HIGH SCALE GENOMIC ANALYSIS APPLIED TO B CHROMOSOME BIOLOGY

SYED FARHAN AHMAD

Botucatu, May 2019

Universidade Estadual Paulista “Júlio de Mesquita Filho”
Instituto de Biociências de Botucatu
Programa de Pós-Graduação em Ciências Biológicas (Genética)

HIGH SCALE GENOMIC ANALYSIS APPLIED TO B CHROMOSOME BIOLOGY

PhD student: **Syed Farhan Ahmad**
Supervisor: **Prof. Dr. Cesar Martins**

PhD thesis submitted to the Institute of Biosciences, São Paulo State University (Portuguese: Universidade Estadual Paulista "Júlio de Mesquita Filho", UNESP), Campus of Botucatu, to obtain the title of Doctor from the Postgraduate Program in Biological Sciences (Genetics).

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Ahmad, Syed Farhan.
High scale genomic analysis applied to b chromosome biology / Syed Farhan Ahmad. - Botucatu, 2019

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Cesar Martins
Coorientador: Guilherme Targino Valente
Coorientador: Rachel O'Neill
Capes: 20200005

1. Chromosomes. 2. Genomes. 3. Genes. 4. Evolution.

Palavras-chave: chromosome; evolution; genes; genome.

“Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less”

Marie Curie

“Look deep into nature and you will understand everything better”

Albert Einstein.

This thesis is heartily dedicated to my mother who took the lead to heaven before the completion of this work.

Acknowledgment

I thank all who in one way or another contributed in the completion of this thesis.

I gratefully acknowledge the Post-Graduate Program in Biological Sciences (Genetics) of the University São Paulo State University, for the opportunity to expand my knowledge and making it possible for me to obtain PhD here.

This work would not have been possible without the financial support of FAPESP . The complete project of my PhD was funded by **FAPESP (process number: 2014/16477-3 and 2018/03877-4)** for both doctoral scholarship and research internship abroad funds. I express deep gratitude to this prestigious foundation of research.

My special and heartily thanks to my supervisor, Professor Dr. Cesar Martins who encouraged and directed me in this work and significantly contributed to my academic training. During my tenure, he gave me intellectual freedom in my work, supporting my attendance at various conferences, engaging me in new ideas, and demanding a high quality of work in all my endeavors. His challenges brought this work towards a completion. It is with his supervision that this work came into existence. For any faults I take full responsibility.

I give deep thanks to my co-supervisor Professor Dr. Guilherme Targino Valente, for beneficial ideas, training and knowledge regarding my research project.

I am also deeply thankful to Professor. Dr. Rachel O'Neill who accepted me in her laboratory for the accomplishment of my internship abroad, allowing me to experience advanced training. I also acknowledge the Department of Molecular and Cell biology, University of Connecticut, USA for hosting me during the six months training.

My appreciation to the Prof. Dr. Diogo Cavalcanti Cabral-de-Mello and Prof. Dr. Vladimir Pavan Margarido for their collaboration in providing the samples.

Additionally, I would like to thank my committee members for their interest in my work.

This thesis was accomplished with the help and support of my fellow lab mates and collaborators, Adauto Lima Cardoso, Erica Ramos, Bruno Fantinatti, Jordana Oliveira, Rafael Coan, Rafael Nakajimae, Maryam Jehangir, Natália Bortholazzi Venturelli and Ivan Wolf. I greatly benefited from their keen scientific insight, their knack for solving seemingly intractable practical difficulties, and their ability to put complex ideas into simple terms.

I also thank my family who encouraged me and prayed for me throughout the time, and my wife, Maryam, for her continued support, encouragement and collaboration in my research. Also not forgetting my daughter, Inshirah, her cute smile gives me more courage.

Syed Farhan Ahmad

Table of Contents

Highlights

- The genomes of three species containing B chromosomes were sequenced in this project.
- The repetitive and gene contents of the B chromosomes in diverse species were investigated.
- In contrast to theories that B chromosomes are gene poor, the present study found that they are gene rich and contain many protein-coding genes.
- In all the species analyzed here, it seems that B chromosomes tend to gain sequences in first preference that are crucial for their own establishment inside the cell.
- Besides the genes that give transmission advantage to Bs, there are others coding for many important biological processes, indicating the contribution of Bs in genome function.
- Evidences were found that considerable amount of genomic portions have been migrated from A chromosomes to B via duplications and rearrangements events.

Abstract

One of the biggest challenges in chromosome biology is to understand the occurrence and complex genetics of extra, non-essential karyotype elements, commonly known as supernumerary B chromosomes (Bs). Bs are present in diverse species of eukaryotes and their molecular characterization remains elusive for years. A distinguished feature that makes them different from the normal chromosomes (called A chromosomes) is their way of inheritance in irregular fashion. Over the last decades, their genetic composition, function and evolution have remained an unresolved query, although a few successful attempts have been made to address these phenomena. The non-Mendelian inheritance and unpairing/non-recombining abilities make the B chromosomes immensely interesting for genomics studies, thus arising different questions about their genetic composition, survival, maintenance and role inside the cell. This study aims to uncover these phenomena in different species. Here, we sequenced the genomes of three model organisms including fish species *Astyanax mexicanus* and *Astyanax correntinus*, and grasshopper *Abracris*

flavolineata with (B+) and without Bs (B-) to identify the B-localized sequences, called B chromosome blocks (“B-blocks”). We established approaches for this analysis that comprised of steps such as comparative genomics analysis and annotation of B chromosomal genes and DNA repeat types. The next generation sequencing (NGS) analyses identified thousands of genes fragments as well as a few complete genes to be present on the Bs. The repetitive DNA analysis showed that the Bs harbor different types of transposable elements (TEs) with domination of Tc1-pogo, hobo-activator and Gypsy DNA transposons, and L2/rex and Jockey retroelements. The functional annotation revealed that the Bs have gained copies of many genes coding for diverse set of functions related to important biological phenomena such as cellular processes, metabolism, development, response to stimulus, immune response, localization, morphogenesis and biological regulation. Our results showed that the Bs are enriched with genes associated to cell cycle and chromosome formation, which might be important for the establishment of Bs in the cell. We further detected different patterns of genomic evolution such as segmental duplications and inversions associated with Bs and highlighted their multi A chromosomal origin. Based on these findings, we corroborate our primary hypothesis that the accumulation of genes on B might have played a key part in driving its transmission, escape, survival and maintenance inside the cell. The B-localized contents, as revealed in our study, provide insights for theories of B chromosome evolution.

Keywords: chromosome, genome, genes, evolution, next generation sequencing.

1. Introduction

1.2 B chromosomes

B chromosomes (Bs) are extra non-essential karyotypic components which show non-Mendelian features and lack the ability of recombination or pairing with the normal A chromosomes (Longley et al. 1927). Bs were firstly discovered in plant bug insect *Metapodius*, now called *Acanthocephal* (Wilson, 1906) and in coleopteran insects *Diabrotica soror* and *D. punctata* (Stevens, 1908). Approximately 2,080 plants and 736 animals' species are currently known to carry B chromosomes (Ahmad and Martins, 2019). The occurrence of Bs in multiple numbers is probably related to their strength of accumulation mechanism and the degree to which a specific species can tolerate these extra elements. In some cases of plants, the high level of tolerance is probably related due to their domestication, for example corn plants have been reported to tolerate as many as 34 B chromosomes (involving a 155% increase in nuclear DNA content; see Jones and Rees, 1982). Similarly, up to 20 Bs have been reported in *Allium schoenoprasum* plants (Bougourd et al. 1995). While some wild plants, for instance the *Lolium perenne* (Jones and Rees, 1982) and *B. dichromosomatica* (Carter, 1978), the frequency in individuals remains as low as three B chromosomes. The existence of supernumeraries in animal species also varies broadly, such as grasshopper *Eyprepocnemis plorans* (Camacho et al. 1997b) and the flatworm *Phyllostachys nigra* (Beukeboom et al. 1996) have carry up to three Bs, while the endemic New Zealand frog *Leiopelma hochstetteri* can acquire up to 15 mitotically stable B chromosomes (Green et al. 1993).

The comparison of size and centromeric position between As and Bs was performed by karyotype analysis (Jones, 1995). Various morphological forms of B chromosomes are reported such as isochromosomes in *Crepis capillaris* (Jones et al. 1991), subtelocentric or telocentrics in *Hypochoeris maculate* (Parker, 1976). Generally, supernumerary chromosomes have smaller size as compared to A

chromosomes. In approximately 40% of B carrying species of angiosperm, Bs were estimated to attain an average size of 1/4 to 3/4 of As (Jones, 1995). While in some species, Bs are categorized as very small microchromosomes such as in *Campanula rotundifolia* (Böcher, 1960), *Linanthus pachyphyllus* (Patterson, 1980) (Lewis, 1951), *Sorghum nitidum* (Raman, and Krishnaswami, 1960) and *Erianthus munja* (Sreenivasan, 1981). “Large” B chromosomes are also reported in flowering taxa *Rumex thyrsiflorus* (Zuk, 1969), *Calycadenia oppositifoli*a, *C. ciliosa* (Carr et al. 1982) and *Plantago serraria* (Frost, 1951).

A well-known and typical concept is that B chromosomes are the derivatives of As (Jones and Rees, 1982), which has been experimentally explained in diverse species (Jamilena et al. 1994 ; Wilkes et al. 1995; Stark et al. 1996; Jin et al. 2005; Valente et al. 2014). As a result of expanding documentations about the genomics contents of Bs, it is now inferred that B chromosomes, once considered as entirely heterochromatic and genetically inert, not only constitute repetitive contents but distinct processed pseudogenes and protein-coding genes. These gene sequences have been localized and identified by utilization of recent techniques in molecular biology such as AFLP, FISH, real-time qPCR and genome sequencing (Yoshida et al. 2011; Valente et al. 2014; Makunin et al. 2014; Banaei-Moghaddam et al. 2015; Huang et al. 2016; Navarro-Domínguez et al. 2017). The revelation of numerous multiple autosomal genes on Bs starts a new debate about their evolutionary role, their complex interactions with host genome and their possible effects ranging from sex determination to fitness and adaptation (Alexey et al. 2014). The evolutionary role of Bs in genome is not clearly understood. How do they originate? Why do they occur more frequently in some species than in others? Are they short term events or do they persist in genomes for a long time? Further analysis of the molecular content of the B chromosomes can answer these questions.

Note: Please refer to our recent paper (Ahmad and Martins, 2019) attached as a supplement for more detailed literature on B chromosomes.

1.2 The application of cytogenetics and genomics to B chromosome analysis

The science of cytogenetics was founded with the beginning of study related to chromosomal behavior during cell division at the end of the nineteenth century. The field gained reputation with the development of new techniques at second half of twentieth century. After 1980, the major breakthrough in molecular biology happened, and modern cytogenetics came out as a result of combination with molecular biology techniques, thus allowing significant advances in understanding genomes through chromosome studies. The first hybridization of nucleotides to chromosomes and nucleus (Pardue and Gall, 1969; Gall and Pardue, 1969), followed by several experiments to use radioactively labeled repetitive DNAs (rRNA genes and satellite DNAs) and finally fluorescent in situ hybridization (FISH) (Pinkel et al. 1986) techniques revolutionized the area of cytogenetics. The scope of cytogenetics further improved with the rise of availability of several completely sequenced eukaryotic genomes in the last decade. As a result, progress is being made to enhance the efficiency of comparative analysis and physical chromosomal mapping of genes. Nevertheless, the application of modern genomics or cytogenetics alone was not satisfactory to accomplish chromosomes related projects with complete scientific outcomes. Both areas needed to depend on one another like: genomics would require significant information from fundamental cytogenetic studies involving the identification of chromosome number and morphology and mapping; similarly, cytogenetics would have to rely on modern genomics to complete the goals. Thus, an integration of both fields was required. The marriage of genomics and cytogenetics gave birth to a new branch of chromosome biology known as cytogenomics. The arrival of genome sequencing and exponential growth in bioinformatics technologies further advanced the cytogenomics studies. This modern field has proven very effective in chromosome biology. Moreover, latest improvement in high-scale DNA, RNA and proteins analysis has allowed biologists to answer the questions regarding the molecular mechanisms involved in the evolution and origin of chromosomes (see review, Valente et al. 2017).

The applications of cytogenetics have significantly contributed to understand the origin and evolution

of B chromosomes. FISH is a useful tool for ascertaining the origin of B chromosomes (Silva et al. 2016). Cytogenetics approaches coupled with large-scale genomics sequencing have effectively unraveled the structure and composition of B chromosomes.

1.3 The *Astyanax* fish as model organisms to study B chromosomes

The *Astyanax* group belongs to fish family Characidae, Characiformes order and is regarded as one of the prevalent genera in South America and reported to encompass around 90 valid species (Ge´ry, 1977; Lima et al. 2003). This genus represents an interesting biological model for chromosomal analysis due to results obtained from the location of ribosomal cistron and satellite DNAs, and also because of characterization of supernumerary chromosome (Mestriner et al. 2000). *Astyanax* has emerged as an excellent model for general studies concerning evolutionary mechanisms (Langecker et al. 1991; Jeffery, 2001). The reason we chose *Astyanax* for our analysis is due to the high prevalence of Bs in the group Astyanax (Silva et al. 2016). Our survey of the Bs literature indicate around a total of 14 species of this genus hitherto reported to carry the supernumeraries.

1.3.1 *Astyanax mexicanus*

Astyanax mexicanus, commonly recognized as blind tetra in the aquarium trade, is one of the 86 fish species that inhabit cave regions and present troglomorphic traces (Romero and Paulson, 2001). The species was considered as a subspecies of *Astyanax fasciatus* (Melo et al. 2001), a group with an expressive karyotypic variability in which many cytotypes (2n 1⁄4 45 to 2n 1⁄4 48) are observed living in sympatry, with no apparent hybridism (Pazza et al. 2006). Cytogenetical studies in *A. mexicanus* were carried out in the 1960s, 1970s, and 1980s, describing the diploid number in three populations, one of which was mentioned as *Astyanax jordani*, an old synonym of *A. mexicanus*. These studies reported two different diploid numbers, 2n 1⁄4 48 (Post, 1965) and 2n 1⁄4 50 (Kirby et al. 1977; Vasil’ev, 1980). Additional detail on the chromosomal structure such as localization of genes and DNA sequences of this species is very limited. *A. mexicanus* rapidly developed into an attractive

model of evolutionary biology and eye development studies after the suppressed Pax6 gene was found to be involved in the absence of sight (Tian, 2005; Jeffery, 2001). The cave fish *A. mexicanus* exhibits certain unique behavior and distinguished morphological and physiological features such as loss of pigments, degeneration of eyes, efficient metabolism (Dowling et al. 2005) and ultra-sensitivity to chemicals and mechanicals stimuli (Panaram and Borowsky, 2005), making it more exciting model to answer evolutionary questions. Phenomenal findings from many studies concluded the identification of genes related to eyes development (Jeffery et al. 2003), isolation of the quantitative trait loci type and detection of (Protas et al. 2006), population studies about the natural hybridism between the surface and cave forms (Mitchell et al. 1977) and indication of low levels of heterozygosity by genetic and biochemical studies in the subterranean populations (Panaram and Borowsky, 2005; Avise and Selander, 1972; Borowsky and Wilkens, 2002). Phylogeography results gathered from the mitochondrial DNA sequences of cave and surface populations of *A. mexicanus* indicated two events of colonization in the North American continent. As a result of biogeographical studies, the cave populations can be classified into two main categories: strongly eye and pigment reduced (SEP) and variable eye size and pigmentation (VEP) (Wilkens, 1988). Recently, the genome of *A. mexicanus* transcriptome was assembled by McGaugh et al. (2014) and Hinaux et al. (2013) respectively and different genes associated with eyes degeneration were revealed. Based on the evolutionary significance and occurences of two Bs, we explored the genome of *A. mexicanus* to gain insights on their anonymous nature.

1.3.2 *Astyanax correntinus*

A. correntinus was reported for the first time by (H. Olmberg, 1891) more than one hundred years ago. This species was reviewed with newly collected material from the Rio Paraná near Corrientes city, northeast of Argentina (Mirande et al. 2006). Although these studies provide the fundamental characteristics for the taxonomy, no further research was conducted to demonstrate the genetic features. A later study (Paiz et al. 2015) revealed the basic cytogenetics and physical mapping of

ribosomal genes. We propose that the under studied *A. correntinus* can be a valuable model for B-chromosomal analysis and anticipate that current project using this species as a model will open new perspectives for the advance analysis in terms of understanding the evolution of B chromosome biology.

1.4 The grasshopper *Abracris flavolineata* as model organisms to study B chromosomes

Grasshoppers (Orthoptera: Acrididae) are among the most recognizable and familiar insects in terrestrial habitats around the world and represent a useful model system in entomology. Grasshoppers are particularly interesting for studying genome evolution because of their gigantic genome size, for example, 6.5 Gb of *Locusta migratoria* which is the largest animal genome sequenced so far (Wang et al. 2014). Here we present the grasshopper *Abracris flavolineata* as a model system in the present study, that has 2n=24/23 (females/males), with the XX/X0 sex chromosome system (Cella and Ferreira 1991). Seven subtelocentric, two metacentric and two submetacentric pairs, and the subtelocentric chromosome X make up the karyotype of *A. flavolineata*. In addition, this species displayed B chromosomes (Cella and Ferreira 1991, Bueno et al. 2013). Molecular markers were applied to understand their molecular composition and mechanisms of evolution (Bueno et al. 2013, Milani and Cabral-de-Mello 2014, Palacios-Gimenez et al. 2014). The two well-known grasshopper species used as model species to study B chromosomes are *E. plorans* and *Locust migratoria*. Over the years, information has been accumulated in these species regarding B chromosome population dynamics, their possible origin. However, limited knowledge has been obtained about the molecular composition of B chromosomes in these species. The model *A. flavolineata* was selected in the present project in order to understand the evolutionary genomics and functional mechanisms of B chromosomes in the grasshopper group of insects.

2. Hypothesis

B chromosomes tend to gain sequences which may play an important role in driving their own transmission, survival and maintenance inside the cell.

3. Objectives

3.1 General Objectives

To perform large scale genomic analysis of B chromosomes in diverse organisms.

3.2 Specific Objectives

- To generate high coverage sequencing data based on Illumina next generation sequencing of B+ and B- genomes;
- To perform comparative analysis through whole genome alignments against reference genomes of the B+ and B- sequence data;
- To identify genomic regions localized on the Bs;
- To understand the genetic composition and function of Bs;
- To study the origin and evolution of Bs;
- To report a comprehensive list of B chromosome genes.

4. Material and Methods

4.1 Model Organisms

The specimens of *Astyanax mexicanus* were obtained from a stock established from the local trade in Sao Paulo, Brazil and deposited in the fish room of the Integrative Genomics Laboratory of UNESP—Sao Paulo State University. *Astyanax correnrinus* were collected from natural habitat in the Iguassu River, in the stretch with around 25 km between downstream of the Iguassu Falls and its mouth on the Paraná River. These species were selected based in the previous description (Beuno et al. 2013) of B chromosomes occurrence. *A. flavolineata* adult individuals were collected in Rio Claro/SP, Brazil.

The experimental research on animals here employed agree with ethical principles in animal research adopted by the Brazilian College of Animal Experimentation and was approved by the Biosciences Institute/UNESP, Sao Paulo State University ethic committee on use of animals (Protocol no. 769-2015).

4.2 Karyotyping and genomic DNA extraction

All the specimens were anesthetized and dissected by an overdose of Benzocaine. The chromosome preparations of *Astyanax* were obtained from anterior kidney cells using 0.02% colchicine treatment for 40 to 30 minutes. The *A. flavolineata* chromosomes were obtained from male testis follicles and female gastric caeca, which contain mitotic chromosomes, according to the procedure described by Castillo et al. (2011). The procedure involved classical cytogenetics using a Giemsa stain. Thirty metaphases spreads from each individual were analyzed and ten the best mitotic metaphases were used to measure karyotypes. The chromosomes were classified as metacentric, submetacentric, telocentric or acrocentric. Individuals carrying B (B+) and those without B (B-) chromosomes were identified by karyotypic analysis.

Genomic DNA was isolated by phenol chloroform method as proposed by Sambrook and Russel (2001). These DNA samples were then analyzed on agarose gel to verify the intactness, and quantified by nanovue spectrophotometer, Qubit Flourometer to obtain information about concentration as required for sequencing.

4.3 Illumina Next-Generation Sequencing

After performing quality control (QC), qualified samples were proceeded to library construction. Males samples (including 0B, 1B and 2B) of model organism (Table 1) were sequenced using HiSeq Illumina. The sequencing library was prepared by random fragmentation of the DNA samples, followed by 5' and 3' adapter ligation. Separate libraries were constructed for each individual using the TruSeq DNA PCR-Free kit and sequencing was done with paired readings of 151 bp.

Table 1. Biological **s**amples used for next generation sequencing in the current work

Model species	**Gender**	**Tissue (for DNA extraction)**	**Sequencing samples**		
A. mexicanus	Male	Muscle	0B	1B	2B
A. correntinus	Male	Liver	0B	1B	
A. flavolineata	Male	Muscle	0B	1B	2B

4.4 Pre-processing and quality control of NGS data

Raw data from Illumina's HiSeq machine was processed with Illumina software to generate Fastq files. Fastq files contain read sequences and quality scores. Sequencing reads were analyzed by quality control tool FastQC followed by quality filtering based on the sequence quality score, adaptors trimming, filtering out short or unpaired sequences and trimming low quality bases using the Trimmomatic tool (Bolger et al. 2014). Specific parameters were set in the commands according to requirements for removing adapters and poor quality reads as per

FastQC report. Filtering of reads was performed by FASTx toolkit (Hannon, 2010) using parameters set to quality number 28 and percentage value 80 for alignments.

4.5 Genome assemblies and alignments

The Illumina reads were aligned to reference genomes using Bowtie2 (Langmead and Salzberg, 2012) with the "sensitive" option. The *A. mexicanus* genome (http://www.ensembl.org/Astyanax_mexicanus/Info/Annotation) was used as reference genome for alignments of *A. mexicanus* reads datasets. The genome assembly of *A. mexicanus* was downloaded from Ensembl and used as a reference genome for alignments of *A. mexicanus* and *A. correntinus* B+ and B- reads. To create reference genomes of *A. flavolineata* and *A. correntinus*, we assembled their Illumina reads using SoapDenovo (Lou et al. 2012). The evaluation of the generated scaffolds was obtained using QUAST software (Gurevich et al. 2013) by computing several metric values (length, number, length variation, N50, gap length). The B+ and B- reads were mapped against *De novo* assemblies. The output aligned files in sam (Sequence Alignment Map) format were manipulated using samtools (Li et al. 2000) for visualization the coverage tracks on Jbrowse (Skinner et al. 2009). The Coverage, read tracks and alignment for sequencing data was uploaded at the Sacibase database (www.sacibase.ibb.unesp.br).

4.6 Coverage based identification of B blocks

The identification of sequences present on B chromosomes was performed using statistical parameters and aligned reads coverage comparison between B+ (genomes with B) and B- (genomes without Bs), as proposed by Valente et al (2014). The sites with at least 15X reads coverage in B+ and B- genomes were selected, and the mean B+/B- coverage (MC) was calculated. Then, normalized coverage (NC) was obtained as: NC = (Raw coverage / Region size) / MC. The mean ratio (MR) and standard deviation (MRSD) for the genome region with most similar size and raw coverage, apparently not containing B sequences, were obtained with

NC. Next, the B+/B- regions with coverage bellow MC / 2 were removed and B+ ratio (BPR) was calculated with NC B+ / NC B-. Regions with BPR >= MR + (SD * N) were selected, were N is the number of SD required to determine a block. This way we were able to set an estimated threshold for detecting the extra copy of A chromosome sequences in B+ genome which can be regarded as putative B chromosomal sequences known as “B-blocks”. We used a script in house for identification of these blocks. The B-blocks were constructed using two levels (100 bp and 1kb) of tolerances; for example a level of 100 bp means that the B sequences within 100 bp regions and that have meant ratio greater than or equal to established value can be considered one part of the same block. In this way, four different sets of B-blocks (0 stdv and plus 2 stdv, both with tolerance of 100 bp and 1 kb) were obtained for further analysis. The B-blocks were manually visualized using J-browser (Skinner et al. 2009, https://jbrowse.org/) and comparative plots were created using Bioconductor (https://www.bioconductor.org/) package in R.

4.7 Analysis of protein coding genes located in the B chromosome.

To search for protein-coding genes residing in the B chromosome, we performed the following analysis. Firstly the transcriptome assemblies of *A. mexicanus* (used as reference for *A. mexicanus* and *A. correntinus*) and *Locust migratoria* (used as reference for *A. flavolineata*) were retrieved from NCBI database. Against these reference transcriptomes, we mapped the reads obtained from the B- and B+ genomes, using Bowtie2 and used an available custom script (https://github.com/fjruizruano/ngs-protocols/blob/master/count_reads_bam.py) to count the number of mapped reads as a measure of abundance. We selected coding sequences (CDS) putatively being located in the B chromosome on the basis of the following two criteria: 1) the sum of mapped reads (adding those from B- and B+ gDNA libraries) should be 40 or higher, and 2) log 2 of the quotient between the number of mapped reads in the B+ and B- gDNA libraries (B+/B-) was equal to or higher than 1. For example, if a single-copy gene would have two copies in a diploid 0B genome, whereas, if each B chromosome would carry one copy then the 2B

genome would carry four copies, i.e. two times more copies than the 0B, so that log 2 (2) = 1. We extracted those CDS in the *A. mexicanus* and *L. migratoria* transcriptomes that were aligned to the B+ and B- reads. We searched for homologous sequences with BLASTN using the extracted CDS as queries against the references gene annotations. These CDS that did not map to genes were then searched for repeats using RepeatMasker. The remaining CDS that were neither aligned to genes nor repeats were termed as non annotated or unknown sequences. We merged the complete genes and repeats annotations and produces graphics using R scripts. The CDS present on the B chromosome were extracted applying a cutoff value of log 2.

4.8 Repeats and Genes Identification and Annotation of B-blocks

In order to acquire an overview about the types of transposable elements, the “B blocks” were annotated using RepeatMasker 3.3.0 (Smit el al. 2013) software. The annotation follows the universal classification (Wicker et al. 2007). The repeats were masked using the reference database of metazoa. We assayed for under-representation of TE superfamilies using equation: (percentage of the TE family in the genome * 100)/Total repeat content in the genome) as described by (Mcgaugh et al. 2014). We also performed a comparative repeats composition analysis between A and B chromosomes. Results were parsed by perl script to depict the relative abundance of repeat classes using the RepeatMasker outfiles. The repeat landscapes were generated with the RepeatMasker “calcDivergence-FromAlign.pl” and “createRepeatLandscape.pl” utility scripts.

We also annotated B-blocks for genes by comparing them to the reference gene sets of close species downloaded from NCBI databases. The reference genes sets consists of *A. mexicanus* for *Astyanax* species and *Drosophila melanogaster* for *A. flavolineata.* These references were selected on the base of the complete representation of genes and high quality chromosomes level assembly. We calculated the integrity and gene length of all B-genes found in the B-blocks by combining all DNA pieces related to the same gene (each “piece” is a different gene length in each list of blocks). The total gene length was recovered by the sum of all pieces. The integrity percentage of each B-gene was calculated

comparing its length to the corresponding gene length in the annotation of the reference genomes. Finally, the integrity percentage for each gene was determined and the genes were categorized in different groups (from 0-100%) on the basis of integrity percent. The BLAST2GO pipeline (Conesa et al. 2005) was applied over both the references and B-genes and Fisher's exact test (using corrected P values to control the false discovery rate) was used to compare the set of Gene ontology (GO) terms from B-genes for a GO enrichment analysis. GO-enriched terms and the complete GO annotation from the genes with an integrity range (from 50% to 100%), were analyzed. Uniprot and NCBI function descriptions, as well as reports available in the literature, were also used to understand the function of each gene. The enrichment graphs were created using Revigo (Supek et al. 2011).

4.9 Primers Designing

Primers for selected repeat elements were designed by NCBI/Primer-Blast (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) and PrimerQuest (https://www.idtdna.com/Primerquest/Home/Index) tools online. Primers quality was evaluated by PCR Primer Stats program (http://www.bioinformatics.org/sms2/pcr_primer_stats.html). Primers designed for FISH probes experiments are listed in supplementary table 1. The primers were suspended and homogenized in DNAase free water according to recommended concentration. We selected some of TEs, explored by bioinformatics analysis and did polymerase chains reaction (PCR) and used the amplified PCR products as probes FISH mapping. DNA fragments obtained by PCR were sequenced (Sanger et al. 1977) using an ABI Prism 3100 automatic DNA sequencer (Applied Biosystems, Foster City, CA, USA) with a DynamicTerminator Cycle Sequencing Kit (Applied Biosystems) as per the manufacturers' instructions. Nucleic acid sequences were subjected to BLAST (Altschul et al. 1990) searches at the NCBI website (http://www.ncbi.nlm.nih.gov/blast) to check for similarities to other previously deposited sequences to confirm if they correspond to the expected genomic regions.

4.10 Fluorescent in situ hybridization (FISH) and quantitative Real-Time PCR (qPCR)

Probes for FISH were obtained via PCR directly from the genome DNA using both forward and reverse primers. The probes were labeled with Digoxigenin-11-dUTP (sigma) and checked by agarose gel to confirm binding. FISH was performed under high stringency conditions using the method described by Pinkel et al. (1986). Chromosomes suspension was dropped on cleaned and sterilized slides and slides were then incubated with RNAse (50 µg/ml) for 20 minutes at 37°C. The pre-hybridization conditions were different according to the probes used and the chromosomal DNA was denatured in 70% formamide for 15 seconds at 65°C. For each slide, 30 µl of hybridization solution (containing 200 ng of each labelled probe, 50% formamide, 2x SSC and 10% dextran sulphate) was denatured for 10 minutes at 70°C, then dropped onto the slides and allowed to hybridize overnight at 37°C in a moist chamber containing 2x SSC. Post-hybridization, all slides were washed in 50% formamide for 20 min at 42°C, followed by a second wash in 0.1x SSC for 15 min at 60°C and a final wash at room temperature in 4x SSC, 0.5% Tween for 10 min. Probe detection was carried out with anti-digoxigenin-rhodamine (Roche), and the chromosomes were counterstained with DAPI (4’,6-diamidino-2-phenylindole, Vector Laboratories) and analyzed using an optical photomicroscope (Olympus BX61). The images were captured using an Olympus DP72 system. and were optimized for brightness and contrast using Adobe Photoshop CS2.

Genomic block characteristics of the B genome were selected for the construction of qPCR primers (supplementary table 1) to screen *A. mexicanus* genomes for the presence of B chromosome and to quantify their abundance. The qPCR of gDNA was used to calculate the gene dose by a #Ct method of relative quantification (Nguyen et al. 2013). Gene dosage ratios (GDR) of the target genes were compared with 45S rDNA gene used as a reference (that has same number of copies in B- and B+ genomes). RT-qPCR was carried on StepOne Real-Time PCR Systems (Life Technologies, Carlsbad, CA). The target and reference genes were analyzed simultaneously in triplicates of three

independent samples. The cycling conditions were 95 # C for 10 min; 45 cycles of 95 # C for 15 s, and 60 # C for 1 min. Specificity of the PCR products was confirmed by analysis of the dissociation curve.

4.11 Analysis of microdissected B chromosomes.

In addition, to the aforementioned analysis of B chromosomes in the initially proposed model species, we further added two more species to test the hypothesis given in this project. For this purpose, the NGS Illumina data for the microdissected B-chromosomes of the additional species *E. plorans* (grasshopper) (Montiel et al. 2014), *Lates calcarifer* (Asian seabass fish) (Vij, et al. 2016), *Apodemus flavicollis* and Apodemus *peninsulae* (mouse) (Rajicic et al. 2017) were downloaded from the public NCBI-SRA database. We chose to analyze these Bs because the complete genomic composition of these microdissected Bs has not been understood and genes search, gene ontologies as well as repeat annotation has not been performed. The NGS reads with quality score less than 20 bp were removed using FASTX-Toolkit, adapter sequences and low quality bases were trimmed using cutadapt pipeline (Martin, 2011) and Trimmomatic (Bolger et al. 2014). Clean reads were mapped to the respective assembled reference genomes using Bowtie2 with default parameters. The references genomes consist of *L. calcarifer* (Vij et al. 2016), *Migratory locust* (Wang et al. 2014) and *Mus musculus* genome assemblies which were retrieved from the NCBI/Genome database. Successfully mapped reads were chained together across gaps less than 10 kb to form B chromosome pseudo-scaffolds. Pseudo-scaffolds were assembled using cap3 (Haung et al. 1999) to remove redundancy and the generated contigs were manually checked to reduce potential mis-assemblies. The micro dissected Bs assemblies were performed on the basis of pseudoscaffolding strategy as proposed by Vij et al. (2016). The assembled B microdissected chromosomes were then functionally annotated and gene enrichment analysis was performed as described in the earlier steps. The reference set of genes and proteins were retrieved from Ensembl browser (https://www.ensembl.org/index.html) and we used BLASTn and BLASTx for homologous genes annotation . The references consisted of

Gasterosteus aculeatus, *Drosophila melanogaster,* and *Mus musculus* for Bs analysis of *L. calcarifer*, *E. plorans* and *Apodemus* respectively selected on the bases of completeness of genes annotations.

4.12 Comparative and evolutionary genomics

In attempt to identify the A chromosomes that the B chromosome were derived from, we mapped the B blocks against the reference chromosome-level genome assemblies and used Circos (Krzywinski et al. 2009) to generate graphics representation. To find putative regions of homology between ancestral sequences of B blocks, we identified collinear regions of sequence similarity to infer synteny and generated a dotplot of the results. For this analysis we chose the largest blocks with size greater than 1,500 bp and did a comparison between the ancestral A and B sequences. We used CoGE SynMap (Haug-Baltzell et al. 2017) to generate a syntenic dotplot and identify the pattern of genomic evolution.

An overview of the methodology described above is illustrated in the figure 1.

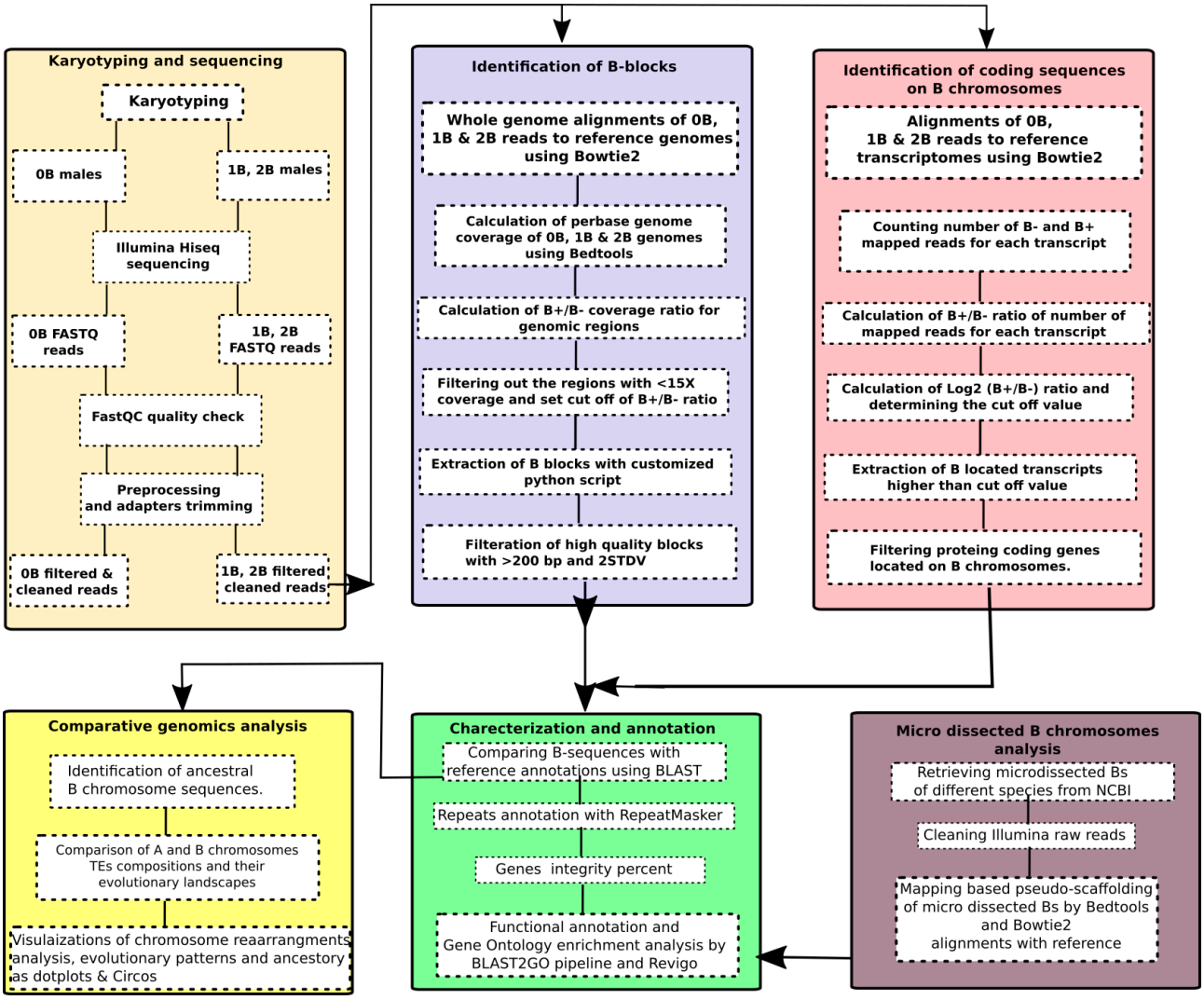


Figure 1. A workflow of steps applied in the present study during the procedure of genomics analyses in understanding B chromosome biology

5. Results

5.1 Karyotypes and Illumina NGS data

A large sub-metacentric B chromosome similar to the A chromosome pair was found in *A. correntinus* while a tiny, apparently dot shaped acrocentric type of B microchromosome was observed from the karyotypics analaysis of *A. mexicanus* (figure 2). The 2n (without B) was 36 and 50 chromosomes for *A. correntinus* and *A. mexicanus*, respectively. In some individuals of *A. mexicanus* we observed 2B microchromosomes. Interestingly, all the karyotyped data we analyzed of around 50 samples showed that B in *A. mexicanus* is more prevalent in males, therefore indicating its possible specificity in males since no B+ female was found. For the grasshopper *A. flavolineata*, we considered data from Bueno et al., (2013) that reported 1 or 2 submetacentric B chromosomes in individuals sampled at the Rio Claro/SP population.

A total of 8 samples of all three model species were sequenced which generated data approximate total of 124x, 82x and 28.1 coverage for *A. mexicanus*, *A. correntinus* and *A. flavolineata* respectively as given in the table 2. The paired end Illumina reads size was 151 with an average insert size of 550 bp. The mapping of filtered cleaned reads with high quality after pre-processing steps to reference genomes resulted an overall alignment rate of around 92%, 91% and 95% for *A. mexicanus*, *A. correntinus* and *A. flavolineata* reads mapped to respective genomes respectively.

Table 2. The NGS data generated for the analysis of B chromosomes in the present project

IDs	Sample	Number of raw reads	Reads length	Coverage	Number of filtered reads	Coverage after filtration	Genome size estimated
ame-1475-0b	*A. maxicanus* (0B)	265,761,096	151 bp	40.1x	265,374,050	40.07	1 Gb
ame-1465-	*A. maxicanus*	265,789,908	151 bp	40.13x	265,402,766	40.07	

1b	(1B)						
ame-1466-2b	*A. maxicanus* (2B)	358,144,230	151 bp	54.07	357,697,838	54.01x	
aco-2220-0b	*A. correntinus* (0B)	388,448,208	151 bp	31.46x	387, 817,038	31.1x	1.8 Gb
aco-2749-1b	*A. correntinus* (1B)	506,354,302	151 bp	41.01x	505,608,854	40.95x	
afl-1h-2h-4h-0b	*A. flavolineata* (0B)	836,480,126	151 bp	19.95x	834,822,322	19.91x	6.3 Gb
afl-5h-6h-1b	*A. flavolineata* (1B)	453,566,904	151 bp	10.81x	452,662,368	10.79x	
afl-7h-8h-2b	*A. flavolineata* (2B)	431,189,476	151 bp	10.28x	430,174,870	10.26x	

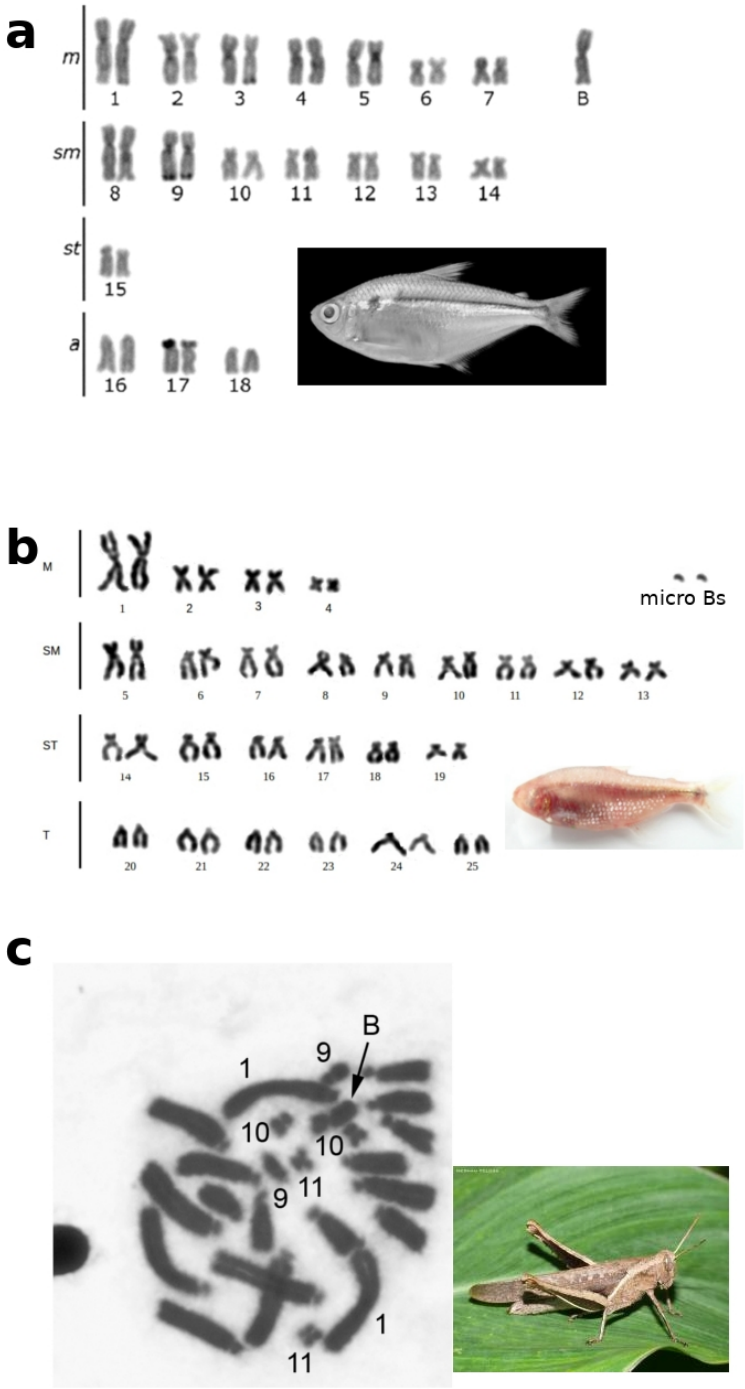


Figure 2. Classical Karyotypes of (a) *A. correntinus*, (b) *A. mexicanus* and (c) *A. flavolineata* (Bueno et al., 2013) after Geimsa staining. The characterization of regular A and extra B

chromosomes are shown. (Species images were downloaded from google).

5.2 Identification of B chromosome sequences: Genomic characterization, structure and composition of B chromosome.

Regions along the genomes of three model species that had significantly higher B+ coverage than B- coverage were identified (see Materials and Methods). Due to the homology between A and B chromosome sequence, most sequence reads derived from the B chromosome will align to their A chromosome homologs present in the reference genome. As a result, alignments of reads from a genome with a B chromosome will have regions of increased coverage compared to an alignment from a genome lacking a B. To improve the rate of alignments, and create a reference genome each for *A. correntinus* and *A. flavolineata* we performed the *de novo* genome assemblies which spanned a total size of 1.8 Gb and 6.3 Gb, respectively (supplementary table 2). The exact number of B chromosome blocks, as well as their exact boundaries in the reference sequence, varied depending on the parameters adopted for the analysis. The graphics of representative B-blocks as a few examples are shown as (supplementary figures 1 and 2) figures 3 and 4. A total of 509,028 and 257,784 and 9,845 number of B-blocks were recorded for *A. correntinus*, *A. mexicanus* and *A. flavolineata*, respectively. To test the abundance of these regions in B+ genome and validate the effectiveness of our coverage based identification of B chromosome sequences, we used qPCR for relative copy number quantification of selected B sequence blocks in *A. mexicanus* with 0, or 1 B genomes. The GDR was determined using qPCR results, which resulted the higher GDR in 1B genome as compared to 0B for all the total 10 representative blocks that were selected for this analyses thus confirming our NGS analyses. (figure 4).

For genes annotations of *A. correntinus* and *A. mexicanus*, the smallest blocks under 200 bp were filtered out leaving a total of 64,627 and 18,340, respectively. B-blocks ranged in length from 200bp to 10 kb, although there were multiple regions in the genome with multiple B blocks in close proximity, suggesting that a larger region was transferred to the B chromosome as a whole. A

comparison between the number of blocks for both species with respective size in base pairs is shown as figure 5a. The lower number of blocks in *A. flavolineata* was observed because of the low genome sequencing coverage. The sequencing of *A. flavolineata* did not yield a deep reads coverage due to its enormous genome size (6.7 Gb). Although we were not able to obtain a comprehensive list of B-blocks and high integral genes in this model species, because of the lower coverage problems, still we could find a considerable amount of sequences its B chromosome (supplementary figure 3). The repeats characterization using RepeatExplorer (Novak, 2013) pipeline concluded a total of 65% and 35% comprising of repeats in *A. correntinus* and *A. mexicanus* genomes respectively. A search for repetitive DNAs in the B chromosome blocks found different types of repeats distributed across various blocks. In the blocks of *A. correntinus*, approximately 5.78% are retroelements, 3.38% are LTR elements with higher amount of Gypsy/DIRS1 and 2.94% are DNA transposons dominated by Tc1-IS630-Pogo elements. In *A. mexicanus*, the blocks comprise of 7.7% retroelements, 3.14% of LTR and 15.67% of DNA transposons. A comparison between the repeat contents of *A. mexicanus* and *A. correntinus* is given as figure 5b. Although, the B blocks of *A. flavolineata* is not complete due to low quality reads coverage, we were able to detect the abundance of R1/LOA/Jockey, Tc1-IS630-Pogo and Gypsy/DIRS1 possibly present on its B chromosome (supplementary figure 5c).

To identify the intact genes on the B chromosomes, we calculated an integrity score for each gene sequence annotated in the B-blocks. The majority of B-located genes of *A. correntinus* (91%) and *A. mexicanus* (93%) have integrity scores less than 50% (figure. 5c). The remaining 7-9% of genes are with integrity scores more than 50%. The NGS data analysis indicating the higher number of B-blocks and number of repeats and genes for *A. correntinus* as compared to *A. mexicanus* (figure 5b) clearly coincides with the karyotype data. The karyotyping has concluded that the macro B chromosome of *A. correntinus* is bigger in size than the micro B of *A. mexicanus.* Therefore the higher genomic contents is indicative to the size of B chromosome and another support of NGS analyses.

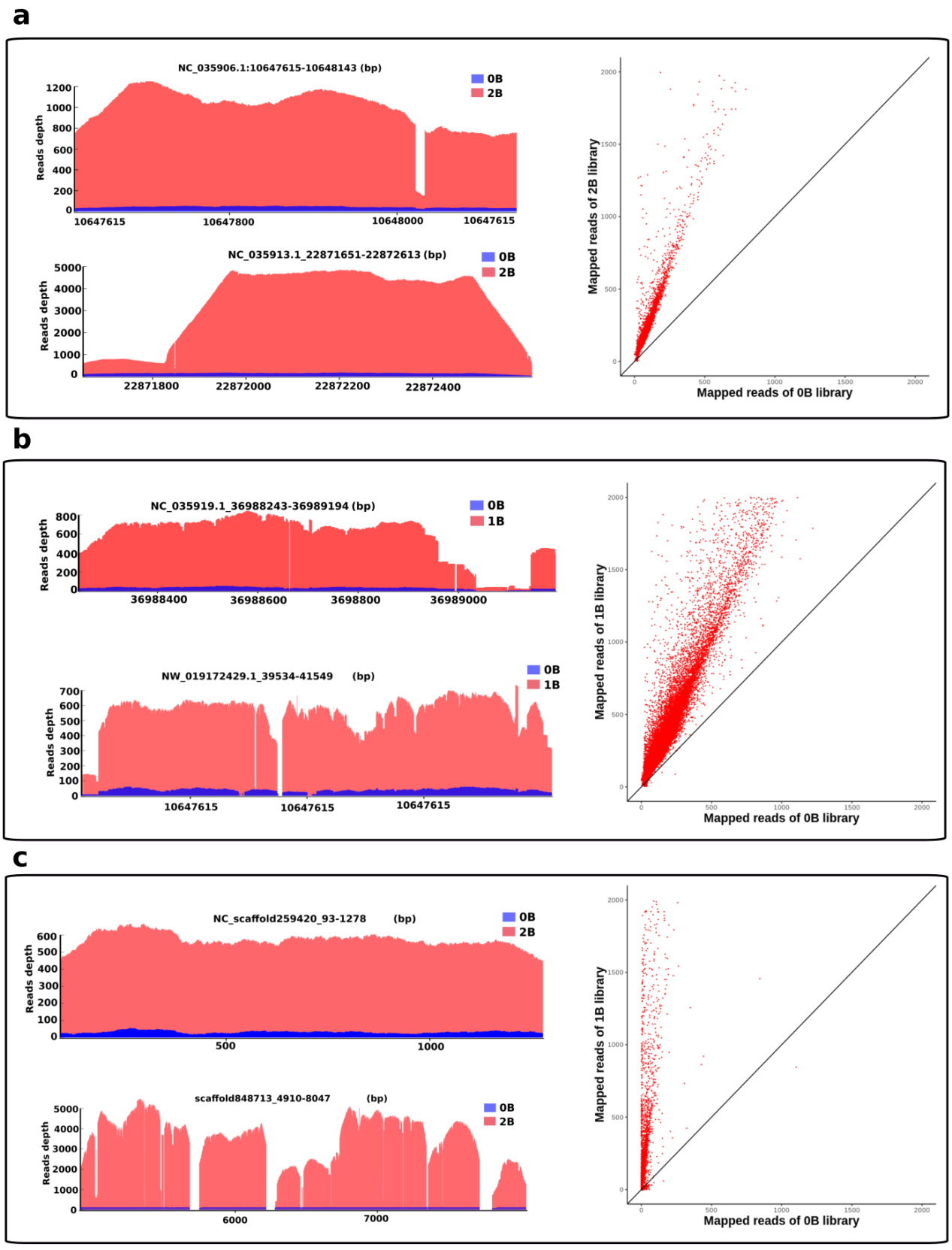

a
NC_035906.1:10647615-10648143 (bp)
0B
2B
Reads depth
1200
1000
800
600
400
200
0
10647615
10647800
10648000
10647615
NC_035913.1_22871651-22872613 (bp)
0B
2B
Reads depth
5000
4000
3000
2000
1000
0
22871800
22872000
22872200
22872400
Mapped reads of 2B library
Mapped reads of 0B library
2000
1500
1000
500
0
b
NC_035919.1_36988243-36989194 (bp)
0B
1B
Reads depth
800
600
400
200
0
36988400
36988600
3698800
36989000
NW_019172429.1_39534-41549 (bp)
0B
1B
Reads depth
700
600
500
400
300
200
100
0
10647615
10647615
10647615
Mapped reads of 1B library
Mapped reads of 0B library
c
NC_scaffold259420_93-1278 (bp)
0B
2B
Reads depth
600
500
400
300
200
100
0
500
1000
scaffold848713_4910-8047 (bp)
0B
2B
Reads depth
5000
4000
3000
2000
1000
0
6000
7000
Mapped reads of 1B library
Mapped reads of 0B library

Figure 3. The B-block plots showing the comparison between B- and B+ coverage. The panels (a), (b) and (c) represent *A. mexicanus*, *A. correntinus* and *A. flavolineata* respectively. On the right side the graphics are shown of representative examples of B-blocks depicting the significant difference between the coverage of B- (blue peaks) and B+ (red peaks). Notice the difference between the coverage of B+ is remarkably greater than B- indicating the amplified genomic region on the B chromosome. The X-axis and Y-axis represent reads depth and genomic position of the B-block. The blocks are named according to their position in respective genomes. The higher reads depth of B+ might be indicative to the duplicated copies of this region on the B chromosomes. On the left side the scatterplots are shown for corresponding species. Each red dot in these plots is a single block, with X-axis and Y-axis representing the number of mapped reads for B- and B+ genomic libraries. Notice that the approximately all dots (blocks) above the diagonal lines inclining towards the Y-axis provide strong evidence to the extracted B-blocks with higher reads coverage of B+ as compared to B-, thus confirming extra copies of these genomic sequences on B chromosomes. The term “0B”, “1B” and “2B” denote to absence, presence of one and two B chromosomes in the respective genomes.

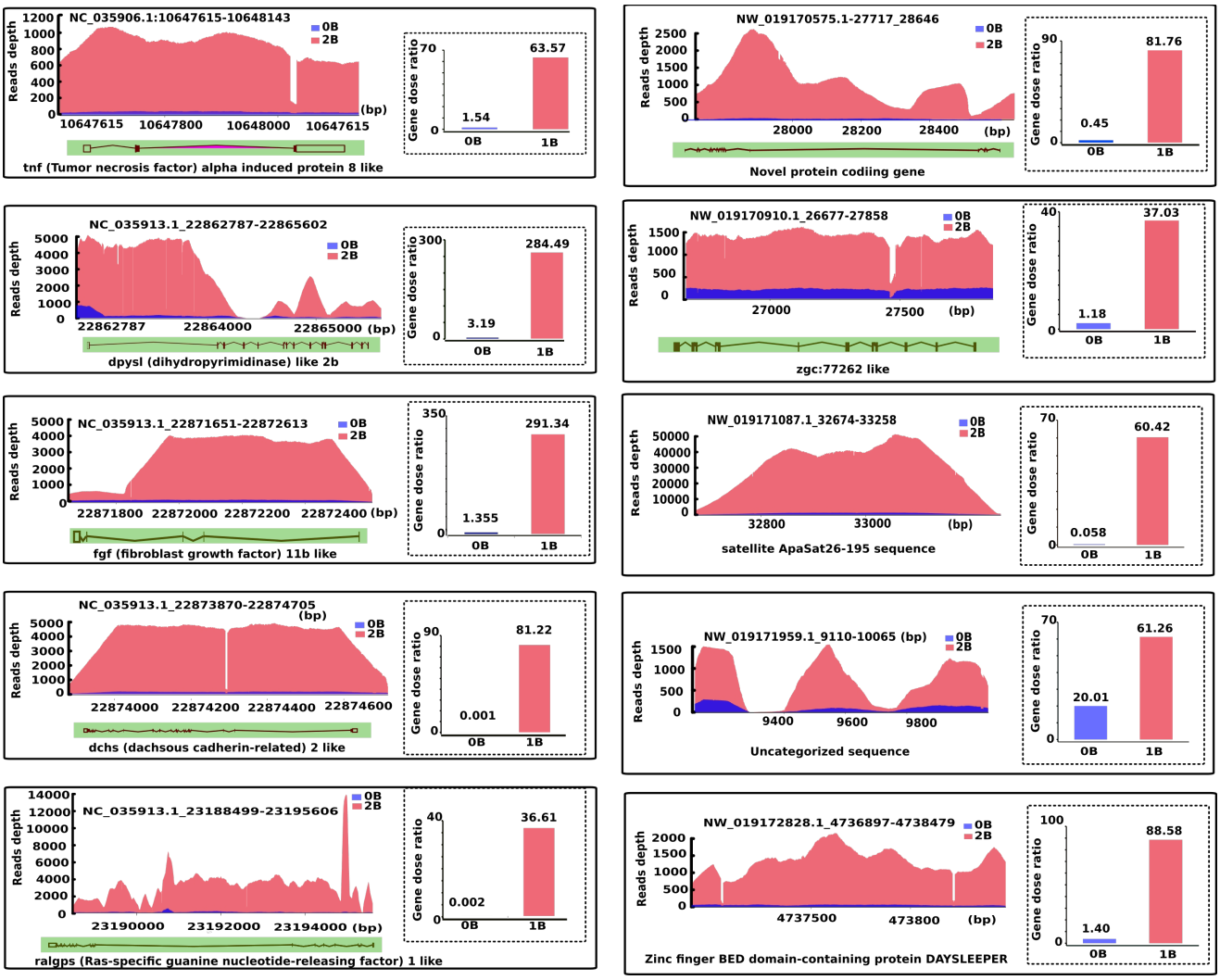


Figure 4. Coverage plots of representative B-blocks of *A. mexicanus* along with the respective GDR qPCR results comparison between B- and B+ genomes. The higher coverage and GDR in B+ genome indicates the duplicated copies of these sequences on the B chromosome. The annotation is also shown for each block. The coverage is correlated to the GDR ratio, as the significant difference between B- (blue) and B+ (red) genomes can be clearly noticed for the blocks shown here. The annotation bars of protein coding genes are also shown. The BLASTn alignments of these representative blocks to Ensemble annotation databases, resulted in several overlapping genes, such

as tnf (function: cell death), dpysl2b (function: microtubules binding activity), fgf11b (function: development, morphogenesis, other broad mitogenic and cell survival activities), Zinc finger BED domain daysleeper like (function: chromatin remodelling), zgc:77262 (function: mRNA splicing) ralgps1 (function: cytoskeleton organization) and dchs (function: cell adhesion). One of the important findings was the detection of satellite Apasat26-195 sequence which has also been recently reported on the B chromosome of *Astyanax paranae* (Duílio et al. 2017).

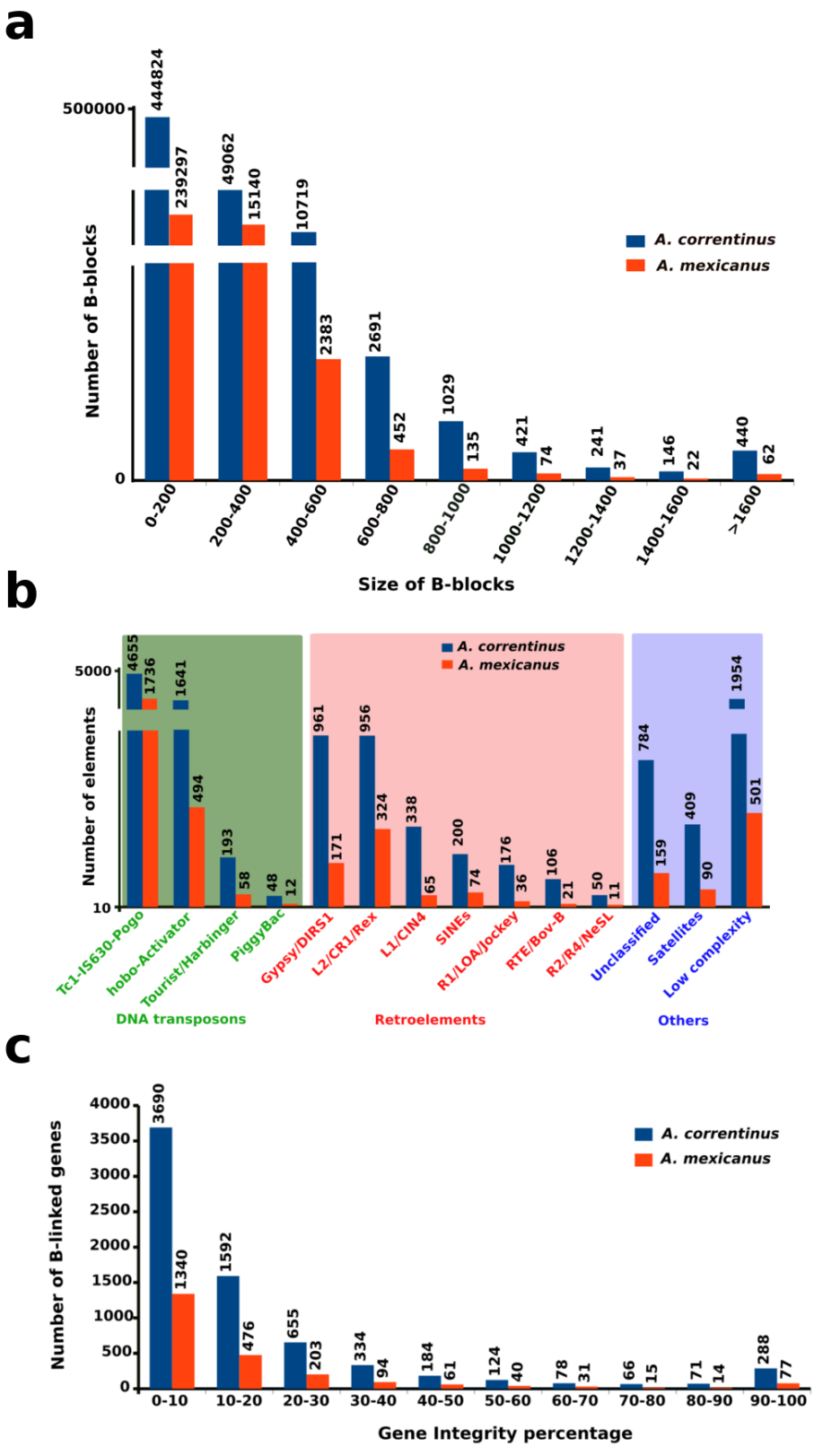
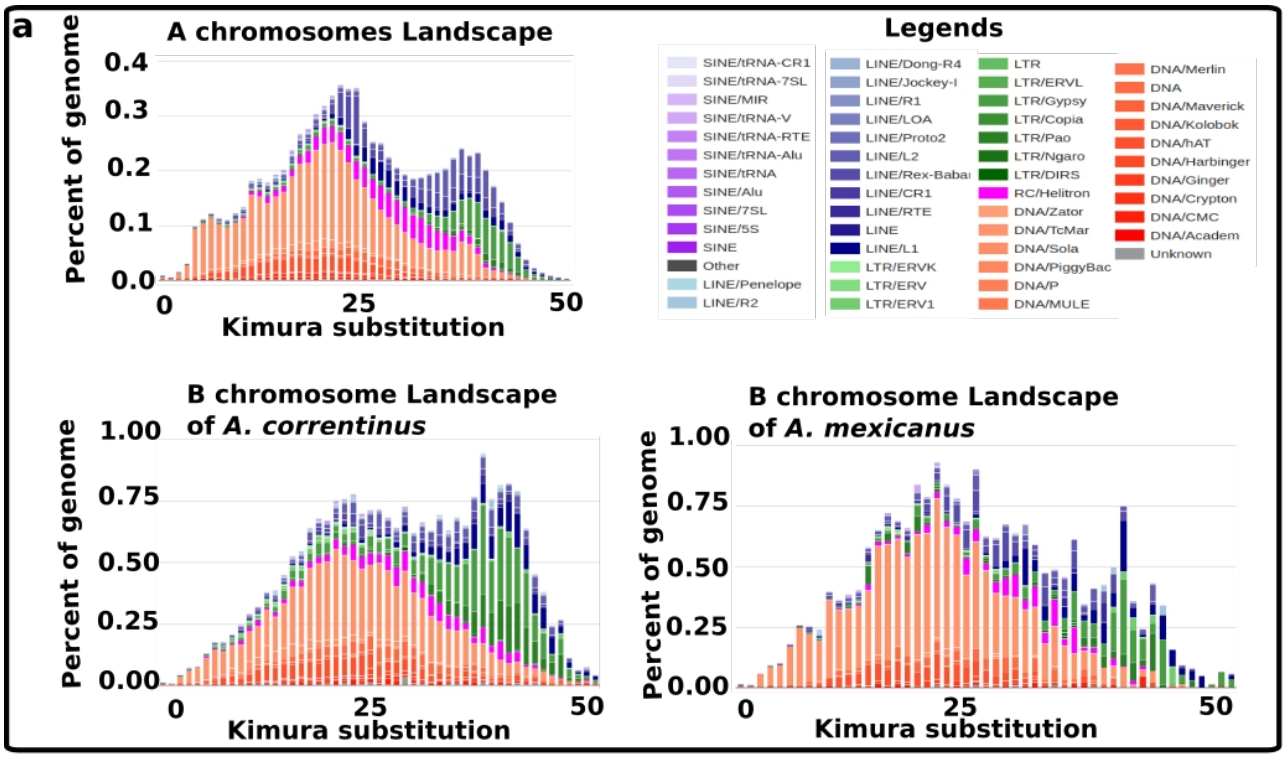

a
Number of B-blocks
500000
0
444824
239297
49062
15140
10719
2383
2691
452
1029
135
421
74
241
37
146
22
440
62
A. correntinus
A. mexicanus
0-200
200-400
400-600
600-800
800-1000
1000-1200
1200-1400
1400-1600
>1600
Size of B-blocks
b
Number of elements
5000
10
4655
1736
1641
494
193
58
48
12
961
171
956
324
338
65
200
74
176
36
106
21
50
11
784
159
409
90
1954
501
A. correntinus
A. mexicanus
Tc1-IS630-Pogo
hobo-Activator
Tourist/Harbinger
PiggyBac
Gypsy/DIRS1
L2/CR1/Rex
L1/CIN4
SINEs
R1/LOA/Jockey
RTE/Bov-B
R2/R4/NeSL
Unclassified
Satellites
Low complexity
DNA transposons
Retroelements
Others
c
Number of B-linked genes
4000
3500
3000
2500
2000
1500
1000
500
0
3690
1340
1592
476
655
203
334
94
184
61
124
40
78
31
66
15
71
14
288
77
A. correntinus
A. mexicanus
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
Gene Integrity percentage

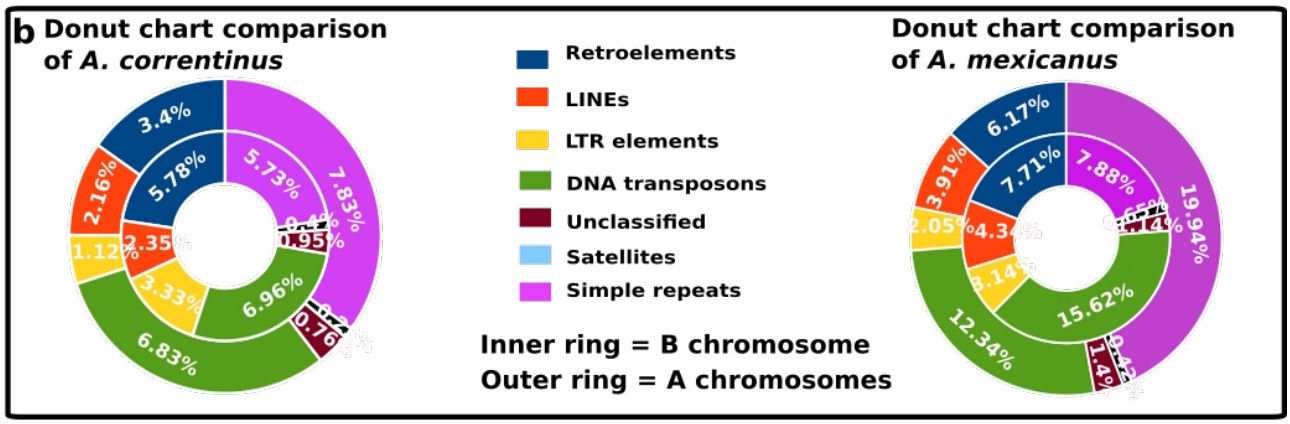
Figure 5. Characterization of genomic regions detected on the B chromosomes of *A. mexicanus* and *A. correntinus* and genomic contents of macro B versus micro B. *(*a). A comparison between both *Astyanax* species of B-blocks number versus blocks size range. The highest number of blocks were recorded for the size range of 0-200 base pairs. (b). Repeats annotations of b-blocks showing the abundance of DNA transposons on the B chromosome of *Astyanax*. (c). Results of gene integrity analysis showing the number of genes (Y axis) in each integrity percentage group (X axis). Notice the higher number of genomic contents in macro B (blocks, genes and repeats elements) of *A. correntinus* (blue bars) than micro B in *A. mexicanus* (red bars) that shows its larger size as detected experimentally. The graphs are scaled and breaks are inserted due to differences of data.

We also performed a comparative analysis to investigate the relative TEs abundance and detect any possible difference in their contents between the regular A chromosomes and B chromosomes in *Astyanax* species. Interestingly, we found that the Bs recorded a higher percentage of TEs, especially DNA transposons and LTR elements as compared to A genome (figure 6 a and b). The repeat landscapes of both A and B chromosomes has a noticeably larger amount of recent TE insertions as can be seen two peeks in the figure 6a reflecting a wave of transposition has occurred during the genome evolution. But the Bs in both species have gained higher copies as compared to A genomes, specially retroelements. Remarkably, the FISH mapping of some these representative elements (we tested), also confirmed our NGS analysis. We found dispersed signals of hybridization for the respective FISH probes of Gypsy and Tc-Mariner on the B chromosomes (figure 6c). Both FISH mapping and bioinformatics analyses showed that these elements are scattered throughout the genome of *Astyanax*. These elements appear widely distributed throughout all the chromosomes with some specific concentrations on certain regions. The plentiful marks of these element dispersed on almost any chromosome indicate a series of transposition events happened during the chromosomal evolution. Our results of FISH provides a cytogenetics validation of bioinformatics

analysis which concluded the high abundance of these elements and demonstrated the copious nature of these sequences in the genome as can be noticed in the repeat landscapes.

We also analyzed rRNA cluster by FISH mapping to check for its organization in *A. mexicanus* genome. Although the NGS annotation of B-blocks in *A. mexicanus* did not detected any 45S rRNA cluster indicating its absence on the micro B chromosomes. However, we performed FISH experiment for confirmation of NGS result and it was confirmed that no micro B markings were identified and therefore micro B is missing 45S rRNA cluster (supplementary figure 4). FISH probe of 45S rRNA demonstrated that *A. mexicanus* has eight site-bearing regular chromosomes, and its distribution was preferentially terminal in the short arms of a few A chromosome pairs. The clusters of 45S rRNA probe in the nucleolus organizer regions (NORs), suggest the intense accumulation. Furthermore, the concentrated localization of this sequence on the sites of A complement confirmed the previously reported patterns and added valuable information regarding its evolutionary importance. Also, our NGS analysis of B chromosome composition showing, no sign for the presence of 45S rRNA, were found to be valid according to these cytogenetics experiments. For confirmation of the localized probes (constructed from amplified PCR products,), we performed Sanger sequencing (see materials and methods), and ultimately was searched using BLASTn for the confirmation. The Sanger sequences of each probe were also annotated by RepeatMasker. Both approaches gave a similarity hits of 98-100% identity.


a
A chromosomes Landscape
Percent of genome
Kimura substitution
Legends
SINE/tRNA-CR1
SINE/tRNA-7SL
SINE/MIR
SINE/tRNA-V
SINE/tRNA-RTE
SINE/tRNA-Alu
SINE/tRNA
SINE/Alu
SINE/7SL
SINE/5S
SINE
Other
LINE/Penelope
LINE/R2
LINE/Dong-R4
LINE/Jockey-I
LINE/R1
LINE/LOA
LINE/Proto2
LINE/L2
LINE/Rex-Baba
LINE/CR1
LINE/RTE
LINE
LINE/L1
LTR/ERVK
LTR/ERV
LTR/ERV1
LTR
LTR/ERVL
LTR/Gypsy
LTR/Copia
LTR/Pao
LTR/Ngaro
LTR/DIRS
RC/Helitron
DNA/Zator
DNA/TcMar
DNA/Sola
DNA/PiggyBac
DNA/P
DNA/MULE
DNA/Merlin
DNA
DNA/Maverick
DNA/Kolobok
DNA/hAT
DNA/Harbinger
DNA/Ginger
DNA/Crypton
DNA/CMC
DNA/Academ
Unknown
B chromosome Landscape of A. correntinus
B chromosome Landscape of A. mexicanus


b
Donut chart comparison of A. correntinus
Donut chart comparison of A. mexicanus
Retroelements
LINEs
LTR elements
DNA transposons
Unclassified
Satellites
Simple repeats
Inner ring = B chromosome
Outer ring = A chromosomes

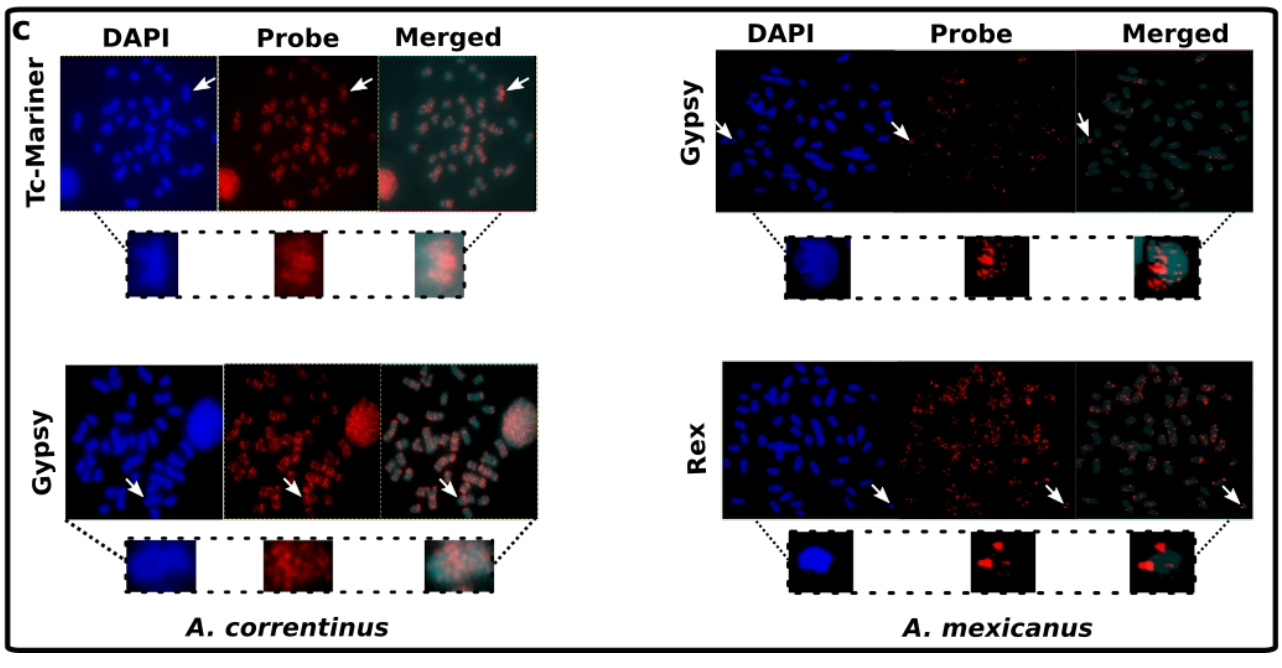

c
DAPI
Probe
Merged
Tc-Mariner
Gypsy
Gypsy
Rex
A. correntinus
A. mexicanus

Figure 6. Comparative analyses of TEs composition between A and B chromosomes. (a). Comparison of repeats landscapes of TEs provide insights on their evolutionary history in both A genome and the B chromosomes of *A. correntinus* and *A. mexicanus.* The X-axis shows the percent of TEs in the genome while Y-axis represents the Kimura distances ranged from value 0 representing recent TE copies to 50 for the old TE insertions. Two peaks of insertions can be noticed that indicate the recent wave of DNA transpositions in the genome of Astyanax genus. The higher abundance of LTRs (colored green) and retroelements (colored blue) in the B chromosomes landscapes can also be observed. (b). Donuts charts show the comparison of repeats composition between the As and B. The outer and inner rings depict A and B chromosomes respectively. Again, the higher percentage of LTRs and retroelements confirm their relative abundance on the B as compared to A chromosomes. (c). FISH mapping of representative elements validated the comparative NGS analysis. Metaphases of *A. mexicanus* and *A. correntinus* with B chromosomes were analyzed for organization of Tc Mariner, Gypsy and Rex which identified a dispersed pattern among diverse chromosomes including Bs. Magnified view of B chromosomes are shown with presence of markings of corresponding elements. The abundant signals of these TEs are indicative of their copious nature in *Astyanax* genome and parallel with the landscapes analyses.

5.3 Protein coding genes detected on B chromosomes

The mapping of the Illumina reads from the B- and B+ genomes on the coding sequences (CDS) of transcriptomes revealed a total of 38,071 for *A. mexicanus*, 34,301 for *A. correntinus* and 3,916 for *A. flavolineata* with more than 40 reads mapped for both genomes. Graphical representation of the B- and B+ showed the presence of some CDSs being over-represented in the B+ genomes (figures 7, 8 and 9). Remarkably, a total 100 and 53 CDSs showed a $\log_2$ 2B/0B quotient >1.5 for *A. mexicanus* and *A. flavolineata* respectively, and 436 CDS for *A. correntinus* with a $\log_2$ 1B/0B > 1 i.e. the expected value if each B chromosome carried at least one copy of the CDS (see materials and methods). Annotation revealed that most of these CDSs were orthologous to different protein-

coding genes in the reference set of genes while others were also identified as repeat elements. Some of the CDS did not align to the references, therefore we termed them as non-annotated or unknown. The annotation detected several novel genes on the B chromosomes of *Astyanax* species. The protein coding genes with the highest $\log_2$ quotient representing highly confidence of B chromosome presence are listed a table 3, table 4 and table 5 for each model species. The coverage pattern for these genes were also visualized and confirmed the higher peaks for B+ as compared to B- genomes hence providing evidence of extra copies present on B chromosomes.

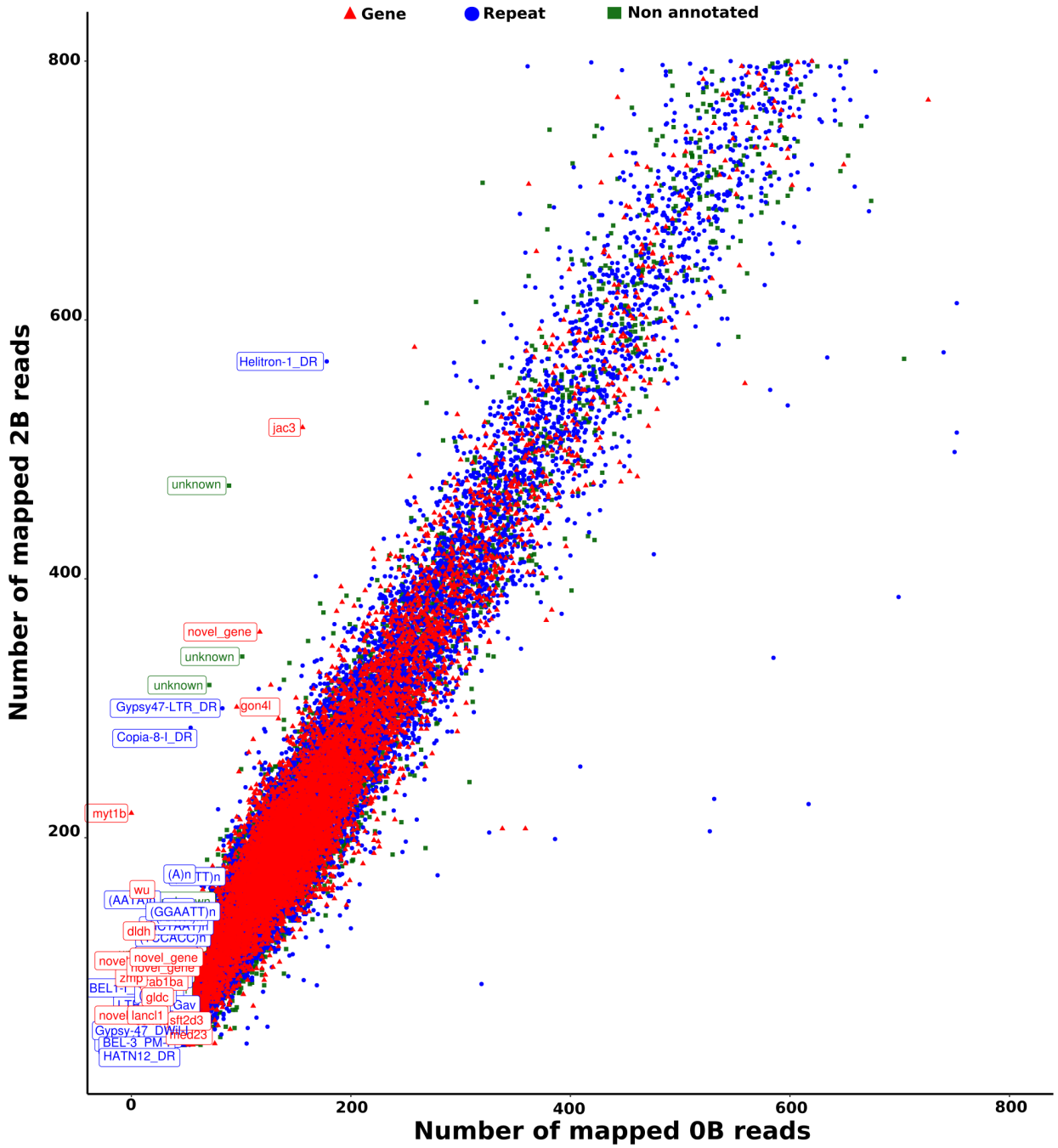


Figure 7. Identification of protein-coding genes located in B chromosomes of the *A. correntinus*, using the number of mapped reads that map to the CDSs found in the transcriptome, in the 0B (X axis) and 1B (Y axis). Each dot represents a coding sequence with only those labeled that recorded the $\log_2$ greater than 1. The plot is limited for 800 mapped reads to optimize the visualizations.

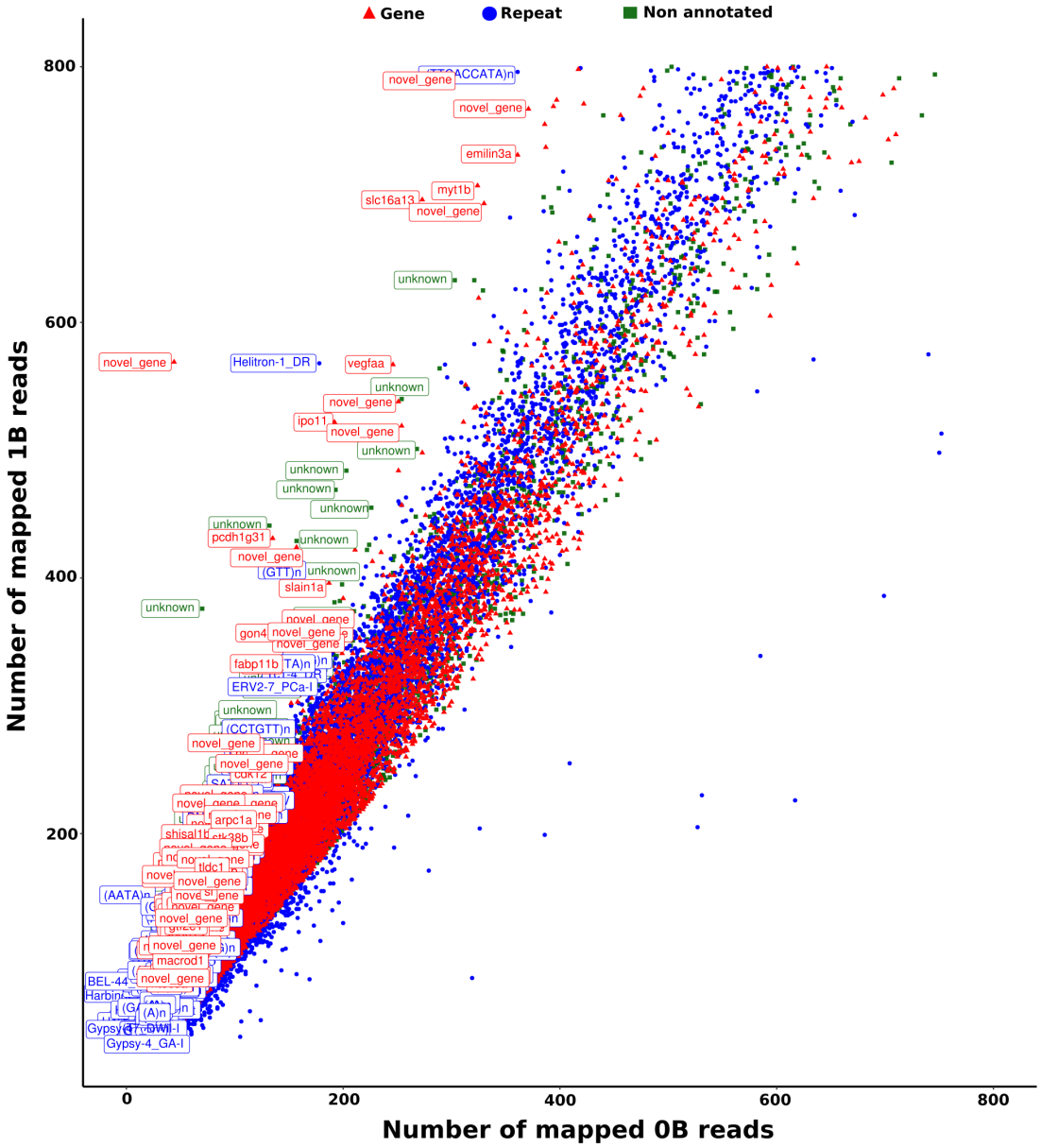


Figure 8. Identification of protein-coding genes located in B chromosomes of the *A. mexicanus*, using the number of mapped reads that map to the CDSs found in the transcriptome, in the 0B (X axis) and 2B (Y axis). Each dot represents a coding sequence with only those labeled that recorded the $\log_2$ greater than 1.5. The plot is limited for 800 mapped reads to optimize the visualizations.

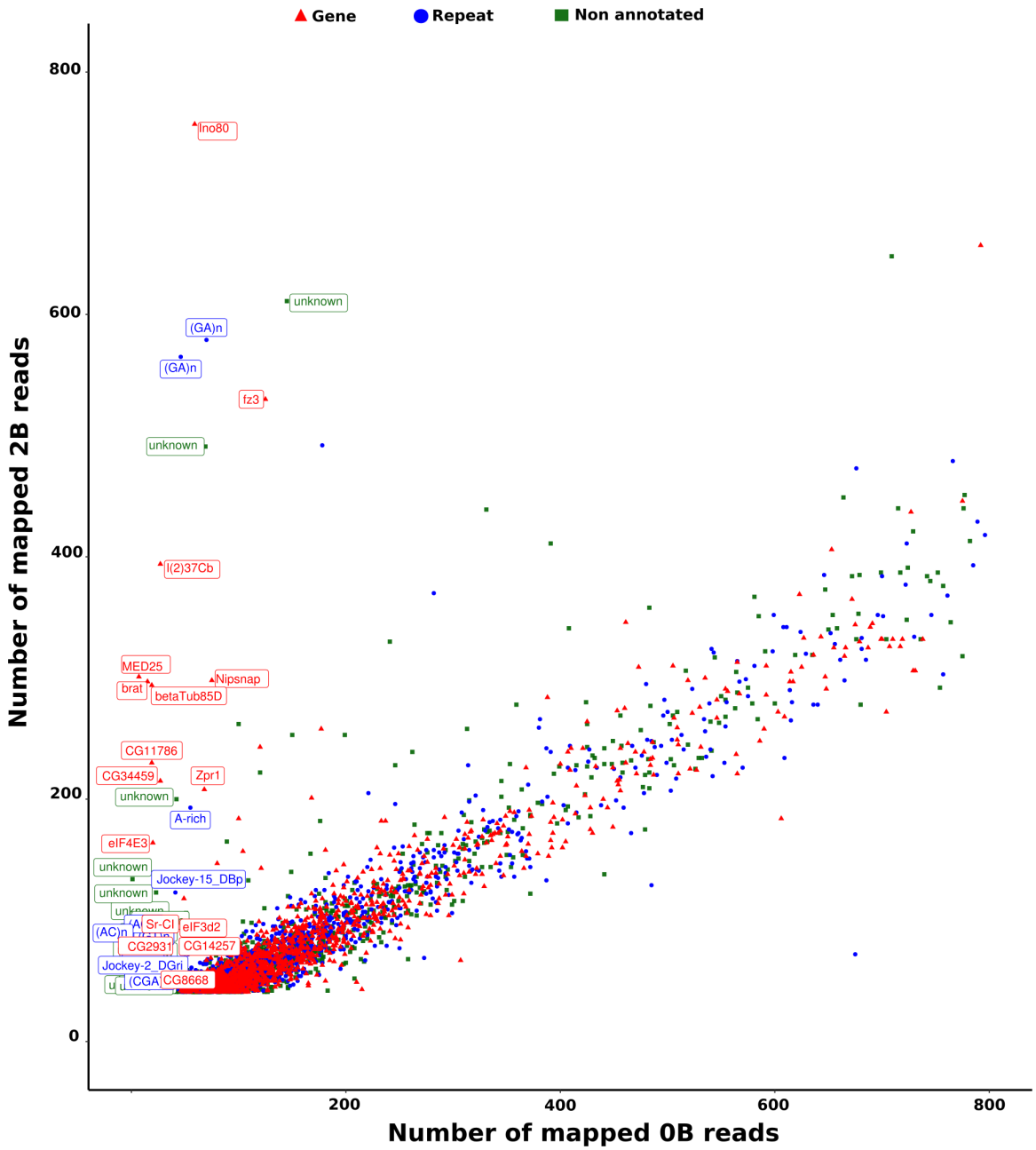


Figure 9. Identification of protein-coding genes located in B chromosomes of the *A. flavolineata* using the number of mapped reads that map to the CDSs found in the transcriptome, in the 0B (X axis) and 2B (Y axis). Each dot represents a coding sequence with only those labeled that recorded

the log_2 greater than 1.5. The plot is limited for 800 mapped reads to optimize the visualizations.

Remarkably, despite the lesser coverage of raw data of 2B (10x) as compare to 0B (20x), as indicated by the inclination of a high percent of dots towards X axis (0B mapped reads) in this plot, some dots can be seen inclined towards Y axis (2B reads) that show the B-localized sequences.

The lists of high confident genes located on the Bs using Log2 ratio analysis are given as table 3, table 4 and table 5.

Table 3. Results of mapping gDNA reads from 0B and 1B males of *A. correntinus* on transcriptome CDS. The higher Log_2 ratio indicates the duplicated copies of CDSs on B chromosome.

Transcript ID	**0B reads**	**1B reads**	**1B/ 0B**	**Log_2 ratio**	**Annotation**
ARA0AAA105YI15EM1.b.am.1	277	1496	5.40072	2.43315	tars
ARA0AAA15YI22EM1.b.am.1	44	569	12.9318	3.69285	novel_gene
ARA0AAA117YG08EM1.b.am.1	51	145	2.84314	1.50749	scn12aa
ARA0AAA12YF15EM1.b.am.1	251	968	3.85657	1.94732	exosc7
ARA0AAA15YI22EM1.b.am.1	44	569	12.9318	3.69285	novel_gene
ARA0AAA6YO21EM1.b.am.1	255	816	3.2	1.67807	novel_gene
ARA0AAA75YN07EM1.b.am.1	40	142	3.55	1.82782	zgc
ARA0ABA106YI03EM1.b.am.1	48	165	3.4375	1.78136	novel_gene
ARA0ABA11YM14EM1.b.am.1	61	224	3.67213	1.87662	zgc
ARA0ABA21YO23EM1.b.am.1	90	271	3.01111	1.5903	novel_gene
ARA0ABA22YA20EM1.b.am.1	57	200	3.50877	1.81097	shisal1b
ARA0ABA3YF16EM1.b.am.1	49	178	3.63265	1.86102	ndufa11
ARA0ABA43YF05EM1.b.am.1	135	431	3.19259	1.67473	pcdh1g31
ARA0ABA65YL16EM1.b.am.1	64	189	2.95312	1.56224	novel_gene
ARA0ABA73YP22EM1.b.am.1	76	224	2.94737	1.55943	novel_gene
ARA0ABA93YE09EM1.b.am.1	48	168	3.5	1.80735	novel_gene
ARA0AFA5YA16EM1.b.am.1	119	357	3	1.58496	gon4l

Table 4. Results of mapping gDNA reads from 0B and 2B males of *A. mexicanus* on transcriptome

CDSs.

Transcript ID	0B reads	1B reads	1B/ 0B	Log_2 ratio	Annotation
ARA0AAA39YG16EM1.b.am.1	727	14308	19.681	4.29873	novel_gene
ARA0AAA40YD05EM1.b.am.1	156	517	3.314	1.72857	jac3
ARA0ABA10YE08EM1.b.am.1	586	1661	2.834	1.50284	arih1l
ARA0ABA90YO24EM1.b.am.1	117	359	3.068	1.6173	novel_gene
ARA0AFA5YA16EM1.b.am.1	96	301	3.135	1.64847	gon4l
ARA0AHA14YK02EM1.b.am.1	50	151	3.02	1.59455	unknown
ARA0AGA13YG24EM1.b.am.1	89	472	5.303	2.40681	unknown
ARA0ABA7YH17EM1.b.am.1	44	140	3.182	1.66993	unknown
ARA0ABA25YA20EM1.b.am.1	44	126	2.864	1.51803	unknown
ARA0ABA107YI10EM1.b.am.1	42	119	2.833	1.50233	unknown
ARA0ABA103YB02EM1.b.am.1	101	340	3.366	1.75104	unknown
ARA0AAA70YD16EM1.b.am.1	71	318	4.479	2.16318	unknown

Table 5. Results of mapping gDNA reads from 0B and 2B males of *A. flavolineata* on transcriptome CDSs.

Transcript ID	0B reads	1B reads	1B/ 2B	Log_2 ratio	Annotation
GDIO01012001.1	68	208	3.059	1.61306	Zpr1
GDIO01014672.1	75	298	3.973	1.99023	Nipsnap
GDIO01044471.1	59	757	12.831	3.68156	Ino80
GDIO01045346.1	125	530	4.24	2.08406	fz3
GDIO01030872.1	69	491	7.116	2.83107	unknown
GDIO01018848.1	42	200	4.762	2.25157	unknown
GDIO01009316.1	145	611	4.214	2.07519	unknown

5.4 Functions of B chromosomes

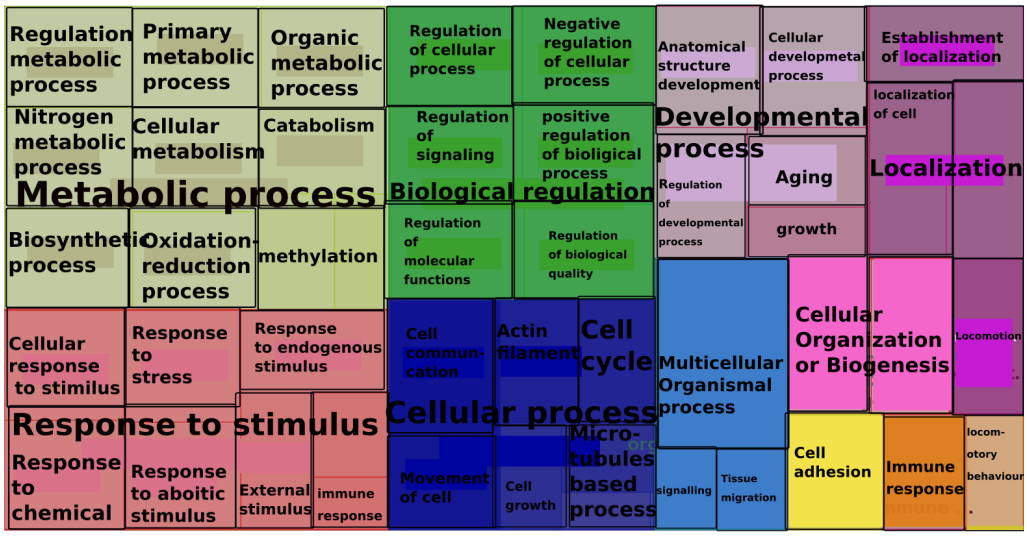
The functional annotation of genes detected on the B chromosomes were determined for high integral genes (>50% integrity) as well as the CDSs putative located on B chromosomes for *Astyanax* species. The statistical graphs of GOs annotations for biological processes functions

category of B-blocks are shown as supplementary figures 5 and 6. We considered the complete list of genes detected in the blocks for *A. flavolineata*. Gene Ontology (GO) enrichment analysis for these genes revealed potential implication in cellular processes likely advantageous for the B chromosome, such as the microtubules process, regulation of transcription, cell division, actin filaments, apoptotic processes and cell death. Apart from these functions, there were also detected enrichment of diverse GO terms for important biological roles such as developmental process, metabolism, cell adhesions, reproduction, immune response, localization, morphogenesis and response to stimulus. The enrichment plots of GO are shown as treemaps graphs (figure 11). The B chromosomes of *Astyanax* species indicated a similar pattern of functions. While the Bs of *A. flavolineata* were more enriched with processes associated with chromosomes and cell division functions (supplementary figure 6). We also performed a comparison of B chromosomes functions in our all three analyzed model species to check if they share common functions. Although there we did not find any same gene shared among the Bs of the three species, however we were able to detect common GO that were enriched in three Bs. For this analysis, we extracted three lists of IDs that contained a total of 1,470, 1,148 and 350 GOs enriched in the B blocks of *A. correntinus, A. mexicanus* and *A. flavolineata*. Interestingly, the Venn diagram (figure 11) of this comparison recorded a total of 267 GOs shared among all species, and 814 GOs common between Astyanax species. These findings suggest that in general the B chromosomes of different species may have different genetic composition, but their behavior to acquire the genes of particular functions remain the same.

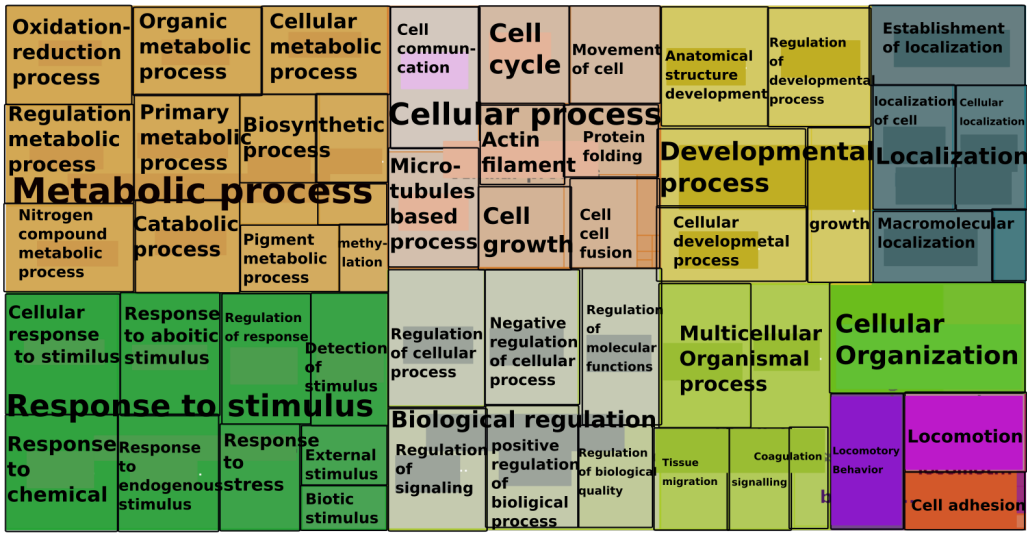
We retrieved a list of genes detected on the Bs of our three models, these genes are directly involved in the chromosome formation and cell cycle related functions (table 6), that are regarded as advantageous for their successful evolution. These kinds of genes were found in each of our model species indicating their importance in the establishment of Bs. A total of at least 35 genes in *A. correntinus* and 7 genes in *A. mexicanus* with higher integrity (>50%), while a minimum of 25

genes in *A. flavolineata* were detected to play a direct role in these functions. These findings corroborate our primary hypothesis that motivated the beginning of the present PhD study. The hypothesis stated that “B chromosomes tend to accumulate the cell cycle genes that might play an important role in their transmission”.

a


Regulation metabolic process
Primary metabolic process
Organic metabolic process
Regulation of cellular process
Negative regulation of cellular process
Anatomical structure development
Cellular developmetal process
Establishment of localization
Nitrogen metabolic process
Cellular metabolism
Catabolism
Regulation of signaling
positive regulation of biological process
Developmental process
localization of cell
Metabolic process
Biological regulation
Regulation of developmental process
Aging
Localization
Biosynthetic process
Oxidation-reduction process
methylation
Regulation of molecular functions
Regulation of biological quality
growth
Cellular response to stimulus
Response to stress
Response to endogenous stimulus
Cell commun-cation
Actin filament
Cell cycle
Multicellular Organismal process
Cellular Organization or Biogenesis
Locomotion
Response to stimulus
Cellular process
Response to chemical
Response to aboitic stimulus
External stimulus
immune response
Movement of cell
Cell growth
Micro-tubules based process
signalling
Tissue migration
Cell adhesion
Immune response
locom-otory behaviour

b


Oxidation-reduction process
Organic metabolic process
Cellular metabolic process
Cell commun-cation
Cell cycle
Movement of cell
Anatomical structure development
Regulation of developmental process
Establishment of localization
Regulation metabolic process
Primary metabolic process
Biosynthetic process
Cellular process
localization of cell
Cellular localization
Actin filament
Protein folding
Developmental process
Localization
Micro-tubules based process
Metabolic process
Nitrogen compound metabolic process
Catabolic process
Pigment metabolic process
methy-lation
Cell growth
Cell cell fusion
Cellular developmetal process
growth
Macromolecular localization
Cellular response to stimulus
Response to aboitic stimulus
Regulation of response
Detection of stimulus
Regulation of cellular process
Negative regulation of cellular process
Regulation of molecular functions
Multicellular Organismal process
Cellular Organization
Response to stimulus
Biological regulation
Response to chemical
Response to endogenous stimulus
Response to stress
External stimulus
Biotic stimulus
Regulation of signaling
positive regulation of bioligical process
Regulation of biological quality
Tissue migration
Coagulation
signalling
Locomotory Behavior
Locomotion
Cell adhesion

c

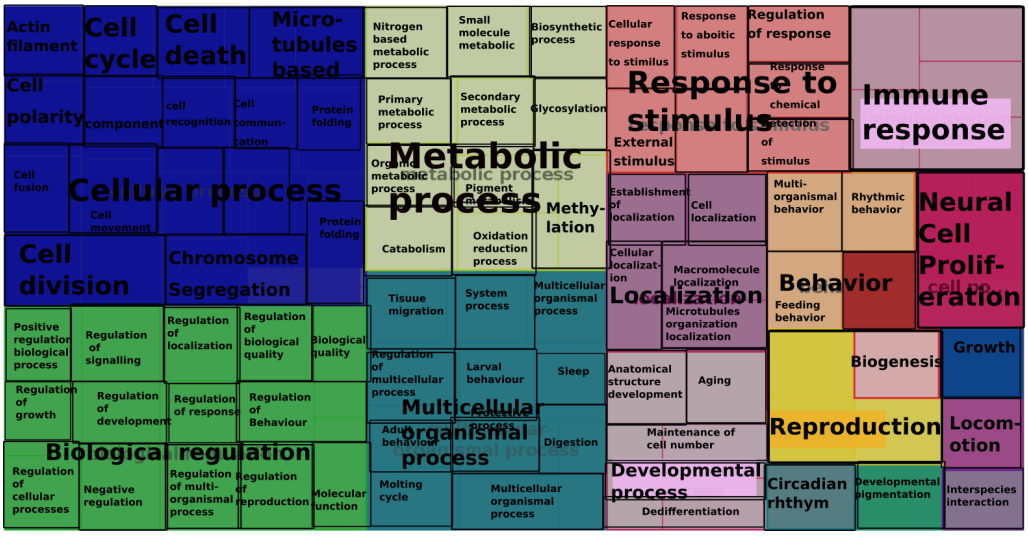

Actin filament
Cell cycle
Cell death
Micro-tubules based
Nitrogen based metabolic process
Small molecule metabolic
Biosynthetic process
Cellular response to stimulus
Response to aboitic stimulus
Regulation of response
Response to
Cell polarity
component
cell recognition
Cell commun-cation
Protein folding
Primary metabolic process
Secondary metabolic process
Glycosylation
Response to stimulus
Immune response
External stimulus
chemical
Cell fusion
Cellular process
Metabolic process
Cell movement
Protein folding
Catabolism
Oxidation reduction process
Methy-lation
Establishment of localization
Cell localization
Multi-organismal behavior
Rhythmic behavior
Neural Cell Prolif-eration
Cell division
Chromosome Segregation
Tisuue migration
System process
Multicellular organismal process
Cellular localizat-ion
Macromolecule localization
Localization
Behavior
Feeding behavior
Positive regulation biological process
Regulation of signalling
Regulation of localization
Regulation of biological quality
Biological quality
Microtubules organization localization
Biogenesis
Growth
Regulation of growth
Regulation of development
Regulation of response
Regulation of Behaviour
Regulation of multicellular process
Larval behaviour
Sleep
Anatomical structure development
Aging
Multicellular organismal process
Reproduction
Locom-otion
Biological regulation
Maintenance of cell number
Regulation of cellular processes
Negative regulation
Regulation of multi-organismal process
Regulation of reproduction
Molecular function
Molting cycle
Multicellular organismal process
Developmental process
Circadian rhthym
Developmental pigmentation
Interspecies interaction
Dedifferentiation

Figure 10. The GO enrichment analysis of B chromosomes. The treemaps graphics visualization of enriched terms detected in the B blocks of (a). *A. mexicanus***,** (b). *A. correntinus* and (c). *A. flavolineata.* The different size of box represents the abundance of enrichment with different colors showing respective functions. Notice the enrichment of functions associated with cellular processes (including cell cycle, cell death, microtubules-based process, actin filament process, cell division, chromosome segregation) in all species, that we believe might play an important role in the evolution of B chromosomes, thus corroborating our principal hypothesis. Except these functions, the enrichment of diverse list of important biological functions were also detected that emphasize the role of B chromosomes inside the cell.

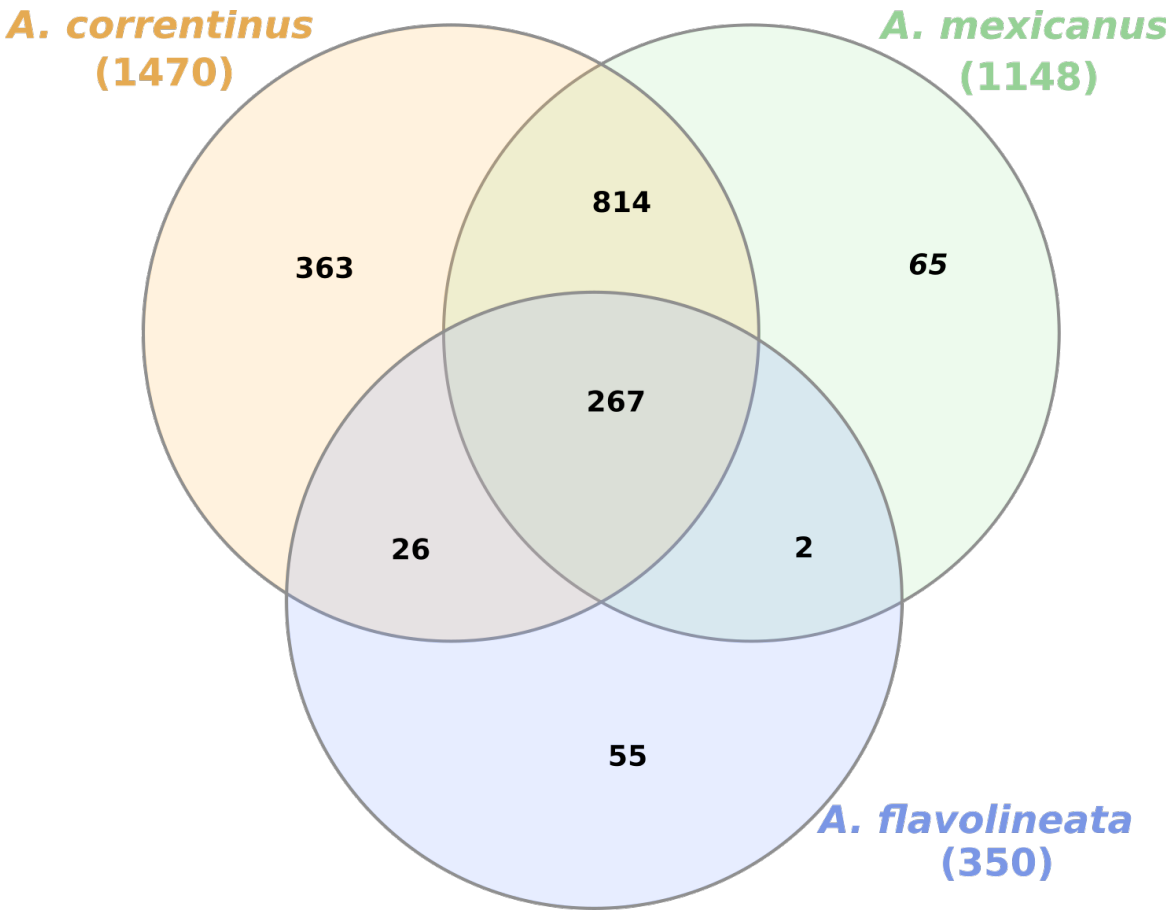


Figure 11. Venn diagram shows the comparison of GOs enriched among the B chromosomes of three model species *A. correntinus*, *A. mexicanus* and *A. flavolineata.* The co-occurrence of 267 enriched ontologies may be indicative to the B chromosome behavior for the acquirement of similar functions in diverse species.

Table 6. A list of genes detected on the B chromosomes of three model species *A. correntinus*, *A. mexicanus* and *A. flavolineata*, involved in chromosomes organization and cell cycle processes. These genes may involve creating favorable circumstances that increase the chances of B chromosome success.

Species	Ensembl ID	Protein name	Gene name	Gene description	Function related to cell cycle/chromosome
A. correntinus	ENSAMXT00000047800.1	ATP-dependent DNA helicase Q1	RECQL	DNA helicase activity	Chromosome organization
	ENSAMXT00000008370.2	Structural maintenance of chromosomes 2	smc2-201	Structural maintenance of chromosomes 2	Chromosome organization
	ENSAMXT00000050221.1	Histone-lysine N-methyltransferase	ASHH2	Histone lysine methylation	The modification of a histone by addition of one or more methyl groups to a lysine residue.
	ENSAMXT00000055649.1	E3 ubiquitin-protein ligase	RNF8	Ubiquitination of histones H2A and H2AX	Chromatin decondensation
	ENSAMXT00000043499.1	dkey-16p21.7	dkey-16p21.7	Protein kinase activity	Chromatin remodeling
	ENSAMXT00000048305.1	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily E member 1	SMARCE1	Alteration of DNA-nucleosome topology	Chromosome remodeling
	ENSAMXT00000039111.1	ch211-255f4.7	Novel gene	unkonwn	Chromatin

				remodeling
ENSAMXT00000039111.1	CCCTC-binding factor (Zinc finger protein)	CTCF	Chromatin biding	Regulation of genne expression
ENSAMXT00000001525.2	transducin beta like 1 X-linked	TBL1X	Signal transduction	Microtubules based process
ENSAMXT00000034475.1	Zgc:85722	Zgc:85722	Unkonown	Establishment of mitotic spindle orientation, microtubule-based process, cell cycle
ENSAMXT00000043819.1	Microtubule actin crosslinking factor 1	MACF1	Form bridges between different cytoskeletal elements	Actin-microtubule interactions at the cell periphery and couples the microtubule network to cellular junctions
ENSAMXT00000030346.1	Tubulin alpha chain	TUBA	It binds two moles of GTP, one at an exchangeable site on the beta chain and one at a non-exchangeable site on the alpha chain.	Major constituent of microtubules
ENSAMXT00000043650.1	Kinesin-like protein	KIN-14A	ATP-dependent microtubule motor activity	Microtubule-based movement
ENSAMXT00000031496.1	Dynein heavy chain 5, axonemal	Dnah5	ATPase activity	Produces force towards the minus ends of microtubules
ENSAMXT00000052127.1	myosin heavy chain 10	MYH10	Regulation of cytokinesis, cell motility, and cell polarity.	Microtubule-based movement, microtubule binding
ENSAMXT00000030102.1	microtubule-associated proteins	MAPs	Microtubules cytoskeleton organization	Stability and regulate microtubules,
ENSAMXT00000026549.2	MAP7 domain-	MAP7D1	Microtubule	Stability and

		containing protein 1		cytoskeleton organization	regulate microtubules,
	ENSAMXT00000026649.2	Rabenosyn, RAB effector	RBSN	Regulate membrane trafficking	Microtubule cytoskeleton organization
	ENSAMXT00000057964.1	Histone H4	HIST1H4J	Core component of nucleosome. Nucleosomes wrap and compact DNA into chromatin	Chromosome stability
A. mexicanus	ENSAMXT00000054831.1	Eukaryotic translation initiation factor 4 gamma 1	EIF4G1	Recognition of the mRNA cap	Mitochondrion organization
	ENSAMXT00000043499.1	dkey-16p21.7	dkey-16p21.7	Protein kinase activity	Chromatin remodeling
A. flaveniota	FBtr0070762	DNA replication licensing factor MCM3	Mcm3	Putative replicative helicase	Cell cycle' and DNA replication initiation and elongation in eukaryotic cells
	FBtr0076253	General transcription and DNA repair factor IIH helicase subunit XPB	hay	ATP-dependent 3'-5' DNA helicase	Transcription-coupled nucleotide excision repair (NER) of damaged DNA
	FBtr0079901	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A containing DEAD/H box 1 homolog	Etl1	ATP-dependent nucleosome-remodeling activity	DNA repair and heterochromatin organization
	FBtr0083193	Meiotic recombination protein	rec	Separation of sister chromatids and homologous chromosomes	Chromosomes separation during meisis
	FBtr0085875	205 kDa	Map205	Phosphorylation	Microtubule

	microtubule-associated protein			assembly and interaction
FBtr0087461	Structural maintenance of chromosomes protein 3	SMC2	chromosome cohesion during the cell cycle	Cell division
FBtr0087904	Anastral spindle 1, isoform A	ana1	Nucleation of microtubules	Spindle fiber formation
FBtr0339505	Meiotic central spindle, isoform B	Meics	Nucleation of microtubules	Spindle Assembly and Chromosome Segregation
FBtr0304801	Mini spindles, isoform D	msps	Nucleation of microtubules	Integrity of the Mitotic Spindle
FBtr0299509	Doublecortin-domain-containing echinoderm-microtubule-associated protein, isoform G	DCX-EMAP	Assembly dynamics of microtubules	Microtubules cytoskeleton segregation
GDIO01044471.1	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1	SMARCAL1	ATP-dependent annealing helicase	Catalyzes the rewinding of the stably unwound DNA
GDIO01001775.1	Histone-lysine N-methyltransferase eggless	egg	Histone methyltransferase	Histones methylation
GDIO01013035.1	Dosage compensation regulator	mle	Compensation of X chromosome linked genes	Sex differenciation
GDIO01001774.1	Histone-lysine N-methyltransferase eggless	egg	Histone methyltransferase	Histone H3-K9 methylation, heterochromatin organization, positive regulation of methylation-

					dependent chromatin silencing
	GDIO01030872.1	GH07148p	Dmel\CG1582	unknown	Chromatin silencing, mitotic chromosome condensation
	GDIO01009316.1	Tubulin beta-1 chain	mec-7	Microtubules formation	Attachment of spindle microtubules to kinetochore involved in meiotic chromosome segregation
	GDIO01018464.1	Gamma-tubulin complex component 3	GCP3	Microtubule nucleation	Spindle assembly, regulation of cell cycle, meiotic nuclear division

5.5 Comparative genomics analysis reveals the pattern of segmental duplication and inversions in B chromosomes.

The comparative analysis of B chromosomes to their ancestral sequences coming from A chromosomes detected several duplication and inversions events that might have happened during their evolutionary period. These events were traced through syntenic dotplot analysis (figure 12, figure 13 and figure 14). Synteny, in genomic terms, is defined as two or more genomic regions that are derived from a common ancestor. For clear visibility of chromosomal rearrangements patterns, we selected only the B blocks greater than 2kb size and were mapped to corresponding regions that come from A chromosomes, as a result the alignments blocks were displayed as green dots in the dotplot. Looking closely at the syntenic dotplot, there is an overlap of the lines when the lines are projected to one axis or the other. This is because a given region on B chromosome is syntenic to the region in the A genome. In order to validate and access the types and patterns of change at these genomic loci, high-resolution analysis of these syntenic regions were also performed for some regions (see an example, focused view in figure 12). The patterns of segmental duplications and inversions as

visualized in these dotplots point toward the chromosomal rearrangements thus providing insights on their evolutionary history. Remarkably, these evidences also corroborated our previous hypothesis and model of B chromosome origin, reported in Ahmad and Martins, (2019).

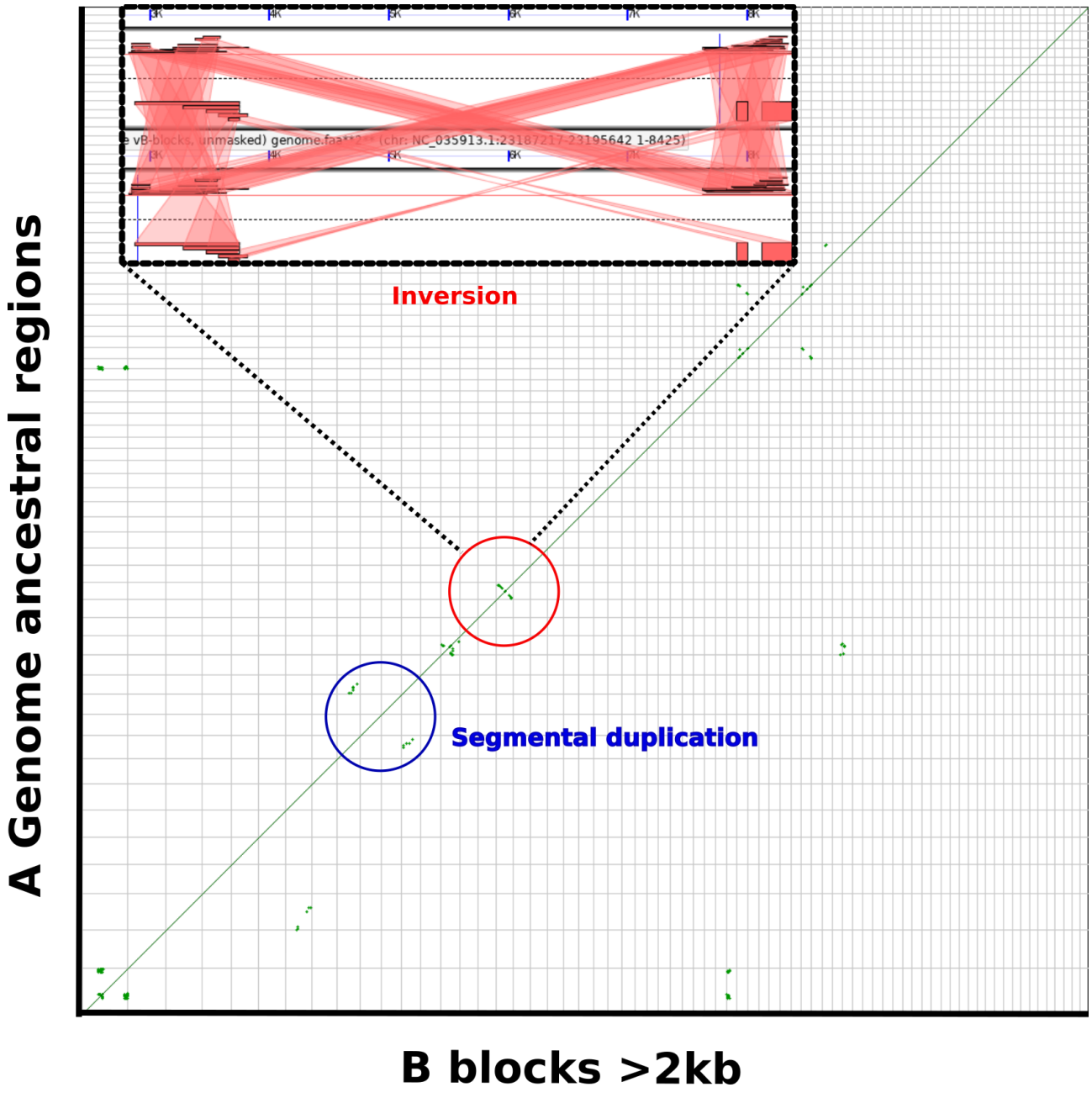


Figure 12. Dotplot shows pattern of genomic rearrangements in the B blocks of *A. mexicanus.* X axis of the dotplot display the B blocks and Y axis represents the aligned A genomic regions laid end to end. Each small square in the graph is an alignment block. The lower-left corner represents the start of

each genome (A and B chromosomal regions). Each putative homologous genomic-pair is drawn as a green dot on the dotplot with its x and y position corresponding to the aligned syntenic genomic position of A and B chromosomes separated by a diagonal line. Noticeably, the syntenic green dots in some (representative circled) areas can be seen as groups forming a big synteny, that indicate a larger portion of B chromosome has been derived from the corresponding A genome as a result of chromosomal rearrangements. The big syntenies on the dotplot are plotted as a result of larger alignments of A and B blocks sequences. The horizontally oriented syntenies faced one to one across the diagonal line reflect a segmental duplication while transverse syntenies can be interpreted as inversion events. Examples of these events are labeled as red and blue circles showing an inversion and duplication respectively. A high-resolution of inversion region is also shown in the box. Two panels, one for each A and B genomic region are shown. The dashed line in the middle of each panel separates the top and bottom strands of DNA. Transparent red wedges are drawn connecting it to its partner B chromosome region in the A genome

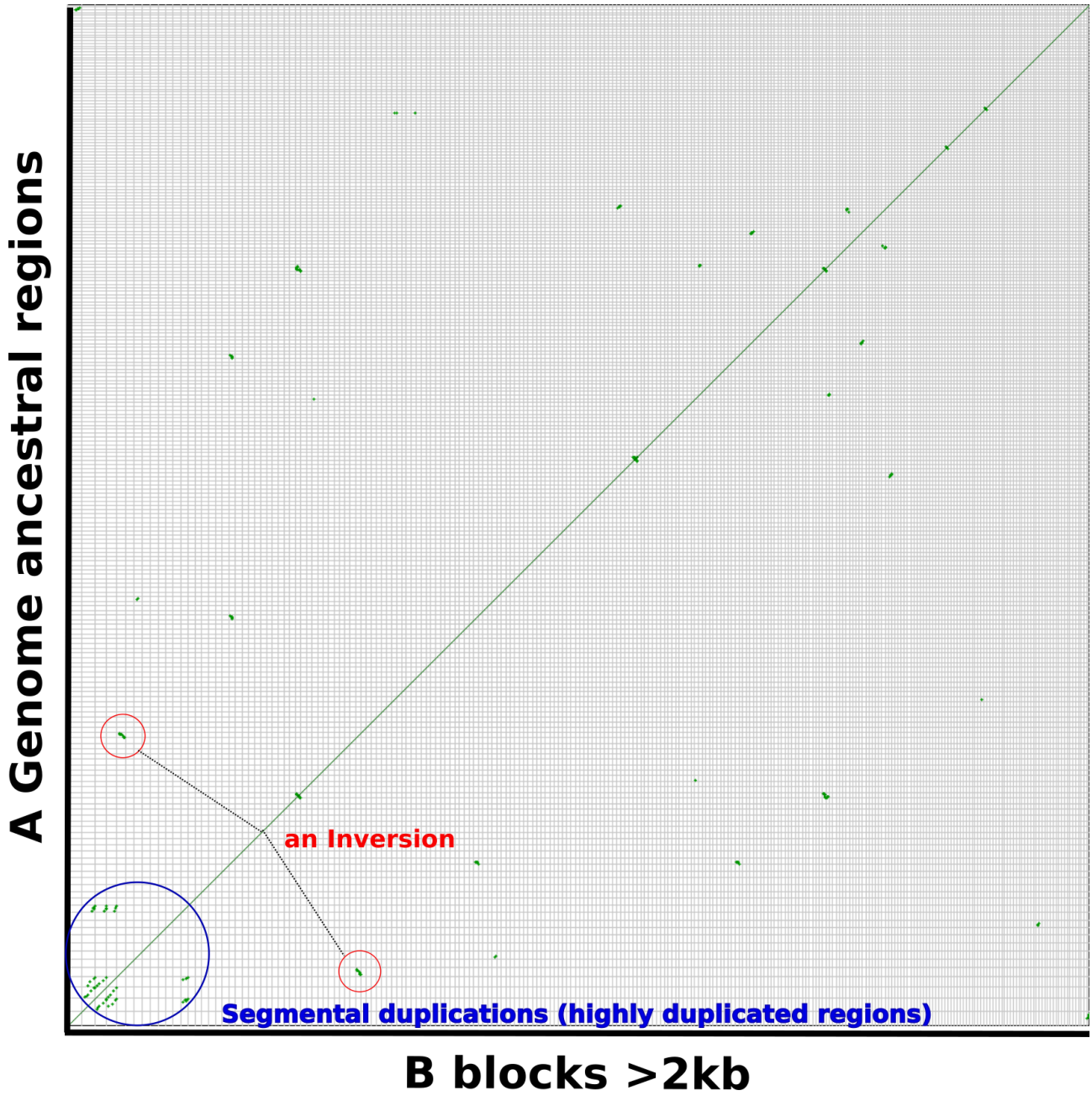


Figure 13. Dotplot shows pattern of genomic rearrangements in the B blocks of *A. correntinus*. High level of duplicated regions and inversions can be seen as evidence that B chromosome accumulated sequences from the A genome due to frequent duplications. Refer to the the Previous figure 12 captions for details.

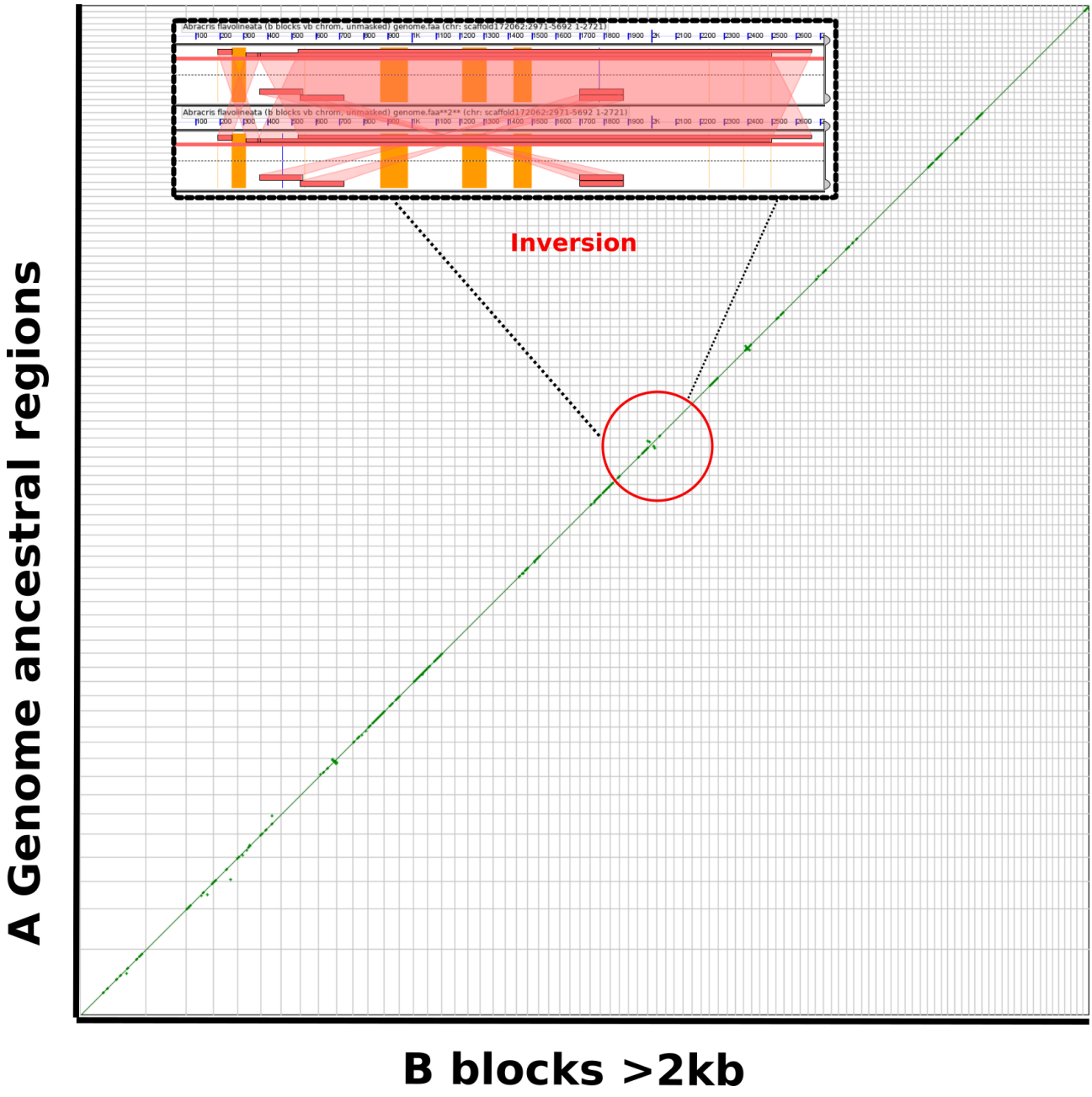


Figure 14. Dotplot shows patterns of genomic rearrangements in the B blocks of *A. flavolineata*. The patterns of duplications and inversions are shown. Refer to the figure 12 captions for more details.

We also compared the B-blocks of *A. correntinus* to detect their origin on A chromosomes of the reference genome. This analysis demonstrated that the B chromosome contains DNA sequences derived from almost all chromosomes. The largest number of blocks map to chromosome 4 (10.99% of B), 5 (5.02% of B), 7 (4.92%), 6 (4.32%), 3 (4.29%) and 2 (4.26% of the B) (figure 15).
This shows the B chromosome of *A. correntinus* has multiple chromosome origin.

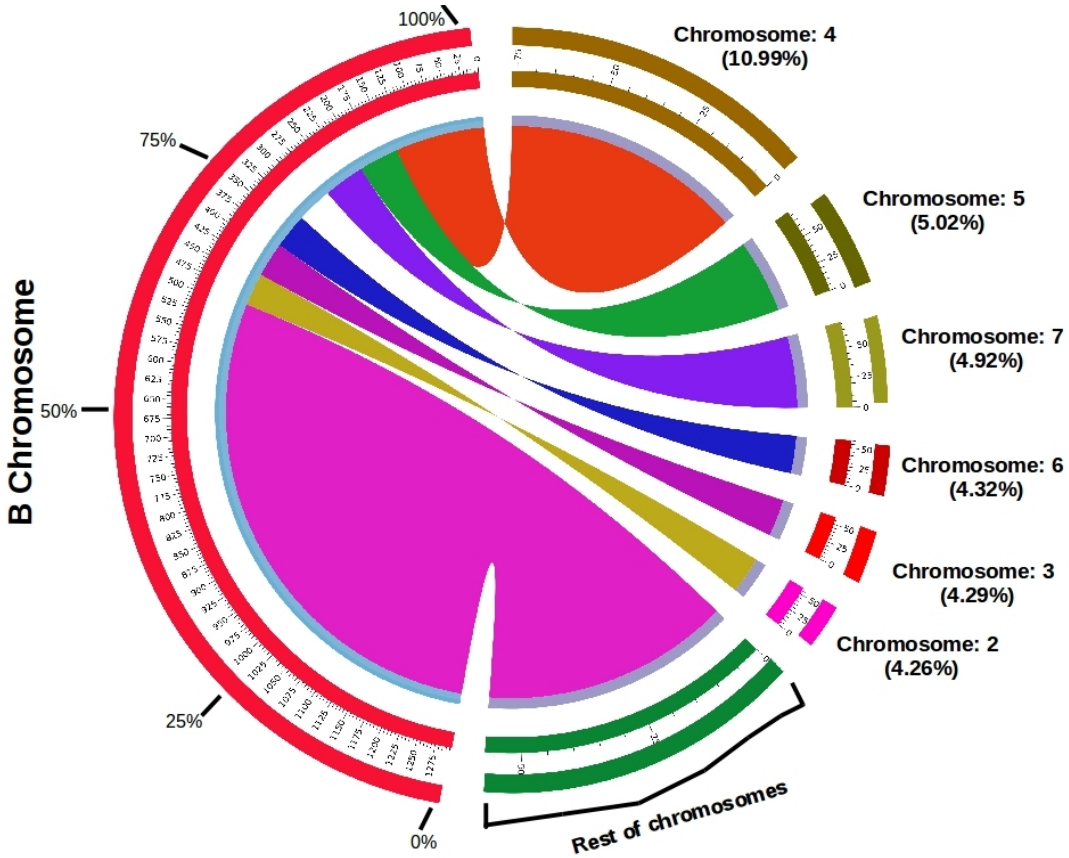


Figure 15. Relationship of the B blocks of *A. correntinus* to the A chromosomes of reference. B blocks are shown on the left and corresponding chromosomes on the right with percentage of content similarity.

5.6 Analysis of microdissected Bs of additional species

A total of 87,6260, 765,254, 69,279,084, 938,990, 2,630,300 and 1,590,728 number of raw reads previously generated by Illumina next generation sequencing, were retrieved from NCBI/SRA database (reference) B1 (accession ID: SRX2041358) and B2 (accession ID: SRX2041352) of *Lates calcarifer*, B3 (accession ID: SRX1484569) of *Eyprepocnemis. plorans*, B4 (accession ID: SRX3412298) and B5 (accession ID: SRX3412297) of *Apodemus peninsulae*, and B6 (accession ID: SRX3412293-SRX3412299) of *A. flavicollis* respectively. A total number of 78,611, 65,780, 67,469,719, 660,346, 113,844 and 783,058 decontaminated trimmed reads for the aforementioned

respective Bs were aligned with reference genomes. Statistics of microdissected Bs analysis is given in the table 7. The pseudo scaffolding based strategy for assembling these chromosomes with spacer length 10kb was considered for annotation and Gene ontologies analysis.

Table 7. Summary of analyzed data used for microdissected Bs assemblies.

Microdissected B/species/ Acession ID:	**Raw reads/Decontaminated trimmed reads**	**Assembly**	**Spacer length**	**Fragments**	**Total length (bp)**	**% of reference genome**
B1/ Asian Seabass fish/ Acession ID: Species: *L. calcarifer*	876,260/78,611	ChB1 psuedo-scaffolds	10 bp	1,764	59,065	0.010
			100 bp	1,743	603,58	0.010
			2 kb	1,662	963,69	0.016
			10 kb*	1,412	434,598	0.074
			100 kb	1213	21,578,827	3.67
B2/ Asian Seabass fish/ Acession ID: Species: *L. calcarifer*	765,254/65,780	ChB2 psuedo-scaffolds	10bp	1,552	71,373	0.012
			100 bp	1530	72,775	0.012
			2 kb	1412	164,958	0.028
			10 kb*	1,601	378,367	0.064
			100 kb	1,097	16,912,139	2.9
B3/ Grasshopper/ Acession ID: Species: *E. plorans*	69,279,084/67,469,719	ChB3 psuedo-scaffolds	10 bp	5,006,153	5,188,127	0.08
			100 bp	4,586,192	7,471,409	0.11
			2 kb	1,460,148	23,634,021	2.3
			10 kb*	969,779	25,972,513	2.5
			100 kb	725,481	383,082,448	38.3
B4/ Mouse/Acession ID: Species: *Apodemus peninsulae*	938,990/660,346	ChB4 psuedo-scaffolds	10 bp	94	14,095	-
			100 bp	93	14,127	-
			2 kb	84	19,715	-
			10 kb*	82	29,798	-
			100 kb	64	800,520	-
B5/	2630300/113844	ChB5	10 bp	40	6,183	-

Mouse/Acession ID: Species: *Apodemus peninsulae*		psuedo-scaffolds	100 bp	40	6,183	-
			2 kb	40	6,183	-
			10 kb*	39	10,445	-
			100 kb	36	132,972	-
B6/ Mouse/Acession ID: Species: *A. flavicollis*	1590728/783058	ChB6 psuedo-scaffolds	10 bp	901	70,503	-
			100 bp	889	71,210	-
			2 kb	854	92,874	-
			10 kb*	773	457,733	-
			100 kb	375	18,833,514	-

10 kb* considered for annotation and gene ontologies

Although the NGS data of Bs does not cover the complete sequences, we were able to present an overview of their genes and repeats contents. The repeat annotations of these Bs showed that they have different level of each repeat content in different species (figure 16). The Bs of fish species (B1 and B2) mainly comprised of simple repeats. Other repeats types such as low complexity and DNA transposons were also detected in abundance for Bs of fish but lacking the retroelements, SINEs and satellites. Similarly, the Bs of grasshopper (B3 of *E. plorans*) was also enriched with simple repeats but notably, the second highest number of repeats sequences were retroelements (SINEs and LINEs) which were not abundant in B1 and B2 of fish. On other hand the Bs of both mouse species (B4, B5 and B6) contained an abundance of SINEs and LINEs but lack the amount of satellite sequences which are found in higher number in grasshopper species. In spite of the fact that the analyzed micro dissected Bs data is not sufficient to draw a conclusion about their repeat contents, however the comparative analysis enabled us to hypothesize that the B chromosomes repeat contents can vary among different species also depending upon the A genomic repeat types enrichment.

The gene annotation recorded of all microdissected Bs detected several genes fragments. Due to low

coverage of data, the number of genes is substantially underrepresented. contents. Interestingly, the GO and enrichment analysis of these genes fragments also detected a diverse set of terms such as regulation of cellular process, metabolism, development, morphogenesis, reproduction, immune response, cell cycle, microtubules processes, cell adhesion, nucleotide binding, single organism process, actin filament binding and regulation (supplementary figure 8).

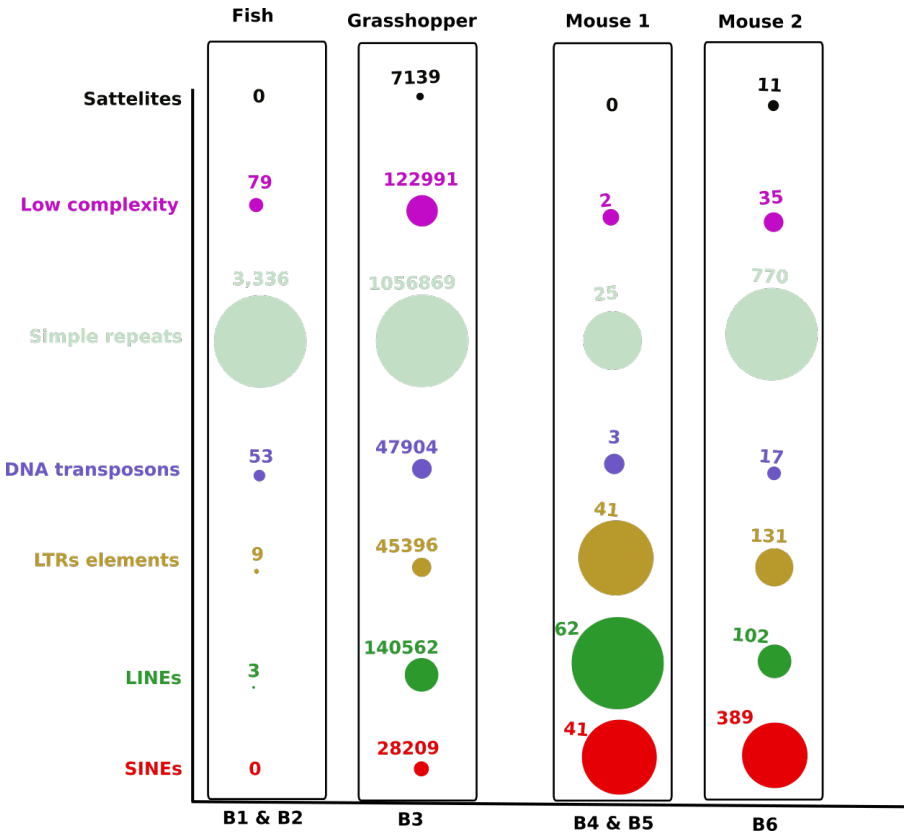


Figure 16. Repeat contents comparison of our analyzed micro dissected B chromosomes from diverse species. The bubble charts have been merged for all Bs showing the type of content in different colors. Each bubble is a repeat type while each bar indicates a species. The differences between repeats abundance among species suggest that amount of these elements in Bs is subject to their abundance in A genome dependent of species. For example, the Bs of mouse species (B4, B5 and B5) have acquired a higher amount of SINEs as depicted by red bubbles and lack abundance of satellite

DNA. While on other hand, grasshopper Bs (B3) gained a considerable amount of satellite DNA apart from the domination of simple repeats and other elements.

6. Discussion

The current work demonstrates a high throughput genomic analysis of B chromosomes in three candidate model species. We present a comprehensive analysis of *A. correntinus*, *A. mexicanus* and *A. flavolineata* genomes both B+ and B- individuals with the aim of unveiling the genetic structure, composition, function and evolution of B chromosomes in these species. Applying a comparative coverage technique, we detected a total of 43.82 Mb and 15.41 Mb of the A chromosomes of *A. correntinus* and *A. mexicanus* respectively that has contributed to the B chromosomes composition. These findings are consistent with the size of their respective B chromosomes observed in karyotype data. This shows that coverage ratio technique was successful in identifying an appreciable amount of sequences on the B chromosome. Our downstream characterization and annotation of B blocks in *Astyanax* species featured a higher amount of gene contents and number of blocks for *A. correntinus* as compared to *A. mexicanus* which is also in agreement to karyotype data. As evident in the karyotype data, the macro B of *A. correntinus* and micro B of *A. mexicanus* provide a validation to our NGS analyses of detection sequences on the Bs. The number of detected blocks of *A. flavolineata* were underrepresented due to low quality coverage in comparison to its giant genome. Nevertheless, we could detect at least a total of 2.05 Mb of A genomes copied into the B chromosome.

An important point about the limitation of our coverage-based identification of B sequences is that this approach cannot detect any sequence entirely unique to the B chromosome (not aligned to homologous A sequence). While unique sequences are mechanistically entirely undetectable with a coverage ratio analysis, we suggest this issue can be resolved by recovering the entire B chromosome using highly accurate chromosomes scale *de novo* assemblies.

To perform a deep survey of DNA repeats, we applied a combination of approaches to predict major

TEs and their abundance in the genomes and to perform general comparison of B+ and B- repetitive composition. Our analysis related to repeats composition showed that *A. correntinus* and *A. mexicanus* genomes are constituted by 66% and 35% repeats respectively with domination of DNA transposons, which is comparable to most published fish genomes (Venkatesh et al. 2007; Star et al. 2011; Jones et al. 2012;Howe et al. 2013; Smith et al. 2013; Mcgaugh et al. 2014). We identified the main repeats in the B chromosomes, and our annotations show TEs and retroelements on the B chromosomes of *A. correntinus, A. mexicanus, A. flavolineata, L. calcarifer, E. plorans, A. flavicollis* and *A. peninsulae* species. In general, B chromosomes are rich in several classes of repetitive DNA, including 5S and 45S ribosomal DNA (rDNA), satellite DNA, histone genes, small nuclear DNA, mobile elements, and organellar sequences (Camacho, 2005; Friebe et al. 1995; Cabrero et al. 2003; Bugrov et al. 2007; Teruel et al. 2010; Oliveira et al. 2011; Bueno et al. 2013; Kour et al. 2014; Houben et al. 2013; Coan and Martins 2018). DNA transposons and Retrotransposons are one of the most common migrants into the clonally inherited B chromosome and could be responsible for the insertion of sequences into the B. Gene sequences could have been inserted as fragments, or they could have been broken after insertion on the B by subsequent TE insertions (Valente et al 2014).

The genes annotations of B blocks in our model species concluded that the Bs include sequences homologous to known genes. Our integrity analysis showed that Bs contain many fragmented genes which possibly are pseudogenes and might have formed from their parental genes on A chromosomes during their incorporation into B chromosomes. These putative pseudogenes might have lost their functional ability after duplication from the parental A genes. However, previously Banaei-Moghaddam et al. (2013) reported that the B of rye harbor the pseudogene-like fragments which are expressed in tissue specific manner. Apart from the genes fragments, we also found some intact genes that remained preserved, probable because of their importance in B chromosome maintenance and transmission. The table 6 enlists B-localized genes of our sequenced models directly involved in cell cycle and chromosomes formation including proteins coding for a variety of functions such as

chromosomes segregation, spindles fibers, microtubules, chromatin organization, chromosome condensation and regulation of cell cycle. These are the vital proteins that played a key role in the formation of B chromosome. Remarkably, the GO enrichment analyses of 6 micro dissected Bs in different species, that we investigated, also revealed a similar set of terms, thus providing solid evidence to corroborate our initial hypothesis that Bs accumulate cell cycle genes for their own advantage and establishment.

Another important set of genes enriched on the B chromosome are involved with developmental processes, mainly morphogenesis. The gene, *indian hedgehog b* (*ihhb*), involved in morphogenesis, was previously identified and highly duplicated on B chromosome in cichlid fishes (Yoshida et al. 2011; Jehangir et al. in press, 2019). An interesting concern arising from our genes enrichment results, why the B is so enriched with high abundance of morphogenesis related genes? Although the reason is unclear, we argue that the B might have accumulated these genes for its successful 'drive' during cell divisions processes. According to the definition given by Houben (2017) 'drive' chromosome transmission advantage that occurs as a result of non-Mendelian segregations, when transmissions rates are higher than 0.5. Drive is the key for understanding most B chromosomes and occurs in many ways at different stages of cell divisions (see review Houben, 2017). We assume that genes involved morphogenesis related process might play key role in the B chromosome transmission (drive) occurring during the progression of cell cycle. Morphogenesis and the cell cycle are somehow coordinated, and numerous subsequent studies have established that the core cell-cycle machinery both regulates morphogenetic events and is in turn regulated by progression of (or defects in) morphogenesis (see review Howell and Lew, 2012). A series of formative and coordinated oriented cell divisions are the principal determinants of morphogenesis (see review; Smolarkiewicz and Dhonukshe, 2013).

Among other genes found on the Bs (including the three model species and micro dissected Bs), we detected enrichment of GOs for various functions such as metabolism, cell adhesions, DNA binding,

response, localization, immune response, regulation of transcriptions, genes regulation and expressions. Genes with such kind of functions were also reported in previous studies to be located on B chromosomes (see a comprehensive and updated list of B genes in our latest review, Ahmad and Martins, 2019). The enrichment of these highly common functions among the Bs of different species suggest that B chromosome posses a common behavior to acquire a similar role inside the cell, although their genetic makeup may vary largely across taxa. Notably, the higher level of metabolism and response related B-genes in *A. mexicanus* are interesting because this fish species has been reported to have efficient metabolism as compared to other fishes. The cavefish *A. mexicanus* typically cope with the scarcity of food by evolving more sensitive tactile and chemical senses and slower or more efficient metabolisms (Protas el al. 2007). Compensatory changes like these probably evolve because of strong selection, we speculate that this may have also facilitated enrichment of metabolic and response related genes detected on the Bs of *A. mexicanus.* This suggests that Bs might have played some role in shaping the genome evolution for effective adaptation in cave environment.

Besides the genes discussed above, there are genes that are related to reproduction, that we detected on the Bs of some species e.g *A. flavolineata*. These suggest that Bs can also have a functional impact on sex determination, such as previously described by Yoshida et al. (2011) in cichlids. Our karyotype data of around all 50 individuals in *A. mexicanus*, revealing a higher prevalence of Bs occurrence in males and no B+ female, also point towards this phenomenon that the presence of Bs may have some role to determine the sex.

The comparative analysis of the B blocks to their ancestral A genome regions allowed us to unravel the evolutionary history of Bs. The presence of B blocks on almost every chromosome supports the idea that once a proto-B forms, it somehow acquires sequence from the rest of the genome. How these sequences make their way to the B, and which types of sequences are most likely to do so, is still unknown. A mechanistic model (figure 17) discussing the B origin and evolution has been proposed (Ahmad and Martins, 2019) which discuss the origin and transfer of sequences from A and eventually

forming B (For detailed discussion of this model see our review attached as supplement: Ahmad and Martins, 2019). Our findings regarding the duplication and rearrangements in the Bs provide some insights to show how Bs could evolved. Other ways of B sequences acquisition can also be possible, e.g nonhomologous recombination. The different sizes of the B blocks ranging from few hundred to thousands bp indicate that the larger regions might have migrated to Bs after the formation of a proto B. The abundance of TEs in these blocks suggests that transposition facilitated the movements and incorporation of these sequences, followed by duplications events as detected in our analyses.

Although the identification of both fragmented and complete genes on the Bs provide interesting insights, the query remains unanswered whether these genes are active in functions or not? Our predictions about the fragments to call them pseudogenes may be tempting, but further analysis at transcription level will assist to understand their exact structure and function. The occurrence of thousands of gene fragments on Bs suggest they may be part of a gene fusion, actively transcribed but might have altered their function. In addition, these partially deleted or truncated genes might have some transcriptional activity to function by interfering with the activity of the original gene. A recent study has confirmed the transcriptional activity of five genes related to cell cycle, located on the Bs of *E. plorans* (Navarro-Domínguez et al. 2017). There are few other studies that have recently confirmed the transcription expression of B-located genes (Huang et al. 2016 ; Ma et al. 2016). Before to draw any conclusion about the functionality of these genes, further evaluation of our detected genes is also needed. An analysis to test the function of our B detected sequences will serve to find out the active genes playing a role in controlling B chromosome behavior such as sex bias and drive. A better understanding about the structure and function of the B, instead of fragmented B blocks, is needed which can be achieved by a complete high-quality B chromosome assembly.

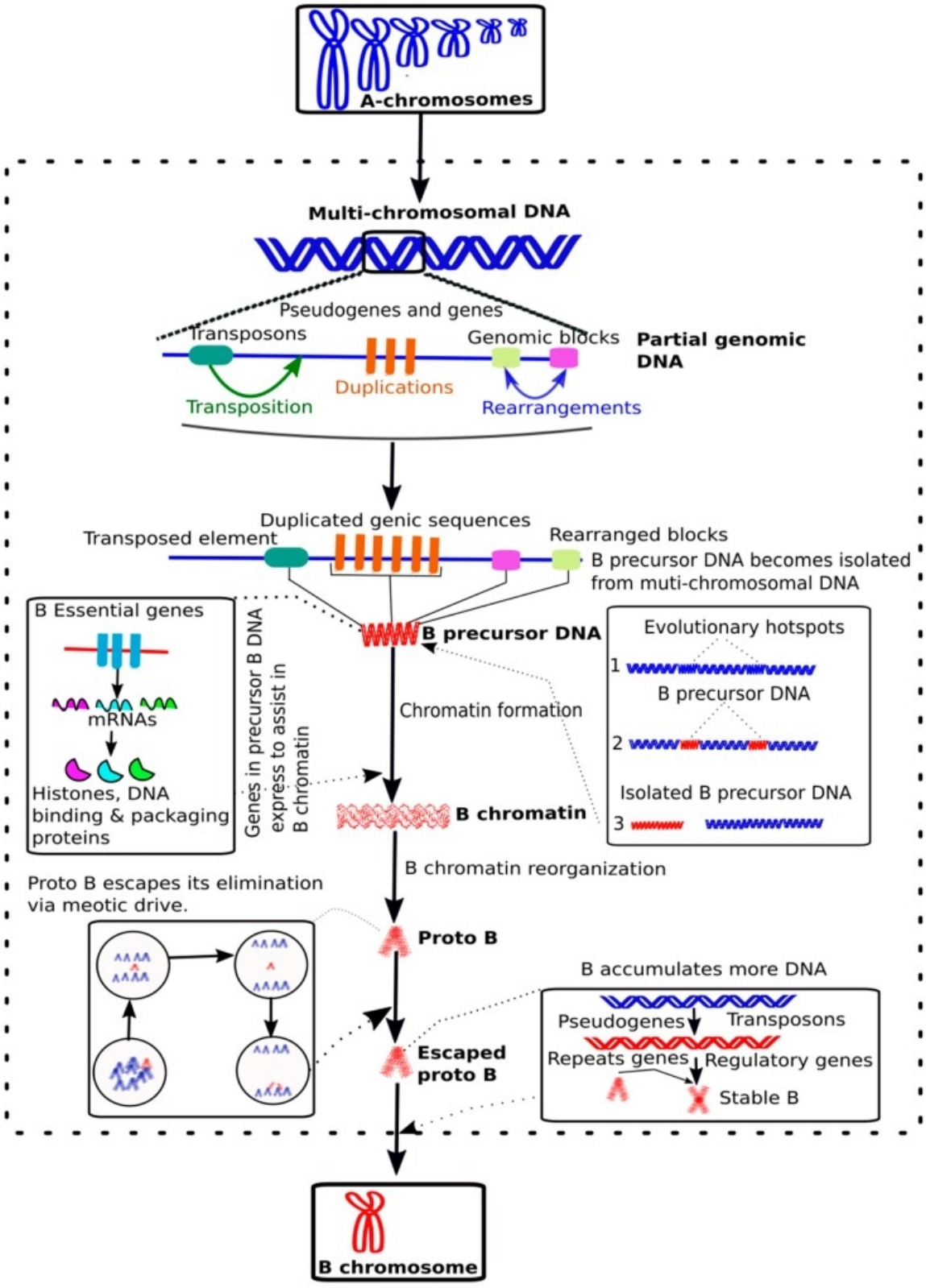

A-chromosomes
Multi-chromosomal DNA
Pseudogenes and genes
Transposons
Genomic blocks
Partial genomic DNA
Duplications
Transposition
Rearrangements
Duplicated genic sequences
Transposed element
Rearranged blocks
B precursor DNA becomes isolated from muti-chromosomal DNA
B Essential genes
mRNAs
Histones, DNA binding & packaging proteins
Genes in precursor B DNA express to assist in B chromatin
B precursor DNA
Chromatin formation
Evolutionary hotspots
1
B precursor DNA
2
Isolated B precursor DNA
3
B chromatin
B chromatin reorganization
Proto B escapes its elimination via meotic drive.
Proto B
B accumulates more DNA
Pseudogenes
Transposons
Repeats genes
Regulatory genes
Stable B
Escaped proto B
B chromosome

Figure 17. A new model of B evolution. We propose an updated model to illustrate B origin and evolution. Initially, the evolutionary hotspots in multi-chromosomal genomic DNA undergo different events such as transposition, duplications and/or genomic rearrangements; which can be considered as principal evolutionary forces. These events cause the origin of “B precursor DNA” that becomes isolated from the A chromosomes by any genomic rearrangement. The B precursor DNA could contain B essential genes (for example histones, DNA binding and packaging proteins), which are expressed to form B chromatin and its reorganization followed by the accumulation of additional DNA including repeats and genes. This ultimately results in the formation of a nascent proto-B through a series of evolutionary events as depicted. See the topic 2 in the review paper: “Genome composition, origin and evolution of B”, for details.

7. Conclusion

This study provides significant contribution and novelties towards understanding the genomic composition structure, function and evolution of B chromosome in diverse species. Applying a coverage based comparative approach we detected a considerable amount of B chromosome portion that contain thousands of gene fragments, several complete genes and an abundance of TEs along with other repeat types. We identified interesting set of genes, including but not limited to functions of chromosome and cell cycle that are crucial in maintaining of B chromosome polymorphism. The gene ontology enrichment analyses of our three model organisms and microdissected Bs genomics data further reported a diverse set of functions which shed light on the evolutionary role and fate of B inside the cell. We further discuss how B-localized contents may provide insights for theories of B chromosome evolution. Finally, we also found several patterns of genomic evolution such as duplications and rearrangements events that might have possibly shaped the evolution of B chromosome. Taken together our findings, we conclude that the Bs, which were believed as silenced elements in past, are underestimated for their participation in genome function and evolution. Our

present research opens new avenues for future research this interesting prospect and we therefore encourage further studies to investigate the expression of our detected B-localized genes for their role in the cell.

8. Supplementary data

Supplementary Table 1. A list of *A. mexicanus* primers designed for qPCR and FISH experiments of the representative B-blocks. This experimental work is currently under analyses.

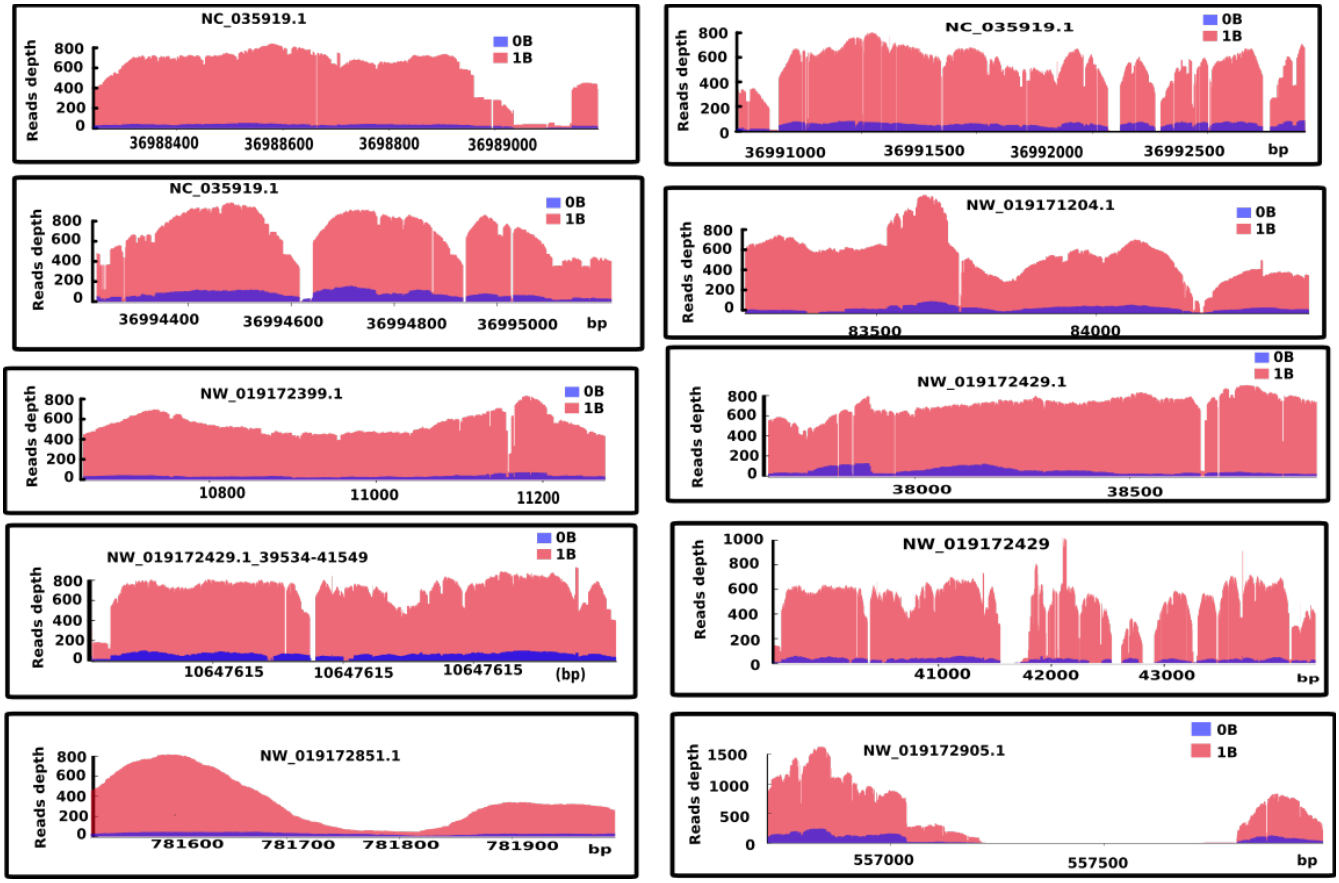
Block Name	Block ID	Forward primer	Reverse Primer	Product size (BP)	Melting temperature	Primers type
NC_035906.1_10647615-10648143	Block-1-FISH	ACATTTCCTGCCTTTTAGTTGGG	ATGACCAATTCATAGACAAGTGTCAC	336	59.3 (F) 59.6 (R)	FISH
NC_035906.1_10647615-10648143	Block-1-qPCR	ACATTTCCTGCCTTTTAGTTGGG (same)	TCATGTCACGTACACTTCACAGT	151	59.4 (F) 60 (R)	qPCR
NC_035913.1_22862787-22865602	Block-2-FISH	CCATCAGTCAGCATCCTGTTTC	cTGGCACCACTTTAAAATAAGGCT	267	59.1 (F) 59.8 (R)	FISH
NC_035913.1_22862787-22865602	Block-2-qPCR	CCATCAGTCAGCATCCTGTTTC (same)	gcgACTAAGCACTTCATAAACACA	164	59.32 (F) 59.8 (R)	qPCR
NC_035913.1_22871651-22872613	Block-3-FISH	CACACCTATTCACTCCCTTGGT	GCAGTACACAAGCAAATGACAAGA	329	59.7 (F) 60.3 (R)	FISH
NC_035913.1_22871651-22872613	Block-3-qPCR	CACACCTATTCACTCCCTTGGT	gtCCTAAGCGTGGTTAAAGTATTACC	152	59.7 (F) 59.7 (R)	qPCR
NC_035913.1_22873870-22874705	Block-4-FISH	AGTTGTTACCTGTTTACCCAACATTAG	ACGAATCATGCACTCTCACCA	406	59.8 (F) 59.7 (R)	FISH
NC_035913.1_22873870-22874705	Block-4-qPCR	AGTTGTTACCTGTTTACCCAACATTAG	ACATGAGGTAGTACATTAACAACGC	180	59.8 (F) 59.4 (R)	qPCR
NC_035913.1_23188499-23195606	Block-5-FISH	GGAGGGTAGATGGGGAAGGA	tCAGTTAGATAGAAGCTGGATGTTGT	249	59.7 (F) 59.8 (f)	FISH
NC_035913.1_23188499-23195606	Block-5-qPCR	GGAGGGTAGATGGGGAAGGA	AAACCCTGGTCTCTGCCCTA	180	59.73 (F) 60.18 (R)	qPCR
NW_019170575.1-27717_28646	Block-6-	AAGAGGCGATTCTGCGGAA	CGCCTCCTCAAGAACAGCAA	536	60.0 (F) 60.6 (R)	FISH

	FISH					
NW_019170575.1-27717_28646	Block-6-qPCR	AAGAGGCGATTCTGCGGAA	ACCCTCGGATGTGGTGATTC	205	60.0 (F) 59.46 (R)	qPCR
NW_019170910.1_26677-27858	Block-7-FISH	ACAGTTCCACCTCAGTCTCAC	CTTGGGTTCTGTAGCAGGACA	537	59.3 (F) 59.7 (R)	FISH
NW_019170910.1_26677-27858	Block-7-qPCR	ACAGTTCCACCTCAGTCTCAC	AGACACAGCAGAGCTACGGT	157	59.3 (F) 60.9 (R)	qPCR
NW_019171087.1_32674-33258 Satellite like	Block-8-FISH	GGGAATGCTCACAGGCTGA	ACACGGGTTTGAAAAATACCAACT	103	59.70 (F) 59.60 (R)	FISH
NW_019171087.1_32674-33258	Block-8-qPCR	SAME	SAME	SAME	SAME	qPCR
NW_019171959.1_9110-10065	Block-9-FISH	ATCGCGGGCAAATGTCCAAT	gACATAGGTGATACCCGGAAACA	323	61.0 (F) 60 (R)	FISH
NW_019171959.1_9110-10065	Block-9-qPCR	ATCGCGGGCAAATGTCCAAT	GGGTCACAGCAGTCATACGG	183	61.0 (F) 60.5 (R)	qPCR
NW_019172828.1_4736897-4738479	Block-10-FISH	GCGGACATTTCCTGCCTTTT	TCCATAGTGAAACGGCGTGA	617	59.40 (F) 59.40 (R)	FISH
NW_019172828.1_4736897-4738479	Block-10-qPCR	GCGGACATTTCCTGCCTTTT	TGACAGCACAGAGACAATCATCA	117	59.40 (F) 60.0 (F)	qPCR
NW_019172877.1_4865-5684	Block-11-FISH	gACTTTCAATCCCATGTTTGCCA	CCCTGAATACCCAGTTTTTCCC	493	60 (F) 60.5 (R)	FISH
NW_019172877.1_4865-5684	Block-11-qPCR	gACTTTCAATCCCATGTTTGCCA	aGCACCCCATTGCATCATGG	173	60 (F) 61 (R)	qPCR

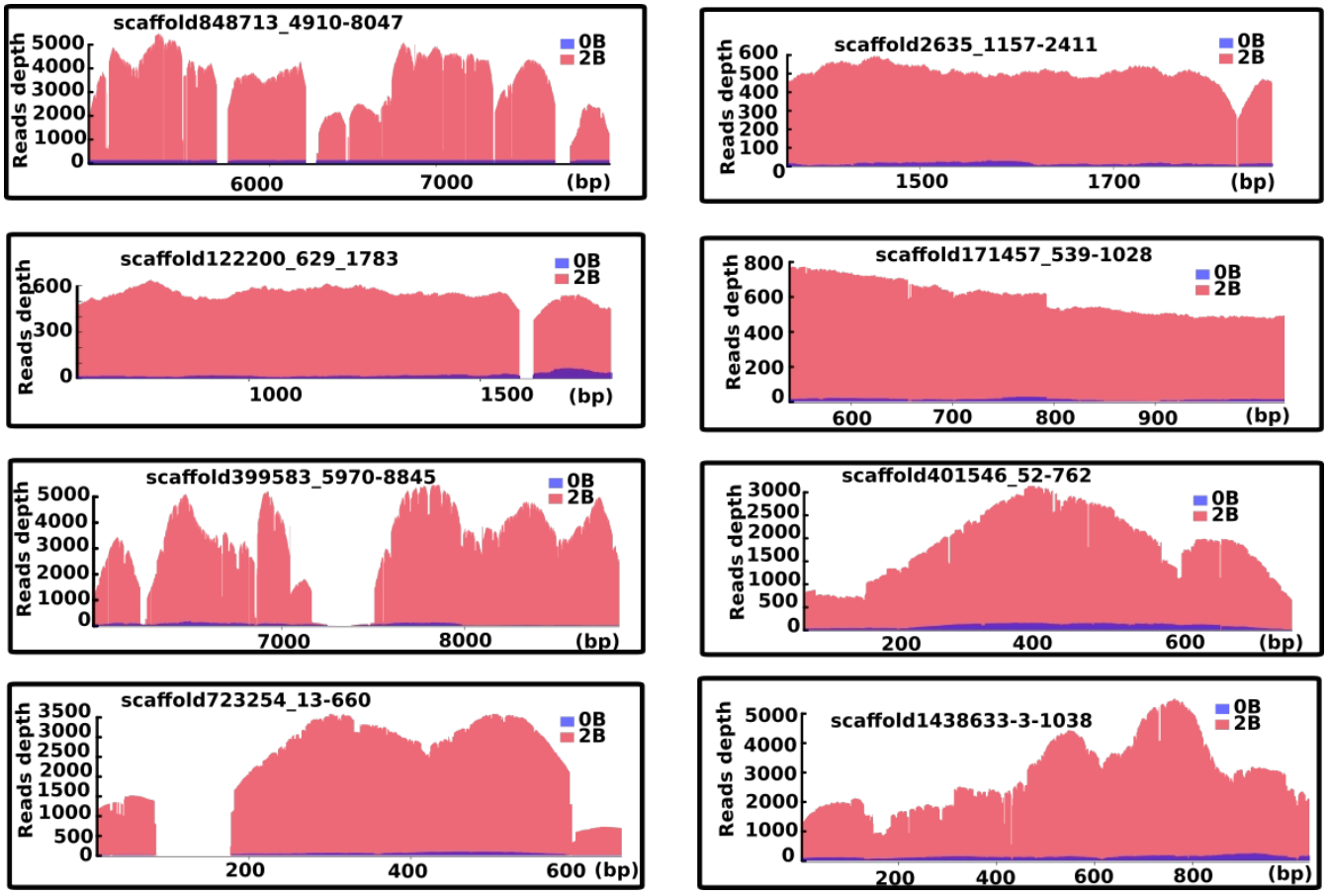
Supplementary table 2. Statistics of *de novo* assemblies performed for creating references for genomic alignments

Assembly	**Assembly *A. correntinus***	**Assembly *A. flavolineata***
# contigs (>= 0 bp)	12577502	13568784
# contigs (>= 1000 bp)	619918	1588359
# contigs	1439858	3000997
Largest contig	70304	92214
Total length	1863957204	6330161131

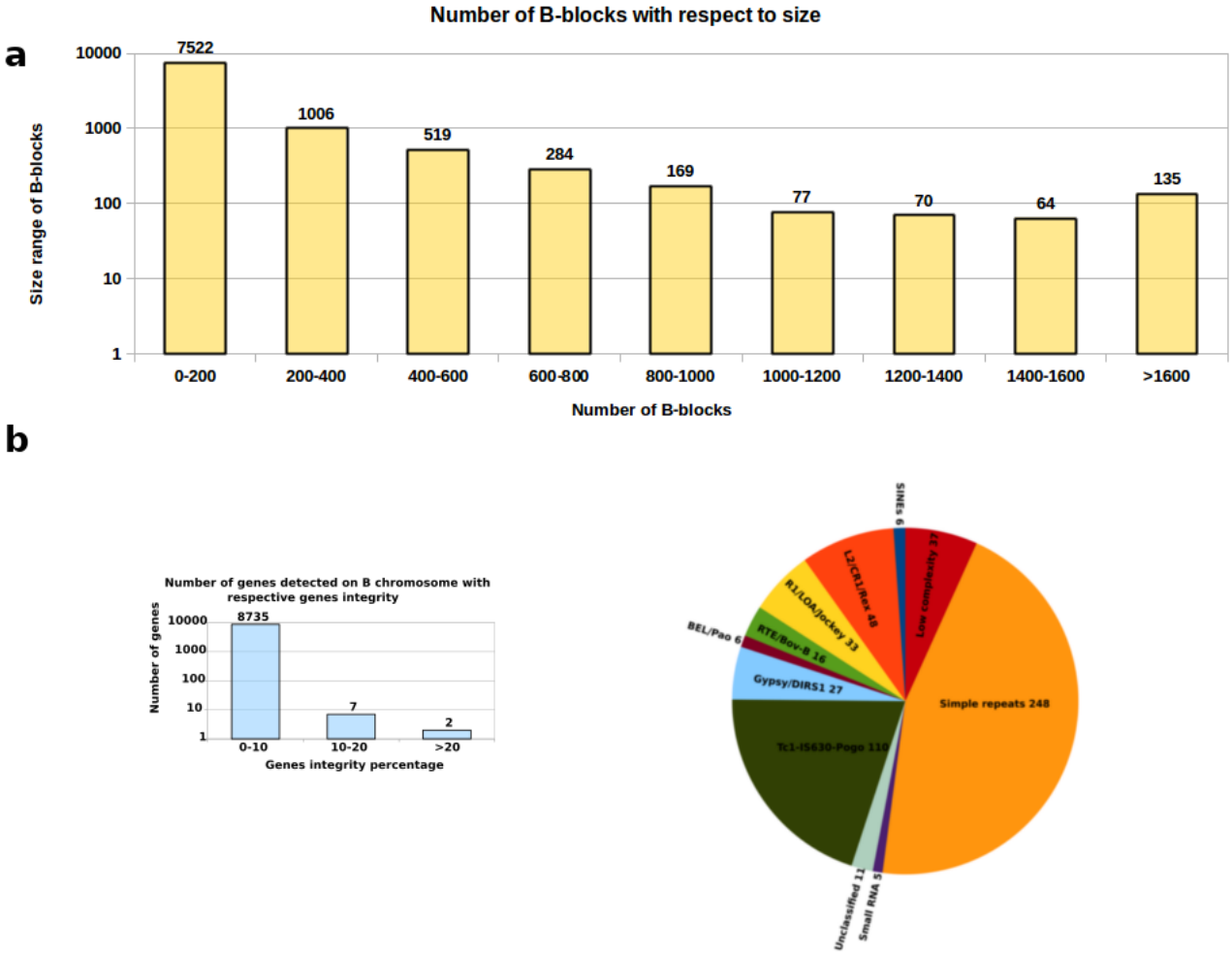
GC (%)	37.31	41.06
N50	1570	3395
N75	917	1529
L50	313997	471635
L75	712047	1174783
# N's per 100 kbp	33506.89	9432.12



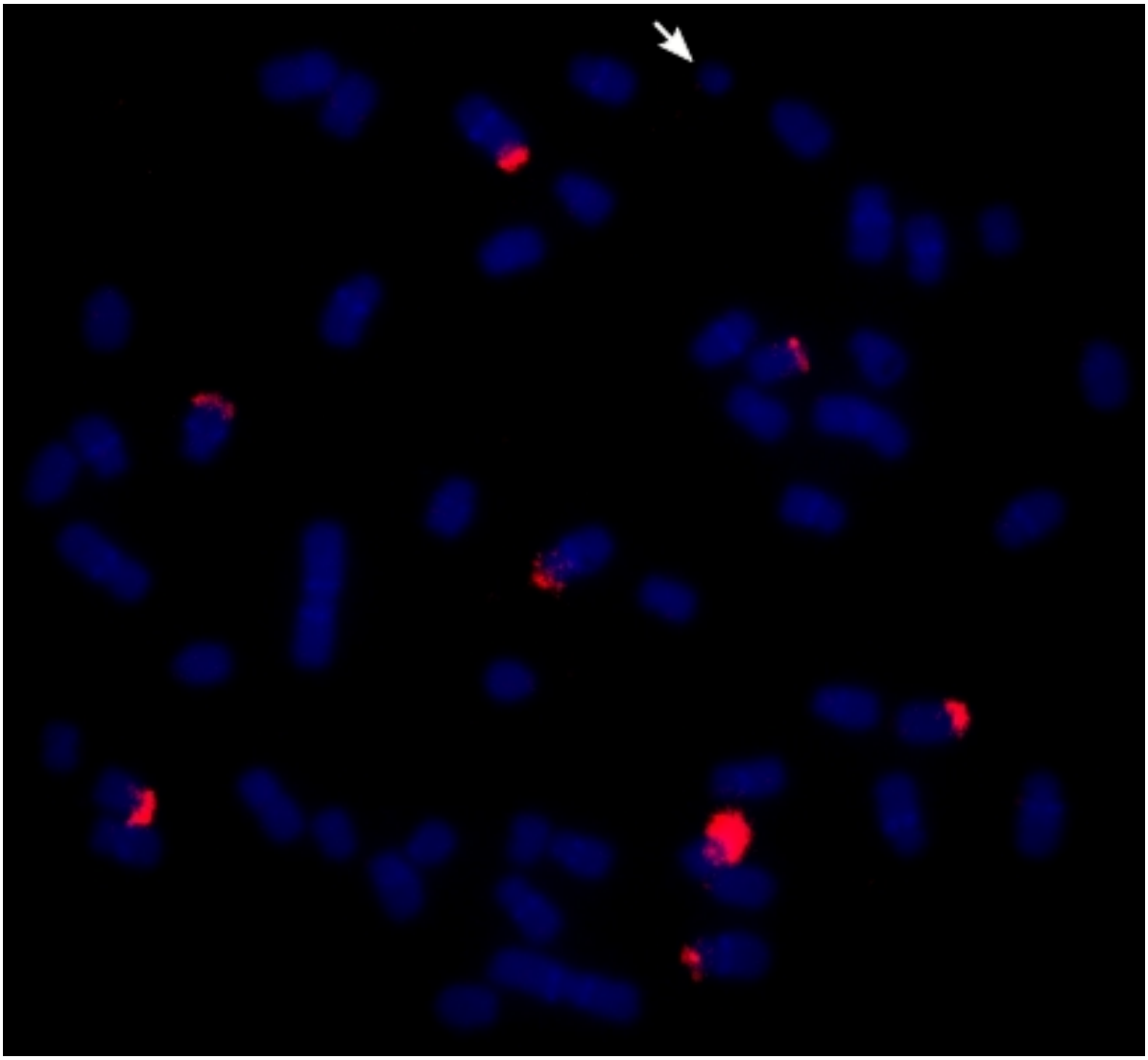
Supplementary figure 1. Coverage plots of B-blocks of *A. correntinus* with remarkable difference in the reads coverage between 0B and 1B.



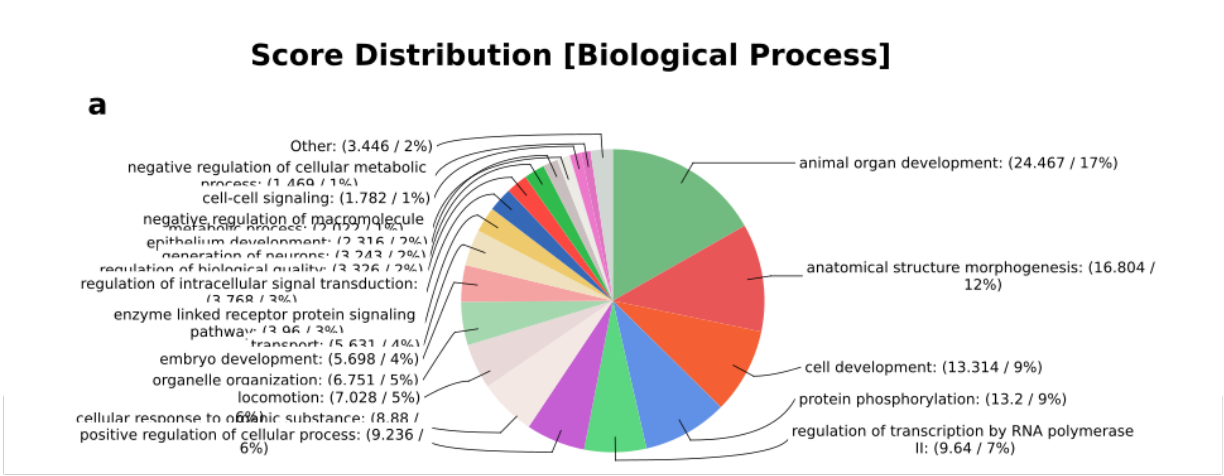
Supplementary figure 2. Coverage plots of B-blocks of *A. flavolineata* with remarkable difference in the reads coverage between 0B and 2B.

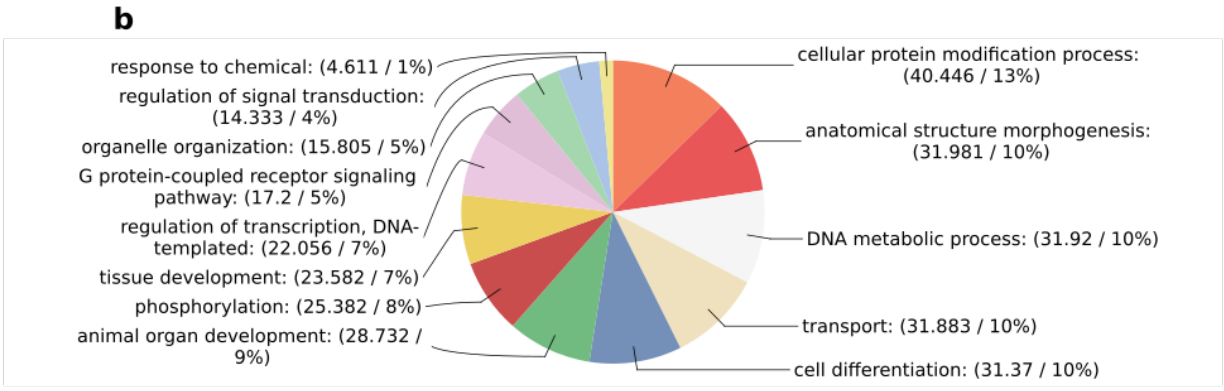


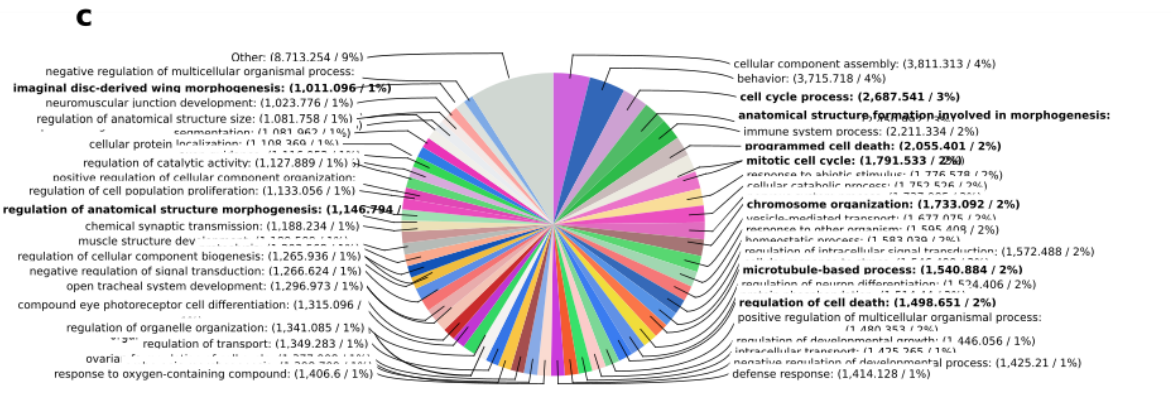
Supplementary figure 3. The B-blocks characterization of *A. flavolineata* and the repeat composition. Because of the low-quality coverage compared to the giant genome size of *A. flavolineata* the lower integrity and number repeats and genes were resulted.



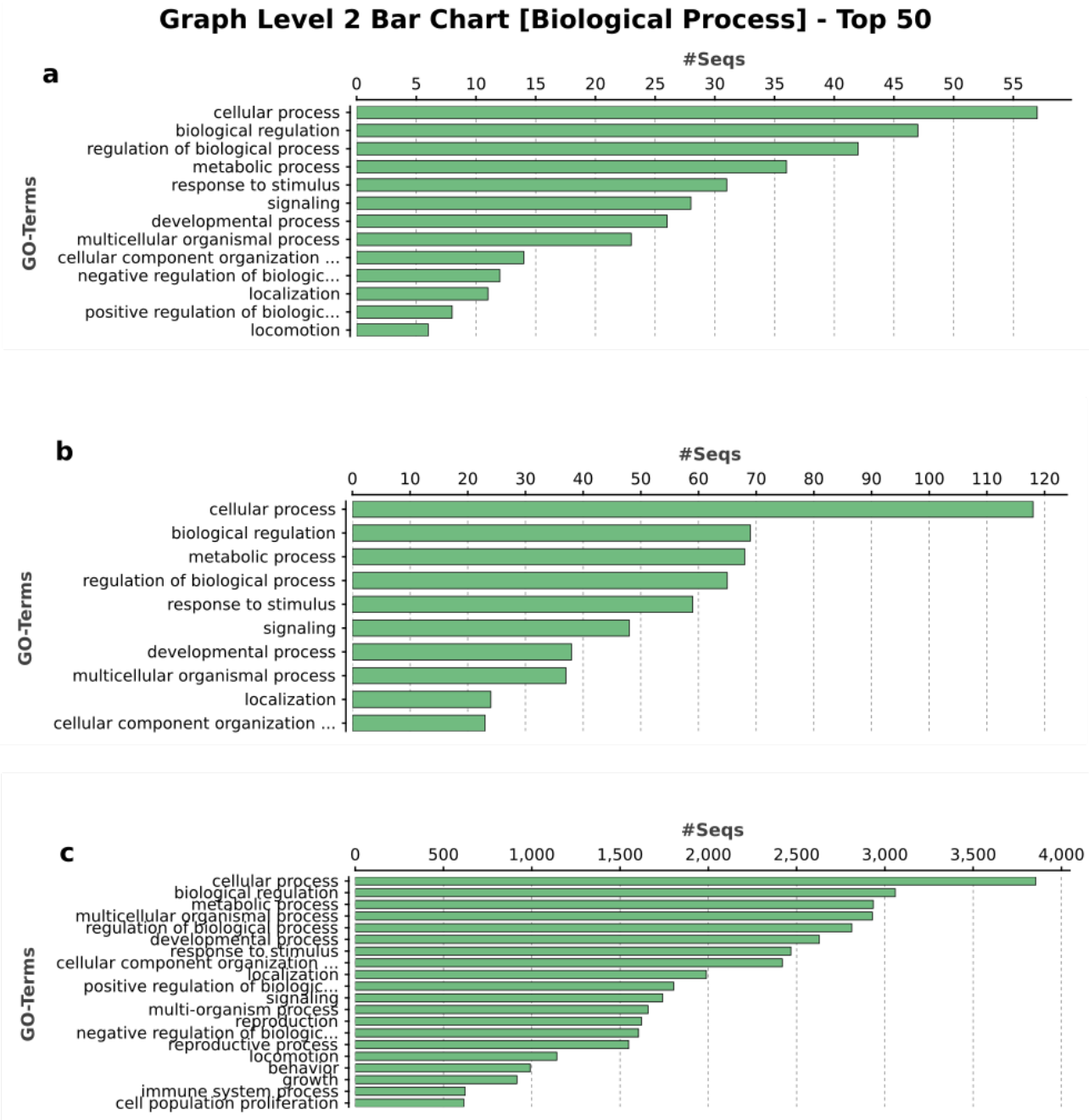
Supplementary figure 4. FISH mapping of 45S rDNA in *A. mexicanus.* The micro B is indicated with an arrow showing no sign of 45S rDNA.



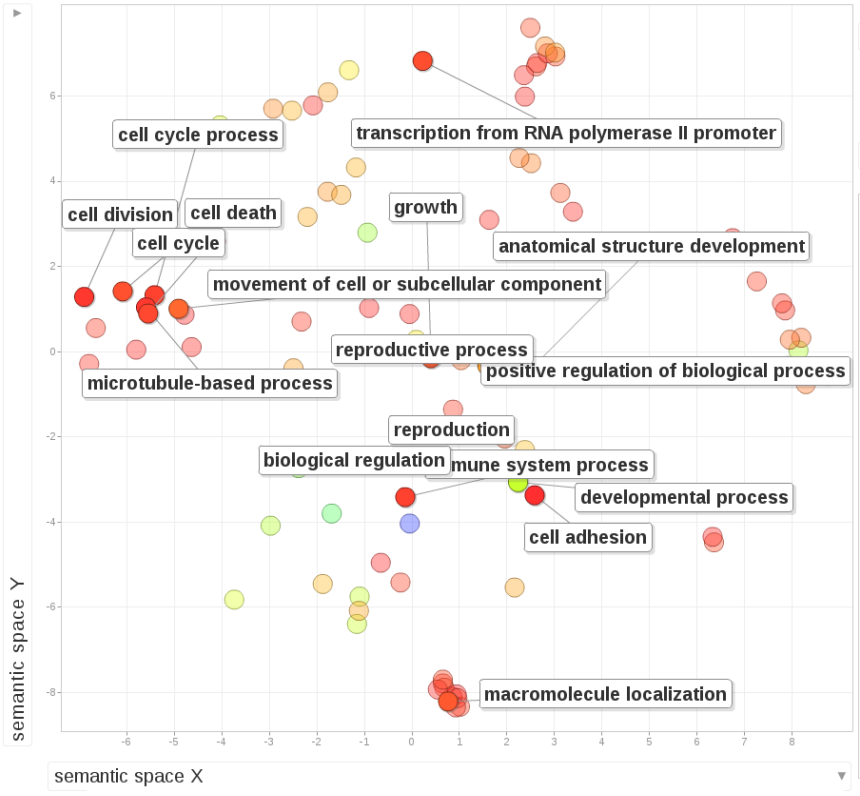




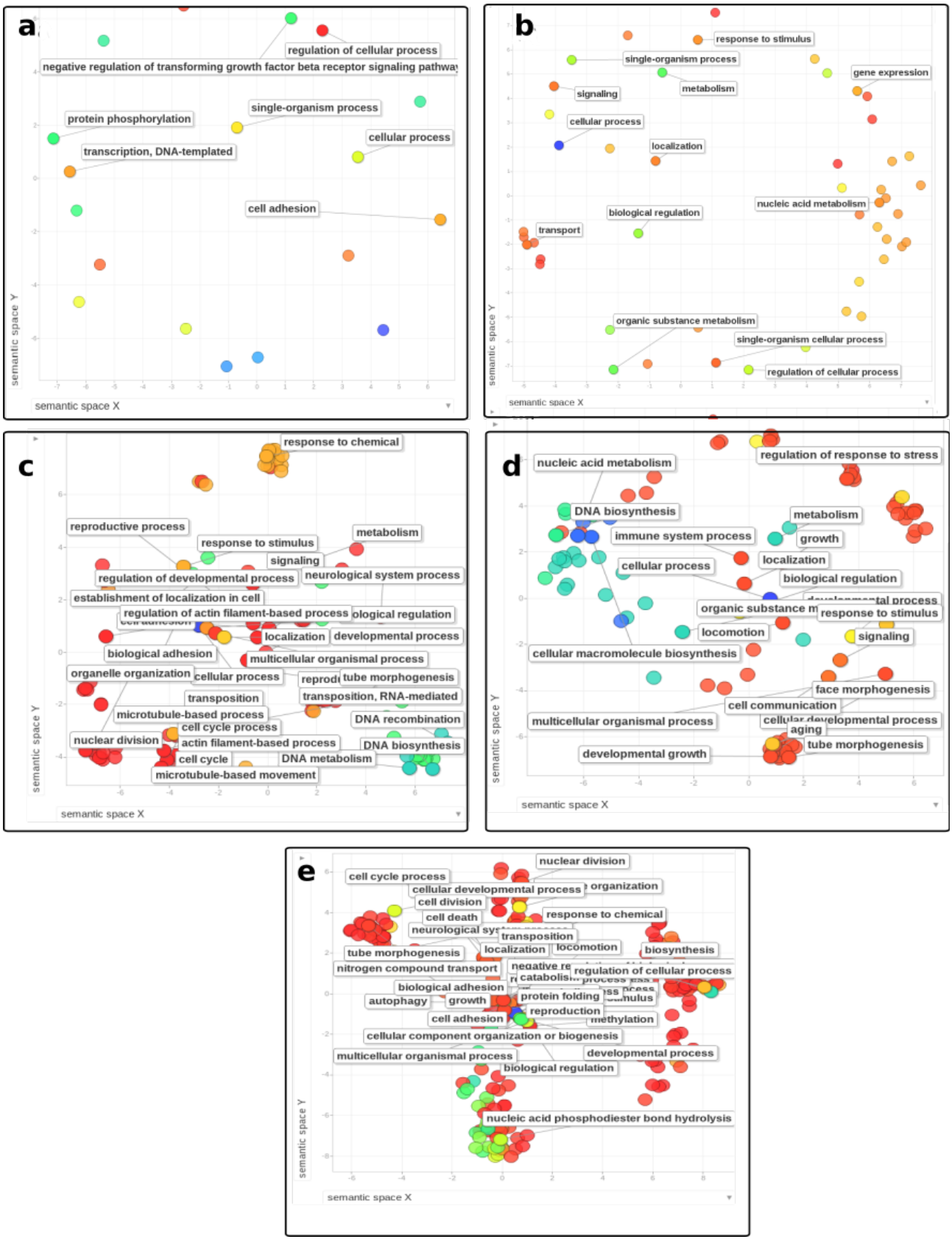
Supplementary figure 5. The GO distribution of B-blocks is shown as Pie graphs. The panels a, b and c represent the distribution of functions percentage for each species including *A. mexicanus*, *A. correntinus* and *A. flavolineata* respectively.



Supplementary figure 6. Bar charts of GO annotations recorded for B blocks of (a). *A. mexicanus*, (b). *A. correntinus* and (c). *A. flavolineata.*



Supplementary figure 7. The bubble plot of enrichment of functions in the B chromosome of *A. flavolineata* is shown. Notice the enrichment of cell cycle and chromosome related functions.



Supplementary figure 8. Gene ontology (GO) enrichment analysis of B-linked genes reported in the micro dissected Bs of diverse species. The panel **a, b, c, d** and **e** show the enrichment of B1&B2, B3, B4, B5 and B6 respectively. Enriched terms are shown as the bubble plot. Different

colors of bubbles show the intensity of enrichment for labeled function based on the log10 of P values, ranging from dark blue with highest level of enrichment. The X and Y axis do not have any intrinsic meaning.

Besides the supplementary figures and table, we could not include the large datasets as excel which will be available with scientific paper publication.

9. References

Ahmad, S. F., & Martins, C., The Modern View of B Chromosomes Under the Impact of High Scale Omics Analyses, Cells, 2019, vol. 8, pp. 156.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. "Basic local alignment search tool." J. Mol. Biol. 1990, vol. 215, pp. 403-410.

Avise, J.C., Selander, R.K., Evolutionary genetics of cave-dwelling fishes of the genus Astyanax, Evolution, 1972, vol. 26, pp. 1–19.

Banaei-Moghaddam, A.M., Meier, K., Karimi-Ashtiyani, R., Houben, A., Formation and expression of pseudogenes on the B chromosome of rye, Plant Cell Online, 2013, vol. 25, pp. 2536–2544.

Beukeboom, L.W., Bewildering Bs: an impression of the first B chromosome conference, Heredity, 1994, vol. 73, pp. 328–336.

Beukeboom, L.W., Seif, M., Mettenmeyer, T., Plowman, A.B., Michiels, N.K., Paternal inheritance of B chromosomes in a parthenogenetic hermaphrodite, Heredity, 1996, vol. 77, pp. 646-654.

Böcher, T.W., Experimental and cytogenetical studies on plant species. 5. The Copanula rotundifolia complex, Biol Skr Dan Vid Selsk, 1960, vol. 11, pp. 1–69.

Bolger, A.M., Lohse, M., Usadel, B., Trimmomatic: A flexible trimmer for Illumina sequence data, Bioinformatics, 2014, vol. 30, pp. 2114-2120.

Borowsky, R., Wilkens, H., Mapping a cave fish genome: polygenic systems and regressive evolution, J Hered, 2002, vol. 93, pp. 19–21.

Bougourd, S.M., Plowman, A.B., Ponsford, N.R., Elias, M.L., Holmes, D.S., Taylor, S., The case for unsel¢sh B-chromosomes: evidence from Allium schoenoprasum. In Kew Chromosome Conference IV, Brandham P.E., Bennet, M.d., 1995, pp. 21-34.

Bueno, D., Palacios-Gimenez, O.M., Cabral-de-Mello, D.C., Chromosomal Mapping of Repetitive DNAs in the Grasshopper Abracris flavolineata Reveal Possible Ancestry of the B Chromosome and H3 Histone Spreading, PlosONE, 2013, vol. 8, pp. E66532.

Bugrov, A.G., Karamysheva, T.V., Perepelov, E.A., Elisaphenko, E.A., Rubtsov, D.N., et al., DNA content of the B chromosomes in grasshopper Podisma kanoi Storozh. (Orthoptera, Acrididae), Chromosome Res, 2007, vol. 15, pp. 315–325.

Cabrero, J., Bakkali, M., Bugrov, A., Warchalowska-Sliwa, E., Lo´pez-Le´on. M.D., et al., Multiregional origin of B chromosomes in the grasshopper Eyprepocnemis plorans. Chromosoma, 2003, vol. 112, pp. 207–211.

Camacho, J.P.M., B Chromosomes. In: Gregory TR editor: The evolution of the genome, Elsevier, San Diego, 2005, pp. 223–286.

Camacho, J.P.M., Sharbel, T.F., Beukeboom, L.W., B-chromosome evolution, Phil Trans R Soc Lond, 2000, vol. 355, pp. 163-178.

Camacho, J.P.M., Shaw, M.W., Lo¨pez-Leo¨n, M.D., Pardo, M.C., Cabrero, J., Population dynamics of a selfish B chromosome neutralized by the standard genome in the grass- hopper Eyprepocnemis plorans, Am Nat, 1997b, vol. 149, pp. 1030-1050.

Carr, G.D. and Carr, R.L., Micro and nucieolar organizing B chromosomes in Calycadenia ciliosa, Cytologia, 1982, vol. 47, pp. 79–87.

Carter, C.R., The cytology of Brachycome. 8. The inheritance, frequency and distribution of B chromosomes in B. dichromosomatica (n ˆ 2), formerly in B. lineariloba, Chromosoma, 1978, vol. 67, pp. 109-121.

Cella, D.M., Ferreira, A., The cytogenetics of Abracris flavolineata (Orthoptera, Caelifera, Ommatolampinae, Abracrini). Revista Brasileira Genética, 1991, vol. 14, pp. 315–329.

Conesa, A., & Götz, S., Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics.

International Journal of Plant Genomics, 2008, vol. 2008, pp. 619832.

Dowling, T.E., Martasian, D.P., Jeffery, W.R., Evidence for multiple genetic forms with similar eyeless phenotypes in the blind cavefish, Astyanax mexicanus, Mol Biol Evol, 2005, vol.19, pp. 446–455.

Friebe, B., Jiang, J., Gill B., Detection of 5S rDNA and other repeated DNA on supernumerary B chromosomes of Triticum species (Poaceae), Pl Syst Evol, 1995, vol. 196, pp. 131–139.

Frost, S., The cytological behaviour and mode of transmission of accessory chromosomes in Plantago serraria, Hereditas, 1959, vol. 45, pp. 191–210.

Gall, J.G., Pardue, M.L., Formation and detection of RNA-DNA hybrid molecules in cytological preparations. Proc. Natl. Acad. Sci. USA, 1969, vol. 63, pp. 378-383.

Ge´ry J., Characoids of the World, TFH Publications, Neptune City, NJ, 1977.

Green, D.M., Zeyl, C.W., Sharbel, T.F., The evolution of hypervariable sex and supernumerary (B) chromosomes in the relict New Zealand frog, Leiopelma hochstetteri, J Evol Biol, 1993, vol. 6, pp. 417-441.

Haug-Baltzell, A., Stephens, S.A., Davey, S., Scheidegger, C.E., Lyons, E., SynMap2 and SynMap3D: web-based whole-genome synteny browsers, Bioinformatics, 2017, vol. 33, pp. 2197–2198.

Hinaux, H., Poulain, J., Da Silva, C., Noirot, C., Jeffery, W.R, Casane, D., et al. De Novo Sequencing of Astyanax mexicanus Surface Fish and Pachón Cavefish Transcriptomes Reveals Enrichment of Mutations in Cavefish Putative Eye Genes. PLoS ONE, 2013, vol. 8, pp. e53553

Holmberg, E. L. 1891. Sobre algunos peces nuevos o poco conocidos de la República Argentina. Revista Argentina de Historia Natural, Buenos Aires, 1: 180-193.

Houben, A., Banaei-Moghaddam, A.M., Klemme, S., Timmis, J.N., Evolution and biology of supernumerary B chromosomes, Cell Mol Life Sci, 2013, vol. 71, pp. 467–478.

Houben, A.B., Chromosomes - A Matter of Chromosome Drive. Front Plant Sci. 2017, vol. 8, pp. 210.

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., et al., The zebrafish reference genome sequence and its relationship to the human genome, Nature, 2013, vol. 496, pp. 498–503.

Howell, A.S, Lew, D.J., Morphogenesis and the cell cycle. Genetics. 2012, vol. 190, pp. 51–77.

Huang, W., Du, Y., Zhao, X., & Jin, W., B chromosome contains active genes and impacts the transcription of A chromosomes in maize (Zea mays L.). BMC Plant Biology, 2016, vol. 16, pp. 775-777.

Jamilena, M., Ruiz, RejonC., Ruiz, RejonM., A molecular analysis of the origin of the Crepis capillaries B chromosome, J Cell Sci, 1994, vol. 107, pp. 703–708.

Jeffery, W.R., Cavefish as a model system in evolutionary developmental biology, Dev Biol, 2001, vol. 231, pp. 1–12.

Jeffery, W.R., Strickler, A.G., Yamamoto, Y., To see or not to see: evolution of eye degeneration in Mexican blind cavefish, Integr Comp Biol, 2003, vol. 43, pp. 531–541.

Jehangir, M., Ahmad, S.F., Cardoso, A.L., Ramos, E.;,Valente, G.T., Martins, C., De novo genome assembly of thecichlid fish Astatotilapia latifasciata reveals a higher level of genomic polymorphism and genes related to B chromosomes. Chromosoma, 2019. (Accepted)

Jin, W., Lamb, J.C., Vega, J.M., Dawe, R.K., Birchler, J.A., Jiang, J., Molecular and functional dissection of the maize B chromosome centromere, Plant Cell, 2005, vol. 17, pp. 1412–1423.

Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., et al., The genomic basis of adaptive evolution in threespine sticklebacks, Nature, 2012, vol. 484, pp. 55-61.

Jones, G.H., Albini, S.M., Whitehorn, J.A.F., Ultrastructure of meiotic pairing in B chromosomes of Crepis capillaris. 2. 4B pollen mother cells, Chromosoma, 1991, vol. 100, pp. 193–202.

Jones, R.N. and Rees, H., B Chromosomes, London Acad Press, 1982.

Jones, R.N., B chromosomes in plants, New Phytol, 1995, vol. 131, pp. 411–434.

Jones, R.N., B-Chromosome Drive, The American Naturalist , 1991, vol. 137, pp. 3.

Kirby, R.F., Thompson, K.W., Hubbs, C.L., Karyotypic similarities between the Mexican and blind tetras, Copeia, 1977, vol. 1977, pp. 578–580.

Kour, G., Kaul, S., Dhar, M.K., Molecular Characterization of Repetitive DNA Sequences from B Chromosome in Plantago lagopus L, Cytogenet Genome Res, 2012, vol. 142, pp. 121–128.

Krzywinski, M., Schein, J., Birol, J., Connors, J., et al. Circos: An information aesthetic for comparative genomics, Genome Res, 2009, 19: 1639-1645.

Langecker, T.G., Wilkens, H., Junge, P., Introgressive hybridization in the Pachon cave population of Astyanax fasciatus, Ichthyol Explor Freshw, 1991, vol. 2, pp. 209–212.

Langmead, B., Salzberg, S., Fast gapped-read alignment with Bowtie 2, Nat Methods, 2012, vol. 9, pp. 357–359.

Lewis, H., The origin of supernumerary chromosomes in natural populations of Clarkia elegans, Evolution, 1951, vol. 5, pp. 142–157.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.,The Sequence alignment/map (SAM) format and SAMtools, 1000 Genome Project Data Processing Subgroup Bioinformatics, 2009, vol. 25, pp. 2078–2079.

Lima, F. C. T., Malabarba, L.R., Buckup, P.A., Pezzi da Silva, J.F.,Vari, R.F., Harold, A., Benine, R., Oyakawa,

O.T., Pavanelli, C.F., Menezes, N.A., Lucena, C.A.S., Malabarba, M.C.S.L., Lucena, Z.M.S., Reis, R.E., Langeani, F., Casatti, L., Bertaco, V.A., Moreira, C., & Lucinda, P.H.F., Genera Incertae Sedis in Characidae. 2003, Pp. 106-169.

Luo, R., Lium B., Xie, Y., et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler, Gigascience, 2015, vol. 30, pp. 1-18.

Makunin, A.I., Dementyeva, P.V., Graphodatsky, A.S., Volobouev, V.T., Kukekova, A.V., Trifonov, V.A., Genes on B chromosomes of vertebrate, Molecular Cytogenetics, 2014, vol. 7, pp. 99.

Marinn, M., Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 2011, vol. 17, pp. 10-12.

McGaugh, S.E., Gross, J.B., Aken, B., Blin, M., Borowsky, R., Chalopin, D., Hinaux, H., Jeffery, W.R., Keene, A., Ma, l., Minx, P., Murphy, D., O'Quin, k.E., Rétaux, S., Rohner, N., Searle, S.M.J., Stahl, B.A., Tabin, C., Volff, J.N., Yoshizawa, M., Warren, W.C., The cavefish genome reveals candidate genes for eye loss, Nat Commun, 2014, vol. 20, pp. 5307.

Melo, F.A.G., Revisa˜o taxonoˆmica das espe´cies do geˆnero Astyanax Baird e Girard, 1854 (Teleostei: Char´rga ˜os.aciformes: Characidae) da regia˜o da Serra dos O Rio de Janeiro, Arq Mus Nac, 2001, vol. 59, pp. 1–46.

Mestriner, C.A., Galetti, P.M.Jr., Valentini, S.R., Ruiz, I.R.G., Abel, L.D.S., Moreira-Filho, O., Camacho, J.P.M., Structural and functional evidence that B chromosome in the characid fish Astyanax scabripinnis is an isochromosome, Heredity, 2000, vol. 85, pp. 1–9.

Milani, D., Cabral-de-Mello, D.C., Microsatellite organization in the grasshopper Abracris flavolineata (Orthoptera: Acrididae) revealed by FISH mapping: remarkable spreading in the A and B chromosomes. PLoS One, 2014, vol 9, pp. e97956.

Mirande, J.M., 1 , Azpelicueta, M.L.M., Aguilera, G., Redescription of Astyanax correntinus (H OLMBERG , 1891) (Teleostei: Characiformes: Characidae), more than one hundred years after original description, Zoologische Abhandlungen (Dresden), 2006, vol. 55, pp. 9–15.

Mitchell, R.W., Russel, W.H., Elliott, W.R., Mexican eyeless characin fishes, genus Astyanax: environment, distribution and evolution, Spec Publ Mus Tex Tech Univ, 1977, vol. 12, pp. 1–89.

Navarro-Domínguez, B., Ruiz-Ruano, F. J., Cabrero, J., Corral, J. M., López-León, M. D., Sharbel, T. F., & Camacho, J. P. M., Protein-coding genes in B chromosomes of the grasshopper Eyprepocnemis plorans. Scientific Reports. 2017,vol. 7, pp. 45200.

Novák, P., Neumann, P., & Macas, J., Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics, 2010, vol. 11, pp. 378 (2010).

Oliveira, N.L., Cabral-de-Mello, D.C., Rocha, M.F., Loreto, V., Martins, C., et al., Chromosomal mapping of rDNAs and H3 histone sequences in the grasshopper Rhammatocerus brasiliensis (acrididae, gomphocerinae): extensive chromosomal dispersion and co-localization of 5S rDNA/H3 histone clusters in the A complement and B chromosome, Molecular Cytogenetics, 2011, vol. 4, pp. 24.

Paiz, L.M., Baumgärtner, L., Júnio da Graça, W., Margarido, V.P., Basic cytogenetics and physical mapping of ribosomal genes in four Astyanax species (Characiformes, Characidae) collected in Middle Paraná River, Iguassu National Park: considerations on taxonomy and systematics of the genus, CompCytogen, 2015, vol. 9, pp. 54–65.

Palacios-Gimenez, O.M., Bueno, D., Cabral-de-Mello, D.C., Chromosomal mapping of two Mariner-like elements in the grasshopper Abracris flavolineata reveals enrichment in euchromatin. Eur J Entomol, 2014, vol 111, pp. 329-334.

Panaram, K., Borowsky, R., Gene flow and genetic variability in cave and surface populations of the Mexican tetra, Astyanax mexicanus (Teleostei, Characidae), Copeia, 2005, vol. 2005, pp. 409–416.

Pardue, M.L., Gall, J.G., Molecular hybridization of radioactive DNA to the DNA of cytological preparations. Proc Natl Acad Sci U S A. 1969, vol. 64, pp. 600–604.

Parker, J.S., The B chromosome system of Hypochoeris maculata. 1. B distribution, meiotic behavior and inheritance, Chromosoma, 1976, vol. 59, pp. 167–177.

Patterson, R., The occurrence of B chromosomes in Linanthus pachyphyllus, Caryologia, 1980, vol. 33, pp. 141–149.

Pazza, R., Kavalco, K.F., Bertollo, L.A.C., Chromosome polymorphism in Astyanax fasciatus (Teleostei, Characidae). I. Karyotypic analysis, Ag-NORs and mapping of the 18S and 5S ribosomal genes in sympatric karyotypes and their possible hybrid forms, Cytogenet Genome Res, 2006, vol. 112, pp. 313–319.

Pinkel, D., Straume, T., Gray, J.W., Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization, Proc Nati Acad Sci USA, 1986, vol. 83, pp. 2934-2938.

Protas, M., Conrad, M., Gross, J.B., Regressive Evolution in the Mexican Cave Tetra, Astyanax mexicanus, current biology, 2007, vol. 24, pp. 452-454.

Protas, M.E., Hersey, C., Kochanek, D., Zhou, Y., Wilkens, H., Jeffery, W.R., Zon, L.I., Borowsky, R., Tabin, C.J., Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism, Nat Genet, 2006, vol. 38, pp. 107–111.

Raman, V.S. and Krishnaswami, D., Accessory chromosomes in Sorghum nitidum Pers, J Ind Bot Soc, 1960, vol. 39, pp. 278–280.

Romero, A., Paulson, K.M., It's a wonderful hypogean life: a guide to the troglomorphic fishes of the world,

Environ Biol Fishes, 2001, vol. 62, pp. 13–41.

Sanger, F., Nicklen, S., Coulson, A.R., DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977, vol. 74, pp. 5463–5467.

Silva, D.M., Pansonato-Alves, J.C., Utsunomia, R., Araya-Jaime, C., Ruiz-Ruano, F.J., Daniel, S.N., Hashimoto, D.T., Oliveira, C., Camacho J.P.M., Porto-Foresti, F., Foresti, F., Delimiting the Origin of a B Chromosome by FISH Mapping, Chromosome Painting and DNA Sequence Analysis in Astyanax paranae (Teleostei, Characiformes), PLOS ONE, 2014, vol. 9, pp. E94896.

Skinner, M.E., Uzilov, a.V., Stein, L.D., Mungall, C.J., Holmes, I.H., JBrowse: A next-generation genome browser, Genome Res, 2009, vol. 19, pp. 1630–1638.

Smit, A.F.A., Hubley, R., & Green, P., RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>.

Smith, J.J., Kuraku, S., Holt, C., Sauka-Spengler, T., Jiang, N., et al., Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution, Nature Genetics, 2013, vol. 45, pp. 415-421.

Smolarkiewicz, M., Dhonukshe, P., Formative cell divisions: principal determinants of plant morphogenesis. Plant Cell Physiol, 2013, vol. 54, pp. 333–342.

Sreenivasan, T.V., Cytogenetical studies in Erianthus: meiosis and behaviour of B chromosomes in 2n = 20 forms, Genetics, 1981, vol. 55, pp. 129–132.

Star, B., Nederbragt, A.J., Jentoft, S., Grimholt, U., Malmstrøm, M., et al., The genome sequence of Atlantic cod reveals a unique immune system, Nature, 2011, vol. 477, pp. 207-210.

Stark, E.A., Connerton, I., Bennett, S.T., Barnes, S.R., Parker, J.S., and Forster, J.W., Molecular analysis of the structure of the maize B chromosome, Chromosome Res, 1996, vol. 4, pp. 15–23.

Stevens, N.M., The chromosomes in Diabrotica vittata, Diabrotica soror and Diabrotica 12 punctata. A contribution to the literature on heterochromosomes and sex determination, J Exp Zool, 1908, vol. 5, pp. 453–470.

Supek, F., Bošnjak, M., Škunca, N., Šmuc, T., "REVIGO summarizes and visualizes long lists of Gene Ontology terms" PLoS ONE, 2011, vol. 7, pp. E21800.

Teruel, M., Cabrero, J., Perfectti, F., Camacho, J.P.M., B chromosome ancestry revealed by histone genes in the migratory locust, Chromosoma, 2010, vol. 119, pp. 217–225.

Tian, N.M.M.L., Price, D.J., Why cavefish are blind, Bioessays, 2005, vol. 27, pp. 235–238.

Valente, G.T., Conte, M.A., Fantinatti, B.E., Cabral-de-Mello, D.C., Carvalho, R.F., Vicari, M.R., Kocher, T.D., Martins, C., Origin and evolution of B chromosomes in the cichlid fish Astatotilapia latifasciata based on integrated genomic analyses, Mol Biol Evol, 2014, vol. 31, pp. 2061–2072.

Valente, G.T., Nakajima, R.F., Fantinatti, E.A., Marques, D.F., Almeida, R.O., Rafael, P., Martins, C., B chromosomes: from cytogenetics to systems biology, Chromosoma, 2016, vol. 126, pp. 73-81.

Vasil'ev, VP., Chromosome numbers in fish-like vertebrates and fish, J Ichthyol, 1980, vol. 20, pp. 1–38.

Venkatesh, B., Kirkness, E.F., Loh. Y-H., Halpern, A.L., Lee, A.P., et al., Survey sequencing and comparative analysis of the elephant shark (Callorhinchus milii) genome, PLoS Biology, 2007, vol. 5, pp. e101.

Vij, S., Kuhl, H., Kuznetsova, I.S, Komissarov, A., Yurchenko, A.A, et al., Correction: Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence Reads and Multi-layered Scaffolding. PLOS Genetics, 2016, vol. 12, pp. e1006500.

Wang, X., Yang, P., Jiang, F., Zhao, D., Li, B., Cui, F., Wei, J., Ma, C., Wang, Y., He, J., et al. The locust genome provides insight into swarm formation and long-distance flight. Nature Communications, 2014, vol. 5, pp. 2957.

Wicker, T., Sabot, F., Hua-Van, A., Jeffrey, L.B, Capy, P., Chalhoub, B., Flavell, F., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman, A.H., A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007, vol. 12, pp. 973–982.

Wilkens, H., Evolution and genetics of epigean and cave Astyanax fasciatus (Characidae, Pisces), Evol Biol, 1988, vol. 23, pp. 271–367.

Wilkes, T.M., Francki, M.G., Langridge, P., Karp, A., Jones, R.N., Forster, J.W., Analysis of rye B chromosome structure using fluorescence in situ hybridization (FISH), Chromosome Res, 1995, vol. 3, pp. 466–472.

Wilson, E.B., Studies on chromosomes. 5. The chromosomes of Metapodius. A contribution to the hypothesis of the genetic continuity of chromosomes, J Exp Zool, 1906, vol. 6, pp. 147–205.

Yoshida, K., Terai, Y., Mizoiri, S., Aibara, M., Nishihara, H., Watanabe, M., Kuroiwa, A., Hirai, H., Hirai, Y., Matsuda, Y., Okada, N., B chromosomes have a functional effect on female sex determination in Lake Victoria cichlid fishes, PLoS Genet, 2011, vol. 7, pp. E1002203.