




De novo genome assembly of the cichlid fish *Astatotilapia latifasciata* reveals a higher level of genomic polymorphism and genes related to B chromosomes

Maryam Jehangir¹ · Syed F. Ahmad¹ · Aduino L. Cardoso¹ · Erica Ramos¹ · Guilherme T. Valente² · Cesar Martins¹ 

Received: 11 January 2019 / Revised: 27 February 2019 / Accepted: 7 May 2019 / Published online: 21 May 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

Supernumerary B chromosomes (Bs) are accessory elements to the regular chromosome set (As) and have been observed in a huge diversity of eukaryotic species. Although extensively investigated, the biological significance of Bs remains enigmatic. Here, we present de novo genome assemblies for the cichlid fish *Astatotilapia latifasciata*, a well-known model to study Bs. High coverage data with Illumina sequencing was obtained for males and females with 0B (B⁻), 1B, and 2B (B⁺) chromosomes to provide information regarding the diversity among these genomes. The draft assemblies comprised 771 Mb for the B⁻ genome and 781 Mb for the B⁺ genome. Comparative analysis of the B⁺ and B⁻ assemblies reveals syntenic discontinuity, duplicated blocks and several insertions, deletions, and inversions indicative of rearrangements in the B⁺ genome. Hundreds of transposable elements and 1546 protein coding sequences were annotated in the duplicated B⁺ regions. Our work contributes a list of thousands of genes harbored on the B chromosome, with functions in several biological processes, including the cell cycle.

Keywords Cichlid fish · Next-generation sequencing · Chromosome rearrangement · Extra chromosome · Supernumerary chromosome · Genome evolution

Introduction

B chromosomes (Bs) are accessories to the normal chromosome set (As) that is present in some individuals of a large number of diverse eukaryotic species. These extra chromosomes usually do not recombine with members of the A chromosomes and do not follow the rules of Mendelian segregation (Jones 2018). They are mostly heterochromatic, composed of a large amount of repetitive DNAs, are not needed for survival or reproduction of individuals, and are maintained through a drive-parasitic mechanism (Camacho 2005). Drive is a specialized process by which Bs can escape elimination

during the cell cycle and be transmitted at a higher rate than the normal Mendelian transmission frequency of As (Houben 2017; Jones 2018). The combination of next generation sequencing (NGS) and modern bioinformatics technologies has provided new methods to identify B-located sequences (Ruban et al. 2017). Recent works suggest that Bs are comprised of fragmented or integral sequences derived from different As and from organelle DNA (Houben et al. 2014; Valente et al. 2014; Banaei-Moghaddam et al. 2015; Ruban et al. 2017). Despite the increase of knowledge about the genomic content, origin, and pattern of evolution of Bs, the biological significance of these elements still remains unclear. Recent studies have shown that B chromosomes carry transcriptionally active DNA sequences and also influence the transcription of other sequences in the cell that could play over a variety of cellular functions (Carmello et al. 2017; Houben 2017; Navarro-Domínguez et al. 2017, 2019; Ramos et al. 2017; Valente et al. 2017).

Among the African cichlid fishes, B chromosomes were first described in *Astatotilapia latifasciata* from Lake Nawampasa, a satellite lake of the Lake Kyoga system (Poletto et al. 2010). Previous analyses of the B chromosome in *A. latifasciata* were based on cytogenetics and comparative

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00412-019-00707-7>) contains supplementary material, which is available to authorized users.

✉ Cesar Martins
cesar.martins@unesp.br

¹ Department of Morphology, Institute of Bioscience at Botucatu, São Paulo State University – UNESP, Botucatu, SP 18618-689, Brazil

² Bioprocess and Biotechnology Department, Agronomical Science Faculty, Sao Paulo State University – UNESP, Botucatu, SP, Brazil

genomics studies. Cytogenetics confirmed that both sexes of *A. latifasciata* can carry either no, one, or two similar B chromosomes enriched with many repetitive DNA sequences (Poletto et al. 2010; Fantinatti et al. 2011). The genomic content of *A. latifasciata* Bs was assessed by comparative coverage analysis of 0B and 2B individuals, followed by experimental validation of qPCR and fluorescence in situ hybridization (FISH) mapping as well as sequencing of microdissected Bs (Valente et al. 2014) that detected an enrichment of transposons, thousands of degenerative genic sequences, and a few complete genes. The complete genes were found to be related to functional terms such as “cell cycle” and “chromosome-associated.” Furthermore, it was revealed that some of the intact genes are potentially transcriptionally active. Recent functional studies in *A. latifasciata* have revealed transcriptional variations of diverse DNA sequence classes including protein code genes, non-coding RNAs, and repetitive sequences (Carmello et al. 2017; Ramos et al. 2017; Valente et al. 2017; Coan and Martins 2018) related to B chromosome presence. However, a better comprehension of genes and functional sequence organization in the genome has been hindered due to lack of an assembled genome for this species. Among the genes discovered in the B chromosome of diverse species (see for review Ahmad and Martins 2019), there is a morphogenesis-related gene named Indian hedgehog b (*ihhb*) detected in the B chromosome of the cichlid *Lithochromis rubripinnis* (Yoshida et al. 2011). Development and morphogenesis have appeared as enriched terms based on gene ontologies (GO) of B chromosome genes described for diverse species (Ahmad and Martins 2019). The hedgehog (hh) gene family was firstly reported in *Drosophila*, and it is involved with embryo polarity and code proteins with two types of domains that function during the embryonic development of skeletal and nerve systems (Nüsslein-Volhard and Wieschaus 1980). The hh family of vertebrates has three genes called sonic hedgehog (*shh*), desert hedgehog (*dhh*), and Indian hedgehog (*ihh*), each with different pattern of expression and playing important roles in diversification and complexity of vertebrates (Pereira et al. 2014). It has been suggested that the paralogous of these hh genes have evolved under different evolutionary rates after the whole genome duplication (WGD) events in vertebrates (Pereira et al. 2014). The presence of *ihhb* copies in the B chromosome of a cichlid species (Yoshida et al. 2011) opens the discussion of the possible effects of B chromosomes over important biological features.

Here, we present de novo assemblies and annotation for *A. latifasciata* genomes with and without B chromosomes, and characterize the genomic diversity of samples containing B chromosomes. These genome assemblies reveal important aspects of B chromosome biology. We detected extensive genomic rearrangements related to the B chromosome presence and identified thousands of coding genes harbored in the B chromosome. In addition, we performed an analysis of

sequence coverage, coupled with FISH mapping, which revealed the existence of high copy number of inactive *ihhb* gene emerging as a major structural component of the B chromosome.

Materials and methods

Chromosome preparation, DNA sampling, and genome sequencing

Astatotilapia latifasciata samples were karyotyped by classical chromosome preparation protocols to check the number of B chromosomes and the metaphasic chromosome suspensions were stored for FISH mapping. The samples were named as B− (absence of B chromosomes) or B+ (presence of 1 or 2 B chromosomes). FISH probes were obtained from genomic DNA via polymerase chain reaction (PCR) (Fantinatti and Martins 2016).

High quality genomic DNA (gDNA) samples from five karyotyped *A. latifasciata* specimens (males and females) with 0B and 1B chromosome were selected for next generation sequencing (NGS). We also reanalyzed two additional read datasets from Valente et al. (2014) (Table 1). The Illumina libraries were sheared to an average size of 350–550 bp using an S220 focused ultrasonicator (Covaris Inc., Woburn, USA) and prepared using the TruSeq DNA sample preparation kit ver.2 rev. C (Illumina Inc., San Diego, USA). Paired-end sequencing was performed using the Illumina HiSeq 1000 and MiSeq platforms. Read quality was checked using the FastQC software (Andrews 2010). Data filtering was performed using the FASTX toolkit (Gordon and Hannon 2010), retaining only reads with a minimum of 90% of nucleotides showing at least 30 in Phred quality score. Reads containing Illumina adapter sequences were eliminated using BLASTn search (E -value $\leq 10e-5$ and 90% of identity as cutoff parameters) and customized Python programming script. Reads without a mate (singletons) were discarded by Pairfq software (<https://github.com/sestaton/Pairfq>). The coverage was calculated for all samples by the equation $Cov = (rc \times rl)/S$, where rc is the read count, rl is the read length, and S the genome size. We considered *A. latifasciata* genome size comparable to *Metriaclima zebra* genome (O’Quin et al. 2013) since this is the most closely related cichlid genome to *A. latifasciata*. All datasets are available at Sacibase (sacibase.ibb.unesp.br).

Genome assembly and quality evaluation

The reads passing pre-processing filters were sorted into two groups: B− and B+ data. They were then passed to KMERGENIE (Chikhi and Medvedev 2014) to obtain the optimal kmer values for genome assembly. The separate datasets were used to produce two de novo assemblies: B−

Table 1 Illumina sequencing data obtained for *A. latifasciata*. M, male; F, female; 0B, sample without B chromosome; 1B, sample with 1B chromosome; 2B, sample with 2B chromosomes

Sample ID	Reads length	Raw data coverage	Coverage after filtration	Total reads	Remained reads after filtration	Reference
M1-0B	101	47.7 ×	38 ×	401,017,570	323,226,972(80%)	Valente et al. 2014
F1-0B	191	75.5 ×	42.5 ×	337,349,994	190,214,688 (56%)	Present work
M2-1B	35–250	2 ×	1.6 ×	716,030	591,126 (82%)	Present work
M3/4-1B	101	16.8 ×	13.4 ×	143,441,264	114,064,786 (79%)	Present work
M4-1B	101	43.1 ×	34.8 ×	366,602,572	296,161,988 (80%)	Present work
F2-1B	191	70.2 ×	40.1 ×	313,818,884	179,286,280 (57%)	Present work
M5-2B	101	43.6 ×	30 ×	306,823,512	254,993,955 (83%)	Valente et al. 2014

and B+ genomes, using the Velvet assembler (v 1.2.08) (Zerbino and Birney 2008). This assembler was recommended by Assemblathon 2 competition (Bradnam et al. 2013). The parameters used were “-ins_length 500, exp-cov auto” and “-unused_reads yes, read_trkg yes.” To close the assembly gaps within scaffolds, we ran the GAPPILLED algorithm (Boetzer and Pirovano 2012) using parameters (-‘m’ = 80, ‘-t’ = 10, ‘-g’ = 5). We computed several metric values (length, number, length variation, N50, gap length) of each assembly using QUAST software (Gurevich et al. 2013). To evaluate the completeness of the B- and B+ assemblies, we searched for a set of 453 core eukaryotic genes using the CEGMA (version 2.4) pipeline (Parra et al. 2007). The assembled genomes are available at Sacibase database.

Genome annotation

We used three methodologies for gene annotation of B- reference assembled genome: identity to known genes available in current databases, de novo prediction, and transcript sequences-alignment using the MAKER v2.31.8 pipeline (Cantarel et al. 2008). We annotated repetitive elements using a custom fish and a general metazoan database. We used coding (CDS) and protein sequences from *Danio rerio* (http://www.ensembl.org/Danio_rerio) to annotate genes. We also used transcriptomes from *A. latifasciata* (Nakajima et al. in preparation) to inform the annotation. We used the lamprey training set available in Augustus software (Stanke et al. 2004) for gene prediction.

The annotated virtual genes were extracted from the assembled genome using customized unix scripts and mapped to a *D. rerio* protein database, retrieved from ensemble (ftp://ftp.ensembl.org/pub/release-91/fasta/danio_rerio/pep/). The BLASTx mapping at NCBI was performed with a minimum *E*-value of 1×10^{-5} and the output xml formatted file was imported into BLAST2GO (Götz et al. 2008). We mapped the resulting aligned proteins of the corresponding query genes against the GO database to obtain the functional information. The functional annotation was processed using default parameters for all gene functions.

Comparative genomics and genome diversity analysis

Genomic rearrangements were identified comparing the two genome assemblies (B+ and B-) based on whole genome pairwise sequence alignments approach. For this purpose, we used minimap2 (Li 2018) mapping and the output results were plotted as dotplots using an R script called dotPlotly (<https://github.com/tpoorten/dotPlotly>). To better interpret the dotplots patterns of synteny, we also perform self-alignments between B+ with B+, and B- with B- genomes. Finally, we analyzed the rearranged blocks identified from the alignments of B+ with B- at several filtering steps of query length and mapping parameters. The filtering of mapped blocks was considered for better visualization of synteny to explore the differences between the two genomes.

We also annotated genes and repeats that were duplicated in the B+ genome. This analysis consisted of several steps. First, we identified these regions by the number of times they were repeatedly aligned to the query sequence of the B- genome. If multiple scaffolds of the B+ assembly align to the same B- target, then the segment is considered duplicated in the B+ genome. To ensure the identification of duplicated block, we compared the total scaffold size to its respective alignment size and calculated the duplicated copy ratio by the following equation, $DR = (al \times ql)/l$, where *DR* is the duplicated copy ratio, *al* is the alignment block length, and *ql* is the total query sequence length; *l* represents a regular single copy in the genome. The *DR* for the same alignment entries was then summed for each query sequence to yield a total value. The total *DR* of a specific scaffold with values greater than 2 indicates the recurrence of two copies of a given sequence. We expect for a regular, not duplicated sequence *DR* value of 1. We extracted the duplicated blocks by customized bash scripts using a threshold of at least twice repeated alignments of a similar scaffold. This extraction was done from the B+ and B- alignments file generated by minimap2 (paf, Pairwise mApping Format). We then mapped these blocks to the reference coding sequence (CDS) as well as proteome database of zebrafish by BLAST. The blocks were also subjected to RepeatMasker program (Smit et al. 2013-

2015) to identify transposable elements. Finally, the functional annotation and GO enrichment of the blocks was done using BLAST2GO pipeline. The enrichment of GOs was plotted using Revigo (Supek et al. 2011).

The read datasets of *A. latifasciata* (males and females with 0B, 1B, and 2B chromosomes, Table 1), were aligned to our B– genome using Bowtie2 (Langmead and Salzberg 2013) with “–very-sensitive” option. Nucleotide polymorphisms were identified using SAMtools (Li et al. 2009) to search for genome variations among the samples. The output variant call format (VCF) files were passed to VCFtools (vcf-stats and vcf-compare) (Danecek et al. 2011) for statistical analysis to discover the frequency of SNPs and insertions/deletions (indels). We compared these variant calls to identify shared, unique, or B chromosome–related variations. Filtering was carried out to eliminate the lower quality ($Q \leq 20$ and $DP \geq 100$) SNPs and indels using vcflib (Garrison 2012). A similar approach was followed to detect variations in the genomic data aligned to de novo transcriptome assembly of *A. latifasciata* (Nakajima et al. in preparation).

The structural variations (translocations) were analyzed based on 1B sequencing data using Delly (Rausch et al. 2012). This pipeline was applied to both B+ and B– reads in Bam format (generated using bowtie2 tool) taking B+ as the query and the B– de novo genome assembly of *A. latifasciata* as reference to locate the variations on the scaffold regions. Translocations (breakpoints) were visualized by ClicO (Cheong et al. 2015), an online web-service (<http://codoncloud.com:3000/home>) based on Circos (Krzywinski et al. 2009).

Structural variations (SVs) such as deletions, insertions, transversions, inversions, and duplications in genomic regions related to the B chromosomes were analyzed by inGAP-sv tool (Qi and Zhao 2011). InGAP-sv detects SVs on the basis the pattern and coverage of mapped paired-end reads. We applied this pipeline to the 2B SAM file generated using BWA (Li and Durbin 2009). The B– assembled genome was used as a reference. After the SAM file was loaded into inGAP-sv, a threshold of mapping quality (default value: 20) was applied to filter non-uniquely mapped reads. Illustrations of paired-end mapping (PEM) patterns for different types of identified SVs were generated according to Qi and Zhao (2011).

Analysis of B-specific sequences

The nucleotide sequences of several genes previously identified on B chromosomes in vertebrates were retrieved from NCBI (Table S1). Consensus sequences constructions were obtained using Geneious v. 4.8.5 software (Drummond et al. 2009) for genes with more than one sequence available. The final sequences were used as queries against the B– assembly in a standard BLASTn search. The number of BLAST hits, *E*-

values, and percent identity were evaluated before proceeding further with B chromosome–related analyses. Although most of the genes were either absent, or had only partial sequence in the *A. latifasciata* genome, the *45S rRNA* and *ihhb* (Indian Hedgehog B) genes were considered for future analysis because of their high level of integrity. The Illumina high coverage reads of all B– (0B male and female) and B+ (1B males, 1B female, and 2B male) samples were aligned to both reference *45S rRNA* and *ihhb* gene copies described for the cichlids *Oreochromis aureus* and *Lithochromis rubripinnis* respectively, using the paired-end mode of Bowtie2 (very sensitive alignment option). The outputs were converted to binary format and indexed using samtools. Each file was normalized using RPKM (reads per kilobase per million mapped reads) package of deeptools (Ramírez et al. 2014) to correct bias in initial coverage. These files were then visualized using the integrated genome browser IGB (<http://bioviz.org/>) to compare coverage of both genes in B– and B+ samples. SNPs at different sites of reads were detected. The transcription of *45S rRNA* and *ihhb* genes was assessed based on the available reads datasets of *A. latifasciata* transcriptomes (Nakajima et al. in preparation). The uploaded transcriptomic and genomic data (aligned files) were visualized and manually evaluated in Sacibase. BLASTn searches of *45S rRNA* and *ihhb* genes in *A. latifasciata* transcriptome assembly (Table S2) (Nakajima et al. in preparation) were conducted to locate those genes in specific scaffolds.

Primer designing, probes construction, and fluorescence in situ hybridization mapping of *ihhb* and *45S rRNA* genes

Primers were designed for the *ihhb* gene and the *45S rRNA* cistron (including the transcribed segments for the 5.8S, 18S, and 28S rRNAs) (Table S3) using PrimerQuest (<http://www.idtdna.com/primerquest/home/index>). Specificity of the primers was checked using Primer-Blast (<http://www.ncbi.nlm.nih.gov/tools/primer-blast/>) and evaluated by Primer Stat (Stothard 2000). Genomic DNA was subjected to PCR to obtain DNA probes for 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, USA) with a Dynamic Terminator Cycle Sequencing Kit (Applied Biosystems) as per the manufacturers’ instructions. The sequences obtained were subjected to BLAST (Altschul et al. 1990) searches at the NCBI to confirm if they correspond to the annotated genes. Probes were labeled with digoxigenin-11-dUTP (Roche Applied Science, Penzberg, Germany) using PCR and the signal was detected with anti-digoxigenin-rhodamine (Roche Applied Science). FISH was performed using the protocol described by Pinkel et al. (1988) with modifications (Cabral-de-Mello

et al. 2012). The slides were denatured in 70% formamide/2xSSC, pH 7, for 36 s, and dehydrated in an ice-cold ethanol series (70%, 85%, and 100%). The images were captured with an Olympus DP71 digital camera coupled to a BX61 Olympus microscope (Olympus Corporation, Tokyo, Japan) and were optimized for brightness and contrast using Adobe Photoshop CS2.

Results

Assembly and gene annotation of the *A. latifasciata* draft genome

A total of seven B⁻ and B⁺ Illumina sequenced samples were analyzed including six individual data sets obtained with the Hiseq and one with the Miseq platform (Table 1). After filtering, a total of 513,441,660 B⁻ reads with 80.5× coverage and 850,418,274 B⁺ reads with 119.9× coverage were recovered for the analysis (Table 1). The de novo assembly of *A. latifasciata* B⁻ reference genome was 771,316,069 bp long. Based on the size of *M. zebra* genome, we estimate that about 84% of the *A. latifasciata* genome was recovered. The B⁻ assembly yielded 218,259 scaffolds with N50 of 18,640 bp, being the longest scaffold with 233,669 bp (Table 2; Table S4). The B⁺ assembly was 781,068,509 bp long, in 197,652 scaffolds with an N50 of 25,546 bp. The longest scaffold was 238,637 bp (Table S5).

The total number of complete and partial core eukaryotic genes (CEGs) recovered in the B⁻ assembly were ~73% and ~94% respectively (Table S6). The annotation pipeline found 24,907 genes in the B⁻ *A. latifasciata* genome assembly (Fig. S1; Table S7), comprising 22.4% of the total assembly. The annotated genes were also screened for the GO level distribution in three categories: biological processes (P), molecular function (F), and cellular component (C). Moreover, the highest number of GO terms was detected for biological processes (Fig. S2), and the most abundant sub-categorized functions are cellular process, biological regulation, and multicellular-organism process. Structural annotations as general feature format (GFF) file have been uploaded to Sacibase.

B chromosome polymorphism

We identified a total of 2,395,658 SNPs and 888,060 indels in the six genomes (B⁻ and B⁺ individuals) when compared to B⁻ assembly (Fig. S3). After filtering, we detected a total of 17,875 high quality SNPs in all genomes (Fig. S4a) and 1B female reads showed the highest number of SNPs (5,181) shared with all other individuals (Fig. S4b). However, in B⁺ individuals, a total 11,978 SNPs were identified relative to the B⁻ reference genome (Fig. S4b). The SNP frequencies of

genic sequences (CDS, exons and introns) in *A. latifasciata* were also screened by comparing coding sequences from B⁺ and B⁻ reads datasets to the *A. latifasciata* transcriptome assembly (Nakajima et al. in preparation), which detected a total of 16,839 SNPs in four samples, two from each B⁺ and B⁻ both male and female (Fig. S4c, S4d). However, male samples, irrespective of B chromosome, had a higher frequency of SNPs in genic sequences (10,508) as compared to female samples (6,331). Comparative analysis of different SNP combinations from B⁺ and B⁻ samples confirmed unique and shared sets of SNP variation. We identified a total of 687 genomic (Fig. S4b) and 167 transcriptomic (Fig. S4d) SNPs shared among all B⁺ samples, suggesting these SNPs are located on a B chromosome. Similarly, the B-specific SNPs were also identified in the B-located genes (see “Sequence analysis and physical mapping of B sequences”).

Whole genome rearrangements and structural variations

The whole genome comparative analysis of B⁺ with B⁻ assemblies using pairwise minimap2 alignments generated a total of 849,600 alignments. The conserved synteny between the two genomes were detected under the significance threshold *E*-value of < 1E⁻⁵⁰, and the alignments results were visualized as dotplots. The expected high proportion of homologous sequences confirmed is apparent as diagonal lines of synteny (Fig. 1a), indicating highly similar conserved contents between the two genomes derived from the same species. However, we also observed breaks in synteny, duplicated blocks including ancient WGD, and several insertions, deletions, and inversions, which signal genome rearrangements. Many duplicated regions were visually observed in the dotplot comparison analysis in B⁺ genome (Fig. 1b). A total of 1,717 duplicated blocks were identified and retrieved from B⁺ genome. The genes annotation of these blocks detected 1,546 protein coding sequences, 8 pseudo-processed, and 3 unknown genes (Dataset S1). Selected alignments with Phred score ≥ 30 (Table S8) were used to determine the number of indels of the whole genome alignments, showing a few amounts of large indels between B⁺ and B⁻ genomes (Fig. 2a). The total of indels comparing both genomes is 21,505,536 bp, which is ~2.78% of B⁻ genome. Moreover, considering the BLAST-like alignment identity calculated from paf file, it is possible to observe that most of aligned sequences have high identity (Fig. S5). Since we want to highlight the divergences between B⁺ and B⁻ genome, we selected the most dissimilar alignments using an identity ≤ 0.5 or ≤ 0.8 to perform the Circos plots (Fig. 2b, c). The Circos plots highlight several genomic blocks rearranged in the B⁺ genome (Fig. 2b, c).

The repeats annotation found the highest number of retrotransposons (a total of 444 elements) including SINES,

Table 2 Comparison of the current assembly of B- *A. latifasciata* to other African cichlids and other fish species assemblies (statistics data sourced from NCBI)

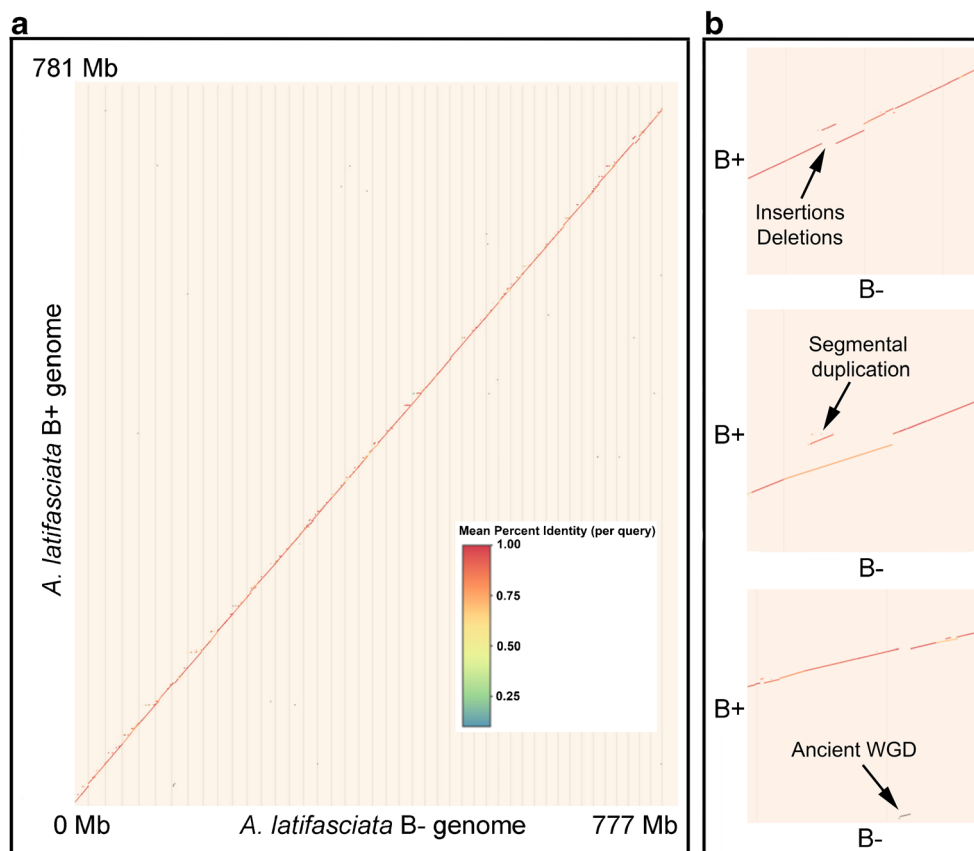
Species	<i>Astatotilapia latifasciata</i>	<i>Astatotilapia burtoni</i>	<i>Pundamilia nyererei</i>	<i>Neolamprologus brichardi</i>	<i>Oreochromis niloticus</i>	<i>Metriaclima zebra</i>	<i>Latimeria chalumnae</i>	<i>Gasterosteus aculeatus</i>
Genome size (Gb)	0.771	0.831	0.830	0.847	0.927	0.957	2.73	0.446
% GC	40.5	40.5	40.6	42.0	39.9	41.1	42.55	44.6
Protein coding gene count	23,391	24,094	22,960	25,018	29,550	25,898	21,021	20,787
Genome coverage	123 ×	131 ×	126 ×	171 ×	269 ×	16.5 ×	299.48 ×	9.0 ×
Sequencing technology	Illumina HiSeq	Illumina HiSeq	Illumina HiSeq	Illumina HiSeq	Illumina	PacBio	Illumina HiSeq	454; Sanger dideoxy sequencing

L1, L2, Gypsy, and BEL/pao followed by DNA transposons (a total of 241 elements) including hobo activator and Tc1 (Fig. 3). The Fisher exact test resulted the GO enrichment of functions related to development, morphogenesis, cell cycle, binding, transport, immune system, and regulation of gene expression (Fig. 4 and Dataset S2).

We applied high-throughput and massive paired-end mapping (PEM) to identify structural variants (SVs) in B+ genomic data over the B- genome. A total of 625 interchromosomal translocations (breakpoints) were

detected in the whole B+ genome (Fig. 1a). We annotated a few of these regions (between 515 and 520 Mb coordinates, Scaffold NODE_552876) with identified translocations; most of them were fragmented genes and non-coding RNAs (Fig. S6 and S7). Genomic regions related to B chromosome (regions showing coverage higher than × 15 in the Illumina datasets) were also subjected to reads-orientation based on SVs detection method. Interestingly, we found duplications, insertions, and inversions at different sites in the B+ blocks (Fig. 5).

Fig. 1 Comparative whole genomics analysis of assembled B- and B+ *A. latifasciata* genomes. **a** Whole genome alignments between B+ and B- genomes assemblies are shown as dotplot depicting the total number of 1664 post filtered aligned blocks. B+ genome and B- genome assemblies are represented as Y- and X-axis respectively. The breaks in the colored diagonal line show the syntenic discontinuity pointing towards genomics rearrangements. The small dots slightly above the diagonal line represent those genomics blocks signaling duplicated regions in B+ genome. **b** Examples of specific genomic rearranged regions between both assemblies confirming segmental duplication, insertions deletions, and ancient WGD event. The diagonal line shows the similar regions between the two genomes with mean percent identity given according to the different colors indicated in the insert in **a**



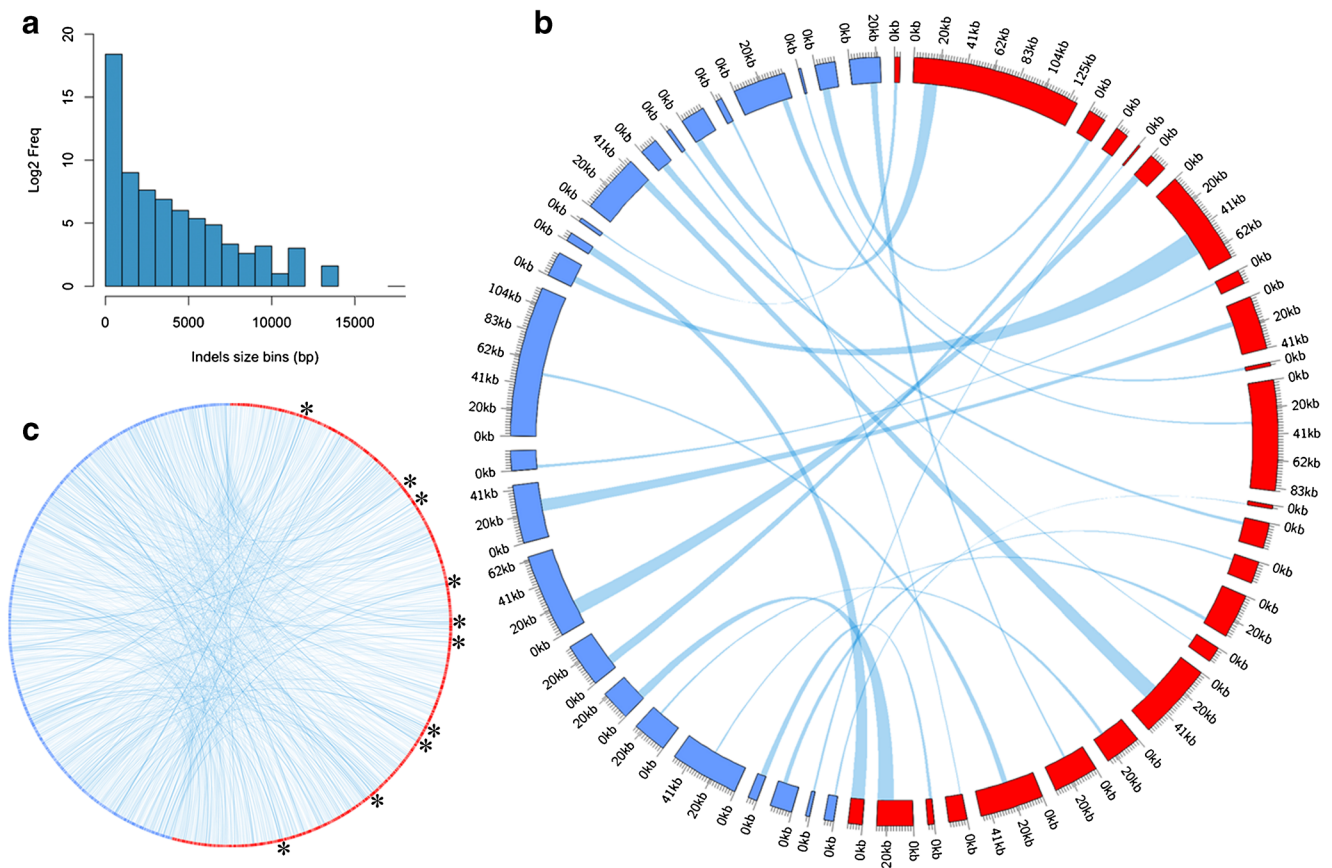


Fig. 2 Alignment analysis between B+ blocks and their counterpart in the B- genome. Graphical distribution (a) of indels size comparing B+ and B- genomes. Circos plots based on cutoff of ≤ 0.5 (b) and cutoff of ≤ 0.8 (c) of most dissimilar alignments comparing B+ and B- genomes. The

outermost rings of b and c represent B+ (red) and B- (blue) sequences. Asterisks (*) in c indicate hotspots with larger segments aligned between both genomes. The blue lines connecting the red and blue outermost rings indicate genomic blocks conserved between B+ and B- genomes

Sequence analysis and physical mapping of B sequences

Among the rearranged and duplicated blocks of *A. latifasciata*, we analyzed those containing genes previously

identified on B chromosomes of diverse species (Table S1). BLASTn results indicate a higher number of hits for *ihhb* and *45S rRNA* genes in the genome of *A. latifasciata*. The remaining genes were not considered in the analysis because of partial or complete absence in the *A. latifasciata* genome. The

Fig. 3 Repeat elements composition of identified duplicated blocks in the B+ genome. The Y- and X-axis of the bar chart indicates the copy-number and type of elements respectively

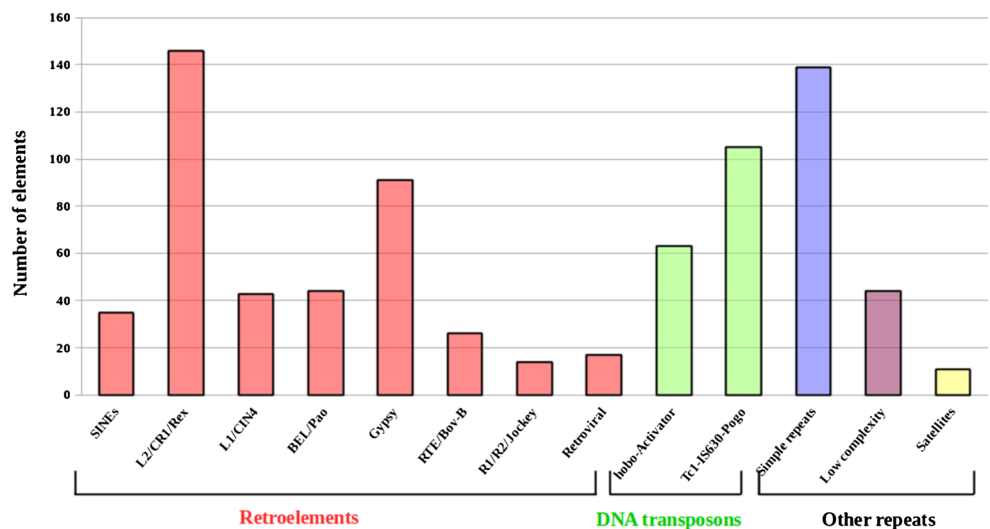
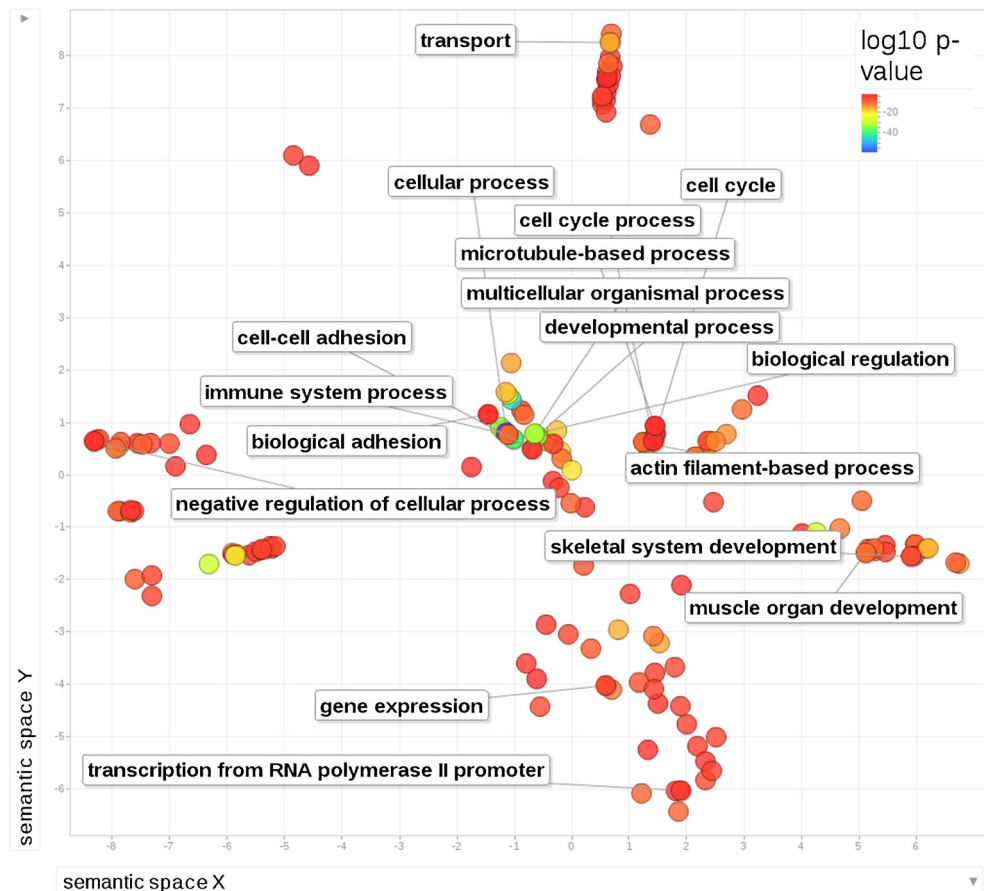


Fig. 4 Enrichment GO terms for the duplicated blocks in the B+ genome. Each circle represents a specific function. Only functions previously identified (Valente et al. 2014; Huang et al. 2016; Navarro-Domínguez et al. 2017) as related to B chromosome are highlighted. The Y- and X-axis have no essential meaning and represent just the graphical space. The bubbles are colored on the basis of $\log_{10} P$ values. Dark red circles indicate less enriched and dark blue indicates more enriched terms



B+/B− alignment shows higher coverage of the B+ than the B− samples for both for *ihhb* and *45S rRNA* genes (Fig. 6; Fig. S8, S9, S10). The higher reads coverage of these genes in B+ genome evidences their duplicated copies in the B chromosome. We also conducted a survey to screen SNPs and indels. The *ihhb* gene has encountered few B-specific SNPs and indels (Fig. 7 and Fig. S9), while a high number of non-B-related SNPs were found in 45S rRNA cluster (Fig. 6 and Fig. S10). The available RNA-seq data was analyzed for both genes to evaluate their transcriptional level and we did not detect any transcripts of *ihhb* and *45S rRNA* genes among B+ and B tissues of *A. latifasciata*. We found only few reads in some tissues but there were no whole transcripts for most of the tissues in both B− and B+.

The FISH mapping revealed extensive hybridization of *ihhb* over the B chromosomes in 2B metaphases and scattered signals over the A chromosomes (Fig. 7). We FISH-mapped each of the 18S, 5.8S, and 28S rRNA transcriptional regions individually to investigate if the complete cluster of 45S rRNA has moved from A complement to the B chromosome (Fig. 7). Positive sites of 18S rRNA gene were observed over the pericentromeric and subtelomeric areas of the B chromosome and on pericentromeric regions and scattered over few A chromosomes. The 5.8S rRNA probe hybridized in the

pericentromeric regions of the B chromosome and in telomeric and subtelomeric regions of autosomes. The 28S rRNA gene probe produced signals on telomeric and centromeric regions of the B chromosome and also in the short arms of few chromosomes. The amplified PCR products of *ihhb* and *45S rRNA* genes used for the constructions of FISH probes were subjected to Sanger sequencing, and the sequences aligned against NCBI database by BLAST. The results found 98–99% identity, with the highest number of hits to *ihhb* and *45S rRNA* genes.

Discussion

The *A. latifasciata* assemblies add a draft genome reference (based on N50, genome size, genes number, and % of GC statistics) in correspondence to other African cichlids (Brawand et al. 2014) and also novelties on the genomic content of B chromosomes. One of the key aspects in the B chromosome studies resides in understanding the genetic polymorphism/variations and their impact on the origin and evolution of B chromosomes. We checked the SNPs rate to evaluate the selective pressure levels in the genic sequences applying a similar approach suggested by Martis et al. (2012). According to

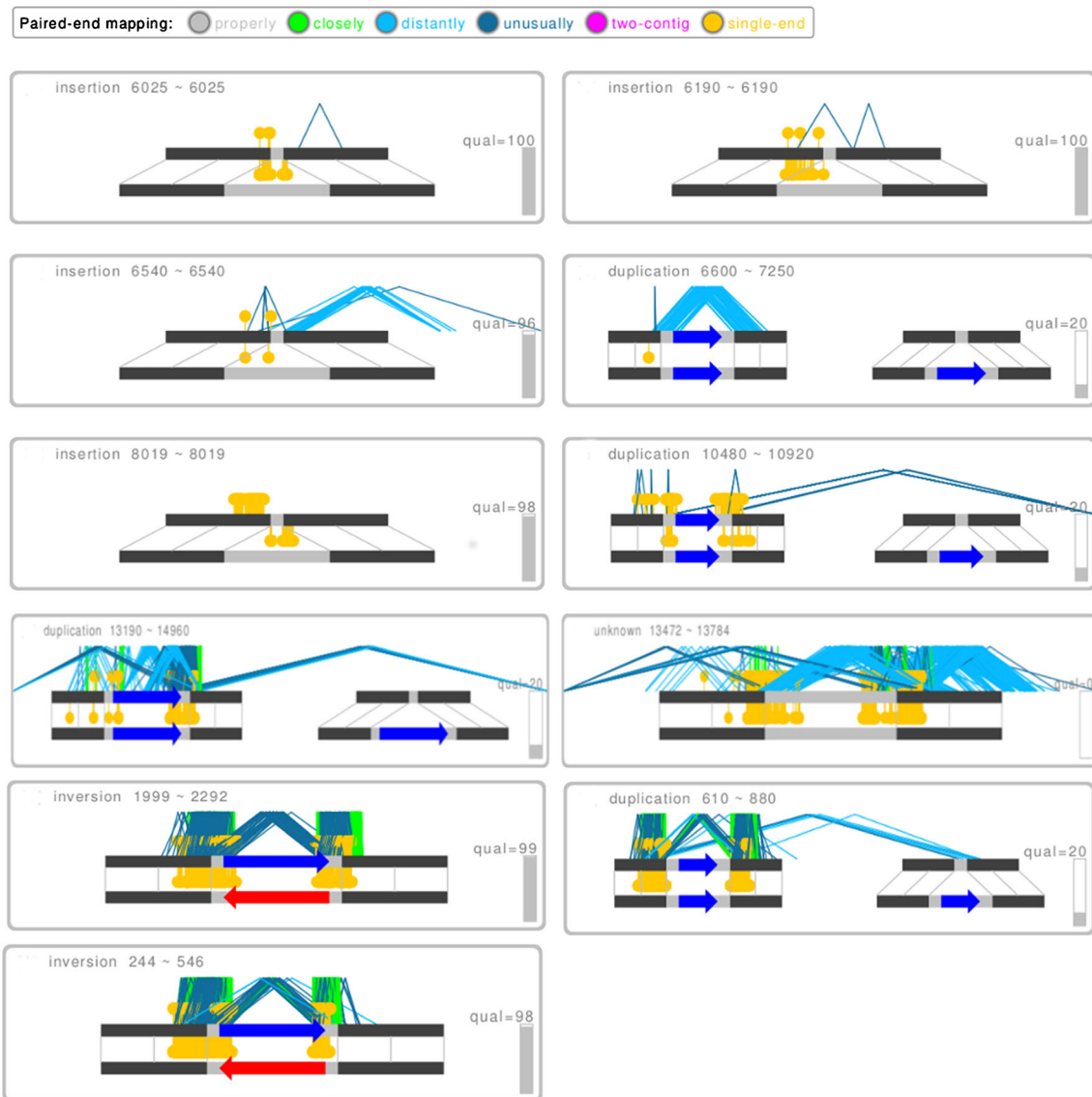


Fig. 5 Structural variations including duplications, deletions, inversions, and transversions in a randomly selected B block (lower line) against the reference genome (upper line). Different color of links represent different orientations of the paired end reads mapped against the reference region indicating specific types of SVs as illustrated below: Gray links, normal

reads; light blue, deletion event; green links, insertions; dark blue, translocations. An inversion causes the paired reads to change the orientation, and both ends will map to the same strand. Segmental tandem duplication is represented by one set of distantly mapped reads and one set of inverted mapped reads

Martis et al. (2012), the SNPs in the B genic sequences could indicate the action of different levels of selective pressure. In this context, our results identified a higher encounter of SNP frequencies in males suggesting they could be under lower selective pressure than female. Furthermore, the comparative analysis of genome and transcriptome datasets recorded shared SNPs among the B+ individuals, evidencing the accumulation of B-specific polymorphisms among B+ genomes.

Different types of SVs including complex rearrangements, duplications, inversions, large deletions, and insertions can be detected using a variety of computational approaches (Keane et al. 2014). A remarkable contribution to understand the evolutionary biology of chromosomes was achieved by revelation

of SVs in diverse organisms (Bickhart and Liu 2014; Keane et al. 2014). The identification of many rearrangements has also been applied to explain the mechanism of sex chromosomes evolution (Rogers 2015). In this sense, we have carried out the whole genome rearrangement analysis to understand the molecular mechanisms of B chromosome evolution. A significant finding of our study was the identification of genomic rearrangements and synteny as a result of comparison between B- and B+ de novo assemblies. Notably, the B+ assembly is 9,752,440 bp larger than B- assembly which may reflect the extra amount of B chromosome genomic contents thus being useful to trace the genomic changes which might have happened due to additional B chromosomal

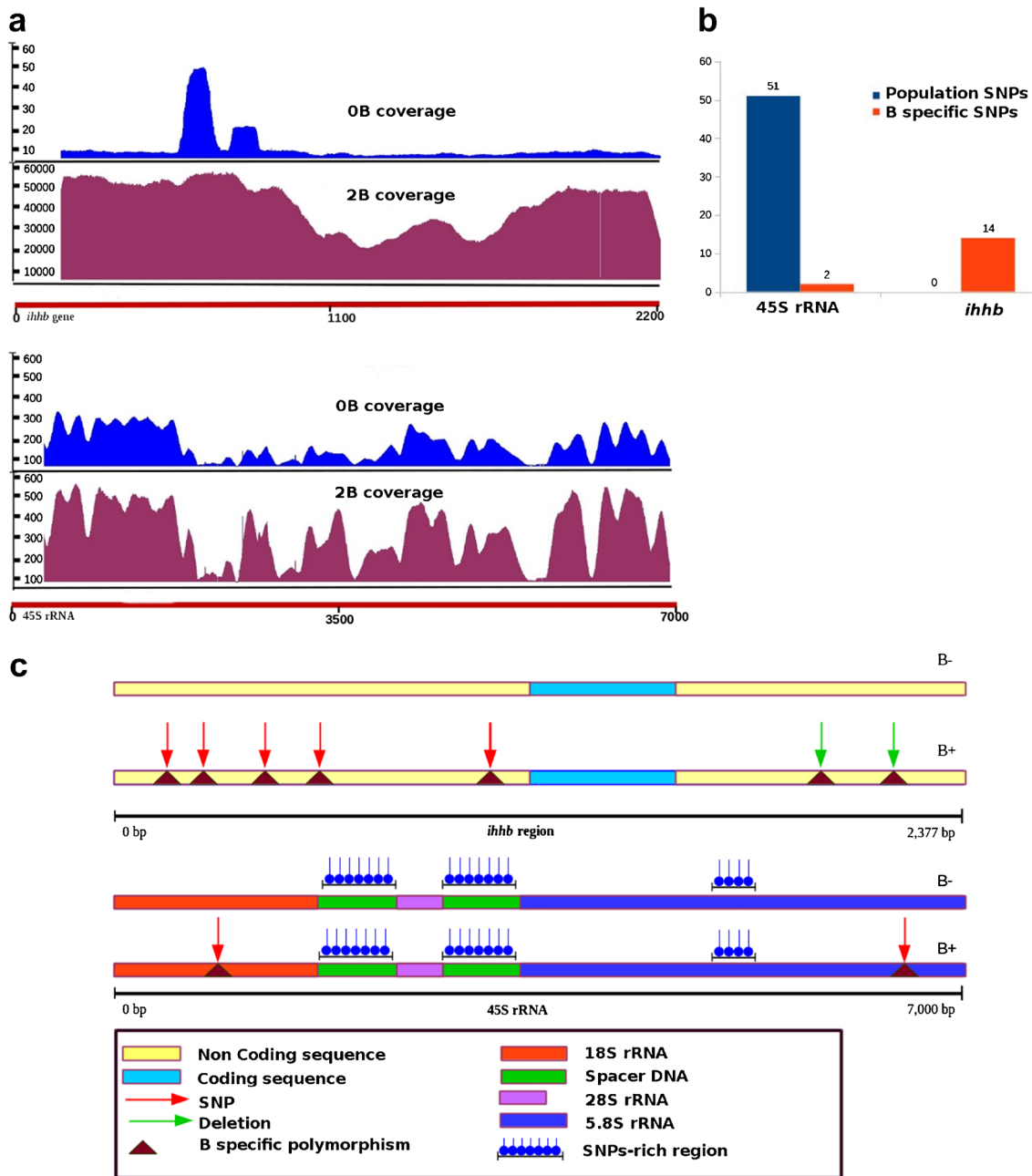


Fig. 6 Sequence analysis of B related genes. **a** Reads coverage of *ihhb* and *45S rRNA* gene sequences. Notice the scale bar on left to differentiate the coverage rate between OB and 2B samples. **b** Number of population and B-specific SNPs are shown as red and blue respectively in the bars for

both genes. **c** Model representation of genes to elaborate their structure and localize the positions of SNPs in regions with comparison of B+ and B- individuals

contents. A large B+ assembly could also be a consequence of the differences in the B+ and B- number of sequence reads used for assemblies. Although the comparative genomics approach using whole genome alignment indicated that both genomes mostly share a similar pattern of mapped blocks, as expected, still there are some discontinuities (chromosomal breaks) between the two genomes. We observed that despite both genomes belonging to the same species, we have encountered a variety of genomic changes. These changes are related

to B chromosome presence and point towards significant events such as duplications, deletions, and insertions occurring during the evolution of B chromosome. Remarkably, we also found a few weak syntenies that could reflect relics of the teleost ancient WGD, a significant process of teleost evolution (Santini et al. 2009; Glasauer and Neuhaus 2014). Vertebrates have been characterized by two rounds of WGD occurring in their early evolutionary history, followed by a third round of duplication exclusive of the teleost fish lineage.

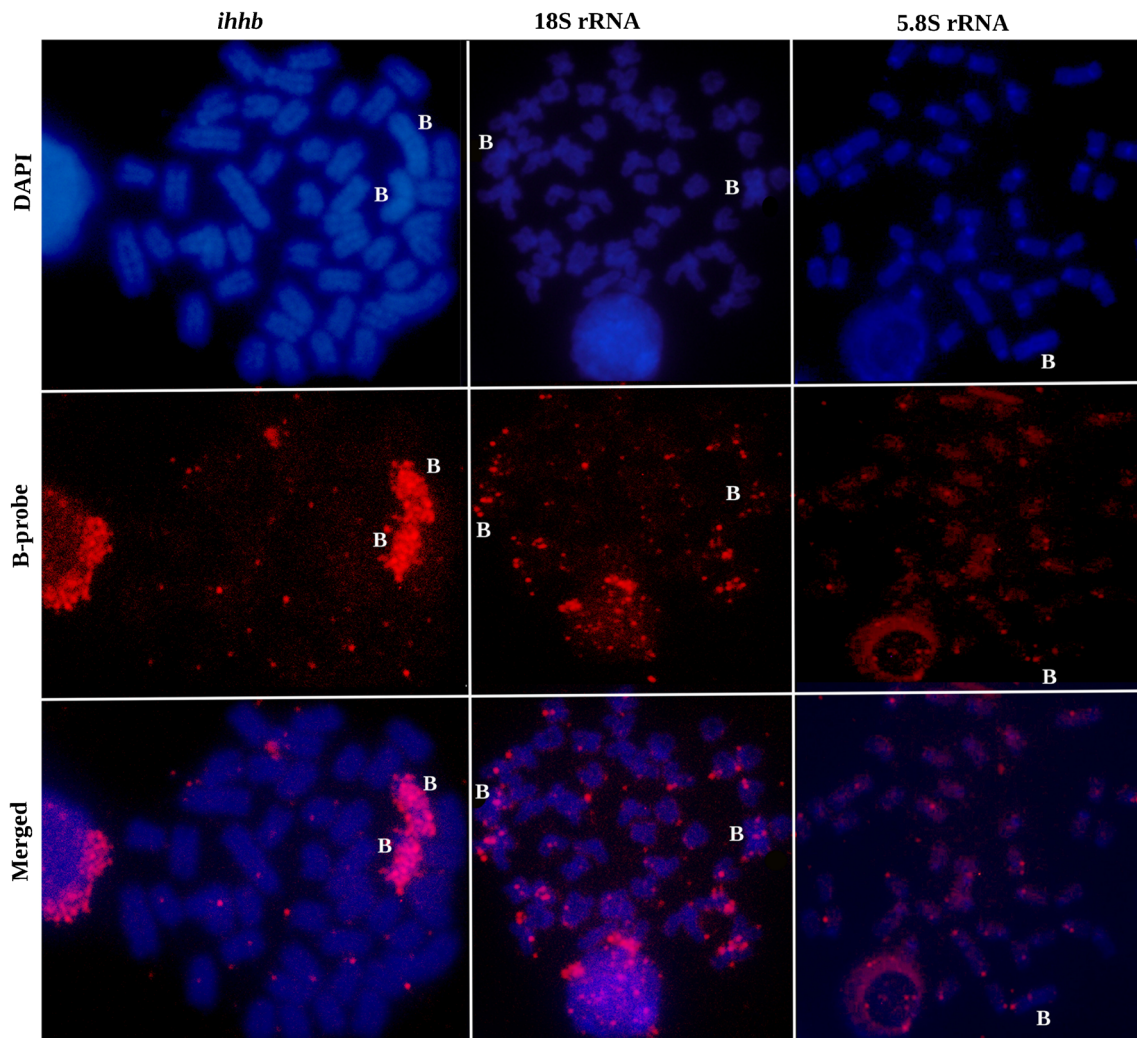


Fig. 7 FISH mapping of *ihhb* and *45S rRNA* (5.8S rRNA, 18 rRNA, and 28S rRNA) genes. Metaphases stained with DAPI, B-probes, and merged are shown for each sequence. The *ihhb* FISH mapping was conducted in 2B metaphases. The B chromosomes are indicated

Apart from this analysis, we also discovered a large number of chromosomal breakpoints such as translocations extensively distributed throughout the whole B+ genome of *A. latifasciata*. Several sequence duplications detected in the B blocks indicated that these sequences originated from A chromosomes and accumulated on B chromosome due to frequent duplication events. The findings of our study support the previous observation (Valente et al. 2014) that most autosomes have contributed sequences to the B chromosome of *A. latifasciata* through gene duplication, as has been documented for the origin of supernumerary elements of diverse species (Teruel et al. 2010; Martis et al. 2012; Utsunomia et al. 2016). The discovery of a large set of SVs in B+ genome provoked the hypothesis that the B chromosome presence could offer additional genetic material to evolution and, thus, directly contributes to evolutionary processes in the populations. To trace these duplication events, the duplicated blocks analysis conducted in B+ genome enabled us to reveal different set of genes and repeats. The annotated repeats of the

duplicated blocks enlisted all the retrotransposons (Gypsy, Bel/Pao/ and L2) that have recently been discovered on the B chromosome of *A. latifasciata* by combination of bioinformatics, qPCR and FISH mapping (Coan and Martins 2018). In addition, the higher amount of retrotransposons and DNA transposons in these B chromosome blocks suggests that these elements were major players in the duplication of B-located genes. B chromosomes are rich in several classes of repetitive DNAs, including mobile elements and derived sequences from rDNA, satellite DNA, histone genes, and small nuclear RNA genes (Friebe et al. 1995; Teruel et al. 2010; Bueno et al. 2013; Ruiz-Estévez et al. 2014; Silva et al. 2014; Marques et al. 2018). Furthermore, cytogenomics studies have also identified repetitive DNAs derived from regular genes and organelle sequences expanded over the B chromosome (Martis et al. 2012; Valente et al. 2014).

Among all genes described in B chromosomes of vertebrates (Table S1), two of them were identified in the duplicated B chromosome blocks: *ihhb* and *45S rRNA* genes. Our

study focused on uncovering B-specific SNPs and indels of both *ihhb* and *45S rRNA* genes since they were found enriched in the B chromosome. *Ihhb* was also detected in the genome of the cichlid *Lithochromis rubripinnis* (Yoshida et al. 2011), with more than 40 copies of *ihhb* paralogs on B chromosomes and a single copy of *ihhb* ortholog on the A chromosomes. Our results suggest that the *ihhb* gene remains highly conserved among different vertebrate species as no population-wise polymorphism was detected. The higher number of B-specific SNPs and indels found in the *ihhb* gene, and absence of transcripts, reveal such copies on Bs have become pseudogenes. Several papers have emphasized the evolutionary importance of the hedgehog gene family and outlined the process of duplication events related to the members of this gene family including *ihhb* in many vertebrates (Holland 1992; Carroll 1995; Ekker et al. 1995; Kumar et al. 1996; Zardoya et al. 1996). We suggest that the observed increase level of polymorphism in B-located copies of *ihhb* gene is an interesting phenomenon to elucidate the mechanism of gene duplication and neofunctionalization and to understand the molecular evolution of B chromosome. During duplication, modifications such as insertion, deletions, and mutation might have occurred in the gene sequence, as in the case of *ihhb*; therefore, we term *ihhb* as a “non-processed or duplicated pseudogene.” More importantly, the *ihhb* gene is involved in several functions related to vertebrate morphogenesis and regulates the PTCH2 genes which have been reported to express in testis tissues (Carpenter et al. 1998) thus being involved in sexual development. Although the abundant *ihhb* copies in the B chromosome of *A. latifasciata* seems to be inactive, we can not rule out that *ihhb* expansion in the B chromosome of other cichlids could have found any function. The recent studies on B chromosomes of cichlids have outlined their role in the sex-related functions mainly sex determination (Yoshida et al. 2011; Clark et al. 2017). Our genomic analyses of this gene in *A. latifasciata* followed by FISH mapping document the novelty of its organization and expansion on the B chromosome, and raise questions about its role in the B evolution and function.

Based on previous descriptions of ribosomal genes on the B chromosome of *A. latifasciata* (Poletto et al. 2010) there rises a hypothesis that the complete 45S ribosomal DNA (rDNA) cluster may have started amplifying from the A complement and moved to B chromosomes during initial stages of evolution of the proto-B chromosome. From the sequencing data analysis, we found that some regions of the cluster indeed showed a higher coverage in B+ samples as compared to B–; however, no distinct differences were found in overall coverage of the 45S cluster. Our FISH mapping results of 18S rRNA and 28S rRNA confirmed that extra rRNA gene copies have accumulated on the B chromosome. The weaker chromosomal signal of the 5.8S rRNA segments seems to be related to its small DNA size compared to the 18S and 28S

rRNA transcribing segments. The 5.8S, 18S, and 28S rRNA marks on B chromosome enabled our hypothesis about the duplication and expansion of the whole rRNA cluster from A to B chromosome. We did not find transcripts of *45S rRNA* gene cistron in the *A. latifasciata* genome. Previous analysis of nucleolus organizer regions (NORs) activity did not detect transcriptional evidences of B chromosomes rRNA copies (Poletto et al. 2010).

In addition to the presence of repeated DNAs, recent studies have added an extensive list of genes or fragments of genes present in B chromosomes (Ahmad and Martins 2019; Houben et al. 2019). The identification of a few genes on B chromosomes has been achieved in the last three decades through classical genetics studies (Dherawattana and Sadanaga 1973; Miao et al. 1991a, b; Graphodatsky et al. 2005). Bioinformatics and genomics tools developed in the last decade have revealed a higher number of B-located genes and functional sequences and started a new debate about the evolutionary role of B chromosomes, their complex interactions with the host genome, and their possible effects ranging from sex determination to development and adaptation. These studies have been extensively conducted in diverse organisms as fungus (Coleman et al. 2009; Goodwin et al. 2011; Bertazzoni et al. 2018), plants (Banaei-Moghaddam et al. 2013; Ma et al. 2017), insects (Akbari et al. 2013; Navarro-Domínguez et al. 2017, 2019) and vertebrates (Trifonov et al. 2013; Valente et al. 2014; Clark et al. 2018; Makunin et al. 2018) and have highlighted a long list of genes present in the B chromosomes. These analyses identified several high integral putative B genes related to functions/structures such as pathogenicity (exclusive for fungus), cell cycle, chromosome organization, cytoskeleton, development, and neural system. There is a strong correspondence in the identification of cell cycle genes in the B chromosomes reported by several of these studies. The gene annotation of *A. latifasciata* B+ blocks showed significant GO enrichment of functions related to diverse biological process, including cell cycle. Based in the large scale accumulated data for *A. latifasciata* (Valente et al. 2014, 2017; present study) it seems plausible that B chromosomes can modulate the cell physiology in a very complex way, including the control of cell-cycle regulatory mechanisms of the B drive. The B is in a constant co-evolutionary battle with the A genome and a drive seems to be the first step to the B stability and survival. Furthermore, we can not rule out that the B chromosome offers an independent genome compartment for genome adaptation and innovation.

Conclusion

Our analysis brings contributions including (1) generation of a genome draft for *A. latifasciata* useful in future analysis involving evolutionary and applied genomics; (2) screening the

high coverage sequencing data of different individuals for polymorphism, gene diversity, and B-specific structural variations, the nucleotide polymorphism identified in B sequences are useful to track evolutionary history and, also, functional aspects of Bs; (3) discovery of B chromosome linked genes/sequences; (4) duplication events generated higher level of structural variations associated with B chromosome and a higher number of copies of gene/sequence variants in the B chromosome; and (5) our data provoke the hypothesis that supernumerary chromosome presence adds new evolutionary genomic components to the cells and organisms.

Assembly and data files

Genomic and transcriptomic datasets are available at Sacibase Database (sacibase.ibb.unesp.br). Sequencing data of PCR products of *28S rRNA*, *18S rRNA*, *5.8S rRNA*, and *ihhb* gene sequences are available in NCBI database Genbank accession numbers MK182936, MK182937, MK185008, and grp6845354.

Acknowledgments We thank Thomas D Kocher for the critical review of the manuscript.

Funding information This work was financially supported through grants from the São Paulo Research Foundation (FAPESP) (2011/03807-7; 2013/04533-3; 2014/17683-6; 2015/16661-1) and the National Counsel of Technological and Scientific Development (CNPq) (474684/2013-0; 305321/2015-3).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving animals were in accordance with the ethical standards of the institution or practice at which the studies were conducted.

References

- Ahmad SF, Martins C (2019) The modern view of B chromosomes under the impact of high scale omics analyses. *Cells* 8:156. <https://doi.org/10.3390/cells8020156>
- Akbari OS, Antoshechkin I, Hay BA, Ferree PM (2013) Transcriptome profiling of *Nasonia vitripennis* testis reveals novel transcripts expressed from the selfish B chromosome, paternal sex ratio. *G3* 3:1597–1605. <https://doi.org/10.1534/g3.113.007583>
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410
- Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Banaei-Moghaddam AM, Meier K, Karimi-Ashtiyani R, Houben A (2013) Formation and expression of pseudogenes on the B chromosome of rye. *Plant Cell* 25:2536–2544. <https://doi.org/10.1105/tpc.113.111856>
- Banaei-Moghaddam AM, Martis MM, Macas J, Gundlach H, Himmelbach A, Altschmid L, Mayer KF, Houben A (2015) Genes on B chromosomes: old questions revisited with new tools. *Biochim Biophys Acta* 1849:64–70. <https://doi.org/10.1016/j.bbaggm.2014.11.007>
- Bertazzoni S, Williams AH, Jones DA, Syme RA, Tan KC, Hane JK (2018) Accessories make the outfit: accessory chromosomes and other dispensable DNA regions in plant pathogenic fungi. *Mol Plant-Microbe Interact* 318:779–788. <https://doi.org/10.1094/MPMI-06-17-0135-FI>
- Bickhart DM, Liu GE (2014) The challenges and importance of structural variation detection in livestock. *Front Genet* 5:37. <https://doi.org/10.3389/fgene.2014.00037>
- Boetzer M, Pirovano W (2012) Toward almost closed genomes with GapFiller. *Genome Biol* 13:R5. <https://doi.org/10.1186/gb-2012-13-6-r56>
- Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou WC, Corbeil J, del Fabbro C, Docking TR, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapathy G, Gibbs RA, Gnerre S, Godzaridis É, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB, Ho IY, Howard J, Hunt M, Jackman SD, Jaffe DB, Jarvis ED, Jiang H, Kazakov S, Kersey PJ, Kitzman JO, Knight JR, Koren S, Lam TW, Lavenier D, Laviolette F, Li Y, Li Z, Liu B, Liu Y, Luo R, MacCallum I, MacManes MD, Mailliet N, Melnikov S, Naquin D, Ning Z, Otto TD, Paten B, Paulo OS, Phillippy AM, Pina-Martins F, Place M, Przybylski D, Qin X, Qu C, Ribeiro FJ, Richards S, Rokhsar DS, Ruby JG, Scalabrin S, Schatz MC, Schwartz DC, Sergushichev A, Sharpe T, Shaw TI, Shendure J, Shi Y, Simpson JT, Song H, Tsarev F, Vezzi F, Vicedomini R, Vieira BM, Wang J, Worley KC, Yin S, Yiu SM, Yuan J, Zhang G, Zhang H, Zhou S, Korf IF (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* 2:2047–217X. <https://doi.org/10.1186/2047-217X-2-10>
- Brawand D, Wagner CE, Meyer A et al (2014) The genomic substrate for adaptive radiation in African cichlid fish. *Nature* 513:375–381. <https://doi.org/10.1038/nature13726>
- Bueno D, Palacios-Gimenez OM, Cabral-de-Mello DC (2013) Chromosomal mapping of repetitive DNAs in the grasshopper *Abracris flavolineata* reveal possible ancestry of the B chromosome and H3 histone spreading. *PLoS One* 8:e66532. <https://doi.org/10.1371/journal.pone.0066532>
- Cabral-De-Mello DC, Valente GT, Nakajima RT, Martins C (2012) Genomic organization and comparative chromosome mapping of the U1 snRNA gene in cichlid fish, with an emphasis in *Oreochromis niloticus*. *Chromosom Res* 20:279–292. <https://doi.org/10.1007/s10577-011-9271-y>
- Camacho JPM (2005) B chromosomes. In: Gregory T (ed) *The evolution of the genome*, 1st edn. Elsevier, San Diego, pp 223–286
- Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 18:188–196. <https://doi.org/10.1101/gr.6743907>
- Carmello BO, Coan RLB, Cardoso AL, Ramos E, Fantinatti BEA, Marques D, Oliveira RA, Valente GT, Martins C (2017) The hnRNP Q-like gene is retroinserted into the B chromosomes of the cichlid fish *Astatotilapia latifasciata*. *Chromosom Res* 25:277–290. <https://doi.org/10.1007/s10577-017-9561-0>
- Carpenter D, Stone DM, Brush J, Ryan A, Armanini M, Frantz G, Rosenthal A, de Sauvage FJ (1998) Characterization of two patched receptors for the vertebrate hedgehog protein family. *Proc Natl Acad*

- Sci U S A 95:13630–13634. <https://doi.org/10.1073/pnas.95.23.13630>
- Carroll SB (1995) Homeotic genes and the evolution of arthropods and chordates. *Nature* 376:479–485. <https://doi.org/10.1038/376479a0>
- Cheong WH, Tan YC, Yap SJ, Ng KP (2015) ClicO FS: an interactive web-based service of Circos. *Bioinformatics* 31:3685–3687. <https://doi.org/10.1093/bioinformatics/btv433>
- Chikhi R, Medvedev P (2014) Informed and automated k-mer size selection for genome assembly. *Bioinformatics* 30:31–37. <https://doi.org/10.1093/bioinformatics/btt310>
- Clark FE, Conte MA, Ferreira-Bravo IA, Poletto AB, Martins C, Kocher TD (2017) Dynamic sequence evolution of a sex-associated b chromosome in Lake Malawi cichlid fish. *J Hered* 108:53–62. <https://doi.org/10.1093/jhered/esw059>
- Clark FE, Conte MA, Kocher TD (2018) Genomic characterization of a B chromosome in Lake Malawi cichlid fishes. *Genes* 9:610. <https://doi.org/10.3390/genes91020610>
- Coan RLB, Martins C (2018) Landscape of transposable elements focusing on the B chromosome of the cichlid fish *Astatotilapia latifasciata*. *Genes* 9:269. <https://doi.org/10.3390/genes9060269>
- Coleman JJ, Rounsley SD, Rodriguez-Carres M, Kuo A, Wasmann CC, Grimwood J, Schmutz J, Taga M, White GJ, Zhou S, Schwartz DC, Freitag M, Ma LJ, Danchin EGJ, Henrissat B, Coutinho PM, Nelson DR, Straney D, Napoli CA, Barker BM, Gribskov M, Rep M, Kroken S, Molnár I, Rensing C, Kennell JC, Zamora J, Farman ML, Selker EU, Salamov A, Shapiro H, Pangilinan J, Lindquist E, Lamers C, Grigoriev IV, Geiser DM, Covert SF, Temporini E, VanEtten HD (2009) The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. *PLoS Genet* 5:e1000618. <https://doi.org/10.1371/journal.pgen.1000618>
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genome Project Analysis Group (2011) The variant call format and VCFtools. *Bioinformatics* 27:2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- Dherawattana A, Sadanaga K (1973) Cytogenetics of a crownrust-resistant hexaploid oat with 42+2 fragment chromosomes. *Crop Sci* 13:591–594. <https://doi.org/10.2135/cropsci1973.0011183X001300060002x>
- Drummond AJ, Ashton B, Buxton S, et al (2009) Geneious v4.8.5. Available online at: <http://www.geneious.com>. Accessed 5 Jan 2018
- Ekker SC, Ungar AR, Greenstein P, von Kessler DP, Porter JA, Moon RT, Beachy PA (1995) Patterning activities of vertebrate hedgehog proteins in the developing eye and brain. *Curr Biol* 5:944–955. [https://doi.org/10.1016/S0960-9822\(95\)00185-0](https://doi.org/10.1016/S0960-9822(95)00185-0)
- Fantinatti BEA, Martins C (2016) Development of chromosomal markers based on next-generation sequencing: the B chromosome of the cichlid fish *Astatotilapia latifasciata* as a model. *BMC Genet* 17:119. <https://doi.org/10.1186/s12863-016-0427-9>
- Fantinatti BEA, Mazzuchelli J, Valente GT, Cabral-De-Mello DC, Martins C (2011) Genomic content and new insights on the origin of the B chromosome of the cichlid fish *Astatotilapia latifasciata*. *Genetica* 139:1273–1282. <https://doi.org/10.1007/s10709-012-9629-x>
- Friebe B, Jiang J, Gill B (1995) Detection of 5S-rDNA and other repeated DNA on supernumerary B-chromosomes of *Triticum* species (Poaceae). *Plant Syst Evol* 196:131–139. <https://doi.org/10.1007/BF00982954>
- Garrison E (2012) Vcflib: a C++ library for parsing and manipulating VCF files. GitHub. <https://github.com/vcflib/vcflib>. Accessed 19 Aug 2017
- Glasauer SMK, Neuhauss SCF (2014) Whole-genome duplication in teleost fishes and its evolutionary consequences. *Mol Gen Genomics* 289:1045–1060. <https://doi.org/10.1007/s00438-014-0889-2>
- Goodwin SB, Ben M'Barek S, Dhillon B, Wittenberg AHJ, Crane CF, Hane JK, Foster AJ, van der Lee TAJ, Grimwood J, Aerts A, Antoniw J, Bailey A, Bluhm B, Bowler J, Bristow J, van der Burgt A, Canto-Canché B, Churchill ACL, Conde-Ferràez L, Cools HJ, Coutinho PM, Csukai M, Dehal P, de Wit P, Donzelli B, van de Geest HC, van Ham RCHJ, Hammond-Kosack KE, Henrissat B, Kilian A, Kobayashi AK, Koopmann E, Kourmpetis Y, Kuzniar A, Lindquist E, Lombard V, Maliepaard C, Martins N, Mehrabi R, Nap JPH, Ponomarenko A, Rudd JJ, Salamov A, Schmutz J, Schouten HJ, Shapiro H, Stergiopoulos I, Torriani SFF, Tu H, de Vries RP, Waalwijk C, Ware SB, Wiebenga A, Zwiers LH, Oliver RP, Grigoriev IV, Kema GHJ (2011) Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensable structure, chromosome plasticity, and stealth pathogenesis. *PLoS Genet* 7:e1002070. <https://doi.org/10.1371/journal.pgen.1002070>
- Gordon A, Hannon GJ (2010) FASTX-Toolkit. FASTQ/A short-reads pre-processing tools. http://hannonlab.cshl.edu/fastx_toolkit/
- Götz S, Garcia-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talóon M, Dopazo J, Conesa A (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* 36:3420–3435. <https://doi.org/10.1093/nar/gkn176>
- Graphodatsky AS, Kukekova AV, Yudkin DV, Trifonov VA, Vorobieva NV, Beklemisheva VR, Perelman PL, Graphodatskaya DA, Trut LN, Yang F, Ferguson-Smith MA, Acland GM, Aguirre GD (2005) The proto-oncogene C-KIT maps to canid B-chromosomes. *Chromosom Res* 13:113–122. <https://doi.org/10.1007/s10577-005-7474-9>
- Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUASt: quality assessment tool for genome assemblies. *Bioinformatics* 29:1072–1075. <https://doi.org/10.1093/bioinformatics/btt086>
- Holland P (1992) Homeobox genes in vertebrate evolution. *Bioessays* 14:267–273. <https://doi.org/10.1002/bies.950140412>
- Houben A (2017) B chromosomes – a matter of chromosome drive. *Front Plant Sci* 8:210. <https://doi.org/10.3389/fpls.2017.00210>
- Houben A, Banaei-Moghaddam AM, Klemme S, Timmis JN (2014) Evolution and biology of supernumerary B chromosomes. *Cell Mol Life Sci* 71:467–478. <https://doi.org/10.1007/s00018-013-1437-7>
- Houben A, Jones N, Martins C, Trifonov F (2019) Evolution, composition and regulation of supernumerary B chromosomes. *Genes* 10:161. <https://doi.org/10.3390/genes10020161>
- Huang W, Du Y, Zhao X, Jin W (2016) B chromosome contains active genes and impacts the transcription of A chromosomes in maize (*Zea mays* L.). *BMC Plant Bio* 16:88. <https://doi.org/10.1186/s12870-016-0775-7>
- Jones RN (2018) Transmission and drive involving parasitic B chromosomes. *Genes* 9:388. <https://doi.org/10.3390/genes9080388>
- Keane TM, Wong K, Adams DJ, Flint J, Reymond A, Yalcin B (2014) Identification of structural variation in mouse genomes. *Front Genet* 5:192. <https://doi.org/10.3389/fgene.2014.00192>
- Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19:1639–1645. <https://doi.org/10.1101/gr.092759.109>
- Kumar S, Balczarek KA, Lai ZC (1996) Evolution of the hedgehog gene family. *Genetics* 142:965–972. <https://doi.org/10.1111/j.1742-481X.2011.00861.x>
- Langmead B, Salzberg SL (2013) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359. <https://doi.org/10.1038/nmeth.1923>
- Li H (2018) Minimap2: versatile pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>

- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Ma W, Gabriel TS, Martis MM, Gursinsky T, Schubert V, Vrána J, Doležel J, Grundlach H, Altschmied L, Scholz U, Himmelbach A, Behrens SE, Banaei-Moghaddam AM, Houben A (2017) Rye B chromosomes encode a functional Argonaute-like protein with in vitro slicer activities similar to its A chromosome paralog. *New Phytol* 213:916–928. <https://doi.org/10.1111/nph.14110>
- Makunin AI, Demytyeva PW, Graphodatsky AS, Volobouev VT, Kukekova AV, Trifonov VA (2014) Genes on B chromosomes of vertebrates. *Mol Cytogenet* 7:99. <https://doi.org/10.1186/s13039-014-0099-y>
- Makunin A, Romanenko S, Beklemisheva V, Perelman P, Druzhkova A, Petrova K, Prokopov D, Chernyaeva E, Johnson J, Kukekova A, Yang F, Ferguson-Smith MA, Graphodatsky AS, Trifonov VA (2018) Sequencing of supernumerary chromosomes of red fox and raccoon dog confirms a non-random gene acquisition by B chromosomes. *Genes* 9:405. <https://doi.org/10.3390/genes9080405>
- Marques A, Klemme S, Houben A (2018) Evolution of plant B chromosome enriched sequences. *Genes* 9:515. <https://doi.org/10.3390/genes9100515>
- Martis MM, Klemme S, Banaei-Moghaddam AM, Blattner FR, Macas J, Schmutzer T, Scholz U, Gundlach H, Wicker T, Simkova H, Novak P, Neumann P, Kubalaková M, Bauer E, Haseneyer G, Fuchs J, Doležel J, Stein N, Mayer KFX, Houben A (2012) Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proc Natl Acad Sci U S A* 109:13343–13346. <https://doi.org/10.1073/pnas.1204237109>
- Miao VP, Covert SF, Vanetten HD (1991a) A fungal gene for antibiotic resistance on a dispensable (B) chromosome. *Science* 254:1773–1776
- Miao VP, Matthews DE, Vanetten HD (1991b) Identification and chromosomal location of a family of cytochrome P-450 genes or pisatin detoxification in the fungus *Nectria haematococca*. *Mol Gen Genet* 226:214–223. <https://doi.org/10.1007/BF00273606>
- Navarro-Domínguez B, Ruiz-Ruano FJ, Cabrero J, Corral JM, López-León MD, Sharbel TF, Camacho JPM (2017) Protein-coding genes in B chromosomes of the grasshopper *Eyprepocnemis plorans*. *Sci Rep* 7(45200). <https://doi.org/10.1038/srep45200>
- Navarro-Domínguez B, Martín-Peciña M, Ruiz-Ruano FJ, Cabrero J, Corral JM, López-León MD, Sharbel TF, Camacho JPM (2019) Gene expression changes elicited by a parasitic B chromosome in the grasshopper *Eyprepocnemis plorans* are consistent with its phenotypic effects. *Chromosoma*. <https://doi.org/10.1007/s00412-018-00689-y>
- Nüsslein-Volhard C, Wieschaus E (1980) Mutations affecting segment number and polarity in *Drosophila*. *Nature* 287:795–801
- O'Quin CT, Drilea AC, Conte MA, Kocher TD (2013) Mapping of pigmentation QTL on an anchored genome assembly of the cichlid fish, *Metriaclima zebra*. *BMC Genomics* 14:287. <https://doi.org/10.1186/1471-2164-14-287>
- Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23:1061–1067. <https://doi.org/10.1093/bioinformatics/btm071>
- Pereira J, Johnson WE, O'Brien SJ, Jarvis ED, Zhang G, Gilbert MTP (2014) Evolutionary genomics and adaptive evolution of the hedgehog gene family (Shh, Ihh and Dhh) in vertebrates. *PLoS One* 9:e74132. <https://doi.org/10.1371/journal.pone.0074132>
- Pinkel D, Landegent J, Collins C, Fuscoe J, Segraves R, Lucas J, Gray J (1988) Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proc Natl Acad Sci U S A* 85:9138–9142. <https://doi.org/10.1073/pnas.85.23.9138>
- Poletto AB, Ferreira IA, Cabral-de-Mello DC, Nakajima RT, Mazzuchelli J, Ribeiro HB, Venere PC, Nirchio M, Kocher TD, Martins C (2010) Chromosome differentiation patterns during cichlid fish evolution. *BMC Genet* 13(2):50. <https://doi.org/10.1186/1471-2156-11-50>
- Qi J, Zhao F (2011) InGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. *Nucleic Acids Res* 39:W567–W575. <https://doi.org/10.1093/nar/gkr506>
- Ramírez F, Dünder F, Diehl S, Grüning BA, Manke T (2014) DeepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* 42:W187–W191. <https://doi.org/10.1093/nar/gku365>
- Ramos E, Cardoso AL, Brown J, Marques DF, Fantinatti BE, Cabral-de-Mello DC, Oliveira RA, O'Neill RJ, Martins C (2017) The repetitive DNA element BncDNA, enriched in the B chromosome of the cichlid fish *Astatotilapia latifasciata*, transcribes a potentially noncoding RNA. *Chromosoma* 126:313–323. <https://doi.org/10.1007/s00412-016-0601-x>
- Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28:i333–i339. <https://doi.org/10.1093/bioinformatics/bts378>
- Rogers RL (2015) Chromosomal rearrangements as barriers to genetic homogenization between archaic and modern humans. *Mol Biol Evol* 32:3064–3078. <https://doi.org/10.1093/molbev/msv204>
- Ruban A, Schmutzer T, Scholz U, Houben A (2017) How next-generation sequencing has aided our understanding of the sequence composition and origin of B chromosomes. *Genes* 8:294. <https://doi.org/10.3390/genes8110294>
- Ruiz-Estévez M, Badisco L, Broeck JV, Perfectti F, López-León MD, Cabrero J, Camacho JP (2014) B chromosomes showing active ribosomal RNA genes contribute insignificant amounts of rRNA in the grasshopper *Eyprepocnemis plorans*. *Mol Gen Genomics* 289:1209–1216. <https://doi.org/10.1007/s00438-014-0880-y>
- Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74:5463–5467. <https://doi.org/10.1073/pnas.74.12.5463>
- Santini F, Harmon LJ, Carnevale G, Alfaro ME (2009) Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. *BMC Evol Biol* 9:194. <https://doi.org/10.1186/1471-2148-9-194>
- Silva DM, Pansonato-Alves JC, Utsunomia R, Araya-Jaime C, Ruiz-Ruano FJ, Daniel SN, Hashimoto DT, Oliveira C, Camacho JP, Porto-Foresti F, Foresti F (2014) Delimiting the origin of a B chromosome by FISH mapping, chromosome painting and DNA sequence analysis in *Astyanax paranae* (Teleostei, Characiformes). *PLoS One* 9:e94896. <https://doi.org/10.1371/journal.pone.0094896>
- Smit AFA, Hubble R, Green P (2013–2015) RepeatMasker Open-4.0. <http://www.repeatmasker.org>
- Stanke M, Steinkamp R, Waack S, Morgenstern B (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res* 32:W309–W312. <https://doi.org/10.1093/nar/gkh379>
- Stothard P (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* 28:1102–1104. <https://doi.org/10.2144/00286i01>
- Supek F, Bošnjak M, Škunca N, Šmuc T (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6:e21800. <https://doi.org/10.1371/journal.pone.0021800>
- Teruel M, Cabrero J, Perfectti F, Camacho JP (2010) B chromosome ancestry revealed by histone genes in the migratory locust. *Chromosoma* 119:217–225. <https://doi.org/10.1007/s00412-009-0251-3>
- Trifonov VA, Demytyeva PV, Larkin DM, O'Brien PC, Perelman PL, Yang F, Ferguson-Smith M, Graphodatsky AS (2013) Transcription of a protein-coding gene on B chromosomes of the Siberian roe deer

- (*Capreolus pygargus*). BMC Biol 6(90). <https://doi.org/10.1186/1741-7007-11-90>
- Utsunomia R, Silva DM, Ruiz-Ruano FJ, Araya-Jaime C, Pansonato-Alves JC, Scacchetti PC, Hashimoto DT, Oliveira C, Trifonov VA, Porto-Foresti F, Camacho JP, Foresti F (2016) Uncovering the ancestry of B chromosomes in *Moenkhausia sanctaefilomenae* (Teleostei, Characidae). PLoS One 11:e0150573. <https://doi.org/10.1371/journal.pone.0150573>
- Valente GT, Conte MA, Fantinatti BEA, Cabral-de-Mello DC, Carvalho RF, Vicari MR, Kocher TD, Martins C (2014) Origin and evolution of B chromosomes in the cichlid fish *Astatotilapia latifasciata* based on integrated genomic analyses. Mol Biol Evol 31:2061–2072. <https://doi.org/10.1093/molbev/msu148>
- Valente GT, Nakajima RT, Fantinatti BEA, Marques DF, Almeida RO, Simões RP, Martins C (2017) B chromosomes: from cytogenetics to systems biology. Chromosoma 126:73–81. <https://doi.org/10.1007/s00412-016-0613-6>
- Yoshida K, Terai Y, Mizoiri S, Aibara M, Nishihara H, Watanabe M, Kuroiwa A, Hirai H, Hirai Y, Matsuda Y, Okada N (2011) B chromosomes have a functional effect on female sex determination in Lake Victoria cichlid fishes. PLoS Genet 7:e1002203. <https://doi.org/10.1371/journal.pgen.1002203>
- Zardoya R, Abouheif E, Meyer A (1996) Evolutionary analyses of hedgehog and Hoxd-10 genes in fish species closely related to the zebrafish. Proc Natl Acad Sci U S A 93:13036–13041. <https://doi.org/10.1073/pnas.93.23.13036>
- Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829. <https://doi.org/10.1101/gr.074492.107>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.