RESSALVA

Atendendo solicitação do(a) autor(a), o texto completo desta dissertação será disponibilizado somente a partir de 30/05/2021.

HIGH SCALE GENOMIC ANALYSIS APPLIED TO B CHROMOSOME BIOLOGY

SYED FARHAN AHMAD

Botucatu, May 2019

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

HIGH SCALE GENOMIC ANALYSIS APPLIED TO B CHROMOSOME BIOLOGY

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

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

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

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

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

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

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

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

Marie Curie

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

Albert Einstein.

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

Acknowledgment

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

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

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

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

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

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

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

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

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

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

Syed Farhan Ahmad

Table of Contents

Highlights

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

Abstract

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

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

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

1. Introduction

1.2 B chromosomes

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

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

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

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

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

1.2 The application of cytogenetics and genomics to B chromosome analysis

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

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

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

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

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

1.3.1 *Astyanax mexicanus*

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

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

1.3.2 *Astyanax correntinus*

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

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

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

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