat abundance (Fig. App.4). The most unsatisfactory phylogenetic trees we generated were from the datasets that included only one

individual per taxon (Figs. App.2-4), in which *Homo* is either misplaced or the tree is generally unresolved with respect to other hominids. This did not vastly improve even after filtering of clusters with homoplasious distributions (Fig. App.4) suggesting that whilst this may eliminate the issue of (some) homoplasy, it does not negate the caveat of inter-individual variation in repeat abundance.

Homoplasious repeats obscure true phylogenetic signal

A phylogenetic hypothesis reflecting the currently accepted Hominidae phylogeny was obtained only using two or three individuals per taxon when their libraries were filtered out for the “disturbing” clusters of repetitive DNA (Fig. App.4g): a satellite DNA and an endogenous retrovirus, which showed differences in abundance between closely related species (e.g. *Homo* and *Pan*). These repetitive elements thus distorted the phylogenetic signal yielding a falsely close relationship between *Homo* and *Pongo*. Removing these sequences from the libraries substantially improves the phylogenies obtained (Fig. App.4). We believe this is a case of "random homoplasy" generated by the chance amplification of the satellite DNA (and spread of the retrovirus) in *Pan* and *Gorilla* but not in *Homo*, which makes the latter more similar to *Pongo* in this respect. As Fig. App.3 shows, the homoplasious satellite DNA was the third repetitive element in order of decreasing abundance in *Pan* (2.7-4.2%) and *Gorilla* (4.8-11.9%), such that its influence on phylogenetic signal appears to be logical. However, the endogenous retrovirus was only the 140th cluster most-abundant cluster (0.003-0.009% in *Pan* and 0.007-0.009% in *Gorilla*), but the trees built with this repeat included failed to fit the accepted phylogeny even after filtering out the abundant satellite (data not shown). This poses a serious problem for phylogenetic reconstruction through this approach as not only the most-abundant repeats can distort the phylogenetic signal but also others that show much lower abundance in the genomes.

Methods of phylogenetic inference that adequately handle continuous data as phylogenetic characters are currently limited but could be improved upon (e.g. model-based solutions) and in this case would aid phylogenetic inference from repeat abundance data. The maximum parsimony (MP) algorithm implemented in TNT is similar to ordinary MP, and therefore homoplasious repeats with large differences in abundance (such as those two identified for hominids) have an adversely large effect on tree length and therefore the most parsimonious phylogenetic tree that is reconstructed. This effect can sometimes be minimized by the use of different transformations on the data matrix,

in order to make the abundances between 0-65. For example, square root or other root transformations retain the abundance differences between taxa but minimize the overall abundance (length) differences for any particular cluster (character), as used in Dodsworth et al., 2016b and tested here as well (data not shown). However, these approaches do not alleviate the problem in the worst cases, as is the one showed in the present study for the Hominidae family, and it is advised that these clusters (repeats) are identified and removed from the dataset prior to phylogenetic inference. In cases without previous knowledge of phylogenetic relationships for the taxa involved, discarding every cluster showing large differential abundances that might be homoplasious, i.e. being absent or present in only two taxa, could be an option. We tried to do this for the present dataset, but it eliminated some clusters that were important for grouping the two *Pan* species, as they included repeats specific to that clade of two species (data not shown). More adequate model-based methods for inferring the phylogeny would also help to overcome the homoplasious nature of some repeat types, but these methods require further development. Therefore, the homoplasy problem might not be easy to solve, as repetitive DNA rarely shows a static path along the tree of life (Kuhn GC et al., 2008; Feliciello et al., 2014; Rojo et al., 2015; Barghini et al., 2015; Ferreira de Carvalho et al., 2016).

Conclusions

Here we tested the abundance of repetitive elements as phylogenetic characters to infer the phylogenetic relationships of hominid primates, the family Hominidae. In general, we were able to recover a phylogenetic hypothesis close to the accepted topology, i.e. that which was recovered from much previous genomic sequence data. We discovered two important caveats when exploring this type of data, that should be borne in mind for future analyses of repeat abundances as phylogenetic characters: (i) individual variation in repeat abundance suggests that multiple samples per taxon should be included if at all possible, and (ii) particular repeats can have highly homoplasious distributions such that they distort the phylogenetic signal in the overall dataset. We suggest that without *a priori* knowledge of the expected phylogenetic topology, researchers are cautious and check for unusual signals yielded by repetitive elements irregularly distributed in the genomes of the tested organisms.

Data accessibility

Icons appearing in the phylogenetic trees were freely download from www.freepik.es and designed by the own website developers. All data matrices, raw RE cluster abundances and processed input matrices, and phylogenetic trees built in the present study can be found in figshare under the accession URL https://figshare.com/s/c2ccda047dd502890dcb. All the scripts used are available from https://github.com/mmarpe/phyl_rep_Hominidae.

References

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

Arnason U, Gullberg A, Burguete AS, Janke A. (2000). Molecular estimates of primate divergences and new hypotheses for primate dispersal and the origin of modern humans. *Hereditas*, 133(3), 217–228.

Arnold C, Matthews LJ, Nunn CL. (2010). The 10kTrees website: A new online resource for primate phylogeny. *Evolutionary Anthropology*, 19, 114–118.

Bao W, Kojima KK, Kohany O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6(1), 11.

Barghini E, Natali L, Giordani T, Cossu RM, Scalabrin S, Cattonaro F, et al. (2015). LTR retrotransposon dynamics in the evolution of the olive (*Olea europaea*) genome. *DNA Research*, 22(1), 91–100.

Bolger AM, Lohse M, Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120.

Cai Z, Liu H, He Q, Pu M, Chen J, Lai J, et al. (2014). Differential genome evolution and speciation of *Coix lacryma-jobi* L. and *Coix aquatica* Roxb. hybrid guangxi revealed by repetitive sequence analysis and fine karyotyping. *BMC Genomics*, 15(1), 1025.

Ciochon RL. (1983). Hominoid cladistics and the ancestry of modern apes and humans. In: Corruccini RS, Ciochon RL, ed. *New interpretations of ape and human ancestry*. New Yor:K Academic Press, 783–843.

Dean D, Delson E. (1992). Second gorilla or third chimp? *Nature*, 359(6397), 676–677.

Dodsworth S, Chase MW, Kelly LJ, Leitch IJ, Macas J, Novák P, et al. (2015). Genomic repeat abundances contain phylogenetic signal. *Systematic Biology*, 64(1), 112–126.

Dodsworth S, Chase MW, Särkinen T, Knapp S, Leitch AR. (2016a). Using genomic repeats for phylogenomics: a case study in wild tomatoes (*Solanum* section Lycopersicon: Solanaceae). *Biological Journal Linnean Society*, 117(1), 96–105.

Dodsworth S, Jang TS, Struebig M, Chase MW, Weiss-Schneeweiss H, Leitch AR. (2016b). Genome-wide repeat dynamics reflect phylogenetic distance in closely related allotetraploid *Nicotiana* (Solanaceae). *Plant Systematics and Evolution*, 33(8), 1013–1020.

Feliciello I, Akrap I, Brajković J, Zlatar I, Ugarković Đ. (2014). Satellite DNA as a driver of population divergence in the red flour beetle *Tribolium castaneum*. *Genome Biology and Evolution*, 7(1), 228–239.

Felsenstein J. (1989). PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics*, 5, 164–166.

Felsenstein J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Seattle: Department of Genome Sciences, University of Washington.

Ferreira de Carvalho J, de Jager V, van Gurp TP, Wagemaker NC, Verhoeven KJ. (2016). Recent and dynamic transposable elements contribute to genomic divergence under asexuality. *BMC Genomics*, 17(1), 884.

Goloboff PA, Farris JS, Kallersjo M, Oxelman B, Ramirez MJ, Szumik CA. (2003). Improvements to resampling measures of group support. *Cladistics*, 19, 324–332.

Goloboff PA, Mattoni CI. (2006). Continuous characters analyzed as such. *Cladistics*, 22, 589–601.

Goloboff PA, Farris JS, Nixon KC. (2008). TNT, a free program for phylogenetic analysis. *Cladistics*, 24, 774–786.

Goodman M, Braunitzer G, Stangl A, Schrank B. (1983). Evidence on human origins from haemoglobins of African apes. *Nature*, 303(5917), 546–548.

Grehan JR, Schwartz JH. (2009). Evolution of the second orangutan: phylogeny and biogeography of hominid origins. *Journal of Biogeography*, 36, 1823–1844.

Grehan JR, Schwartz JH. (2011). Evolution of human-ape relationships remains open for investigation. *Journal of Biogeography*, 38, 2397–2404.

Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, Finstermeier K, et al. (2013). Next-generation museomics disentangles one of the largest primate radiations. *Systematic Biology*, 62(4), 539–554.

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(13), e129.

Harkess A, Mercati F, Abbate L, McKain M, Pires JC, Sala T, et al. (2016). Retrotransposon proliferation coincident with the evolution of dioecy in *Asparagus*. *G3 (Bethesda)*, 6(9), 2679–2685.

Hayasaka K, Gojobori T, Horai S. (1988). Molecular phylogeny and evolution of primate mitochondrial DNA. *Molecular Biology and Evolution*, 5(6), 626–644.

Holmquist R, Miyamoto MM, Goodman M. (1988). Higher-primate phylogeny--why can't we decide? *Molecular Biology and Evolution*, 5(3), 201–216.

Kohany O, Gentles AJ, Hankus L, Jurka J. (2006). Annotation, submission and screening of repetitive elements in Repbase: Repbase Submitter and Censor. *BMC Bioinformatics*, 7, 474.

Koop BF, Tagle DA, Goodman M, Slightom JL. (1989). A molecular view of primate phylogeny and important systematic and evolutionary questions. *Molecular Biology and Evolution,* 6(6), 580–612.

Kuhn GC, Sene FM, Moreira-Filho O, Schwarzacher T, Heslop-Harrison JS. (2008). Sequence analysis, chromosomal distribution and long–range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the *Drosophila buzzatii* cluster. *Chromosome Research*,

16(2), 307–324.

Margush T, McMorris FR. (1981). Consensus n-trees. *Bulletin of Mathematical Biology*, 43, 239–244.

Mascagni F, Barghini E, Giordani T, Rieseberg LH, Cavallini A, Natali L. (2015). Repetitive DNA and plant domestication: Variation in copy number and proximity to genes of LTR-retrotransposons among wild and cultivated sunflower (*Helianthus annuus*) genotypes. *Genome Biology and Evolution*, 7(12), 3368–3382.

Novák P, Neumann P, Macas J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, 11, 378.

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.

Novák P, Hřibová E, Neumann P, Koblížková A, Doležel J, Macas J. (2014). Genome-wide analysis of repeat diversity across the family Musaceae. *PLoS One*, 9(6), e98918.

Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MA, et al. (2011). A molecular phylogeny of living primates. *PLoS Genetics*, 7(3), e1001342.

Polavarapu N, Bowen NJ, McDonald JF. (2006). Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses. *Genome Biology*, 7(6), R51.

Piednoël M, Carrete-Vega G, Renner SS. (2013). Characterization of the LTR retrotransposon repertoire of a plant clade of six diploid and one tetraploid species. *Plant Journal*, 75(4), 699–709.

Pons J, Bruvo B, Petitpierre E, Plohl M, Ugarkovic D, Juan C. (2004). Complex structural features of satellite DNA sequences in the genus *Pimelia* (Coleoptera: Tenebrionidae): random differential amplification from a common 'satellite DNA library'. *Heredity (Edinb)*, 92(5), 418–427.

Popadin K, Gunbin K, Peshkin L, Annis S, Kraytsberg G, Markuzon N, et al. (2017). Mitochondrial pseudogenes suggest repeated interspecies hybridization in hominid evolution. *BioRxiv*, 134502.

Prado-Martínez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdós B, et al. (2013). Great ape genetic diversity and population history. *Nature*, 499(7459), 471–475.

Purvis A. (1995). A composite estimate of primate phylogeny. *Philosophical Transaction of the Royal Society of London, Series B, Biological Science*, 348(1326), 405–421.

Ren Z, Harris AJ, Dikow RB, Ma E, Zhong Y, Wen J. (2017). Another look at the phylogenetic relationships and intercontinental biogeography of eastern Asian – North American *Rhus* gall aphids (Hemiptera: Aphididae: Eriosomatinae): Evidence from mitogenome sequences via genome skimming. *Molecular Phylogenetics and Evolution*, 117, 102–110.

Ribeiro T, Dos Santos KG, Richard MM, Sévignac M, Thareau V, Geffroy V, et al. (2017). Evolutionary dynamics of satellite DNA repeats from *Phaseolus* beans. *Protoplasma*, 254(2), 791–801.

Ricci M, Luchetti A, Bonandin L, Mantovani B. (2013). Random DNA libraries from three species of the stick insect genus *Bacillus* (Insecta: Phasmida): repetitive DNA characterization and first observation of polyneopteran MITEs. *Genome*, 56(12), 729–735.

Richter S, Schwarz F, Hering L, Böggemann M, Bleidorn C. (2015). The utility of genome

skimming for phylogenomic analyses as demonstrated for glycerid relationships (Annelida, Glyceridae). *Genome Biology and Evolution*, 7(12), 3443–3462.

Rojo V, Martínez-Lage A, Giovannotti M, González-Tizón AM, Nisi Cerioni P, Caputo Barucchi V, et al. (2015). Evolutionary dynamics of two satellite DNA families in rock lizards of the genus *Iberolacerta* (Squamata, Lacertidae): different histories but common traits. *Chromosome Research*, 23(3), 441–461.

Roos C, Zinner D. (2010). Primate phylogeny. In: A Fuentes, ed. *The International Encyclopedia of Primatology*. Wiley–Blackwell, Hoboken, NJ.

Royle NJ, Baird DM, Jeffreys AJ. (1994). A subterminal satellite located adjacent to telomeres in chimpanzees is absent from the human genome. *Nature Genetics*, 6(1), 52–56.

Schmieder R, Edwards R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS On*e, 6(3), e17288.

Seuánez HN. (1982). Chromosome Banding and Primate Phylogeny. Inaugural Address. In: Chiarelli AB, Corruccini RS, ed. *Advanced Views in Primate Biology. Proceedings in Life Sciences*. Springer, Berlin, Heidelberg.

Sibley CG, Ahlquist JE. (1984). The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. *Journal of Molecular Evolution*, 20(1), 2–15.

Skaletsky H, Hughes JF, Page DC. (2004). Consensus sequence of an endogenous retrovirus CERV1. *Repbase Reports*, 4(7), 189.

Straub SC, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. (2012). Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany*, 99(2), 349–364.

Sveinsson S, Gill N, Kane NC, Cronk Q. (2013). Transposon fingerprinting using low coverage whole genome shotgun sequencing in cacao (*Theobroma cacao L.*) and related species. *BMC Genomics*, 14, 502.

Tetreault HM, Ungerer MC. (2016). Long terminal repeat retrotransposon content in eight diploid sunflower species inferred from next-generation sequence data. *G3 (Bethesda)*, 6(8), 2299–2308.

Timmermans MJ, Lees DC, Thompson MJ, Sáfián S, Brattström O. (2016). Mitogenomics of 'Old World Acraea' butterflies reveals a highly divergent 'Bematistes'. *Molecular Phylogenetics and Evolution*, 97, 233–241.

Usai G, Mascagni F, Natali L, Giordani T, Cavallini A. (2017). Comparative genome-wide analysis of repetitive DNA in the genus *Populus L. Tree Genetics & Genomes*, 13, 96.

Ustyantsev K, Blinov A, Smyshlyaev G. (2017). Convergence of retrotransposons in oomycetes and plants. *Mobile DNA*, 8, 4.

Wall JD, Bustamante CD, Hammer MF, Eichler EE, Marques-Bonet T. (2013). Great ape genetic diversity and population history. *Nature*, 499(7459), 471–475.

Yohn CT, Jiang Z, McGrath SD, Hayden KE, Khaitovich P, Johnson ME, et al. (2005). Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. *PLoS Biology*, 3(4), e110.

Supplementary Information for Appendix

Technical replicates analysis

We analyzed 15 technical replicates (one per each biological sample used in this study, outgroup excluded, see table below SApp.1) in order to check if the resulting phylogenetic tree was similar to the ones previously found under this study or the influence of Illumina sequencing bias could be strong enough to distort the phylogenetic signal of the same biological sample. The phylogenetic analyses were performed as previously described in the Materials and Methods section of the main text of the article without filtering out of homoplasious repeats, just to check the performance of technical replicates.

Results of the analysis are shown in Figure SApp.1, in the case of the phylogenetic reconstruction using one individual per species 2 out of 5 OTUs changed between initial and technical replicates tree in Figure SApp.1a and there were no changes between initial and technical replicate trees shown in Figures SApp.1b and c. When we used two individuals per species, 2 out of 10 OTUs changes in comparison shown in Figures SApp.1d and SApp.1f, 4 out of 10 changed in SApp.1e. The phylogenetic reconstruction using three individuals per species showed that 2 OTUs out of 15 changed between the initial tree and the technical replicates tree (Figure SApp.1g). Finally, the two technical replicates of each biological sample group together after a global phylogenetic analysis of genomic repeat abundances using 30 OTUs (two technical replicates per individual) as shown in Figure SApp.1h. Total clustered reads after RE analyses of every technical replicate is indicated in supplementary table SApp.2.

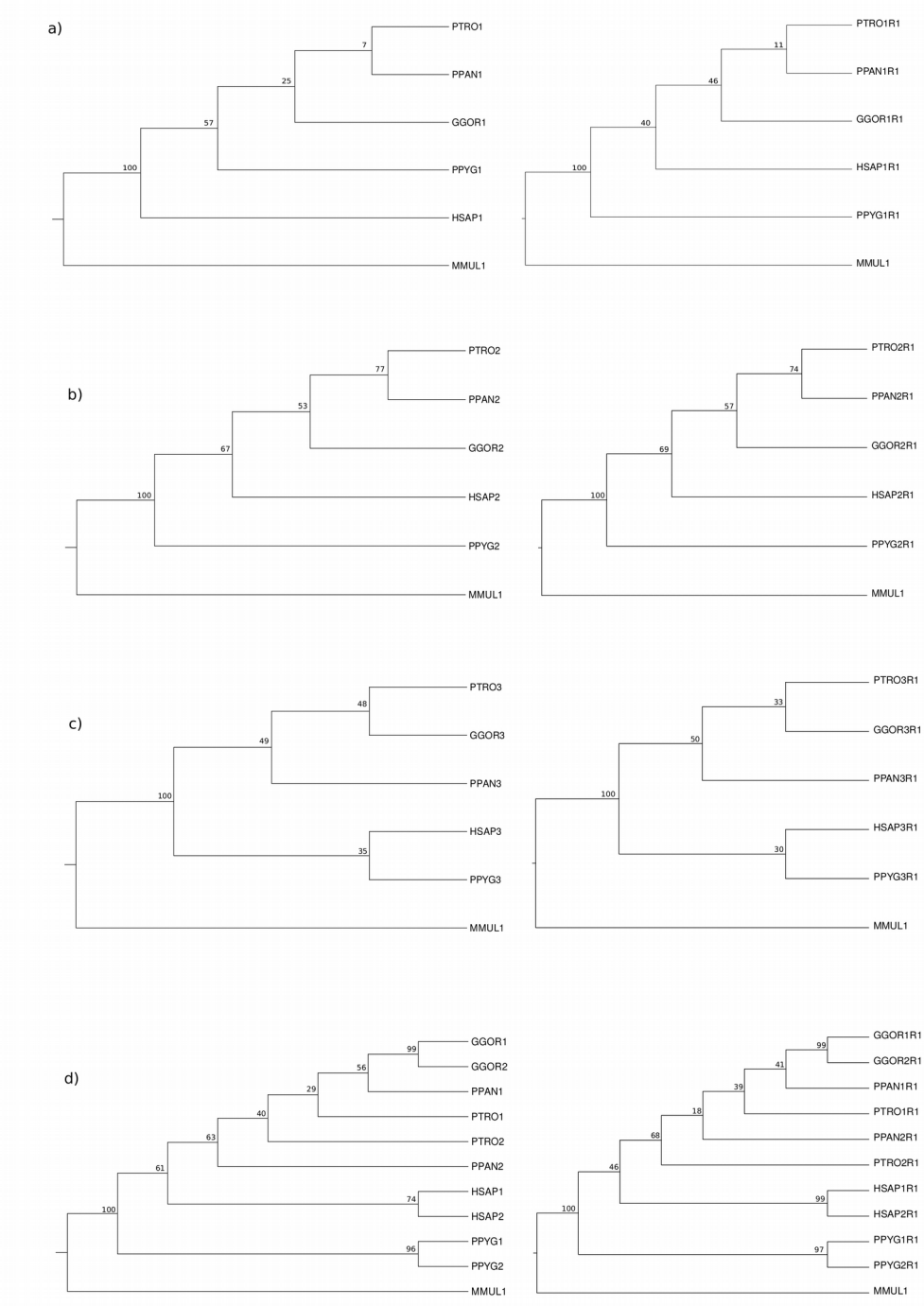
The changes of OTUs reported above took place in low supported branches of the phylogenetic trees so we suspect that these changes are due to the low support (random process) rather than the Illumina sequencing bias between technical replicates. Therefore, we can conclude than, although some Illumina sequencing bias could not be discarded, in this case it is not significant enough to extremely distort the phylogenetic analysis and the advice to use multiple individuals per taxon arise from the inter-individual variation found for genomic repetitive DNA content of each sample.

Table SApp.1. Sampling of technical replicates from NCBI SRA.

Species	Phylogeny ID	SRA RUN ID	BioSample ID
Homo sapiens	HSAP1R1	ERR056987	SAMN00263022
	HSAP2R1	ERR050097	SAMN00014366
	HSAP3R1	ERR050116	SAMN00014365
Pan troglodytes	PTRO1R1	SRR748071	SAMN01920536
	PTRO2R1	SRR748061	SAMN01920534
	PTRO3R1	SRR748057	SAMN01920533
Pan paniscus	PPAN1R1	SRR740805	SAMN01920509
	PPAN2R1	SRR740796	SAMN01920508
	PPAN3R1	SRR740770	SAMN01920506
Gorilla gorilla	GGOR1R1	SRR748090	SAMN01920490
	GGOR2R1	SRR748094	SAMN01920491
	GGOR3R1	SRR748098	SAMN01920492
Pongo pygmaeus	PPYG1R1	SRR748018	SAMN01920551
	PPYG2R1	SRR748003	SAMN01920547
	PPYG3R1	SRR748006	SAMN01920548
Macaca mulatta	MMUL1	SRR1944168	SAMN03264679

Table SApp.2. Reads in clusters after RE clustering of two technical replicates biological sample.

Species	Initial dataset	Reads in repetitive clusters (%)	Technical replicate	Reads in repetitive clusters (%)
Homo sapiens	HSAP1	30	HSAP1R1	31.3
	HSAP2	33.8	HSAP2R1	33.3
	HSAP3	32.9	HSAP3R1	32.8
Pan troglodytes	PTRO1	33.6	PTRO1R1	33.5
	PTRO2	32.9	PTRO2R1	32.3
	PTRO3	32.8	PTRO3R1	32.7
Pan paniscus	PPAN1	33	PPAN1R1	33.1
	PPAN2	31.9	PPAN2R1	31.6
	PPAN3	31.5	PPAN3R1	31.1
Gorilla gorilla	GGOR1	47.7	GGOR1R1	46.8
	GGOR2	40.3	GGOR2R1	39.3
	GGOR3	37.9	GGOR3R1	34.4
Pongo pygmaeus	PPYG1	30	PPYG1R1	28.9
	PPYG2	29.1	PPYG2R1	28.8
	PPYG3	29.6	PPYG3R1	29%
Macaca mulatta	MMUL1	47.5	-	-


a)
PTRO1
PPAN1
GGOR1
PPYG1
HSAP1
MMUL1
7
25
57
100
PTRO1R1
PPAN1R1
GGOR1R1
HSAP1R1
PPYG1R1
MMUL1
11
46
40
100
b)
PTRO2
PPAN2
GGOR2
HSAP2
PPYG2
MMUL1
77
53
67
100
PTRO2R1
PPAN2R1
GGOR2R1
HSAP2R1
PPYG2R1
MMUL1
74
57
69
100
c)
PTRO3
GGOR3
PPAN3
HSAP3
PPYG3
MMUL1
48
49
100
35
PTRO3R1
GGOR3R1
PPAN3R1
HSAP3R1
PPYG3R1
MMUL1
33
50
100
30
d)
GGOR1
GGOR2
PPAN1
PTRO1
PTRO2
PPAN2
HSAP1
HSAP2
PPYG1
PPYG2
MMUL1
99
56
29
40
63
61
74
100
96
GGOR1R1
GGOR2R1
PPAN1R1
PTRO1R1
PPAN2R1
PTRO2R1
HSAP1R1
HSAP2R1
PPYG1R1
PPYG2R1
MMUL1
99
41
39
18
68
46
99
100
97

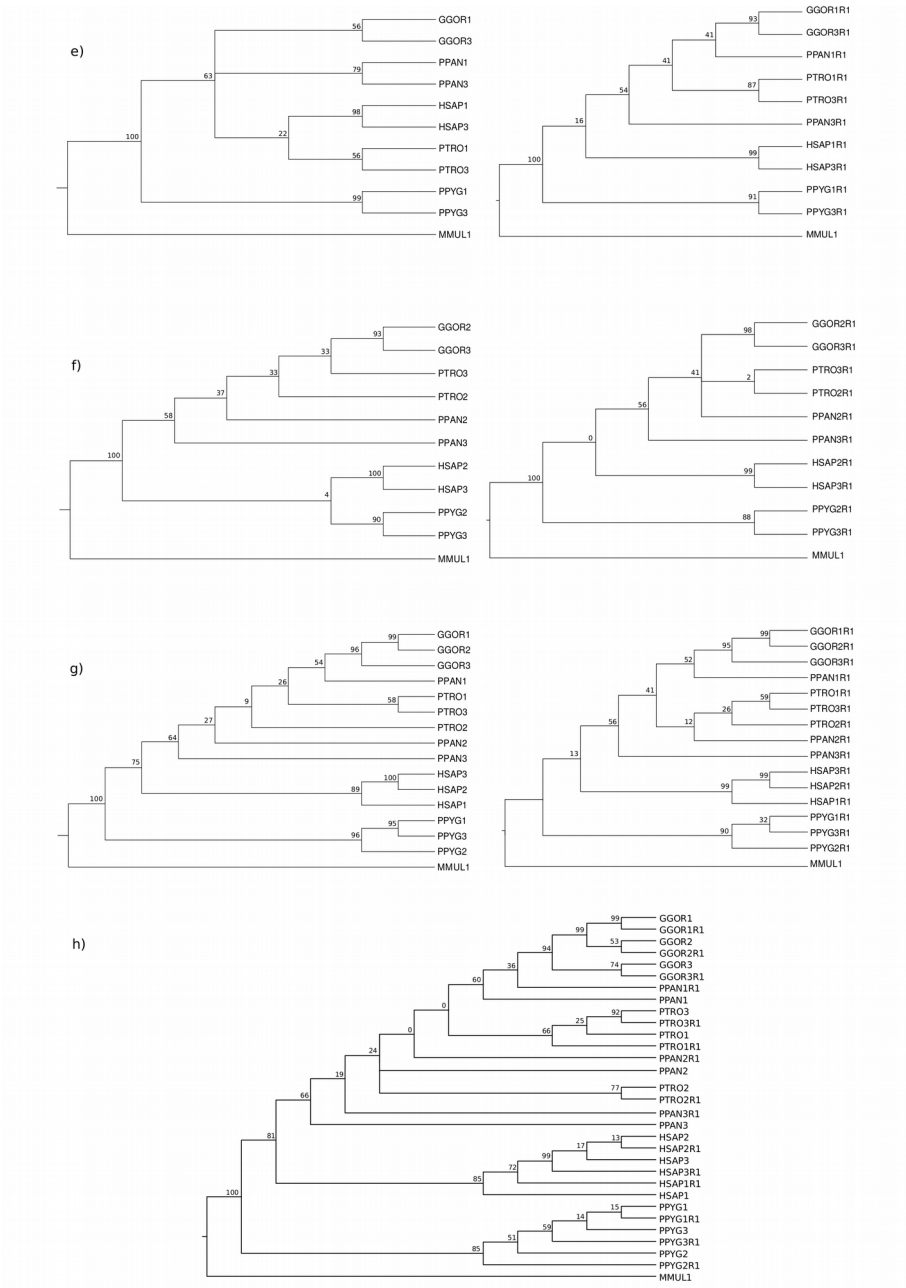


Figure SApp.1. Comparison between the phylogenetic reconstruction of samples used in this study (right) and a technical replicate for each of them (left), in both cases samples were not filtered out of the two homoplasious clusters. a) b) c) three combinations of one individual per species, d) e) f) two individuals per species and g) three individuals per species (all samples). The phylogenetic tree of all samples together (initial sample + technical replicate) is shown in figure h).

Brief discussion and perspectives

Throughout this doctoral thesis we have revealed several molecular and cytogenetic details about the B chromosomes of the grasshopper *E. plorans* and the intragenomic conflict they cause in the host genome. However, as expected in every research, the arising of new questions usually emerges when trying to solve challenges; here we propose some of them.

In Chapters 1 and 2 of this thesis we exhaustively characterized the repetitive DNA fraction of the *E. plorans* genome, which highlights the complexity of these elements that populate a great proportion of the molecular content of B chromosomes. In fact, most of the B chromosomes that have been described so far are rich in repetitive DNA, predominantly satellite DNA (Camacho et al., 2005). In Chapter 1 we described the *E. plorans* satellitome to reveal the congruence between the molecular and cytogenetic properties of these elements. We also show that there exists tandem repeats that do not show FISH signal but exhibit molecular properties resembling those claimed for satDNA. On the contrary, some of them yielded FISH signal but were poorly polymerized. Furthermore, we demonstrated the relationship between transposable elements and satellite DNA. Therefore, in light of these findings on satDNA, a conceptual revision of the term satellite DNA, not based solely on FISH pattern, would be required. It will be just a matter of time that satDNA definition takes into account other molecular properties, including also short tandem repeats. In fact, these latter sequences could develop until reaching higher degrees of polymerization according the evolutionary dynamics they follow and depending also on the genomic context in which they are found.

One of the *E. plorans* satDNA families was found at the centromeric region of all chromosomes, making it a good candidate to behave as centromeric DNA. To validate this idea, a ChIP-seq experiment with antibodies against centromeric proteins, such as

CENP-A, could be carried out. The sequencing of the precipitated DNA could give us the answer about what sequences make up the centromeres of chromosomes in *E. plorans* and also whether there are some differences between the centromeric sequences of the A and B chromosomes that could intervene in the B-drive as described for selfish centromeres (Henikoff and Malik, 2002; Lampson and Black, 2018).

Repetitive DNA has become a very useful tool to identify chromosomes, their origin and relationships between them. In Chapter 2, we described the repetitive content of the B chromosomes of *E. plorans* from distant populations, which has allowed us to delimit the origin of Bs on the chromosome 9 of standard complement, thus from an intragenomic origin. Furthermore, the positive correlation between the repetitive content of B chromosomes from different populations of *E. plorans* would support the hypothesis of a common origin of B in this species. However, the presence of specific B repetitive sequences such as EplTR112-11, mapping against very few reads from the 0B libraries, leads us to propose a possible origin of this satDNA within the B chromosome or even its arrival at Bs after an interspecific hybridization process, which would point to an interspecific origin of the B chromosome. In this sense, it would be interesting to analyze gDNA libraries from populations of *E. plorans* in which the B chromosome is absent to search for these specific satDNA. If we found this satDNA in genomic libraries from B-lacking populations, we would consider that their appearance in the B chromosomes could have been produced by local amplification in Bs of sequences that were previously in the host genome. This would take us away from thoughts about a *de novo* origin of satDNA in the B chromosomes or through a hybridization process that involved B chromosomes.

In Chapter 3 we identified 42 protein-coding genes in the B chromosomes of *E. plorans* from Torrox, many of them were also found in the chromosomes of populations located in Tanzania, Egypt and Armenia. This finding again suggests a common origin for the B chromosomes in *E. plorans*. In fact, these B-genes also shared several specific polymorphisms (SNPs) from the B-carrying libraries. We also demonstrated, by means of the expression analysis of B-specific SNPs, that copies of genes found on the B chromosomes of *E. plorans* are actively expressed, even to a greater extent than the B-located repetitive DNA does. Then, the role of B-genes could be crucial for their own maintenance and evolution. In fact, several B-genes involved in cell cycle function harbored a high number of B-specific SNPs. In this thesis, according to Lahn and Page

(1999) and Kinsella et al. (2019), we propose that genes showing the highest number of SNPs could be ancestral on the B chromosomes assuming a model of neutral evolution. However, we are also aware that there are alternative readings about this issue. Some interpretations could assert that the high number of SNPs in some B-genes indicates a selection against certain variants, perhaps harmful for the host genome. In fact, the *ndl* B-gene (showing the highest number of SNPs in *E. plorans*) presents a ratio of non-synonymous and synonymous substitution higher than 1 (i.e. 1.32; see Table 3.3), which would also support some selection against intact copies of that gene in the B chromosomes. These outcomes about B-genes depict a new paradigm in which the B chromosomes could contain the necessary genes for their own maintenance in populations, thus involved in their drive. Therefore, the B chromosome would appear as a specialized chromosome based on its genetic content, the latter being the ultimate responsible for the selfish nature of this chromosome. This view is supported by recent findings in other species about key B-genes that cause their own effects in the host. Among some gene-based specialized chromosomes, we should mention the Y chromosome that contains the *sry* gene that determines the carriers to be males (Berta et al., 1990), the PSR in *N. vitripennis* harboring the *haploidizer* gene that causes the deletion of the paternal genome (Dalla Benetta et al., 2019) or the GRC chromosome that contain germline determining genes which is congruent with the lack of this chromosome in somatic cells (Kinsella et al., 2019).

One of the first obstacles that we should overcome in order to demonstrate the functionality of the B-copies of protein-coding genes would comprise the study of their translation to protein. For this purpose, it would be interesting to carry out a western blot study, protein immunoprecipitation and sequencing using antibodies either specific to polymorphisms found in B chromosomes or including also copies from the As provided that we could distinguish between them (truncated genes in B chromosomes, amino acid changes...). Another approach to reveal the functions of B-genes in carriers individuals would consist of using the RNA interference pathway (RNAi) to avoid translation of the B-genes copies and thus their effects on individuals.

On the other hand, the gene content of the B chromosomes could also help to elucidate the origin of these genomic elements in *E. plorans*. In this sense, it would be interesting to find out if the B-genes are linked in the standard complement and come from a single chromosome. With this aim, we could set up controlled crosses to track

polymorphism inheritance as well as design a FISH probe including several of the B-genes (e.g. BAC-FISH) to identify the chromosome of the host genome from which B chromosome could have arisen.

Chapters 4 and 5 bring to light several key aspects of the intragenomic conflict caused by the B chromosomes of *E. plorans*. We demonstrated the elimination of B chromosomes in *E. plorans* males and we evidenced the transcriptional crosstalk between A and Bs through a comprehensive RNA-seq experiment including different developmental stages and sexes of the species. These results point to an intense effect of the B chromosome in gonads of *E. plorans*, especially in ovary. Interestingly, we detected over expression of genes involved in silencing mechanism associated with the presence of B chromosomes in testis, which would be congruent with the elimination of B in the spermiogenesis of *E. plorans*. In addition, we found a down-regulation of several cell cycle genes in ovaries when the B chromosome is present, which could promote the drive of the Bs and its accumulation through offspring. To go deeper into this intragenomic conflict, it would be interesting to consider a similar RNA-seq experiment, but directed towards crosses between females with Bs and males belonging to populations in which the B chromosome is absent. Herrera et al. (1996) described the reactivation of drive after controlled crosses of this nature, so a study of RNA-seq in that offspring could reveal fundamental details in order to understand how the drive of B chromosomes occurs at transcriptional level. In addition, the new single-cell sequencing technologies could be a good approach to know what specifically happens in a testis and an oocyte when the elimination of Bs or their drive, respectively, takes place in *E. plorans* and we expected the B chromosomes to play their best cards.

References

Berta P, Hawkins JR, Sinclair AH, Taylor A, Griffiths BL, Goodfellow PN, et al. (1990). Genetic evidence equating SRY and the testis-determining factor. *Nature*, 348(6300), 448–50.

Camacho JPM. (2005). B chromosomes. The evolution of the genome (Gregory TR, ed.), 223–286.

Dalla Benetta E, Antoshechkin I, Yang T, Nguyen HQM, Ferree PM, Akbari OS. (2020). Genome elimination mediated by gene expression from a selfish chromosome. *Science Advances*, 6(14), eaaz9808.

Henikoff S, Malik HS. (2020). Centromeres: selfish drivers. *Nature*, 417(6886), 227.

Herrera JA, López-León MD, Cabrero J, Shaw MW, Camacho J. (1996). Evidence for B chromosome drive suppression in the grasshopper *Eyprepocnemis plorans*. *Heredity*, 76(6), 633.

Kinsella CM, Ruiz-Ruano FJ, Dion-Côté AM, Charles AJ, Gossmann TI, Cabrero J, et al. (2019). Programmed DNA elimination of germline development genes in songbirds. *Nature Communications*, 10(1), 5468.

Lahn BT, Page DC. (1999). Four evolutionary strata on the human X chromosome. *Science*, 286, 964–967.

Lampson MA, Black BE. (2017). Cellular and Molecular Mechanisms of Centromere Drive [published correction appears in Cold Spring Harb Symp Quant Biol. 2018 Feb 26;:]. *Cold Spring Harbor Symposia on Quantitative Biology*, 82, 249–257.

Conclusions

1. We have found and characterized 1,726 repetitive sequences in the genome of *Eyprepocnemis plorans* considering transposons, satDNA, rDNA, mtDNA, histones, tRNA and snRNA. In particular, repetitive elements represent about a 65% of the genome content in this species being LINE retrotransposons the most abundant class.

2. The 8.4% of the *E. plorans* genome is dedicated to satDNA and tandem repeats, classified in 112 TR families. In deep analysis of molecular and cytological properties of each family allowed us to prove the congruence between all these features in the satellitome of *E. plorans*. We demonstrated that long repeat units, a high degree of polymerization, array homogenization and genomic independence appear together in several TR families, suggesting well-polymerized satDNAs.

3. SatDNA and tandem repeats are dynamic repetitive elements of the genome that could appear in different polymerization stages, associated or not with transposons and yielding different FISH patterns. This finding points to the existence of different evolutionary stages in a cyclic model of polymerization, dissemination and degradation of satDNA in which TEs could play an important role.

4. In the B chromosomes of *E. plorans* there are plenty of repetitive elements, which constitute around the 86.3% of their molecular content. The most abundant class of repetitive DNA found in Bs is satDNA making up the 65% of these chromosomes which is consistent with their heterochromatic nature.

5. B-carrying individuals of *E. plorans* express some repetitive elements that they contain. The B chromosome show a high expression of several transposons while the mtDNA is the outstanding repetitive class more intensely expressed in 0B samples. However, absolute expression of B-located copies is insignificant compared to that from the A chromosomes, suggesting silencing mechanisms against some possible harmful effects of repetitive DNA activity from B chromosomes.

6. The B chromosome of *E. plorans* contains protein-coding genes, here we identified 42, several of them related to cell cycle functions. One of them is the gene *cdc16* that controls cell division, as it does the gene *apc1* found in the B chromosomes of *Locusta migratoria*, both genes involved in the assembly of the APC/C complex.

7. Individuals of *E. plorans* harboring B chromosomes show expression of B-specific copies of protein-coding genes. In particular, the *ndl* gene, involved in cell cycle functions, shows higher expression of B-copies than that from A chromosomes, especially in ovaries, the tissue in which the drive of B chromosomes takes place in this species. This fact suggests that genes located in the B chromosomes could play a crucial role for their own maintenance and evolution.

8. The location of repetitive elements found in the B chromosomes of *E. plorans* suggests an intragenomic origin from the chromosome 9 of the standard set. Furthermore, the high correlation between the repetitive and gene content of Bs from individuals belonging to distant populations supports a common origin of B chromosomes in this species.

9. The B chromosome of *E. plorans* is postmeiotically eliminated from testis during spermiogenesis, thus suggesting a dynamic for maintenance and balance of B chromosomes in nature through elimination in males and accumulation by female meiotic drive.

10. B chromosome presence is associated with greater differences in gene expression for *E. plorans* embryos than those produced by sex, probably because sexual differentiation in these embryos is still emerging. Several DEGs in the presence of B chromosomes are transposons, which would represent a response to the stress caused by these chromosomes in embryos. Furthermore, we found more differentially expressed genes in embryos that received a maternal B chromosome than in case of paternal inheritance, which suggests a kind of imprinting for the B chromosome in this species.

11. The gonad, in particular the ovary, is the tissue in which we found more changes in gene expression associated with the presence of a B chromosome, followed by embryos and legs. Most of DEGs in gonads are protein-coding genes. Moreover, in ovary we found a down-regulation for many of them, which could entail a response to manage the B drive that takes place in this organ for *E. plorans*.

Conclusiones

1. Hemos identificado y caracterizado 1.726 secuencias repetitivas en el genoma de *Eyprepocnemis plorans* considerando transposones, ADNsat, ADNr, ADNmt, histonas, ARNt y ARNsn. En particular, los elementos repetitivos representan aproximadamente el 65% del genoma en esta especie, siendo los retrotransposones LINE la clase más abundante.

2. El 8,4% del genoma de *E. plorans* está dedicado a ADN satélite y repeticiones en tándem, agrupadas en 112 familias. El análisis detallado de las propiedades moleculares y citológicas de cada familia nos permitió probar la congruencia entre estas características en el satelitoma de *E. plorans*. Además, demostramos que las unidades de repetición más largas, un grado de polimerización alto, la homogeneización de arrays, la independencia genómica y las señales de FISH evidentes aparecen al mismo tiempo en varias familias, lo que sugiere satélites bien polimerizados.

3. El ADNsat y las repeticiones en tándem son elementos repetitivos dinámicos del genoma que aparecen en diferentes etapas de polimerización, asociados o no con transposones y produciendo diferentes patrones de FISH. Este hallazgo sugiere la existencia de diferentes etapas evolutivas en un modelo cíclico de polimerización, diseminación y degradación de ADNsat en el que los elementos transponibles podría tener un papel importante.

4. Los cromosomas B de *E. plorans* están repletos de elementos repetitivos que constituyen alrededor del 86,3% de su contenido. La clase más abundante de ADN repetitivo que se encuentra en los Bs es el ADNsat, que constituye el 65% de estos cromosomas, lo que es consistente con su naturaleza heterocromática.

5. Los individuos de *E. plorans* portadores de cromosomas B expresan algunos elementos repetitivos contenidos en éstos. El cromosoma B muestra una alta expresión de varios elementos transponibles, mientras que el ADN mitocondrial es la clase repetitiva expresada con mayor intensidad en individuos 0B. Sin embargo, la expresión absoluta de las copias de repetitivo localizadas en los cromosomas B, es insignificante en comparación con la de los As, lo que sugiere mecanismos de silenciamiento contra posibles efectos perjudiciales de la actividad del ADN repetitivo de los cromosomas B.

6. El cromosoma B de *E. plorans* contiene genes codificantes de proteínas, aquí identificamos 42, varios de los cuales están relacionados con funciones del ciclo celular. Uno de ellos es el gen *cdc16* que controla la división celular, al igual que el gen *apc1* identificado en los cromosomas B de *Locusta migratoria*, ambos involucrados en el ensamblaje del complejo APC/C.

7. Los individuos de *E. plorans* que albergan cromosomas B expresan copias de genes codificantes de proteína específicas de los Bs. En particular, el gen *ndl*, involucrado en funciones del ciclo celular, muestra una mayor expresión de copias provenientes del B que de los As, especialmente en ovarios, el tejido en el que tiene lugar el impulso y la acumulación de los cromosomas B en *E. plorans*. Este hecho sugiere un papel crucial de los genes que se encuentran en los cromosomas B para su propio mantenimiento y evolución.

8. La localización de los elementos repetitivos que se encuentran en los cromosomas B de *E. plorans* sugiere un origen intragenómico del mismo a partir del cromosoma 9 del conjunto estándar. Además, la alta correlación entre el contenido de repetitivo y genes en los Bs de individuos pertenecientes a poblaciones distantes sugiere un origen común de los cromosomas B en esta especie.

9. El cromosoma B de *E. plorans* se elimina postmeióticamente en los testículos durante la espermiogénesis, lo que sugiere una dinámica para el mantenimiento y equilibrio del cromosoma B en las poblaciones a través de su eliminación en los machos y su acumulación mediante el impulso meiótico en las hembras.

10. La presencia de un cromosomas B se asocia en embriones de *E. plorans* con mayores diferencias de expresión génica que las producidas por el sexo, probablemente porque en estos embriones la diferenciación sexual es todavía muy incipiente. La mayor parte de los DEGs en presencia de cromosomas B son transposones lo que representaría una respuesta al estrés provocado por estos cromosomas en embriones. Además, en los embriones que recibieron el B por vía materna encontramos más genes diferencialmente expresados que en los que lo recibieron por vía paterna, lo que sugeriría algún tipo de impronta en el cromosoma B de esta especie.

11. La gónada, y en particular el ovario, es el tejido en el que encontramos más cambios de expresión génica asociados a la presencia de un cromosoma B por delante de embriones y patas. La mayor parte de los genes diferencialmente expresados en gónada son codificantes de proteínas. Además, en ovario encontramos una infraexpresión de muchos de estos genes lo que podría representar una respuesta para gestionar el impulso del B que tiene lugar en este órgano.

Agradecimientos/Acknowledgements

¡Qué paradójico el tiempo! Parece que fue ayer cuando llegué a Granada por primera vez pero han pasado tantas cosas desde entonces que la metamorfosis es inevitable, y precisamente de eso se trata. El camino hasta aquí podría haber sido mucho más duro sin todas las personas que me han acompañado en esta travesía, no hay palabras para expresar lo agradecida que me siento.

Gracias a mi director de tesis, Juan Pedro, tus valiosas enseñanzas no las voy a olvidar, he aprendido enormemente de ti en estos años. A Lola y Pepi, gracias por acogerme en el grupo con los brazos abiertos, hacéis del laboratorio un lugar sorprendentemente cómodo. A todos los miembros del Departamento de Genética, ha sido un placer compartir con vosotros consejos de departamento, saludos por los pasillos o celebraciones navideñas, gracias por cargar de energía el departamento cada día. En particular, quiero agradecer a todas las personas del laboratorio A105 del CIBM, especialmente a Miriam y David, por vuestra buena predisposición cuando necesité un Qubit o un termociclador, por abrirme las puertas de vuestro laboratorio con tanta amabilidad.

Mi profundo agradecimiento a Bea y Paquillo por ser mis hermanos mayores en esto de hacer una tesis, por vuestra inestimable ayuda científica, pero sobre todo gracias por los buenísimos ratos en el laboratorio y fuera de él, eso deja poso. Gracias también a Rubén, Merce, Javi, Marco, Modesto, Tati, Helena, Sandra, Diego, Raquel, Alicia, Ricardo, Dilamm, Judit y Álex (seguro que me dejo alguien) con los que he tenido el gusto de compartir quejas y risas, ambas eternas. A Ester y Julio, no pude tener mejor compañía durante mi último año en el laboratorio, gracias por el soplo de aire fresco, por vuestra pureza, sois personas estupendas.

I would also like to thank Stein Aerts for hosting me in his group as if I had always been a part of it. Thanks for totally involving me in the everyday environment of the lab in such a kind and friendly way. I am also truly grateful to Valerie for introducing me to ChIP-seq and technical support. My greatest thanks to the students in the group

(especially to Hana, Jelle, Dmitry and Gert) for the science, the flies, the concerts, the beers and, why not?, the typical Spanish dancing. I had an amazing time in Belgium thanks to you. Thanks to Frank for the inspiring and crazy chats about evolutionary biology and economics.

Agradezco también a los participantes del EMPSEB24 y a todos los que contribuyeron a que este encuentro fuera un éxito, no pudo ser más motivadora y enriquecedora esa semana de ciencia intensiva. Este reto que Caro y yo asumimos no se hubiera materializado sin Adrián, Antón, Diego de Miguel, Diego Salazar y Jorge, ¡muchas gracias equipo!

Durante este año en el CIC he conocido a cinco personas maravillosas: César, Delia, Marta, Paloma y Sara; gracias por los indispensables cafés, por vuestra sonrisa en la cara, por vernos fuera del trabajo, por todos los planes que no han llegado a término pero lo harán. Gracias también a *los pinturicas* granaínos, y a Andreia, Bea, Nacha y Sandro, el tiempo compartido durante estos años ha sido muy estimulante. También a los colegas pegados: Álex, Beli, Carrillo, Esther, Estela, Lidia, Magda y Rafa, por vuestros brazos siempre abiertos y por descubrirme rincones maravillosos de Granada desde que llegué. Gracias a Jorge, un científico excepcional y mejor persona.

Y esta tesis directamente no sería sin todas esas personas que hacen que mi hogar sea cada vez más grande, soy muy afortunada por teneros en mi vida.

Gracias a Mónica, Jessica, Antowan, Laura y Dayana, será cierto eso de que el Erasmus te cambia la vida porque la mía me gusta más con vosotros, gracias por hacer que la distancia se esfume en una charla.

A los *Doctores de USAL y tiral*: Becky, Diego, Elisa, Marcos, Mario, Noe, Patri, Sandra, Usa y Víctor, gracias por detener el tiempo, por recordarme que al lugar donde has sido feliz también puedes volver. Un inmenso gracias a Cynthia y a Marisa, mis *cabesas páharo* favoritas, porque Salamanca fue mi sitio gracias a vosotras, por el *¡aunque sea unos pintxitos!*, por estar ahí cuando yo no estaba.

Un millón de gracias a las chicas *pin!* e *hiper-tankes* (me ahorro el megalistado), la cuadrilla, por hacer familia. Es increíble que un grupo enorme de personas tan dispares sea el lugar idóneo para encontrarse. Gracias por hacer que el pueblo sea la mejor opción y que unos tercios con doritos se conviertan en la cena más exquisita. Sois maravillosas.

A mi familia y amigos de Coín, en especial a Jane y Pepe, gracias por hacerme sentir una más desde el primer momento, por preocuparos y cuidarme, por vuestro cariño y cercanía.

A Caro, una persona excepcional. Gracias infinitas por tu amistad, por toda la ayuda, por estar siempre, por tu fuerza descomunal, tu alegría y tus locuras serenas, por pensar que otra ciencia es posible. Has sido un pilar imprescindible en estos años, ¡qué suerte cruzarnos en el camino! Sea como fuere, este tiempo habrá merecido la pena gracias a ti.

A mi familia, tíos y primos (y muchos más), gracias por vuestro cariño, por estar siempre ahí para reconfortarme, no puedo sentirme más privilegiada. A mi abuela Angelines, yo te echo de menos aunque tu calidez no se vaya, me abrigarás toda la vida.

A Adrián, el mejor compañero de vida, por caminar junto a mi pese a lo tortuoso del camino. Gracias por quererme sin reservas ni condiciones, por conocerme, entenderme y, aun así, quedarte. Me haces muy feliz.

Por último, a mis padres, Montse y Ángel, fundamentales en mi vida, os admiro. Sois los verdaderos artífices de esta tesis, sin vuestro apoyo no estaría aquí. No soy sin vosotros, gracias por enseñármelo todo, por hacerlo tan bien, por la confianza y la libertad, por darme alas y aplastar los miedos, por ser mi raíz. Una vida no será suficiente para compensar todo lo que me dais, estaré en deuda con vosotros, ¡os quiero!