Article

Genomic Characterization of a B Chromosome in Lake Malawi Cichlid Fishes

Frances E. Clark [1], Matthew A. Conte [1] and Thomas D. Kocher [1,*]

[1] Department of Biology, University of Maryland, College Park, Maryland 20742
* Correspondence: tdk@umd.edu

Abstract: B chromosomes (Bs) were discovered a century ago, and since then most studies have focused on describing their distribution and abundance using traditional cytogenetics. Only recently have attempts been made to understand their structure and evolution at the level of DNA sequence. Many questions regarding the origin, structure, function and evolution of B chromosomes remain unanswered. Here we identify B chromosome sequences from several species of cichlid fish from Lake Malawi by examining the ratios of DNA sequence coverage in individuals with and without B chromosomes. We examine the efficiency of this method, and compare results using both Illumina and PacBio sequence data. The B chromosome sequences detected in 13 individuals from 7 species were compared to assess the rates of sequence replacement. B-specific sequence common to at least 12 of the 13 datasets are identified as the "Core" B chromosome. The location of B sequence homologs throughout the genome provides further support for theories of B chromosome evolution. Finally, we identified candidate genes located on the B chromosome which may regulate the segregation and maintenance of the B chromosome.

Keywords: supernumerary chromosomes; B chromosomes; next-generation sequencing, coverage ratio analysis

1. Introduction

The genomes of eukaryotic species are typically organized into linear chromosomes, and each species has a characteristic number of chromosomes pairs referred to as the A chromosomes (As). The genomes of 10-20% of eukaryotic species contain additional chromosomes commonly referred to as B chromosomes (Bs). These supernumerary B chromosomes are not essential, and are found in some, but not all individuals of a population [1,2,3,4]. Among species, the number of B chromosomes in each cell has been found to vary from 1-32 [5,6]. B chromosomes are thought to manipulate the normal mechanisms of cell division in order to increase their transmission to the next generation, a process known as drive [3,6,7].

B chromosomes often contain large amounts of highly repetitive DNA [4,8,9,10], and are frequently either partially or completely heterochromatic [2,4,6]. In several species it has been shown that B chromosomes share homology with sequences from all or many of the A chromosomes [11], (the grasshopper *Podisma kanoi* [9], the fish *Astatotilapia latifasciata* [12], rye *Secale cereale* [13] and maize *Zea mays* [8]). This suggests that sequences on B chromosomes are derived from the A chromosomes through as yet uncharacterized mechanisms of gene duplication [14]. Theoretically, because they are non-essential, B chromosomes should experience relaxed selective pressures [14,15]. 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 [16,17,18,19].

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 [2,6,20,21,22]. 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 [14,23]. They have been thought to be composed of non-functional "junk" DNA together with one or two genes contributing to drive [9].

Recent advances in next-generation sequencing and bioinformatic analyses of genomic data have begun to contradict this long-standing view. These technological and analytical improvements make it possible to address many questions about B chromosome biology, including how Bs acquire sequence from the As, how these sequences evolve once on the B chromosome, whether and to what extent the B contains functional sequence, and finally the identification of the gene(s) controlling drive. The C-KIT gene in two *Canidae* species [24], rRNA genes and thousands of genes and gene fragments in the fish *Astatotilapia latifasciata* [12,25], and protein coding genes in the grasshopper *Eyprepocnemis plorans* [26] are examples of genic sequences detected on B chromosomes. Furthermore, transcription has been characterized for an rRNA gene in the smooth hawksbeard *Crepis capillaris* [27], rRNA genes and a pseudogene in the grasshopper *E. plorans* [28,29,30], pseudogenes in rye *Secale cereale* [31], and protein coding genes in maize *Zea mays* [32,33].

Current approaches to identifying B sequence can be categorized into two types: direct and indirect [16]. Direct methods, such as the sequencing of B chromosomes isolated through flow sorting or microdissection, have a high rate of contamination [12,16] and are only possible in a few organisms with large B chromosomes. Indirect methods, such as the comparison of whole genome sequence data between samples with and 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 B sequence by its alignment to homologous portions of the A genome. While Illumina sequencing has dramatically lowered costs, there are considerable limitations to Illumina sequence data [34]. 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.

Among cichlid fishes, B chromosomes were first identified in species from South America [35,36]. More recently they have been identified also in species from lakes Victoria and Malawi in East Africa [37,38]. B chromosomes have been found in at least 7 species from Lake Malawi. In all 7 species, B chromosomes are found only in females, but not all females have a B chromosome. The females that do possess a B chromosome have only a single B (haploid) per cell [39]. A karyotype of one of these species, *Metriaclima lombardoi*, shows the B chromosome is one of the largest chromosomes, representing approximately 4.5% of the genome when present. In Lake Victoria cichlids, B chromosomes are found in both males and females, and individuals carry as many as 3 B chromosomes per cell [37,40]. Whole-genome resequencing and the sequencing of a microdissected B chromosome was used to identify B chromosome sequences in the Lake Victoria species *Astatotilapia latifaciata* [12]. Mapping of whole genome sequencing reads to a reference assembly 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. Presumably, the sequences contributing to drive are among the genes and gene fragments identified in this study.

Here we perform a sequence coverage ratio analysis of multiple individuals and species from Lake Malawi to systematically detect sequence from the B chromosome. From this, we identify genes and gene fragments on the B chromosome. We characterized the location and length of the homologous A-located sequences to understand the origin and dynamics of sequence accumulation on the B chromosome. Finally, we analyzed the proportion of B chromosome sequences that are shared among species to estimate the rate of sequence turnover on these unique chromosomes.

2. Materials and Methods

All procedures involving live animals were approved by the University of Maryland IACUC and conducted in accordance with protocol #R-13-58. The *Metriaclima lombardoi* individuals used were collected from stocks maintained at the Tropical Aquaculture Facility at the University of Maryland. These stocks were originally sourced from Lake Malawi, Africa in 2014, 2015 and 2016. The remaining individuals were collected directly from Lake Malawi in 2005, 2008, and 2012. Individuals were euthanized using tricaine methanesulfonate (MS-222) and inspected for testes or ovaries to confirm sex. Standard phenol chloroform methods were used in conjunction with phase-lock gel tubes (5Prime, Gaithersburg, Maryland) for DNA extraction from fin tissue. Genotyping of B-specific SNPs was performed according to [39] to identify individuals carrying a B chromosome. Blood was collected from a *M. lombardoi* individual in order to prepare the high molecular weight DNA necessary for Pacific Biosciences SMRT (PacBio) sequencing.

Illumina sequencing was performed from the fin clips of 12 female individuals with a B chromosome. To provide a comparison lacking B chromosome DNA, sequence from pooled male individuals, previously collected and sequenced with Illumina, was used. Each pooled sample contained between 10-20 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 *M. zebra* 'Boadzulu' males for the scaled coverage analysis. Sequence from 2 samples of pooled NoB female individuals, previously collected and sequenced with Illumina, were used as controls. These 2 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 1.

Table 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

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., San Diego, CA). 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. 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. 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 9 SMRT cells and on the PacBio Sequel platform using 9 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 [41], (publicly available on NCBI, Accession: GCA_000238955.4, [42]) with BWA [43] and NGM-LR [44], respectively. BWA alignments were then run through Picard version 2.1.0 'MarkDuplicates' to identify PCR duplicates.

After alignment to the reference genome, all genomic samples were analyzed with samtools version 0.1.18 mpileup 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.

Equation 1:

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

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.

Equation 2:

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

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, a binomial test p-value ≤ 0.001 and within 300 bp of another such position were merged into a block feature with Bedtools version 2.26.0 merge function [45]. 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 remaining block features are referred to as "B blocks". The B blocks of all individuals were then processed with Bedtools Intersect 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 Malawi "core" blocks.

The sum of the lengths of all B blocks was calculated as an estimate of total B sequence length in 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, which 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 data set. For NoB male scaled coverage ≥1, we used Equation 3:

$$(SCR * 2)\text{-}2$$

In Equation 3, 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 Equation 4:

$$Female\ Scaled\ Coverage * 2$$

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. Example scripts for our B block identification analysis are provided in Appendix A (Directory A1).

3. Results

3.1. Identification of B Chromosome Sequence

3.1.1. Characterization of B Blocks

Due to the homology between A and B chromosome sequence, most sequence reads derived from the B chromosome will align to their A chromosome homologs present in the reference genome. As a result, alignments of reads from a genome with a B chromosome will have regions of increased coverage compared to an alignment from a genome lacking a B. Our 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 2). In comparison, the same analysis in our 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 are 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 1). The largest such regions are located on LG4 (~120 kb), LG9 (~250 kb), LG17 (~260 kb) and LG23 (~420 kb).

Table 2. B Block Sizes

	% of A genome passing both thresholds	Number of blocks	Mean block size (bp)	Standard deviation of block size (bp)	Maximum block size (kb)
L. trewavasae 2005-1306	0.59	3517	1554.8	2592.9	42.9
M. auratus 2008-1601	0.34	2476	1415.7	1859.6	30.2
M. greshakei 2012-3493	0.69	4392	1395.1	2618.8	52.8
M. lombardoi 2014-1018	1.31	10918	1285.1	1845.8	63.2
M. lombardoi 2014-1021	1.04	8251	1298.0	1954.9	63.3
M. lombardoi 2014-1108	1.10	8684	1274.9	1902.3	63.2
M. zebra mbenji 2012-3997	0.68	4147	1344.7	2519.2	63.2
M. zebra 'Boadzulu' 2005-0976	0.85	5907	1264.9	2002.4	42.9
M. zebra 'Boadzulu' 2005-0983	0.79	5369	1238.5	2402.4	100.6
M. zebra 'Boadzulu' 2005-0986	0.84	5986	1228.8	2293.8	100.1
M. zebra 'Nkhata Bay' 2012-5340	0.64	4869	1419.0	2856.6	99.9
M. zebra 'Nkhata Bay' 2012-5347	0.89	7162	1420.9	2821.8	100.0
M. lombardoi 2016-1012 (PacBio)	0.59	1904	2971.1	4569.8	98.6
L. trewavasae Maison XX females (control)	0.44	2125	714.0	243.4	3.6
L. trewavasae Maison WZ females (control)	0.06	343	819.5	478.5	5.2
Core blocks	N/A	622	2194.6	3582.8	32.7

Figure 1. Read Coverage and B Blocks

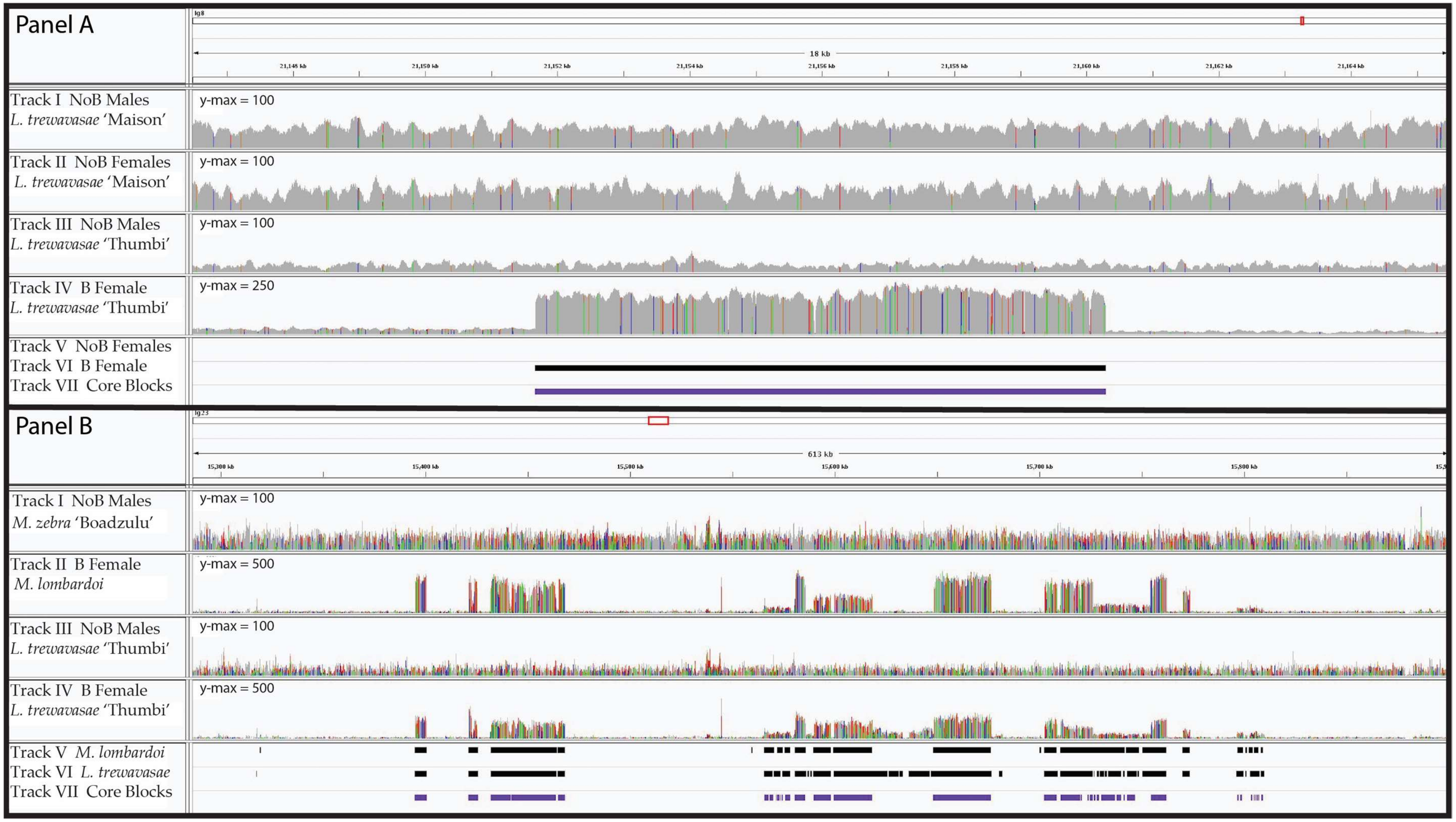


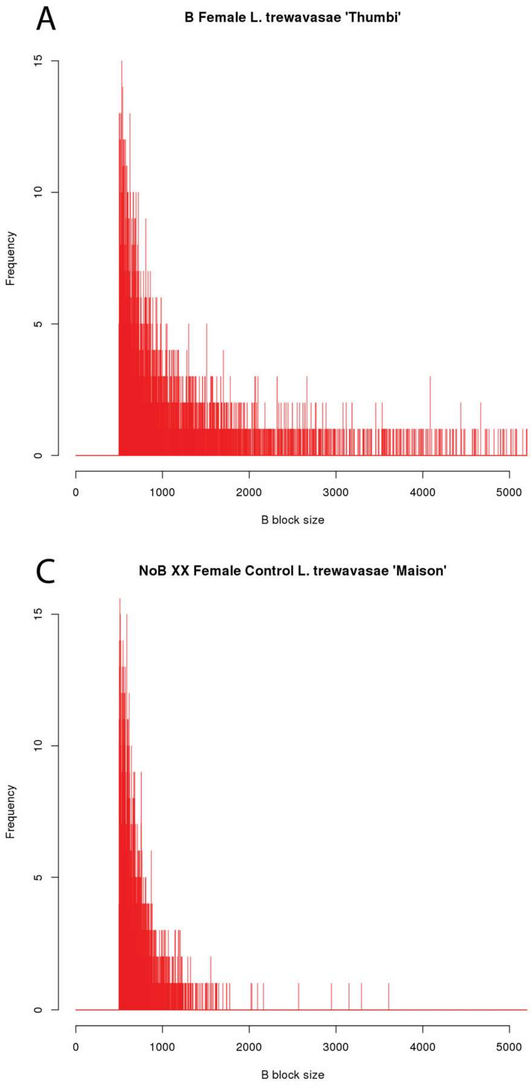
Figure 1 Caption: B blocks from two genomic regions are shown with the corresponding read coverage.

Panel A depicts an 18kb region of LG8 with a typical B block. Tracks I and III are the male coverage (*L. 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 ~8.5kb 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 613kb 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 *M. 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 (tracks 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 3.6-5.2kb (Table 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 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 1 are often fragmented into smaller block calls, we opted not to remove the shorter block calls at this stage of the analysis. The B block information for each data set, including block location, coverage details and length, is provided in Appendix A (Directory A2).

Figure 2. Block Length Histograms



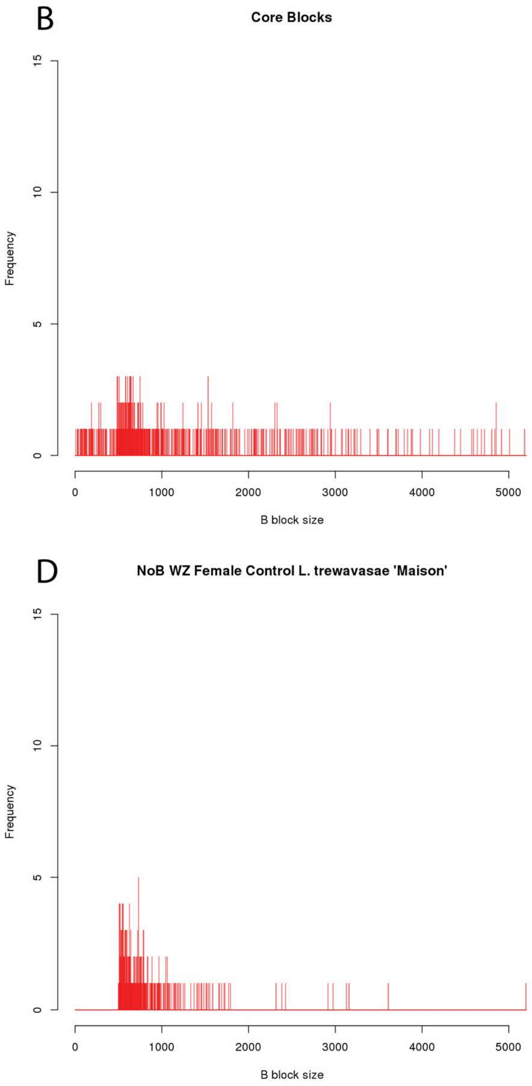


Figure 2 Caption: Histograms of B block length for four data sets. B block size along the x-axis is reported in bp, and only blocks 5000 bp or smaller were included in the figure to more easily view the majority of the data. Because blocks shorter than 500 bp were removed during analysis, the B female (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). All other B block length histograms are included in the Appendix as Figure A1.

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 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.

The total length in A space ranges from 3.51-14.06 Mb among the B females and only 0.28-1.52 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 23.19-99.69 Mb among B females and only 0.39-2.15 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 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 set), 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 3 individuals may be inflated by false positive blocks, while the total length of core sequence is B space is slightly 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. [40]. Notably, *M. auratus* consistently has the least amount of sequence detected by our analysis. The 12.31 Mb, in B space found in *M. auratus*, compared to the 30.84 Mb found in *M. greshakei*, suggests that the B chromosome of *M. greshakei* may be twice as large as the B chromosome of *M. auratus*.

Table 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. zebra 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	-	-	-

3.1.2. 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 averaged ~100 bp in length and the PacBio reads averaged 8,295 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 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 2) and a depletion of shorter blocks can also be seen in the block size histogram of the PacBio data (Appendix A Figure A1). 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 error. However, even when using the conservative core block set, the PacBio data identifies only 15.02 Mb of core sequence in B space compared to the 17.67-24.09 Mb identified in the three Illumina data sets, 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 3 panel A). The Illumina reads in these short regions all aligned to several other locations (as indicated with white reads in Figure 3) and these regions were annotated as various repeats. Our interpretation is that these regions represent 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 Illumina data was able to detect these B-specific sequences while 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.

A second difference between the sequence data types was in the detection of retrogene insertions (Figure 3 panel B). Again, since the PacBio reads are much longer than the retro-inserted 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 retro-inserted exons and therefore do align well to the reference. This means that standard alignment software and parameters will detect retro-inserted sequences on the B chromosome with short read data, but not with long read data. Proper alignment of retro-inserted 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 retro-insertion with the PacBio data by aligning with GMAP, but the majority of A genome reads did not map. Alignment software that accounted 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 retro-inserted genes (Figure 3 panel C). The Illumina data showed increased coverage in the exons, but not the introns, of some genes, suggesting it was another retro-inserted 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 retro-inserted gene. We were only able to distinguish between "true" and "false" retro-inserted genes on the B chromosome by comparing the Illumina data with PacBio data.

Figure 3. Comparisons of Illumina and PacBio Read Alignment in B Blocks

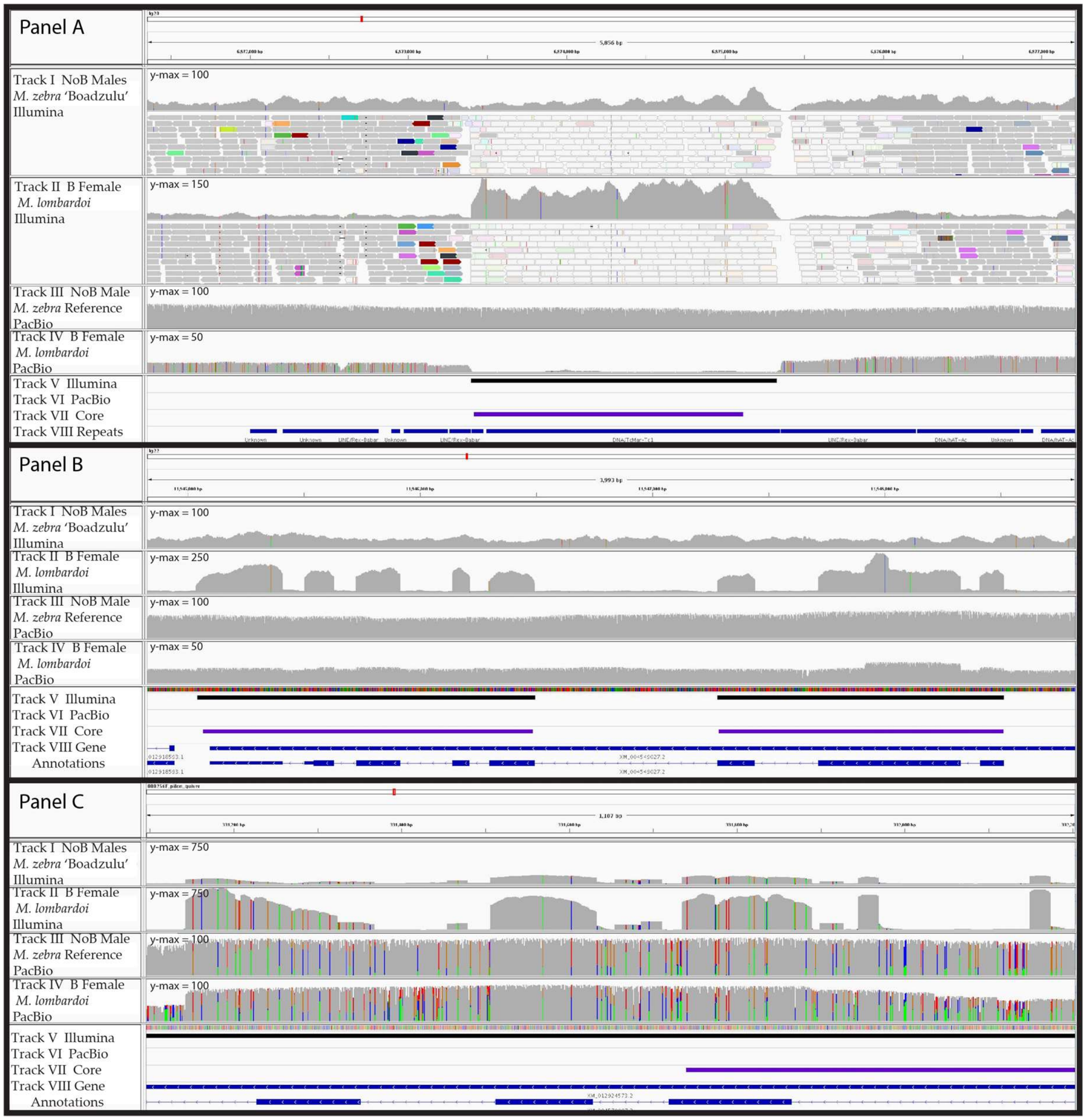


Figure 3 Caption: Differences in the alignment of Illumina and PacBio reads affect the B block identification analysis. Panels A, B and C represent regions of LG20, LG22 and the unanchored scaffold 000256F_pilon_quiver, respectively. Panel A demonstrates a failure to identify a B block with PacBio data. 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). In panels A, B and 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, B and C, the B female *M. lombardoi* block sets for Illumina and PacBio are shown in tracks Vand 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 and 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.

3.2. B Block Turnover

B chromosomes are thought to have a high rate of sequence turnover because they experience little purifying selection [14,15]. Because the Lake Malawi cichlid species studied here diverged less than 1MY ago [46] 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 sequence that was lost from a common ancestor and new sequence acquired during the evolution of particular lineages.

3.3. 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 7 cichlid species. Figure 4 depicts the position of core B blocks on the chromosome-scale assembly of the *M. 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 [11,12,13]. No trend was observed between B block position and centromere position. There is no readily visible pattern suggesting certain regions are more likely than others to be 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 1). The SCR of the core blocks varies among individuals. The largest difference in SCR between these two individuals is shown on LG8.

Figure 4. Karyoplot showing the A genome origins of the B chromosome.

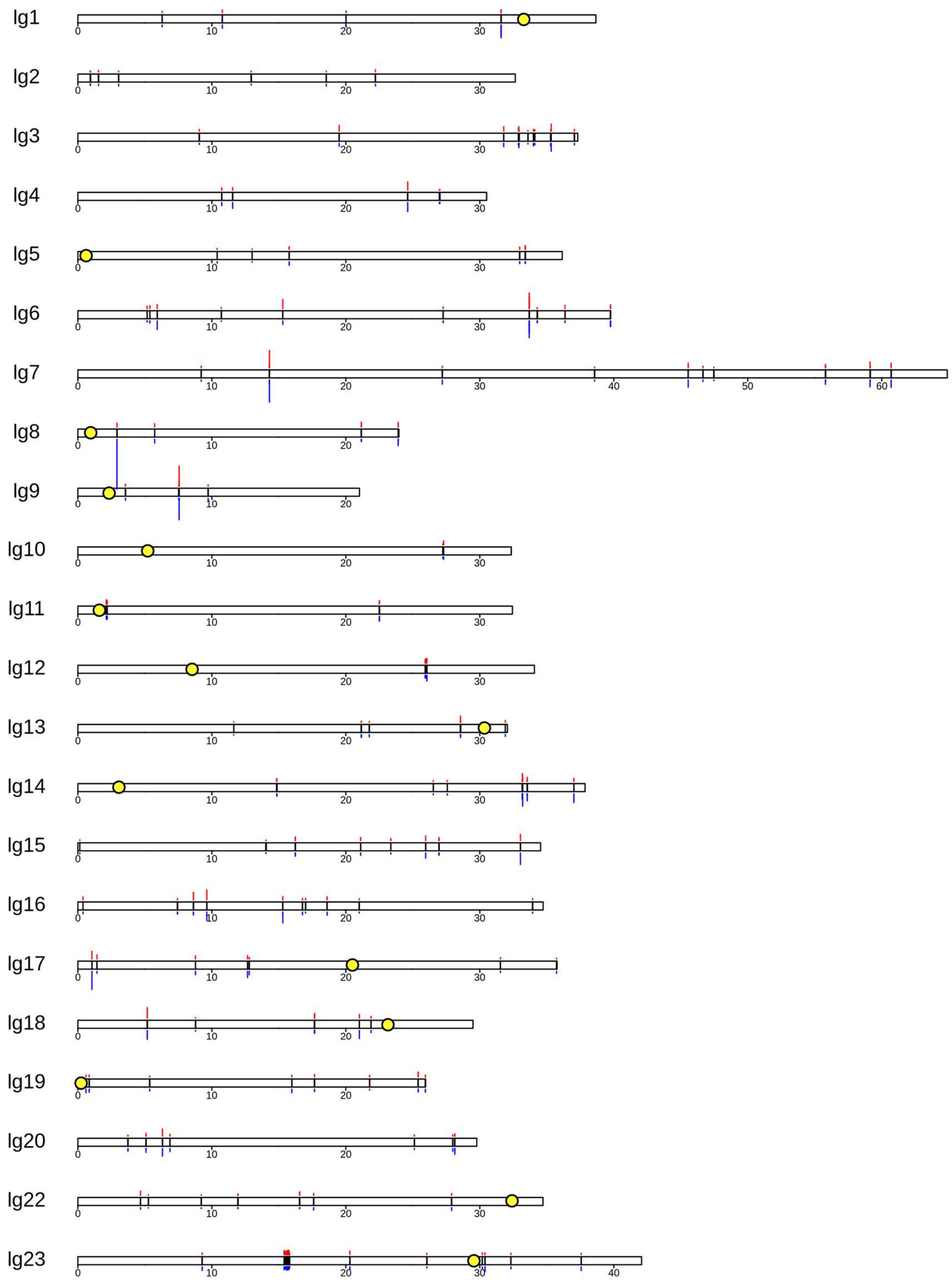


Figure 4 Caption: The position of B blocks (black bars) is superimposed on a karyoplot of the *M. zebra* 'Mazinzi' reference genome. The A genome consists of 22 chromosomes. For simplicity, unanchored scaffolds of the genome assembly were not included. Physical distances are noted beneath each LG in 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 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.

3.4. 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. The complete list of genes and gene fragments identified in each data set is provided in Appendix A (Directory A3).

Table 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. zebra 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

Figure 5 includes two Venn diagrams of B-located genes shared among the three *M. zebra* ‘Boadzulu’ individuals (0976, 0983, 0986) and among the three *M. lombardoi* individuals (1018, 1021, 1108). In both cases, the individuals from the same population share most of their presumed B-located genes, though there are still several hundred unique to each individual.

Figure 5. Shared B-located Genes and Gene Fragments Among Individuals of the Same Population

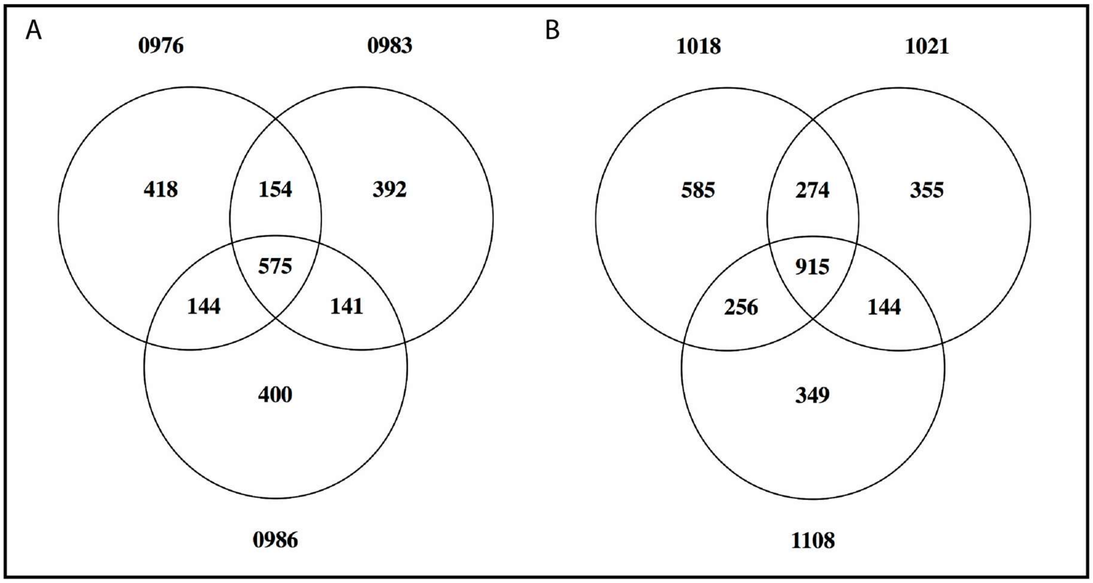


Figure 5 Caption: The number of B chromosome gene segments shared among individuals. The Venn diagrams show the number of genes and gene fragments shared by A) three *M. zebra* 'Nkhata Bay' individuals and B) three *M. lombardoi* individuals.

4. Discussion

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 Mb 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 percent 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 4 B-located copies (a SCR of 3).

Across the 13 datasets, the core blocks represented 30.22-53.63% of the total sequence identified in B space, leaving 46.37-69.78% as variable among individuals. As discussed above, we believe an appreciable amount of this unshared sequence is actually type 1 error, and therefore 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.

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 [7]. It is possible 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 [47]. 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). 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 remains, 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.

The number of genes and gene fragments overlapping with B blocks ranged from 516-2030 among datasets. Only 132 were common to at least 12 of the 13 datasets. When comparing individuals of the same population, (Figure 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.

Our 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, we do not know the structure of these sequences on the B chromosome. 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 there of, 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 female sex bias. 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 better able to assemble the repetitive sequence of the B chromosome.

5. Conclusions

Using a coverage ratio analysis, we were able to identify the sequence of a significant portion of the B chromosome in several cichlid species from Lake Malawi. An evaluation of this approach, including the comparison of sequence data types, has provided crucial insight for the future application of coverage ratio analysis to study B chromosomes in other taxa. 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. Both the rate of turnover and pattern of B blocks across the A genome provide important caveats to efforts to characterize the origin of B chromosomes. Finally, we identify a list of candidate genes and gene fragments located on the B chromosome that may include the gene(s) responsible for drive and female sex bias in Lake Malawi cichlids.

Author Contributions: DNA samples were collected, extracted and libraries were made by all authors. PacBio alignment was performed by M. A. Conte. Block identification scripts were written by F. E. Clark, with oversight from M. A. Conte and T. D. Kocher. All authors contributed to writing of the manuscript.

Funding: This research was funded by National Science Foundation, grant number DEB-1143920.

Acknowledgments: We would like to thank the National Science Foundation for funding this study.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

Directory A1 – Scripts

Directory A2 – Block information

Figure A1 – Histograms of Block Length

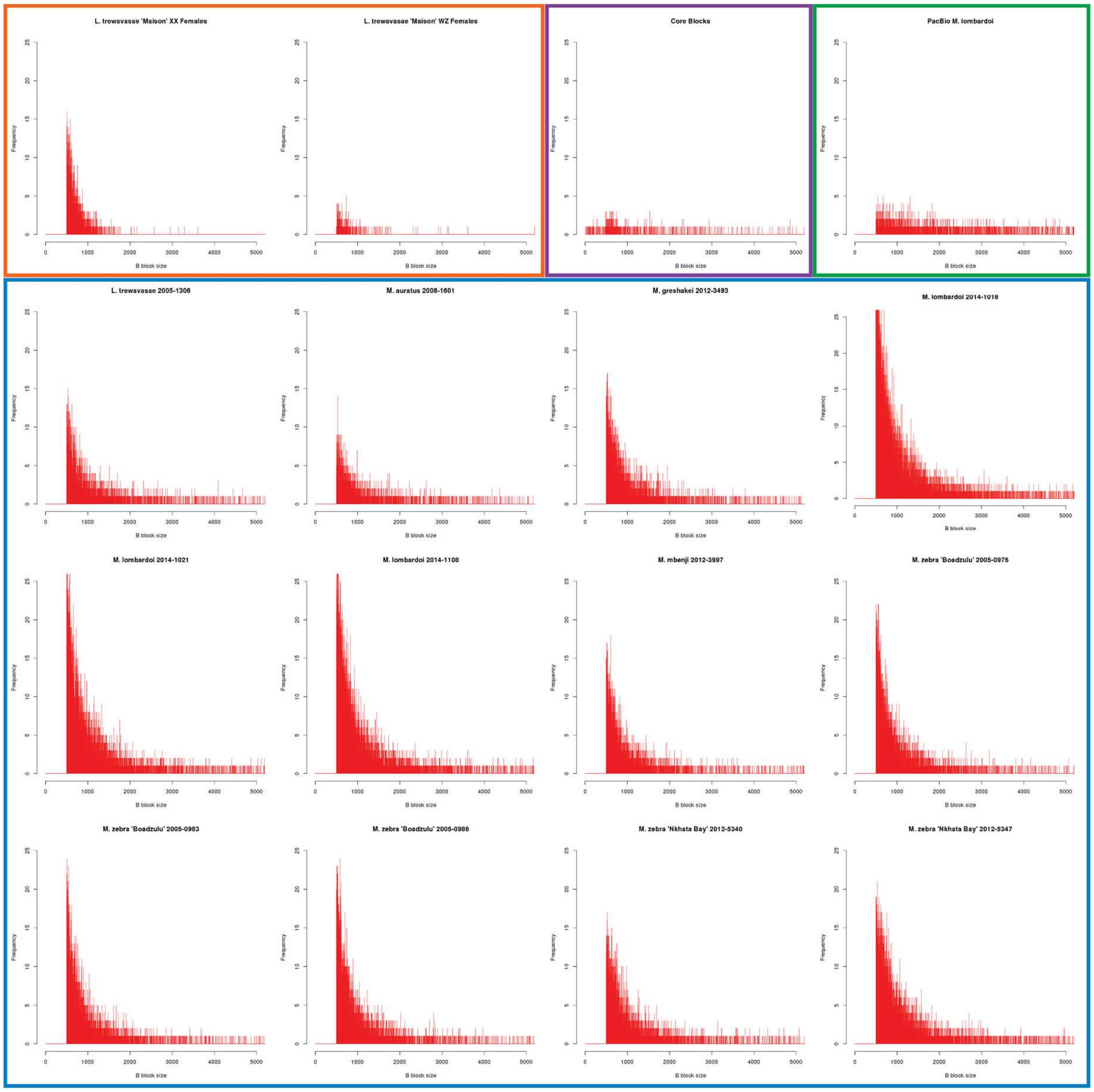


Figure A1 Caption: Histograms of B block length are provided for all data sets. B block size along the x-axis is reported in bp, and only blocks 5000 bp or smaller are shown. The histograms of the two controls are outlined in orange. The core block set in outlined in purple. The histogram of the PacBio data set is outlined in green. Finally, the histograms of the Illumina B data sets are outlined in blue.

Directory A3 – List of genes

References

1. Wilson, E. B. The supernumerary chromosomes of Hemiptera. *Science.* 1907, 26, 870-871.
2. Jones, R. N.; Rees, H. B chromosomes. Academic Press, New York USA, 1982.
3. Jones, R. N. B-chromosome drive. *Am Nat* 1991, 137, 430-442, doi: 10.1086/285175.
4. Camacho, J. P.; Sharbel, T. F.; Beukeboom, L. W. B-chromosome evolution. *Philos Trans R Soc B Biol Sci* 2000, 355, 163-178, doi: 10.1098/rstb.2000.0556.
5. Randolph, L. F. Genetic characteristics of the B chromosomes in maize. *Genetics* 1941, 26(6), 608-631.
6. Burt, A.; Trivers, R. Genes in conflict: the biology of selfish genetic elements. Belknap Press, Cambridge England, 2008, 325-380.
7. Houben, A. B chromosomes – a matter of chromosome drive. *Frontiers in Plant Sci* 2017, 8(210), doi: 10.3389/fpls.2017.00210.eCollection2017.
8. Cheng, Y. M.; Lin, B. Y. Cloning and characterization of maize B chromosome sequences derived from microdissection. *Genetics* 2003, 164(1), 299-310.
9. Bugrov, A. G.; Karamysheva, T. V.; Perepelov, E. A.; Elisaphenko, E. A.; Rubtsov, D. N.; Warchalowska-Sliwa, E.; Tatsuta, H. Rubtsov, N. B. DNA content of the B chromosomes in grasshopper *Podisma kanoi* Storozh. (Orthoptera, Acrididae). *Chromosome Res* 2007, 15(3), 315-326, https://doi.org/10.1007/s10577-007-1128-z.
10. Ruiz-Ruano, F. J.; Cabrero, J.; Lopez-Leon, M. D.; Sanchez, A.; Camacho, J. P. M. Quantitative sequence characterization for repetitive DNA content in the supernumerary chromosome of the migratory locust. *Chromosoma* 2018, 127, 45-57, doi: 10.1007/s00412-017-0644-7.
11. Jones, N.; Houben, A. B chromosomes in plants: escapees from the A chromosome genome? *TRENDS in Plant Science* 2003, 8(9), 1360-1385, doi: 10.1016/S1360-1385(03)00187-0.
12. Valente, G. T.; Conte, M. A.; Fantinatti, B. E. A.; Cabral-de-Mello, D. C.; Carvalho, R. F.; Vicari, M. R.; Kocher, T. D.; Martins, C. Origin and evolution of B chromosomes in the cichlid fish *Astatotilapia latifasciata* based on integrated genomic analyses. *Molecular Biology and Evolution* 2014, 31(8), 2061-2072, doi: 10.1093/molbev/msu148.
13. Martis, M. M.; Klemme, S.; Banaei-Moghaddam, A. M.; Blattner, F. R.; 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. F. X.; Houben, A. Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proc Natl Acad Sci USA* 2012, 109(33), 13343-13346, doi: 10.1073/pnas.1204237109.
14. Houben, A.; Banaei-Moghaddam, A. M.; Klemme, S.; Timmis, J. N. Evolution and biology of supernumerary B chromosomes. *Cell Mol Life Sci* 2014, 71(3), 467-478, doi: 10.1007/s00018-013-1437-7.
15. Klemme, S.; Banaei-Moghaddam, A. M.; Macas, J.; Wicker, T.; Novak, P.; Houben, A. High-copy sequences reveal distinct evolution of the rye B chromosome. *New Phytol* 2013, 199, 550-558, doi: 10.1111/nph.12289.
16. Ruban, A.; Schmutzer, T.; Scholz, U.; Houben, A. How next-generation sequencing has aided our understanding of the sequence composition and origin of B chromosomes. *Genes* 2017, 8(11), 294, doi: 10.3390/genes8110294.
17. Makunin, A. I.; Dementyeva, P. V.; Graphodatsky, A. S.; Volobouev, V. T.; Kukekova, A. V.; Trifonov, V. A. Genes on B chromosomes of vertebrates. *Mol Cytogenetics* 2014, 7(99), doi: 10.1186/s13039-014-0099-y.
18. Banaei-Moghaddam, A. M.; Martis, M. M.; Macas, J.; Gundlach, H.; Himmelbach, A.; Altschmied, L.; Mayer, K. F. X.; Houben, A. Genes on B chromosomes: Old questions revisited with new tools. *Biochim Biophys Acta* 2015, 1849(1), 64-70, doi: 10.1016/j.bbagrm.2014.11.007.
19. Makunin, A. I.; Kichigin, I. G.; Larkin, D. M.; O'Brien, P. C. M.; Ferguson-Smith, M. A.; Yang, F.; Proskuryakova, A. A.; Vorobieva N. V.; Chernyaeva, E.N.; O'Brien, S. J.; Graphodatsky, A. S.; Trifonov, V. A. Contrasting origin of B chromosomes in two cervids (Siberian roe deer and grey brocket deer) unraveled by chromosome-specific DNA sequencing. *BMC Genomcis* 2016, 17(618), doi: 10.1186/s12864-016-2933-6.
20. Jones, R. New species with B chromosomes discovered since 1980. *Nucleus* 2017, 60, 263-281, doi: 10.1007/s13237-017-0215-6.
21. Zhou, Q.; Zhu, H.; Huang, Q.; Zhao, L.; Zhang, G.; Roy, S. W.; Vicoso, B.; Xuan, Z.; Ruan, J.; Zhang, Y.; Zhao, R.; Ye, C.; Zhang, X.; Wang, J.; Wang, W.; Bachtrog, D. Deciphering new-sex and B chromosome evolution by the draft genome of *Drosophila albomicans*. *BMC Genomics* 2012, 13(109), doi: 10.1186/1471-2164-13-109.

22. Gonzalez-Sanchez, M.; Chiavarino, M.; Jimenez, G.; Manzanero, S.; Rosato, M.; Puertas, M. J. The parasitic effects of rye B chromosomes might be beneficial in the long term. *Cytogenet Genome Res* 2004, 106(2-4), 386-393, doi: 10.1159/000079316.
23. Jones, R. N. Gonzalez-Sanchez, M.; Gonzalez-Garcia, M.; Vega, J. M.; Puertas, M. J. Chromosomes with a life of their own. *Cytogenet Genome Res* 2008, 120, 265-280, doi: 10.1159/000121076.
24. Graphodatsky, A. S.; Kukekova, A. V.; Yudkin, D. V.; Trifonov, V. A.; Vorobieva, N. V.; Beklemisheva, V. R.; Perelman, P. L.; Graphodatskaya, D. A.; Trut, L. N.; Yang, F.; Ferguson-Smith, M. A.; Acland, G. M.; Aguirre, G. D. The proto-oncogene C-KIT maps to canid B-chromosomes. *Chromosome Res* 2005, 13, 113-122.
25. Poletto, A. B.; Ferreira, I. A.; Martins, C. The B chromosomes of the African cichlid fish *Haplochromis obliquidens* harbor 18S rRNA gene copies. *BMC Genet* 2010, 11(1), doi: 10.1186/1471-2156-11-1.
26. Navarro-Dominguez, B.; Ruiz-Ruano, F. J.; Cabrero, J.; Corral, J. M.; Lopez-Leon, M. D.; Sharbel, T. F.; Camacho, J. P. M. Protein-coding genes in B chromosomes of the grasshopper *Eyprepocnemis plorans*. *Scientific Reports* 2017, 7(45200), doi: 10.1038/srep45200.
27. Leach, C. R.; Houben, A.; Field, B.; Pistrick, K.; Demidov, D.; Timmis, J. N. Molecular evidence for transcription of genes on a B chromosome in *Crepis capillaris*. *Genetics* 2005, 171, 269-278, doi: 10.1534/genetics.105.043273.
28. Ruiz-Estevez, M.; Lopez-Leon, M. D.; Cabrero, J.; Camach, J. P. M. B-chromosome ribosomal DNA is functional in the grasshopper *Eyprepocnemis plorans*. *PLOSone* 2012, 7(5), e36600, doi: 10.1371/journal.pone.0036600.
29. Ruiz-Estevez, M.; Badisco, L.; Broeck, J. V.; Perfectti, F. Lopez-Leon, M. D.; Cabrero, J.; Camacho, J. P. M. B chromosomes showing active ribosomal RNA genes contribute insignificant amounts of rRNA in the grasshopper *Eyprepocnemis plorans*. *Mol Genet Genomics* 2014, doi: 10.1007/s00438-014-0880-y.
30. Navarro-Dominguez, B.; Ruiz-Ruano, F. J.; Camacho, J. P. M.; Cabrero, J.; Lopez-Leon, M. D. Transcription of a B chromosome CAP-G pseudogene does not influence normal Condensin Complex genes in a grasshopper. *Scientific Reports* 2017, 7, 17650, doi: 10.1038/s41598-017-15894-5.
31. Banaei-Moghaddam, A. M.; Meier, K.; Karimi-Ashtiyani, R.; Houben, A. Formation and expression of pseudogenes on B chromosome of rye. *The Plant Cell* 2013, 25, 2536-2544, doi: 10.1105/tpc.113.111856.
32. Huang, W.; Zhao, Y. D. X.; Jin, W. B chromosome contains active genes and impacts the transcription of A chromosomes in maize (*Zea mays* L.). *BMC Plant Biology* 2016, 16(88), doi: 10.1186/s12870-016-0775-7.
33. Lin, H. Z.; Lin, W. D.; Lin, C. Y.; Peng, S. F.; Cheng, Y. M. Characterization of maize B-chromosome-related transcripts isolated via cDNA-AFLP. *Chromosoma* 2014, 123, 597-607, doi: 10.1007/s00412-014-0476-7.
34. Treangen, T. J.; Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 2011, 13(1), 36-46, doi: 10.1038/nrg3117.
35. Feldberg E.; Bertollo, L. A. C. Discordance in chromosome number among somatic and gonadal tissue cells of *Gymnogeophagus balzanii* (Pisces: Cichlidae). *Braz J Genet* 1984, 4, 639-645.
36. Feldberg, E.; Porto, J. I. R.; Alves-Brinn, M. N.; Mendonca, M. N. C.; Benzaquem, D. C. B chromosomes in Amazonian cichlid species. *Cytogenet Genome Res* 2004, 106, 195-198, doi: 10.1159/000079287.
37. Poletto, A. B.; Ferreira, I. A.; Cabral-de-Mello, D. C.; Nakajima, R. T.; Mazzuchelli, J.; Ribeiro, H. B.; Venere, P. C.; Nirchio, M.; Kocher, T. D.; Martins, C. Chromosome differentiation patterns during cichlid fish evolution. *BMC Genet* 2010, 11(50), doi: 10.1186/1471-2156-11-50.
38. Pires, L. B.; Sampaio, T. R.; Dias, A. L. Mitotic and meiotic behavior of B chromosomes in *Crenicichlia lepidota*: New report in the family Cichlidae. *Journal of Heredity* 2015, doi: 10.1093/jhered/esv007.
39. Clark, F. E.; Conte, M. A.; Ferreira-Bravo, I. A.; Poletto, A. B.; Martins, C.; Kocher, T. D. Dynamic sequence evolution of a sex-associated B chromosome in Lake Malawi cichlid fish. *Journal of Heredity* 2017, 108(1), 53-62, doi: 10.1093/jhered/esw/059.
40. Yoshida, K.; Terai, Y.; Mizoiri, S.; Aibara, M.; Nishihara, H.; Watanabe, M.; Kuroiwa, A.; Hirai, H.; Hirai, Y.; Matsuda, Y.; Okada, N. B chromosomes have a functional effect on female sex determination in Lake Victoria cichlid fishes. *Plos one Genetics* 2011, 7(8), e1002203, doi: 10.1371/journal.pgen.1002203.
41. Conte, M. A.; Kocher, T. D. An improved genome reference for the African cichlid, *Metriaclima zebra*. *BMC Genomics* 2015, 16, 724, doi: 10.1186/s12864-015-1930-5.
42. Conte, M. A.; Joshi, R.; Moore, E. C.; Nandamuri, S. P.; Gammerdinger, W. J.; Roberts, R. B.; Carleton, K. L.; Lien, S.; Kocher, T. D. Chromosome-scale assemblies reveal the structural evolution of African cichlid genomes. *bioRxiv* preprint, doi: 10.1101/383992.

43. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, 25(14), 1754-1760, doi: 10.1093/bioinformatics/btp324.
44. Sedlazeck, F. J.; Rescheneder, P.; Smolka, M.; Fang, H.; Nattestad, M.; Haeseler, A. V.; Schatz, M. C. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods* 2018, 15, 461-468.
45. Quinlan, A. R.; Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010, 26(6), 841-842, doi: 10.1093/bioinformatics/btq033.
46. Kocher, T. D. Adaptive evolution and explosive speciation: The cichlid fish model. *Nature Reviews Genetics* 2004, 5, 288-298, doi: 10.1038/nrg1316.
47. Feschotte, C.; Pritham, E. J. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet*. 2007, 41, 331-368, doi: 10.1146/annurev.genet.40.110405.090448.
48. Gel, B.; Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* 2017, 31-33, doi: 10.1093/bioinformatics/btx346.