UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Botucatu

Genome assembly of the cichlid fish A*statotilapia latifasciata* with focus in population genomics of B chromosome polymorphism

Maryam Jehangir

Botucatu, July, 2017

UNIVERSIDADE ESTADUAL PAULISTA

“Julio de Mesquita Filho”

INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU

GENOME ASSEMBLY OF THE CICHLID FISH *ASTATOTILAPIA LATIFASCIATA* WITH FOCUS IN POPULATION GENOMICS OF B CHROMOSOME POLYMORPHISM

MARYAM JEHANGIR

DR. CESAR MARTINS

Dissertation presented to the Institute of Biosciences, Campus of Botucatu, UNESP, to obtain the title of Master in the Postgraduate Program -University graduate in Biological Sciences (Genetics).

Botucatu- SP

2017

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Jehangir, Maryam.
Genome assembly of the cichlid fish astatotilapia latifasciata with focus in population genomics of B chromosome polymorphism / Maryam Jehangir. - Botucatu, 2017

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Cesar Martins
Coorientador: Guilherme Targino Valente
Capes: 20204000

1. Peixe - Genética. 2. Ciclídeos. 3. Cromossomos - Polimorfismo. 4. Genoma. 5. Genômica. 6. Citogenética.

Palavras-chave: Cromossomos B; Evolução; Montagem do genoma; Peixe cíclido; Polimorfismo.

Acknowledgements

In the name of Allah, the Most Gracious and the Most Merciful, for the strengths and His blessings in completing this thesis.

I would like to thank all the people who contributed in some way to the work described in this thesis. First and foremost, I thank my academic advisor, Professor Cesar Martins, for accepting me into his group. 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. My appreciation to my co-advisor professor Guilherme Targino Valente, for beneficial ideas and knowledge regarding my research project.

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

Every result described in this thesis was accomplished with the help and support of fellow lab mates and collaborators especially, Syed Farhan Ahmad, Erica Ramos and Adauto Lima Cardoso. 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 am indebted to UNESP administration, for providing much needed assistance with administrative tasks, reminding us of impending deadlines, and keeping our work running smoothly.

I am grateful for the FAPESP (process number: 2014/17683-6), which provides me research resources to pursue my master project.

I would like to acknowledge the Department of Genetics for arrangement of different courses and and the high-quality seminars, benefited greatly from these.

I must express my gratitude to Syed Farhan Ahmad, my husband, for his continued support, encouragement and collaborating in my research. Finally, I would like to acknowledge my mother, father, sisters and brother for their love and prayers. Also not forgetting my daughter, Inshirah Farhan, her cute smile gives me more courage.

Maryam Jehangir

(Master student)

Contents list

Abstract

B chromosomes (Bs) are additional to the standard regular chromosome set (As), and present in all groups of eukaryotes. A reference genome is key to understand genomics aspects of an organism. Here, we present the *de novo* genome assembly of the cichlid fish *A. latifasciata*: a well known model to study Bs. The assembly of *A. latifasciata* genome has not been performed so far. The main focus of this study is to analyze and assemble the *A. latifasciata* genome with no B (B-) and with B (B+) chromosomes. The assembled draft B- and B+ genomes comprised of 774 Mb and 781 Mb with 1.8 Mb and 2.5Mb of N50 value of scaffolds respectively, and spanning 23,391 number of genes. High coverage data with Illumina sequencing was obtained for males and females with 0B, 1B and 2B chromosomes to provide information regarding the population polymorphism of these genomes. We observed a high scale genomic diversity in all analyzed genomes showing a high rate/frequency of population polymorphism with no evident effect of B chromosome presence. However, the B specific single nucleotide polymorphisms were found in the sequences that were located on B chromosome. While, the whole-genome rearrangements (inter chromosomal translocations) were detected in B+ genome, and structural variations including insertions, deletions, inversions and duplications were predicted in a representative genomic region of B chromosome. These results bring an evidence that existence of Bs in a genome should favour the accumulations of mutations and structural polymorphisms in the amlified genomic regions present on B chromosmes. In addition, we also performed the coverage based sequence study coupled with FISH mapping which revealed: 1) the existence of high copy number of inactive Indian Hedgehog b (*Ihhb*) gene on B chromosome emerging as a pseudogene after series of duplication events ultimately becoming a major structural component of B; 2) B chromosome have incorporated the entire 45S RNA cluster (18S ribosomal RNA, internal transcribed spacer 1, 5.8S ribosomal RNA, internal transcribed spacer 2, and 28S ribosomal RNA) from the A complement set. The assembly of *A. latifasciata* genome will serve as a reference for genetic analysis and the approach presented in this paper opens the perspective to advance understanding B chromosomes biology.

Keywords: Genome Assembly, Cichlid fish, B chromosome, Genome, Sequencing, polymorphism, evolution

1. Introduction

1.1. B chromosomes

B chromosomes (Bs) were first reported more than a century ago by E. B. Wilson in the leaf-footed plant bug insect (*Metapodius*) (Wilson, 1907). Bs are accessories to the standard regular chromosome set (As) and also named as extra chromosomes that are present in some individuals of species. B chromosomes behave different from normal set of chromosomes because they do not pair or undergo recombination with the A chromosomes during meiosis. They are inherited clonally and do not obey Mendelian law (Jones and Houben, 2003). Two key factors are involve in maintenance of Bs, its transmission rate (i.e., drive) and effects on fitness. It is believed unlikely that young extra chromosomes lacking drive or beneficial effects (even being neutral) might invade a population and become B chromosomes (Camacho et al. 1997; Camacho, 2005).

Today more than 15% of eukaryotic species including 500 animal species have been reported to posses B chromosomes (Bs) (Camacho, 2005). Within the same population not all individuals carry B chromosomes and their number can differ between individuals (e.g. *Vulpes vulpes*, 2n = 34 As + 0–8 Bs; *Rattus rattus*, 2n = 42 + 0–5 Bs). Some species have different morphological types of Bs exist within a single species . They can surprisingly exceed the number of As in some species (e.g. Zea mays, 2n = 20 As + 0–34 Bs) (Liehr et al. 2008).

In most species which carry Bs, the mitotic transmission of Bs during growth and development is normal and hence all cells carry the same number of Bs within the individual. However, several exceptional studies stated that some Bs are mitotically unstable and can vary in numbers in specific tissues and/or organs. For example, in the grasses (*Aegilops speltoides* and *A. mutica*), Bs exist in aerial organs but not in roots (Mendelson and Zohary, 1972; Ohta, 1996). Bs can also be distributed differently between genders from species to species. Normally Bs exist in both gender, but sometime the frequency of Bs are higher in one sex. In some species the Bs are present either in males only (eg. *Moenkhausia sanctaefilomenae)* (Portela-Castro et al. 2000) or females only (eg. *Astyanax scabripinnis paranae)* (Maistro et al. 1992; Mizoguchi and Martins-Santos, 1997).

B chromosomes are composed of repetitive DNAs such as rRNA genes (Houben et al. 2005; Ruiz-Estévez et al. 2012), tandemly arranged repetitive elements (Potapov et al. 1990), LINEs (long interspersed nuclear elements), SINEs (short interspersed nuclear elements) (Peppers et al. 1997), interstitial telomeric sequences (Wurster et al. 1988; Szczerbal et al. 2003). The heterochromatic

nature of B chromosomes gives the idea that these elements were mediated genetically inert and their presence were not needed for survival or reproduction of the individuals (Camacho, 2005). But recently comparative sequence analysis of the A and B chromosomes of rye plant (*Secale cereale*) and cichid fish *(Astatotilapia latifaciat*a) claim the inert nature of supernumerary chromosomes. These studies concluded that B chromosome has gained a diverse range of repeat sequences and protein-coding genes (Martis et al. 2012; Valente et al. 2014).

Different studies have revealed that B chromosomes keep transcriptionally active DNA sequences that could play some role in variety of functions, such as the discovery of proto-oncogenes and tumor-suppressor genes in the B chromosomes of canid species(Graphodatsky et al. 2005; Makunin et al. 2014), H3 and H4 histone genes in those of the migratory locust (Teruel et al. 2010), other protein-coding genes in the B chromosomes of a cichlid fish and the Siberian roe deer (*Capreolus pygargus)* (Trifonov et al. 2013; Yoshida et al. 2011). In addition, Valente *et al*. investigated the gene content of B chromosomes in cichlid fish (*Astatotilapia latifasciata)* accomplice with different functions.

The numerical frequency of Bs carrying individuals shows phenotypic effects in some species (Jones, 1982; Green, 1990). The small number of the B chromosomes seems to have no impact on the phenotype while in a high number they can effect the phenotype (Bosemark, 1957b; Gonzalez-Sanchez. et al. 2004). B chromosomes are linked with both negative and positive effects. In grasshopper (*Myrmeleotettix maculatus)*, they likely prevent animal development (Harvey and Hewitt, 1979) and sperm dysfunction (Hewitt et al. 1987). A positive behavior was also found in some organisms for example as in rice (*Oryza sativa*), Bs play role on plant height, weight of grain, and length of its panicle (Cheng et al. 2000). The presence of B chromosomes in maize (*Zea mays L*), alters the recombination frequency of A chromosomes (Rhoades, 1968).

The B chromosome originated as a by-product of A chromosome evolution of either the same or related species (Camacho et al. 2000). The study on Bs origin of *Canidae* identified that it carry several chromosomal regions of domestic dog that show co-hybridization to wild canid B chromosomes (Becker et al. 2015). Recently, the rise of next generation sequencing confessed that the B chromosomes of fish species *Astatolilapia latifasciata* and *Astyanax paranae* were evolved from multiple As (Silva et al. 2014; Valente et al. 2014). Sequencing of rye B chromosome showed that the B chromosome was originated from A chromosomes 3R and 7R(Martis et al. 2012). Kao et al. (2015) using the Random Amplified Polymorphic DNA (RAPD) technology, revealed that four short repetitive sequences were found to locate on both A and B chromosomes.

1.2. Cichlid fish as a model organism

Cichlid fishes belong to the most rich species family of Perciformes and represent one of the best model for studying different genetic fields, such as evolution and cytogenetics (Kocher, 2004; Ijiri et al. 2007; Poletto et al. 2010a, b). There is approximately 3000 species of cichlid fishes spread through out the different regions, from Central and South America, Africa to Madagascar, the Middle East, and Southern India. An intense level of adaptive radiation has characterized the evolution of cichlids, where closely 2,000 species have been diversified in the last 10 million years (Stiassny, 1991; Kocher, 2004).

The evolutionary tree of this family are divided into four subfamilies: Etroplinae (cichlid from Madagascar and South Asia - India and Sri Lanka), Ptychochrominae (Madagascar), Cichlinae (species of the Neotropics) and Pseudocrenilabrinae (cichlids from African). Further more, Cichlinae and Pseudocrenilabrinae comprise of more diverse species (Stiassny, 1991; Sparks and Smith, 2004; Genner, 2005). The African-Neotropical cichlids represent an ancestral group of Madagascar and Indian Cichlids which is regarded as the most basal group of divergence (Smith et al. 1994).

Although more than 60% of the species present a karyotype with 2n = 48, the diploid number ranges from 2n = 32 to 2n = 60. African cichlids have a modal diploid number of 44 chromosomes, whereas the Neotropical cichlids have 2n = 48 chromosomes (Poletto et al. 2010). In the genomic level, different cichlids genomes (*Oreochromis niloticus*, *Astatotilapia burtoni*, *Neolamprologus Brichardi, Pundamilia nyererеie* and *Metriaclima zebra*) have been sequenced to interpret its role and evolutionary pathways (Brawand et al. 2014). These genomes are freely available online in different databases on (available http://cichlid.umd.edu/cichlidlabs / kocherlab / bouillabase.html).

The presence of B chromosomes in cichlids is of particular interest. In our field of concern, *Astatotilapia latifasciata* is one of the appropriate model for examination of B chromosomes function and evolution.

1.3. *Astatotilapia latifasciata* as a model species for B chromosomes studies

Cichlid fishes are considered a well known model to study B chromomsomes. Bs have been determined in 7 South Americans (Feldberg and Bertollo, 1984; Martins-Santos et al. 1995;

Feldberg et al. 2004) and 14 African (Poletto et al. 2010b; Fantinatti et al. 2011; Yoshida et al. 2011; Kuroiwa et al. 2014) species of cichlids. Among the African species, B chromosomes were first described in *Astatotilapia latifasciata* from Lake Nawampasa, a satellite lake of the Lake Kyoga, part of Lake Victoria system (Poletto et al. 2010a).

B chromosomes in *A. latifasciata* have been studied with a focus on cytogenetics. Different cytogenetic techniques such as mapping ribosomal DNA sequences, the repetitive element SATA, BACs (Bacterial Artificial Chromosomes), genomic DNA fraction containing highly repeating (C0t-1 DNA) and comparative genomic hybridization (CGH) indicated that no differences specific to sex were detected and these chromosomes have many repetitive DNA sequences. One or two similar Bs were observed in both sexes of *A. latifasciata* (Poletto et al. 2010a; Fantinatti et al. 2011).

Large scale genomic analysis in the cichlid fish *A. latifasciata* (Valente et al. 2014) studied the B chromosome origin of the species. A better concept was drawn to understand the gene content, genome pattern and biological role of B chromosome by the application of NGS. The B chromosome of *A. latifasciata* composed of mostly fragmented genes with a few largely intact. Among these high intact sequences, genes involved with cell cycle (microtubule organization, kinetochore structure, recombination and progression through the cell cycle) were detected and may play a role in driving the transmission of the B chromosome in their hosts (Valente et al. 2014).

1.4. Next generation sequencing (NGS) as a powerful tool to reconstruc*t the Astatotilapia latifas*ciata

The automated Sanger method is considered as a ‘first-generation’ technology, and newer methods are referred to as next-generation sequencing (NGS). Next-generation sequencing (also known as massively parallel sequencing) technologies can be used as a means for in-depth genomic analysis (Lin Liu et al. 2012; Wold et al. 2008).

There are few common methodologies to be performed during sequencing such as preparation of template, imaging and sequencing and analysis of data. Single DNA molecules templates or clonally amplified templates are required for the preparation of NGS process. The two well known approaches “sequencing by synthesis” and “sequencing by ligation” are applied to sequence DNA. The former is based on numerous DNA polymerase-dependent step while later describe the use of DNA ligase instead of DNA polymerase. The integration of imaging techniques and these sequencing approaches range from measuring bioluminescent signals to four-colour imaging of single molecular events. Due to generation of huge amount of data by NGS, information

technology is becoming a substantial demand in terms of quality control, tracking and data storage (Metzker, 2005; Fan et al. 2006; Pop et al. 2008).

Many challenges have arisen for bioinformatics to analyze the short-reads and gene variation effectively due to short-read strategy of NGS (Wold and Myers, 2008; Yang et al. 2009). Advances in bioinformatics field such as assembly and alignments would increase the chance of accurate and error free interpretation of short reads (Pop and Salzberg, 2008). Normally a run of typical NGS platform produces tens to hundreds of Gbp short reads. Finally, a raw data of terabytes is generated in an average next generation sequencing experiment, hence causing problems to manage and analyze the data. Creation of bioinformatics tools, advancement in management and development of data storage can lead to successful and useful application of NGS. Another important problem for bioinformatics analysis is the variabilities among the different NGS platforms. Because of differences in read length and data format, many factors such as data processing, sequence quality scoring, assembly and alignment of softwares need to be diversified accordingly etc. Various features of NGS is the reason of coexistence of multiple platforms in the marketplace with some having prominent benefits for specific applications over others. Commercially available technologies of NGS are Roche/454, Illumina/Solexa, Life/APG and Helicos BioSciences, the Polonator instrument and the near-term technology of Pacific Biosciences, who aim to bring their sequencing device to the market (Branton et al. 2008).

NGS technologies have rapidly changed our minds in respect to various scientific aspects such as clinical, basic and applied research. In some aspects, the potential of NGS is comparable to the early days of PCR, with one’s imagination being the primary limitation to its use. In some cases NGS can exceed of one billion short reads per instrument run (Wold et al. 2008; Wang et al. 2009). Due to innovation in sequencing technology, progress in genomics has been progressed at a rapid pace. With more and more organisms being sequenced, a big stream of genetic data is overloading the world every day. Not only do these studies provide the knowledge for basic research, but also they afford immediate application benefits (Lin Liu et al. 2012). Next generation sequencing is now a useful tool in routine applications to be used for wide areas of biology, enabling researchers to uncover important biological mysteries (Soon et al. 2012). The arrival of different NGS platforms is opening opportunities to produce large number of sequence reads at low cost with wide range of applications. These include variant discovery by resequencing targeted regions of interest or whole genomes, *de novo* assemblies of bacterial and lower eukaryotic genomes, cataloguing the transcriptomes of cells, tissues and organisms (RNA–seq), genome-wide profiling of epigenetic marks and chromatin structure using other seq-based methods (ChIP–seq, methyl–seq and DNase–

seq), and species classification and/or gene discovery by metagenomics studies (Petrosino et al. 2009).

Although large-scale genomic analyzes have already been explored in understanding B chromosome biology, such approach can be applied to a wide range of research questions as sequencing genomes, comparative biology studies, public health, epidemiology, physiology, and gene expression. High-scale analysis provides detailed information on the composition of genomes and representing important tools to advance over structural and functional analysis of cells, tissues and organisms.

1.4.1. Illumina sequencing

The Illumina platform employs cyclic reversible termination (CRT) chemistry for DNA sequencing. The process relies on growing nascent DNA strands complementary to template DNA strands with modified nucleotides, while tracking the emitted signal of each newly added nucleotide. Each nucleotide has a 3’ removable block and is attached to one of four different fluorophores depending on its type. Sequencing occurs in repetitive cycles, each consisting of three steps: (a) extension of a nascent strand by adding a modified nucleotide; (b) excitation of the fluorophores using two different lasers, one that excites the A and C labels and one that excites the G and T labels; (c) cleavage of the fluorophores and removal of the 3’ block in preparation for the next synthesis cycle. Using this approach, each cycle interrogates a new position along the template strands. Illumina provides a large number of intermediate input files that can be used to perform a fine-scale analysis of distortion factors. The most useful among them are imaging files, intensity files, and sequencing files (Naiara and Michael, 2012).

Illumina Sequencing increasing throughput exponentially over first-generation Sanger sequencing. There are certain drawbacks of second generation instruments despite of several useful benefits. Amplification of template DNA is required before sequencing which result in biased coverage. Another major disadvantage of these technologies is the production of short length reads making assembly and related analysis problematic. Average length of Illumina reads is 100 bp is not appropriate for an ideal assembly as a theoretical model assumes that reduction of read lengths from 1,000 bp to 100 bp may result a sixfold or more decline in continuity (Kingsford et al. 2010). The new chemistry for Illumina equipment promises reads of 300bp, but the quality of longer reads is lower.

1.4.2. *De novo* genome assembly

In bioinformatics, *de novo* genome assembly refers as the procedure of aligning short DNA fragments(reads) together into longer segments (contigs or scaffolds) and further joined into linkage groups or placed on chromosomes (Ellegren, 2014; Simpson and Pop, 2015). Adequate overlap between the sequence reads in the genome at every position is essential for an accurate assembly. Longer reads may cause more overlap hence decreasing the depth of unnecessary raw reads. Paired end sequencing technology is more efficient in assembly because of its accurate placement of reads (Robert et al. 2014; Ellergren, 2014). A well assembled genome is needed to detect vital functional and structural genomic elements and essential for better analysis of genetic variations (Mahul et al. 2015).

Large scale genomic data were previously obtained for the cichlid fish *A. latifascita* using Illumina NGS platform (Valente et al. 2014). Although it was generated a high coverage (over 30x of coverage) of 0B (genomes with no B chromosome) and 2B (genome with 2B chromosomes) genomes, the data were not enough to reconstruct the genome of the species. Although the generation of high amount of nucleotide sequence data becomes available with the NGS technologies, the assembly of genomes continues to be one of the central problems of bioinformatics. This owes, in large part, to the technologies that provide ‘reads’ of short sequences of DNA, from which the genome is inferred. Larger sets of data, and changes in the properties of reads such as length and errors, bring with them new challenges for assembly (Henson et al. 2014).

1.5. DNA polymorphism

DNA polymorphisms are broad type of genetic variations among individuals of same species or a different species. In another words, if a few alleles of a polymorphic site has the most widely recognized variation among them happening with under 99% recurrence in the population(Cavalli-Sforza, 1971; Schork et al. 2000). Substantial variation can be found at different points in the genome while performing a comparative analysis of genomic DNA sequences of two individuals on equivalent chromosome. Genes can comprise of many polymorphisms which have impact on many phenotypes traits such as hair color and height. Some of these polymorphisms may cause disease susceptibility, affect drug responses and considered important in other medical aspects. Although several polymorphisms occur outside of genes and do not show any effect (Bentley, 2000; Barnes and Grey, 2003).

DNA polymorphism emerges as a result of mutation in the DNA level. The different types of polymorphism are specified according to the type of mutation that they are generated in the sequence. Polymorphism is broadly divided into different types, include single nucleotide polymorphism (SNPs), insertions and deletions (INDELs), and other larger rearrangements such as (inversion, translocation and transversion) (Botstein et al. 1980; Schork et al. 2000).

1.5.1. Single Nucleotide Polymorphism (SNP)

Single nucleotide polymorphism(SNP) is simplest type of polymorphism results from alteration of a single nucleotide base adenine (A), thymine (T), cytosine (C) or guanine (G) by other nucleotide and bring mutation in genome. SNPs are the most common form of genetic variation, accounting for 90 percent of all human genetic variations. For instance, the nucleotide diversity (the SNP rate) between humans is about 0.1%, while this rate can be as high as 3-5% in some fish and other sea organisms (Riva et al. 2002). Few studies have predicted an average range of 0.3-1,000 kb of genome can detect SNPs, however most of the data was based on specific genes and therefore fail to address the question of SNP frequency. There is no clear cut evidence for the correct estimation of SNPs occurrence in the rest of genomes (Halushka et al. 1999; Cargill et al 1999). Most SNPs have only 2 alleles (biallelic SNP) and they may occur in both coding and non coding regions of the genome (99% not in genes). The SNPs located in coding regions are little harmful and inherit changes to DNA while the remaining non-coding SNPs become silent, harmless and does not effect (Strittmatter et al. 1996).

1.5.2. Insertion or deletion Polymorphism(INDELs)

Insertions or deletions (INDELs) is another common form of genetic polymorphism, as the name indicates it result when a section of DNA is either deleted or inserted. This type of polymorphism mostly occurs due to the existence of nucleotide pattern or variable number of repeated base (Cooper et al. 1999). The size of repeat base pattern vary from few (two, three or four) nucleotides repeats known as 'micro-satellites' up to several hundreds base pairs called 'mini-satellites or 'variable tandem repeats' (VNTRs). Many alleles are referred as 'highly polymorphic' due to frequent number of repeat polymorphisms having several different repeat sizes. This is very much helpful to study population genetics because the probability of having the same number of

repeats in two individuals from, suppose, different population (healthy vs diseased, ethnic groups) may be much lower (Shriver et al. 1997).

1.5.3. Genome rearrangement

A large amount of DNA segments are displaced as a result of a unique category of mutation known as genome rearrangement. This happens during the process of chromosomal breakdown at certain sites also called breakpoints, followed by reassembling of chromosomal pieces in a incorrect order. The genome rearrangements can have deleterious effects or even lethal if they occur in the coding sequence, making the gene inactive. On other hand, a rearrangement whose breakpoint fall in non-functional region is likely thought to be neutral in effects (Mathieu, 2001).

The related studies conducted on genome rearrangement until now has concluded two categories: (i) intra-chromosomal rearrangements, have subtypes: transpositions, reversals/inversions, and block-interchanges/generalized transpositions, and (ii) inter-chromosomal rearrangements, have subtypes: translocations, fusions and fissions. Transpositions shift a segment from one location to another on a chromosome or, equivalently, exchange two adjacent and non-overlapping segments on the chromosome. Inversions or Reversals reverse a chromosomal segment and also exchange its strands. Exchanging two non-overlapping segments which are not necessary adjacent on chromosome is called generalized transposition or Block-interchanges. The exchange of a telomeric segment of one chromosome with that of other chromosome is termed as translocations. Fusions combine two separate smaller chromosomes into a single bigger one and fissions involve breakage of a chromosome into two smaller ones (Alekseyev, 2008; Alekseyev and Pevzner, 2008).

Advanced and low-error analytical strategies should be designed to use DNA polymorphism for latest genetic applications. Approximately all DNA polymorphisms can be captured now with the help of modern development of methods for detection of structural variants (Svs) and SNPs including NGS. Still, there are many challenges to meet for highly accurate identification of DNA polymorphisms since short-read sequencing strategy is difficult for analysis to infer chromosomal context (Kidd et al. 2008; Graubert et al. 2007; Emerson et al. 2008).

DNA polymorphism analysis were also studied in B chromosomes of rye (*Secale cereale*) (Martis et al. 2012) and Grasshopper (*Eyprepocnemis plorans*) (Muñoz-Pajares et al. 2011). Numerical polymorphisms have been described in a large number of maize landraces (Mcclintock et al. 1981; Chiavarino et al. 1995; Naranjo et al. 1995; Rosato et al. 1998), with differences in the number of B's being one of the main factors contributing to intraspecific genome-size variation.

Silva & Yonenaga-Yassuda (1998) reported a conspicuous heterogeneity of size, morphology, constitutive heterochromatin patterns and localization of telomeric sequences of B chromosomes for the rodent Nectomys, which allowed them to suggest differences in the composition of these chromosomes. A considerable amount of chromosome variability involving the presence of Bs and polymorphisms of autosome pairs was detected microteiid lizard (*Nothobachia ablephara) (*Pellegrino et al. 1999*)*. Reports on the nucleotide level polymorphism of Bs are still extremely scarce , therefore our work is a novel contribution to understand the association and significance of polymorphism in B chromosome.

2. Objectives

2.1. General Objective

- Reconstruction of *A. latifasciata* genomes (B+ and B-) to analyze B chromosome polymorphism and genomic diversity

2.2. Specific Objectives

- Generation of high coverage data using Illumina sequencing
- Assembly of B+ and B- genomes of *A. latifasciata*
- Prediction of genes (structural annotation)
- Analysis of genome diversity
- Study of B chromosomes polymorphism
- Search of B specific genes in *A. latifasciata* genome

3. Material and Methods

3.1. The model organism, *Astatotilapia latifasciata*

The cichlid fish *A. latifasciata* (Pseudocrenilabrinae) is native from the lakes Kyoga and Nawampasa (East Africa) and classical and molecular cytogenetic studies detected the presence of one or two B chromosomes in some individuals of populations farmed in Brazil for aquarium hobbyist market (Poletto et al. 2010a; Fantinatti et al. 2011).

The presence of B chromosomes in *A. latifasciata* associated to the release of whole genome sequences of several other African cichlids (The International Cichlid Genome Consortium, 2006) (Brawand et al. 2014) makes this species an excellent model for genomic studies. Furthermore, *A. latifasciata* can be easily maintained in captivity allowing directed crosses and sampling of tissues for chromosome, DNA, RNA and protein analysis. Currently, populations of *A. latifasciata* without B chromosomes (0B) and with 1 (1B) or 2 B chromosomes (2B) have being maintained at the fish facility of the Integrative Genomics Laboratory at Institute of Biosciences/UNESP, Botucatu, SP, Brazil. The use of these animals for biological samples were held according to ethical conventions utilized by Brazilian College for animal experiments accepted by ethic committee of the Biosciences Institute, UNESP-Sao Paulo State University.

3.2. Chromosome preparation, DNA sampling and sequencing

The liver tissues of *A. latifasciata* samples were collected and karyotyped by classical chromosome preparation protocols employed for fish to check the presence of 0, 1 or 2B chromosomes. Additionally, these samples from males and females with 0B, 1B or 2 B chromosome genomes were checked via polymerase chain reaction (PCR) and real time PCR (Fantinatti et al. 2016). High quality genomic DNA (gDNA) from liver tissues samples of *A. latifasciata* males and females with 0B, 1B and 2B chromosomes selected for next generation sequencing (NGS).

Seven individuals including males and females with 0B, 1B and 2B chromosomes were subject to Illumina sequencing using the service sequencing facility of University of Maryland, USA ("http://www.ibbr.umd.edu/facilities/sequencing) . Additionally NGS were also performed in a MiSeq Illumina equipment available to our use in the Institute of Biosciences/UNESP, Botucatu. All the generated data are already available at Sacibase (www.sacibase.ibb.unesp.br).

3.3. Illumina Next-Generation Sequencing

For the Illumina libraries, gDNA from the individuals were sheared to an average size of 350-550 bp using an S220 focused ultrasonicator (Covaris Inc., Woburn, MA). Separate libraries were constructed for the gDNA samples using the TruSeq DNA sample preparation kit ver.2 rev.C (Illumina Inc., San Diego, CA). Paired end (100-190 bp) reads sequencing of each library were performed in separate lanes on an Illumina HiSeq 1000 sequencer.

3.4. Pre-processing step of NGS data

At the point when the sequences data are prepared, its need to checks whether a set of sequence reads in a .fastq file exhibit any bizarre qualities (which might indicate either low sequence quality, or interesting biological features in sample). The reads quality was checked by the software FastQC (website: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and the ambiguous data need to filtered out, just leaving only a small fraction of usable, unique read pairs for assembly. The filtration of the data was done according to the requirement of genome assemblies by FASTX (http://hannonlab.cshl.edu/fastx_toolkit/) tool kit (-q 30, -p 90 parameters). The illumina adapters checking using blastn search (e-value ≤10e -5 and 90% of identity were the cutoff parameters), which reads with some hit were eliminated through customized python programming script. The not paired reads (singletons) were discarded by Pairfq software for *de novo* assembling. Libraries related to one fragment paired-end were merged and make two datasets for *de novo* assembling. While the sequences coverage(defines as a the average number of reads per site) of all samples were calculated by coverage formula given as:coverage = (read count x read length) / total genome size, Read Counts= Total reads of genome, Read length= Expected length of reads, Total genome= Total genome size / close related genome size (*Metriaclima. Zebra*)

3.5. Genome assemblies and quality evaluation

For choosing appropriate assembly software, it is important to consider both the amount of sequencing data and which computational resources are available (Schatz et al. 2010a). For that

reason we used Assamblathon2 (Bradnam et al. 2013) recommended strategy for fish genome, using the Velvet (v 1.2.08) (Zerbino and Birney, 2008) and SOAPdenovo2 softwares (Luo et al. 2012).

To estimates the best k-mer length for genome *de novo* assembly, *we* ran KMERGENIE (Chikhi and Medvedev, 2014) on pre-processed reads with putative k values of 31-91. The optimal values of k were predicted to be 71-mer and using this k value for assembling. The clean reads comes from pre-proceesing step consisting of all B- and B+ samples having male and female were inserted in to Velvet (Zerbino and Birney, 2008) and SOAPdenovo2 assemblers to construct scaffold level *de novo* assemblies. The two different assembler were used to compare both assembler generated assemblies and at the end choose the best possible assembly. Velvet assembler with following parameter settings for command line run; "-ins_length 500 , exp-cov auto, -unused_reads yes, read_trkg yes" were used for both assemblies. Other assembler SOAPdenovo was also set for genome assemblies by using parameters: “insert length 350, 500 and 550, asm_flags=3” were setted up to high coverage cleaned reads. Both assemblers generated a scaffold level assemblies (B+ and B-). To close the gaps within scaffolds of B+ and B- assemblies to be more accurate assemblies, GAPFILLER software (Boetzer and Pirovano, 2012) was operated. The program was executed by ('–m’ = 80, ‘-t’= 10, '-g' = 5) settings. The final gap closed scaffolds assemblies were constructed and used for further analysis**.** QUAST software (Gurevich et al. 2013) was applied for evaluation of the generated scaffolds before and after gaps filled to computing several metric values (length, number, length variation, N50, gap length). This software was executed in two seperate runs, one for independent evaluation and other with to compare our genomes with closest reference, *M. zebra* genome. The program was executed for eukaryotic genes comparison. We therefore, evaluated assembly using two common quality measures: the contig N50 length, the assembly size. Finally, the B- scaffold level genome assembly was selected for future analysis and used as a reference.

3.6. Structural annotation of genes

The GeneMark-ES (Lukashin and Borodovsky, 1998) was used for eukaryotic ab-initio gene prediction, and identification of a set of 453 core genes that are supposed to be highly conserved in all eukaryotes, using CEGMA pipeline (version 2.4) .

For identification of protein coding genes in assembled genome, We used three approaches: homology-based, *de novo* and transcript sequences-based by using a pipeline MAKER v2.31.8

(Cantarel et al. 2008); we used repetitive elements of custom libraries of fish and Metazoan elements, *Danio rerio* CDS and proteome files from (http://www.ensembl.org/Danio_rerio), *A. latifasciata* assembled transcriptomes accessed from (http://sacibase.ibb.unesp.br/), gene prediction based on Lamprey training set and Blastn to NCBI database (National Center for Biotechnology Information – http://www.ncbi.nlm.nih.gov/).

3.7. Genome diversity and genome structural variations analysis

The genomic diversity(Identification of unique and common SNPs & INDELs among different individuals) was determined using the following procedures. The Illumina reads generated for the different *A. latifasciata* genomes (males and females with 0B, 1B and 2B chromosomes) were first filtered for alignments by FASTX (http://hannonlab.cshl.edu/fastx_toolkit/) tool kit (-q 20, -p 80 parameters). The trim reads were aligned against our draft B- assembly as a reference using Bowtie2 (Langmead and Steven, 2012) --very-sensitive" option. Nucleotide polymorphism was identified using SAMtools tool to search for genome variations among the samples. Nucleotide polymorphism was identified using SAMtools (http://samtools.sourceforge.net/) to search for genome variations among the samples. The output files (VCF format) were subjected to VCFtools (vcf-stats and vcf-compare) (http://vcftools.sourceforge.net/) for statistical analysis to discover the frequency of single nucleotide polymorphism (SNPs), insertion/deletion (INDELs) and to compare these factors among indiviuals to find out the shared, unique or B related population polymorphism . Filtration was carried out to eliminate the lower quality ($Q \leq 20$ and $DP \geq 100$) SNPs and INDELs using vcffilter (https://github.com/vcflib/). Parameters for each step of this analysis were established according to the standard requirements. Similar approach was followed to detect polymorphisms in genic sequences (CDS, exons and introns) of *A. latifasciata* B+ and B- genomes aligned against *de novo* transcriptome assembly (Marques et al. unpublished).

The structural variations (translocations) were detected in 1B sequencing data by Delly (Rausch et al. 2012). Delly integrates short insert paired-ends and split-read alignments to accurately delineate genomic rearrangements throughout the genome. This pipeline was applied to both B+ and B- reads in Bam format (generated using bowtie tool) considered as sample and control respectively. *Denovo* genome assembly of *A. latifasciata* was utilized as reference to locate the variations on the scaffolds regions. The detected translocations (breakpoints) were visualized by ClicO (Cheong et al. 2015), an online web-service based on Circos (Martin et al. 2009). To uncover

nature of these transclocation, We extracted randemly a few of these regions (515-520 MB) and subjected to NCBI Blast (https://blast.ncbi.nlm.nih.gov) for annotation.

More specifically, structural variations such as deletions, insertions, transversions, inversions and duplications in genomic regions related to B chromosome (B block: B+ reads having higher coverage than B- reads) were analyzed by inGAP-sv tool (Qi et al. 2011). InGAP-sv detects the structural variations (SVs) on the basis of paired end mapped reads pattern and coverage of depth strategies. We applied this pipeline to B+ sam file generated using BWA (Li and Durbin, 2009) and *A. latifasciata* genome was used as a reference. After the SAM file was loaded into inGAP-sv, user-defined 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 et al (2011).

3.8. Analysis of B localized sequences

The nucleotide sequences of several previously identified B chromosomes genes of vertebrates (Makunin et al. 2014) supplementary (Table 1) were retrieved from the NCBI database (https://www.ncbi.nlm.nih.gov/). Consensus sequences constructions were obtained using Geneious v. 4.8.5 software (Drummond et al. 2009) for genes with more than one sequence available. All final sequences were used as queries against the *A. latifasciata* genome in a standard blastn search. Number of hits, E values and percent identity were considered to further proceed with B related analysis. This analysis confirmed the existence of *Ihhb* (Indian Hedgehog B gene) and 45S rRNA (18S ribosomal RNA, internal transcribed spacer 1, 5.8S ribosomal RNA, internal transcribed spacer 2, and 28S ribosomal RNA) gene sequences in *A. latifasciata* genome while the remaining genes were eliminated from the project because of partial or complete absence. The generated Illumina high coverage reads of all B- (0B male and 0B female) and B+ (1B Males, 1B Female and 2B Male) samples were aligned against both reference genes 45S rRNA and *Ihhb* of the cichlids *Oreochromis aureus* and *Lithocromis rubripinnis* respectively, using paired-end mode of Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml) with the –very sensitive option. The output aligned files were converted to binary format and indexed using samtools. Each file was normalized using RPKM package of deeptools (Ramírez et al. 2014) to fix bias of initial coverage. These files were then visualized by integrated genome browser IGB (http://bioviz.org/igb/index.html) to compare coverage of both genes in B- and B+ samples. SNPs at different sites of nucleotide reads were detected. According to Marques et al. (unpublished) a total

of 36 libraries were generated from 18 samples of transcriptomes of *A. latifasciata*. These transcriptomes were divided into triplicates of gonads, brain and muscle of male and female individuals. The uploaded transcriptomics data (http://sacibase.ibb.unesp.br/) and genomics data (aligned files) were visualized and screened. We did BLAST alignment of the genes against transcriptome assembly to locate them on specific scaffolds. The B specific SNPs were searched for simultaneous viewing all tracks of available data against transcriptome assembly as a reference.

3.9. Primer design, probes construction and FISH mapping of *Ihhb* and 45S rRNA genes

Primers listed in the supplementary (Table 2) were constructed using PrimerQuest (http://www.idtdna.com/primerquest/home/index) tool, checked for specificity by Primer-Blast (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) and evaluated by Primer Stat (http://www.bioinformatics.org/sms2/pcr_primer_stats.html). Genomic DNA was subjected to polymerase chains reaction (PCR) to obtain DNA segments to be used as probes for fluorescent in situ hybridization (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 Dynamic Terminator 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 database to check for similarities to deposited sequences of corresponding genes. Probes were labelled with digoxigenin-11-dUTP (Roche Applied Science) and the signal was detected with anti-digoxigenin-rhodamine (Roche Applied Science). FISH was performed using the protocol described by Pinkel et al. (1986) 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 and were optimized for brightness and contrast using Adobe Photoshop CS2.

4. Results

4.1. Illumina next generation sequencing

A total of seven B- and B+ samples were sequenced using Illumina HiSeq and Miseq platforms respectively. Each sample has different coverage of the corresponding genomes over the entire length (848,776,495bp) of the *M. zebra* genome. Each sample has been given a name such as: “M1-5” having different male samples and “F1-2” for female samples; 0-2 for B-, 1B and 2B, respectivelly; in (Table 1).

Table.1. Illumina data obtained for *A latifasciata* including mapped over *M. zebra* genome.

Samples	Read length	Coverage	Coverage after filtration	Total reads	Remained reads after filtration	Reference
M1-0B	101	47.7x	38x	401,017,570	323226972(80%)	Valente et al. 2014
F1-0B	191	75.5x	42.5x	337,349,994	190214688(56%)	Present work
M2-1B	35-250	2x	1.6x	716,030,6	5911265 (82%)	Present work
M3-1B	101	16.8x	13.4x	143,441,264	114064786(79%)	Present work
M4-1B	101	43.1x	34.8x	366,602,572	296161988(80%)	Present work
F2-1B	191	70.2x	40.1x	313,818,884	179286280(57%)	Present work
M5-2B	101	43.6x	30x	306,823,512	254993955 (83%)	Valente et al. 2014

The raw data was then filtered and illlumina adapters were trimmed and a total of filtrated 513,441,660 (77.05%) of B- reads with 80.5x coverage and 850,418,274 (74.7%) with 120x coverage of B+ reads were remained after filtration (Table 1).

4.2. *De novo* based B- and B+ genome assemblies

The B- and B+ draft assemblies were 771,316,069 bp and 781,068,509 bp of genome size respectively, which indicates that 84% of the B- and 80% of B+ genomes (This values correspond to the total used reads by assembler during assembly) are captured in scaffolds. The *de novo* assembly yielded 197,652 number of scaffolds for the B+ genome, with N50 of 25.54 kb, and 218,259 number of scaffolds for B- genome with N50 of 18.640 kb being the longest one with 23.86 kb and 23.36 kb respectively in (Table 2 and Supplementary Table 3). Our scaffold level genome assemblies corresponding to the closed reference *M. zebra* genome consisting of

848,776,495 bp cumulative length and GC contents of 40.5% given in (Supplementary Figure 1a,b). About (75.329%) and (74.556%) of *M. zebra* sequences had a significant hit in the B+ and B- genome assemblies respectively.

The scaffold assembling gave a total of 3,998 and 3789 gaps per 100-kb of the assembled genomes(B- and B+) filled to 2,076 and 1884 respectively. The comparative analysis by QUAST (Gurevich et al. 2013) indicated that there was an improvement in assemblies after the gaps filling. Total length of B- and B+ genomes increased up to 774,140,944 bp and 782,153,984 bp respectively. Additional statistics is given in Supplementary Table 3 and Table 2.

Table 2. QUAST evaluations statistic of B+ genome.

Statistic	**B+ assembly**	**B+ assembly after gaps filled**
# of contigs	322328	-
Total Length of contigs	767906325	-
N50 of contigs	7997	-
GC (%) contigs	40.49	-
Largest contig	101668	-
# of scaffolds (>=500bp)	78064	78078
# scaffold(>= 0 bp)	197652	197652
Total length of scaffolds(>= 0 bp)	781068509	782153984
Total length of scaffolds(>=500bp)	756378434	757469399
Largest scaffold	238637	238740
GC (%)	40.48	40.51
Scaffolds N50	25546	25546
Scaffolds NG50	21574	21623
Scaffolds N75	11565	11570
Scaffolds NG75	7368	7429
Scaffolds L50	8086	8095
Scaffolds LG50	10075	10059
Scaffolds L75	19087	19109
Scaffolds LG75	26612	26517
# unaligned scaffold	22147	20980
Unaligned length	30818517	4751825
Genome fraction (%)	75.329	76.255
Duplication ratio	1.134	1.125
# N's per 100 kbp	3789.43	1884.55
Largest alignment	173293	180244

Remaining results are given in the manuscript attached as a chapter 1 in this thesis. The attached manuscript will be reviewed and submitted to publication in scientific journal.

5. Discussion

The chapter 1 (manuscript) encompasses the discussion as a separate section.

6. Chapter 1

De novo genome assembly of the cichlid fish *Astatotilapia latifasciata* with focus in B chromosome.

M. Jehangir[a], S. F. Ahmad[a], A. L. Cardoso[a], G. T. Valente[b] , C. Martins[a]

[a]Department of Morphology, Institute of Bioscience, UNESP -São Paulo State University, Botucatu, SP, Brazil

[b]Bioprocess and Biotechnology Department, Agronomical Science Faculty, UNESP Sao Paulo State University, Botucatu SP, Brazil. Email:maryam.bioinfo.unesp@gmail.com

Abstract

B chromosomes (Bs) are additional to the standard regular chromosome set (As), and present in all groups of eukaryotes. The origin and role of Bs have been the objective of genome-wide studies. A reference genome is key to understand genomics aspects of an organism. Here, we present the *de novo* genome assembly of the cichlid fish *A. latifasciata*: a well known model to study Bs. The assembled draft genome comprised of 774 Mb with 1.8 Mb of N50 value of scaffolds and spanning 23,391 protein coding genes. High coverage data with Illumina sequencing was obtained for males and females with 0B, 1B and 2B chromosomes to provide information regarding the population polymorphism of these genomes. We observed a high scale genomic diversity in all analyzed genomes showing a high rate/frequency of population polymorphism with no evident effect of B chromosome presence. However, the B specific single nucleotide polymorphisms were found in the sequences specially located on B chromosome. Only whole-genome rearrangements (inter chromosomal translocations) were detected in B+ genome, and structural variations including insertions, deletions, inversions and duplications were predicted in a representative genomic region of B chromosome. These results bring an evidence that existence of Bs in a genome should favour the accumulations of mutations and structural polymorphisms in the amlified genomic regions present on B chromosomes. In addition, we also performed the coverage based sequence study coupled with FISH mapping which revealed: 1) the existence of high copy number of inactive Indian Hedgehog b (*Ihhb*) gene on B chromosome emerging as pseudogene after series of duplication events ultimately becoming a major structural component of B; 2) B chromosome have incorporated the entire 45S RNA cluster (18S ribosomal RNA, internal transcribed spacer 1, 5.8S ribosomal RNA, internal transcribed spacer 2, and 28S ribosomal RNA) from the A complement

set. The assembly of *A. latifasciata* genome will serve as a reference for genetic analysis and the approach presented in this paper opens the perspective to advance understanding B chromosomes biology.

Keywords: Genome Assembly, Cichlid fish, B chromosome, Genome, Sequencing, polymorphism, evolution

Introduction

B chromosomes (Bs are accessories to the standard regular chromosome set (As) that are present in some individuals of more than 15% of eukaryotic species. These extra chromosomes do not recombine with members of the basic A chromosomes complement and do not pursue the rules of the Mendelian segregation law (Jones and Houben, 2003). They are mostly heterochromatic, comprising of a large amount of repetitive DNA and their presence are not needed for survival or reproduction of the individuals (Camacho, 2005). But recently different studies have revealed that B chromosomes keep transcriptionally active DNA sequences that could play some role in variety of functions (Graphodatsky et al. 2005; Makunin et al. 2014; Teruel et al. 2010; Trifonov et al. 2013; Yoshida et al. 2011). Several studies have also focused to investigate the origin of B chromosome in various species of plants and animals, and found it be a derivative of standard A chromosomes. (Alfenito and Birchler, 1993; Martis et al. 2012; Silva et al. 2014; Valente et al. 2014).

Chromosome biology, has gained remarkable advancement due to the recent development of genomics. The applications of genomics, such as next generation sequencing (NGS) to the study of B chromosome is on the rise. Recently, NGS technology has contributed to many studies involving B chromosome analysis in plants, animals, and fungi (Camacho, 2005). The use of this technology joined with increasing information of genome organization together open a new chapter in B chromosome studies (Makunin et al. 2014). The development of NGS has made a key impact in assembling genomes of diverse organisms and studying their genomics features such as genetic polymorphism and variations (Imelfort et al. 2009; Mahul et al. 2015). Approximately all DNA polymorphisms can now be captured with the help of modern development of methods for detection of structural variants (SVs) and SNPs using NGS data (Kidd et al. 2008).

Despite an increase in many evidences about their origin, the gene content and pattern of evolution of Bs, these genomic elements still remain an open question for the majority of species. One of them is a cichlid fish (*Astatotilapia latifasciata)*. Among the African species, B

chromosomes were first described in *Astatotilapia latifasciata* from Lake Nawampasa, a satellite lake of the Lake Kyoga system (Poletto et al. 2010). Previous analysis applied to the B chromosome in *A. latifasciata* were based on cytogenetics and comparative genomics studies. Cytogenetics confirmed the both sexes of *A. latifasciata* can have either one or two similar B chromosomes enriched with many repetitive DNA sequences with no sex-specific differences (Poletto et al. 2010; Fantinatti et al. 2011). While the cytogenomics revealed that the B chromosome of *A. latifasciata* has intact genes and degenerated sequences derived from most of the A chromosome set; some of the intact genes are potentially active for transcription. (Valente et al. 2014). However, a complete overview regarding certain genomic features of the B chromosomes was not accomplished due to lack of an assembled genome.

Here, we present a *de novo* genome assembly of *A. latifasciata*, and its annotation and genetic level polymorphisms. This genomic assembly not only contributes to uncover the important aspects of B chromosome but will also provide an extra resource for studying genomics features such as variations, genes identification and sequence mapping in closely related species. Further, in this study, we have explored the genomic diversity of different individuals and discussed the association of structural variations with B chromosome. The detection of extensively distributed interchromosomal rearrangements allow us to uncover unknown evolutionary breakpoints that occurred in the *A. latifasciata* genome with B chromosome. We have also performed an analysis of B chromosome sequences originated from A complement and their physical mapping.

Methods

Chromosome preparation, DNA sampling and next generation sequencing data

The kidney tissues of *A. latifasciata* samples were collected and karyotyped by classical chromosome preparation protocols employed for fish to check the presence of 0, 1 or 2B chromosomes. Additionally, these samples from males and females with 0B, 1B or 2 B chromosome genomes were checked via polymerase chain reaction (PCR) and real time PCR (Fantinatti et al. 2016). High quality genomic DNA (gDNA) from liver tissues samples of *A. latifasciata* males and females with 0B, 1B and 2B chromosomes selected for next generation sequencing (NGS). Seven individuals including males and females with 0B, 1B and 2B chromosomes were subjected to Illumina sequencing using the service sequencing facility of University of Maryland, USA (http://www.ibbr.umd.edu/facilities/sequencing). For the Illumina libraries, gDNA from the individuals were sheared to an average size of 350-550 bp using an S220 focused ultrasonicator (Covaris Inc., Woburn, MA). Separate libraries were constructed for the gDNA samples using the TruSeq DNA sample preparation kit ver.2 rev.C (Illumina Inc., San Diego, CA). Paired-end (100-190 bp) sequencing of each library were performed in separate lanes on an Illumina HiSeq 1000 plataform. Additionally, one sample 1B male was sequenced by Miseq paltform equipment available in the Institute of Biosciences/UNESP, Botucatu.

The reads quality was checked by the software FastQC (website: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and filtration of the data was done according to the requirement of genome assemblies by FASTX (http://hannonlab.cshl.edu/fastx_toolkit/) tool kit (-q 30, -p 90 parameters). The illumina adapters checking using blastn search (e-value ≤10e -5 and 90% of identity were the cutoff parameters), which reads with some hit were eliminated through customized python programming script. The not paired reads (singletons) were discarded by Pairfq software for *de novo* assembling. Libraries related to one fragment paired-end were merged and make two datasets for *de novo* assembling.

The coverage of all samples were calculated by coverage formula given as: coverage = (read count x read length)/total genome size, being read count = total reads of genome, read length = expected length of reads and the total genome = total genome size/close related genome size (*Metriaclima zebra*).

Genome assemblies and quality evaluation

For choosing appropriate assembly software, it is important to consider both the amount of sequencing data and which computational resources are available (Schatz et al. 2010a). For that reason we used Assamblathon2 (Bradnam et al. 2013) recommended strategy for fish genome, using the Velvet (v 1.2.08) (Zerbino and Birney, 2008) and SOAPdenovo2 softwares (Luo et al. 2012).

To estimates the best k-mer length for genome *de novo* assembly, *we* ran KMERGENIE (Chikhi and Medvedev, 2014) on pre-processed reads with putative k values of 31-91. The optimal values of k were predicted to be 71-mer and using this k value for assembling. The clean reads comes from pre-proceesing step consisting of all B- and B+ samples having male and female were inserted in to Velvet (Zerbino and Birney, 2008) and SOAPdenovo2 assemblers to construct scaffold level *de novo* assemblies. The two different assembler were used to compare both assembler generated assemblies and at the end choose the best possible assembly. Velvet assembler with following parameter settings for command line run; "-ins_length 500 , exp-cov auto, -unused_reads yes, read_trkg yes" were used for both assemblies. Other assembler SOAPdenovo was also set for genome assemblies by using parameters: “insert length 350, 500 and 550, asm_flags=3” were setted up to high coverage cleaned reads. Both assemblers generated a scaffold level assemblies (B+ and B-). To close the gaps within scaffolds of B+ and B- assemblies to be more accurate assemblies, GAPFILLER software (Boetzer and Pirovano, 2012) was operated. The program was executed by ('–m’ = 80, ‘-t’= 10, '-g' = 5) settings. The final gap closed scaffolds assemblies were constructed and used for further analysis**.** QUAST software (Gurevich et al. 2013) was applied for evaluation of the generated scaffolds before and after gaps filled to computing several metric values (length, number, length variation, N50, gap length). This software was executed in two separated runs, one for independent evaluation and other with to compare our genomes with closest reference, *M. zebra* genome. The program was executed for eukaryotic genes comparison. We therefore, evaluated assembly using two common quality measures: the contig N50 length, the assembly size. Finally, the B- scaffold level genome assembly was selected for future analysis and used as a reference.

Structural annotation of genes

The GeneMark-ES (Lukashin and Borodovsky, 1998) was used for eukaryotic ab-initio gene prediction, and identification of a set of 453 core genes that are supposed to be highly conserved in all eukaryotes, using CEGMA pipeline (version 2.4) (Parra et al. 2007).

For identification of protein coding genes in assembled genome, We used three approaches: homology-based, *de novo* and transcript sequences-based by using a pipeline MAKER v2.31.8 (Cantarel et al. 2008); we used repetitive elements of custom libraries of fish and Metazoan elements, *Danio rerio* CDS and proteome files from (http://www.ensembl.org/Danio_rerio), *A. latifasciata* assembled transcriptomes accessed from (http://sacibase.ibb.unesp.br/), gene prediction based on Lamprey training set and Blastn to NCBI database (National Center for Biotechnology Information – http://www.ncbi.nlm.nih.gov/).

Genome diversity and genome structural variations analysis

The Illumina reads generated for the different *A. latifasciata* genomes (males and females with 0B, 1B and 2B chromosomes) were first filtered for alignments by FASTX (http://hannonlab.cshl.edu/fastx_toolkit/) tool kit (-q 20, -p 80 parameters). The trim reads were aligned against our draft B- assembly as a reference using Bowtie2 (Langmead and Steven, 2012) --very-sensitive" option. Nucleotide polymorphism was identified using SAMtools tool to search for genome variations among the samples. Nucleotide polymorphism was identified using SAMtools (http://samtools.sourceforge.net/) to search for genome variations among the samples. The output files (VCF format) were subjected to VCFtools (vcf-stats and vcf-compare) (http://vcftools.sourceforge.net/) for statistical analysis to discover the frequency of single nucleotide polymorphism (SNPs), insertion/deletion (INDELs) and to compare these factors among indiviuals to find out the shared, unique or B related population polymorphism. Filtration was carried out to eliminate the lower quality ($Q \leq 20$ and $DP \geq 100$) SNPs and INDELs using vcffilter (https://github.com/vcflib/). Parameters for each step of this analysis were established according to the standard requirements. Similar approach was followed to detect polymorphisms in genic sequences (CDS, exons and introns) of *A. latifasciata* B+ and B- genomes aligned against *de novo* transcriptome assembly (Marques et al. unpublished).

The structural variations (translocations) were detected in 1B sequencing data by Delly (Rausch et al. 2012). Delly integrates short insert paired-ends and split-read alignments to accurately delineate genomic rearrangements throughout the genome. This pipeline was applied to both B+ and B- reads in Bam format (generated using bowtie tool) considered as sample and

control respectively. *Denovo* genome assembly of *A. latifasciata* was utilized as reference to locate the variations on the scaffolds regions. The detected translocations (breakpoints) were visualized by ClicO (Cheong et al. 2015), an online web-service based on Circos (Martin et al. 2009). To uncover nature of these transclocation, We extracted randemly a few of these regions (515-520 MB) and subjected to NCBI Blast (https://blast.ncbi.nlm.nih.gov) for annotation.

More specifically, structural variations such as deletions, insertions, transversions, inversions and duplications in genomic regions related to B chromosome (B block: B+ reads having higher coverage than B- reads) were analyzed by inGAP-sv tool (Qi et al. 2011). InGAP-sv detects the structural variations (SVs) on the basis of paired end mapped reads pattern and coverage of depth strategies. We applied this pipeline to B+ sam file generated using BWA (Li and Durbin, 2009) and *A. latifasciata* genome was used as a reference. After the SAM file was loaded into inGAP-sv, user-defined 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 et al (2011).

3.8. Analysis of B localized sequences

The nucleotide sequences of several previously identified B chromosomes genes of vertebrates (Makunin et al. 2014) supplementary (Table 1) were retrieved from the NCBI database (https://www.ncbi.nlm.nih.gov/). Consensus sequences constructions were obtained using Geneious v. 4.8.5 software (Drummond et al. 2009) for genes with more than one sequence available. All final sequences were used as queries against the *A. latifasciata* genome in a standard blastn search. Number of hits, E values and percent identity were considered to further proceed with B related analysis. This analysis confirmed the existence of *Ihhb* (Indian Hedgehog B gene) and 45S rRNA (18S ribosomal RNA, internal transcribed spacer 1, 5.8S ribosomal RNA, internal transcribed spacer 2, and 28S ribosomal RNA) gene sequences in *A. latifasciata* genome while the remaining genes were eliminated from the project because of partial or complete absence. The generated Illumina high coverage reads of all B- (0B male and 0B female) and B+ (1B Males, 1B Female and 2B Male) samples were aligned against both reference genes 45S rRNA and *Ihhb* of the cichlids *Oreochromis aureus* and *Lithocromis rubripinnis* respectively, using paired-end mode of Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml) with the –very sensitive option. The output aligned files were converted to binary format and indexed using samtools. Each file was normalized using RPKM package of deeptools (Ramírez et al. 2014) to fix bias of initial coverage.

These files were then visualized by integrated genome browser IGB (http://bioviz.org/igb/index.html) to compare coverage of both genes in B- and B+ samples. SNPs at different sites of nucleotide reads were detected. According to Marques et al. (unpublished) a total of 36 libraries were generated from 18 samples of transcriptomes of *A. latifasciata*. These transcriptomes were divided into triplicates of gonads, brain and muscle of male and female individuals. The uploaded transcriptomics data (http://sacibase.ibb.unesp.br/) and genomics data (aligned files) were visualized and screened. We did BLAST alignment of the genes against transcriptome assembly to locate them on specific scaffolds. The B specific SNPs were searched for simultaneous viewing all tracks of available data against transcriptome assembly as a reference.

Primer design, probes construction and FISH mapping of *Ihhb* and 45S rRNA genes

Primers listed in the supplementary (Table 2) were constructed using PrimerQuest (http://www.idtdna.com/primerquest/home/index) tool, checked for specificity by Primer-Blast (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) and evaluated by Primer Stat (http://www.bioinformatics.org/sms2/pcr_primer_stats.html). Genomic DNA was subjected to polymerase chains reaction (PCR) to obtain DNA segments to be used as probes for fluorescent in situ hybridization (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 Dynamic Terminator 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 database to check for similarities to deposited sequences of corresponding genes. Probes were labelled with digoxigenin-11-dUTP (Roche Applied Science) and the signal was detected with anti-digoxigenin-rhodamine (Roche Applied Science). FISH was performed using the protocol described by Pinkel et al. (1986) 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 and were optimized for brightness and contrast using Adobe Photoshop CS2.

Results

De novo assembly and gene predictions

A total of seven B- and B+ samples were sequenced using Illumina HiSeq and Miseq platforms respectively. Each sample has different coverage of the corresponding genomes over the entire length (848,776,495bp) of the *M. zebra* genome. Each sample has been given a name such as; “M1-5” and “F1-2” for males and females, respectively; 0-2 for B-, 1B and 2B, respectivelly; S1-3 for different individuals of male samples (Table 1).

Table.1. Illumina data obtained for *A latifasciata* including mapped over *M. zebra* genome.

Samples	Read length	Coverage	Coverage after filtration	Total reads	Remained reads after filtration	Reference
M1-0B	101	47.7x	38x	401,017,570	323226972(80%)	Valente et al. 2014
F1-0B	191	75.5x	42.5x	337,349,994	190214688(56%)	Present work
M2-1B	35-250	2x	1.6x	716,030,6	5911265 (82%)	Present work
M3-1B	101	16.8x	13.4x	143,441,264	114064786(79%)	Present work
M4-1B	101	43.1x	34.8x	366,602,572	296161988(80%)	Present work
F2-1B	191	70.2x	40.1x	313,818,884	179286280(57%)	Present work
M5-2B	101	43.6x	30x	306,823,512	254993955 (83%)	Valente et al. 2014

The raw data was then filtered and illlumina adapters were trimmed and a total of filtrated 513,441,660 (77.05%) of B- reads with 80.5x coverage and 850,418,274 (74.7%) with 120x coverage of B+ reads were remained after filtration (Table 1). The B- assembling to accomplish primary assembly of *A. latifasciata* genome having 771,316,069 bp, which indicates that 84% (This value corresponds to the total used reads by assembler during assembly) of genome was captured in scaffolds. The *de novo* assembly yielded 218,259 scaffolds with N50 of 18.640 kb (Supplementary Table 3) being the longest one with 23.36 kb.

A. latifasciata draft genome has correspond to the *M. zebra* genome consisting of 848,776,495 bp cumulative length, GC contents of 40.5% (Supplementary Figure 1a,b). About 75.329% of *M. zebra* sequences had a significant hit against our performed assembly. The scaffold assembling gave a total of 3,998 gaps per 100-kb of the assembled genome filled to 2,076. The comparative analysis indicated that there was an improvement in assembly after the gaps filling.

Total length of genome slightly increased up to 774,140,944 bp. Additional statistics is given in Supplementary Table 3.

We used the CEGMA pipeline (Parra et al. 2007) to evaluate the completeness of our assembly. The total number of complete core eukaryotic genes (CEGs) are 183 with the percentage of 73%, and number of partially complete CEGs are with the percentage of 94% supplementary data (Table 4). The ab-initio gene model predicted 71,917 genes of eukaryotes, of them 23,391 are protein coding genes. The 23,391 annotated genes in the A. latifasciata genome contain 204,656 exons and 177,219 with longest gene size is 66982 bp. On average, there are 7 exons and 6 introns per gene. Additional information (size, number and length) of major structure genome componets are in (Figure 1; Supplementary Data Table 5). All the structural annotation has been uploaded to Sacibase.

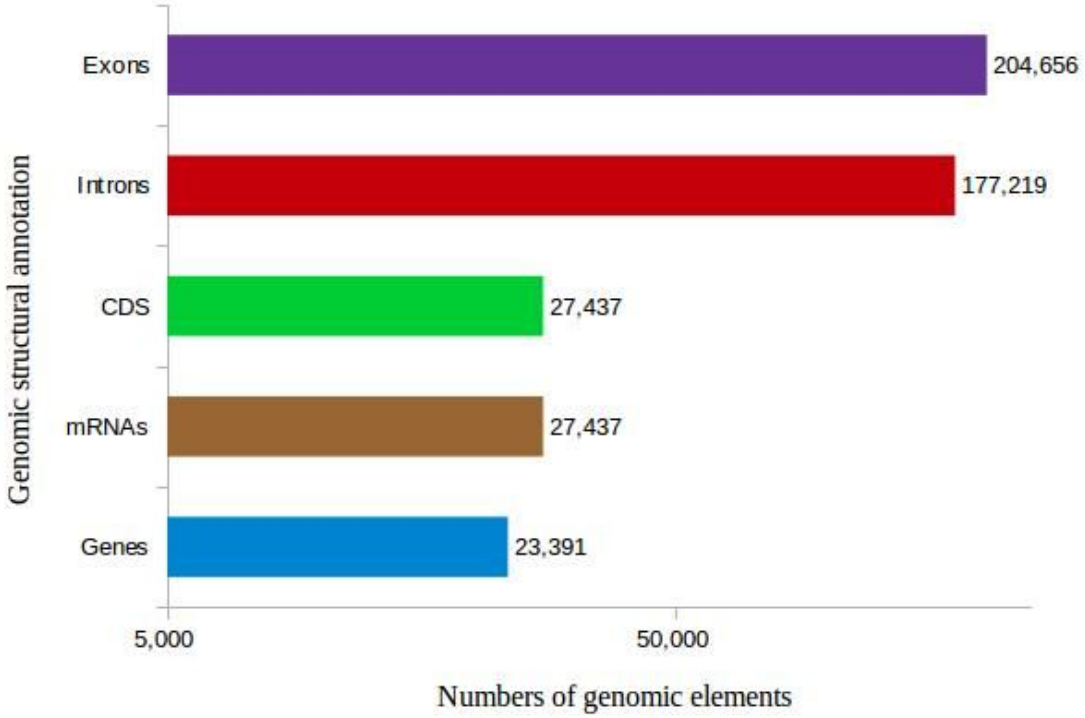


Figure 1. Structural annotations of *A. latifasciata* genome. The bar graph shows the number of major genomic structural component.

Genomic diversity analysis

In the genomes of six individuals (B- and B+), it was identified total 2,395,658 raw SNPs and 888,060 INDELs with respect to reference B- assembly. All the individuals carried higher number of SNPs than INDELs as given in (Supplementary Figure 2). These SNPs were forwarded to filtration and detected total of 17,875 high quality SNPs in all genomes (Figure 2a). Out of these total SNPs, the genome of 1B-female called significantly high number of SNPs (5,181). In case of B+ individuals signaled 11,978 SNPs against B- reference genome. Although we mapped the same reads of B- samples (male and female) against the same reference, still there appeared many SNPs of 0B male and 0B female 3,800 and 2,097 respectively. Furthermore, Comparative analysis of different SNPs combinations from five different samples confirm unique and shared SNPs variation (Figure 2b). We found that 1B-female has shared a higher number of SNPs with all other five individuals genomes and 2,936 SNPs are shared among all individual genomes (Figure 2b). Furthermore, we made use of genic sequences of *Astatotilapia latifasciata* B+ and B- genomes to determine the mutation frequency.

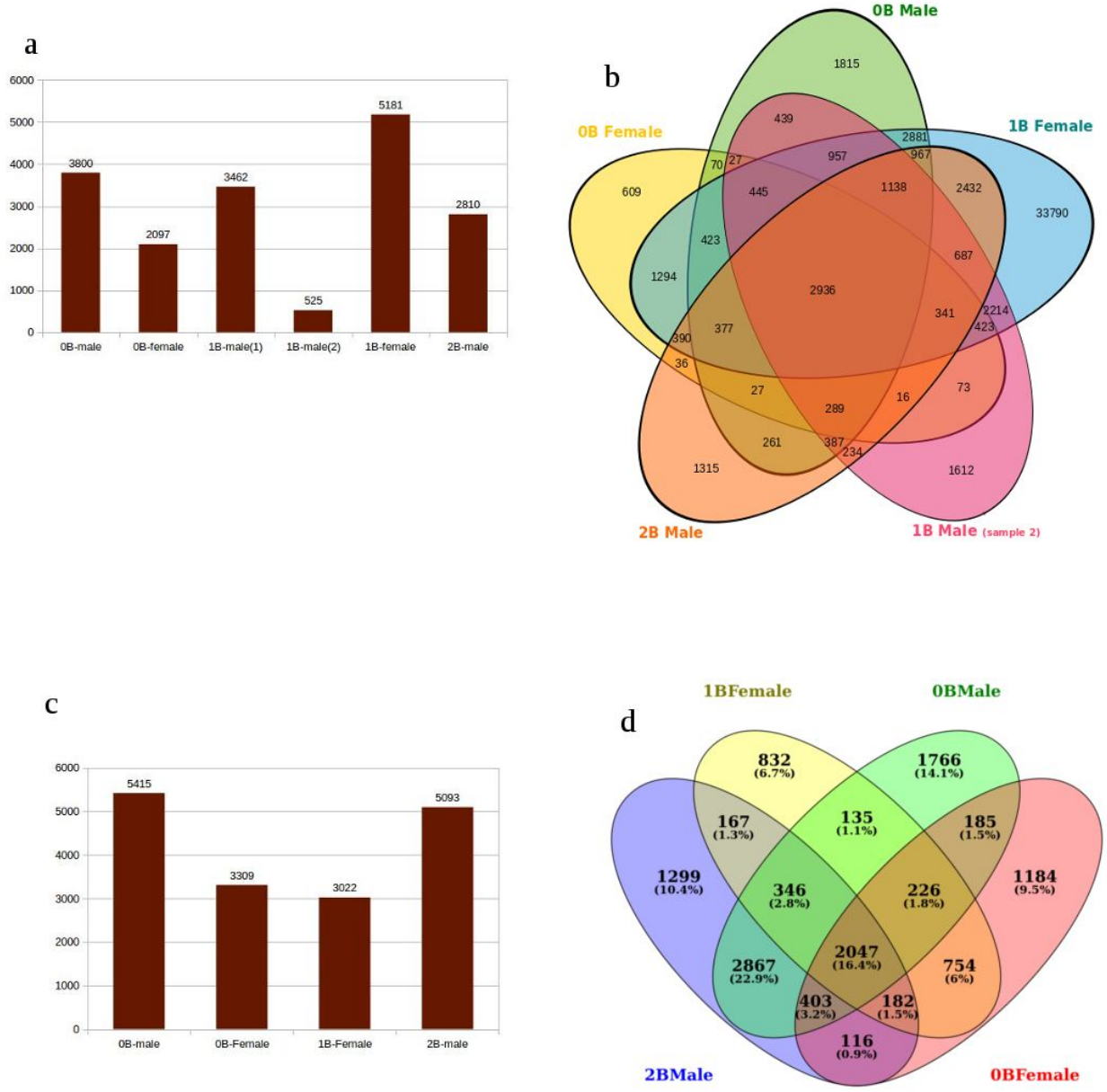


Figure 2. Population genomic polymorphism.Venn diagram using different colors congruent ellipses, each ellipse showing different individual and the overlapping curves among and between samples representing shared SNPs while the points outside the boundary represents unique SNPs. The bar graphs show the frequency of high quality filtered SNPs. Figures (a) and (c) represent the diversity analysis of genomic data aligned against our assembled genome, (b) and (d) represent genomic data aligned against the transcriptome assembly.

To analyze the differences in SNP frequencies of genic sequences, which should reflect the presence or absence of selective pressure (Martis et al. 2012). This analysis was restricted to only 4 samples (two from each B+ and B- both male and female) in (Figure 2c). We found that all four

samples shared a total number of 2,047 SNPs (16.4%) and 0B male recorded the highest 1,766 (14.1%) unique SNPs as represented in the venn diagram (Figure 2d). Our results (Figure 2c) suggest the males were comparatively under lower SNPs rate than female with no evident effect of B sequences.

Whole Genome rearrangement and structural variations

Structural variations of the genome involve kilobase to megabase-sized deletions, duplications, insertions, inversions, and complex combinations of rearrangements (Korbel et al. 2007). A total of 625 interchromosomal translocations (breakpoints) were detected in the whole genome of 1B individual (Figure 3). We annotated a few of these regions with identified translocations, most of them were fragmented genes and non coding RNAs (Supplementary Figures 3, 4). Genomic regions related to B chromosome or B blocks was also subjected to reads-orientation based on SVs detection method. Interestingly, we found duplications, insertions and inversions at different sites in B blocks (Figure 4). These results contribute to understand the mechanism of evolutionary process of B chromosome. However, these polymorphic events might also be noticed in B- genomes in different number and pattern than B+ genomes. Several sequence duplications detected in the B block indicated that these sequences from A chromosome accumulated on B chromosome due to frequent duplication events.

Sequence analysis and physical mapping of B sequences

Blastn result summarized the highest identity and number of hits for *Ihhb* and 45S rRNA genes in complete set of B chromosome genes of vertebrates confirming their existence in the genome of *A. latifaciata* (Supplementary Table 5). The bowtie2 alignments of B + and B- Illumina reads from *A. latifasciata* shows higher coverage from the B+ than the B- samples for both genes (Figures 5a). The higher coverage of these genes in B+ sequence data shows their duplicated copies presence on B chromosomes. The FISH mapping (the probes had 98-99% identity with *Ihhb* and 45S rRNA genes of different vertebrates) revealed extensive markings of *Ihhb* over the two B chromosomes in 2B metaphases, representing its repetitive nature. The duplicated copies of *Ihhb* gene emerged as a major structural component of B chromosomes in FISH results (Figure 6). The available RNA-seq data was analyzed for both genes, the absence of *Ihhb* gene transcripts indicated

that it was not transcribed. We constructed separate probes for 18S, 5.8S and 28S rRNAs to investigate if the complete cluster of 45S rRNA has moved from A complement to B chromosome. Positive sites of 18S rRNA were observed over the pericentromeric and subtelomeric areas of the B chromosome and on pericentromeric regions of some A chromosomes. The 5.8S rRNA probe were marked on pericentomeric regions of B chromosome and in telomeric and subtelomeric regions of autosomes. The 28S rRNA produced signals on telomeric regions and centromeric regions of A chromosomes and that of B chromosomes. The results of mapping all three probes on B chromosome provide evidence that Bs have incorporated the entire 45S rRNA cluster from the A complement set. While transcript analysis indicated that there were some transcript found for 45S rRNS gene.

We also conducted a survey to screen SNPs, INDELs at both genomics and transcriptomics level. The *Ihhb* gene has encountered few B specific SNPs and INDELs. A high number of population SNPs were found in 45S rRNA cluster at genomic level but these population SNPs were located in spacer Dna, the 18S, 28 and 5.8 regions were no SNPs found viewed in Figure 5b,c and Supplementary data Figure 6 & 7.

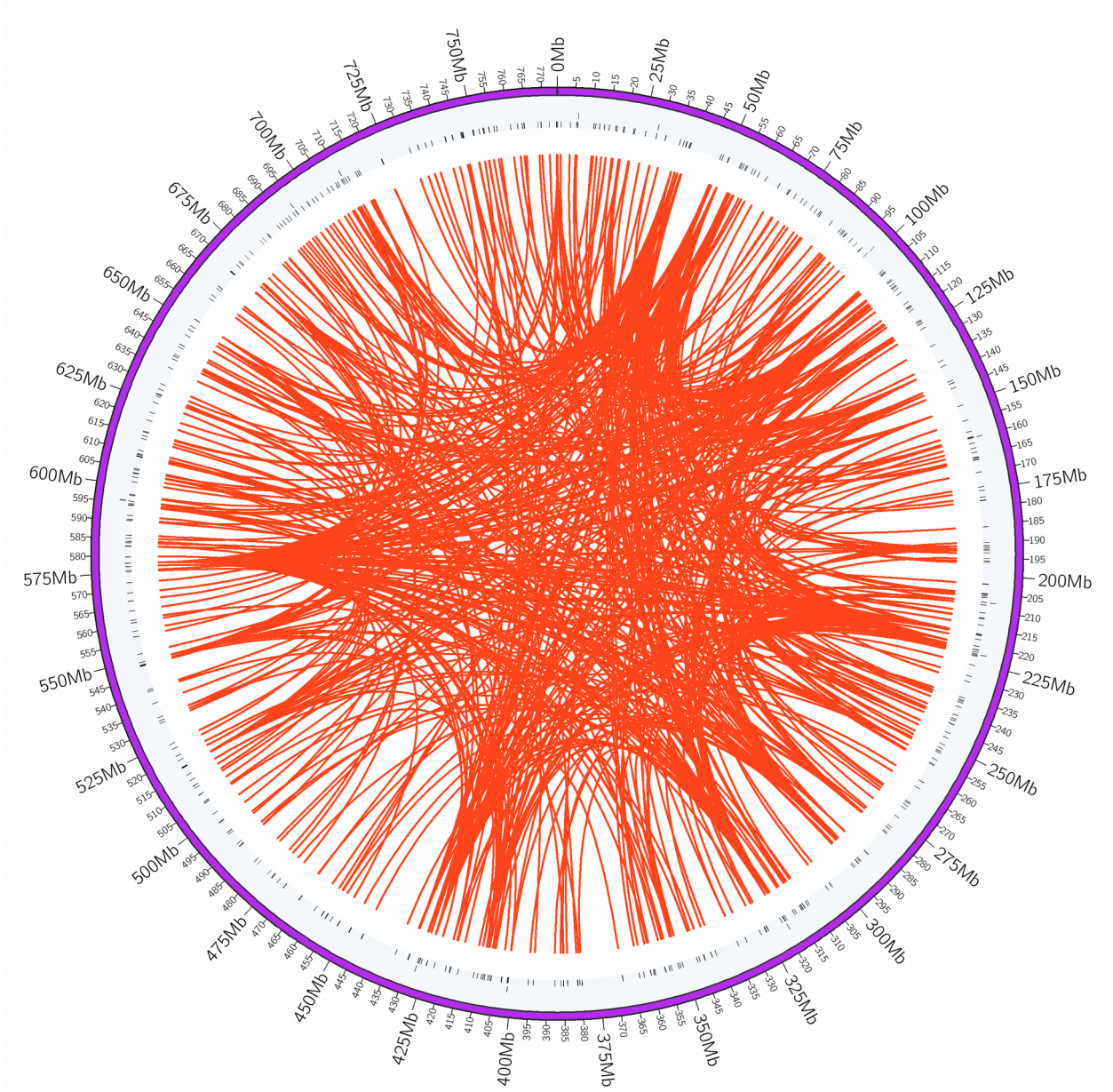


Figure 3. Circos visualization of whole genome rearrangements of B+ genome. The outermost purple ring represents the *A. latifaciata de novo* assembled genome in Mega bases. The second fragmented ring represents those scaffolds with B+ rearrangement. The red links show genomic regions of B+ reads with Intra-chromosomal rearrangements(translocations) or breakpoints.

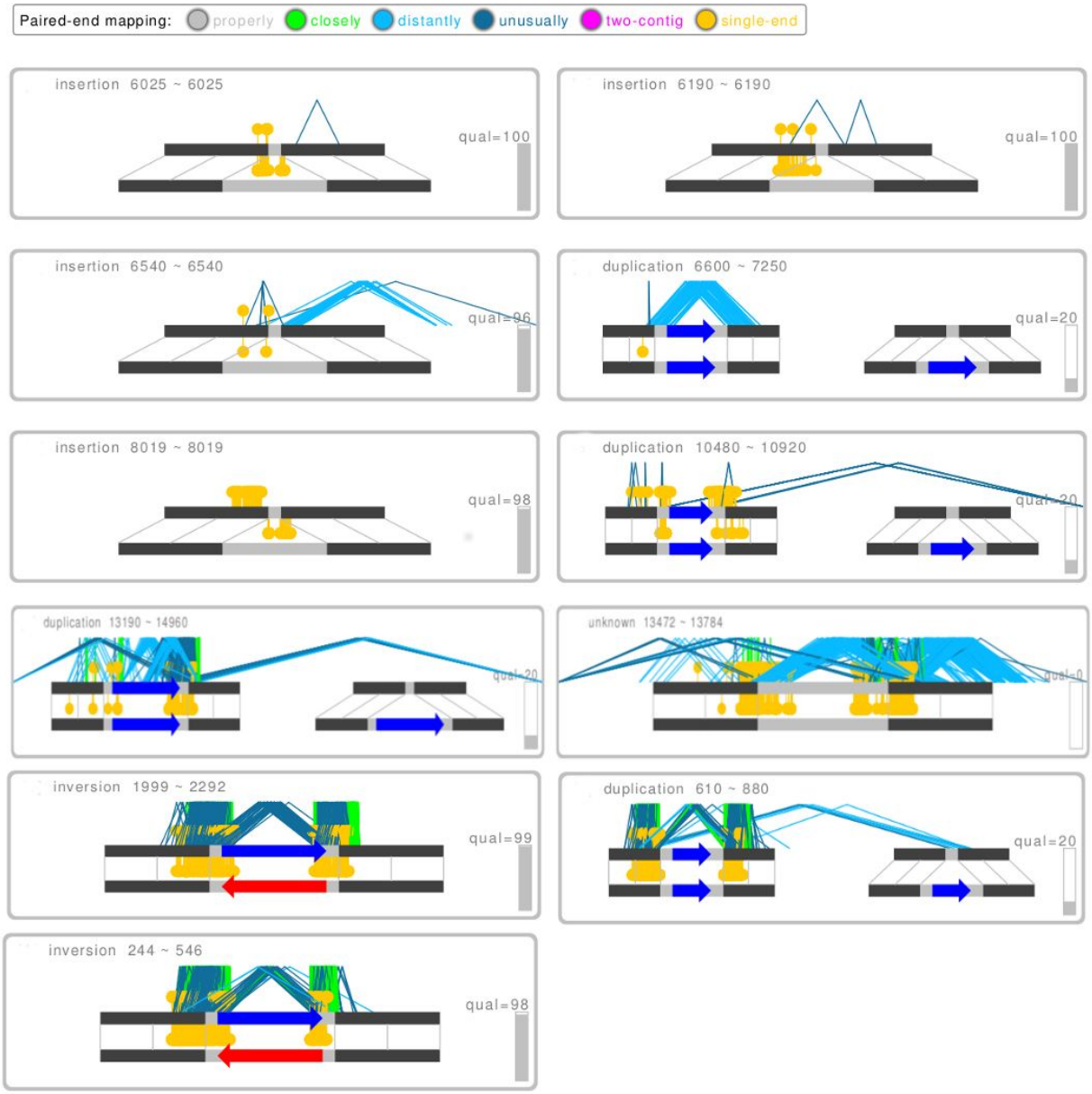


Figure 4. Different structural variations including duplications, deletions, inversions and transversions in a randomly selected B block (lower line) against the reference genome (upper line) Illustrations of paired end mapping (PEM) patterns for different types of SVs are described as follow: Grey links indicate normally mapped read pairs with proper read orientation and distance. Light blue links represent read pairs with proper read orientation but longer distance, which may indicate a deletion event in the query sequence. Green links represent read pairs with proper orientation but shorter distance, and thus indicate an insertion. Dark blue links show read pairs with abnormal orientation, in which paired ends are mapped to the wrong strand(s). Yellow lines indicate single-end mapped reads (SE reads), in which only one of the paired reads is mapped. For a small insertion (< the insert size), a fraction of paired reads (in green) that span the insertion is mapped too closely in the reference. The insertion is surrounded by a set of single-end mapped reads (in yellow). A translocation is represented by two sets of distantly mapped pairs and one set of inverted mapped pair (in dark blue). 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.

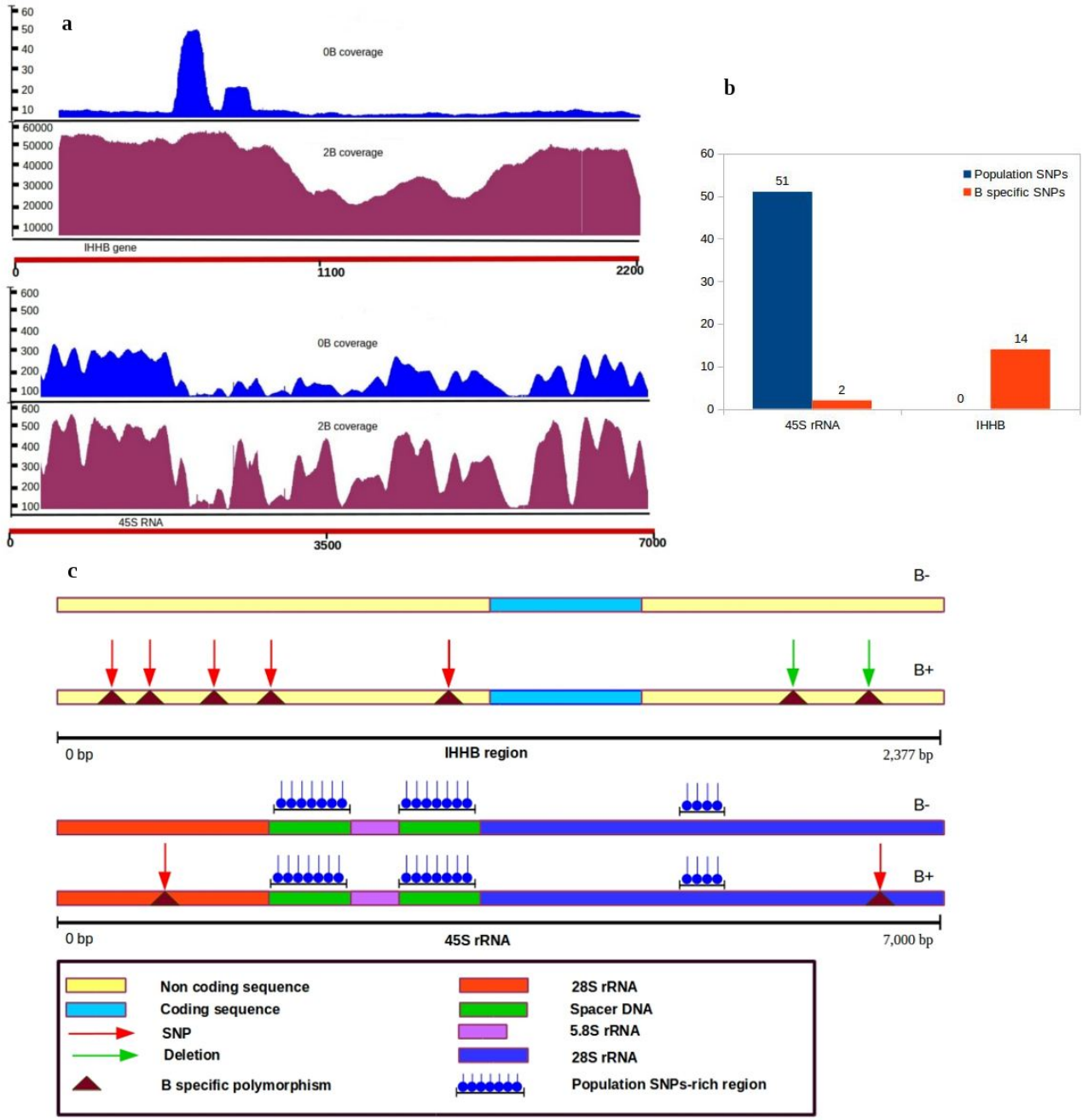


Figure 5. Sequence analysis of B related genes. (a) Reads coverage of *Ihhb* and 45S rRNA gene sequences. Notice the scale bar on left to differentiate the coverage rate between 0B and 2B samples. (b) Number of population and B specific SNPs are shown as red and blue respectively in the bars for both 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.

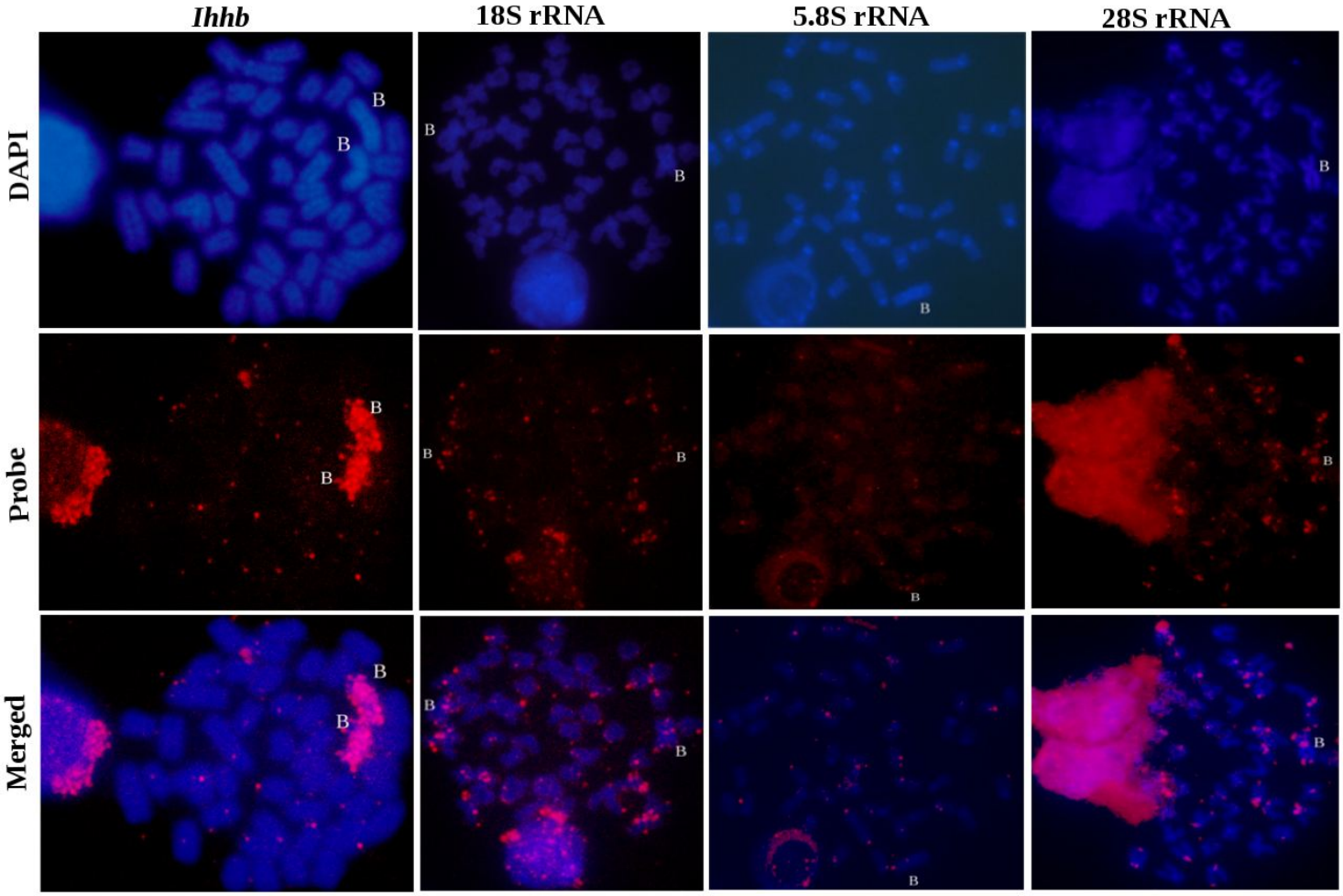


Figure 6. FISH mapping of *Ihhb* and 45S rRNA genes. Images of metaphases stained with DAPI, probes and merged are shown for each sequence. Note the signals markings on the B chromosomes.

Discussion

Our analysis bring several contributions including: 1) generation of a *A. latifasciata* reference genome; 2) screening the high coverage sequencing data of different individuals for population polymorphism, genetic diversity and to discover B specific structural variations; 3) searching different B-linked genes in the *de novo* genome of *A. latifaciata* and show a relative copy number of selected genes between B+ and B- samples by coverage analysis; 4) physically mapping these genes on B chromosomes and validate the sequence analysis.

De novo scaffolding of *A. latifasciata* genome

In the present study, we obtained a high deep dataset of B- and B+ individuals for genome assembling of *A. latifasciata*. The evolutionary aspects of fish genomes, especially cichlids, are being extensively studied (Brawand et al. 2014). Our motivation to sequence the first *A. latifasciata* genome stemmed from the fact that the species is rapidly becoming important for B chromosome evolution and good model for genetic diversity with other cichlid species. The comparison of our work with other african cichlid fish assemblies presented by the International Cichlid Genome Consortium project (Brawand et al. 2014), in terms of N50, genome size, genes number and % of GC in Table 2, showed corresponding values with other cichlids. It indicated that our reconstructed genomes can be assumed as the possible *de novo* draft assembly of our model species performed. According to Assemblathon 2 project (Bradnam et al. 2013) the results should not be only relied on the basis of a single assembly and parameter, therefore we also reconstructed our genomes by SOAPdeno2 (Luo et al. 2012) assembler to assure much better results. However, we neglected the genome given as output by SOAPdenovo2 because of its lower accuracy as compared to Velvet.

Table 2. Comparison of the current assembly to others african cichlid fish assemblies presented by the International Cichlid Genome Consortium project (Brawand et al. 2014).

Locality Species	African Riverine		Lake Victoria	Lake Tanganyika	Lake Malawi			
	O. niloticus	*A. burtoni*	*P. nyererei*	*N. bricherdi*	*M. zebra*	*Latimeria chalumnae*	*Gasteosteus aculeatus*	*A. latifasciata (Our draft assembly)*
Estimated genome size(Gb)	1.01	0.923	0.993	0.98	0.946	2.85	0.53	0.771
Scaffold N50(kb)	2.8	1.2	2.5	4.4	3.7	0.92	10.8	1.8
% GC content	40.42	40.51	40.6	40.44	40.54	41.15	44.6	40.5
Protein coding gene count	24,559	23,436	20,611	20,119	21,673	19,033	20,787	23, 391

Genomic diversity analysis between B- and B+ individual genomes

The Illumina reads coverage obtained for several samples (including males and females and B+ and B- genotypes) were useful to generate a population genomics scenario for the species, and made it possible to analyze the variants specific to B chromosomes and comparative analysis of genomic diversity between B+ and B- genomes. Previous studies have utilized the same strategy to to screen whole genome for distribution of SNPs and INDELs using illumina sequencing data and proposed this method to be useful in understanding genomic diversity (Loh et al. 2008; Gudbjartsson et al. 2015). Our study is in attempt to deal with investigation of the similarities and differences between the two genomes (B+ and B-) of the same species using the same approach as described in aforementioned reports. B chromosome studies might be interesting to analyze different polymorphism such as SNPs, INDELs, duplications and rearrangements. B chromosome is originated by duplications and expansion of DNA copies from A chromosomal set (Kellis et al. 2004; Martis et al. 2012; Valente et al. 2014). We explored the genomics data in order to present a unique brief description about the whole genome polymorphism concerned with B chromosomes, considering that the hypothesis that B+ genome would probably have accumulated many polymorphic DNA sites as compared to B- genome. Our results indicated a higher polymorphism (SNPs and INDELs) in B+ female as compared to B- female at genomic level. However, in other B+ individuals, no association of B chromosome was detected to effect the frequency of SNPs. We applied a similar approach suggested by (Martis et al. (2012) to find out the presence or absence of selective pressure on the genic sequences. Our analysis concluded no effect either on selective

pressure with respect to the B chromosome existence in the genome (B is not related to the higher frequency of SNPs in *A. latifasciata*). This is in contrast to the findings of Martis et al (2012), which identified a higher frequency of SNPs due to the presence of B chromosome signifying the lower selective pressure on B genic sequences. For better and precise understanding the relationship of polymorphism and variation with supernumerary chromosomes, we needed a more specific analysis, which rather focus on individuals DNA sequences or genes. This way we carried out an analysis specific to *Ihhb* gene and interestingly found B specific SNPs as discussed laterin next section. The whole genome diversity analysis was a novel strategy to study the B related and population based polymorphism in *A. latifasciata* and has never been performed. This approach was aimed to generalize an overall view about the mutual similarities and differences between the B+ and B- genomes.

Exploring duplicates of *Ihhb* and 45S rRNA genes

The findings of our study are helpful to add another evidence in explaining chromosomal origin of the supernumerary elements. As previously described by numerous reports that the B is derived from autosomes (Alfenito and Birchler 1993; Houben et al. 2001; Peng et al. 2005; Martis et al. 2012), it has been proposed that all autosomes have contributed sequences to the B chromosome of *A. latifasciata* through gene duplication (Valente et al. 2014). Among the genes discovered on B chromosome of vertebrates, there is a morphogenesis-related gene named indian hedgehog b (*Ihhb*) gene detected in the genome of the cichlid *Lithocromis rubripinnis* (Yoshida et al. 2011). This study found more than 40 copies of the *Ihhb* paralogs on B chromosomes and a single copy of *Ihhb* ortholog on the A chromosomes by qPCR in *L. rubripinnis*. Our genome analysis of this gene in *A. latifasciata* followed by FISH mapping provide a novelty of its organization and duplication on B chromosome. 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 (Friebe et al. 1995; Camacho, 2005; Houben et al. 2013). The presence of 18S rRNA gene sequences in the B chromosomes suggests a possible origin from the A chromosomes, which carries rRNA genes. The B chromosomes of *A. latifasciata* have accumulated 18S rRNA gene sequences on the subtelomeric and pericentromeric regions (Fantinatti et al. 2011). Based on previous description of ribosomal genes there rises a hypothesis that the complete 45S ribosomal DNA (rDNA) cluster would have started amplifying its copies from A complement and moved to B chromosomes during initial stages

of its evolution and ultimately lead to formation of 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. Our FISH mapping results of 18S rRNA and 28S rRNA confirmed the previous findings made by Poletto et al. (2010) and Fantinatti et al. (2011) . Further, the 5.8S rRNA marks on B chromosome enabled our hypothesis about the movement of whole cluster from A to B chromosome to be correct.

A search for the transcripts of both *Ihhb* and 45S rRNA genes was performed in RNA-seq data of different tissues including brain, gonads and muscle to check if their B copies are expressed or inactive. We find transcripts of 45S rRNA in some tissues, which prove that the cluster is transcribed in *A. latifasciata* genome. Similarly, this analysis suggests that *Ihhb* gene is not transcribed and evolved as a pseudogene as a result of its series of sequence duplication. Previous studies also confirmed the occurence of pseudogenization of B-located gene-like fragments with only 15% of these pseudogenes transcribed in rye plant (Banaei-Moghaddam et al. 2013).

Genome-wide rearrangements, structural variations and Polymorphisms in B+ genome.

One of the key aspects in the B chromosomes studies is to understand the nature of genetic polymorphisms/variations and their evolutionary importance in the origin of B chromosome (Martis et al. 2012; Muñoz-Pajares et al. 2011). Although, most of the genomic studies are focused to identify single-nucleotide polymorphisms (SNPs) or small insertion deletion (INDELs), different types of structural variations (SVs) including complex rearrangements, duplications, inversions, large deletions and insertions can also be detected using a variety of computational approaches (Keane et al. 2014). Indeed, the SVs play a major role in evolutionary processes such as adaptation (Iskow et al. 2012) and speciation (Noor et al. 2001). A study reported by Fan and Meyer (2014) gives a detailed overview of different kinds of genomic variation across four recently diverged cichlid lineages and postulate on the correspondence of SVs for one of the largest adaptive radiations in vertebrates. A remarkable contribution to understand the evolutionary biology of chromosome is achieved by revelation of SVs in many 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). There are several reports that superpose B and sex chromosomes, with evidences of sex chromosomes origin from domesticated B chromosomes (Martins et al. 2011) and also the B presence influencing the sex development

(Yoshida et al. 2011). Therefore, B and sex chromosomes have become one of the major focuses in chromosome biology studies because they offer an excellent model to investigate mechanisms of chromosome changes during evolution that, in some cases, are incipient. In this sense, we have carried out the whole genome rearrangement analysis to understand the molecular mechanism of B chromosome evolution.

In addition to the classical application of NGS to recover whole genomes, the paired-end approach can locate chromosomal breakpoints (Chen et al. 2010). Several reports of polymorphism, both structural and numerical, were published among B containing species. Frequent structural and numerical variations noted among B containing species possibly reflects that they do not possess any constrain upon the cellular structure and organization as do the A chromosome type (Datta et. al 2016). However very little is known about the polymorphism of Bs at molecular or gene level. We have found a large number of chromosomal breakpoints such as translocations extensively distributed throughout the whole genome of B+ individual of *A. latifasciata*. The detection of these genomic patterns are important in studying the molecular basis of evolutionary processes involved in the formation of B chromosomes. To understand this mechanism, we carried out a refined analysis to a B block (genomic region present on B chromosome) in order to determine the association of SVs with B chromosomes. The discovery of these B associated SVs provide an evidence of how genetic changes in these sequences might have occurred and originated the B chromosome. Our study is an initiative to uncover B specific and population SNPs and INDELs of both *Ihhb* and 45S rRNA genes. Our results suggest that *Ihhb* gene remains highly conserved among different vertebrates species as no population-wise polymorphism was detected. The higher number of B specific SNPs and INDELs found in *Ihhb* gene and absence of transcripts revealed such copies on Bs have become pseudogenes. The higher rate of B-specific SNPs found in *Ihhb* gene suggests how duplicated copies of this gene accumulated on B chromosomes. We assume that these polymorphisms might have played a key role in evolutionary course of duplication event. However we may not conclude the clear findings since *Ihhb* genes is linked with morphogenesis and should be studied at embryonic level to draw a better concept of its expression. The Hedgehog (Hh) gene family encodes certain proteins that play a key role in embryogenesis. The *Ihhb* gene expresses itself at the time of specification in zebrafish embryos (Ingham et al. 2001; Chung et al. 2013). Several papers have emphasized the evolutionary importance of hedgehog gene family and outlined the process of duplication events related to the members of this gene family including *Ihhb* in many vertebrates (Kumar et al. 1996, Zardoya et al. 1996; Ekker et al. 1995; Holland, 1994; Carroll, 1995). 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. The identification of these B specific SNPs in the duplicated copies of *Ihhb* gene is biologically significant to understand the molecular structural of B chromosome. During duplication, modifications such as insertion, deletions and mutation occur in the gene sequence, as in *Ihhb* in this case, therefore we term *Ihhb* as a "non-processed or duplicated pseudogene". Pseudogenes are vital to give an understanding of how genomic DNA has been changed without such evolutionary pressure and can be used as a model to determine the rate of polymorphism. Our inter-specific analyses of 45S rRNA gene cluster identified spacer DNA enriched with many SNPs and conserved coding region; is consistent with Seo et al. 2013 study to sequence polymorphisms in ribosomal RNA genes and variations in chromosomal loci of *Oenothera odorata* and *O. laciniata.* From the inter-specific variation, spacer regions of species were distant from each other, whereas the remaining coding regions were highly conserved.

Conclusion

The Illumina reads coverage obtained for several samples (including males and females with B+ and B- genotypes) will be useful to generate a population genomics scenario for the species, including the B chromosome polymorphism. The development of the present research provides the first *de novo* genome assembly of *A. latifasciata*. This genome is to be explored as a reference for future analysis involving evolutionary and applied genomics. We have identified nucleotide polymorphism to search for genome variations among B+ and B- individuals. We observed a high scale genomic diversity in all analyzed genomes showing a high rate of population polymorphism. This analysis suggests that the males were comparatively under lower selective pressure than female with no evident effect of B sequences. We also identified that duplication events generate a higher number of copies for the Indian Hedgehog b (*Ihhb*) gene and 45S rRNA gene-cluster in B+ genome. Single nucleotide polymorphism for the corresponding genes were detected in B+ sequencing data. The B chromosome does not influence the frequency of SNPs in the genome, however the presence of B specific polymorphisms in duplicated B copies of *Ihhb* gene are vital in its duplication. Structural variations associated with B chromosomes were reported which provoke a hypothesis that supernumerary arise due to the series of these polymorphic events.

Acknowledgement

This work was supported by grants from Sao Paulo Research Foundation (FAPESP: process number: 2014/17683-6).

Assembly and Data files

Assembly and sequencing data of PCR products will be made available in NCBI data base soon after the acceptance of paper.

Supplementary Files

Additional files to support our results are organised as the attached PDF file.

References

(include in thesis, will be organised according to Journal format)

7. Thesis conclusion

- The Illumina reads coverage obtained for several samples (including males and females with B+ and B- genotypes) will be useful to generate a population genomics scenario for the species, including the B chromosome polymorphism.
- The development of the present research provides the first *de novo*, both B+ and B- genome, assemblies of *A. latifasciata.* These genomes are to be explored as a reference for future analysis involving evolutionary and applied genomics. Both genomes also open the perspective to advance understanding the role of B chromosomes in evolution of African cichlids
- We have identified nucleotide polymorphism to search for genome variations among B+ and B- individuals. We observed a high scale genomic diversity in all analyzed genomes showing a high rate of population polymorphism. This analysis suggests that the males were comparatively under lower selective pressure than female with no evident effect of B sequences.
- We also identified that duplication events generate a higher number of copies for the Indian Hedgehog b (*Ihhb*) gene and 45S rRNA gene-cluster in B+ genome. Single nucleotide polymorphism for the corresponding genes were detected in B+ sequencing data. The B chromosome does not influence the frequency of SNPs in the genome, however the presence of B specific polymorphisms in duplicated B copies of *Ihhb* gene are importenat in its duplication.
- Structural variations associated with B chromosomes were reported which provoke a hypothesis that supernumerary arise due to the series of these polymorphic events.

8. Recommendations

Our research has produced useful assemblies, containing a significant representation of their genes and overall genome structure. However, much improvement can be made by integration of PacBio and Illumina assemblies as mentioned in our initial proposal plan. We further suggest that more accurate assembly of B+ genome can resolve various problems faced to researchers in B chromosomes analysis. This can be achieved by generating long reads and mate pairs libraries using appropriate NGS platforms such as PacBio and Illumina respectively. This argument is also supported by a recent paper aimed to study the Y chromosome of *Drosophila melanogaster (*Carvalho et al. 2015) that accurately reconstructed a very challenging region of the Y chromosome due to the presence of very high identity of the repeats. Unfortunately PacBio sequencing is no longer under consideration for this project, considering higher costs and time consuming.

9. Supplementary materials

Supplementary Table 1. List of B related genes selected from different vertebrates with detail of BLAST result**.**

Gene name	Species name	Accession number	Total Gene length	No. of Blast hits in B+ genome	Alignment. Length of 1st blast hit	Identity %	E-value	Bit score
C-KIT	*Canis lupus familiaris*	Gene ID: 403811	82018 bp	32	131 bp	80.15	4e-23	120
45S rRNAs	*Oureochromiaureus*	GenBank: GU289229.1	6949 bp	210	1056 bp	99.43 %	0.0	1871
KDR (kinase insert domain receptor)	*Metriaclima zebra*	Gene ID: 796537	61182 bp	6211	1211 bp	74 %	0.0	771
LRIG1 (leucine-rich repeats and immunoglobulin-like domains 1)	*Canis lupus familiaris*	Gene ID: 484698	107294 bp	44	398 bp	70.35 %	1e-42	185
Ihhb (Indian hedgehog homolog b)	Lithocromis rubripinnis	GenBank: AB601494.1	2381 bp	565	640 bp	99.69 %	0.0	1142
Lysosomal alpha-mannosidase	*Metriaclima zebra*	Gene ID: 541519	17958 bp	755	201 bp	82.09 %	3e-48	201
VPS10 domain receptor protein SORCS 3 like	*Metriaclima zebra*	Gene ID: 100004230	86392 bp	5029	886 bp	73.59 %	3e-138	502
TNNI3K (TNNI3 Interacting Kinase)	*Capreolus Carpeolus*	GenBank: JN871289.1 to JN871291	212 bp	55	26 bp	92.31 %	0.23	39.2
FPGT (Fucose-1phosphate guanylyltransferase)	*Capreolus pygargus*	GenBank: JN871287, JN871288	1461 bp	0	0	0	0	0
Rnasel 2 (Ribonuclease-like 2)	*Metriaclima zebra*	Gene ID: 100003397	822 bp	80	36 bp	88.89 %	0.002	48.2
LRRIQ3 (Leucine-Rich Repeats And IQ Motif Containing 3	*Capreolus pygargus*	GenBank: JN871294, JN871295	220 bp	0	0	0	0	0

Supplementary Table 2. PCR primer list of *Ihhb* and 45Sr RNA cluster genes.

Genes	**Primers No**	**Forward primer**	**Reverse primer**	**Product length in bp**
Ihhb gene	1	CTCACCACAGAGAGCAAACA	ACCTTGCTTCTTACTTCCTGTG	568
28S rRNA Gene	1	CTGAGAAACGGCTACCACATC	GAACCTCCGACTTTCGTTCTT	608
5.8S rRNA Gene	1	ACTCTTAGCGGTGGATCACT	GCGTTCGAAGTGTCGATGAT	105
5.8S rRNA	2	ACTCTTAGCGGTGGATCACT	AAGGTGCGTTCGAAGTGT	110
28S rRNA Gene	1	CTGTAACGCAGGTGTCCTAAG	TATCCGAGACCAACCGAAGA	550
28S rRNA Gene	2	GCCAAATGCCTCGTCATCTA	CCGCTTTCACGGTCTGTATT	478

Table 3. QUAST evaluations statistic of *Astatotilapia latif*asciata assembly.

Statistic	Assembly	Assembly after gaps filled
# of contigs	413663	-
Total Length of contigs	741426874-bp	-
N50 of contigs	5603	-
GC (%) contigs	40.43	-
Largest contig	57508-bp	-
# of scaffolds (>=500bp)	91192	91144
# scaffold(>= 0 bp)	218259	218259
Total length of scaffolds(>= 0 bp)	771316069	774140944
Total length of scaffolds(>=500bp)	749397988	749419706
Largest scaffold	233669	234344
GC (%)	40.46	40.48
Scaffolds N50	18640	18659
Scaffolds NG50	15597	15712
Scaffolds N75	8759	8778
Scaffolds NG75	5409	5529
Scaffolds L50	10970	10988
Scaffolds LG50	14001	13921
Scaffolds L75	25544	25585
Scaffolds LG75	36680	36288
# local misassemblies	108473	105510
# unaligned scaffold	30070	27208
Unaligned length	36126952	30803824
Genome fraction (%)	74.556	75.877
Duplication ratio	1.122	1.115
# N's per 100 kbp	3998.54	2076.68
Largest alignment	166120	167605
# predicted genes (unique)	71917	76480

Supplementary Table 4. CEGMA statistic of presence of core eukaryotic genes within *Astatotilapia latif*asciata assembly.

Out of 248 CEGs	***Astatotilapia latif*asciata genome**
# of fully represented	183
# of at least partially represented	235
% of fully represented	73.79
% of at least partially represented	94.76
Average number of orthologs per CEG	1.62
% of detected CEGs with more than 1 ortholog	40

Supplementary Table 5. Statistic of gene annotation of *Astatotilapia latif*asciata genome**.**

Ab-initio Gene models statistic	**Numbers**
# predicted genes (unique)	71917
# predicted genes (>= 0 bp)	377769
# predicted genes (>= 1500 bp)	1390
# predicted genes (>= 3000 bp)	230
Protein coding genes structure prediction	
Number of genes	23391
Number of mRNAs	27437
Number of introns	177219
Number of exons	204656
Number of CDS	27437
Total gene length	147793588
Total mRNA length	187112381
Total intron length	138340497
Total exon length	49126322
Total CDS lenght	28549080
Longest gene	66982
Longest intron	16907
Longest exon	17286
Longest CDS	40644
Longest mRNA	6820
mean mRNAs per gene	1
Mean exon per gene	7
mean intron per gene	6

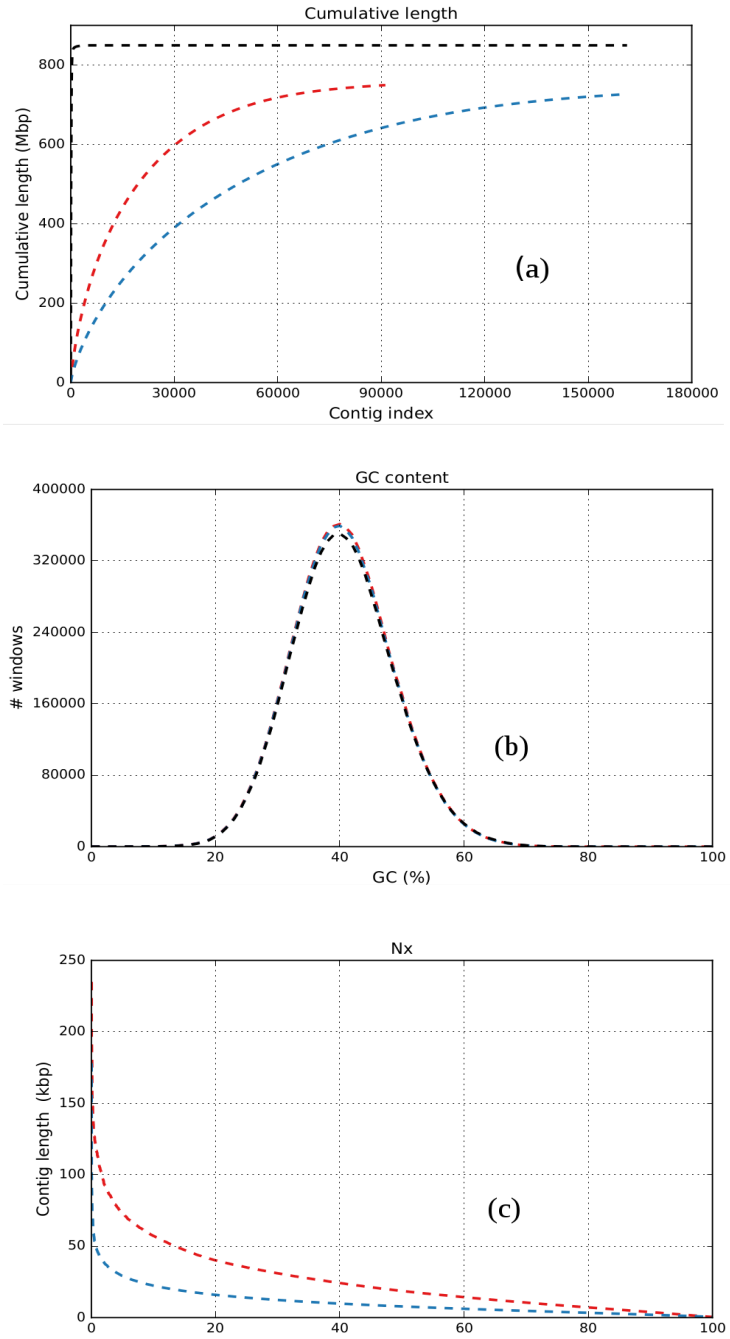
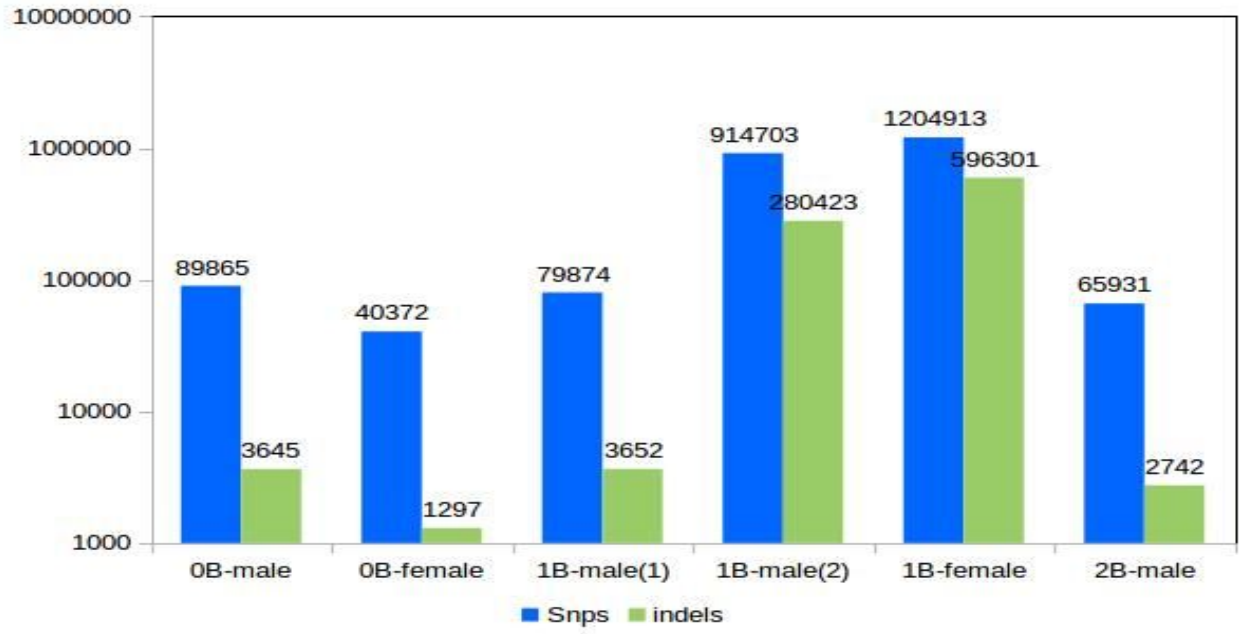
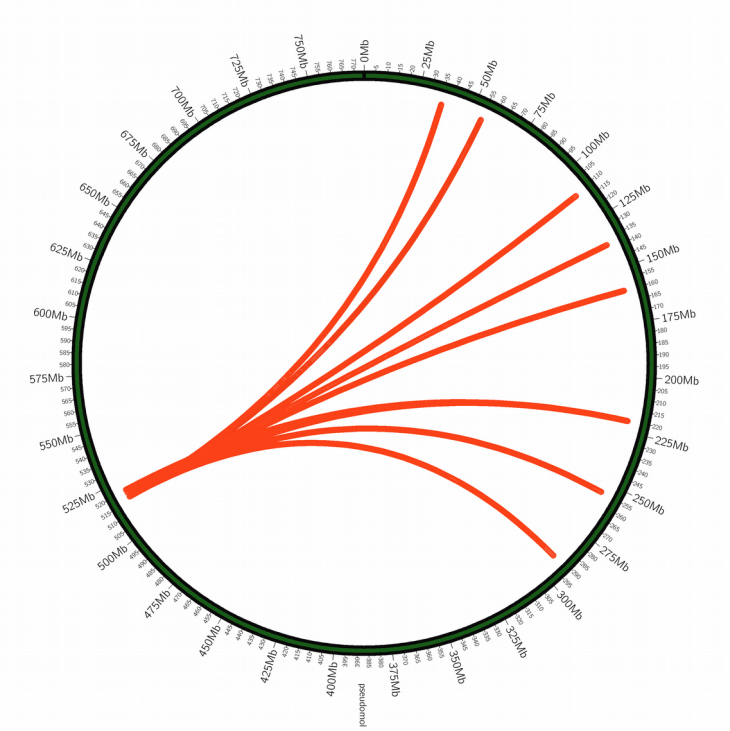


Figure.1. QUAST evaluation statistics of *A. latifasciata* genome assembly. (a) Cumulative length plot shows the increase of contig and scaffolds length, on the x-axis, contigs and scaffolds are ordered from the largest to smallest. The y-axis gives the size of x largest contigs in the assembly. (b) GC content plot shows the distribution of GC content in the contigs and scaffolds, The X value is the GC percentage (0-100%), The Y value is the number of non-overlapping 100bp windows. (c) Nx plot represents Nx values as x values from 0 to 100%. Scaffold (red dot line) and contig (blue dot line) with respect to reference genome *M. zebra* (black dot line) is shown.



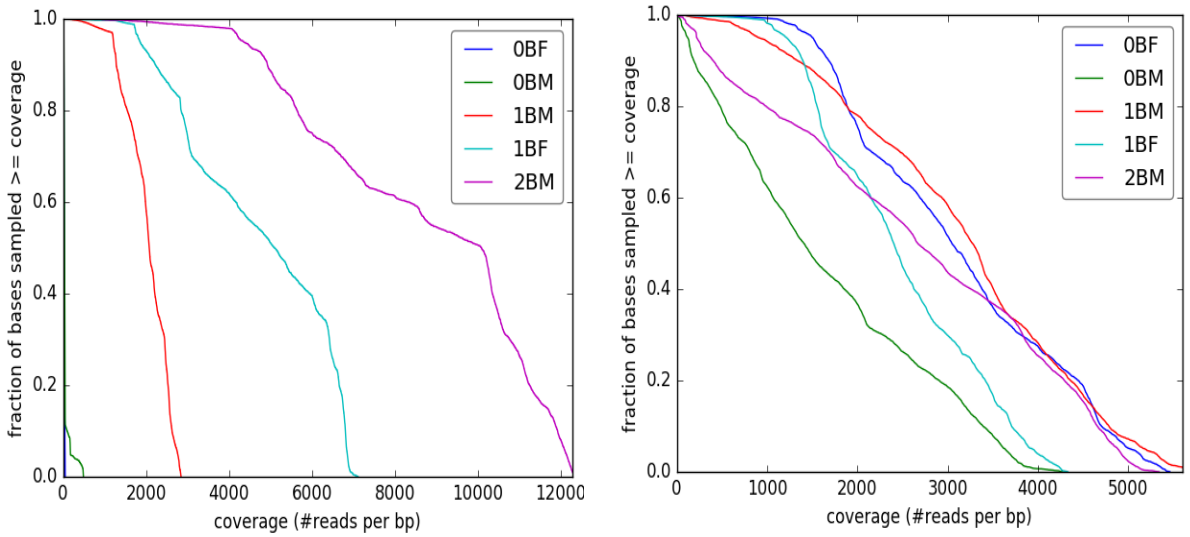
Supplementary Figure 2. Genomic diversity analyses. Different number of SNPs and INDELs in the whole genomes of all six individuals with a reference.



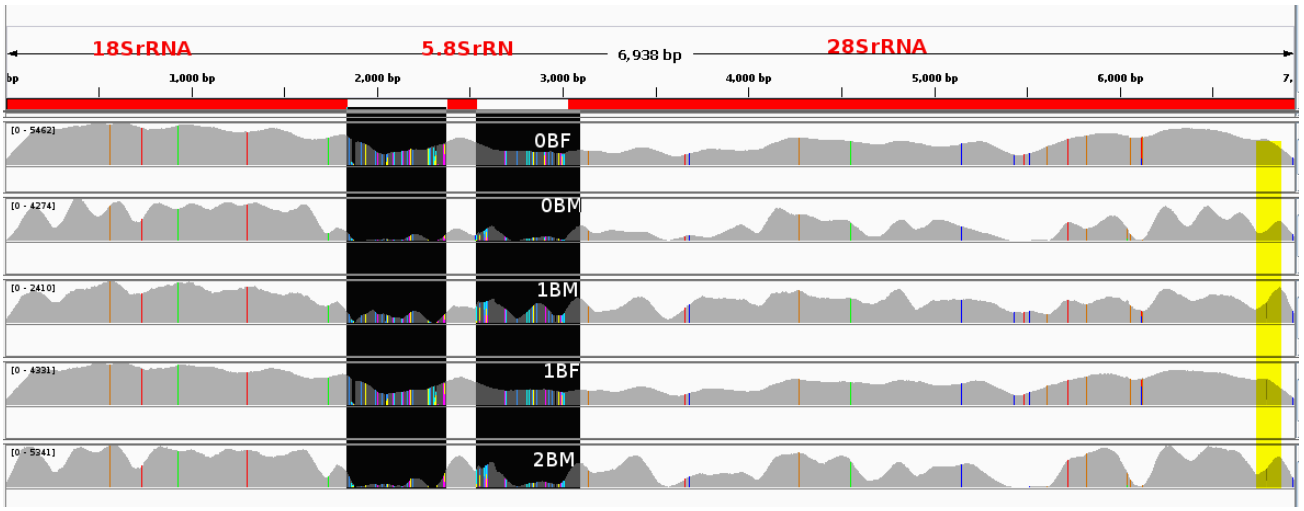
Supplementary Figure 3. circos representation of annotated region. The outermost ring shows the genome in green. Red links show the pattern of translocation in annotated region

Description	Max score	Total score	Query cover	E value	Ident	Accession
Dicentrarchus labrax chromosome sequence corresponding to linkage group 1, top part, complete sequence	959	1226	23%	0.0	84%	FQ310506.3
PREDICTED: Pundamilia nyererei uncharacterized LOC102200262 (LOC102200262), mRNA	752	1163	16%	0.0	89%	XM_005752029.2
PREDICTED: Haplochromis burtoni uncharacterized LOC102290459 (LOC102290459), transcript variant X2, ncRNA	676	676	9%	0.0	89%	XR_001336914.1
PREDICTED: Maylandia zebra RNA-directed DNA polymerase homolog (LOC106675602), mRNA	599	599	7%	3e-166	94%	XM_014410540.1
Haplochromis chilotes DNA, containing V2R gene cluster region, clone: 44K17	468	468	6%	8e-127	90%	AB780552.1
Haplochromis chilotes DNA, containing V2R gene cluster region, clone: 42K2	468	468	6%	8e-127	90%	AB780551.1
PREDICTED: Maylandia zebra uncharacterized LOC106676074 (LOC106676074), ncRNA	451	831	13%	9e-122	85%	XR_001342185.1
PREDICTED: Haplochromis burtoni XK-related protein 8-like (LOC102308239), mRNA	420	420	6%	2e-112	88%	XM_005943583.2
PREDICTED: Oreochromis niloticus uncharacterized LOC102082544 (LOC102082544), transcript variant X5, mRNA	353	353	4%	2e-92	90%	XM_019367097.1
PREDICTED: Oreochromis niloticus uncharacterized LOC102082544 (LOC102082544), transcript variant X4, mRNA	353	353	4%	2e-92	90%	XM_019367096.1

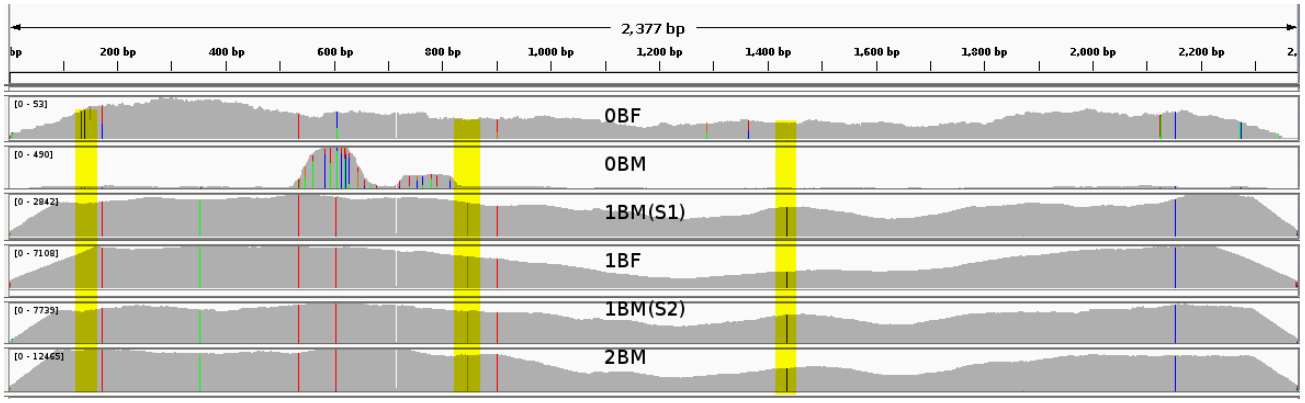
Supplementary Figure 4. Blast hits of B+ genome transloacation regions.



Supplementary Figure 5. The sequencing depth of the samples are shown. (a) The gene *Ihhb*: Sample 2B male having maximum and 0B female having minimum coverage. (b) 45S rRNA gene Sample 1B male having maximum and 0B male having minimum coverage.



Supplementary Figure 6. B-specific polymorphism of 45S rRNA cluster in B+(1B,2B) and B-(0B) having males(M) and female(F). The red fragmented line showing different RNAs while white space in between are interspersed DNA. Different types of SNPs (gender specific, specie related and B-specific,) are seen in the complete cluster and examples of B specific and population SNPs are shown on yellow and black highlighted regions respectively.



Supplementary Figure 7. B-specific polymorphism of *Ihhb* gene in B+(1B, 2B) samples(S) and B-(0B) having males(M) and female(F). The yellow highlighted region showed B specific SNPs at different locations.

Supplementary Figure 8. Agarose gel electrophoresis bands show amplified 45SrRNA cluster and *Ihhb* genes by PCR.

10. References

Alekseyev MA (2008). Multi-break rearrangements and breakpoint re-uses: from circular to linear genomes. Journal of Computational Biology. 15:1117–1131.

Alekseyev MA and Pevzner PA (2008). Multi-break rearrangements and chromosomal evolution. Theoretical Computer Science. 395: 193–202.

Alfenito MR and Birchler JA (1993). Molecular characterization of a maize B chromosome centric sequence. Genetics. 135: 589–597.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. J Mol Biol. 215: 403–410.

Amos A and Dove G (1981). The distribution of repetitive DNAs between regular and supernumerary chromosomes in species of Glossina (tsetse): a two-step process in the origin of supernumeraries. Chromosoma. 81:673-690.

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

Banaei-Moghaddam AM, Meier K, Karimi-Ashtiyani R, Houben A. (2013). Formation and Expression of Pseudogenes on the B Chromosome of Rye. *The Plant Cell*. 25:2536-2544.

Barnes MR, Gray IC (2003). Bioinformatics for geneticists. United Kingdom: Wiley.

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

Bentley DR (2000). The Human Genome Project - an overview. Med Res Rev. 20: 189-96.

Bickhart DM, Liu GE (2014). The challenges and importance of structural variation detection in livestock. Front Genet. 5:37

Boetzer M and Pirovano W (2012). Toward almost closed genomes with GapFiller. Genome Biology. 6:R56.

Bosemark NO (1957b). On accessory chromosomes in Festuca pratensis 5. Influence of accessory chromosomes on fertility and vegetative development. Hereditas. 43:211-235.

Botstein D, White RL, Skolnick M, Skolnick M, Davis RW (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 32 (3):314– 331.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I et al (2013). Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. GigaScience. 2:10.

Branton D, Deamer DW, Marziali A, Bayley H, Benner SA et al (2008). The potential and challenges of nanopore sequencing. Nature Biotechnology. 26:1146–1153.

Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S et al (2014). The genomic substrate for adaptive radiation in African cichlid fish. Nature. 513:375-381.

Burt A and Trivers RL (2006). Genes in Conflict : The Biology of Selfish Genetic Elements. Belknap Press, Harvard.

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*. Chromosome Research. 20: 279-292.

Cabrero J, López-León M, Ruíz-Estévez M, Gómez R, Petitpierre E, Rufas J, and Massa B (2014). B1Was the Ancestor B Chromosome Variant in the Western Mediterranean Area in the GrasshopperEyprepocnemis plorans. Cytogenetic and Genome Research. 142: 54-58.

Cabrero, J. Lo´pez–Leo´n MD, Go´mez R et al (1997). Geographical distribution of B chromosomes in the grasshopper Eyprepocnemis plorans, along a river basin, is mainly shaped by non-selective historical events. Chromosome Res. 5:194–198.

Camacho JP, Sharbel T, Beukeboom L (2000). B-chromosome evolution. Phil Trans R Soc Lond B. 355:163–178.

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

Camacho JPM (2005). B Chromosomes. In: Gregory TR, editor. The evolution of the genome. Amsterdam: Elsevier Academic Press. 223–286.

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research., 18(1):, 188–196.

Cargill M, Altshuler D, Ireland J, Irland J, Sklar J, Ardlie K, Patil N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet. 22:231 – 238.

Carroll SB (1995). Homeotic genes and the evolution of chordates. Nature (London). 376:479–485.

Carvalho AB, Vicoso B, Russo CAM, Swenor B, Clark AG (2015). Birth of a new gene on the Y chromosome of Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America. 112:12450–12455.

Cavalli-Sforza LL, Bodmer WF (1971). The Genetics of Human Populations. San Francisco: W.H. Freeman and Company.

Chen W, Ullmann R, Langnick C, Menzel C, Wotschofsky Z, Hu H, Döring A, Hu Y, Kang H, Tzschach A, Hoeltzenbein M, Neitzel H, Markus S, Wiedersberg E, Kistner G, van Ravenswaaij-Arts CMA, Kleefstra T, Kalscheuer VM, Ropers HH (2010). Breakpoint analysis of balanced chromosome rearrangements by next-generation paired-end sequencing. Eur J Hum Genet. 18:539–43.

Cheng ZK, Yu HX, Yan HH, Gu MH, and Zhu LH (2000). B chromosome in a rice aneuploid variation. Theoretical and Applied Genetics. 101: 564-568.

Cheong WH, Tan YC, Yap SJ, Ng K (2015). ClicO FS: an interactive web-based service of Circos. Bioinformatics. 31:3685-3687.

Chiavarino AM, Rosato M, Naranjo CA, Camara Hernandez H and Poggio L (1995). B chromosome polymorphism in N. Argentine populations of maize. Maize Genet Coop News Lett. 69: 94.

Chung A, Kim S, Kim E, Kim D, Jeong I, cha YR, Bae YK, park SW, Lee J, Park HC (2013). Indian hedgehog B function is required for the specification of oligodendrocyte progenitor cells in the zebrafish CNS. J Neurosci. 33:1728-1733.

Danecek P, Auton A, Abecasis G, Albers A.C, Bank E, Depristo M.A, Handsaker R.E, Lunter G (2011). The variant call format and VCFtools. Bioinformatics. 27:2156-2158.

Datta AK, Mandal A, Das D, Gupta A, Saha R, Paul S (2016). B chromosomes in angiosperm—a review. Cytol Genet. 50: 60.

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

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.

Ellegren H (2014). Genome sequencing and population genomics in non model organisms. Trends in Ecology & Evolution. 29:51–63.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M (2008). Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. Science. 320:1629–163.

Fan JB, Chee MS, and Gunderson KL (2006). Highly parallel genomic assays. Nature Review Genetics. 7:632–644.

Fan S, Meyer A (2014). Evolution of genomic structural variation and genomic architecture in the adaptive radiations of African cichlid fishes. *Front Genet*. 5:163.

Fantinatti BEA and 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 Genetics. 17:119.

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.

Feldberg E, Bertollo LAC (1984). Discordance in chromosome number among somatic and gonadal tissue cells of Gymnogeonhagus balzani. Brazilian Journal of Genetics. 4:639-645.

Feldberg E, Porto JIR, Brinn MNA, Mendonça MNC, Benzaquem DC (2004). B- Chromosomes in Amazonian cichlid species. Cytogenetic and Genome Research. 106:195-198.

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

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

Genner MJ, Turner GF (2005). The mbuna cichlids of Lake Malawi: a model for rapid speciation and adaptive radiation. Fish Fish. 6:1-34.

Gonzalez-Sanchez M, Chiavarino M, Jimenez G, Manzanero S, Rosato M, Puertas MJ (2004). The parasitic effects of rye B chromosomes might be beneficial in the long term. Cytogenetic and genome research. 106:386-393.

Goodwin SB, M'Barek SB, Dhillon B, Wittenberg AHJ, Crane CF, Hane JK et al (2011). Finished genome of the fungal wheat pathogen Mycosphaerella graminicola reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genetics. 7: e1002070.

Graphodatsky AS, Kukekova AV, Yudkin DV, Trifonov VT, Vorobieva NV, Beklemisheva VR, Perelman PL, Graphodatskaya DA, Trut LN, Yang F, Ferguson-Smith MA, Acland GM and Aguirre GD (2005). The proto-oncogene C-KIT maps to canid B-chromosomes. Chromosome Res. 13:113-12.

Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, Eis PS, Shannon WD, Li X, McLeod DL, Chevraud JM, Ley TJ (2007). A high-resolution map of segmental DNA copy number variation in the mouse genome. PLoS Genet. 3(1):e3.

Green DM (1990). Muller Ratchet and the evolution of supernumerary chromosomes. Genome. 33:818-824.

Gudbjartsson DF, Sulem P, Helgason H, Gylfason A, Gudjonsson SA , Zink F , Oddson A, Magnusson G, Halldorsson BV, Hjartarson E, Sigurdsson GT1, Kong A, Helgason, Masson G, Magnusson OT, Thorsteinsdottir, Stefansson K (2015). Sequence variants from whole genome sequencing a large group of Icelanders. Scientific Data. 2:150011.

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013). QUAST: quality assessment tool for genome assemblies.Bioinformatics. 29:1072-1075.

Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet. 22 (3):239 – 247.

Harvey AW and Hewitt GM (1979). B chromosomes slow development in a grasshopper. Heredity. 42:397-401.

Henriques-Gil N and Arana P (1990). Origin And Substitution Of B Chromosomes In The Grasshopper Eyprepocnemis Plorans. Evolution. 44:747-753.

Henson J, Tischler G, Ning Z (2012). Next-generation sequencing and large genome assemblies. Pharmacogenomics. 13(8):901–915.

Hewitt GM, East TM, Shaw MW (1987). Sperm dysfunction produced by B chromosomes in the grasshopper Myrmeleotettix maculatus. Heredity. 58:59-68.

Holland PWH (1994). Homeobox genes in vertebrate evolution . BioEssays .14:267–273.

Houben A, Banaei-Moghaddam AM, Klemme S, Timmis JN (2013). Evolution and biology of supernumerary B chromosomes. Cell Mol Life Sci. 71:467-478.

Houben A, Verlin D, Leach CR, Timmis JN (2001). The genomic complexity of micro B chromosomes of Brachycome dichromosomatica. Chromosoma. 110: 451–459.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M et al (2013). The zebrafish reference genome sequence and its relationship to the human genome. Nature. 496:498-903.

Ijiri S, Kaneko H, Kobayashi T, Wang DD, Sakai F, Paul-Prasanth B, Nakamura M, Nagahama Y (2007). Sexual Dimorphic Expression of Gene in Gonads during early differentiation of a Teleost Fish, the Nile Tilapia Oreochromis niloticus. Bio Reprod. 78:333–341.

Ingham PW, McMahon AP (2001). Hedgehog signaling in animal development: paradigms and principles. Genes Dev. 15: 3059–3087.

Iskow RC, Gokcumen O, Lee C (2012). Exploring the role of copy number variants in human adaptation. *Trends Genet*. 28, 245–257.

Jones JDG and Flavell RB (1982). The structure, amount and chromosomal localization of defined repeated DNA sequences in species of the genus Secale. Chromosoma. 86:613-641.

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

Jones NR, Gas WD, Ben AH (2008). A Century of B Chromosomes in Plants: So What?. Annals of Botany. 101: 767 –775, 2008.

Jones RN and Rees H (1982). B chromosomes. (London: Academic Press).

Kao KW, Lin CY, Peng SF, Cheng YM (2005). Characterization of four B-chromosome-specific RAPDs and the development of SCAR markers on the maize B-chromosome. Mol Genet Genomics. 2015;290(2):431–41.

Kasahara M, Naruse K, Sasaki S, Nakatan Yi, Qu W , Ahsan B et al (2007). The medaka draft genome and insights into vertebrate genome evolution. Nature. 447:714-719.

Keane TM, Wong K, Adams DJ, Flint J, Reymond A, Yalcin B (2014). Identification of structural variation in mouse genomes. Front Genet. 5:192.

Kellis M, Birren BW, Lander ES (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature. 428:617–24.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P et al (2008). Mapping and sequencing of structural variation from eight human genomes. Nature. 453:56–64.

Kingsford C, Schatz M, Pop M (2010). Assembly complexity of prokaryotic genomes using short reads. BMC Bioinformatics. 11:21.

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

Konerat JT, Thums J, Vanessa B, Lucas B, Martins-Santos IC, Margarido VP (2014). B chromosome and NORs polymorphism in Callichthys callichthys (Linnaeus, 1758) (Siluriformes: Callichthyidae) from upper Paraná River, Brazil. Neotropical Ichthyology. 12:603-609.

Korbel JO, Urban AE, Affourtit JP, et al (2007). Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science (New York, NY)*. 318:420-426.

Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009). Circos: An information aesthetic for comparative genomics. Genome Res. doi: 10.1101/gr.092759.109.

Kumar S, Balczare KA, Zhi-Chun L (1996). Phylogenetic analyses of Sonic hedgehog (Shh) and Hoxd-10 genes from 18 cyprinid fish species closely related to the zebrafish. Genetics 142:965–972.

Kumar S, Balczarek K, Lai Z (1996). Evolution of the hedgehog gene family. Genetics. 142:965–1037.

Kuroiwa A, Terai Y, Kobayashi N, Yoshida K, Suzuki M, Nakanishi A, Matsuda Y, Watanabe M, Okada N (2014). Construction of chromosome markers from the Lake Victoria cichlid Paralabidochromis chilotes and their application to comparative mapping. Cytogenet Genome Research. 142:112–120

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature methods*. 2012;9(4):357-359.

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

Liehr T, Mrasek K, Kosyakova N, Ogilvie CM, Vermeesch J, Trifonov V, Rubtsov N (2008). Small supernumerary marker chromosomes (sSMC) in humans; are there B chromosomes hidden among them. Mol Cytogenet. 1:12.

Lin G, Chai J, Yuan S, Mai C, Cai L, Murphy R.W, Zhou W, Luo J (2016). VennPainter: A Tool for the Comparison and Identification of Candidate Genes Based on Venn Diagrams. PLoS One. 11: e0154315.

Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin H, Pong R, Lin D, Lu L, Law M (2012). Comparison of Next-Generation Sequencing Systems. Journal of Biomedicine and Biotechnology. 2012:251364.

Loh YHE, Katz LS, Mims MC, Kocher TD, Yi SV, Streelman JT (2008). Comparative analysis reveals signatures of differentiation amid genomic polymorphism in Lake Malawi cichlids. Genome Biology, 9:R113.

López-León M, Cabrero J., Dzyubenko, V, Bugrov A, and Karamysheva T (2008). Differences in ribosomal DNA distribution on A and B chromosomes between eastern and western populations of the grasshopper Eyprepocnemis plorans plorans. Cytogenetic and Genome Research. 121:260-265.

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

Lukashin AV and Borodovsky M (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 26:1107–1115.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J et al (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 1:18.

Mahul C, James GBB, Anthony DL, Emerson JJ (2015). A practical guide to de novo genome assembly using long reads. BioRxiv.

Maistro EL, Foresti F, Oliveira C, Toledo LFD (1992). Occurrence of macro-B chromosomes in Astyanax scabripiMartins-Santos IC, Portela-Castro ALB, Julio HF Jr (1995). Chromosome analysis of 5 species of the Cichlidae family (Pisces, Perciformes) from the Paraná River. Cytologia. 60:223–231.

Makunin AI, Dementyeva PW , Graphodatsky AS, Volobouev VT , Kukekova AV, Trifonov VA (2014). Genes on B chromosomes of vertebrates. Molecular Cytogenetics. 7:99.

Marques DF, Conte MA, Fantinatti BEA, Valente GT, Nakajima RT, Coan RLB, Kocher TD, Martins C. B chromosomes effects on gene transcription in the cichlid fish Astatotilapia latifasciata (unpublihed).

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, Kubalakova M, Bauer E, Haseneyer G, Fuchs J, Dolezel J, Stein N, Mayer KF, Houben A (2012). Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. Proc Natl Acad Sci USA. 109:13343–13346.

Mathieu Blanchette (2001). Evolutionary Puzzles: An Introduction to Genome Rearrangement. Computational Science – ICCS. 2074:1003-1011.

Matzke M, Varga F, Berger H, Schernthaner J, Schweizer D, Mayr B, Matzke AJM (1990). A tandemlyrepeated sequence isolated from nuclear envelopes of chickenerythrocytes is located predominantly on microchromosomes. Chromosoma. 99: 131-137

Mazzuchelli J, yang F, Kocher TD, Martins C (2011). Comparative cytogenetic mapping of *Sox*2 and *Sox14* in cichlid fishes and inferences on the genomic organization of both genes in vertebrates. Springer Link. 19: 657–667.

McClintock B, T. Kato TA, Blumenschein A (1981). Chromosome Constitution of the Races of Maize. Colegio de Postgraduados, Chapingo, México.

Mendelson D and Zohary D (1972). Behavior and transmission of supernumerary chromosomes in Aegilops speltoides. Heredity. 29:329–339.

Metzker ML (2005). Emerging technologies in DNA sequencing. Genome Research. 15:1767-1776.

Meyer A, Van de Peer Y (2005). From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). Bioessays. 27:937–945.

Mizoguchi SMHN and Martins-Santos IC (1997). Macro- and microchromosomes B in females of Astyanax scabripinnis (Pisces, Characidae). Hereditas. 127:249-253.

Muñoz-Pajares AJ, Martínez-Rodríguez L, Teruel M, Cabrero J, Camacho J PM, Perfectti F (2011). A Single, Recent Origin of the Accessory B Chromosome of the Grasshopper Eyprepocnemis plorans. Genetics. 187:853–863.

Naiara RE, Michael H, Ana MA (2012). Bioinformatics for High Throughput Sequencing. Springer. DOI 10.1007/978-1-4614-0782-9.

Naranjo CA, Chiavarino AM, Rosato M, Quintela Fernández E, Poggio L (1995). Tamaño del genoma y polimorfismo para cromosomas B en razas nativas argentinas y bolivianas de maiz. 969–980.

Noor MA, Grams KL, Bertucci LA, Reiland J (2001). Chromosomal inversions and the reproductive isolation of species. Proc Natl Acad Sci U.S.A. 98:12084–12088.

O'Quin CT, Drilea AC, Conte MA, Kocher TD. Mapping of pigmentation QTL on an anchored genome assembly of the cichlid fish, Metriaclima zebra. BMC Genomics.14: 287.

Ohta S (1996). Mechanisms of B-chromosome accumulation in Aegilops mutica Boiss. Genes Genet Syst. 71(1):23–29.

Oliveira C, Saboya SMR, Foresti F, Senhorini JA, Bernardino G (1997). Increased B chromosome frequency and absence of drive in the fish Prochilodus lineatus. Heredity. 79:473-476.

Parra G, Bradnam K, Korf I (2007). CEGMA: a pipeline to accurately annotate core genes eukaryotic genomes. Bioinformatics. 23 :1061-1067.

Pauls E, Bertollo LAC (1983). Evidence for a system of supernumerary chromosome in Prochilodus scrofa (Pisces, Prochilodontidae). Caryologia. 36:307-314.

Pellegrino KCM, Rodrigues MC, Yonenaga-Yassuda Y (1999). Chromosomal polymorphisms due to supernumerary chromosomes and pericentric inversions in the eyelidless microteiid lizard Nothobachia ablephara. Chromosome Research.7:247-254.

Peng SF, Lin YP, Lin BY (2005). Characterization of AFLP sequences from regions of maize B chromosome defined by 12 B-10L translocations. Genetics. 169:375-388.

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

Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J (2009). Metagenomic pyrosequencing and microbial identification. Clinical Chemistry. 55:856–866.

Petrov DA (2002). Mutational equilibrium model of genome size evolution. Theor Popul Biol. 61:533–46.

Pinkel D, Landegent J, Collins C, et al (1988). Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. Proceedings of the National Academy of Sciences of the United States of America. 85: 9138-9142.

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. 11: 50.Poletto AB, Ferreira IA, Martins C (2010a). The B chromosome of the cichlid fish Haplochromi sobliquidens harbors 18S rRNA genes. BMC Genetics. 11:1.

Pop M and Salzberg SL (2008). Bioinformatics challenges of new sequencing technology. Trends Genetics. 24:142–149.

Portela-Castro ALD, Juilo HF, Nishiyama PB (2000). New occurrence of microchromosomes B in Moenkhausia sanctaefilomenae (Pisces, Characidae) from the Parana River of Brazil: analysis of the synaptonemal complex. Genetica 110:277-28.

Potapov VA, Solov'ev VV, Romashchenko AG, Sosnovtsev SV, Ivanov SV (1990). Features of the structure and evolution of complex, tandemly organized Bsp-repeats in the fox genome. I. Structure and internal organization of the BamHI-dimer. Mol Biol. 24:1649–1665.

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(Web Server issue):W567-575.

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

R. Chikhi, P. Medvedev, Informed and automated *k*-mer size selection for genome assembly, *Bioinformatics* (2014) 30 (1): 31-37.

Ramirez J, dundar F, Diehl s, Gruning B, Manke T (2014). deepTools: a flexible platform for exploring deep-sequencing data. doi: 10.1093/nar/gku365.

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.

Rhoades M (1968). Studies on the cytological basis of crossing over. In: Peacock W, Brock P, editors. Replication and Recombination of Genetic Material. Canberra: Australian Acad Sci. 229–41.

Riva A and Kohane IS (2002). SNPper: retrieval and analysis of human SNPs. Bioinformatics. 18:1681-1685.

Robert E and Jochen BWW (2014). A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications. 7:1026–1042.

Rosato M, Chiavarino AM, NaranjoCA, Cámara Hernández J, Poggio (1998). Genome size and numerical polymorphism for the B chromosome in races of maize (Zea mays ssp. mays, Poaceae). Am J Bot. 85: 168–174.

Rubtsov NB, Karamysheva TV, Andreenkova OV, Bochkaerev MN, Kartavtseva IV, Roslik GV, Borissov YM (2004). Comparative analysis of micro and macro B chromosomes in the Korean field mouse Apodemus peninsulae Rodentia, Murinae) performed by chromosome microdissection and FISH Cytogenet. Genome Res 2004. 106:289–294.

Ruíz-Estévez M, López-León MD, Cabrero J, Camacho JPM (2012) B-Chromosome Ribosomal DNA Is Functional in the Grasshopper Eyprepocnemis plorans. PLoS ONE. 7:e36600.

Sandery MJ, Forster JW, Blunden R, Jones RN (1990). Identi¢cation of a family of repeated sequences on the rye B chromosome. Genome. 33: 908-913.

Sanger F; Nicklen S; Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci. 74: 5463–5467.

Schartl M, Walter RB, Shen Y, Garcia T, Catchen J et al (2013). The genome of the platyfish, Xiphophorus maculatus, provides insights into evolutionary adaptation and several complex traits. Nat Genet. 45:567–572.

Schatz MC, Delcher AL, Salzberg SL (2010). Assembly of large genomes using second-generation sequencing. Genome. 9:1165-1173.

Schmid M, Ziegler CG, Steinlein C, Nanda I, Haaf, T (2002). Chromosome banding in amphibia - XXIV. The B chromosomes of Gastrotheca espeletia (Anura, Hylidae). Cytogenetic and genome research. 97:205-21.

Schork NJ, Fallin D, Lanchbury S (2000). Single nucleotide polymorphisms and the future of genetic epidemiology. Clin Genet .58:250 – 264.

Seo JH, Bae GH, Park DH, Kim BS, Lee JW, Lee JI, Kim KH, Lee SK, Seo BB (2013). Sequence polymorphisms in ribosomal RNA genes and variations in chromosomal loci of Oenothera odorata and O. laciniata. Genes Genom. 35:117–124.

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE (1997). Ethnic affiliation estimation by use of population-specific DNA markers. Am J Hum Genet. 60:957 – 964.

Silva DM, Pansonato-Alves JC, Utsunomia R, Araya-Jaime C, Ruiz-Ruano FJ, Daniel SN, Hashimoto DT, Oliveira C, Camacho JPM, Porto-Foresti F, Foresti F (2014). Delimiting the Origin of a B Chromosome by FISH Mapping, Chromosome Painting and DNA Sequence Analysis in (Teleostei, Characiformes). PloS One. 9:e94896.

Silva MJJ, Yonenaga-Yassuda Y (1998). Heterogeneity and meiotic behaviour of B and sex chromosomes, banding patterns and localization of (TTAGGG)n sequences by FISH, in the Neotropical water rat Nectomys (Rodentia, Cricetidae). Chromosome Res. 6: 455-462.

Simpson JT and Pop M (2015). The Theory and Practice of Genome Sequence Assembly. Annu Rev Genomics Hum Genet. 16:153-72.

Sims DI, Sudbery N E, IlottA, Heger CP, Ponting (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics. 15:121–132

Smit AF (1996). The origin of interspersed repeats in the human genome. Curr Opin Genet Dev. 6:743–748.

Soon WW, Hariharan M and Michael P Snyder (2013). High-throughput sequencing for biology and medicine. Molecular System Biology. 9:640.

Sparks JS and Smith WL (2004). Phylogeny and biogeography of cichlid fishes (Teleostei: Perciformes: Cichlidae). Cladistics. 20:501-517.

Stiassny MLJ (1991). Phylogenetic intrarelationships of the family Cichlidae: an overview. Cichlid fishes. Behaviour, ecology and evolution. London: 1-35.

Stitou S, de La Guardia RD, Jiménez R, Burgos M (2000). Inactive ribosomal cistrons are spread throughout the B chromosomes of Rattus rattus (Rodentia, Muridae). Implications for their origin and evolution. Chromosome Res. 8:305–311.

Strittmatter WJ and Roses AD (1996). Apolipoprotein E and Alzheimer's disease. Annu Rev Neurosci. 19:53-77.

Szczerbal I and Switonski M (2003). B chromosomes of the Chinese raccoon dog (Nyctereutes procyonoides procyonoides Gray) contain inactive NOR-like sequences. Caryologia. 56:213–216.

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

Treangen TJ and Salzberg SL (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews. Genetics, 13:36–46.

Trifonov VA, Dementyeva 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 Biolology. 6:90.

Valente GT, Conte MA, Fantinatti BEA, Cabral-de-melo DC, Carvalho R, Vicari MR, Kocher TD, Martins C (2014). Origin and Evolution of B Chromosomes in the Cichlid Fish Astatotilapia latifasciata Based on Integrated Genomic Analyses. Molecular Biology Evolution. 31:2061-72.

Vicente VE, Moreira-Filho O, Camacho JP (1996). Sex-ratio distortion associated with the presence of a B chromosome in Astyanax scabripinnis (Teleostei, Characidae). Cytogenetics and cell genetics. 74:70-75.

Wang Z, Gerstein M, Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 10:57–63.

Wheat CW (2010). Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. Genetica. 138:433–451.

Wilkes TM, Francki MG, Langidge P, Karp A, Jones RN, Forster JW (1995). Analysis of rye B-chromosome structure using fluorescence in situ hybridization (FISH). Chromosome Res. 3:466 -472.

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

Wold B and Myers RM (2008). Sequence census methods for functional genomics. Nature Methods. 5:19–21.

Wurster-Hill DH, Ward OG, Davis BH, Park JP, Moyzis RK, Meyne J (1988). Fragile sites, telomeric DNA sequences, B chromosomes, and DNA content in raccoon dogs, Nyctereutes procyonoides, with comparative notes on foxes, coyote, wolf, and raccoon. Cytogenet Genome Res. 49:278–281.

Yandell M and Ence D (2012). A beginner's guide to eukaryotic genome annotation. Nat Rev Genet, 13:329–342.

Yang MQ, Athey BD, Arabnia HR, Sung AH, Liu Q, Yang JY, Deng Y (2009). High-throughput next generation sequencing technologies foster new cutting-edge computing techniques in bioinformatics. BMC Genomics. 10:I1.

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.

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. USA. 93:13036–13041.

Zerbino DR and Birney E (2008). Velvet: Algorithms for de novo short read assemblyusing de Bruijn graphs. Genome Research. 18:821–829

Zeyl CW, Green DM (1992). Heteromorphism for a highly repeated sequence in the New Zealand frog Leiopelma hochstetteri. Evolution. 46:1891-1899.

Zurita, S, Cabrero J, Lopez-Leon MD, Camacho JP (1998). Polymorphism Regeneration for a Neutralized Selfish B Chromosome. Evolution. 52: 274.