ABSTRACT

Title of Dissertation: GENETIC CONFLICT IN LAKE MALAWI CICHLIDS: B CHROMOSOMES AND SEX DETERMINATION

Frances Elizabeth Clark, Doctor of Philosophy, 2019

Dissertation directed by: Professor Thomas D. Kocher, Biology

B chromosomes (Bs) are selfish genetic elements known to manipulate various cellular processes. These manipulations increase their transmission to the next generation, a process known as drive. After the recent discovery of Bs in African cichlid fish, sequence amplification methodologies were used to quantify B chromosome distribution in 7 species of Lake Malawi cichlids. In these species, Bs are limited to females and are haploid in the diploid genome. Considering various possible drive mechanisms, I propose this B chromosome drives by manipulating meiosis I in females. Genetic crosses quantifying B transmission in *Metriaclima lombardoi* confirmed transmission above Mendelian expectations. The transmission of this B also skews the sex ratio among progeny towards females. *M. lombardoi* individuals lacking Bs were shown, via a genetic linkage analysis, to have a male heterogametic (XY) sex determination system. A similar linkage analysis of families

segregating B chromosomes indicated only the progeny lacking a B were influenced by this XY system. This substantiates the hypothesis that this B is a female sex determiner. Individuals of all 7 species were re-sequenced with short-reads and read coverage across the genome was compared in a coverage ratio analysis that resulted in the detection of 1.37 Mb in the reference genome with copies on the B, shared by all 7 species. Accounting for copy number of each sequence, 12-44 Mb of shared B sequence was identified. Amongst this sequence were 144 loci containing genes and gene fragments. A differential expression analysis found hundreds to thousands of differentially expressed loci between individuals with and without Bs, biased towards decreased expression in B individuals. Transcriptomes were analyzed for B-specific SNPs revealing 53 loci transcribed from the B chromosome and six candidate genes that might contribute to drive. I have described the distribution and behavior of the Lake Malawi cichlid B as well as captured a large portion of its sequence. This, combined with the genomic resources available for cichlids, makes this model system a valuable tool for future studies of the molecular mechanisms of drive, sequence structure and evolution of B chromosomes, and the association between B and sex chromosomes.

GENETIC CONFLICT IN LAKE MALAWI CICHLIDS: B CHROMOSOMES AND SEX DETERMINATION

by

Frances Elizabeth Clark

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Professor Thomas Kocher, Chair
Professor Karen Carleton
Professor Eric Haag
Professor Stephen Mount
Professor Gerald Wilkinson

Foreword

The work presented in Chapter 2 was previously published in the *Journal of Heredity*. Author contributions are as follows. DNA samples were collected and extracted by all authors. Karyotypes were made by I. Ferreira-Bravo, A. Poletto and C. Martins. DNA libraries were made by M. Conte and T. Kocher. Primers were designed by F. Clark. PCR, qPCR and copy number analysis was performed by F. Clark. All authors contributed to writing of the manuscript.

> Clark, F., Conte, M., Ferreira-Bravo, I., Poletto, A., Martins, C. and Kocher, T. (2017). Dynamic Sequence Evolution of a Sex-Associated B Chromosome in Lake Malawi Cichlid Fish. *Journal of Heredity*, 108(1), pp.53-62.

The work presented in Chapter 4 was previously published in the *Genes* special issue, Evolution, Composition and Regulation of Supernumerary B chromosomes. Author contributions are as follows. DNA samples were collected, extracted, and libraries were made by all authors. PacBio alignment was performed by M. Conte. Block identification scripts were written by F. Clark, with oversight from M. Conte and T. Kocher. All authors contributed to writing of the manuscript.

> Clark, F., Conte, M. and Kocher, T. (2018). Genomic Characterization of a B Chromosome in Lake Malawi Cichlid Fishes. *Genes*, 9(12), p.610.

Acknowledgements

I'd like to acknowledge the current and past members of the cichlid lab community at the University of Maryland. I truly believe the best science happens when fostered by a supportive and communicative culture. I am particularly grateful to Dr. Matthew Conte for patiently helping me to learn bioinformatics. I would like to offer my special thanks to my advisor, Dr. Thomas Kocher not only for his valuable guidance and constructive feedback, but also for providing me with the environment in which to develop as a scientist. Finally, I would like to express my very great appreciation to my husband and family for being the best source of support and encouragement throughout this endeavor.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Bs – B chromosomes or Supernumerary chromosomes

As – typical “A” chromosomes found in every member of a species

B, 2B – an individual that possesses a B chromosome or 2 B chromosomes

NoB – an individual that does not possess a B chromosome

XY – male heterogametic sex chromosomes

ZW – female heterogametic sex chromosomes

MYA – million years ago

PCR – polymerase chain reaction

qPCR – quantitative PCR

RT-PCR – real time PCR

cDNA – complimentary DNA

FISH – Fluorescent *in situ* hybridization

CNV – copy number variation

MS-222 - tricaine methanesulfonate

SCR – scaled coverage ratio

Chapter 1: Introduction to B Chromosome Biology

B Chromosomes

This chapter reviews the literature on B chromosomes to provide context for the work described in chapters 2 through 5. I begin by describing what B chromosomes are, their occurrence in other taxa, and their unusual transmission or drive. This is pertinent for Chapter 2, where I quantify the distribution of B chromosomes in cichlids from Lake Malawi and propose a mechanism of drive. This portion of the introduction, along with the section discussing B chromosomes' association with sex is also relevant for Chapter 3, in which I provide data explaining the female-limited presence of the Lake Malawi cichlid B chromosome. As a primer for Chapter 4, I discuss what is known about the DNA sequence of B chromosomes in other taxa. Finally, I review what is known about transcription of sequences from the B chromosomes of other species, as context for my characterization of transcription in Chapter 5.

Every species has its own typical set of chromosomes referred to as the A chromosomes (As). Many species across a wide taxonomic range, including plants, animals and fungi, can possess an additional chromosome called a B chromosome (B) (Burt and Trivers 2008; Camacho 2000; D'Ambrosio et al. 2017). B chromosomes are defined by their existence in some but not all members of a population, their non-

essential quality, and their non-Mendelian inheritance (Jones 1991; Jones and Rees 1982; Camacho 2011).

B chromosomes are found in varying numbers within cells (Randolph 1941; Jones 1991; Burt and Trivers 2008). In natural populations, individuals may carry 1, 2, 3, 4 or more copies of a B chromosome. The chive, *Allium schoenoprasum,* has been shown to carry as many as 20 Bs, maize have been shown to carry up to 34 Bs and succulents of the genus *Pachyphytum* can carry more than 50 Bs (Camacho 2000; Uhl et al. 1973). The number of Bs can vary from one individual to the next within a population as well as from one cell to the next within an individual (Jones 1991; Burt and Trivers 2008). In a stock of *Drosophila melanogaster* recently shown to contain an average of 10 B chromosomes, the number of Bs present varied from cell to cell within the brain tissue (Bauerly 2014). Furthermore, B chromosome number can differ between tissue types and often differs between the gonadal tissues and somatic tissues in many organisms, including several species of grasshopper (Burt and Trivers 2008).

In addition to varying B number, qualitatively different B chromosomes have been known to segregate within a population. Cichlids from Lake Victoria carry two distinct B chromosomes, referred to as B1 and B2, which differ in size (Yoshida 2011). In the fish *Astyanax scabripinnis*, three morphologically distinct chromosomes are found at different frequencies within the same population, along an altitudinal cline (Neo 2000). B chromosomes are not always present in more than one copy or in

even numbers, and they do not necessarily pair with other chromosomes during meiosis. While Bs sometimes pair with other Bs within the cell, they very rarely pair (partially) with an A chromosome during meiosis (Burt and Trivers 2008; Alfenito and Birchler 1993). In fact, by Jones' definition, Bs cannot be entirely homologous to any A chromosome (Jones 1991).

Since their discovery approximately a century ago (Wilson, 1907), Bs have been considered selfish genetic elements that share a parasitic relationship with the rest of the genome (Jones, 1991; Houben, 2013). It is now known that while most Bs are truly parasitic, cases of neutral or even beneficial B chromosomes do exist which can show near-Mendelian transmission (Burt and Trivers, 2008). For example, in a strain of *Drosophila albomicans* it was observed that individuals with 1 or 2 B chromosomes produced a higher number of offspring than those with either no B chromosomes or those with more than 2 B chromosomes. (He 2000). The two most common effects of B chromosomes are on fertility and overall fitness, where fertility typically decreases with B chromosome presence and fitness is usually only affected when the individual has a great many B chromosomes (Camacho 2011; Yoshida 2011; Zhou 2012; Randolph 1941) For the most part, phenotypes are difficult to perceive and are usually detrimental (Yoshida 2011). B chromosomes have been shown to have other effects resulting from sequence located on the B chromosome, including increasing recombination rates (Burt and Trivers 2008; Rhoades 1968), reducing chiasma frequency in males (Camacho 2011), and increased occurrence of nondisjunction in females of *D. melanogaster* where individuals with Bs show 27.0%

nondisjunction of the 4th chromosome over the 0.3% in individuals lacking a B (Bauerly 2014).

B chromosomes are non-essential and typically univalent, and as such, they require special mechanisms to be maintained in populations (Jones et al. 2008). B chromosomes often take advantage of pre-existing meiotic and mitotic machinery to increase their rate of transmission, a process known as drive. Drive can occur in nuclear divisions before, during, or after meiosis as well as at fertilization (Jones 1991; Burt and Trivers 2008). Cytological studies have revealed numerous types of B chromosome drive in plants and animals (Jones and Rees 1982; Jones 1991; Burt and Trivers 2008). However, the molecular basis of these drive mechanisms remains largely unknown, with the notable exception of the rye B chromosome (Banaei-Moghaddam et al. 2012).

Drive and Transmission Mechanisms

The following discussion will focus on the detrimental, or at least neutral, B chromosomes that exhibit non-Mendelian patterns of inheritance. Whether the B chromosome is neutral and present in odd numbers, or whether it is detrimental and under negative selection, some form of drive or self-accumulation mechanism is required for the persistence of such a chromosome (Burt and Trivers, 2008). The form of drive varies greatly among taxa and many examples in various organisms have been examined cytologically (Burt and Trivers, 2008; Camacho, 2011).

A very significant and widespread mechanism utilized by B chromosomes is nondisjunction (Beukeboom 1994; Burt and Trivers, 2008). Normally, when a cell divides, chromosome pairs (homologous chromosomes and sister chromatids) will separate from one another (in meiosis I, and anaphase/meiosis II, respectively) into separate daughter cells. This separation is referred to as disjunction. Nondisjunction causes the pair of homologous chromosomes (meiosis I), or 2 sister chromatids (meiosis II or mitosis), to end up in the same daughter cell. It is important to note that nondisjunction does not increase the overall number of Bs in the population; a parent cell with a single B chromosome still produces 2 daughter cells with a total of 2 B chromosomes between them. If those 2 daughter cells are gametes, or have an equal chance of producing gametes, then the number of B chromosomes will not increase in frequency in the next generation. The only difference is that those 2 Bs are contained in a single cell, rather than 2. In order for the population frequency of the B chromosome to increase via nondisjunction at a meiotic or mitotic division, the cell with 2 Bs must outcompete the cell without Bs or have a better chance of ending up in the offspring, by either preferential segregation or a B-induced increase in cell division within the germline (Burt and Trivers 2008). Preferential segregation is the increased likelihood of a chromosome or pair of chromosomes segregating into a specific daughter cell. It can be particularly important during female meiosis, where a B chromosome risks ending up in a polar body rather than the egg-generative nucleus. Preferential segregation can skew the typical 1:1 ratio of inheritance and increase the transmission of a B (or nondisjoined Bs) to the cells that are more likely to produce offspring. Examples of known drive mechanisms are discussed below.

There are 4 crucial times for self-accumulation through transmission: Pre-meiosis, meiosis, post-meiosis and fertilization (Burt and Trivers, 2008 and Jones, 1991).

Pre-meiotic mechanisms of drive manipulate mitotic divisions. Since even a single B that goes through normal mitosis will result in two daughter cells with one B each, typically, a self-accumulation mechanism exploiting this time frame would require nondisjunction. For this nondisjunction to contribute to B accumulation, it needs a) to use preferential segregation, and b) to occur during a pivotal mitotic division. The differentiation between germ and somatic cells occurs early in embryonic development. If the cell prior to this differentiation carries at least one B, and if that B goes through the mitotic division with nondisjunction and segregates to the daughter cell that will develop into the germ line, then the B will not only be transmitted to the offspring, but at a higher frequency. This is depicted in Figure 1.1. The combination of nondisjunction and preferential segregation into the germ line cell makes this possible, but exactly how this is accomplished at the molecular level is still unknown. This is seen in the grasshopper *Calliptamus palaestinesis,* which has a varying number of mitotically unstable B chromosomes within and between tissue types. While most of the somatic tissue has a consistent number of Bs (only one), the germ line cells possess either 2 Bs (2B) or 0 Bs (NoB). The ratio between these two possibilities is 15:1, respectively (Jones, 1991; Burt and Trivers, 2008). The 2B is more prevalent because nondisjunction is followed by preferential segregation

towards the germline. There are still NoB germ line cells because the process is not 100% effective.

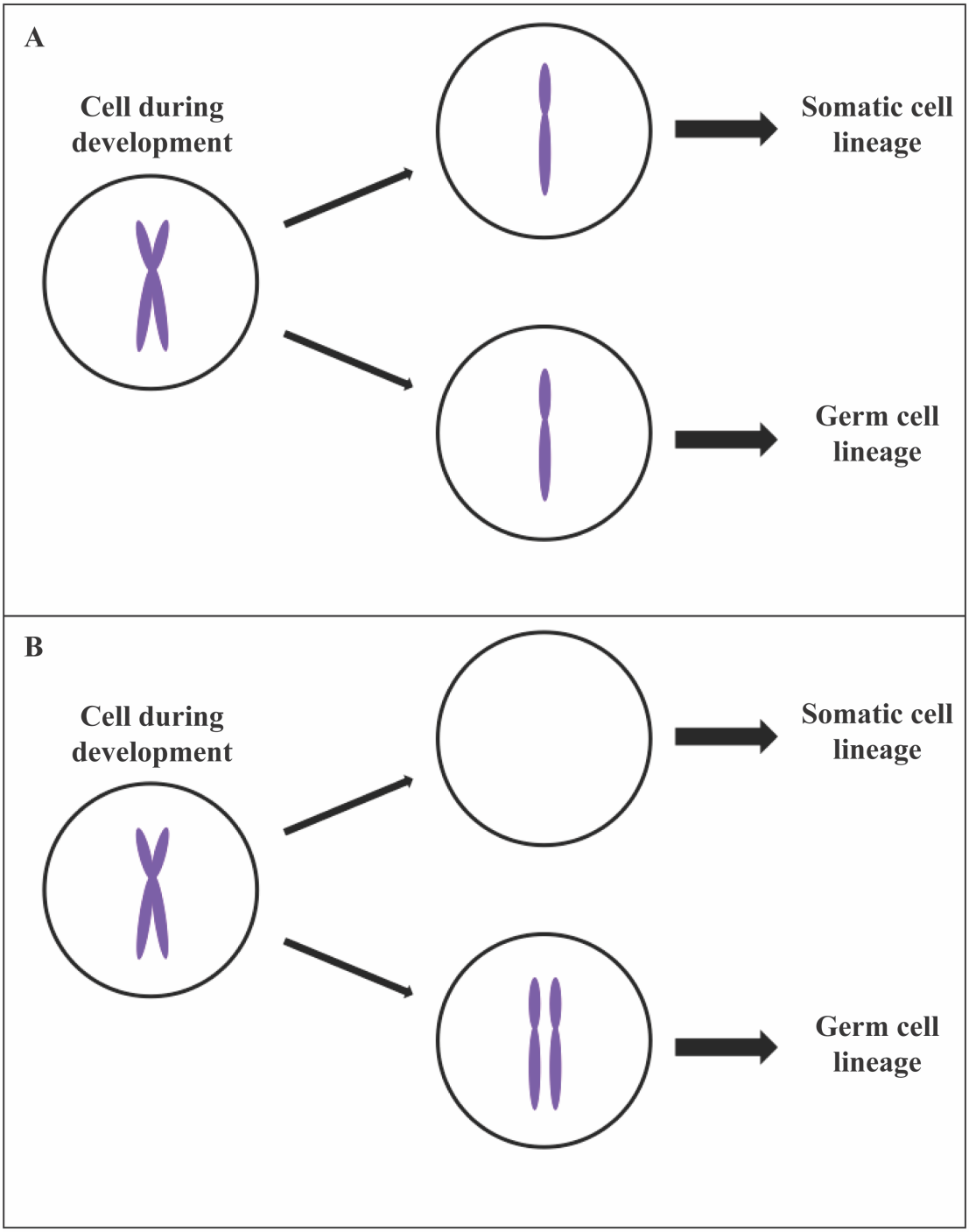

Figure 1.1: Pre-meiotic drive. Mitotic divisions typically result in the segregation of sister chromatids. This disjunction is shown in **A**. Failure to segregate, or nondisjunction in mitosis results in one daughter cell lacking the non-disjoined

chromosome, and the other with both sister chromatids. As portrayed in **B**, preferential segregation can ensure the non-disjoined sister chromatids are transmitted to the daughter cell that will generate the germline, resulting in drive.

Meiotic drive can occur via a variety of mechanisms. These mechanisms can exploit sex-specific traits so that accumulation occurs only in one sex. During meiosis in females, preferential segregation can lead to Bs segregating to the egg and avoid being lost in the polar body, as shown in Figure 1.2. This is the mechanism seen in the lily *Lilium callosum*, the first organism studied to show B chromosome drive (Kayano 1957; Jones 1991) as well as *Myrmeleotettix maculatus,* (Jones 1991). In *Lilium callosum,* crosses were conducted between 1B males and NoB females as well as NoB males and 1B females. The 1B male and NoB female crosses showed Mendelian inheritance, but the NoB male and 1B female crosses produced 80% 1B and 20% NoB progeny, pointing to a female-based accumulation mechanism (Kayano 1957; Jones 1991). Upon cytogenetic investigation, it was found that this observed drive is caused by a high frequency of ovules where the B is located in the megaspore destined for fertilization, as opposed to a triploid nucleus destined to degenerate. Similarly, in *M. maculatus*, a high frequency of cells with a higher frequency of Bs located on the micropylar side (egg side) of the spindle during meiosis was observed. Meiotic manipulations can also occur through the male sex, as seen in *Rhinocola aceris* and *Psylla foersteri*. Here, the B chromosome will pair with and segregate away from the X chromosome, leading to its accumulation in males (Nokkala, 2000). While this is not in itself sufficient for drive, these B chromosomes are believed to have higher fitness in or transmission through males. As a result, this mechanism

increases their transmission by ensuring they are always in the sex in which they have the highest fitness.

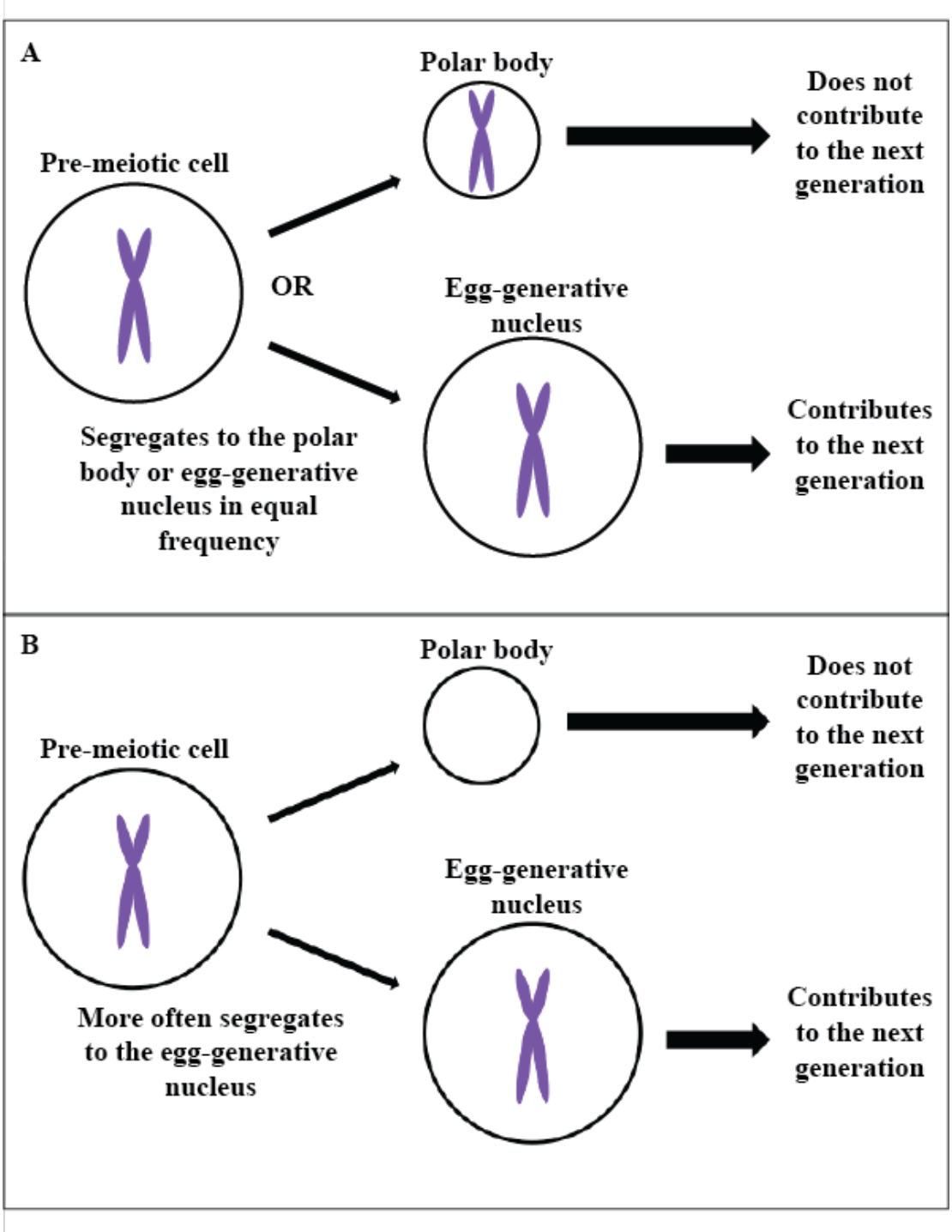

Figure 1.2: Example of meiotic drive in female meiosis. The first meiotic cell division is shown above. A single, unpaired chromosome is expected to segregate to either the

polar body or the egg-generative nucleus in equal proportions in female meiosis I. If it segregates to the polar body, it is lost. If it segregates to the egg-generative nucleus, it will be transmitted to the next generation. This is depicted in **A**. Preferential segregation can lead to the B segregating more frequently to the egg-generative nucleus, as shown in **B**, resulting in drive.

Yet another, unusual meiotic method is seen in mealybugs, *Pseudococcus affinis*, where accumulation also occurs through males. In males, the paternally inherited chromosome set is heterochromatic through meiosis and degenerates leaving the euchromatic maternal set to be inherited by the next generation. B chromosomes in males become euchromatic during prophase I, allowing them to be inherited with the maternal set, guaranteeing that Bs from males are passed to 90% of the offspring (Jones, 1991).

A common type of B drive in angiosperms is accomplished by post-meiosis directional nondisjunction. After meiosis, there is a division between the generative nucleus and cells that will contribute to the pollen tube, but will not contribute to the next generation. Nondisjunction followed by preferential segregation into the generative nucleus prevents the B from being lost in the pollen tube. Maize (*Zea mays*) exhibits a post-meiotic method that involves nondisjunction during mitosis II in males and a higher rate of success in fertilization by the B-containing sperm (Jones, 1991).

These cover the majority of known mechanisms, but Bs are quite diverse and evolve mechanisms particular to their systems. There are potentially even more unique methods of self-accumulation and persistence not yet described.

B Chromosomes and Sex

Associations between B chromosomes and sex have long been recognized in many species (Camacho et al. 2000; Camacho et al. 2011). Many B chromosomes drive in one sex, but not the other (Burt and Trivers 2008, Jones 2018). The plants *Lilium callosum* (Kimura and Kayano 1961), *Phleum nodosum* (Frost 1969), *Plantago serraria* (Frost 1959), and *Trillium grandiflorum* (Rutishauser 1956) as well as the animal species *Pseudococcus obscurus* (Nur and Brett 1985) and several species of grasshopper (Hewitt 1976; Nur 1977; Cano and Santos 1989; Santos et al. 1993) experience drive only in females. The plant species *Haplopappus validus* (Jones 1982), *Clarkia elegans* (Jones 1982), *Iseilema laxum* (Jones 1982), and *Briza humilis* (Murray 1984) as well as the bushrat *Rattus fuscipes* (Thompson et al. 1984) drive only in males. Furthermore, while a B may drive in one sex, it can also exhibit reduced transmission, or drag, in the other sex (Jones 2018).

Most B chromosome systems exhibit similar frequencies of B-carriers in both sexes, though some have a higher frequency in one sex or the other, and others lack B-carriers in one sex altogether. A common example of this is the jewel wasp, *Nasonia vitripennis.* The sex of a jewel wasp is determined by its ploidy such that unfertilized eggs produce haploid males and fertilized eggs produce diploid females. The jewel wasp B chromosome benefits from being in a male individual and has evolved an interesting method of ensuring that it does not end up in a female. When sperm possessing a B chromosome fertilizes an egg, the B chromosome causes the chromatin remodeling of the paternal set of A chromosomes resulting in their loss in

early mitotic divisions. The B chromosome is the sole remaining paternal chromosome and is incorporated into the maternal haploid set, resulting in a haploid male individual (Beukeboom and Werren, 1993). Among the characid fishes, *Astyanax scabripinnis* demonstrates a system in which B chromosomes are found more frequently in females. Additionally, several intersex individuals were identified, all of which possessed B chromosomes (Vicente et al. 1996; Neo et al. 2000). In one population of the characid fish *Moenkhausia sanctaefilomenae*, B chromosomes are found in males, but not females (Camacho et al. 2011). B chromosomes in the fairy shrimp *Branchipus schaefferi*, are also found solely in males and the number of B chromosomes is associated with the sex ratio of the population (Beladjal et al. 2002; Burt and Trivers 2008). Bs have been shown to influence sex ratio in other species as well (Nur, 1966; Camacho, 2011; Yoshida, 2011; Zhou, 2012). In each of these examples, it is believed that the difference in B-carrier sex ratio is a result of the drive mechanism or a secondary mechanism meant to further increase the opportunity to drive. Either the drive mechanistically results in one sex having more B chromosomes, or a mechanism evolves to ensure the B is more frequently found in the sex in which it drives.

B chromosomes have been observed to associate with sex chromosomes during meiosis, resulting in their transmission more frequently to a certain sex. This has been seen in B-carrier males of some Orthoptera, where drive has been observed to occur in females, not males. In the grasshopper *Tettigidea lateralis* the B chromosome associates and segregates with the X chromosome during male meiosis, ensuring that

the B chromosome is more frequently transmitted to females where it has the benefit of drive (Burt and Trivers 2008; Fontana and Vickery 1973). The opposite has been observed in two other grasshopper species, *Phaulacridium vittatum* (Jackson and Cheung 1967) and *Euprepocnemis plorans* (Lopez-Leon et al. 1996) where the B chromosome segregates away from the X and is therefore transmitted more frequently to males. B chromosome segregation away from the X has also been observed in two species of Hemiptera, *Rhinocola aceris* and *Psylla foesteri* (Nokkala et al. 2000).

Beyond meiotic association of B chromosomes and sex chromosomes, several studies have suggested an evolutionary transition between these two chromosome types. In the grasshopper *E. plorans* the B chromosome was shown through double fluorescent *in situ* hybridization to share sequence with the X chromosome, suggesting the X was the chromosomal origin of the B (Lopez-Leon et al. 1994). In Drosophila species, sex is determined by the ratio of autosomes to X chromosomes. The Y chromosome is not necessary for or involved in sex determination. Because the Y does not experience crossing over and has very little sequence homology with the X, it was proposed that the Y in these species is not a degenerated homolog of the X (Hackstein et al. 1996). Rather, these authors use the Y chromosome's similarities to B chromosomes (i.e. its highly repetitive sequence, increased rDNA content and the variability in size and shape across population) to suggest the Drosophila Y evolved from a B chromosome. Yet another example among fruit flies, *Drosophila albomicans* has neo-sex chromosomes that resulted from the fusion of a sex chromosome with an autosome. The B chromosomes of *D. albomicans* were found to share sequence with

subcentromeric regions of the ancient X and neo-X chromosome (Zhou et al. 2012). It is believed that the B of this species originated from a chromosomal byproduct of the autosome-sex chromosome fusion. And finally, in the frog *Leiopelma hochstetteri*, sequence homology between the univalent W chromosome and B chromosomes of this species was demonstrated with Southern hybridizations, suggesting either the B originated from the W or the W originated from the B (Green 1988; Green et al. 1993; Sharbel et al. 1998).

B Chromosome Sequence

B chromosomes are thought to arise from the A chromosomes (Burt and Trivers 2008; Martis et al. 2012; Houben et al. 2013). Fluorescent in situ hybridization (FISH) has revealed that Bs often share homologous sequence with at least one A chromosome (Martis et al. 2012; Silva et al. 2014). An evolutionary model from a Lake Victorian cichlid suggests that B chromosomes can arise as a segmental duplication of an A chromosome that includes a centromere but relatively few genes so that it avoids negative selective pressures arising from unbalanced gene dosage. This proto-B might then undergo an internal duplication, producing an isochromosome with two nearly identical arms (Valente et al. 2014). This model also includes subsequent accumulation of A sequences. Once these A sequences are inserted onto the B, they can undergo duplication and reach high copy number on the B. Most of these sequences eventually undergo decay because they experience little, if any, purifying selection. Regardless of the origin of the proto-B, the continued

accumulation and subsequent duplication of sequences leads to a high repeat content on B chromosomes (Martis et al. 2012; Zhou et al. 2012; Valente et al. 2014).

Approximately half of all known Bs are highly heterochromatic (Jones 1975; Tanic et al. 2005; Camacho et al. 2000). Despite the fact that B chromosomes add significant amounts of genetic material to the genome, B chromosomes have rarely been associated with novel phenotypes, the most frequent exception being an effect on fertility discussed above (Jones and Rees 1982; Burt and Trivers 2008; Jones 2017; Zhou et al. 2012; Gonzalez-Sanchez et al. 2004). With a limited list of known B-specific sequences and few or no visible phenotypes beyond drive, the prevalent view has been that B chromosomes carry few genes and are largely unexpressed and inactive (Houben et al. 2014; Jones et al. 2008). They have been thought to be composed of nonfunctional "junk" DNA together with one or two genes contributing to drive (Bugrov et al. 2007). More recent findings have been contradicting some of these notions as access to next-generation sequencing technologies has led to a surge in B chromosome research. The past few decades have provided us with several insights into the sequence and activity of B chromosomes, though progress still remains limited by technical challenges.

B chromosomes often contain large amounts of highly repetitive DNA (Camacho et al. 2000; Cheng and Lin 2003; Bugrov et al. 2007; Ruiz-Ruano et al. 2018) and are frequently either partially or completely heterochromatic (Jones and Rees 1982; Camacho et al. 2000; Burt and Trivers 2008). In several species, it has been shown

that B chromosomes share homology with sequences from all or many of the A chromosomes (Jones and Houben 2003) (the grasshopper *Podisma kanoi* (Bugrov et al. 2007), the fish *Astatotilapia latifasciata* (Valente et al. 2014), rye *Secale cereale* (Martis et al. 2012), and maize *Zea mays* (Cheng and Lin 2003)). This suggests that sequences on B chromosomes are derived from the A chromosomes through as yet uncharacterized mechanisms of gene duplication (Houben et al. 2014). Theoretically, because they are nonessential, B chromosomes should experience relaxed selective pressures (Houben et al. 2014; Klemme et al 2013). For this reason, they might be expected to experience high rates of sequence turnover. B chromosomes are continuously acquiring new sequences. Sequences already on the B collect mutations at a high rate, and most are eventually lost. It has been difficult to produce sequence assemblies of B chromosomes due to their repetitive nature and their high levels of homology with sequences in the A chromosomes (Ruban et al. 2017; Makunin et al. 2014; Banaei-Moghaddam et al. 2015; Makunin et al. 2016; Ma et al. 2017).

Examples of genic sequences detected on B chromosomes include the C-KIT gene in two canid species (Graphodatsky et al. 2005), ribosomal RNA (rRNA) genes and thousands of genes and gene fragments in the fish *Astatotilapia latifasciata* (Valente et al. 2014; Poletto et al. 2010), protein coding genes in the grasshopper *Eyprepocnemis plorans* (Navarro-Dominguez et al. 2017), and protein-coding genes in two mouse species from the genus *Apodemus* (Makunin et al. 2018).

Current approaches to identifying B sequences can be categorized into two types: direct and indirect (Ruban et al. 2017). Direct methods, such as the sequencing of B chromosomes isolated through flow sorting or microdissection, have a high rate of contamination (Valente et al. 2014; Ruban et al. 2017) and are only possible in a few organisms. Indirect methods, such as the comparison of whole genome sequence data between samples with or without a B chromosome, can be performed on any species. For many species, the sequence reads can be aligned to a reference genome assembled from an individual lacking a B chromosome, allowing a characterization of a B sequence by its alignment to homologous portions of the A genome. While Illumina sequencing has dramatically lowered costs, there are significant limitations to Illumina sequence data (Treangen and Salzberg 2011). Namely, Illumina reads are very short and are not very useful for assembling the repetitive sequence of B chromosomes. However, the extent to which short reads can be used to identify B chromosome sequence has not been fully explored.

With the discovery of genic sequences on B chromosomes came the question of function; were these just pseudogenes or were they expressed and could they contribute to phenotypic change? Towards this end, transcription of B chromosomes has been investigated through various techniques. The conversion of RNA to complimentary DNA (cDNA) followed by quantification with quantitative PCR (qPCR) or real time PCR (RT-PCR) was used to look for differential expression between individuals with and without B chromosomes, such as in the yellow-necked mouse *Apodemus flavicollis* (Tanic et al. 2005), and was the first method used to

confirm transcription from a B chromosome in the smooth hawksbeard *Crepis capillaris* (Leach et al. 2005). In *C. capillaris*, two ribosomal RNA gene families were transcribed from the B chromosome (Leach et al. 2005). In *A. flavicollis*, 3 genes were found to be differentially expressed between individuals with and without a B, with higher expression in the B individual suggesting the transcription came from the B chromosome (Tanic et al. 2005). Following this, these methods were applied to various other taxa. In maize *Zea mays*, analysis of cDNA fragment length polymorphism identified 6 transcripts, 2 of which had expression correlated to the number of B chromosomes present (Lin et al. 2014). In rye *Secale cereale*, RT-PCR revealed 15% of pseudogene-like fragments on the B were transcribed in a tissue-specific manner among roots, leaves and anthers (Banaei-Moghaddam et al. 2013). The authors also demonstrated differential expression of genes not known to be on the B chromosome. In the grasshopper *Eyprepocnemis plorans*, cytological examination of nucleoli coupled with PCR of B specific sequence demonstrated transcription of rDNA (Ruiz-Estevez et al. 2012). However, qPCR revealed the B contributed proportionally very little of the rRNA in the cell suggesting this transcription from the B had little effect (Ruiz-Estevez et al. 2014). Later, qPCR in *E. plorans* detected transcription of a single pseudogenized gene, the CAP-G subunit of condensing I (Navarro-Dominguez et al. 2017). In the Siberian roe deer *Capredus pygargus*, karyotype analysis led to the detection of a 2 Mb region on the B chromosome homologous to chromosome 3, containing a partial copy of one gene and complete copies of two other genes (Trifonov et al. 2018). The authors identified B-specific mutations among these genes that, combined with qPCR, confirmed the

complete gene copy of FPGT was expressed from the B chromosome. Among fish species, RT-PCR was used to demonstrate differential expression of a masculinizing gene, DMRT1, in B individuals of *Astyanax scabripinnis* (Castro et al. 2018) and B chromosome transcription and differential RNA-processing of a single non-coding gene in the cichlid fish *Astatotilapia latifasciata* (Ramos et al. 2017). These studies represent a tremendous advancement in our understanding of B chromosome biology. However, these methods present a bottleneck as they only result in the detection of transcription for a handful of genes.

Even more recently, RNA sequencing and transcriptome analysis have been applied to studies of B chromosome transcription. In the grasshopper *E. plorans*, a differential expression analysis detected 188 genes (Navarro-Dominguez et al. 2019). The authors removed all known B transcripts and found 46 genes among the A chromosomes with differential expression, 30 up-regulated and 16 down-regulated in B individuals. Transcriptomes were also analyzed to identify 10 protein coding genes on the B chromosome with differential expression (Navarro-Dominguez et al. 2017). In the cichlid *A. latifasciata*, a differential expression analysis revealed that even though the B chromosome was enriched for transposable elements, few were differentially expressed between individuals with and without a B and those that were showed low expression (Coan and Martins 2018). In maize, 130 genes were found to be differentially expressed with an effect proportional to the number of B chromosomes present (Huang et al. 2016). Fluorescent in situ hybridization was used to confirm 4 of these genes were located on the B chromosome. Flow-sorted rye B chromosomes

were used to construct a gene-centered assembly of B sequences to which transcriptomes were aligned resulting in the detection of 1954 B-specific transcripts in germline tissue and 1218 B-specific transcripts in somatic tissue (Ma et al. 2017). The gene-focused assembly analysis used by these authors captured transcripts that mapped uniquely to B sequence and not homologous A sequence, a high percentage of which were short and perhaps transcribed from pseudogenes. The authors continued to examine one B-transcribed protein-coding gene AGO4B to demonstrate the production of a functional protein, confirming the rye B chromosome carries both pseudogenized genes as well as functional genes (Ma et al 2017).

African Cichlids

East African cichlids have been identified as a powerful model system for studying adaptive evolution and speciation (Kocher, 2004). The combination of using cichlids as a model system to address genetic-based evolutionary questions and the recent dramatic decrease in sequencing costs has lead to the sequencing and mapping of 5 cichlid genomes, including a Lake Malawi cichlid, *Metriaclima zebra*. These mapped genomes, established cichlid lab protocols, and more budget-friendly sequencing technologies are tools that have provided a unique opportunity for studying the B chromosomes found in several cichlid species.

B chromosomes were first identified in cichlid species from South America (Feldberge and Bertollo 1984; Feldberg et al. 2004; Pires et al. 2015). More recently, they have been identified also in several species from Lake Victoria and one species

from Lake Malawi in East Africa (Poletto et al. 2010). *Astatotilapia latifasciata*, an African species from Lake Nawampasa in the Lake Victoria basin, carried either 1 or 2 metacentric B chromosomes in 38 of 96 individuals, both male and female. All of the kidney cells analyzed from B-carrying individuals contained a B chromosome, suggesting mitotic stability. In individuals with 2 B chromosomes, the Bs did not appear to pair during meiosis. Instead, they formed ring-like univalents, consistent with their isochromosomal structure (Poletto et al. 2010). B chromosomes were subsequently found in each of 12 cichlid species analyzed from Lake Victoria (Yoshida et al. 2011; Kuroiwa et al. 2014). Two morphologically distinct Bs that share repetitive sequences were found in Lake Victoria, where they were found in both sexes in most species (Fantinatti et al. 2011; Yoshida et al. 2011). In one species, *Lithochromis rubripinnis*, all of the B-carrying individuals were female, but not all females carried the B (Yoshida et al. 2011). Yoshida and his collaborators performed a series of crosses to examine *L. rubripinnis.* These crosses examined B transmission to offspring, sex ratio, and ratio of male to female B-carriers. They concluded that the B chromosome in this species had a functional effect on sex, i.e. it was acting as a feminizing sex determiner. Sex determination among cichlids has experienced a dramatic amount of turnover. There have been more than a dozen sex determination systems mapped within African cichlids alone (Gammerdinger et al. 2018). It has been hypothesized that this turnover is fueled by genomic conflict such as sexually antagonistic selection of alleles genetically linked to a sex determiner (Doorn and Kirkpatrick 2007; Roberts 2009).

B chromosomes were also identified in *Metriaclima lombardoi*, a cichlid species from Lake Malawi (Poletto et al. 2010). Karyotype data for *Metriaclima lombardoi* shows the B chromosome is one of the largest chromosomes, representing approximately 4.5% of the genome when present. Divergence of the cichlid flocks of Lake Victoria and Lake Malawi occurred no longer than 8 million years ago (MYA) and all cichlids within Lake Malawi share a common ancestor no more than 1 MYA (Sturmbauer et al. 2001). Of 22 *M. lombardoi* individuals analyzed, 9 females carried a single large B chromosome, but no males were found with a B. The individuals examined were collected from the aquarium trade in Brazil and the Tropical Aquaculture Facility at the University of Maryland. Thus, it is unclear whether the frequency of Bs in this stock accurately reflects the frequency in wild populations.

The sequence of a B chromosome from *A. latifasciata* was further examined by whole genome sequencing (Valente et al. 2014). *Astatotilapia latifasciata* individuals with 2 Bs and without B chromosomes as well as a microdissected B were sequenced and the resulting reads were aligned to a closely related reference genome. Thousands of regions across the genome showed significantly higher coverage in the individual with the B chromosome. These regions representing B chromosome sequences are referred to as “B chromosome blocks” or “B blocks.” Copy number for several of these B blocks (estimated by qPCR) was tightly correlated with the B chromosome numbers determined by karyotype. These data support the idea that large portions of the B chromosome originate from A chromosome material, and many of these sequences are found in high copy numbers on the B. This analysis identified

thousands of gene fragments and tens of complete genes on the B chromosome. Sequence of microdissected B chromosomes detected only a small portion of the overall B chromosome in this study. Whole-genome sequencing data also revealed that the reference genome assembly of *Pundamilia nyererei* (Brawand et al. 2014) from Lake Victoria contained a B chromosome highly similar to the B chromosome of *A. latifasciata* (Valente et al. 2014). Analysis of the *P. nyererei* transcriptome data (Brawand et al. 2014) also revealed B chromosome-specific transcription of several genes in multiple tissues (Valente et al. 2014).

Chapter 2: Quantifying the Presence of B Chromosome in Lake Malawi Cichlids

This chapter is published as:

Clark, F., Conte, M., Ferreira-Bravo, I., Poletto, A., Martins, C. and Kocher, T. (2017). Dynamic Sequence Evolution of a Sex-Associated B Chromosome in Lake Malawi Cichlid Fish. *Journal of Heredity*, 108(1), pp.53-62.

Context and Motivation

B chromosomes were previously detected with cytogenetic methods in twelve species of East African cichlid, including one from Lake Malawi, *Metriaclima lombardoi* (Poletto et al 2010; Yoshida et al. 2011; Fantinatti et al. 2011; Kuroiwa et al. 2014). In this species, all 9 B-carrying individuals discovered were female and each had a single B chromosome per diploid genome. B chromosomes are known to vary from individual to individual within a population. Indeed, the first African cichlid species, *Astatotilapia latifasciata*, found to carry B chromosomes had male and female individuals with 0, 1, or 2 Bs. Another cichlid, *Lithochromis rubripinnis* in Lake Victoria has individuals with 0, 1, 2 or 3 Bs. With further sampling, would this also be the case in *M. lombardoi* or were the results of this small sampling indicative of a difference in distribution in this species? To address this question of B chromosome distribution in *M. lombardoi*, a larger sample size, from *M. lombardoi* and other Lake Malawi cichlid species, was investigated. Genotyping for presence or absence of a B

chromosome was carried out by means of the polymerase chain reaction (PCR). Since every species of cichlid in Lake Victoria inspected to date was found to possess B chromosomes, additional species in Lake Malawi were also genotyped for B presence.

Methods

The 18 male and 18 female *M. lombardoi* used for cytogenetic analysis were obtained from stocks maintained at the Tropical Aquaculture Facility at the University of Maryland and the aquarium trade in Brazil. Individuals were euthanized using tricaine methanesulfonate (MS-222) and inspected for testes or ovaries to confirm sex. Mitotic chromosome preparations were obtained from kidney tissue according to (Bertollo et al. 1978), with modifications (Poletto et al. 2010). DNA was extracted from kidney tissue for these karyotyped individuals using standard phenol chloroform methods. Male and female *M. lombardoi, Metriaclima zebra* "Boadzulu," *Metriaclima greshakei, Metriaclima mbenji, M. zebra* "Nkhata Bay," *Labeotropheus trewavasae*, and *Melanochromis auratus* fin clips were collected from the wild in 2005, 2008, 2012, and 2014. DNA was extracted from fin tissue using standard phenol chloroform methods. Purified genomic DNA was quantified on a BioTek FLx800 using Pico-green and normalized to a concentration of 0.5ng/μL. These samples were used for sequencing, PCR and qPCR analysis. DNA was extracted from kidney tissue for the previously karyotyped individuals using standard phenol chloroform methods. Purified genomic DNA was quantified on a BioTek FLx800 using Pico-green and normalized to a concentration of 0.5ng/μL. The resulting DNA

samples were used in PCR and qPCR analysis.

DNA libraries were prepared from the pooled DNA of 20 male or 20 female *M. zebra* "Boadzulu" individuals. The TruSeq DNA sample preparation kit ver.2 rev.C (Illumina) was used for library construction. Libraries were sheared to an average size of 500 bp and 100 bp paired-end reads were sequenced on Illumina's HiSeq 1500 platform. Raw sequencing reads were evaluated with FastQC (Babraham Bioinformatics) to remove reads of poor quality. Reads were aligned to the unmasked *M. zebra* reference genome ("M_zebra_v0" available at www.bouillabase.org, Brawand et al. 2014) using Bowtie2 v2.02 with the parameter "--very-sensitive" (Langmead and Salzberg 2012).

Read coverage was compared between males and females in a genome browser and blocks of sequence with 10-fold or higher difference in coverage between females versus males were found. These blocks, similar in pattern to those found in *A. latifasciata* (Valente et al. 2014), are referred to as B chromosome blocks or B blocks.

Primers were designed using *Metriaclima zebra* "Boadzulu" sequence. SNPs identified within the B blocks were incorporated into the primers so that the primers would distinguish between homologous A and B sequence and amplification would

be B-specific. SNPs were incorporated such that 1–3 B-specific SNPs were present in at least one primer, forward or reverse, for each pair. Primer3, v0.4.0, was used to calculate the expected melting temperature and evaluate the primer sequences (Untergrasser et al. 2012). PCR reactions contained 5 µL of Life Technologies' Dream Taq, 0.5 µL (10 µM/L) forward primer, 0.5 µL (10 µM/L) reverse primer, 3 µL water and 1 µL (0.5ng/µL) DNA. PCR products were separated on 2% agarose gels.

Real-time amplifications were recorded on a Roche LightCycler LC480 thermocycler. Quantitative real-time PCR (qPCR) reactions contained 10 µL of Life Technologies' Maxima SYBR Green/ROX, 1 µL (10 µM/L) of forward primer, 1 µL (10 µM/L) of reverse primer, 2 µL water and 6 µL (0.5 ng/µL) DNA. Each sample was amplified with 3 technical replicates, the average of which was used for all future calculations. Starting template quantities (*T*) were calculated using the following equation:

$$T = \frac{1}{E^{Ct}}$$

where *E* is the PCR efficiency of the primer set used and C_t is the critical cycle number from the qPCR reaction. Relative copy number was calculated using a control primer set. The control primer set amplifies the single copy cichlid SWS1 (UV) opsin locus which is present in the A genome, but not the B chromosome. Relative copy number was calculated by using the ratio of starting template of each B block primer pair over the control primer pair. Hierarchical cluster analysis of individuals by

sequence copy number was performed using SPSS Statistics 23 software.

Identification of High Coverage Blocks

Pools of 20 wild-caught male and female *Metriaclima zebra* "Boadzulu" were sequenced and the resulting reads were aligned to the *M. zebra* reference genome (M_zebra_v0). An example of the coverage differences between male and female pools is shown in Figure 2.1. Males and females show similar coverage across most of the genome, roughly 25.5×, but as shown in Figure 2.1 there are blocks of sequence with much higher coverage in females than in males. There are thousands of these short blocks of high sequence coverage, distributed across all linkage groups. The blocks were found exclusively in females. The female-limited presence of these blocks is consistent with the karyotypic detection of Bs in females, but not males, of *M. lombardoi* (Poletto et al. 2010). These blocks are similar in pattern (i.e., length, coverage increase, variability of coverage across block) to those found in *A. latifasciata* (Valente et al. 2014), but their identities and locations are different (data not shown). There is very little overlap among the sequence blocks found in the 2 species. These blocks may represent repetitive B chromosome sequence, a hypothesis tested below.

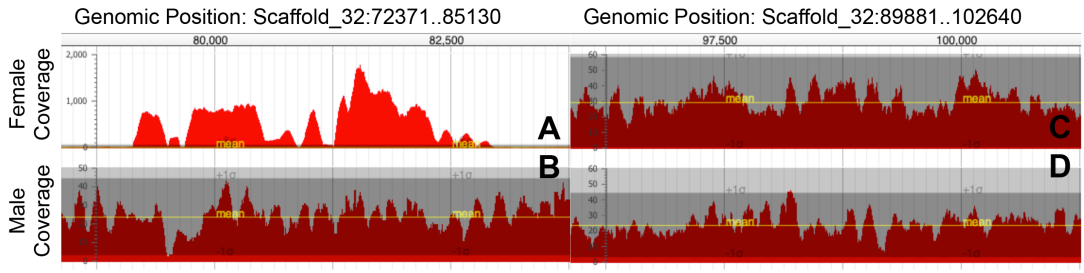

Figure 2.1: Read coverage of female and male *Metriaclima zebra* "Boadzulu" at a B block. Read coverage at 2 locations on scaffold_32 for female (**A** and **C**) and male (**B** and **D**) pooled samples of 20 *Metriaclima zebra* "Boadzulu" individuals. The location in plots **A** and **B** (72371–85130 bp on Scaffold 32) includes a B block, visible in plot **A**. Plots **C** and **D** represent sequence without a B block. Reads were aligned to the *M. zebra* reference genome. *Note:* the *y* axis differs between plots.

High coverage sequences from *M. zebra* "Boadzulu," located in 11 separate scaffolds of the *M. zebra* A genome reference assembly (M_zebra_v0), were used to design a total of 21 primers sets for PCR amplification. Sequence corresponding to the B was distinguishable from A sequence by high frequency SNPs found within the high coverage blocks from the female pool but not found in the sequences from the male pool. These SNPs were incorporated into the forward primer, the reverse primer, or both. While the sequences appear to be continuous on the A chromosomes, the homologous sequences on the B may have undergone structural rearrangement, preventing efficient amplification with some primer sets. Of the 21 primer sets designed, 7 amplified the expected fragments, 6 amplified products too large to use for qPCR, 4 amplified sequence not specific to the B chromosome, 3 amplified a complex set of fragments, and 1 failed to amplify any sequence at all. Five primer sets amplifying the fragments of the expected size were selected for further analyses.

The 5 primer sets were used to amplify DNA from *M. lombardoi* individuals that had been karyotyped and found to either carry the B ($N = 8$) or not to carry a B ($N = 5$). The karyotypes of a female with a B chromosome, a female without a B chromosome and a male without a B chromosome are shown in Figure 2.2. A subsample of the PCR data is shown in Figure 2.3. Amplification of all 5 primer sets was observed in each karyotyped individual with a B. No amplification was observed in any of the karyotyped individuals without a B. These data demonstrate that the primer sets are B-specific. The high coverage sequence blocks are hereinafter referred to as B blocks.

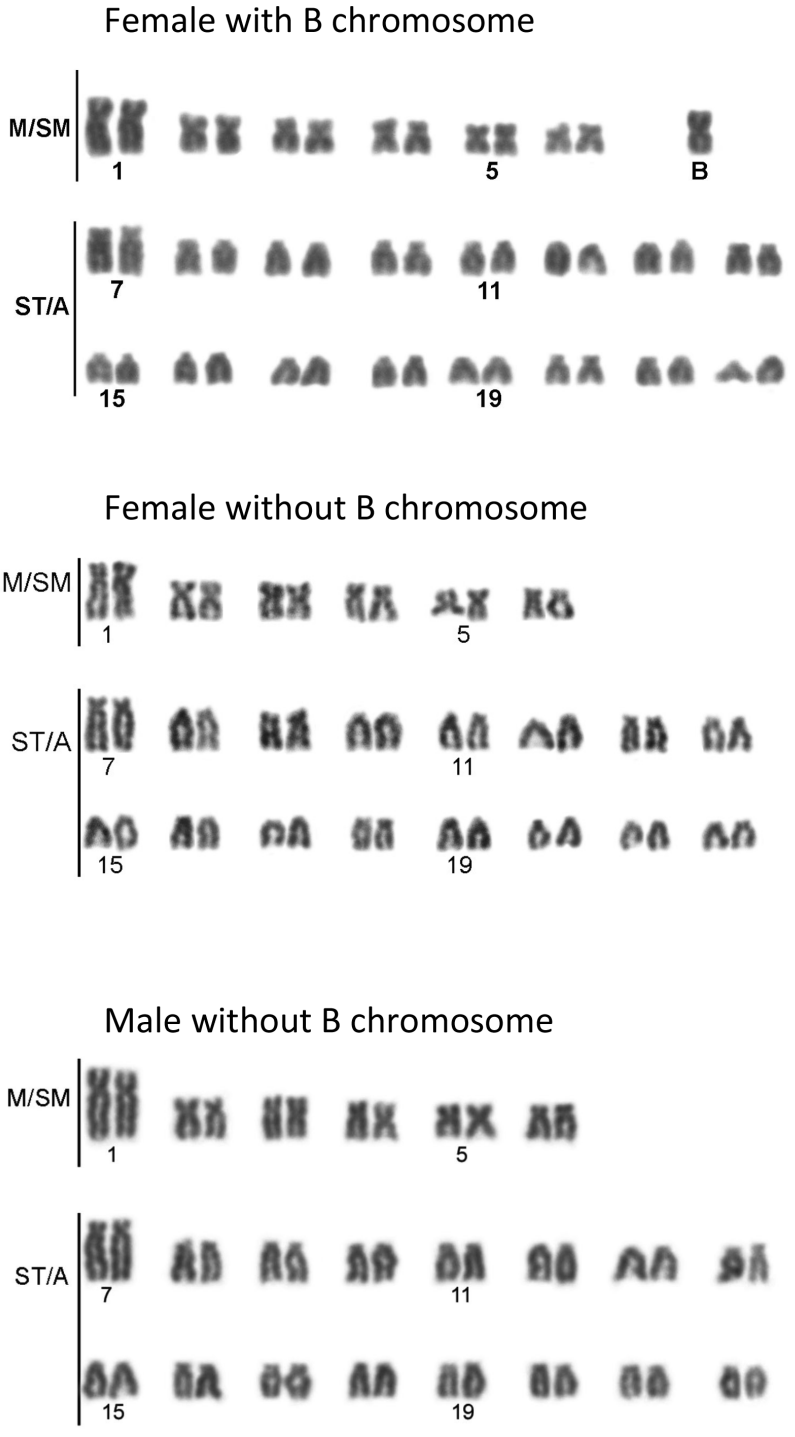

Figure 2.2: Karyotypes. Giemsa-stained karyograms of an *Metriaclima* [SEP] *lombardoi* female B-carrier, an *M. lombardoi* female without a B and an [SEP] *M. lombardoi* male without a B.

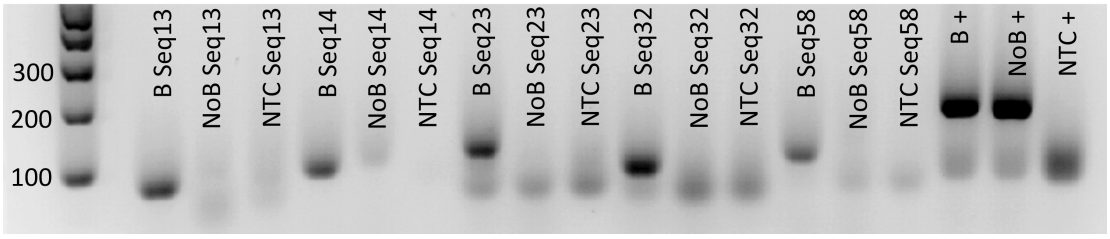

Figure 2.3: B-specific amplification. Agarose gel (2%) of PCR product resulting from amplification with either control or B-specific primers. The control primer (+) set amplifies the single copy cichlid SWS1 (UV) opsin locus which is present in the A genome, not the B chromosome. Amplification with the control primer is a positive control to indicate amplifiable DNA. DNA from 2 individuals was used, 1 female known cytogenetically to carry a B chromosome (B) and 1 female known cytogenetically to not carry a B (NoB). Amplification of a non-template control (NTC), containing no DNA, was also used for each primer set as a negative control.

Prevalence of B Chromosomes in Lake Malawi Cichlids

Because these primer sets are B-specific, we can use them to assay for the presence/absence of B chromosomes in additional individuals collected from the wild whose karyotypes are unknown. B chromosomes were identified in 6 additional species: *Metriaclima zebra* 'Boadzulu,' *M. zebra 'Nkhata Bay', M. greshakei, M. mbenji, Labeotropheus trewavasae* and *Melanochromis auratus* (Table 2.1). There remains some debate as to whether some of these listed species (*M. zebra* 'Boadzulu' and *M. zebra* 'Nkhata Bay') are in fact different species or simply different populations of the same species (Kocher 2004; Turner et al. 2001). These taxonomic groups were originally described according to their coloration; drab females and bright blue males with dark vertical bars (Stauffer et al. 1997). As the coloration of these taxonomic groups are similar, they were classified as populations within the same species. However, various genomic studies have found a number of differences between such populations, including different segregating sex determiners (Ser et al.

2009; Streelman et al 2003; Roberts et al. 2009; Mims et al. 2010). Furthermore, these populations are not known to experience any gene flow. For these reasons, they will be referred to here as species. Together with the previously published identification of B chromosomes in *M. lombardoi*, B chromosomes now have been found in a total of 7 species of Lake Malawi cichlid. In all 7 of these species, B chomosomes have been found only in females.

Table 2.1: Individuals genotyped for B chromosome

Population	Females with B/ total females	Males with B/ total males
Metriaclima lombardoi	10/93	0/43
Metriaclima zebra 'Boadzulu'	21/49	0/30
Metriaclima greshakei	3/26	0/47
Metriaclima mbenji	1/27	0/33
Labeotropheus trewavasae	3/36	0/101
Melanochromis auratus	2/12	0/12
Metriaclima zebra 'Nkhata Bay'	3/80	0/51

The number of individuals, from each population, that were shown to carry a B chromosome via amplification with B-specific primers. Individuals were initially genotyped using primers for sequence 32 (Seq32), which produce the strongest amplification. Positive amplification was then confirmed by amplification for the other 4 B-specific sequences. Of the 43 B-carrying individuals genotyped, 1 individual (sample ID: 2005–0995) amplified with 4 of the 5 primer sets. The other 42 amplified with all 5 primer sets used.

Copy Number Variation of B chromosome Sequences

Now that the female-limited nature of the B had been examined further, the question of how many B chromosomes individuals of these species could carry was assessed. Did all individuals carry a single B, as observed karyotypically in *M. lombardoi*, or could there be variation as observed in species from Lake Victoria? Reason suggests that if individuals had 2 B chromosomes they should have twice as much of the B-specific sequences as individuals with a single B chromosome. Quantitative PCR was performed on the DNAs from 7 of the 8 karyotyped individuals shown to have the B, as well as B-carrying individuals identified by PCR from the population samples. This allowed for the quantification of the copy number of each of the 5 B block repeats studied above in each individual. If individuals possessed 2 or 3 B chromosomes, I expected that the copy number of each B block repeat would roughly double or triple, respectively, compared to individuals known to carry a single B. Individuals with the same number of Bs should consistently cluster in pairwise comparisons of copy number.

Figure 2.4 shows the variation in copy number from each B block amplified, organized by population and then left to right by increasing average copy number. Table 2.2 lists the mean copy number and standard deviation for each species. Each individual used in Figure 2.4 and Table 2.2 has a B chromosome as determined through PCR. Some B chromosome repeats (corresponding to the sequence from B blocks on scaffold_13, scaffold_58, and scaffold_14) tend to show little copy number variation (CNV), while other regions of the B chromosome (corresponding to the

sequence from B blocks on scaffold_23 and scaffold_32) show much higher CNV between individuals.

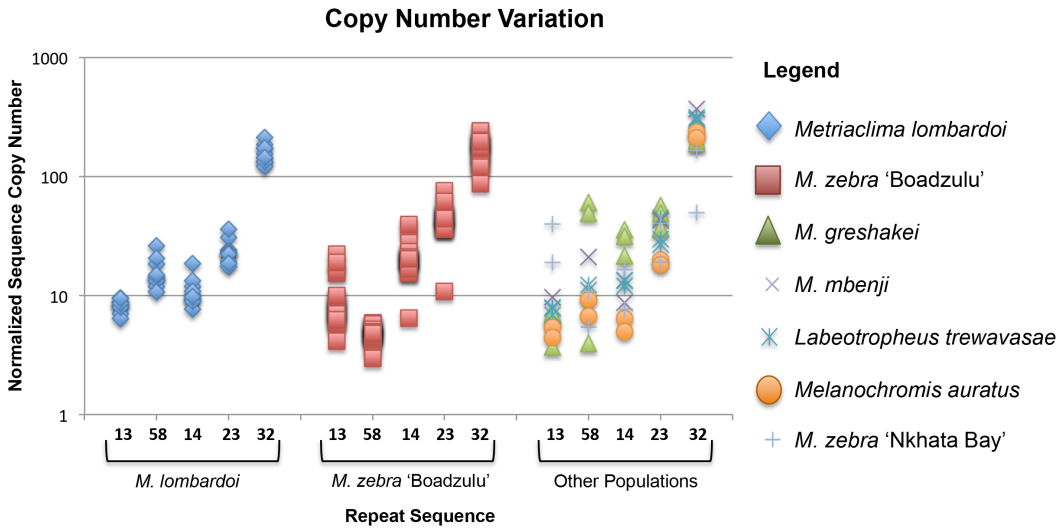

Figure 2.4: Copy number variation. Copy number of each B block repeat is shown in 3 groups; *Metriaclima lombardoi, Metriaclima zebra* "Boadzulu" and all other populations. The *y*-axis depicts copy number and the *x*-axis depicts the specific B block repeat analyzed. B block repeats are ordered left to right (for each of the 3 groups) by increasing average copy number. *Note:* the copy number of sequence 13 is not shown for 1 *M. zebra* "Boadzulu" individual (sample ID: 2005–0995), because the individual did not carry a copy of the sequence.

Table 2.2: Copy number for each B-specific sequence

Population	N	Seq13	Seq14	Seq23	Seq32	Seq58
Metriaclima lombardoi	10	9.21 (2.08, 22.6%)	10.73 (2.58, 24.0%)	23.3 (4.69, 20.1%)	155.98 (36.5, 23.4%)	15.22 (4.18, 27.5%)
Metriaclima zebra 'Boadzulu'	21	9.11 (5.13, 56.3%)	21.17 (6.88, 32.5%)	44.66 (12.21, 27.3%)	169.34 (40.74, 24.1%)	4.86 (0.73, 15.0%)
Metriaclima greshakei	3	5.85 (1.89, 32.3%)	29.42 (7.1, 24.1%)	47.49 (10.15, 21.4%)	214.82 (36.76, 17.1%)	37.73 (29.83, 79.1%)
Metriaclima mbenji	1	9.67 (-)	8.7 (-)	42.92 (-)	369.27 (-)	21.11 (-)
Labeotropheus trewavasae	3	7.53 (0.59, 7.8%)	12.9 (0.76, 5.9%)	33.67 (9.5, 28.2%)	303.18 (9.01, 3.0%)	10.46 (2.21, 21.1%)
Melanochromis auratus	2	4.9 (0.64, 13.1%)	5.65 (1.01, 17.9%)	18.98 (1.1, 5.8%)	222.55 (16.64, 7.5%)	7.93 (1.75, 22.1%)
M. zebra 'Nkhata Bay'	3	29.36 (14.61, 49.8%)	12.06 (6.18, 51.2%)	29.73 (14.91, 50.2%)	107.64 (82.11, 76.3%)	8.15 (3.85, 47.2%)

The mean copy number followed by the standard deviation and the relative standard deviation (in parentheses) for each population and each B-specific primer set. The value *N* represents the number of individuals. There was a significant difference in CNV between sequences, $F(4, 25) = 6.99$, $P \leq 0.001$. Standard deviation is not listed for *Metriaclima mbenji* as only a single individual from this population was found to have a B.

Sequences with a higher average copy number show higher absolute variation in copy number than sequences with lower average copy number. Within species, however, the range of copy number rarely exceeds 2-fold and appears to be a single cluster, consistent with the idea that all individuals carry a single B chromosome. A single individual (sample ID: 2005–0995) from the *M. zebra* "Boadzulu" population appears to be an outlier for each B block sequence (most easily observed in Figure 2.4 for repeat sequences 14 and 23). This individual does not possess a copy of sequence 13 (data point not shown in Figure 2.4). The copy number of each sequence in this individual is not only smaller than any other *M. zebra* "Boadzulu" individual, but it is smaller than all of the karyotyped *M. lombardoi* individuals, which are known to carry a single B chromosome. This suggests this individual (2005–0995) possesses only a fragment of the B chromosome.

To better detect individuals that have an increased copy number for multiple blocks, a pairwise comparison was performed for each pair of sequences. Figure 2.5 shows pairwise comparisons between each pair of B block repeats, organized in a half-matrix fashion. Figure 2.5 includes data from individuals genotyped as having the B chromosome through PCR as well as the 7 *M. lombardoi* individuals shown to have the B chromosome through cytogenetic analysis (as indicated in the legend). Several patterns are apparent. First, individuals of the same species tend to cluster. Second, there appear to be structural (copy number) differences among the B chromosomes of different species. Third, within species, there is no apparent correlation of the copy numbers for different sequence blocks. There are a few individuals that appear to be

outliers with respect to the cluster of individuals for that species. However, these individuals are not consistent outliers for each of the B block repeat classes. Thus, there is CNV of individual B block repeats among B chromosomes, but there is no evidence for correlation of CNV across loci, as would be expected if there was variation in the number of B chromosomes among individuals.

To further analyze repeat number among individuals, a hierarchical cluster analysis of the B block copy number was performed. Three groups emerged (Figure 2.6). One cluster contains all of the samples of *L. trewavasae* and *M. mbenji*. This cluster appears to reflect species-specific differences in the copy number of Seq32 (Figure 2.6). The 2 remaining groups each contain individuals that have been karyotyped and shown to have a single B. We conclude that all individuals in each group have a single B chromosome.

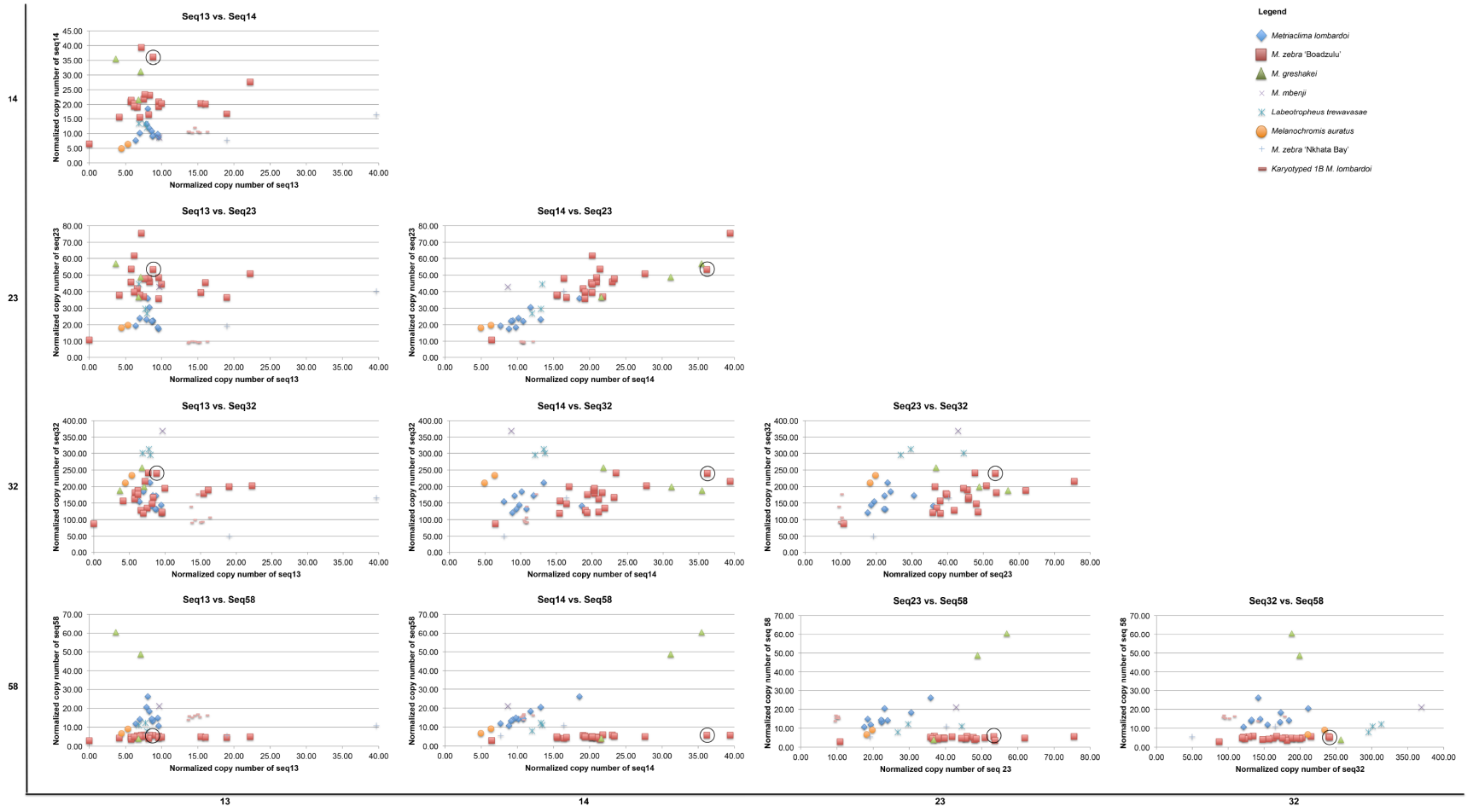

Figure 2.5: Two-way plots of copy number. Two-way plots arrayed in a half-matrix portray the copy number of a B block repeat on each axis. The same individual has been circled in each graph to demonstrate that while an individual may be an outlier for one B block repeat, it is not an outlier for other B block repeats.

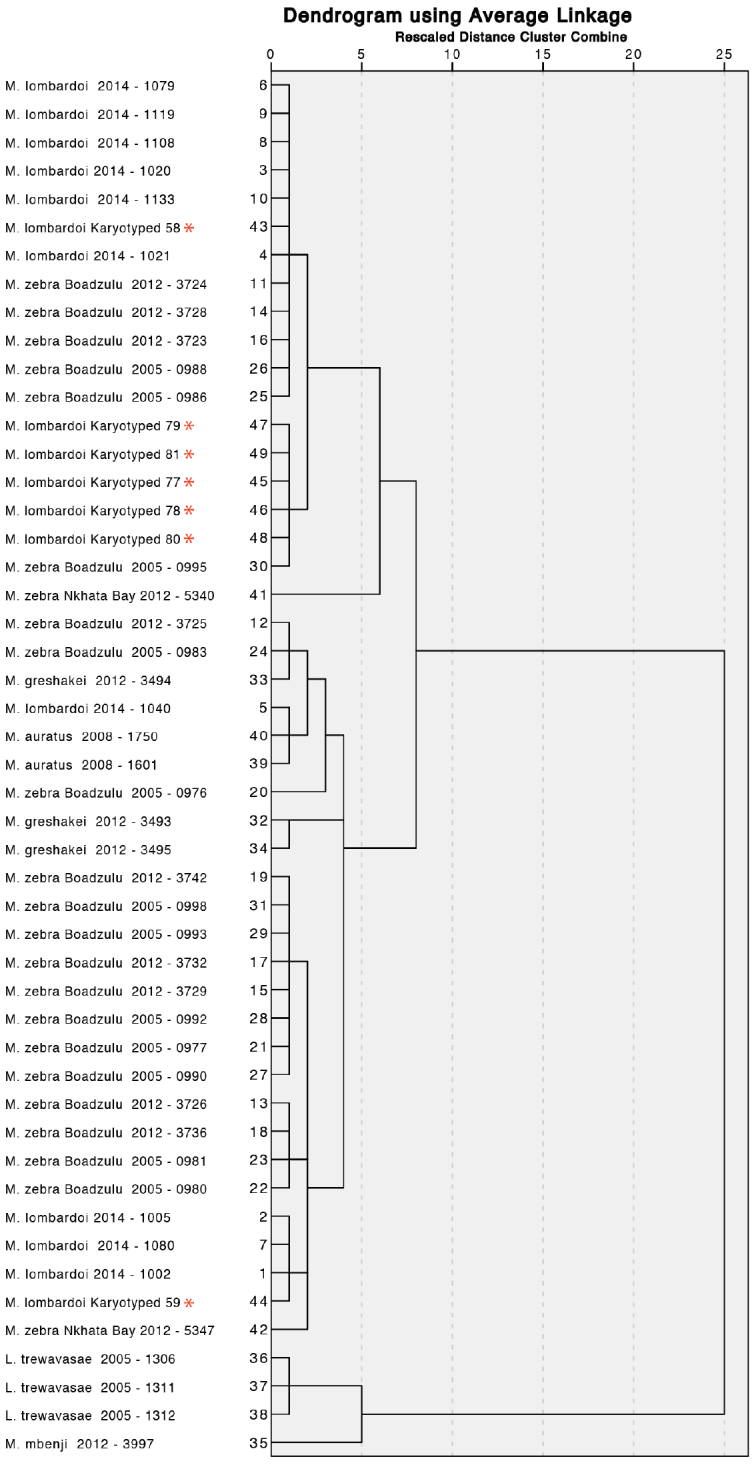

Figure 2.6: Cluster analysis dendrogram. Hierarchical clustering of samples by the CNVs of sequences Seq13, Seq14, Seq23, Seq32, and Seq58. Species and sample ID listed to the left. Karyotyped *Metriaclima lombardoi* individuals, known to have 1 B chromosome, are marked with a red asterisk (*).

Intergenerational Variation in Copy Number

Variation in copy number of B block sequences among siblings of karyotyped, lab-reared *M. lombardoi* was also examined. Since the offspring inherit the B exclusively from their mother, any sibling variation in sequence of the B must have arisen in a single generation. Quantitative PCR on the DNA from these individuals allowed for quantification of sequence CNV among siblings (Table 2.3). Individuals 58 and 59 are from family A002, and individuals 77–81 are from family A024. For most sequences, the copy number of each B block is consistent across individuals of a family, but there was considerable variation in copy number of Seq32. The variation at each locus is independent of the other loci. Since these individuals had been karyotyped and shown to possess only one B, these data reinforce the idea that minor variations in copy number reflect structural variation, not differences in the number of B chromosomes.

Table 2.3: Family member copy number

Sample ID	Family	Seq13	Seq14	Seq23	Seq32	Seq58
58	A002	13.5 (0.30)	10.6 (0.04)	9.1 (0.11)	139.3 (0.08)	13.9 (0.03)
59	A002	14.4 (0.53)	12.0 (0.02)	9.9 (0.03)	176.6 (0.04)	16.2 (0.07)
77	A024	15.1 (0.13)	10.4 (0.08)	9.3 (0.02)	93.8 (0.01)	15.4 (0.08)
78	A024	16.1 (0.20)	10.6 (0.09)	9.9 (0.12)	106.7 (0.03)	16.4 (0.01)
79	A024	14.8 (0.10)	10.7 (0.20)	9.3 (0.06)	92.2 (0.01)	16.8 (0.17)
80	A024	14.0 (0.06)	10.3 (0.08)	10.1 (0.06)	99.7 (0.02)	15.2 (0.03)
81	A024	13.6 (0.03)	10.6 (0.04)	9.8 (0.09)	91.1 (0.03)	15.8 (0.02)

This table lists the B block repeat copy number followed by the standard deviation of the three technical replicates (in parentheses) for each primer set for the individuals of 2 families. Columns 1 and 2 indicate Sample ID and family, respectively. All individuals included in this table have been karyotyped and possess a single B chromosome.

The variation identified by qPCR shows that the copy number of a sequence on the B can change quickly. Not only can sequence copy number vary among individuals in the same population, but it also varies among siblings. It is unclear whether this variation is produced during meiosis or mitosis or both. It is interesting to speculate what duplication mechanism could bring about the copy number changes in a single generation if there is only a single B chromosome present in the cell. Poletto et al. (2010) and Valente et al. (2014) suggest that the Lake Victoria B is an isochromosome (Poletto et al., 2010; Valente et al., 2014). While univalent, the 2

chromosome arms can associate with one another and potentially undergo recombination in meiosis. Unequal crossovers between the chromosome arms of sister chromatids may contribute to CNV among siblings. Various other duplication and deletion mechanisms (long-range slippage, break-induced replication, single strand annealing) may play a role in CNV, but it is difficult to distinguish these mechanisms without knowing the size of the duplications and their arrangement on the B chromosome. Alternatively, a drive mechanism that utilizes non-disjunction and preferential segregation during mitosis would result in 2B tissue prior to meiosis. During prophase I of meiosis, these 2 B chromosomes could then function as homologous chromosomes and undergo unequal crossover. Any of these mechanisms could produce the variation in copy number between single generations seen in the data presented here.

A Model of B Sequence Evolution

We propose the following model of B sequence evolution in Lake Malawi cichlids. Once established, the B chromosome experiences a continuous bombardment of sequences derived from the A genome, through translocation, transposition, and reverse transcription. Once on the B, these sequences are frequently duplicated, leading to a highly repetitive DNA sequence. It is unclear whether related B block repeats remain tandemly arrayed, or become dispersed throughout the B by structural rearrangements such as inversions. Localization of B block repeats via FISH or long read sequencing may resolve this. Because the B chromosome is inherited clonally,

mutations (single nucleotide polymorphisms, duplications, deletions, and structural rearrangements) are expected to accumulate rapidly.

If recombination is an important force of sequence evolution for the B, the location of the sequence along the chromosome arm may be important in determining the frequency and size of mutations. Sequences located in the middle of each arm may experience higher rates of recombination, and thus more rapid changes in copy number, than sequences near the centromere or telomeres. We have shown that the amount of variation in copy number can vary between different B block repeats. It would be interesting to look for correlations between the chromosomal location of a sequence and the rate of change in copy number. Alternatively, copy number may be a factor in copy number variation. B blocks present in higher copy number may be more likely to undergo unequal crossing over or slipped strand mispairing.

If this model of B sequence evolution is correct, then it may be difficult to determine whether a B chromosome arose from a particular A chromosome. The presence of a few homologous sequences found prominently on the B might be evidence of origin from a single or a few A chromosomes. Alternatively, this could indicate the success of those sequences in colonizing an already existing B, and not reflect the origin of the B itself. The abundance of a B block repeat also may not be related to the length of time it has been on the B. If there is rapid turnover of B sequences, it may prove impossible to determine from which A chromosome the B was originally derived.

Insights into the Mechanism of B Chromosome Drive

The data presented above demonstrates that 1) Malawi cichlid B chromosomes occur at appreciable frequency in populations of at least 6 Malawi cichlid species, 2) that all B-carriers are female, and 3) that carriers have only a single B in their somatic tissue. This is in contrast to the B chromosomes found in Lake Victoria cichlids, which are found in both males and females, and in up to 3 copies per individual (Poletto et al. 2010; Yoshida et al. 2011). It is possible that these differences in B chromosome distribution reflect differences in the underlying mechanisms of B chromosome drive.

A drive mechanism may act by increasing the number of B chromosomes that individuals carry, or by increasing the frequency of carriers, or both. The mechanism of drive will, in part, determine the frequency of Bs in tissues, individuals and populations. The efficiency of the mechanism will also contribute to the frequency of B chromosomes, but for simplicity only perfect mechanisms will be considered below. Only 3 combinations of nondisjunction and preferential segregation will produce a population with only 1 B among carriers (Table 2.4). Only 2 of these combinations actually produce drive. If the Lake Malawi B chromosome uses nondisjunction or preferential segregation to drive, it would either use a combination of both during a pivotal mitotic division, or preferential segregation during meiosis I to avoid the polar body. The latter indicates a drive mechanism specific to females, meaning the B chromosome has a higher fitness in females. This may explain its female-specific presence.

Table 2.4: Drive mechanisms utilizing nondisjunction and preferential segregation

Mitosis	Meiosis I	Meiosis II	Outcome
ND			0B, 1B
ND and PS			1B
	PS		1B
		ND	0B, 2B
		ND and PS	0B, 2B
ND	ND		0B, 2B
ND	ND and PS		0B, 2B
ND		ND	0B, 2B
ND		ND and PS	0B, 2B
	PS	ND	0B, 2B
	PS	ND and PS	2B

Possible combinations of nondisjunction (ND) and preferential segregation (PS), with the corresponding outcomes depicted. The possible drive mechanisms, combinations of ND and PS, are indicated in rows. Columns indicate the time when ND or PS occurs: during mitosis, meiosis I or meiosis II. The final column, "Outcome," indicates the expected type of offspring, that is, the number of B chromosomes they possess. Where 2 types of expected offspring are denoted, both types of offspring would be produced. Drive mechanisms considered here utilize combinations of nondisjunction and preferential segregation during at least one of these 3 cellular divisions. Each mechanism is assumed to be perfect (it produces the indicated result 100% of the time). All outcomes are based on individuals that have a single B chromosome and experience the type of drive listed.

While Bs have been shown to be mitotically unstable, or to employ nondisjunction and preferential segregation during mitotic divisions (Nur 1969; Kayano 1971), we are not aware of any such examples where mitotic nondisjunction is controlled in such a manner that it occurs precisely once during development. If the processes of nondisjunction and preferential segregation are recurring in mitosis, it would lead to higher variation of B chromosome number in the gonads. Furthermore, we can see no reason to expect that one mitotic division during development would be so distinct as to allow for this precise control. For this reason, we suggest that preferential segregation during meiosis I is the most likely scenario for the drive mechanism of the B chromosome in Lake Malawi cichlids.

Chapter 3: B Chromosomes and Sex

Context and Motivation

With a hypothesis of the possible mechanism the Lake Malawi cichlid B chromosome might use to achieve drive, it was important to confirm that drive is indeed occurring. While preliminary genetic-linkage data was available for some of the B-carrying species (Ser et al. 2009), no study of transmission of the B chromosome itself had been conducted. Furthermore, a complete association was found between B chromosome presence and the female sex in Lake Malawi cichlids, the mechanism for which remained to be determined. This female-limited distribution differs from the B chromosomes of most Lake Victoria cichlids, which are found in both males and females (Poletto et al. 2010). The notable exception is *Lithochromis rubripinnis* where B chromosomes are also limited to females (Yoshida et al. 2011). The female-specific presence of B chromosomes in Lake Malawi cichlids may be a result of the drive mechanism employed, or a secondary mechanism that benefits the B chromosome. If drive is accomplished through preferential segregation in meiosis I, it is only effective in females, which produce polar bodies. If the B chromosome acquired a method of transmission solely to females, it would increase the fitness of the B chromosome by increasing the opportunity for drive. As described in Chapter 1, some Orthoptera experience B drive in females as well as a secondary mechanism during male meiosis. The B chromosome takes advantage of the XY sex chromosomes. During male meiosis, the B pairs with the XY sex chromosomes and segregates with the X to the same daughter cell. The result is that males pass the B

chromosome more frequently to female offspring. This would not work in the Lake Malawi cichlids, as the B is never found in males. However, the same type of mechanism should work with a WZ sex determination system. Here, the B chromosome could hypothetically pair with the WZ sex chromosomes and segregate with the W to the same daughter cell. As a W chromosome is feminizing, this would ensure the B chromosome is transmitted from a female (WZ) to female offspring. As multiple XY and WZ sex determination systems have been found among African cichlids (Gammerdinger et al. 2018), it is possible these 7 species have both B chromosomes and a WZ sex determination system. Alternatively, with so many genic sex determiners arising in these cichlids, it is possible that the B chromosome itself acquired a feminizing sequence and is acting as a W chromosome in a WO sex determination system. While the two suggested mechanisms above ensure transmission to only females, conflict between the male genome and B chromosomes could also explain the lack of B-carrier males. Just as the jewel wasp, *Nasonia vitripennis*, experiences the elimination of paternal A chromosomes due to the B chromosome (Beukeboom and Werren 1993), the male genome in these cichlids could potentially cause the elimination of the B chromosome during an early mitotic division. And finally, perhaps the simplest explanation is male lethality. An incompatibility between B chromosome sequence and the expression of the male genome could lead to male B-carrier death early in development. An investigation of these hypotheses, summarized in a list below, is necessary to fully understand drive of this B chromosome and its corresponding evolutionary impacts.

1) The B segregates with a W chromosome
2) The B acts as a feminizing sex determiner
3) Elimination of the B from the male genome
4) Male lethality

Towards this end, genetic crosses were carried out and the F1 progeny were examined for B chromosome inheritance, sex ratio and sex chromosome linkage.

Methods

Among the 7 species of Lake Malawi cichlids with B chromosomes the species with the highest frequency of B-carriers, approximately 43% among females as determined in Chapter 2, was *M. zebra* 'Boadzulu'. Unfortunately, this population is located within a National Park, making it difficult to acquire the live individuals necessary for genetic crosses. Accounting for both B-carrier frequency and accessibility, *M. lombardoi*, the species where approximately 11% of females had a B, was selected for use in these genetic crosses. Live, wild individuals were imported from Lake Malawi, Africa in 2014, 2015 and 2016 to establish a laboratory line. This laboratory line is maintained at the Tropical Aquaculture Facility at the University of Maryland. Individuals were genotyped for B chromosome presence with PCR, as described below. Females with a B chromosome were crossed with males lacking a B. For comparison to the B chromosome cross, females without a B were crossed with males lacking a B. Lake Malawi cichlids are maternal mouth brooders, holding the offspring in their mouth for 2-3 weeks. Offspring were collected from the female at 1 week to

be raised together in a tank. A fin clip was taken from the brooding female from which the offspring were collected.

Each family was collected after reaching sexual maturity (approximately 9 months). Individuals were euthanized with tricaine methanesulfonate (MS-222) and inspected for testes or ovaries to confirm sex. *M. lombardoi* is known to have bright yellow males and bright blue females, but we found several yellow females and blue males as well as many individuals equally blue and yellow. This was true for both B and NoB individuals and there is no evidence B chromosome presence is responsible for differences in color. For this reason, we did not use color as an indicator of sex and highly discourage this method of sexing for *M. lombardoi*. Fin clips were collected from each individual.

DNA was extracted from fin tissue with standard phenol chloroform methods and phase-lock gel tubes (5prime, Gaithersburg, MD, USA). The B-specific primers designed and used in Chapter 2 were applied here to genotype for B presence/absence. All individuals were amplified with the B-specific primer set for Seq32 as well as the SWS1 (UV) opsin primer set to confirm that the quality of the DNA was sufficient for amplification. Any sample with poor amplification of the opsin primer set was not able to be genotyped for B presence/absence. Any sample with ambiguous amplification of the Seq32 primer set was then amplified with the remaining four B-specific primer sets.

Several sex-determining systems have been found among cichlids (Gammerdinger and Kocher 2018) and the XY system on LG7 and the WZ system on LG5 have been previously documented in some of the 7 B-carrier species from Lake Malawi as well as related species (Ser et al. 2010). To confirm which sex determination system(s) is acting in *M. lombardoi*, families were examined for linkage between phenotypic sex and microsatellite markers for these two known sex-determination systems. The primers used to amplify the markers on LG5 (UNH2139 and c-Ski) and the markers on LG7 (UNH2086 and UNH2031) are the same as those used in Ser et al. 2010. After PCR amplification of these microsatellite markers with fluorescently labeled primers, amplification product size was determined on an Applied Biosystems 377 DNA sequencer and analyzed with GeneScan v3.1.2.

Sex ratios, reported as male frequency, were analyzed for statistical significance with a binomial test. B transmission, reported as B-carrier frequency, was also analyzed with a binomial test. Finally, a Fisher's exact test was performed on each family assessing the relationship between B chromosome presence and sex.

B Transmission

If the Lake Malawi cichlid B chromosome does not drive, then Mendelian inheritance suggests that half the offspring of a B chromosome cross will inherit the B. If, however, the B chromosome does drive, as most B chromosomes do, then we expect to see more than half the offspring of a B chromosome cross inherit the B. Drive mechanisms do not need to achieve perfect efficiency (100% transmission) in order to

maintain B chromosomes or increase their frequency in a population. For this reason, any transmission rate above 50% is considered drive.

The six families resulting from a “B cross” (a B female crossed with a NoB male) ranged in B transmission from 50% to 95% (Table 3.1). The parents of each cross differ, however the B females of this laboratory line are all descending from just 3 B female founders. As a result, the B females in several of these crosses are related to one another (sisters, aunts/nieces, cousins). Only two of these families (A008 and A040) had a statistically significant deviation from Mendelian expectations. Mortality prior to tissue sample collection could explain this variation in transmission rate. The initial size of each B chromosome family is reported in Table 3.2 to provide a measure of mortality rates. However, much of this mortality is the result of various difficulties establishing a laboratory line (mainly bacterial infections) and we strongly encourage skepticism when evaluating these mortality rates with respect to B chromosome presence/absence. Alternatively, the observed variation in B transmission could reflect actual variation in drive success against different background A genomes. B chromosomes represent a source of genetic conflict. They use drive to selfishly increase their frequency, regardless of the fitness costs exerted on the host genome. For this reason, B and A chromosomes are expected to engage in an evolutionary arms race. As B chromosomes accumulate mechanisms to increase their transmission (drive), A chromosomes accumulate mechanisms to suppress this drive. The range of B transmission in the data presented here could be the result of such an evolutionary arms race. Ideally, crosses repeated with the same individuals,

and therefore the same B and background A chromosomes, could be performed to determine if they result in consistent transmission rates. The average transmission rate among these 6 B families is 67%. Another binomial test, assuming this averaged transmission rate, was performed on each family and was only significant for family A008 (p-value =0.00698). The remaining 5 families were statistically consistent with a transmission rate of 67%. Family A008 was only statistically consistent with transmission rates greater than 75%. Family A035 was only statistically consistent with transmission rates less than 73%.

Table 3.1: B chromosome transmission

B Family	# of B individuals	# of NoB individuals	B transmission rate	Binomial p-value
A008	19	1	0.95	0.00004**
A018	16	10	0.62	0.32690
A035	8	8	0.50	1.00000
A036	18	11	0.62	0.26490
A038	14	10	0.58	0.54130
A040	29	9	0.76	0.00166**

Transmission data is provided for 6 families segregating B chromosomes. For each family, the number of individuals, confirmed via PCR, to have or lack a B is listed in the 2nd and 3rd columns, respectively. B transmission rate is reported as proportion of B individuals. A binomial test, assuming a transmission rate of 0.5, was performed on each family. Families A008 and A040 significantly deviated from the expected 0.5 transmission rate.

Table 3.2: B family sample size

B Family	Initial # of progeny	# of progeny with recorded genotype	# of progeny with recorded sex
A008	32	20	20
A018	43	26	17
A035	17	16	13
A036	41	29	20
A038	35	24	23
A040	41	38	0

The number of individuals initially collected from the mother (dam) for each family is reported in the 2nd column. Individuals were euthanized after reaching sexual maturity (approximately 9 months) and inspected for testes or ovaries to confirm sex. Fin tissue was collected either after euthanasia or discovery of early mortality. Not all fin tissue collected after early mortality yielded viable DNA. Mortality prior to sexual maturation resulted in individuals for which sex could not be determined. The 3rd and 4th columns show the number of individuals of each family for which genotype and sex was determined, respectively.

The four hypotheses described above can first be distinguished by analyzing sex ratios among the different families. The hypothesis that the B chromosome acts as a feminizing sex determiner and the hypothesis of male lethality are expected to alter the sex ratio. The first would feminize males, leading to a lower male frequency. The second would result in more male than female mortalities, also lowering the male frequency. The remaining two hypotheses, segregation with a W chromosome and B elimination in males, are not expected to alter the sex ratio. Comparing the sex ratio of families resulting from a B cross to families resulting from a NoB cross would therefore help to eliminate two of the four hypotheses.

The sex ratio among 15 families resulting from NoB crosses averaged 0.48 (or 48% males) with a median of 0.45 (or 45% males) as shown in Table 3.3 and Figure 3.1. Only 5 of the 6 families resulting from a B cross had sufficient data on sex to analyze. The sex ratio for these families ranged from 0.05 to 0.31 (Table 3.4). Three of these were statistically significant according to a binomial test. The transmission of this B chromosome is skewing the sex ratio towards females. The average sex ratio among the B families was approximately 20% males. Another binomial test was performed on each family assuming a sex ratio of 0.2 and revealed no significant results; each B family was statistically consistent with a sex ratio of 0.2. In conclusion, the hypotheses of segregation with a W chromosome and B chromosome elimination in males are rejected. Additional analyses were conducted to further distinguish between the two remaining hypotheses, B sex determination and male lethality.

Table 3.3: NoB family sex ratio

NoB Family	Initial # of progeny	# of confirmed males	# of confirmed females	Sex Ratio
A003	35	3	4	42.9
A006	39	5	8	38.5
A009	36	5	6	45.5
A011	35	7	9	43.8
A013	48	4	2	66.7
A017	45	2	4	33.3
A022	64	5	5	50.0
A024	39	3	3	50.0
A026	31	5	4	55.6
A027	43	9	12	42.9
A028	40	10	11	47.6
A029	15	4	6	40.0
A032	42	5	6	45.5
A033	29	14	5	73.7
A034	53	9	9	50.0

Table 3.4: B family sex ratio

B Family	# of Males	# of Females	Sex Ratio	Binomial p-value
A008	1	19	0.05	0.00004**
A018	3	14	0.18	0.01273*
A035	4	9	0.31	0.26680
A036	6	14	0.30	0.11530
A038	5	18	0.22	0.01062*

Sex ratio data is provided for 5 families segregating B chromosomes. Sex ratio is reported as the proportion of males. A binomial test, assuming a sex ratio of 0.5, was performed on each family. Families A008, A018 and A038 significantly deviated from the expected 0.5 sex ratio with a bias towards females.

Table 3.5: Association between B chromosome and sex

B Family	# of NoB Males	# of NoB Females	# of B Males	#of B Females	Fishers exact p-value
A008	1	0	0	19	0.05000*
A018	3	1	0	13	0.00588**
A035	4	4	0	5	0.10490
A036	6	3	0	11	0.00217**
A038	5	5	0	13	0.00749**

The association between B chromosome presence and sex ratio is evaluated with a Fishers exact test for each family. Families A008, A018, A036 and A038 had a significant p-value, supporting the association between B chromosomes and the female sex.

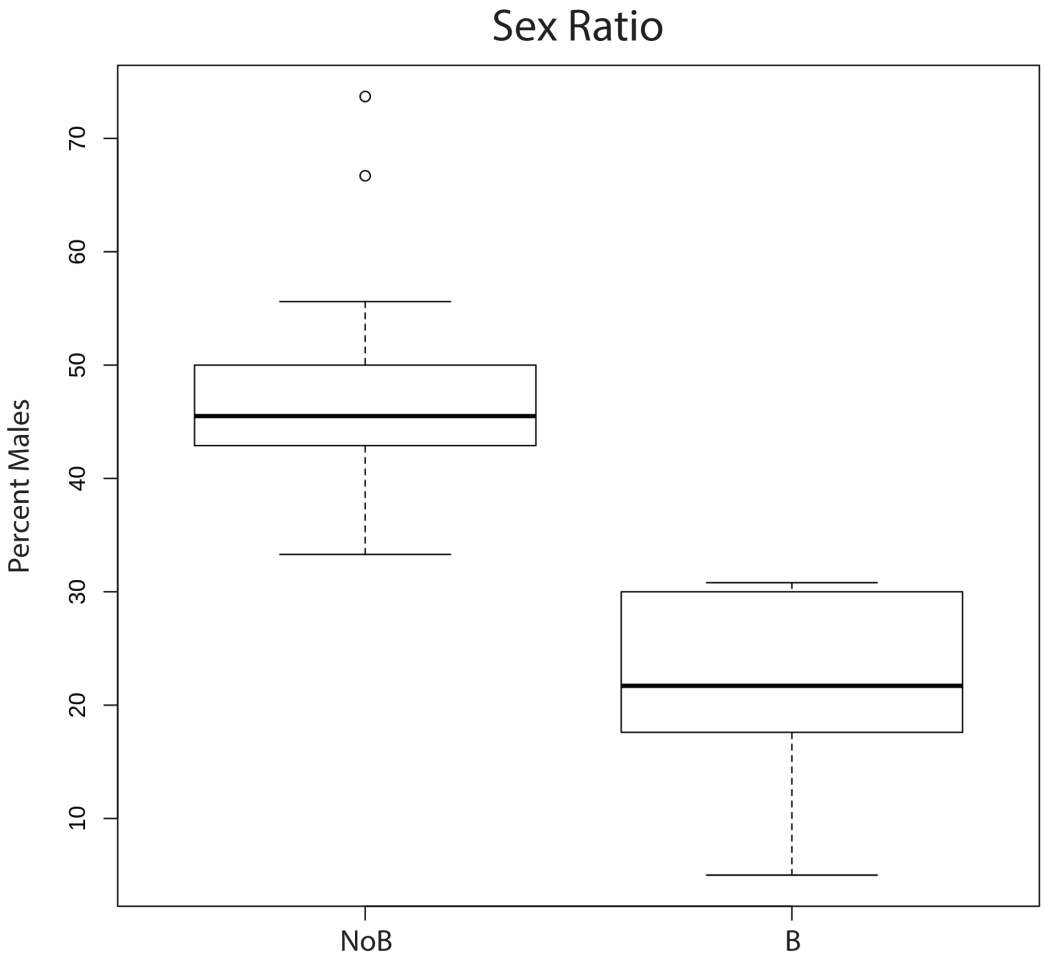

Figure 3.1: Sex Ratio. Box plots are shown for the sex ratio, reported as percent of males, for 15 NoB families and 5 B families.

While mortality rates theoretically could be used to distinguish the two remaining hypotheses, the process of establishing this laboratory line resulted in high mortality rates that varied dramatically. For this reason, our mortality data was not sufficiently controlled and, while reported in tables 3.2 and 3.3, will not be used. Instead, linkage with known sex determination systems was examined. If the B chromosome does not determine sex and Bs are limited to females due to male lethality, then sex-linkage should remain the same between families of B crosses and NoB crosses. Additionally,

a lack of individuals with a B and the male genotype (either XY or ZZ) is expected. However, if the B chromosome does determine sex, then sex linkage is expected to be the same between NoB individuals of a B family and NoB families, but differ between B individuals of a B family and NoB families.

Families resulting from a NoB cross revealed tight linkage with the LG7 XY system. An example of this is portrayed in Figure 3.2. Families resulting from a B cross revealed linkage with LG7 XY only among NoB individuals. The B individuals could possess either paternal haplotype in this region, presumably the X and Y (Figure 3.3).

	Markers			
	UNH2086	UNH2031	UNH2086	UNH2031
	160	165	151	148
	124	123	141	161
Sex				
F	124	123	141	161
F	160	165	141	161
F	160	165	141	161
F	124	123	141	161
F	160	165	141	161
F	124	123	141	161
M	124	123	151	148
M	160	165	151	148
M	124	123	151	148
M	160	165	151	148
M	160	165	151	148

Figure 3.2: Sex-linkage of a NoB family. Two markers (UNH2086 and UNH2031) known to be tightly linked with the LG7 XY sex determination system (Ser et al. 2009) are depicted for a family lacking B chromosomes. The haplotypes inherited from the mother (dam) are shown in light and dark pink while the haplotypes inherited from the father (sire) are shown in light and dark blue. The parental haplotypes are shown above those of the offspring. The offspring are arranged in rows with the following information: sex, maternal haplotype and paternal haplotype. All females inherited the same haplotype (light blue) from the father while all males inherited the other haplotype (dark blue). This is consistent with an XY sex determination system.

		Markers			
		2086	2031	2086	2031
		172	154	167	152
		184	156	190	132
B?	Sex				
B	F	172	154	190	132
B	F	184	156	190	132
B	F	ND	154	167	152
B	F	184	156	190	132
B	F	172	154	167	152
B	F	184	156	167	152
B	F	184	156	190	132
B	F	172	154	167	152
B	F	172	154	190	132
B	F	184	156	167	152
B	F	184	156	167	152
B	F	184	156	190	132
B	F	172	154	167	152
B	ND	184	156	167	152
NOB	F	172	154	167	152
NOB	F	172	154	167	152
NOB	F	172	154	167	152
NOB	F	184	156	167	152
NOB	F	184	156	167	152
NOB	M	184	156	190	132
NOB	M	172	154	190	132
NOB	M	172	154	190	132
NOB	M	172	154	190	132
NOB	M	184	156	190	132

Figure 3.3: Sex-linkage of a B family. Two markers (UNH2086 and UNH2031) known to be tightly linked with the LG7 XY sex determination system (Ser et al. 2009) are depicted for a family with B chromosomes. The haplotypes inherited from the mother (dam) are shown in light and dark pink while the haplotypes inherited from the father (sire) are shown in light and dark blue. The parental haplotypes are shown above those of the offspring. The offspring are arranged in rows with the following information: presence/absence of a B, sex, maternal haplotype and paternal haplotype.

These linkage data demonstrates that most families of *M. lombardoi* segregate an XY sex determination system on LG7. However, the sex of the B individuals is not determined by the LG7 locus alone. The B chromosome is epistatically dominant to the Y haplotype causing a female phenotype. In conclusion, the female-limited distribution of the Lake Malawi cichlid B is the result of an acquired feminizing sex determiner on the B chromosome.

A feminizing ability of the B chromosome might be expected given the hypothesis that drive is achieved through preferential segregation during meiosis I. Preferential segregation causes drive only in female meiosis where an asymmetrical division produces an egg-generative nucleus and a polar body. The egg-generative nucleus will continue through meiosis to produce a single oocyte or egg. The polar body will support this oocyte, but will eventually disintegrate and fail to contribute to the next generation. Male meiosis, by comparison, has symmetrical meiotic divisions that produce 4 sperms cells, all with equal opportunity to contribute to the next generation. If drive is achieved through preferential segregation in meiosis I, then the B would drive in females and not in males. Any mechanism that ensures the B is transmitted to females, such as a feminizing sex determiner or segregation with a W, would then result in an association of the B with the sex in which it drives, increasing the fitness of the B. Elimination of B chromosomes from a male genome or male lethality would also limit Bs to females, however these mechanisms would prevent even Mendelian transmission of Bs in males and would therefore decrease the fitness of the B. Therefore, drive via preferential segregation in female meiosis would be

expected to increase selection for any mechanism that ensures transmission of the B to females. I propose that the drive mechanism of this B evolved first. Subsequently, when a B acquired a feminizing sequence that B rose to higher frequency in the population. Comparative studies could test this hypothesis by identifying related cichlid species with B chromosomes that do not carry an epistatically dominant feminizing gene. The cichlids in lakes Malawi and Victoria are very closely related, diverging less than 8 MYA (Strumbauer et al. 2001). Species from each lake have B chromosomes, but their distribution and behavior differ. If the B chromosomes of the two lakes arose from a common ancestor, this would facilitate crucial studies of comparative evolution. Such studies could resolve not only the evolution of a sex determiner on the B chromosome, but also address questions about the evolution, maintenance, and turnover of drive mechanisms. Resolving the evolutionary relationship, if any, of the Lake Malawi and Lake Victoria cichlid B chromosomes should be a focus of future research.

An epistatically dominant sex determining B chromosome also highlights the potential genetic conflict brought about by B chromosomes. In this case, the B is altering the sex ratio of the population for selfish gain. In a taxonomic group already shown to have a high turnover rate for sex determination systems (Gammerdinger et al. 2018), this could increase selection for a masculinizing sex determiner in the A genome. Interestingly, this scenario also provides the opportunity to diminish genetic conflict. Sexually antagonistic selection often results in levels of gene expression that are a compromise for each sex. The B chromosome has accumulated sequences from

each of the A chromosomes and is limited to females. In the event that the B accumulates a gene under sexually antagonistic selection, the B-located copy of that gene is now free to evolve toward the female optimum.

An association between B chromosomes and sex is well established (see Camacho et al. 2011 or Burt and Trivers 2008 for a review). As discussed in Chapter 1, the frequency of B-carriers can differ between sexes, including the extreme of a sex-limited B, as seen in Lake Malawi cichlids (Beukeboom and Werren, 1993; Vicente et al. 1996; Neo et al. 2000; Camacho et al. 2011; Beladjal et al. 2002). Changes in the frequency of B chromosomes has been correlated with changes in sex ratio of various species, though in some cases further work is required to demonstrate that the B is causing the sex ratio bias (Camacho et al. 2011; Beladjal et al. 2002; Nur 1966; Yoshida et a. 2011; Zhou et al. 2012). B chromosomes have been shown to segregate with or away from the X chromosome of several insect species (Fontana and Vickery 1973; Jackson and Cheung 1967; Lopez-Leon et al. 1996; Nokkala et al. 2000). And finally, a shared evolutionary history between B chromosomes and sex chromosomes has been proposed in several species where either A) an autosome-sex chromosome fusion byproduct resulted in the origin of a B chromosome (Zhou et al. 2012), B) a B chromosome evolved from a sex chromosome (Lopez-Leon et al. 1994; Sharbel et al. 1998) or C) a sex chromosome evolved from a B chromosome (Hackstein et al. 1996). The data provided here demonstrates the B chromosome of Lake Malawi cichlids is a sex chromosome. The hypothesis of drive evolving first followed by the acquisition of feminizing sequence places this B chromosome in category C) a sex

chromosome that evolved from a B chromosome. Unlike the example of the Drosophila Y, the Lake Malawi cichlid B is functioning in the sex determination gene regulatory network, as opposed to pairing with an XX/XO system, making it more like the WO/OO system in the frog *Leiopelma hochstetteri* (Green 1988; Green et al. 1993; Sharbel et al. 1998). As sequencing technologies improve, it should be possible to assemble and analyze the repetitive sequence common to sex chromosomes and B chromosomes. Model systems like these cichlids can be used to elucidate the relationship of these chromosome types. Are sex chromosomes more likely to give rise to B chromosomes than autosomes? Do B chromosomes frequently transition into sex chromosomes? If so, do they more often carry sex determining genes or simply evolve to pair with sex determining chromosomes? Is this transition more likely to happen in taxa with stable heteromorphic sex chromosomes or rapid turnover of homomorphic sex chromosomes, or within a narrow timeframe after sex chromosome turnover? Are established B chromosomes more likely to acquire sequence from a degenerate Y or W chromosome than from less repetitive autosomes? What are the evolutionary consequences of a sex-determining selfish B chromosome? With the valuable genomic resources available in cichlid research and the identification of a feminizing selfish B chromosome in Lake Malawi cichlids, these questions can now be addressed in parallel with other model systems.

Chapter 4: The Identification of B Chromosome Sequence

This chapter is published as:

Clark, F., Conte, M. and Kocher, T. (2018). Genomic Characterization of a B Chromosome in Lake Malawi Cichlid Fishes. *Genes*, 9(12), p.610.

Context and Motivation

Knowing the Lake Malawi cichlid B chromosome drives and determines sex, the next logical question is, which sequences are responsible for these behaviors? Through what molecular mechanisms are these feats accomplished? While the mechanism of drive has been described for several species, studies revealing the causative sequence and molecular mechanism have been hindered by the fact that the sequence of B chromosomes is not well known. A notable exception is found in rye, *Secale cereale* (Banaei-Moghaddam et al. 2012). B chromosomes are highly repetitive, making them a challenge to assemble. There is not, to our knowledge, a single chromosome-scale assembly of a B chromosome for any species. To investigate the B chromosome for sequences underlying drive and sex determination, B sequence must first be identified.

Methods

Fin clips were collected from individuals of all 7 species directly from Lake Malawi in 2005, 2008, and 2012. Sex was determined via gonadal dissection. Standard phenol chloroform methods were used in conjunction with phase-lock gel tubes (5Prime, Gaithersburg, MD, USA) for DNA extraction from fin tissue. Genotyping for B presence/absence was performed with the B-specific primers designed in Chapter 2. Blood was collected from a *M. lombardoi* individual from our laboratory line in order to prepare the high molecular weight DNA necessary for Pacific Biosciences SMRT (PacBio, Menlo Park, CA, USA) sequencing.

Illumina sequencing was performed from the fin clips of 12 female individuals with a B chromosome. To provide a comparison from individuals lacking B chromosomes, the sequence from pooled male individuals, previously collected and sequenced with Illumina (San Diego, CA, USA), was used. Each pooled sample contained between 10 and 20 male individuals of that species. As there was no male *M. lombardoi* sequence data available, the B female *M. lombardoi* data was compared to a pool of *Metriaclima zebra* "Boadzulu" males for the scaled coverage analysis. Sequences from two samples of pooled female individuals lacking a B chromosome, previously collected and sequenced with Illumina, were used as controls. These two NoB samples represent two types of females from the *Labeotropheus trewavasae* "Maison Reef" population, XX females, and WZ females, all lacking a B chromosome. The samples used are summarized in Table 4.1.

The 12 B female individual samples, the 7 NoB male pooled samples, and the 2 NoB female pooled samples were prepared for Illumina sequencing with the TruSeq DNA sample preparation kit ver.2 rev.C (Illumina Inc.). Each DNA sample was sonically sheared and selected to produce libraries of 500 bp fragments. Paired-end reads of 100 bp were obtained using an Illumina HiSeq 1500.

Table 4.1: Sample information

Genus	Species	Locality	Sex	B?	Sample Type (#)	Sample ID	Sequencing Method	Mean Sequencing Depth
Labeotropheus	*trewavasae*	Thumbi	Female	B	Individual	2005-1306	Illumina	15.02
			Male	NoB	Pooled (10)	2005	Illumina	12.66
		Maison	Male	NoB	Pooled (20)	2012	Illumina	36.64
			XX Female	NoB	Pooled (20)	2012	Illumina	38.62
			WZ Female	NoB	Pooled (20)	2012	Illumina	36.63
Melanochromis	*auratus*		Female	B	Individual	2008-1601	Illumina	14.54
			Male	NoB	Pooled (10)	2005	Illumina	13.18
Metriaclima	*greshakei*		Female	B	Individual	2012-3493	Illumina	14.59
			Male	NoB	Pooled (20)	2012	Illumina	24.51
	lombardoi		Female	B	Individual	2014-1018	Illumina	16.21
			Female	B	Individual	2014-1021	Illumina	17.12
			Female	B	Individual	2014-1108	Illumina	11.75
			Female	B	Individual	2016-1012	PacBio	17.08
	mbenji		Female	B	Individual	2012-3997	Illumina	14.57
			Male	NoB	Pooled (20)	2012	Illumina	29.70
	zebra	Boadzulu	Female	B	Individual	2005-0976	Illumina	15.24
			Female	B	Individual	2005-0983	Illumina	14.78
			Female	B	Individual	2005-0986	Illumina	12.48
			Male	NoB	Pooled (20)	2012	Illumina	24.57
		Mazinzi	Male	NoB	Individual	SAMN03890374	PacBio	52.42
		Nkhata Bay	Female	B	Individual	2012-5340	Illumina	13.27
			Female	B	Individual	2012-5347	Illumina	16.20
			Male	NoB	Pooled (20)	2012	Illumina	34.39

Pacific Biosciences SMRT sequencing was performed on one *M. lombardoi* B female. DNA was extracted from nucleated blood cells using the MagAttract HMW DNA kit from Qiagen (Germantown, MD, USA). Pulse-field gel electrophoresis was performed with a Blue Pippin instrument by the University of Maryland Genomics Resource Center to select DNA fragments of the proper size. PacBio sequencing was carried out on the PacBio RS II platform with P6-C4 chemistry using nine SMRT cells and on the PacBio Sequel platform using nine additional SMRT cells.

Illumina and PacBio sequencing reads were aligned to the reference assembly of a *M. zebra* "Mazinzi Reef" NoB male individual sequenced with PacBio (Conte and Kocher, 2015), (publicly available on NCBI, Accession: GCA_000238955.4, Conte et al. 2017) with BWA (Li and Durbin, 2009) and NGM-LR (Sedlazeck et al. 2018), respectively. This reference assembly (M_zebra_UMD2) is an improvement of the assembly used in Chapter 2 (M_zebra_v0). BWA alignments were then run through Picard (v2.1.0) "MarkDuplicates" (http://broadinstitute.github.io/picard) to identify PCR duplicates.

After alignment to the reference genome, all genomic samples were analyzed with samtools (v0.1.18) mpileup (http://samtools.sourceforge.net) to calculate read coverage depth across the genome. The raw coverage depth was scaled by dividing the raw coverage at each position by the average genome-wide coverage depth of the sample. This scaled coverage value was then used to calculate the scaled coverage ratio (SCR) between the B chromosome female and the corresponding NoB pooled

male sample:

$$\text{SCR} = \frac{\text{scaled coverage of the B female}}{\text{scaled coverage of the NoB male pool}}$$

For each base in the genome, a binomial test was performed to check for a statistically significant difference in coverage between the B female dataset and the NoB pooled male dataset:

$$P(X) = \frac{n!}{(n-X)!\,X!} * (p)^X * (q)^{n-X}$$

In this binomial test, X represents the raw coverage depth in the B female sample and n is the sum of the raw coverage depth in the B female sample and the NoB pooled male sample. The expected frequency of B female reads, p, is calculated from the relative genome-wide sequence depth of the B female sample. The expected frequency of NoB pooled male reads, q, is calculated from the relative genome-wide sequence depth of the NoB pooled male sample. Any positions with a SCR ≥ 3 (corresponding to ≥4 B-located copies), a binomial test p-value ≤ 0.001, and within 300 bp of another such position were merged into a block feature with Bedtools (v2.26.0) merge function (Quinlan and Hall, 2010). Requiring a minimum SCR of three fails to detect any sequences with fewer than four copies on the B chromosome and avoids detection of simple A chromosome duplications, which would result in a SCR of 2. These block features were filtered to remove any block feature ≤ 500 bp in length and then any block feature with ≤10% of the positions spanned meeting the SCR ≥ 3 requirement and the p-value ≤ 0.001 requirement. The latter three parameters (merging distance of 300 bp, minimum block length of 500 bp, and

minimum percent positions of 10%) were chosen after manual inspection of several preliminarily identified regions. The remaining block features are referred to as "B blocks." The B blocks of all individuals were then processed with Bedtools (v2.26.0) intersect (Quinlan and Hall, 2010) to find B blocks common among at least 12 of the 13 B individuals (12 Illumina and 1 PacBio). These shared B blocks are referred to as the "core" blocks.

The sum of the lengths of all B blocks was calculated as an estimate of total B sequence length in the reference genome, further referred to as "A chromosome space." To account for copy number of these sequences on the B, the length of each block was multiplied by its estimated copy number, resulting in each block's contribution to the B. This was then summed to estimate the total B sequence length. Estimated copy number was calculated with one of two equations, depending on the average scaled coverage in the NoB male dataset. When the NoB male scaled coverage ≥1, we used the following equation:

$$(SCR * 2) - 2$$

In this equation, SCR was multiplied by 2 to compensate for the fact that we are comparing a haploid B genome to a diploid A genome. The A chromosome copy was then accounted for by subtracting 2. To avoid overestimating the B-located copy number when the NoB male scaled coverage was less than 1, we used the following equation:

$$Female\ Scaled\ Coverage * 2$$

Here, a NoB male scaled coverage of 1 was assumed (accounting for one copy of this sequence in the A genome of the reference), allowing us to use the scaled coverage of the B female to estimate copy number, without having to account for the A chromosome copy by subtracting 2.

Characterization of B Blocks

Due to the homology between A and B chromosome sequence, sequence reads derived from B chromosome regions with retained high sequence similarity will align to their A chromosome homologs present in the reference genome. As a result, alignments of reads from a genome with a B chromosome will have regions of increased coverage compared to an alignment from a genome lacking a B. An analysis of coverage ratios initially identified 0.34%–1.31% of the bases in the genome as having relatively higher coverage in the B female dataset (Table 4.2). In comparison, the same analysis in the controls identified 0.06% and 0.44% of bases in the WZ and XX NoB females, respectively. Further analysis combined these individual bases into features referred to as B blocks, defined as consecutive sequence with increased coverage in B chromosome samples. Thousands of B blocks were identified in each B female individual. B blocks ranged in length from 500 bp to 100 kb, although there were multiple regions in the genome with multiple B blocks in close proximity, suggesting that a larger region was transferred to the B chromosome as a whole (Figure 4.1). The largest such regions were located on LG4 (~120 kb), LG9 (~250 kb), LG17 (~260 kb), and LG23 (~420 kb).

Table 4.2: B block sizes

	% of A Genome Passing Both Thresholds	Number of Blocks	Mean Block Size (bp)	Standard Deviation of Block Size (bp)	Median Block Size (bp)	Maximum Block Size (bp)
L. trewavasae 2005-1306	0.59	3517	1554.8	2592.9	849	42941
M. auratus 2008-1601	0.34	2476	1415.7	1859.6	836	30172
M. greshakei 2012-3493	0.69	4392	1395.1	2618.8	805	52821
M. lombardoi 2014-1018	1.31	10918	1285.1	1845.8	824	63229
M. lombardoi 2014-1021	1.04	8251	1298.0	1954.9	822	63250
M. lombardoi 2014-1108	1.10	8684	1274.9	1902.3	809	63229
M. mbenji 2012-3997	0.68	4147	1344.7	2519.2	783	63230
M. zebra "Boadzulu" 2005-0976	0.85	5907	1264.9	2002.4	793	42941
M. zebra "Boadzulu" 2005-0983	0.79	5369	1238.5	2402.4	769	100567
M. zebra "Boadzulu" 2005-0986	0.84	5986	1228.8	2293.8	771	100079
M. zebra "Nkhata Bay" 2012-5340	0.64	4869	1419.0	2856.6	842	99928
M. zebra "Nkhata Bay" 2012-5347	0.89	7162	1420.9	2821.8	867	100026
M. lombardoi 2016-1012 (PacBio)	0.59	1904	2971.1	4569.8	1723	98620
L. trewavasae "Maison" XX females (control)	0.44	2125	714.0	243.4	642	3607
L. trewavasae "Maison" WZ females (control)	0.06	343	819.5	478.5	686	5198
Core blocks	N/A	622	2194.6	3582.8	937	32721

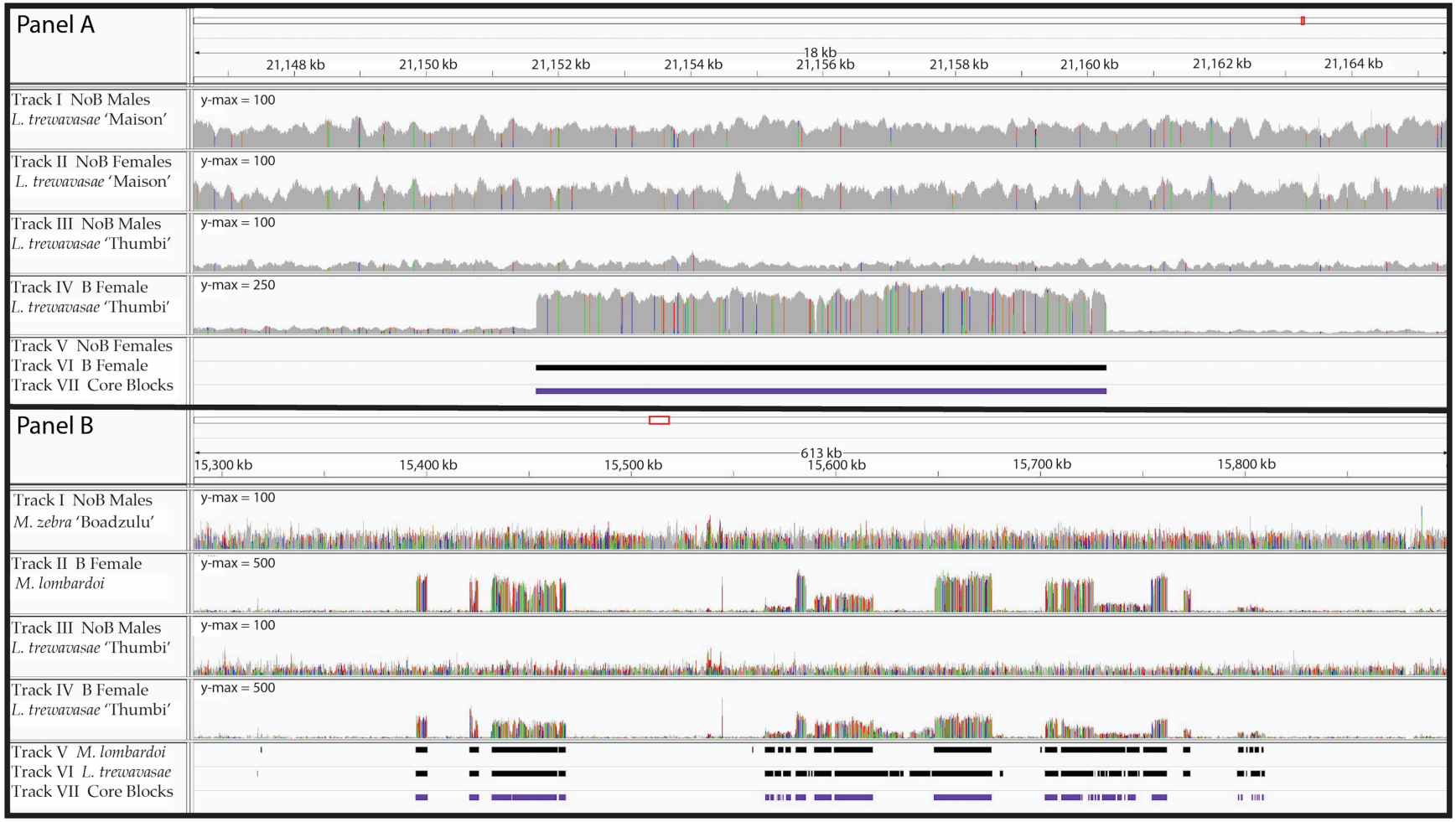
Panel A
18 kb
21,148 kb
21,150 kb
21,152 kb
21,154 kb
21,156 kb
21,158 kb
21,160 kb
21,162 kb
21,164 kb
Track I NoB Males
L. trewavasae 'Maison'
y-max = 100
Track II NoB Females
L. trewavasae 'Maison'
y-max = 100
Track III NoB Males
L. trewavasae 'Thumbi'
y-max = 100
Track IV B Female
L. trewavasae 'Thumbi'
y-max = 250
Track V NoB Females
Track VI B Female
Track VII Core Blocks
Panel B
613 kb
15,300 kb
15,400 kb
15,500 kb
15,600 kb
15,700 kb
15,800 kb
Track I NoB Males
M. zebra 'Boadzulu'
y-max = 100
Track II B Female
M. lombardoi
y-max = 500
Track III NoB Males
L. trewavasae 'Thumbi'
y-max = 100
Track IV B Female
L. trewavasae 'Thumbi'
y-max = 500
Track V M. lombardoi
Track VI L. trewavasae
Track VII Core Blocks

Figure 4.1: Read coverage and B blocks. B blocks from two genomic regions are shown with the corresponding read coverage. **Panel A** depicts an 18 kb region of LG8 with a typical B block. Tracks I and III are the male coverage (*Labeotropheus trewavasae* 'Maison' and *L. trewavasae* 'Thumbi', respectively), while tracks II and IV are the female coverage (NoB XX *L. trewavasae* 'Maison' control and B *L. trewavasae* 'Thumbi', respectively). Please note the *y*-axis maximum is 100 for tracks I, II, and III but 250 for track IV. Beneath the coverage plots are the blocks detected by our analysis; track V shows the NoB XX female blocks, track VI shows the B female *L. trewavasae* blocks, and track VII shows the core blocks. A ~8.5-kb B block can be observed by the increased coverage in the B female *L. trewavasae* (track IV), but no such increased coverage is observable in the other coverage plots. Our B block analysis pipeline identified the B female *L. trewavasae* block (track VI) but did not identify a block in the NoB XX female control data (track V). As this B block was similarly found in at least 12 of the 13 datasets, it is included in the core block set (track VII). **Panel B** depicts several B blocks in close proximity to one another across a 613-kb region of LG23. Tracks I and III are again male coverage but for *M. zebra* 'Boadzulu' and *L. trewavasae* 'Thumbi', respectively. Tracks II and IV both depict B female coverage (B *Metriaclima lombardoi* and B *L. trewavasae* 'Thumbi', respectively). Please note the *y*-axis maximum is 100 for tracks I and III but 500 for tracks II and IV. The block sets detected in the B female *M. lombardoi* (track V), B female *L. trewavasae* 'Maison' (track VI), and the core blocks (track VII) are shown below. B blocks can be observed in the coverage of both B females (track II and IV) and correspond well with the blocks identified through our B block identification analysis (tracks V, VI, and VII). The B blocks span ~420 kb and appear to have migrated to the B as a single unit in the ancestor of *M. lombardoi* and *L. trewavasae*.

In the WZ and XX NoB females controls, we identified 343 and 2125 putative B blocks, respectively, and the longest blocks were only 3.6–5.2 kb (Table 4.2). As neither of these individuals carried a B chromosome, these putative B blocks represent false positives. While actual variation in A genome copy number may explain some of this error, stochastic variation in the coverage depth of Illumina data and regions of poor alignment likely also contribute to these false B block calls. Figure 4.2 provides representative histograms of block length, showing data for a B chromosome female (*L. trewavasae* 2005-1306), the blocks included in the core set, and the XX and WZ NoB females. Both the B female and the core set show enrichment for blocks of longer lengths when compared to the controls. The core set shows a depletion of shorter blocks. An interpretation of this is that false positive B block calls are more likely to be short in length and that a sizable portion of the shorter B blocks may be false positives (type 1 error) and do not represent actual B sequence. However, since large regions, as seen in Figure 4.1, are often fragmented into smaller block calls, we opted not to remove the shorter block calls at this stage of the analysis.

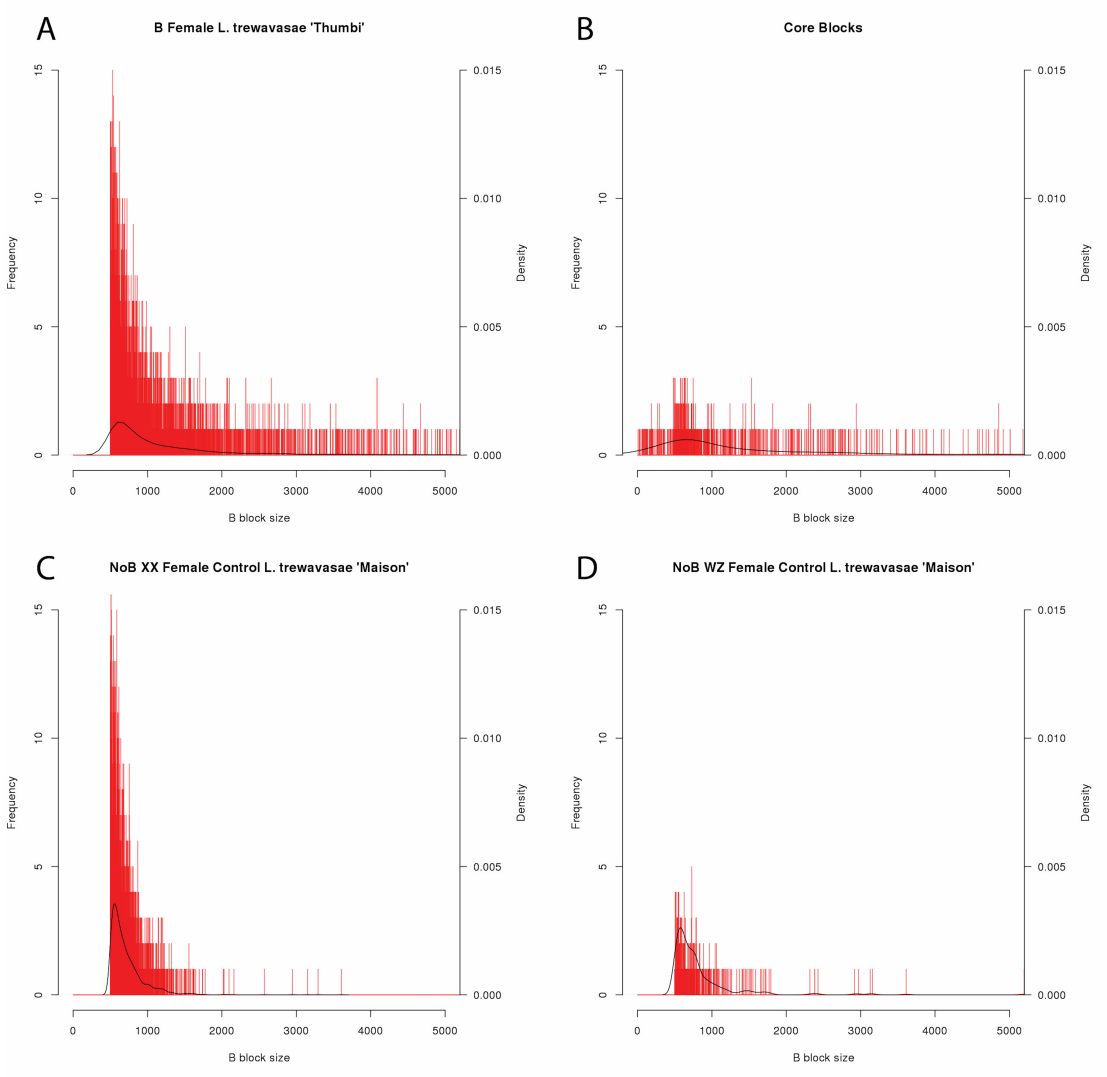

Figure 4.2: Block length histograms. Histograms of B block length for four datasets. B block size along the *x*-axis is reported in bp (bin size = 1 bp), and the *x*-axis maximum is 5000 bp to more easily view the majority of the data. The blocks not visible in this graph, larger than 5000 bp, represent only 10.3% of the core blocks and <5% of the blocks in the B female and two NoB controls. The number of blocks of each length is shown with red bars with the *y*-axis scale on the left and the density is depicted with a black line with the corresponding *y*-axis scale on the right. Because blocks shorter than 500 bp were removed during analysis, the B female (**A**) and the two NoB controls (**C**,**D**) show a lack of these smaller blocks. However, during the identification of core blocks, some larger B blocks were fragmented further, resulting in the smaller B blocks shown in the core block set histogram (**B**).

The lengths of all B blocks were then summed for each sample, as well as for the set of core blocks, producing the total length of B sequence in A chromosome space (Table 4.3). However, since there are multiple copies of these sequences on the B, we multiplied the length of each block by the copy number of that sequence, as estimated by the difference in coverage between the B female dataset and the male dataset. These values were then summed across all blocks to produce the total estimated length of B chromosome sequence (i.e., in B chromosome space). The total length of B sequence from the core block set (not including variable blocks specific to some individuals or species) in B chromosome space was also calculated for each sample.

Table 4.3: Total estimated length of B sequence

	In A Space (Mb)	In B Space (Mb)	Core Blocks in B Space (Mb)	Core Blocks % of Total B in B Space
L. trewavasae 2005-1306	5.48	49.44	25.13	50.84
M. auratus 2008-1601	3.51	23.19	12.31	53.11
M. greshakei 2012-3493	6.14	58.20	30.84	52.98
M. lombardoi 2014-1018	14.06	74.53	24.09	32.31
M. lombardoi 2014-1021	10.73	62.58	23.00	36.76
M. lombardoi 2014-1108	11.09	58.48	17.67	30.22
M. mbenji 2012-3997	5.59	41.30	21.62	52.34
M. zebra "Boadzulu" 2005-0976	7.49	59.48	30.86	51.88
M. zebra "Boadzulu" 2005-0983	6.66	51.06	27.38	53.63
M. zebra "Boadzulu" 2005-0986	7.37	49.55	23.44	47.31
M. zebra "Nkhata Bay" 2012-5340	6.92	48.61	21.58	44.40
M. zebra "Nkhata Bay" 2012-5347	10.19	99.69	44.07	44.21
M. lombardoi 2016-1012 (PacBio)	5.66	35.40	15.02	42.44
L. trewavasae "Maison" XX females (control)	1.52	2.15	0.63	-
L. trewavasae "Maison" WZ females (control)	0.28	0.39	0.80	-
Core blocks	1.37	-	-	-

Total estimated length of B sequence identified in the coverage ratio analysis. The total length of sequence identified in the reference genome, or "in A space", is provided for all B individuals, the two NoB controls, and the core block set (the blocks shared by at least 12 of the 13 B individuals). The total length of these blocks, after accounting for their copy number on the B chromosome (i.e., "in B space") is provided for the 13 B individuals and 2 NoB controls. Similarly, the total length of just the core blocks, accounting for copy number, is provided for the 13 B individuals and 2 NoB controls. Finally, the portion of the total estimated B chromosome length that is represented by the core blocks is provided for the 13 B individuals.

The total length in A space ranges from 3.51 to 14.06 Mb among the B females and only 0.28 to 1.52 Mb in the controls. Only 1.37 Mb (in A space) is shared among at least 12 of the 13 B females. After taking copy number of these sequences into account, the total length in B space ranges from 23.19 to 99.69 Mb among B females and only 0.39 to 2.15 Mb in the controls. The 1.37 Mb of core blocks in A space translates to 12.31–44.07 Mb among B females and as little as 0.63–0.80 Mb in the controls.

The consensus, or core, block set with blocks common to at least 12 of the 13 individuals successfully removed the greatest proportion of false positives (type 1 error). However, the core block set lacks any B chromosome sequence that is specific to only a few individuals or species. The B chromosome of the *M. lombardoi* individuals, sequenced with Illumina, is estimated to be 58.48–74.53 Mb in length. Considering just the most conservative B blocks (the core blocks), the estimated length is 17.67–24.09 Mb in these individuals. Karyotype data, available only for *M. lombardoi*, shows that the B chromosome is one of the largest chromosomes. A tentative estimate of chromosome size from karyotype data suggests a B chromosome of roughly 50 Mb. The total length of B sequence in B space in these three individuals may be inflated by false positive blocks, while the total length of core sequence is B space is smaller than the length estimated from the karyotypes. The variation in estimated B chromosome length across individuals could indicate that B chromosomes vary in size among these species. This is consistent with the finding that B chromosomes vary in length within and among species of Lake Victoria

cichlid. (Yoshida et al., 2011). Notably, *Melanochromis auratus* consistently has the least amount of sequence detected by this analysis. The 12.31 Mb, in B space, found in *M. auratus*, compared to the 30.84 Mb found in *Metriaclima greshakei*, suggests that the B chromosome of *M. greshakei* may be twice as large as the B chromosome of *M. auratus*.

Using an analysis of sequence coverage, we identified 1.37 Mb of the A genome that has been copied to the B chromosome and which is now shared among several Lake Malawi cichlid species as core B chromosome sequence. In addition to this core sequence, there were many additional megabases of B chromosome sequence that were found among various subsets of individuals/species. Because the core B chromosome sequences are found in multiple copies, the total length of B-specific sequence in the three *M. lombardoi* individuals totaled 17.67–24.09 Mb. This is consistent with the size of the *M. lombardoi* B chromosome observed in karyotype data. This suggests that the coverage ratio analysis was successful in identifying an appreciable amount of sequences on the B chromosome. Using all the B blocks identified with each individual dataset (including both variable and core blocks) resulted in a size estimate of 58.48–74.53 Mb (in B space), which is slightly larger than expected from karyotype data. This suggests that some portion of the identified B blocks represent false positives, or type 1 error. Another approach to understanding the amount of type 1 error in this analysis is through the two control datasets. The percentages of individual bases in the genome passing the SCR and binomial thresholds were not markedly different for the XX NoB female control (0.44%) and

the B females (0.34%–1.31%). Our downstream filtering to produce block features helped to further reduce the type 1 error, resulting in an order of magnitude fewer blocks identified in the two controls compared to the B female datasets. Further filtering of short blocks would likely continue to reduce the type 1 error but would simultaneously increase type 2 error. The length of identified B sequence, in B space, of the two controls was 0.39–2.15 Mb. Arguably, we can extrapolate from this to predict that any individual could have at least 1–5 Mb of falsely identified B sequence. Yet, the total amount of variable B sequence, in B space, ranged from 39.58–50.45 Mb for the B female datasets. From this, we conclude that B blocks identified using sequence data from a single individual likely contain some type 1 error but also correctly represent a large number of unique B blocks that are not shared among individuals or species.

In the estimation of total length in B space, proper estimation of B-located copy number is clearly crucial. For most regions, SCR can be used to estimate copy number. However, in regions of poor alignment, scaled coverage can be <1, leading to an overestimate of copy number and therefore an inflated estimate of total length in B space. To avoid this issue, the B-located copy number of any region with a scaled coverage value <1 in the NoB male dataset was instead calculated with the scaled coverage in the female B dataset. The use of multiple individuals and the identification of core sequence greatly reduces the type 1 error and we suggest that multiple individuals, if not species, be used to produce the most conservative identification of B sequence when using a coverage ratio analysis. Notably, our

coverage ratio analysis ignores any sequence entirely unique to the B chromosome (not aligned to homologous A sequence) or any sequence with fewer than four B-located copies (an SCR of 3). While unique sequences are undetectable with a coverage ratio analysis, the detection of less abundant sequences on B chromosomes presents a trade off with type 1 error rates. This can be circumvented by sequencing individuals with a high number of B chromosomes per cell, when possible.

Comparison of Illumina and PacBio Sequence Data

To better understand the differences in B blocks called from Illumina and PacBio datasets, we compared an *M. lombardoi* B female sequenced with PacBio to the three *M. lombardoi* B females sequenced with Illumina. The Illumina reads are 100 bp in length and the PacBio reads averaged 8295 bp. The blocks identified in the individuals sequenced with Illumina ranged in total length in A space from 10.73 to 14.06 Mb, whereas the total length of blocks identified in the individual sequenced with PacBio was only 5.66 Mb in A space. As demonstrated with the block size histograms (Figure 4.2), we believe most falsely identified blocks are short in length. Indeed, the mean length of B blocks identified using the PacBio data was much longer than with the Illumina data (Table 4.2) and a depletion of shorter blocks can also be seen in the block size histogram of the PacBio data (Figure 4.2). This discrepancy in length in A space could be a byproduct of the longer PacBio reads resulting in more consistent coverage and preventing the erroneous identification of shorter blocks. Additionally, longer PacBio reads will have more accurate mapping in repetitive regions than the shorter Illumina reads. These factors suggest that PacBio

data would result in fewer false positives or type 1 errors. However, even when using the conservative core block set, the PacBio data identified only 15.02 Mb of core sequence in B space compared to the 17.67–24.09 Mb identified in the three Illumina datasets, suggesting the Illumina data is able to detect sequences the PacBio data does not.

While inspecting the read alignments and coverage data in detail, a few key patterns emerged. First, there were several regions of high coverage in the Illumina data, which had low coverage in the PacBio data (Figure 4.3, panel A). The Illumina reads in these short regions all aligned to several other locations (as indicated with white reads in Figure 4.3, panel A, tracks I and II) and these regions were annotated as various repeats. Our interpretation is that these regions represent a shorter, highly repetitive sequence, with many copies found on the B chromosome. We hypothesize that the A chromosome in the *M. zebra* reference assembly experienced a recent insertion of this repeat, resulting in a lack of coverage by the *M. lombardoi* PacBio data because it does not have this insertion. Because the Illumina reads are too short to span the length of the repeat, they aligned to this insertion in the reference. This means that the same analysis with Illumina data was able to detect these B-specific sequences while analysis with the PacBio data was not. However, the Illumina data wrongly places the A chromosome origin of these B sequences at the new insertion site when their existence on the B appears to predate this insertion.

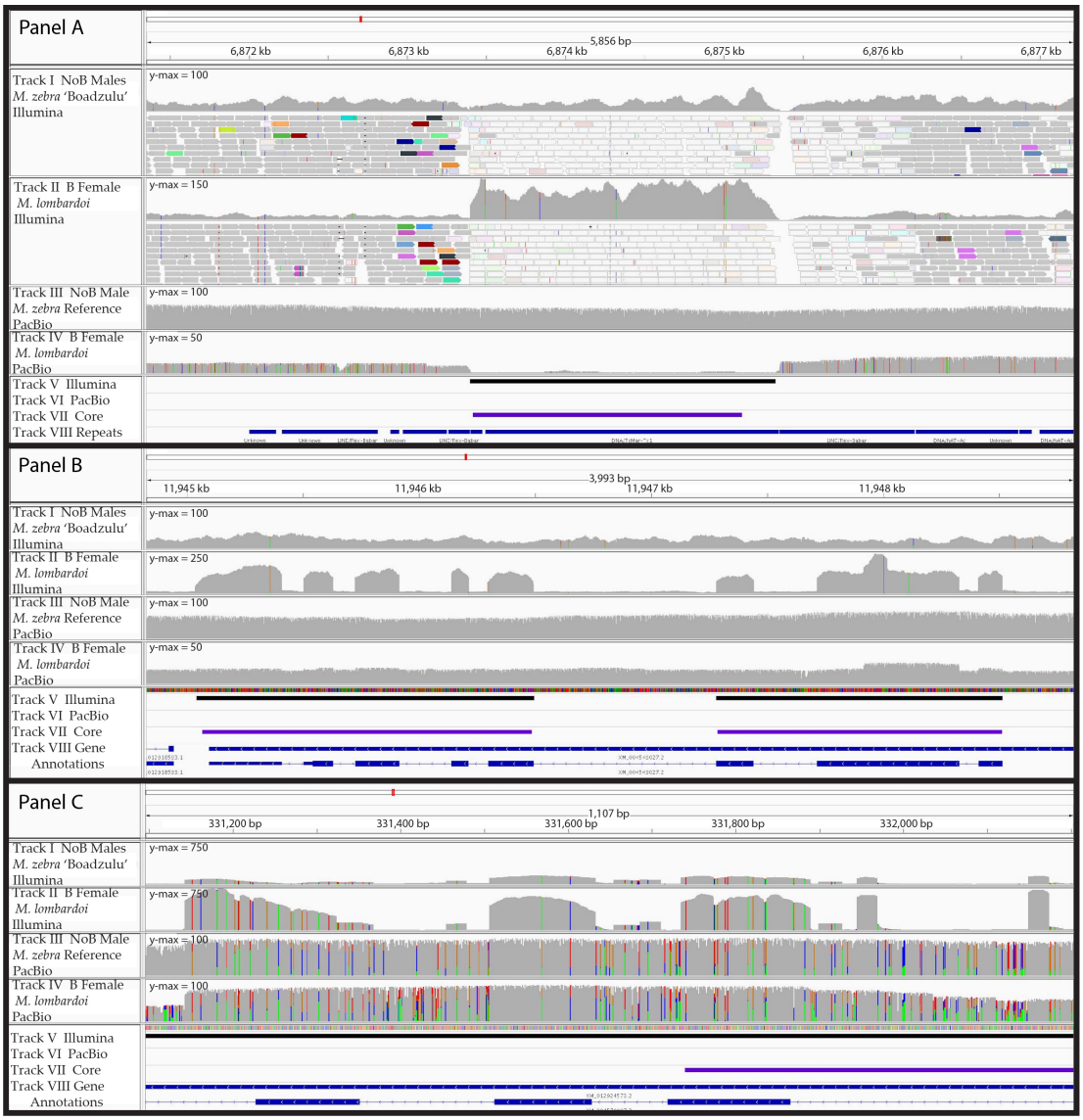

Figure 4.3: Comparisons of Illumina and PacBio read alignment in B blocks. Differences in the alignment of Illumina and PacBio reads affected the B block identification analysis. **Panels A–C** represent regions of LG20, LG22, and the unanchored scaffold 000256F_pilon_quiver, respectively. **Panel A** demonstrates a failure to identify a B block with PacBio data. Additionally, the localization of that block with Illumina data to a recent insertion inaccurately suggests LG20 as the A chromosome origin of this B-located sequence. **Panel B** demonstrates the failure of PacBio data to detect a retrogene. **Panel C** demonstrates a case where Illumina data suggests a retrogene, which the PacBio data reveals to be a complete gene (possessing both exons and introns). This specific example also shows increased coverage in the NoB data, suggesting the region has also experienced a duplication event within the A genome in addition to the copies present on the B chromosome. In **Panels A–C**, tracks I and II represent the coverage of the NoB male and B female sequenced with Illumina, respectively, while tracks III and IV depict the coverage of the NoB male

and B female sequenced with PacBio, respectively. In **Panels A–C**, the B female *M. lombardoi* block sets for Illumina and PacBio are shown in tracks V and VI, respectively, while the core block set is shown in track VII. Tracks I and II in **Panel A** also show a portion of the reads aligning to that region. Reads shown in white have a map quality of 0 indicating multiple mapping to several regions. In **Panel A**, track VIII displays the annotated repeat content of this region. In **Panels B,C,** track VIII displays the gene annotations. Please note the y-axis maximum of the coverage plots varies to best view the variable coverage data of each plot.

A second difference between the sequence data types was in the detection of retrogene insertions (Figure 4.3, panel B). Again, since the PacBio reads are much longer than the retroinserted exons, they do not align well to the A reference using typical PacBio alignment software such as NGM-LR and BLASR with standard alignment parameters. In contrast, Illumina reads are usually shorter than the length of these retroinserted exons and therefore do align well to the reference. This means that standard alignment software and parameters will detect retroinserted sequences on the B chromosome with short read data but not with long read data. Proper alignment of retroinserted genes using PacBio reads requires the use of alignment tools that are splice-site aware, such as GMAP. We were able to recover this particular retroinsertion with the PacBio data by aligning with GMAP, but the majority of A genome reads did not map. Alignment software that accounts for both types of reads is needed, but to our knowledge, such tools do not yet exist.

The third difference between the two sequence data types was in the false detection (type 1 error) of retroinserted genes (Figure 4.3 panel C). The Illumina data showed increased coverage in the exons but not the introns of some genes, suggesting it was another retroinserted gene on the B. However, the PacBio data revealed consistently

high coverage across both introns and exons, with much higher sequence polymorphism in the introns. The higher sequence polymorphism in the introns compared to the exons suggests that the B-located copy of this gene is relatively old and still experiencing purifying selection for the encoded protein. The short reads of the Illumina data failed to align to the divergent introns but did align in the less divergent exons, resulting in what appeared to be a retroinserted gene. We were only able to distinguish between 'true' and 'false' retroinserted genes on the B chromosome by comparing the Illumina data with PacBio data.

B Block Turnover

B chromosomes are thought to have a high rate of sequence turnover because they experience little purifying selection (Houben et al., 2014; Klemme et al., 2013). Because the Lake Malawi cichlid species studied here diverged less than 1 million years (MY) ago (Kocher 2004), we have an opportunity to study the rates and patterns of sequence turnover on the B chromosome. To gauge the amount of sequence turnover that has occurred between these species, we compared the core block set to all B blocks (core and variable) identified in each individual. The core blocks accounted for 30.22%–53.63% of total B sequence (in B space), leaving 46.37%–69.78% B sequence (in B space) variable among individuals. While some of the variable B blocks represent false positives (type 1 error), many represent sequences that are unique to a particular individual or species. These variable B blocks likely represent both sequences that were lost from a common ancestor and new sequences acquired during the evolution of particular lineages. If an appreciable

amount of the variable blocks detected among individuals is actually type 1 error, then more than 30%–54% of the B is shared among these species. Even though the individuals in this study span three genera, they are less than 1 MY diverged from one another. This suggests that the Lake Malawi B chromosome has experienced turnover of roughly half of its sequences in 1 MY. Whether the rate of sequence turnover is constant or varies over time is not yet known.

B Block Origin

A comparison across these 13 B individuals has allowed us to identify sequence (the core blocks) present on the B chromosome of the most recent common ancestor to these seven cichlid species. Figure 4.4 depicts the position of core B blocks on the chromosome-scale assembly of the *Metriaclima zebra* A genome. Notably, each linkage group (LG), and therefore each chromosome, has at least one core B block and most have several, distantly spaced core B blocks. This is consistent with the idea that cichlid B chromosomes continue to collect A chromosome sequences over time (Jones and Houben, 2003; Valente et al., 2014; Martis et al., 2012). No trend was observed between B block position and centromere position. There is no readily visible pattern that would suggest certain regions are more likely than others to be the source of B chromosome sequence. The longest stretch of B chromosome (along A chromosome space) corresponds to a ~420-kb region comprising several neighboring B blocks on LG23 (also shown in Figure 4.1). The SCR of the core blocks varies among individuals. The largest difference in SCR between these two individuals is shown on LG8.

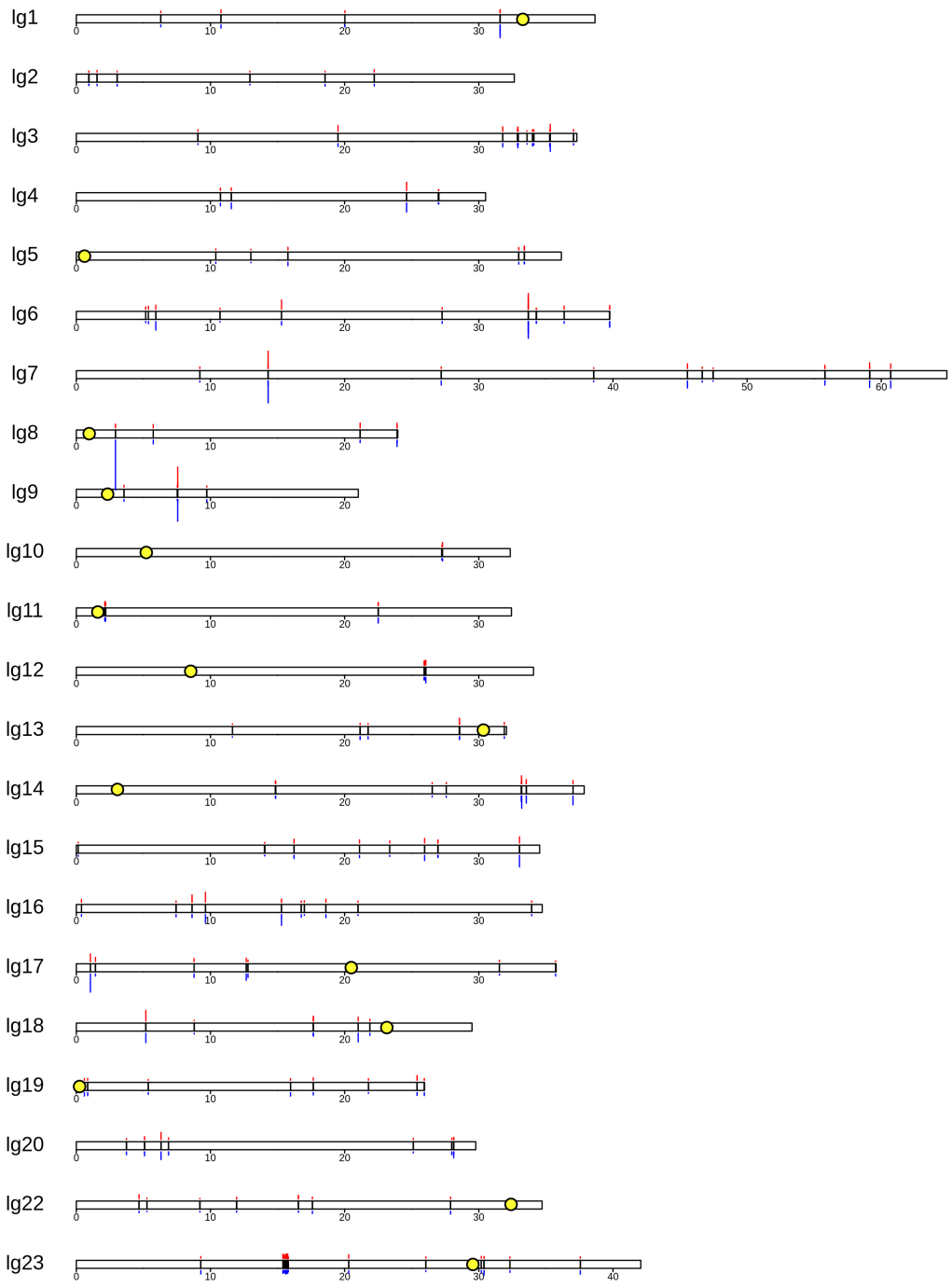

Figure 2: Karyoplot showing the A genome origins of the B chromosome. The position of B blocks (black bars) is superimposed on a karyoplot of the *Metriaclima zebra* "Mazinzi" reference genome. The A genome consists of 22 chromosomes. For simplicity, unanchored scaffolds of the genome assembly were not included. Physical distances are noted beneath each LG in Mb and the locations of centromeres (available for only some LGs) are indicated with yellow circles. Above and below

each LG is a bar graph representing the scaled coverage ratio (SCR) of each core block. Above, in red, is the SCR of *M. lombardoi* 2014-1021. Below, in blue, is the SCR of *L. trewavasae* 2005-1306. The SCR of these individuals ranges from 3 to 234. The two individuals shown are arbitrarily chosen representatives of these two genera.

The length and position of B blocks along the A chromosomes allows us to begin unraveling the history of the B chromosome. The presence of B blocks on every chromosome supports the idea that once a proto-B forms, it somehow acquires sequence from the rest of the genome. How these sequences make their way to the B, and which types of sequences are most likely to do so, is still unknown. Most discussion of mechanisms that transfer sequence to the B involves transposable elements (Houben, 2017). Other mechanisms, such as non-homologous recombination, could also be contributing to the acquisition of B sequence. B blocks range in size from a few hundred to a few hundred thousand bases. Homologous regions larger than 100 kb have been found on each of several chromosomes, suggesting that some, if not all, of these larger regions must have migrated to the B after its origin. So, the mechanisms responsible for the migration of sequences to the B must include a mechanism capable of moving and incorporating sequence blocks greater than 100 kb. While not common, transposable elements are known to move such large regions (Feschotte and Pritham, 2007). These large regions are not restricted to the distal chromosome arms, as would be expected if translocations were responsible. Furthermore, the core blocks appear to be evenly distributed across the LGs, suggesting that location along the A chromosome does not impact likelihood of migration to the B. Of course, if multiple mechanisms are involved in the acquisition

of A sequence, the combination of blocks acquired via these multiple methods might obscure actual patterns in the block location data.

The most extreme divergence in SCR of core blocks between the two individuals is found on LG8 (Figure 4.4). The SCR of this core block is 17.5 in the *M. lombardoi* B female and 234 in the *L. trewavasae* B female. This illustrates that copy number can vary greatly and is not an indication of how long a sequence has been on the B chromosome. We suggest caution in making interpretations about the origins of the B chromosome from observations of the length and position of B blocks or their copy number on the B chromosome. In these cichlids, the longest regions of homology are dispersed over too many chromosomes to suggest they were all involved in the production of the proto-B. Similarly, regions with some of the highest SCR, therefore contributing a significant amount of sequence to the B chromosome, are regions with relatively low SCR in other species. Moreover, the rate of Malawi cichlid B sequence replacement suggests that any B chromosome more than a few million years old may have replaced the original sequence of the proto-B to the point that none remain, making assignment of origin impossible. We suggest that efforts to identify the origin of B chromosomes focus on very young B chromosomes, and then use a combination of basic sequence homology, as performed here, as well as approaches that study chromosomal rearrangements and/or centromere evolution.

Genes

B chromosome gene sequences were identified as overlap between RefSeq annotated genes and B blocks. Annotated genes were either partially or completely encompassed in a B block. The total number of partial or complete genes in the B chromosome blocks is listed in Table 4.4.

Table 4.4: Genes and gene fragments on the B chromosome

Sample	Number of Genes and Gene Fragments
L. trewavasae 2005-1306	702
M. auratus 2008-1601	516
M. greshakei 2012-3493	972
M. lombardoi 2014-1018	2030
M. lombardoi 2014-1021	1688
M. lombardoi 2014-1108	1664
M. mbenji 2012-3997	899
M. zebra "Boadzulu" 2005-0976	1291
M. zebra "Boadzulu" 2005-0983	1262
M. zebra "Boadzulu" 2005-0986	1260
M. zebra "Nkhata Bay" 2012-5340	1094
M. zebra "Nkhata Bay" 2012-5347	1739
M. lombardoi 2016-1012 (PacBio)	678
L. trewavasae "Maison" XX females (control)	595
L. trewavasae "Maison" WZ females (control)	132
Core blocks	132

The number of genes and gene fragments overlapping with B blocks ranged from 516 to 2030 among datasets. Only 132 were common to at least 12 of the 13 datasets. When comparing individuals of the same population (Figure 4.5), the majority of genes identified were shared. However, several hundred genes were still unique to one or two of the individuals. We believe this is the result of the higher amount of type 1 error in the unique, unshared B blocks. Again, the core blocks provide us with

the most conservative estimate of gene number. Furthermore, the comparison between Illumina and PacBio datasets revealed that some blocks, while representing B sequence, are erroneously positioned in the A reference where a recent insertion of that repeat occurred. If such an insertion were to occur in the intron of a gene, our analysis would incorrectly identify that gene as being partially on the B chromosome, leading to an overestimate of gene number. Nevertheless, if the B chromosomes of these different species use the same gene(s) to achieve drive, it is reasonable to believe that gene might be found among the 132 genes common across species.

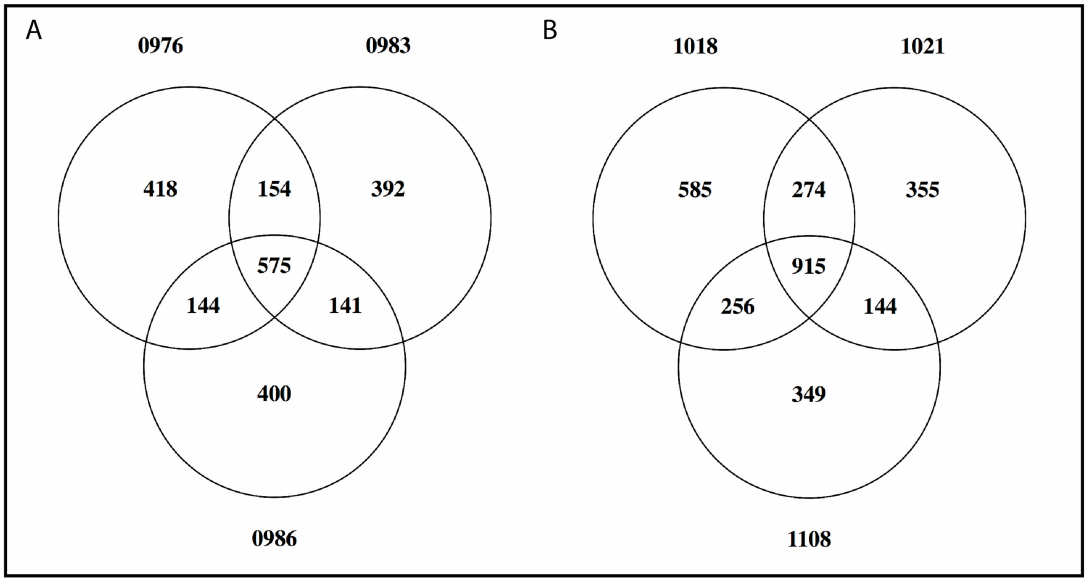

Figure 3: Shared B-located genes and gene fragments among individuals of the same population. The Venn diagrams show the number of genes and gene fragments shared by (A) three *M. zebra* "Nkhata Bay" individuals and (B) three *M. lombardoi* individuals The numbers outside the Venn diagram circles (0976, 0983, 0986, 1018, 1021 and 1108) are sample identifiers. The numbers inside the Venn diagram circles are the number of genes and gene fragments.

This analysis has identified both genes and gene fragments indiscriminately. The question remains whether these genes are functional or merely pseudogenes. While it may be tempting to label the gene fragments as pseudogenes, the structure of these sequences on the B chromosome is unknown. These gene fragments may be part of a gene fusion on the B, active but with an altered function. Moreover, transcription of altered (truncated, or partially deleted) copies of these genes could function by interfering with the activity of the original gene. Further examination of the genes on B chromosomes is needed before any conclusion regarding the functionality, or lack thereof, of B-located genes. A study of B sequence function will also serve to indicate which genes among these 132 could control B chromosome behavior, namely, drive and sex determination. A more complete understanding of the structure of the B, rather than a series of fragmented blocks, would further this goal. Future studies might benefit from using PacBio or other long read sequencing methodologies that are better able to assemble the repetitive sequence of the B chromosome.

Chapter 5: Transcriptomic Analysis of a B chromosome

Context and Motivation

Now that we know the sequence composition of the Lake Malawi B chromosome, we can inspect these sequences for genes contributing to drive. Arguably, the first step in searching for functional genes on the B chromosome is to assay for transcription. For many years, B chromosomes were believed to be transcriptionally silent and to have little function other than their drive. This was consistent with a general lack of associated phenotypes. However, with new technological advances, recent years have seen several reports of genome-wide differential expression in the presence of B chromosomes as well as transcription of B chromosome sequences. This invokes the question; do B chromosomes do more than just drive? Do they cause cryptic phenotypic changes not previously appreciated? Is the differential expression observed in these species due to sequences expressed from the B chromosome itself, or does the presence of the B in the genome actually change the expression of A chromosome genes? These questions will be addressed below by quantifying the amount of differential expression between B and NoB females in *Metriaclima lombardoi* in brain, gonad and liver. Transcripts of sequences found on both the A and B chromosomes will be examined to determine their source. This will simultaneously capture B-specific expression as well as further explain the amount of differential expression observed.

Methods

Brain, gonad (ovary) and liver tissue was collected from lab-reared *Metriaclima lombardoi* females immediately after euthanasia and stored in RNAlater. A total of 6 individuals were used in this study, 3 B females and 3 NoB females. Females were collected as adults at an approximate length 55 mm. In order to choose females at a similar stage of oogenesis, the size of yolk sacs visible in ovaries was size-matched between individuals. As few B females were available, the 3 B individuals selected were full siblings, while the 3 NoB individuals were unrelated. RNA was extracted using the Ambion *mir*Vana extraction kit. Illumina's TruSeq Stranded mRNA library prep kit was used for library prep. Libraries were barcoded and multiplexed during sequencing across 3 flow-cell lanes of an Illumina HiSeq1500. Libraries were arranged across these 3 lanes such that each lane had the brain, gonad and liver samples of one B and one NoB individual. This was done to avoid lane-specific patterns of bias. Paired-end 100bp reads were trimmed with Trimmomatic (v0.32).

Two differential expression analyses were performed on the RNA-Seq data through two different bioinformatic pipelines, each leveraging alignment to a different reference: 1) the reference genome and 2) a *de novo* transcriptome assembly (Figure 5.1). A *de novo* transcriptome assembly was produced for each tissue from both B and NoB sequence data using Trinity (version r20140717). A single transcriptome assembly from all available samples was necessary, but difficult produce with Trinity alone. Instead, these three Trinity assemblies were then merged into a single assembly with EvidentialGene [http://arthropods.eugenes.org/EvidentialGene/trassembly.html].

Salmon (v0.12.0) was used to quantify transcription of the 18 samples using this merged transcriptome assembly. A differential expression analysis was performed with the Wald-test option in the R package Sleuth [https://github.com/pachterlab/sleuth].

For the alignment to the reference genome, reads were aligned to *Metriaclima zebra* (M_zebra_UMD2) using Tophat2 (v2.1.1) and analyzed with the Cufflinks suite (v2.1.1). Cuffmerge takes the assembled transcripts of several samples and merges them together producing one list of loci to be used in downstream analyses. This list of loci does not have a one to one correspondence with annotated genes and for this reason, the results of this study will be reported in terms of loci, not genes. When a locus contains more than one annotated gene, this will be described appropriately. Please note, because Chapter 4 reported findings in terms of genes, and loci are reported here, numbers will vary slightly between the two chapters. For example, the study described in Chapter 4 identified 132 genes present on the core blocks. This is equivalent to 144 loci, as defined by Cuffmerge. The process of identifying genes did not account for multiple copies of the same gene or overlapping annotations. Alternatively, the process of identifying loci does not account for multiple genes combined into a single locus.

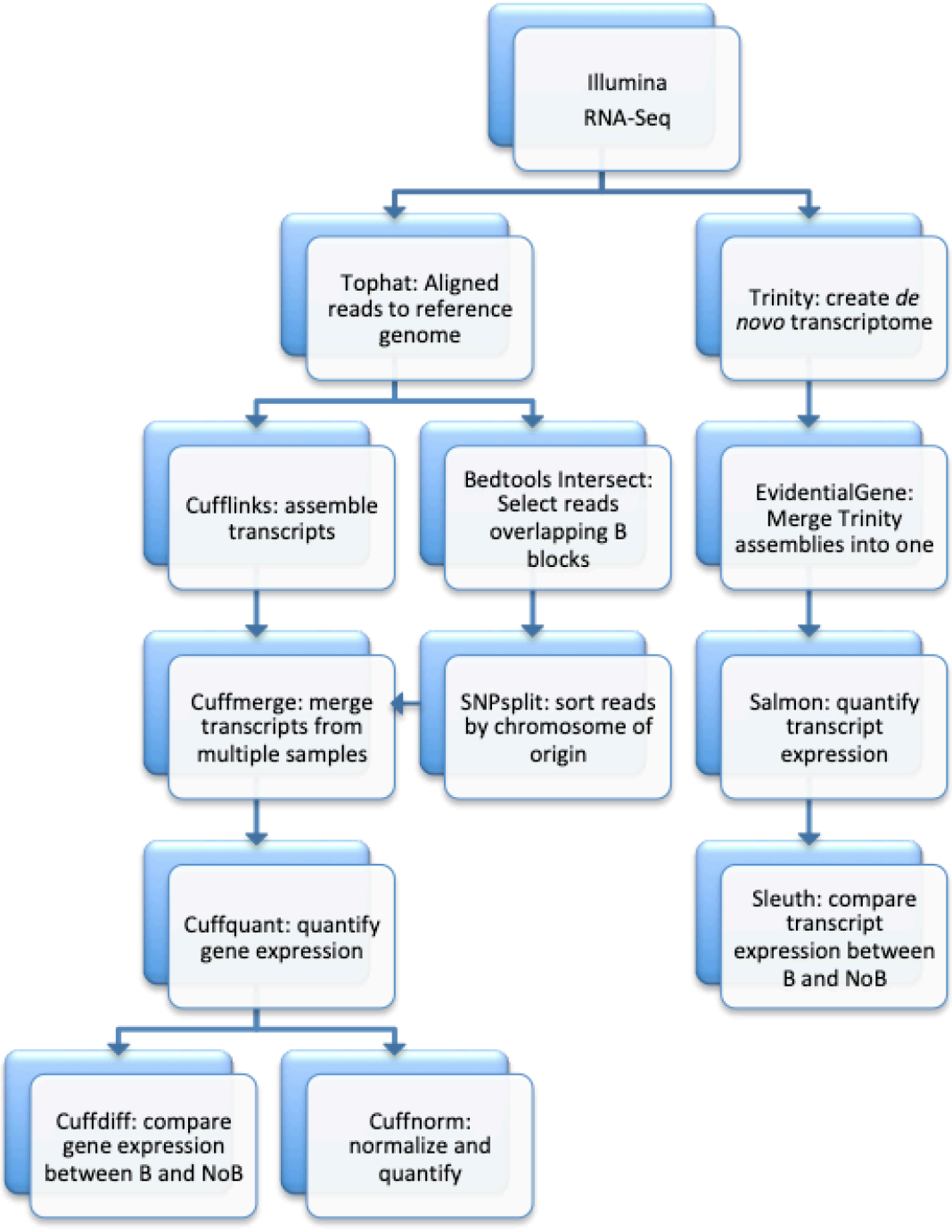

Figure 5.1: Differential expression pipeline. This flow-chart represents the analyses performed on RNA-Seq data from *Metriaclima lombardoi* females with and without a B chromosome. At each stage (box), the software used is listed followed by a simple description of the task that software accomplished.

B blocks were identified with genomic data according to the methods in Chapter 4 (Clark et al. 2018). Two types of B block sets were used for this transcriptome analysis, the "core" blocks and the "*M. lombardoi* core" blocks. Core blocks are regions where a B block was identified in at least 12 out of 13 B individuals across three genera. This strategy of identifying core sequence leads to a lower false-positive (type 1 error) rate and indicates sequence shared by the B chromosomes across three genera. However, B chromosomes are expected to experience a high sequence turnover rate. This faster sequence evolution and the clonal maternal transmission of the Lake Malawi B suggests that the B chromosomes of different species likely have unique sequence in addition to this shared sequence. The core blocks will not include any sequences specific to the B chromosome of *M. lombardoi*. While we expect any sequence of great importance to the maintenance of the B to be preserved and therefore represented in the core blocks, the content of the *M. lombardoi*-specific B sequence may contribute to the differential expression observed in our transcriptome analysis. Towards this end, we have also produced a *M. lombardoi* specific "core" block set, which will be referred to as the "*M. lombardoi* core" block set. The core block set includes 1.37 Mb of sequence in the A genome, while the *M. lombardoi* core block set includes this 1.37 Mb and an additional 3.95 Mb, or 5.32 Mb in total.

RNA-Seq reads aligned to the reference genome were examined for SNPs to determine whether they originated from an A chromosome or B chromosome. The genomic data analyzed in Chapter 4 was used first to identify B-specific SNPs to be used for this analysis. The aligned genomic reads from a *Metriaclima lombardoi*

individual (2014-1018) and the *Metriaclima zebra* ‘Boadzulu’ male pool were processed with Freebayes (v1.0.2-33-gdbb6160-dirty) to call SNPs and indels compared to the reference genome (M_zebra_UMD2). SNPsift (http://snpeff.sourceforge.net) was then used to filter for positions where the B *M. lombardoi* female possessed an “alternate” SNP and the NoB *M. zebra* ‘Boadzulu’ male pool matched the reference. These SNPs were then overlapped with the core blocks and then with the lombardoi core blocks using Bedtools Intersect. This provided two sets of SNPs, core SNPs and lombardoi core SNPs, within regions identified as B blocks and with a B-specific pattern.

Bedtools Intersect was used again to filter the aligned transcriptomic reads, and their corresponding paired ends, that align to a region identified as a B block. This was done for the core blocks and the lombardoi core blocks. The B-specific SNPs were used in conjunction with SNPsplit (v0.3.2) to sort these filtered transcriptomic reads as either B-specific, A-specific, conflicting or unassignable. These sorted reads were normalized and quantified with the Cuffnorm option in the Cufflinks suite.

Differential Expression

The differential expression analysis leveraging the *de novo* transcriptome assembly identified 949 significantly differentially expressed transcripts (q-value <= 0.05) between B and NoB individuals in brain, 189 in gonad and 7298 in liver. These results are depicted in the form of volcano plots (Figure 5.2) and heat maps (Figure 5.3).

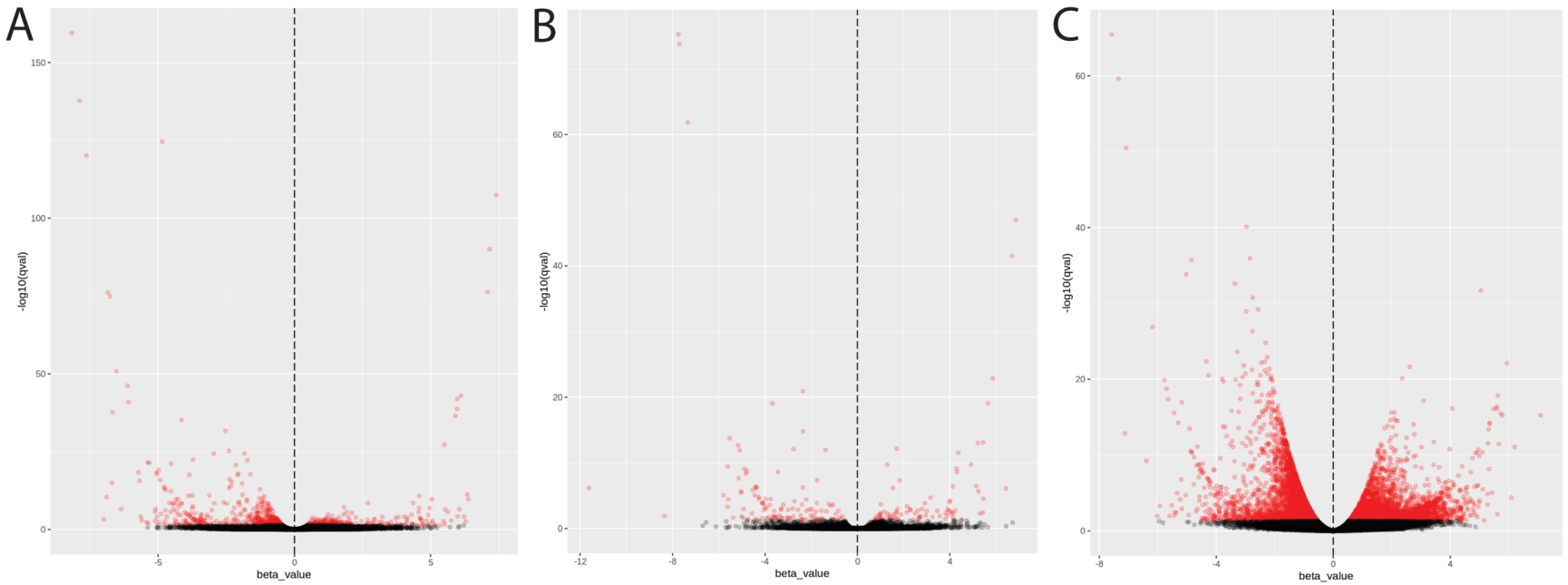

Figure 4.2: Volcano plots: differential expression quantified with Sleuth and based on the *de novo* transcriptome assembly. Differential expression between B and NoB individuals is shown with volcano plots for brain (**A**), gonad (**B**) and liver (**C**). Statistical significance (q-value) is plotted on a –log10 scale on the *y*-axis. Beta values, a measure of fold change in expression, is plotted on the *x*-axis. Each dot represents a transcript in the *de novo* transcriptome. Red dots represent statistically significant differential expression; black dots represent non-statistical significance. Values to the left of the dashed line (beta value < 0) represent transcripts with lower expression levels in B individuals. Values to the right of the dashed line (beta value > 0) represent transcripts with higher expression levels in B individuals.

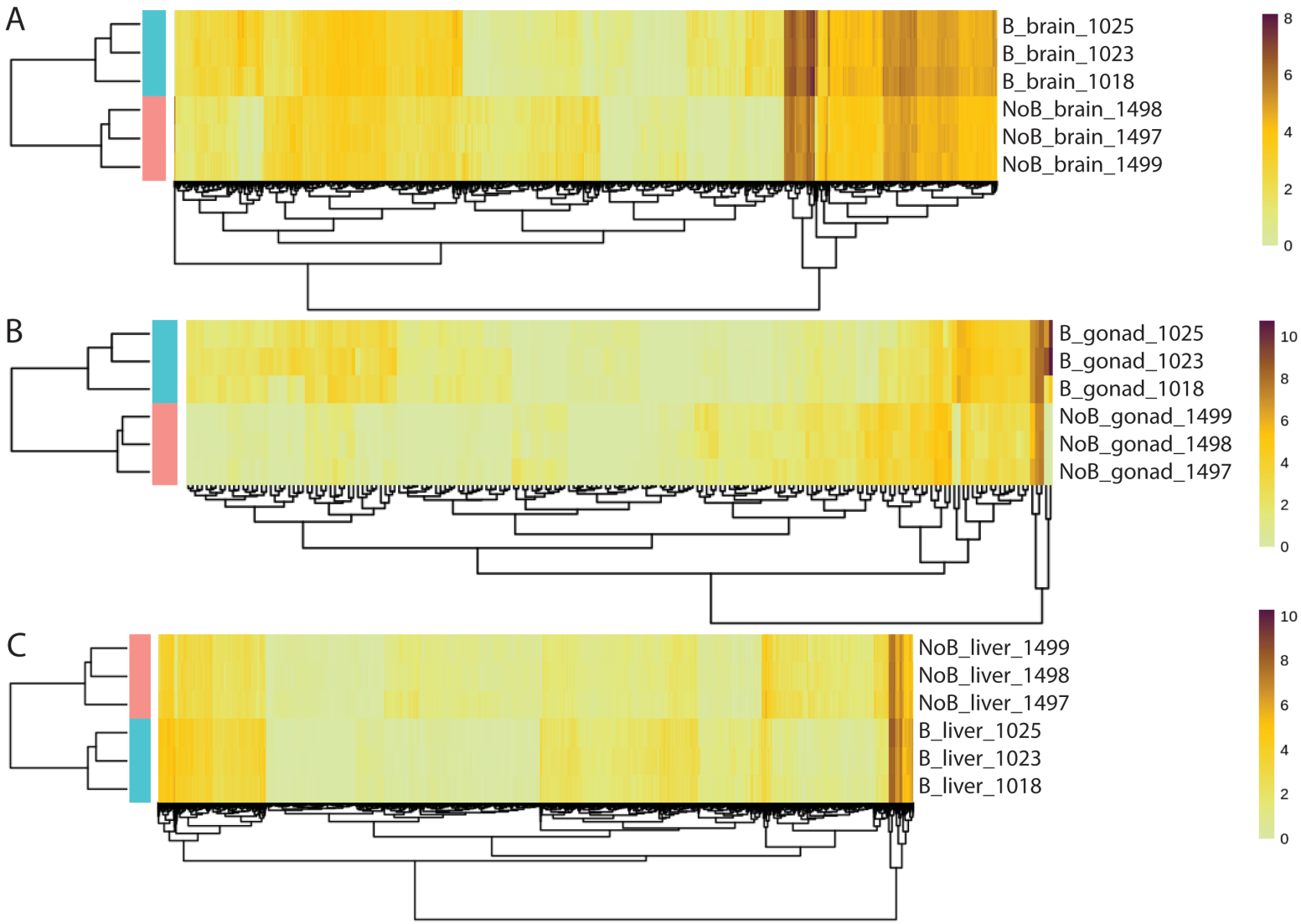
A
B_brain_1025
B_brain_1023
B_brain_1018
NoB_brain_1498
NoB_brain_1497
NoB_brain_1499
8
6
4
2
0
B
B_gonad_1025
B_gonad_1023
B_gonad_1018
NoB_gonad_1499
NoB_gonad_1498
NoB_gonad_1497
10
8
6
4
2
0
C
NoB_liver_1499
NoB_liver_1498
NoB_liver_1497
B_liver_1025
B_liver_1023
B_liver_1018
10
8
6
4
2
0

Figure 5.3: Heat maps: differential expression quantified with Sleuth and based on the *de novo* transcriptome assembly. Differential expression between B and NoB individuals is shown in the form of heat maps of expression in brain (**A**), gonad (**B**) and liver (**C**). Only significantly differentially expressed transcripts have been included. Hierarchical clustering of samples and transcripts is shown to the left and below the heat maps, respectively.

The differential expression analysis leveraging the reference genome found 1266, 540 and 6352 loci to be significantly differentially expressed (q-value <= 0.05) in brain, gonad and liver, respectively, between B and NoB individuals. This represents 4.5%, 1.9% and 22.4%, respectively, of the 28,353 loci in the genome. The results of this differential expression analysis are portrayed in volcano plots (Figure 5.4) and heat maps (Figure 5.5). Only the significantly differentially expressed loci are included in the heatmaps. The majority of these loci in two tissues, 67% in brain and 68% gonad, exhibit higher expression in the NoB individuals. In liver, 46% exhibit higher expression in NoB individuals. Of the 1266 loci significantly differentially expressed in brain, 10 have copies on the B according to the core blocks, and 48 according to the lombardoi core blocks. Of the 540 loci significantly differentially expressed in gonad, 14 and 23 were found to have B copies according to the core and lombardoi core blocks, respectively. Of the 6352 loci significantly differentially expressed in the liver, 23 and 104 were found to have B copies according to the core and lombardoi core blocks, respectively. Of the loci significantly differentially expressed and on the core, 80%, 64% and 65% exhibit higher expression in the B individuals in brain, gonad and liver, respectively.

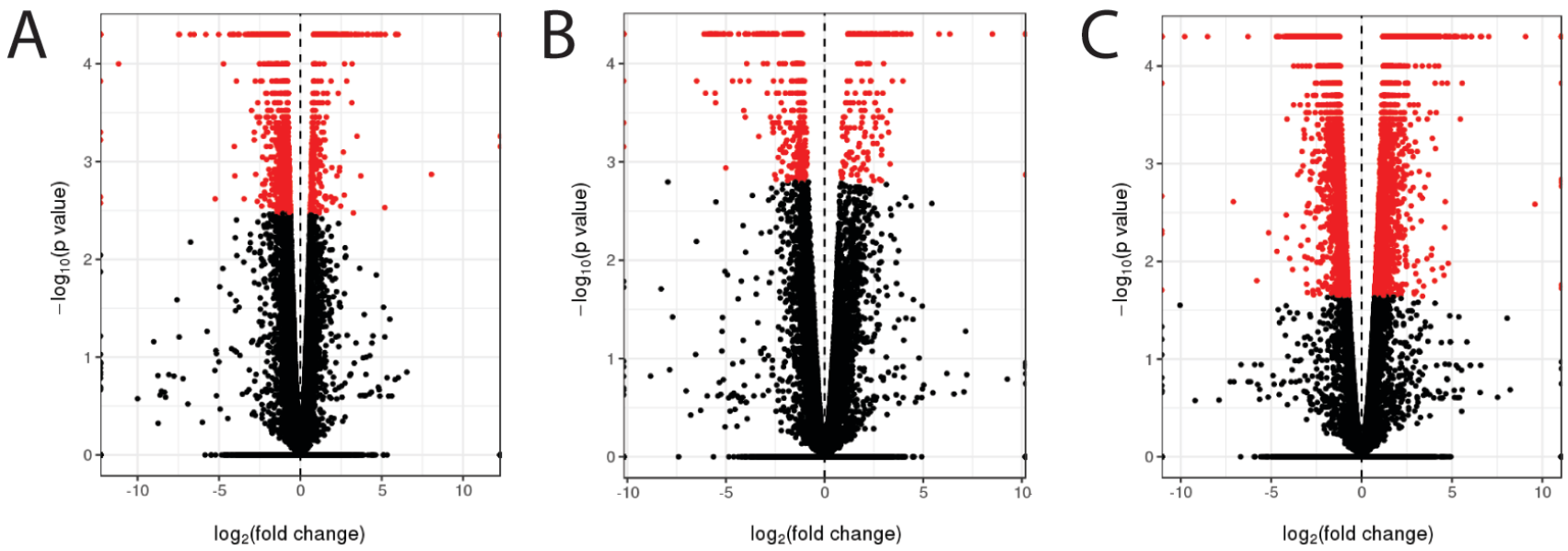

Figure 5.4: Volcano plots: differential expression quantified with Cuffdiff and based on the reference genome assembly. Differential expression between B and NoB individuals is shown with volcano plots for brain (**A**), gonad (**B**) and liver (**C**). Statistical significance (p-value) is plotted on a –log10 scale on the *y*-axis. Change in expression is plotted as $\log_2$ fold change on the *x*-axis. Each dot represents a locus in the reference genome. Red dots represent statistically significant differential expression (q-value < 0.05); black dots represent non-statistical significance. Values to the left of the dashed line represent transcripts with lower expression levels in B individuals. Values to the right of the dashed line represent transcripts with higher expression levels in B individuals.

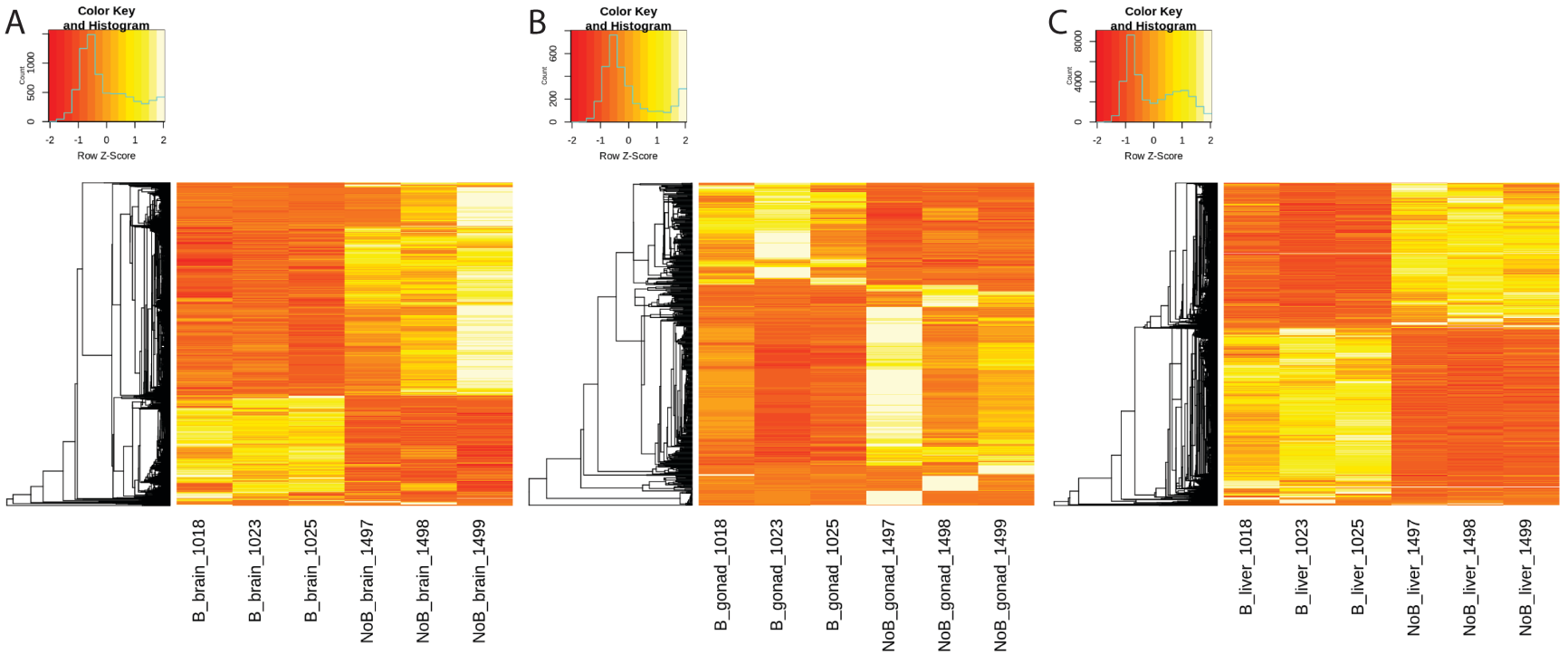

Figure 5.5: Heat maps: differential expression quantified with Cuffdiff and based on the reference genome assembly. Differential expression between B and NoB individuals is shown in the form of heat maps for brain (**A**), gonad (**B**) and liver (**C**). Only significantly differentially expressed transcripts have been included.

Even when using the less stringent B-sequence identification method (that results in the lombardoi core blocks) and thus a larger number of loci being identified as presumably on the B, less than 5% of the significantly differentially expressed loci can be explained by B-specific transcription. This suggests that the B chromosome has an effect on the transcription of the A genome. Whether or not the observed difference in expression of any of these genes is enough to lead to a phenotypic change beyond the molecular phenotype of transcription is unknown. Theoretically, the causal sequence behind the B's effect on sex determination could be among these differentially expressed A genes. It is important to note that the genes involved in differential expression are not limited to highly repetitive sequences such as transposable elements and ribosomal RNA genes, but include many protein-coding genes.

Worthy of note, the majority of loci with copies on the B chromosome exhibit higher expression in B individuals. One interpretation of this is that the differential expression of these loci is likely resulting from having the same level of transcription from the A chromosome copy of each locus with the added transcription from the B chromosome copy or copies. Yet when examining all significantly differentially expressed loci, not just those with copies on the B chromosome, the trend is towards higher expression in the NoB individuals. This suggests that other than the expression of sequences from the B itself, the B chromosome more often causes the down-regulation of A-located sequences, at least in brain and gonad. This could be the result of a genome-wide effect of B chromosome presence, such as altered DNA

methylation patterns or histone modifications. Since B-located sequences only account for less than 5% of the differentially expressed loci, the B must, through some mechanism, change the expression of A-located sequences. RNA or protein expressed from the B chromosome could interact with the A genome in a gene-specific manner to result in this differential expression. Alternatively, the B could cause a more global change by altering chromatin structure or position within the nucleus. It is also possible that the general decrease in gene expression levels among the B individuals is the result of inadequate transcript quantification methods as opposed to a biological phenomenon. Normalization of RNA-seq data is both essential and difficult. Significant up-regulation of a few genes in one sample type could result in the false detection of differential expression among other genes (Robinson and Oshlack, 2010). In addition to accounting for library size, the normalization methods used by Cuffdiff leverage non-differentially expressed genes to normalize read count distributions across samples (Evans et al. 2018). Theoretically, this should prevent false positives due to the additional transcripts produced from the B chromosome. More work is needed to confirm this bias in differential expression, determine how common it is among B chromosomes in other taxa, and examine the molecular cause(s).

Expression of B-located Loci

Of the 144 loci found among the core sequence, only 14 were completely overlapping the core blocks and subsequently are being referred to as “complete” loci. It is important to note that “complete” is not indicative of function nor suggests that non-

complete loci lack function. It is not difficult to imagine that truncated proteins that impede the original A copy or even gene-fusions with an altered function could still be a path for B chromosome function. Similarly, of the 781 loci in the lombardoi core sequence, only 37 were complete.

Among the 144 core loci, 57 were differentially expressed (p-value <= 0.05), of which 35 were significantly differentially expressed (q-value <= 0.05) in at least one of the three tissues. Only three of the 144 loci were found to be both complete on the B and significantly differentially expressed in at least one tissue; XLOC_010072 (nsmce4a), XLOC_003603 (CBL-like), XLOC_004619 (RTEL1). Each of these loci include a single annotated gene, indicated in parentheses following their unique locus identifier. XLOC_003603 (CBL-like) and XLOC_004619 (RTEL1) were significantly differentially expressed in the gonad while XLOC_010072 (NSMCE4A) was significantly differentially expressed in brain and liver. Interestingly, XLOC_010072 (NSMCE4A) was differentially expressed in the gonad, but not significantly (p-value = 0.00425, q-value = 0.08661). Of the 781 loci on the lombardoi core blocks, 240 were differentially expressed, of which 147 were significantly differentially expressed in at least one of the three tissues. Examining the lombardoi core loci that were both complete and significantly differentially expressed in at least one tissue, there are only 3 loci in addition to those found to be complete and significantly differentially expressed among the core loci. These 3 additional loci all contain 3 uncharacterized annotated genes. XLOC_008020 yields a non-coding RNA and was significantly differentially expressed in all three tissues.

XLOC_000637 and XLOC_005025 both yield mRNA and both were significantly differentially expressed in the gonad.

Using the core SNPs to evaluate the 144 core loci, we were able to detect expression (from either A or B sequence) for 123, 124 and 106 loci in brain, gonad and liver, respectively. Of these, 46 in brain, 50 in gonad and 36 in liver have B-specific transcription. A total of 53 loci have B-specific expression in at least one tissue. Of the 781 loci on the lombardoi core, the lombardoi core SNPs allowed us to distinguish B and A expression for 655, 667, 567 loci in brain, gonad and liver, respectively. Of these, 101 in brain, 103 in gonad and 83 in liver have B-specific expression. A total of 126 lombardoi core loci have B-specific expression in at least one tissue.

While the detection of a B-specific origin of reads was possible for the loci mentioned above, other loci did not possess B-specific SNPs within exons, making reads aligning to these loci "unassignable." There are two possible explanations for a lack of B-specific SNPs: 1) these loci have recently migrated to the B chromosome and have yet to accumulate sequence variants or 2) the exons of the loci have experienced purifying selection because they are functional and necessary for B chromosome maintenance. For this reason, it is important to scrutinize these loci that have high exon sequence similarity and unsubstantiated, yet potential expression from the B. In the brain, 55 of the 123 expressed core loci had only unassignable reads. In the gonad, 57 of the 124 expressed core loci had only unassignable reads. And in the liver, 49 of

the 106 expressed core loci had only unassignable reads. Similarly, 393, 370 and 273 of the 655, 667, and 567 expressed lombardoi core loci in brain, gonad and liver, respectively, had only unassignable reads. Counting loci which either had confirmed B-specific expression or only unassignable reads across all three tissues, we were able to account for expression of 118 of the 144 core loci and 582 of the 781 lombardoi-core loci.

Gene Candidates for Drive

Differential expression, B-specific expression and gene ontologies were used to identify 6 candidate genes from among the 144 loci in the B chromosome core sequence. These genes are being evaluated for their possible roles in B chromosome drive. For each, expression was highest in gonad in both B and NoB individuals. The transcripts from the *de novo* transcriptome assembly were aligned to the reference genome and compared with gene annotations and B blocks to assess the B-located copy. This along with expression data for the 6 candidates is described below. Figure 5.6, borrowed from Wood et al. 2008, depicts some of the known roles of the first 3 candidate genes in kinetochore attachment to microtubules.

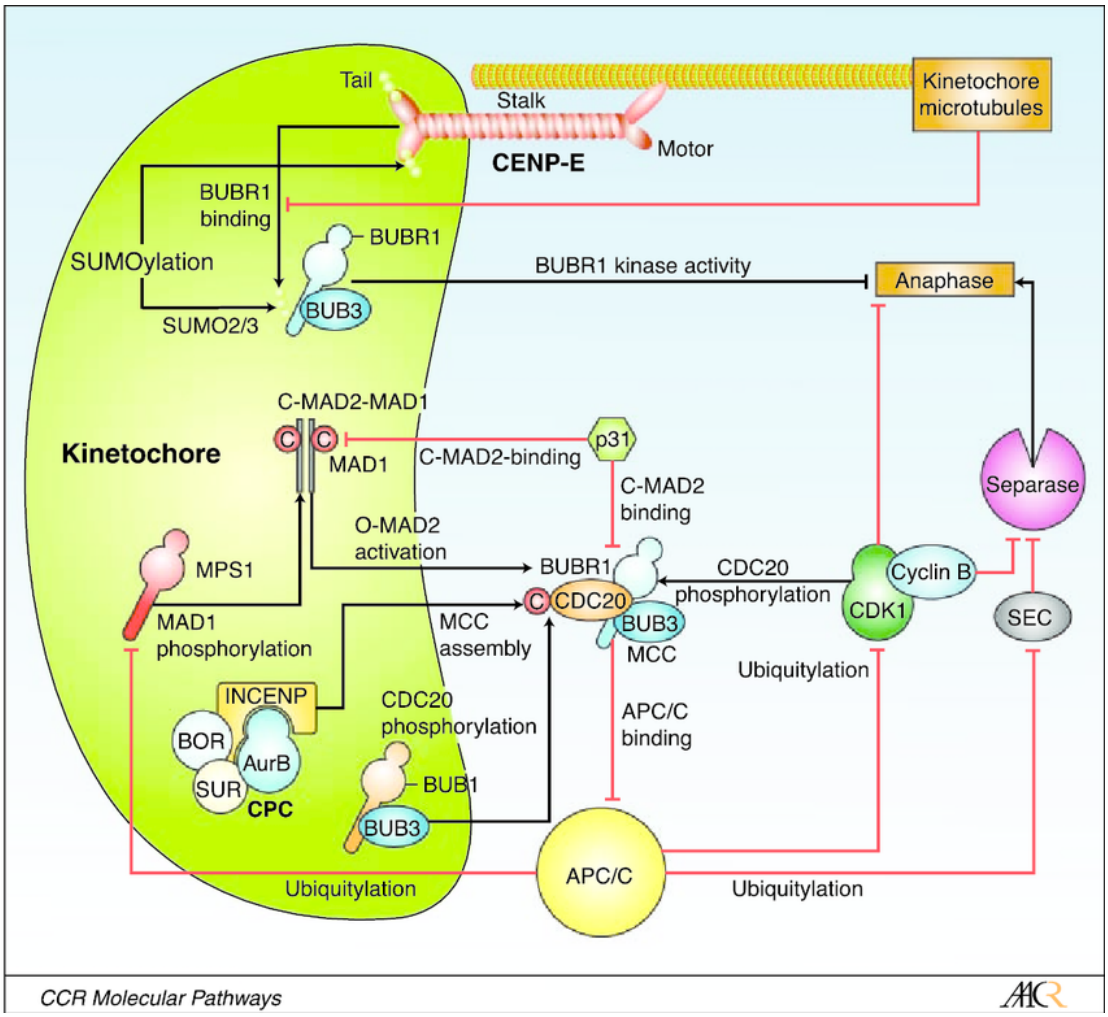

Figure 5.6: CENP-E and kinetochore mitotic checkpoint signaling figure as published in Wood et al. 2008. This figure illustrates some known roles of CENP-E, INCENP and MAD2 in its alternate states, C-MAD2 and O-MAD2.

INCENP

Inner centromeric protein (INCENP) connects the two functional modules of the chromosome passenger complex (Krenn and Musacchio 2015). The N terminal region of INCENP associates with Survivin and Borealin and localizes to centromeres. The C terminal region of INCENP activates Aurora B. This chromosome passenger complex is responsible for many processes, including mitigating unsuccessful attachment of chromosomes and microtubules as well as activation of the spindle

assembly checkpoint (Carmena et al. 2012). An increase in expression of INCENP was observed in all three tissues of B individuals (significantly in gonad and liver). B-specific reads, detected with core SNPs, were found in the transcriptome of each tissue. It appears that INCENP is only partially on the B chromosome (FIGURE 5.7). Consistent with this, a transcript was assembled containing only exons contained within the area of B block overlap. This could represent the truncated RNA transcript expressed from the B. A nearby B block overlaps not only the first exon, but also 1800 bp upstream, presumably including the proximal promoter. If the binding regions on either end of this protein are disrupted by exon deletion, this truncated protein expressed from the B chromosome may hinder INCENP's function, slowing progression through the spindle assembly checkpoint. B chromosomes are known to lag though cell division (Houben 2017; Houben et al. 1997), and a mechanism that slows this process down could give the B chromosome time to reach a more optimal position.

Figure 5.7: INCENP. Region of LG7 viewed in IGV where INCENP overlaps two B blocks. Tracks I and II are the PacBio read coverage depth of the *M. zebra* 'Mazinzi' male individual used to create the reference genome assembly and an *M. lombardoi* B female, respectively. Tracks III and IV show the core blocks (blue) and *M. lombardoi* core blocks (green) respectively. Track V is the alignment of the reference assembly gene annotation (orange). Track VI is the alignment of the *de novo* transcriptome assembly.

CENP-E

Centromere Protein E (CENP-E) is a kinesin-like motor protein that connects the centromere-bound kinetochore to microtubules (Gudimchuk et al. 2013; Ciossani et al. 2018). This association is important for chromosome alignment and progression though the spindle assembly checkpoint. Disruption of this gene results in chromosome instability and aneuploidy (Veneziano et al. 2018). This study found CENP-E to be differentially expressed in gonad, though not significantly (p-value = 0.03515 5, q-value = 0.28237). While there were no core SNPs with which to distinguish reads, the *M. lombardoi* core SNPs were used to detect B-specific expression in brain and gonad and very little (fpkm < 1) in liver. Gene annotations in this region reveal two overlapping open reading frames, one identified as CENP-E, the other CENP-E-like. It is unclear if this region does include two homologous genes, or if this gene was mistakenly divided into two parts when annotating the genome. Overlapping B blocks suggest the B-located copy of CENP-E is missing several exons (FIGURE 5.8). A transcript from the *de novo* transcriptome assembly seems to mirror this, possessing exons of both the CENP-E and CENP-E-like annotations, but missing those exon in the middle where genomic read coverage is low. It is possible the B-located copy of CENP-E, missing several exons, could interfere with the functional A-located copy, again, slowing progression through the spindle assembly checkpoint. Alternatively, this altered composition of this motor protein could make it more effective than the A-located copy, moving the B chromosome to a more advantageous position prior to anaphase.


42 kb
90 kb
100 kb
110 kb
120 kb
Track I. M. zebra 'Mazinzi' Male PacBio Coverage
Track II. M. lombardoi B Female PacBio Coverage
Track III. Core Blocks
Track IV. Lombardoi Core Blocks
Track V. Gene Annotations
Track VI. Transcriptome Assembly
XM_014409329.1
XM_014409327.1
XM_014409328.1
XM_004564729.3
XM_012922879.1
XM_004564721.1
XM_014409307.1
XM_014409308.1

Figure 5.8: CENP-E. Region of the unanchored scaffold 000897F_pilon_quiver. Track I is the PacBio read coverage depth of the *M. zebra* 'Mazinzi' male individual used to create the reference genome assembly. Track II shows the PacBio read coverage depth of an *M. lombardoi* B female. Tracks III and IV show the core blocks (blue) and *M. lombardoi* core blocks (green) respectively. Track V is the alignment of the reference assembly gene annotation (orange). Track VI is the alignment of the *de novo* transcriptome assembly.

MAD2A-like

The mitotic spindle assembly checkpoint protein, Mitotic Arrest Deficient 2A like (MAD2A-ike), interacts with the anaphase-promoting complex (APC) to initiate anaphase (Sironi et al. 2002). MAD2 localization to the centromere is dependent on CENP-E. Both INCENP and MAD2 are part of protein complexes that interact with the mitotic checkpoint complex (MCC) (Wood et al. 2008). There are two copies of MAD2A-like in the A genome identified through these analyses; one on the highly repetitive LG3 and the other among the unanchored scaffolds of the reference genome. The copy on LG3 has highly variable coverage with Illumina data, but comparatively low coverage with PacBio data (Figure 5.9). Examining the PacBio coverage, it seems possible that the identification of a B block overlapping MAD2A-like on LG3 was the result of misaligned reads from the copy on the unanchored scaffold. The copy on the unanchored scaffold does not completely overlap the core or *M. lombardoi* core blocks, but could be a complete in some individuals (Figure 5.10), perhaps in low copy number. The copy on the unanchored scaffold did not experience any differential expression and failed to provide any SNPs with which to identify B-specific reads. The copy on LG3, however, was differentially expressed in brain (p-value = 0.00465, q-value = 0.06161). This copy had core SNP data to detect B-specific expression in all three tissues.

Figure 5.9: MAD2A-like on LG3. Region of LG3 with little evidence of a B block. Track I is the PacBio read coverage depth of the *M. zebra* 'Mazinzi' male individual used to create the reference genome assembly. Track II shows the PacBio read coverage depth of an *M. lombardoi* B female. Tracks III and IV show the core blocks (blue) and *M. lombardoi* core blocks (green) respectively. Track V is the alignment of the reference assembly gene annotation (orange). Track VI is the alignment of the *de novo* transcriptome assembly.

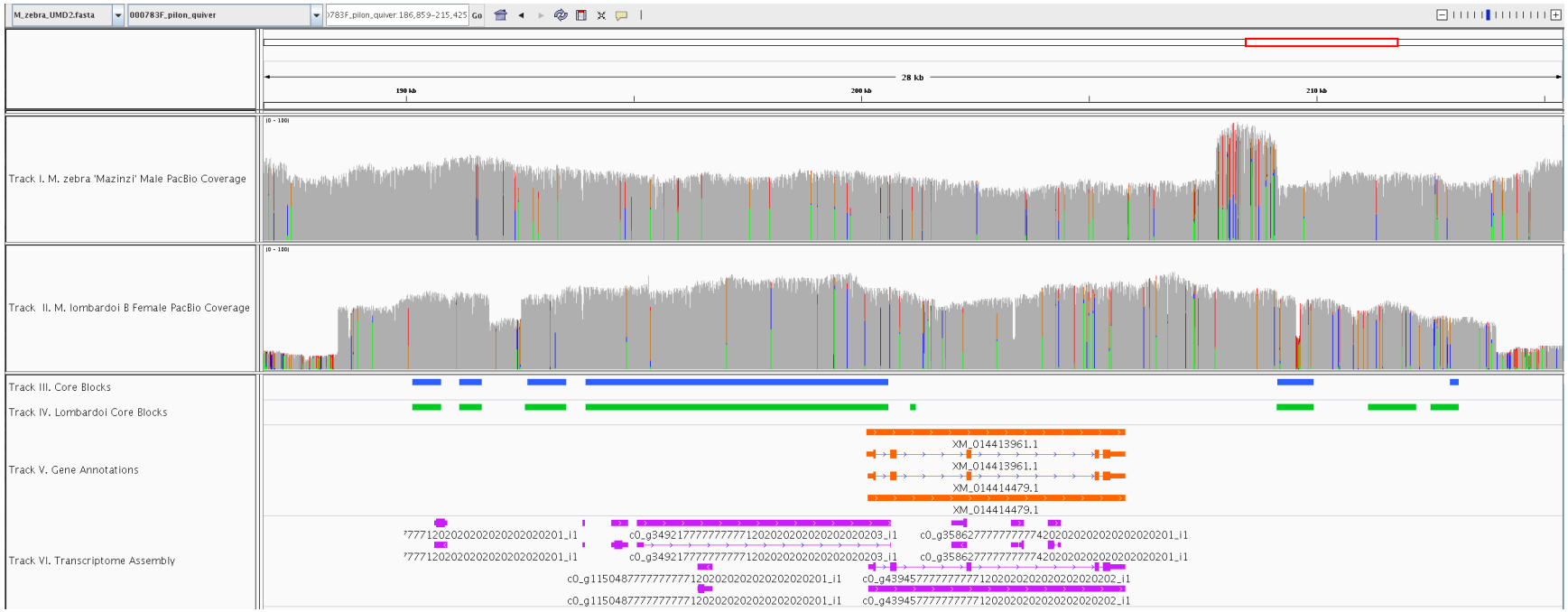

Figure 5.10: MAD2A-like on an unanchored scaffold. Region of the unanchored scaffold, 000783F_pilon_quiver. Track I is the PacBio read coverage depth of the *M. zebra* 'Mazinzi' male individual used to create the reference genome assembly. Track II shows the PacBio read coverage depth of an *M. lombardoi* B female. Tracks III and IV show the core blocks (blue) and *M. lombardoi* core blocks (green) respectively. Track V is the alignment of the reference assembly gene annotation (orange). Track VI is the alignment of the *de novo* transcriptome assembly

SYCE1

Synaptonemal complex central element 1 (SYCE1) forms a protein structure with SYCP1 and CESC1, linking homologous chromosomes during prophase I. This structure is shown in Figure 5.11, borrowed from Costa et al. 2005. SYCE1 plays a role in synapsis and recombination. It is also hypothesized to cause ovarian insufficiency when disrupted (Costa et al. 2005; Vries et al. 2014). SYCE1 is differentially expressed in gonad, but not significantly (p-value = 0.03885, q-value = 0.29795) where it has higher expression in B individuals. While there weren't any core SNPs to distinguish reads aligning to this gene, the *M. lombardoi* core SNPs were used to detect B-specific reads in gonadal tissue. The B-located copy of SYCE1 appears to be missing a ~3kb region of an intron, but has every exon, according to the *M. lombardoi* core blocks (Figure 5.12). Alternatively, the A genome may have experienced a 3kb insertion in this intron; further comparative analyses would have to be performed to determine the evolutionary history. The core blocks suggest this gene is less intact on the B of other species, or in low copy number. It is interesting to speculate how this protein, meant to bind homologous chromosomes, might be manipulated by the univalent B chromosome. A B chromosome in the closely related cichlids in Lake Victoria was shown to be an isochromosome. The metacentric B depicted in the karyotype in Figure 2.2 could also be an isochromosome. It is possible that the B chromosome is forming a synaptonemal-like complex between its two arms to better segregate during meiosis, though this is highly speculative.

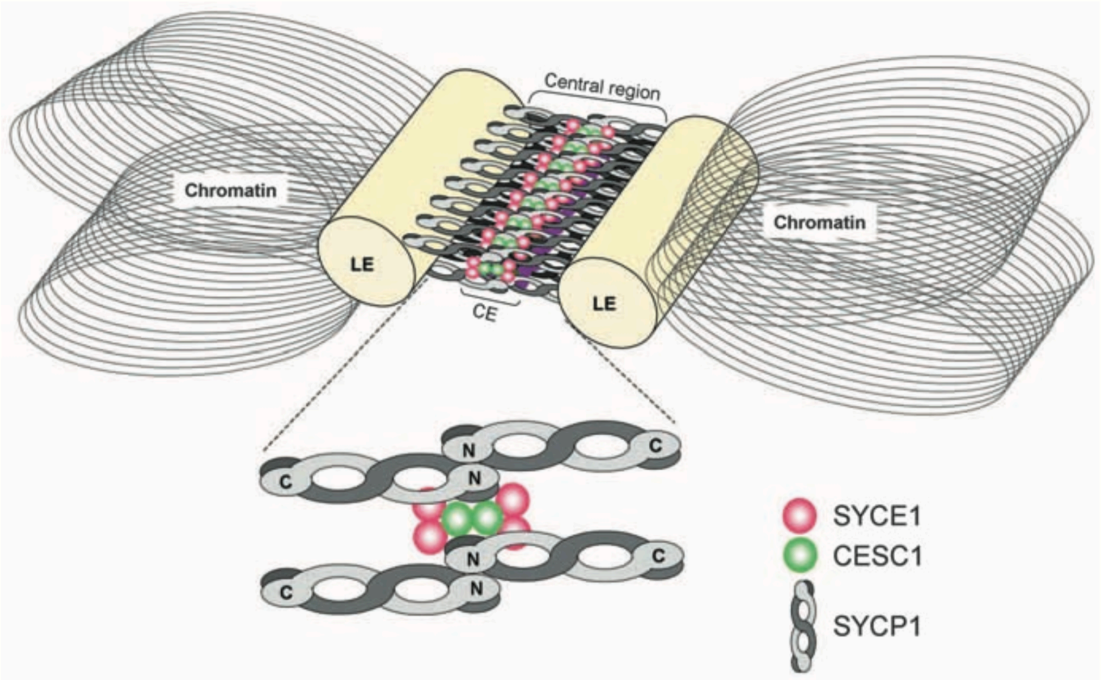

Figure 5.11: Model for the distribution of SYCE1 and CESC1 within the synaptonemal complex, as published in Costa et al. 2005.

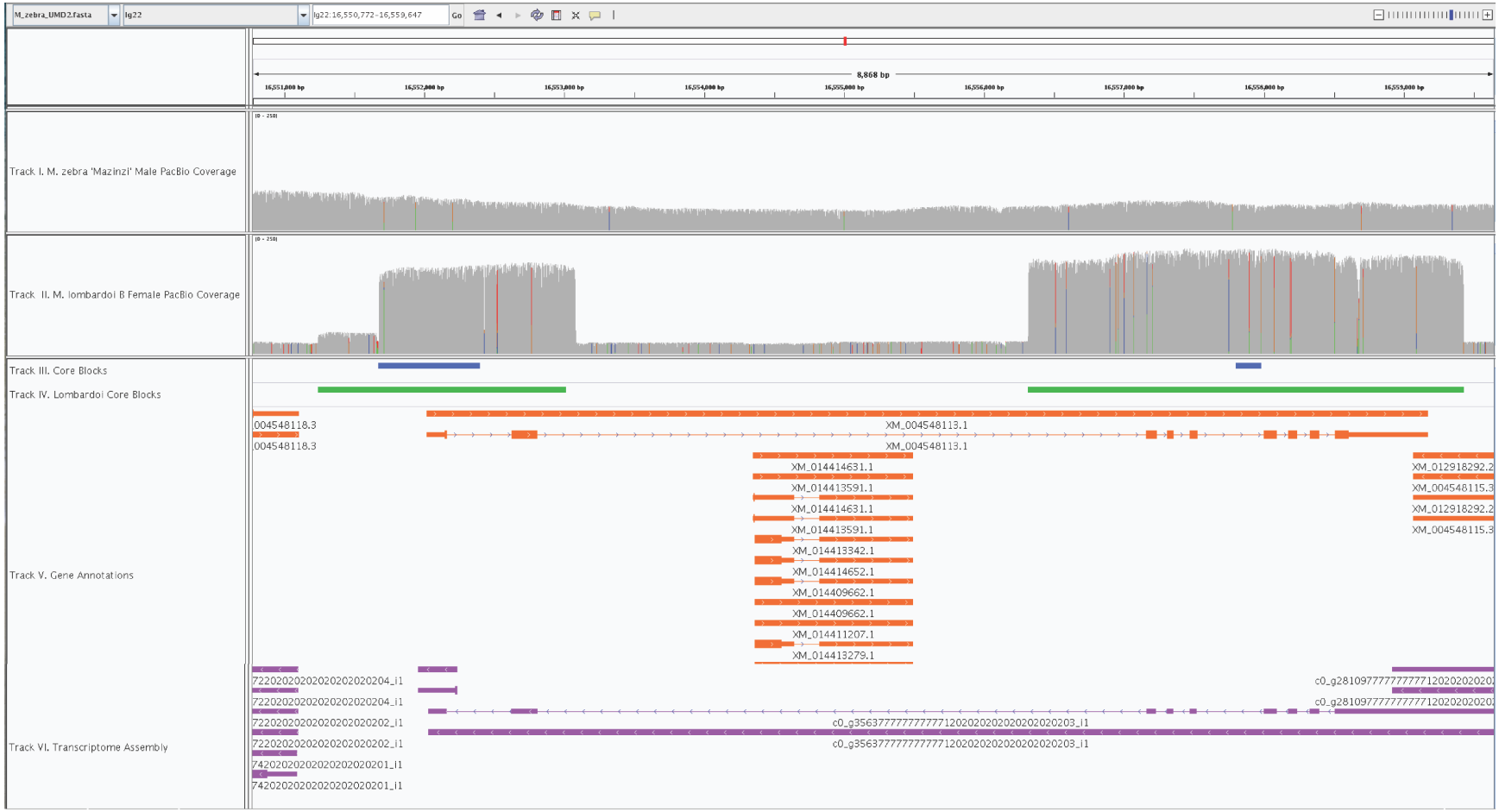

Figure 5.12: SYCE1. Region of LG22 with two B blocks overlapping SYCE1 Track I is the PacBio read coverage depth of the *M. zebra* 'Mazinzi' male individual used to create the reference genome assembly. Track II shows the PacBio read coverage depth of an *M. lombardoi* B female. Tracks III and IV show the core blocks (blue) and *M. lombardoi* core blocks (green) respectively. Track V is the alignment of the reference assembly gene annotation (orange). Track VI is the alignment of the *de novo* transcriptome assembly.

NSMCE4A

The Non-Structural Maintenance of Chromosomes Element 4 (NSMCE4) homolog A of the structural maintenance of chromosomes complex (SMC5-SMC6), NSMCE4A for short, has many roles in chromosome maintenance. Namely, it is involved in homologous recombination, recruiting cohesin to double-stranded breaks, DNA replication and telomere maintenance (Palecek et al. 2006; Hudson et al. 2011). NSMCE4A is significantly differentially expressed in brain and liver. As mentioned previously, NSMCE4A was differentially expressed in the gonad, but not significantly (p-value = 0.00425, q-value = 0.08661). B-specific expression was detected in every tissue with core SNPs. The core blocks overlap the entire gene and also encompass the sequence ~4kb upstream to ~5kb downstream (Figure 5.13). Homologous recombination is essential in meiosis I for proper chromosome segregation. The B chromosome could be altering the frequency of homologous recombination across the genome, or increasing its own ability to do so.

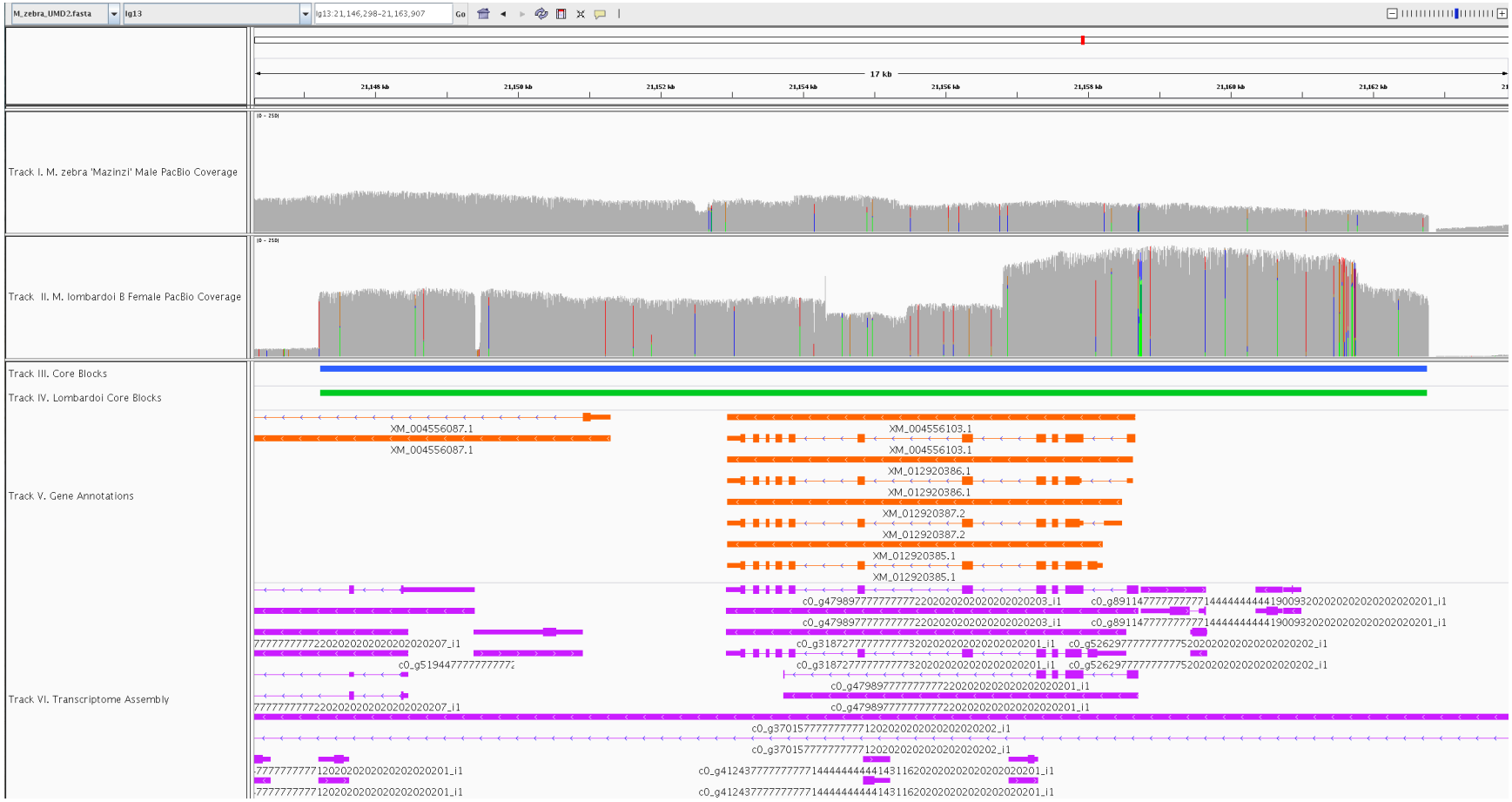

Figure 5.13: NSMCE4A. Region of LG13. Track I is the PacBio read coverage depth of the *M. zebra* 'Mazinzi' male individual used to create the reference genome assembly. Track II shows the PacBio read coverage depth of an *M. lombardoi* B female. Tracks III and IV show the core blocks (blue) and *M. lombardoi* core blocks (green) respectively. Track V is the alignment of the reference assembly gene annotation (orange). Track VI is the alignment of the *de novo* transcriptome assembly.

RTEL1

Regulator of telomere elongation helicase 1 (RTEL1) primarily functions in telomere maintenance, but it also suppresses homologous recombination (Porreca et al. 2018; Barber et al. 2008). RTEL1 is significantly differentially expressed in gonadal tissue. Surprisingly, there is lower expression in B individuals than NoB individuals. The exons of this gene lack B-specific SNPs preventing detection of B-specific expression. The B-located copy of this gene appears to be complete (Figure 5.14), but the A genome seems to have undergone some duplications and structural rearrangements of this gene as well. If there is B-specific expression of this gene, it would be unexpected for B individuals to have, overall, less RTEL1 expression. Perhaps the B-specific transcripts (if there are any) lead to decreased expression of the A-located copy, or increased RNA degradation. But the question remains, to what purpose? Could the decreased expression result in higher rates of homologous recombination? While this gene is being considered as a possible component of drive, it is also interesting to speculate how genes involved in homologous recombination might be useful to the maintenance of B chromosomes in other ways. This particular gene is also found on the B chromosome of Lake Victoria cichlids, where B chromosomes, even when present in multiple copies, do not pair in meiosis.

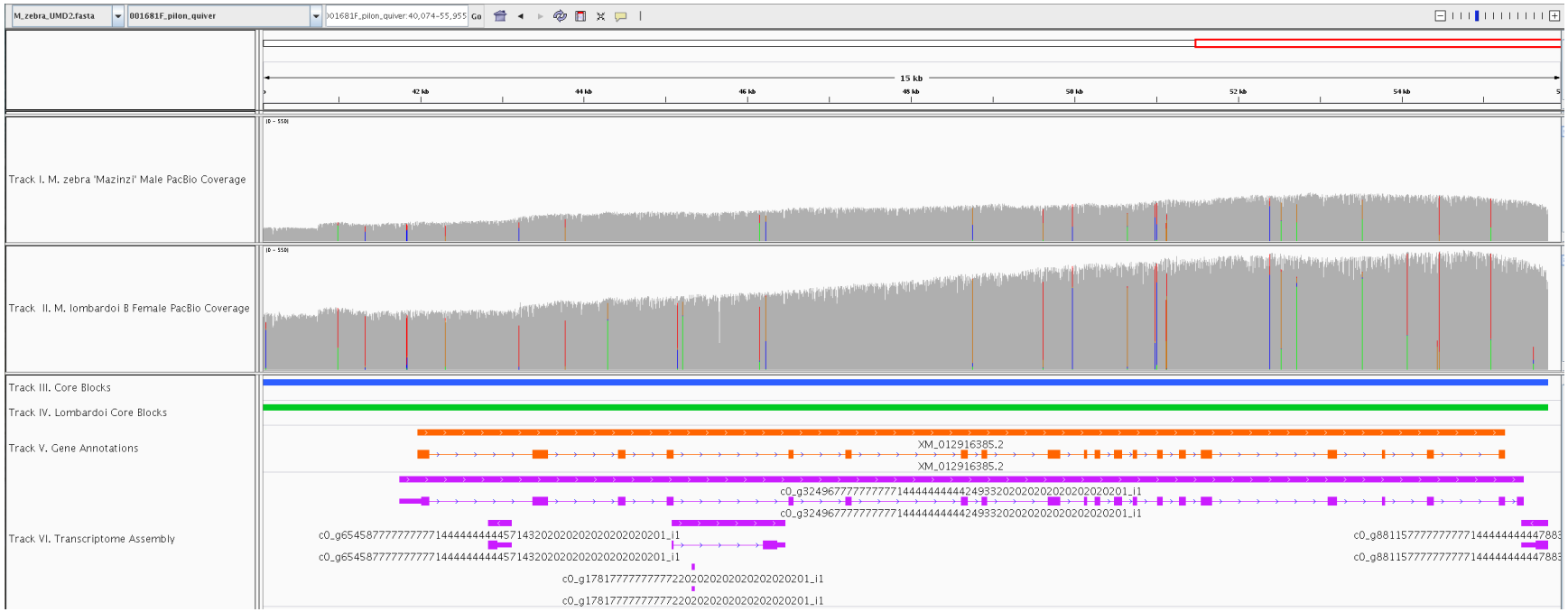

Figure 5.14: RTEL1. Region of the unanchored scaffold, 001681F_pilon_quiver. Track I is the PacBio read coverage depth of the *M. zebra* 'Mazinzi' male individual used to create the reference genome assembly. Track II shows the PacBio read coverage depth of an *M. lombardoi* B female. Tracks III and IV show the core blocks (blue) and *M. lombardoi* core blocks (green) respectively. Track V is the alignment of the reference assembly gene annotation (orange). Track VI is the alignment of the *de novo* transcriptome assembly.

The identification of B chromosome sequence, differential expression analysis and discovery of B-specific transcription was a combination of genome wide approaches meant to sift through the genome to find sequences controlling drive. The detection of 6 candidate genes represents a promising beginning. There is still much research necessary to confirm if any of these genes is involved in drive and if so, their molecular dynamics.

Chapter 6: Summary

By implementing a simple amplification assay of B-specific sequence to examine the presence/absence of B chromosomes, we characterized B chromosomes in seven species of Lake Malawi cichlid fish. Across 7 species, a total of 43 B-carriers were identified among 323 females. B-carriers were exclusively female; no Bs were observed in the 317 males surveyed from these species. Copy number variation (CNV) of the B sequence helped to outline a basic model of B sequence evolution. Quantitative analysis of the copy number variation of B-specific sequence demonstrated that B-carriers possess a single B chromosome, consistent with previous karyotyping of *M. lombardoi*. A single B chromosome in B-carriers is consistent with 2 potential drive mechanisms: one involving nondisjunction and preferential segregation in a mitotic division prior to the germ-line, and the other involving preferential segregation during meiosis I.

Drive was quantified via transmission from a B female to the F1 progeny. The strength of drive varied, possibly reflecting the struggle between B and A chromosomes. Transmission of the Bs in families was also significantly linked with an altered sex ratio. Families segregating a B chromosome had a female-biased sex ratio. Linkage data demonstrated this altered sex ratio is the result of an epistatically dominant sex determiner on the B chromosome. We hypothesize that the sex

determiner evolved in response to the drive mechanism that functions in female meiosis.

We identified B chromosome sequences from several species of cichlid fish from Lake Malawi with a coverage ratio analysis of individuals with and without B chromosomes. We examined the efficiency of this method, and compared results using both Illumina and PacBio sequence data. We found our method identified a significant portion of the B chromosome. The mapping of B blocks to their A chromosome homologs provides further support for the theory that B chromosomes collect sequences from the A genome. A differential expression analysis between B and NoB *M. lombardoi* individuals revealed differences in expression at hundreds to thousands of loci. The identified B chromosome sequence was compared with this differential expression and was found to account for less than 5%, suggesting the differential expression is the result of changes to A chromosome expression in the presence of a B chromosome.

The transcriptome data from B *M. lombardoi* individuals was further scrutinized with B-specific SNPs to detect B-specific and A-specific transcription of sequences found on both the B chromosome and A chromosomes. We found that, using the most conservative method of identifying B sequence, 53 B-located loci were actively transcribed and an additional 65 could still be expressed from the B, but lack any distinguishing sequence variation. Six candidate genes were identified and a preliminary inspection of their sequence and expression was discussed.

B chromosomes of African cichlids represent a promising model for the study of B chromosome evolution, but much work remains to be done characterizing this system. B chromosomes are present in the 7 species presented here, as well as species in Lake Victoria and several other regional water systems. Preliminary research suggests the B chromosomes found among species of Lake Malawi and Lake Victoria have a common ancestor. Yet these chromosomes are characteristically distinct. The Lake Malawi cichlid B is female restricted and always haploid. The Lake Victoria cichlid B is common to males and females and can have 0, 1, 2 or 3 copies per diploid genome. The drive mechanism of the Lake Victoria cichlid B has not been investigated. The B of one Lake Victoria cichlid species, *Lithochromois rubripinnis*, is similar to the Lake Malawi cichlid B in that it may be functioning as a feminizing sex determiner. Future work needs to be done to clarify the evolutionary transitions resulting in these characteristic differences. We hypothesized that the Lake Malawi B chromosome first evolved drive during female meiosis and subsequently acquired a feminizing sex determiner to further increase the success of drive. Could the presence of Lake Victoria cichlid B chromosomes in males and females mean that this B has a different mechanism of drive? Or does it too drive during meiosis I? If so, does *L. rubripinnis* represent parallel evolution in the recruitment of a sex determiner? The presence of 1, 2 or 3 B chromosomes would also be consistent with the hypothesis that the Lake Victoria B chromosome drives during female meiosis but lacks a sex determiner. If it is not restricted to females, the B could be inherited from either parent, or both, increasing the copy number per cell. This presents an unprecedented opportunity to

examine B chromosome evolution with comparative evolution methodologies in very closely related species.

This system will also be a valuable tool for understanding the previously appreciated but poorly understood link between B chromosomes and sex chromosomes. We have demonstrated that the B chromosome of Lake Malawi cichlids is functioning as a W sex chromosome. Comparative studies may even be able to date when this sex determiner arose. In *M. lombardoi* this W (B) is epistatically dominant to the XY system on LG7. But many other sex determiners have been found in African cichlids. How does this B chromosome interact epistatically with those sex chromosomes? Lake Malawi cichlid species will often hybridize in a lab setting, allowing for the construction of genetic crosses spanning genera. Genetic crosses between the established *M. lombardoi* laboratory line and species with any known sex determining system would further our understanding of the interaction of sex chromosomes and the complex sex determination network.

Cytogenetic, genetic and genomic methods have improved our understanding of B chromosomes tremendously. However, there is an important opportunity for cytological approaches now. Female meiosis has not been visualized in these cichlid species. Developing the necessary methodologies to inspect female meiosis would not only confirm (or disprove) the mode of drive suggested here, but would allow these six candidate genes to be scrutinized during meiotic divisions. Characterizing their localization and protein interaction seems the most appropriate next step.

Bibliography

Ahmad, S. and Martins, C. (2019). The Modern View of B Chromosomes Under the Impact of High Scale Omics Analyses. *Cells*, 8(2), p.156.

Alfenito, M. and Birchler, J. (1993). Molecular characterization of a maize B chromosome centric sequence. *Genetics*, 135(2), pp.589-597.

Babraham Bioinformatics FastQC [Internet] Cambridge: Babraham Institute (England). Available from http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/.

Banaei-Moghaddam, A., Schubert, V., Kumke, K., Weiβ, O., Klemme, S., Nagaki, K., Macas, J., González-Sánchez, M., Heredia, V., Gómez-Revilla, D., González-García, M., Vega, J., Puertas, M. and Houben, A. (2012). Nondisjunction in Favor of a Chromosome: The Mechanism of Rye B Chromosome Drive during Pollen Mitosis. *The Plant Cell*, 24(10), pp.4124-4134.

Banaei-Moghaddam, A., Meier, K., Karimi-Ashtiyani, R. and Houben, A. (2013). Formation and Expression of Pseudogenes on the B Chromosome of Rye. *The Plant Cell*, 25(7), pp.2536-2544.

Banaei-Moghaddam, A., Martis, M., Macas, J., Gundlach, H., Himmelbach, A., Altschmied, L., Mayer, K. and Houben, A. (2015). Genes on B chromosomes: Old questions revisited with new tools. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1849(1), pp.64-70.

Barber, L., Youds, J., Ward, J., McIlwraith, M., O'Neil, N., Petalcorin, M., Martin, J., Collis, S., Cantor, S., Auclair, M., Tissenbaum, H., West, S., Rose, A. and Boulton, S. (2008). RTEL1 Maintains Genomic Stability by Suppressing Homologous Recombination. *Cell*, 135(2), pp.261-271.

Bauerly, E., Hughes, S., Vietti, D., Miller, D., McDowell, W. and Hawley, R. (2014). Discovery of Supernumerary B Chromosomes in Drosophila melanogaster. *Genetics*, 196(4), pp.1007-1016.

Beladjal, L., Vandekerckhove, T., Muyssen, B., Heyrman, J., de Caesemaeker, J. and Mertens, J. (2002). B-chromosomes and male-biased sex ratio with paternal inheritance in the fairy shrimp Branchipus schaefferi (Crustacea, Anostraca). *Heredity*, 88(5), pp.356-360.

Benetta, E., Akbari, O. and Ferree, P. (2019). Sequence Expression of Supernumerary B Chromosomes: Function or Fluff?. *Genes*, 10(2), p.123.

Bertollo, L., Takahashi, C., Moreira-Filho, O. (1978). Citotaxonomic consideration on Hoplias lacerdae (Pisces, Erythrinidae). *Braz J Genet*, 1, pp.103-120.

Beukeboom, L. and Werren, J. (1993). Transmission and expression of the parasitic paternal sex ratio (PSR) chromosome. *Heredity*, 70(4), pp.437-443.

Beukeboom, L. (1994). Bewildering Bs: an impression of the 1st B-Chromosome Conference. *Heredity*, 73(3), pp.328-336.

Brawand, D., Wagner, C., Li, Y., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A., Lim, Z., Bezault, E., Turner-Maier, J., Johnson, J., Alcazar, R., Noh, H., Russell, P., Aken, B., Alföldi, J., Amemiya, C., Azzouzi, N., Baroiller, J., Barloy-Hubler, F., Berlin, A., Bloomquist, R., Carleton, K., Conte, M., D'Cotta, H., Eshel, O., Gaffney, L., Galibert, F., Gante, H., Gnerre, S., Greuter, L., Guyon, R., Haddad, N., Haerty, W., Harris, R., Hofmann, H., Hourlier, T., Hulata, G., Jaffe, D., Lara, M., Lee, A., MacCallum, I., Mwaiko, S., Nikaido, M., Nishihara, H., Ozouf-Costaz, C., Penman, D., Przybylski, D., Rakotomanga, M., Renn, S., Ribeiro, F., Ron, M., Salzburger, W., Sanchez-Pulido, L., Santos, M., Searle, S., Sharpe, T., Swofford, R., Tan, F., Williams, L., Young, S., Yin, S., Okada, N., Kocher, T., Miska, E., Lander, E., Venkatesh, B., Fernald, R., Meyer, A., Ponting, C., Streelman, J., Lindblad-Toh, K., Seehausen, O. and Di Palma, F. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, 513(7518), pp.375-381.

Bugrov, A., Karamysheva, T., Perepelov, E., Elisaphenko, E., Rubtsov, D., Warchałowska-Śliwa, E., Tatsuta, H. and Rubtsov, N. (2007). DNA content of the B chromosomes in grasshopper Podisma kanoi Storozh. (Orthoptera, Acrididae). *Chromosome Research*.

Burt, A. and Trivers, R. (2008). *Genes in conflict.* Cambridge, Mass.: Belknap, pp.325-380.

Camacho, J., Shaw, M., Leon, M., Pardo, M. and Cabrero, J. (1997). Erratum: Population Dynamics of a Selfish B Chromosome Neutralized by the Standard Genome in the Grasshopper Eyprepocnemis plorans. *The American Naturalist*, 155(6), p.828.

Camacho, J., Sharbel, T. and Beukeboom, L. (2000). B-chromosome evolution. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 355(1394), pp.163-178.

Camacho, J., Schmid, M. and Cabrero, J. (2011). B Chromosomes and Sex in Animals. *Sexual Development*, 5(3), pp.155-166.

Cano, M. and Santos, J. (1989). Cytological basis of the B chromosome accumulation mechanism in the grasshopper Heteracris littoralis (Ramb). *Heredity*, 62(1), pp.91-95.

Carmena, M., Wheelock, M., Funabiki, H. and Earnshaw, W. (2012). The chromosomal passenger complex (CPC): from easy rider to the godfather of mitosis. *Nature Reviews Molecular Cell Biology*, 13(12), pp.789-803.

Castro, J., Hattori, R., Yoshinaga, T., Silva, D., Foresti, F., Santos, M., Almeida, M. and Artoni, R. (2019). Differential Expression of dmrt1 in Astyanax scabripinnis (Teleostei, Characidade) Is Correlated with B Chromosome Occurrence. *Zebrafish*, 16(2), pp.182-188.

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

Ciossani, G., Overlack, K., Petrovic, A., Huis in 't Veld, P., Koerner, C., Wohlgemuth, S., Maffini, S. and Musacchio, A. (2018). The kinetochore proteins CENP-E and CENP-F directly and specifically interact with distinct BUB mitotic checkpoint Ser/Thr kinases. *Journal of Biological Chemistry*, 293(26), pp.10084-10101.

Clark, F., Conte, M., Ferreira-Bravo, I., Poletto, A., Martins, C. and Kocher, T. (2017). Dynamic Sequence Evolution of a Sex-Associated B Chromosome in Lake Malawi Cichlid Fish. *Journal of Heredity*, 108(1), pp.53-62.

Clark, F., Conte, M. and Kocher, T. (2018). Genomic Characterization of a B Chromosome in Lake Malawi Cichlid Fishes. *Genes*, 9(12), p.610.

Cnaani, A., Zilberman, N., Tinman, S., Hulata, G. and Ron, M. (2004). Genome-scan analysis for quantitative trait loci in an F2 tilapia hybrid. *Molecular Genetics and Genomics*, 272(2).

Cnaani, A., Lee, B., Zilberman, N., Ozouf-Costaz, C., Hulata, G., Ron, M., D'Hont, A., Baroiller, J., D'Cotta, H., Penman, D., Tomasino, E., Coutanceau, J., Pepey, E., Shirak, A. and Kocher, T. (2008). Genetics of Sex Determination in Tilapiine Species. *Sexual Development*, 2(1), pp.43-54.

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

Conte, M. and Kocher, T. (2015). An improved genome reference for the African cichlid, Metriaclima zebra. *BMC Genomics*, 16(1).

Conte, M., Joshi, R., Moore, E., Nandamuri, S., Gammerdinger, W., Roberts, R., Carleton, K., Lien, S. and Kocher, T. (2019). Chromosome-scale assemblies reveal the structural evolution of African cichlid genomes. *GigaScience*, 8(4).

Costa, Y., Speed, R., Ollinger, R., Alsheimer, M., Semple, C., Gautier, P., Maratou, K., Novak, I., Hoog, C., Benavente, R. and Cooke, H. (2005). Two novel proteins recruited by synaptonemal complex protein 1 (SYCP1) are at the centre of meiosis. Journal of Cell Science, 118(12), pp.2755-2762.

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

Erdman, S. and Burtis, K. (1993). The Drosophila doublesex proteins share a novel zinc finger related DNA binding domain. *The EMBO Journal*, 12(2), pp.527-535.

Eshel, O., Shirak, A., Weller, J., Hulata, G. and Ron, M. (2012). Linkage and Physical Mapping of Sex Region on LG23 of Nile Tilapia (Oreochromis niloticus). *G3: Genes|Genomes|Genetics*, 2(1), pp.35-42.

Evans, C., Hardin, J. and Stoebel, D. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, 19(5), pp.776-792.

Fantinatti, B., Mazzuchelli, J., Valente, G., Cabral-de-Mello, D. and Martins, C. (2011). Genomic content and new insights on the origin of the B chromosome of the cichlid fish Astatotilapia latifasciata. *Genetica*, 139(10), pp.1273-1282.

Feldberg, E., Bertollo, L. (1984). Discordance in chromosome number among somatic and gonadal tissue cells of Gymnogeophagus balzanii (Pisces: Cichlidae). *Braz. J. Genet*., 4, pp.639–645.

Feldberg, E., Porto, J., Alves-Brinn, M., Mendonça, M. and Benzaquem, D. (2004). B chromosomes in Amazonian cichlid species. *Cytogenetic and Genome Research*, 106(2-4), pp.195-198.

Feschotte, C. and Pritham, E. (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. *Annual Review of Genetics*, 41(1), pp.331-368.

Fontana, P. and Vickery, V. (1973). Segregation-distortion in the B-chromosome system ofTettigidea lateralis (say) (Orthoptera: Tetrigidae). *Chromosoma*, 43(1), pp.75-98.

Frost, S. (1959). The cytological behavior and mode of transmission of accessory chromosomes in Plantago Serraria. *Hereditas*, 45(2-3), pp.191-210.

Frost, S. (2009). The inheritance of accessory chromosomes in plants, especially in Ranunculus acris and Phleum nodosum. *Hereditas*, 61(3), pp.317-326.

Gammerdinger, W. and Kocher, T. (2018). Unusual Diversity of Sex Chromosomes in African Cichlid Fishes. *Genes*, 9(10), p.480.

Gel, B. and Serra, E. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics*, 33(19), pp.3088-3090.

González-Sánchez, M., Chiavarino, M., Jiménez, G., Manzanero, S., Rosato, M. and Puertas, M. (2004). The parasitic effects of rye B chromosomes might be beneficial in the long term. *Cytogenetic and Genome Research*, 106(2-4), pp.386-393.

Graphodatsky, A., Kukekova, A., Yudkin, D., Trifonov, V., Vorobieva, N., Beklemisheva, V., Perelman, P., Graphodatskaya, D., Trut, L., Yang, F., Ferguson-Smith, M., Acland, G. and Aguirre, G. (2005). The proto-oncogene C-KIT maps to canid B-chromosomes. *Chromosome Research*, 13(2), pp.113-122.

Green, D. (1988). Cytogenetics of the endemic New Zealand frog, Leiopelma hochstetteri: extraordinary supernumerary chromosome variation and a unique sex-chromosome system. *Chromosoma*, 97(1), pp.55-70.

Green, D., Zeyl, C. and Sharbel, T. (1993). The evolution of hypervariable sex and supernumerary (B) chromosomes in the relict New Zealand frog, Leiopelma hochstetteri. *Journal of Evolutionary Biology*, 6(3), pp.417-441.

Gudimchuk, N., Vitre, B., Kim, Y., Kiyatkin, A., Cleveland, D., Ataullakhanov, F. and Grishchuk, E. (2013). Kinetochore kinesin CENP-E is a processive bi-directional tracker of dynamic microtubule tips. *Nature Cell Biology*, 15(9), pp.1079-1088.

Hackstein, J., Hochstenbach, R., Hauschteck-Jungen, E. and Beukeboom, L. (1996). Is the Y chromosome ofDrosophila an evolved supernumerary chromosome?. *BioEssays*, 18(4), pp.317-323.

He, L., Ling, F., Zheng, X., Wang, W., Kuang, R. (2000) The effect of B-chromosome on the reproduction of Drosophila albomicans. *Acta Genet. Sinica,* 27, pp.114–120.

Hewitt, G. (1976). Meiotic drive for B-chromosomes in the primary oocytes of Myrmekotettix maculatus (Orthoptera: Acrididae). *Chromosoma*, 56(4), pp.381-391.

Houben, A., Belyaev, N., Leach, C. and Houben, A. (1997). Differences of histone H4 acetylation and replication timing between A and B chromosomes of Brachycome dichromosomatica. *Chromosome Research*, 5(4), pp.233-237.

Houben, A., Banaei-Moghaddam, A., and Klemme, S. (2013). Biology and evolution of B chromosomes. In Plant Genome Diversity, Vol. 2., J. Greilhuber, J. Dolezel, and J.F. Wendel, eds (Vienna: Springer Press) pp. 149–166.

Houben, A., Banaei-Moghaddam, A., Klemme, S. and Timmis, J. (2013). Evolution and biology of supernumerary B chromosomes. *Cellular and Molecular Life Sciences*, 71(3), pp.467-478.

Houben, A. (2017). B Chromosomes – A Matter of Chromosome Drive. *Frontiers in*

Plant Science, 08.

Huang, W., Du, Y., Zhao, X. and Jin, W. (2016). B chromosome contains active genes and impacts the transcription of A chromosomes in maize (Zea mays L.). *BMC Plant Biology*, 16(1).

Hudson, J., Bednarova, K., Kozakova, L., Liao, C., Guerineau, M., Colnaghi, R., Vidot, S., Marek, J., Bathula, S., Lehmann, A. and Palecek, J. (2011). Interactions between the Nse3 and Nse4 Components of the SMC5-6 Complex Identify Evolutionarily Conserved Interactions between MAGE and EID Families. *PLoS ONE*, 6(2), p.e17270.

IBM Corp. Released 2013. IBM SPSS Statistics for Macintosh, Version 22.0. Armonk, NY: IBM Corp.

Jackson, W. and Cheung, D. (1967). Distortional meiotic segregation of a supernumerary chromosome producing differential frequencies in the sexes in the short-horned grasshopper Phaulacridium vittatum. *Chromosoma*, 23(1), pp.24-37.

Jacobs, F., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A., Katzman, S., Paten, B., Salama, S. and Haussler, D. (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature*, 516(7530), pp.242-245.

Jones, R. (1975). B-chromosome systems in flowering plants and animal species. *Int. Rev. Cytol*., 40, pp. 1-100

Jones, R. and Rees, H. (1982). *B Chromosomes*. London: Academic Press.

Jones, R. (1991). B-Chromosome Drive. *The American Naturalist*, 137(3), pp.430-442.

Jones, N. and Houben, A. (2003). B chromosomes in plants: escapees from the A chromosome genome?. *Trends in Plant Science*, 8, pp.1360-1385.

Jones, R., González-Sánchez, M., González-García, M., Vega, J. and Puertas, M. (2008). Chromosomes with a life of their own. *Cytogenetic and Genome Research*, 120(3-4), pp.265-280.

Jones, R., Viegas, W. and Houben, A. (2008). A Century of B Chromosomes in Plants: So What?. *Annals of Botany*, 101(6), pp.767-775.

Jones, N. (2017). New species with B chromosomes discovered since 1980. *The Nucleus*, 60(3), pp.263-281.

Kayano, H. (1957). Cytogenetic Studies in Lilium callosum. *Proceedings of the Japan Academy*, 33(9), pp.553-558.

Kayano, H. (1971). Accumulation of B chromosomes in the germ line of Locusta migratoria. *Heredity*, 27(1), pp.119-123.

Kimura, M., Kayano, H. (1961).The maintenance of supernumerary chromosomes in wild populations of *Lilium callosum* by preferential segregation. *Genetics*, *46*, pp.1699–1712.

Klemme, S., Banaei-Moghaddam, A., Macas, J., Wicker, T., Novák, P. and Houben, A. (2013). High-copy sequences reveal distinct evolution of the rye B chromosome. *New Phytologist*, 199(2), pp.550-558.

Koboldt, D., Zhang, Q., Larson, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, E., Ding, L. and Wilson, R. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3), pp.568-576.

Kocher, T. (2004). Adaptive evolution and explosive speciation: the cichlid fish model. *Nature Reviews Genetics*, 5(4), pp.288-298.

Kondo, M., Nanda, I., Hornung, U., Asakawa, S., Shimizu, N., Mitani, H., Schmid, M., Shima, A. and Schartl, M. (2003). Absence of the Candidate Male Sex-Determining Gene dmrt1b(Y) of Medaka from Other Fish Species. *Current Biology*, 13(5), pp.416-420.

Krenn, V. and Musacchio, A. (2015). The Aurora B Kinase in Chromosome Bi-Orientation and Spindle Checkpoint Signaling. *Frontiers in Oncology*, 5.

Kuroiwa, A., Terai, Y., Kobayashi, N., Yoshida, K., Suzuki, M., Nakanishi, A., Matsuda, Y., Watanabe, M. and Okada, N. (2013). Construction of Chromosome Markers from the Lake Victoria Cichlid Paralabidochromis chilotes and Their Application to Comparative Mapping. *Cytogenetic and Genome Research*, 142(2), pp.112-120.

Langmead, B. and Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), pp.357-359.

Leach, C., Houben, A., Field, B., Pistrick, K., Demidov, D. and Timmis, J. (2005). Molecular Evidence for Transcription of Genes on a B Chromosome in Crepis capillaris. *Genetics*, 171(1), pp.269-278.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), pp.1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,

Abecasis, G. and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078-2079.

Lin, H., Lin, W., Lin, C., Peng, S. and Cheng, Y. (2014). Characterization of maize B-chromosome-related transcripts isolated via cDNA-AFLP. *Chromosoma*, 123(6), pp.597-607.

López-León, M., Neves, N., Schwarzacher, T., (Pat) Heslop-Harrison, J., Hewitt, G. and Camacho, J. (1994). Possible origin of a B chromosome deduced from its DNA composition using double FISH technique. *Chromosome Research*, 2(2), pp.87-92.

López-León, M., Cabrero, J. and Camacho, J. (1996). Achiasmate segregation of X and B univalents in males of the grasshopperEyprepocnemis plorans is independent of previous association. *Chromosome Research*, 4(1), pp.43-48.

Ma, W., Gabriel, T., Martis, M., Gursinsky, T., Schubert, V., Vrána, J., Doležel, J., Grundlach, H., Altschmied, L., Scholz, U., Himmelbach, A., Behrens, S., Banaei-Moghaddam, A. and Houben, A. (2016). Rye B chromosomes encode a functional Argonaute-like protein within vitroslicer activities similar to its A chromosome paralog. *New Phytologist*, 213(2), pp.916-928.

Makunin, A., Dementyeva, P., Graphodatsky, A., Volobouev, V., Kukekova, A. and Trifonov, V. (2014). Genes on B chromosomes of vertebrates. *Molecular Cytogenetics*, 7(1).

Makunin, A., Kichigin, I., Larkin, D., O'Brien, P., Ferguson-Smith, M., Yang, F., Proskuryakova, A., Vorobieva, N., Chernyaeva, E., O'Brien, S., Graphodatsky, A. and Trifonov, V. (2016). Contrasting origin of B chromosomes in two cervids (Siberian roe deer and grey brocket deer) unravelled by chromosome-specific DNA sequencing. *BMC Genomics*, 17(1).

Makunin, A., Rajičić, M., Karamysheva, T., Romanenko, S., Druzhkova, A., Blagojević, J., Vujošević, M., Rubtsov, N., Graphodatsky, A. and Trifonov, V. (2018). Low-pass single-chromosome sequencing of human small supernumerary marker chromosomes (sSMCs) and Apodemus B chromosomes. *Chromosoma*, 127(3), pp.301-311.

Martis, M., Klemme, S., Banaei-Moghaddam, A., Blattner, F., 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, K. and Houben, A. (2012). Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proceedings of the National Academy of Sciences*, 109(33), pp.13343-13346.

Mims, M., Hulsey, D., Fitzpatrick, B., Streelman, T. (2010). Geography disentangles introgression from ancestral polymorphism in Lake Malawi cichlids. *Molecular*

Ecology, 19(5), pp.940-951.

Murray, B. (1984). The structure, meiotic behaviour and effects of B chromosomes in Briza humilis Bieb. (Gramineae). *Genetica*, 63(3), pp.213-219.

Navarro-Domínguez, B., Ruiz-Ruano, F., Cabrero, J., Corral, J., López-León, M., Sharbel, T. and Camacho, J. (2017). Protein-coding genes in B chromosomes of the grasshopper Eyprepocnemis plorans. *Scientific Reports*, 7(1).

Navarro-Domínguez, B., Ruiz-Ruano, F., Camacho, J., Cabrero, J. and López-León, M. (2017). Transcription of a B chromosome CAP-G pseudogene does not influence normal Condensin Complex genes in a grasshopper. *Scientific Reports*, 7(1).

Navarro-Domínguez, B., Martín-Peciña, M., Ruiz-Ruano, F., Cabrero, J., Corral, J., López-León, M., Sharbel, T. and Camacho, J. (2019). Gene expression changes elicited by a parasitic B chromosome in the grasshopper Eyprepocnemis plorans are consistent with its phenotypic effects. *Chromosoma*, 128(1), pp.53-67.

Néo, D., Filho, O. and Camacho, J. (2000). Altitudinal variation for B chromosome frequency in the characid fish Astyanax scabripinnis. *Heredity*, 85(2), pp.136-141.

Nokkala, S. and Grozeva, S. (2000). Achiasmatic male meiosis in Myrmedobia coleoptrata(Fn.) (Heteroptera, Microphysidae). *Caryologia*, 53(1), pp.5-8.

Nokkala, S., Kuznetsova, V. and Maryańska-Nadachowska, A. (2000). Achiasmate Segregation of a B Chromosome from the X Chromosome in Two Species of Psyllids (Psylloidea, Homoptera). *Genetica*, 108(2), pp.181-189.

Nur U. (1963). A Mitotically Unstable Supernumerary Chromosome with an Accumulation Mechanism in a Grasshopper. *Chromosoma (Berl.)*, 14, pp.407-422.

Nur, U. (1966). Harmful B chromosomes in a mealy bug population. *Genetics*, 54, pp.1225–1238.

Nur U. (1969). Mitotic Instability Leading to an Accumulation of B-Chromosomes in Grasshoppers. *Chromosoma (Berl.)*, 27, pp.1-19.

Nur, U. (1977). Maintenance of a “parasitic” B chromosome in the grasshopper *Melanoplus femur-rubrum*. *Genetics, 87*, pp.499–512.

Nur, U., Brett, B. (1985). Genotypes suppressing meiotic drive in a B chromosome in the mealy bug, *Pseudococcus* [illegible] *obscurus*. *Genetics, 110*, pp.73–92.

Otake, H., Shinomiya, A., Kawaguchi, A., Hamaguchi, S. and Sakaizumi, M. (2008). The medaka sex-determining geneDMYacquired a novel temporal expression pattern

after duplication ofDMRT1. *genesis*, 46(12), pp.719-723.

Palaiokostas, C., Bekaert, M., Khan, M., Taggart, J., Gharbi, K., McAndrew, B. and Penman, D. (2013). Mapping and Validation of the Major Sex-Determining Region in Nile Tilapia (Oreochromis niloticus L.) Using RAD Sequencing. *PLoS ONE*, 8(7), p.e68389.

Palecek, J., Vidot, S., Feng, M., Doherty, A. and Lehmann, A. (2006). The Smc5-Smc6 DNA Repair Complex. *Journal of Biological Chemistry*, 281(48), pp.36952-36959.

Palestis, B., Trivers, R., Burt, A. and Jones, R. (2004). The distribution of B chromosomes across species. *Cytogenetic and Genome Research*, 106(2-4), pp.151-158.

Pardo, M., López-León, M., Cabrero, J. and Camacho, J. (1994). Transmission analysis of mitotically unstable B chromosomes in Locusta migratoria. *Genome*, 37(6), pp.1027-1034.

Parker, J., Taylor, S. and Ainsworth, C. (1982). The B-chromosome system of Hypochoeris maculata. *Chromosoma*, 85(2), pp.299-310.

Patro, R., Mount, S. and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, 32(5), pp.462-464.

Pires, L., Sampaio, T. and Dias, A. (2015). Mitotic and Meiotic Behavior of B Chromosomes in Crenicichla lepidota: New Report in the Family Cichlidae. *Journal of Heredity*, 106(3), pp.289-295.

Poletto, A., Ferreira, I. and Martins, C. (2010). The B chromosomes of the African cichlid fish Haplochromis obliquidens harbour 18S rRNA gene copies. *BMC Genetics*, 11(1), p.1.

Poletto, A., Ferreira, I., Cabral-de-Mello, D., Nakajima, R., Mazzuchelli, J., Ribeiro, H., Venere, P., Nirchio, M., Kocher, T. and Martins, C. (2010). Chromosome differentiation patterns during cichlid fish evolution. *BMC Genetics*, 11(1), p.50.

Porreca, R., Glousker, G., Awad, A., Matilla Fernandez, M., Gibaud, A., Naucke, C., Cohen, S., Bryan, T., Tzfati, Y., Draskovic, I. and Londoño-Vallejo, A. (2018). Human RTEL1 stabilizes long G-overhangs allowing telomerase-dependent over-extension. *Nucleic Acids Research*, 46(9), pp.4533-4545.

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

Ramos, É., Cardoso, A., Brown, J., Marques, D., Fantinatti, B., Cabral-de-Mello, D., Oliveira, R., O'Neill, R. and Martins, C. (2016). The repetitive DNA element BncDNA, enriched in the B chromosome of the cichlid fish Astatotilapia latifasciata, transcribes a potentially noncoding RNA. *Chromosoma*, 126(2), pp.313-323.

Randolph L. (1941) Genetic characteristics of the B chromosome in maize. *Genetics,* 26, pp.608-631.

Rhoades, M. (1968) Studies on the cytological basis of crossing over. pp. 229-241. In: Repli- cation and Recombination of Genetic Material. Edited by W. J. PEACOCK and R. D. BROCK. Austrialian Academy of Science, Canberra.

Roberts, R., Ser, J. and Kocher, T. (2009). Sexual Conflict Resolved by Invasion of a Novel Sex Determiner in Lake Malawi Cichlid Fishes. *Science*, 326(5955), pp.998-1001.

Robinson, M. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3), p.R25.

Ruban, A., Schmutzer, T., Scholz, U. and Houben, A. (2017). How Next-Generation Sequencing Has Aided Our Understanding of the Sequence Composition and Origin of B Chromosomes. *Genes*, 8(11), p.294.

Ruiz-Estévez, M., Cabrero, J. and Camacho, J. (2012). B-Chromosome Ribosomal DNA Is Functional in the Grasshopper Eyprepocnemis plorans. *PLoS ONE*, 7(5), p.e36600.

Ruiz-Estévez, M., Badisco, L., Broeck, J., Perfectti, F., López-León, M., Cabrero, J. and Camacho, J. (2014). B chromosomes showing active ribosomal RNA genes contribute insignificant amounts of rRNA in the grasshopper Eyprepocnemis plorans. *Molecular Genetics and Genomics*, 289(6), pp.1209-1216.

Ruiz-Ruano, F., Cabrero, J., López-León, M., Sánchez, A. and Camacho, J. (2017). Quantitative sequence characterization for repetitive DNA content in the supernumerary chromosome of the migratory locust. *Chromosoma*, 127(1), pp.45-57.

Rutishauser, A. (1956). Genetics of fragment chromosomes in Trillium grandiflorum. *Heredity*, 10(2), pp.195-204.

Santos, J., Cerro, A., Fernández, A. and Díez, M. (1993). Meiotic Behaviour of B Chromosomes in the Grasshopper Omocestus burri: A Case of Drive in Females. *Hereditas*, 118(2), pp.139-143.

Sedlazeck, F., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A. and Schatz, M. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 15(6), pp.461-468.

Ser, J., Roberts, R. and Kocher, T. (2010). MULTIPLE INTERACTING LOCI CONTROL SEX DETERMINATION IN LAKE MALAWI CICHLID FISH. *Evolution*, 64(2), pp.486-501.

Sharbel, T., Green, D. and Houben, A. (1998). B-chromosome origin in the endemic New Zealand frog Leiopelma hochstetteri through sex chromosome devolution. *Genome*, 41(1), pp.14-22.

Silva, D., Pansonato-Alves, J., Utsunomia, R., Araya-Jaime, C., Ruiz-Ruano, F., Daniel, S., Hashimoto, D., Oliveira, C., Camacho, J., Porto-Foresti, F. and Foresti, F. (2014). Delimiting the Origin of a B Chromosome by FISH Mapping, Chromosome Painting and DNA Sequence Analysis in Astyanax paranae (Teleostei, Characiformes). *PLoS ONE*, 9(4), p.e94896.

Sironi, L. (2002). Crystal structure of the tetrameric Mad1-Mad2 core complex: implications of a 'safety belt' binding mechanism for the spindle checkpoint. *The EMBO Journal*, 21(10), pp.2496-2506.

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1).

Stauffer, J., Bowers, N., Kellogg, K., McKaye, K. (1997) A revision of the blue-black Pseudotropheus zebra (Teleostei: Cichlidae) complex from Lake Malawi, Africa, with a description of a new genus and ten new species. *Proceedings of the Academy of Natural Sciences of Philadelphia,* 148, pp.189-230.

Streelman, J., Albertson, R. and Kocher, T. (2003). Genome mapping of the orange blotch colour pattern in cichlid fishes. *Molecular Ecology*, 12(9), pp.2465-2471.

Sturmbauer, C., Baric, S., Salzburger, W., Rüber, L. and Verheyen, E. (2001). Lake Level Fluctuations Synchronize Genetic Divergences of Cichlid Fishes in African Lakes. *Molecular Biology and Evolution*, 18(2), pp.144-154.

Tanić, N., Vujošević, M., Dedović-Tanić, N. and Dimitrijević, B. (2005). Differential gene expression in yellow-necked mice Apodemus flavicollis (Rodentia, Mammalia) with and without B chromosomes. *Chromosoma*, 113(8), pp.418-427.

Thomson, R., Westerman, M. and Murray, N. (1984). B chromosomes in Rattus fuscipes I. Mitotic and meiotic chromosomes and the effects of B chromosomes on chiasma frequency. *Heredity*, 52(3), pp.355-362.

Treangen, T. and Salzberg, S. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1), pp.36-46.

Trifonov, V., Dementyeva, P., Larkin, D., O'Brien, P., Perelman, P., Yang, F., Ferguson-Smith, M. and Graphodatsky, A. (2013). Transcription of a protein-coding gene on B chromosomes of the Siberian roe deer (Capreolus pygargus). *BMC Biology*, 11(1).

Turner, G., Seehausen, O., Knight, M., Allender, C. and Robinson, R. (2008). How many species of cichlid fishes are there in African lakes?. *Molecular Ecology*, 10(3), pp.793-806.

Uhl, C., Moran, R. (1973). The chromosomes of Pachyphytum (Crassulaceae). *Am. J. Bot.* 60, pp.648–656

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B., Remm, M. and Rozen, S. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Research*, 40(15), pp.e115-e115.

Valente, G., Conte, M., Fantinatti, B., Cabral-de-Mello, D., Carvalho, R., Vicari, M., Kocher, T. and Martins, C. (2014). Origin and Evolution of B Chromosomes in the Cichlid Fish Astatotilapia latifasciata Based on Integrated Genomic Analyses. *Molecular Biology and Evolution*, 31(8), pp.2061-2072.

van Doorn, G. and Kirkpatrick, M. (2007). Turnover of sex chromosomes induced by sexual conflict. *Nature*, 449(7164), pp.909-912.

Veneziano, L., Barra, V., Cilluffo, D. and Di Leonardo, A. (2018). Proliferation of aneuploid cells induced by CENP-E depletion is counteracted by the p14ARF tumor suppressor. *Molecular Genetics and Genomics*, 294(1), pp.149-158.

Vicente, V., Moreira-Filho, O., Camacho, J. (1996). Sex-ratio distortion associated with the presence of a B chromosome in Astyanax scabripinnis (Teleostei Characidae). *Cytogenet Cell Genet.,* 74, pp.70–75.

Vries, L., Behar, D., Smirin-Yosef, P., Lagovsky, I., Tzur, S. and Basel-Vanagaite, L. (2014). Exome Sequencing Reveals SYCE1 Mutation Associated With Autosomal Recessive Primary Ovarian Insufficiency. *The Journal of Clinical Endocrinology & Metabolism*, 99(10), pp.E2129-E2132.

Wilson, E. B. (1907) The supernumerary chromosomes of Hemiptera. *Science*, 26,pp.870-871.

Wood, K., Chua, P., Sutton, D. and Jackson, J. (2008). Centromere-Associated Protein E: A Motor That Puts the Brakes on the Mitotic Checkpoint. Clinical Cancer Research, 14(23), pp.7588-7592.

Yoshida, K., Terai, Y., Mizoiri, S., Aibara, M., Nishihara, H., Watanabe, M., Kuroiwa, A., Hirai, H., Hirai, Y., Matsuda, Y. and Okada, N. (2011). B Chromosomes Have a

Functional Effect on Female Sex Determination in Lake Victoria Cichlid Fishes. *PLoS Genetics*, 7(8), p.e1002203.

Zhou, Q., Zhu, H., Huang, Q., Xuan, Z., Zhang, G., Zhao, L., Ding, Y., Roy, S., Vicoso, B., Ruan, J., Zhang, Y., Zhao, R., Mu, B., Min, J., Zhang, Q., Li, J., Luo, Y., Liang, Z., Ye, C., Li, R., Zhang, X., Wang, J., Wang, W. and Bachtrog, D. (2012). Deciphering neo-sex and B chromosome evolution by the draft genome of Drosophila albomicans. *BMC Genomics*, 13(1), p.109.