ABSTRACT

Title of Dissertation:	STRUCTURAL EVOLUTION OF AFRICAN CICHLID GENOMES
	Matthew A. Conte, Doctor of Philosophy, 2018
Dissertation directed by:	Professor Thomas D. Kocher, Department of Biology

An unanswered question in biology is how the evolution of genome structure supports or accompanies diversification and speciation on different time scales. African cichlid fishes are a well-documented system ideal for studying rapid evolution, due to their phenotypic diversity and high number of speciation events over the last several million years. I generated two *de novo* genome assemblies of the riverine cichlid *Oreochromis niloticus* (tilapia) and the Lake Malawi cichlid *Metriaclima zebra* using high-coverage long-read sequencing data and anchored the assemblies to chromosomes using several genetic and physical maps, to produce two high-quality anchored references. By comparing these chromosome-scale assemblies to integrated recombination, transcriptome, and resequencing data of multiple genera and species, I identified and characterized many large novel genome rearrangement events. These rearrangements included multiple novel sex-determination inversions, several metacentric-acrocentric karyotype differences via centromere assembly and placement, and wide regions of suppressed recombination in genera- and species-

level crosses of Lake Malawi cichlids. Karyotype evolution in cichlids was further analyzed with long-read sequencing, specifically revealing the complex structure and content of a highly repetitive supernumerary chromosome present in some but not all individuals of a population across a wide range of eukaryotes, including many cichlid species. These supernumerary “B” chromosomes are shown to be limited to female Lake Malawi cichlids and have a unique evolutionary history with B chromosomes present in Lake Victorian cichlids male and females. This work reveals how structural genomic changes impact a rapidly evolving clade, while providing high-quality resources for the community, a context for previous genetic studies, and a robust platform for future genome research in cichlids.

STRUCTURAL EVOLUTION OF AFRICAN CICHLID GENOMES

by

Matthew A. Conte

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
2018

Advisory Committee:
Professor Thomas D. Kocher, Chair
Professor Karen L. Carleton
Professor Michael P. Cummings
Professor Carlos A. Machado
Professor Mihai Pop, Dean's Representative

Preface

This dissertation is based on the following publications:

Chapter 2

Matthew A. Conte and Thomas D. Kocher. An improved genome reference for the African cichlid, *Metriaclima zebra. BMC Genomics*. 2015;16(1):724.

Chapter 3

Matthew A. Conte, William J. Gammerdinger, Kerry L. Bartie, David J. Penman, Thomas D. Kocher. A high quality assembly of the Nile Tilapia (*Oreochromis niloticus*) genome reveals the structure of two sex determination regions. *BMC Genomics*. 2017;18(1):341.

Chapter 4

Matthew A. Conte, Rajesh Joshi, Emily C. Moore, Sri Pratima Nandamuri, William J. Gammerdinger, Frances E. Clark, Reade B. Roberts, Cesar Martins, Karen L. Carleton, Sigbjørn Lien, Thomas D. Kocher. Chromosome-scale assemblies reveal the structural evolution of African cichlid genomes. *Prepared*.

Chapter 5

Matthew A. Conte, Frances E. Clark, Karen L. Carleton, Cesar Martins, and Thomas D. Kocher. Origin, evolution, and history of B chromosomes in African cichlids. *In preparation*.

Foreword

"Essentially, all models are wrong, but some are useful."

-George Box

Dedication

I dedicate this dissertation to my family for their love and support throughout: Kathryn Walters-Conte and our two children Avery and Marcus.

Acknowledgements

This dissertation would not have been possible without significant contributions from a variety of people. I would first like to recognize my advisor and great mentor, Thomas Kocher, whom I've had the pleasure and opportunity of learning so much from over the years. Karen Carleton has also been an instrumental mentor and supporter of my research. I am fortunate to have them both by my side along the way. I would like to recognize the three additional members of my committee, Michael Cummings, Carlos Machado and Mihai Pop for their time and valuable input towards this dissertation. I would like to recognize many members within the community of the Biological Sciences graduate program, including numerous colleagues, collaborators and staff. The comradery of all the members of the cichlid labs at the University of Maryland, both past and present, and our daily lunch discussions of both science and the world are something I will always fondly remember.

The following acknowledgements are specific to each chapter in this dissertation.

Chapter 2 - We thank Thomas Hackl for the very helpful discussion and guidance in running the Proovread error correction steps. We thank Luke Tallon and Naomi Sengamalay of the Institute for Genome Sciences core facility for providing a high quality PacBio library and sequence reads. We thank the members of the Kocher and Carleton labs at UMD and the UMD Bring Your Own Bioinformatics club for providing thoughtful advice during the course of the project. We thank Karen Carleton, William Gammerdinger and Ian Misner for reading and providing

thoughtful feedback on the manuscript. We acknowledge the University of Maryland supercomputing resources (www.it.umd.edu/hpcc/) made available in conducting the research reported in this paper. This work was supported by the National Science Foundation under grant DEB-1143920.

Chapter 3 - We thank Luke Tallon and Naomi Sengamalay at the Genomics Resource Center, Institute for Genome Sciences for providing a high quality PacBio library and sequence reads. We thank Gideon Hulata of the Agricultural Research Organization in Israel for providing the original stocks of *O. aureus* that we maintained for this study. We thank Christopher Hill at University of Washington for helpful guidance on assembly parameters. We thank the members of the Kocher and Carleton labs at UMD for providing thoughtful advice during the course of the project. We thank Karen Carleton for reading and providing thoughtful feedback on the manuscript. We thank Christos Palaiokostas (Roslin Institute) and Michaël Bekaert (University of Stirling) for providing the marker sequences of the RAD map. We thank Lucile Soler (Bioinformatics Infrastructure for Life Sciences) for BAC-end analysis scripts. We thank the NCBI RefSeq curators for annotation of the O_niloticus_UMD1 genome assembly. We acknowledge the University of Maryland supercomputing resources (www.it.umd.edu/hpcc/) made available in conducting the research reported in this paper.

Chapter 4 - We acknowledge all members of the UMD cichlid labs for their countless discussions on this project. We thank Luke Tallon and Naomi Sengamalay at the Genomics Resource Center, Institute for Genome Sciences for providing a high quality PacBio library and sequence reads. We acknowledge the University of

Maryland supercomputing resources (www.it.umd.edu/hpcc/) made available in conducting the research reported in this paper. We thank many individuals in the cichlid community for their patience while we developed these resources.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AFC - African cichlid specific repetitive element.

ALE – Assembly likelihood estimate.

BAC - Bacterial artificial chromosome.

BLAST – Basic Local Alignment Search Tool.

bp – base pair.

BUSCO - Benchmarking Universal Single-Copy Orthologs.

BWA - Burrows-Wheeler Aligner.

CDS – Coding sequence.

CEG - Core eukaryotic gene.

CEGMA - Core Eukaryotic Genes Mapping Approach.

cM - Centimorgan.

DNA - deoxyribonucleic acid

F_{ST} – Fixation index.

Gb – Gigabase.

GMAP - Genomic Mapping and Alignment Program.

HMW – High molecular weight.

Kb– Kilobase.

L50 - Smallest number of contigs whose length sum produces N50.

LD - Linkage disequilibrium.

LG – Linkage group.

LINE – Long interspersed nuclear element.

LG50 - Smallest number of contigs whose length sum produces NG50.

LGs – Linkage groups.

LWS – Long-wavelength-sensitive opsin.

mRNA - Messenger RNA.

Mb – Megabase.

N50 - Shortest contig/scaffold/read/sequence length at 50% of the genome/read set.

NCBI - National Center for Biotechnology Information.

NG50 - Shortest contig/scaffold/read/sequence length at 50% of the estimated genome/read set size.

ONSATA – *Oreochromis niloticus* satellite A repeat.

ONSATB *Oreochromis niloticus* satellite B repeat

PacBio - Pacific Biosciences.

PCA - Principal component analysis.

PCR - Polymerase chain reaction.

RAD – Restriction site associated DNA.

RefSeq - NCBI Reference Sequence Database.

RH – Radiation hybrid.

RH2 - Blue-green-sensitive opsin.

RNA - ribonucleic acid.

SAM - Sequence Alignment/Map.

SINE – Short interspersed nuclear element.

SMRT - Single Molecule, Real-Time.

SWS – Short-wavelength sensitive opsin.

TE – Transposable element.

TZSAT - *Tilapia zillii* satellite repeat.
XX – Female homogametic sex.
XY – Male heterogametic sex.
ZW – Female heterogametic sex.
ZZ – Male homogametic sex.

Chapter 1: Introduction

1.1 Genome evolution

Evolutionary forces shape the structure and content of genomes over time. The field of comparative genomics has documented changes large and small, including genome duplications, genome reductions, chromosome fusions, chromosome rearrangements, gene duplications and pseudogenization. Genomic conflicts contribute to the invasion of selfish genetic elements such as transposable elements, gene drivers and B chromosomes that also shape genomes. Many changes in genome structure are associated with noticeable differences in gene expression, specific phenotypes and overall organismal fitness. Study of the many genome sequences now available, for species with diverse and unique phenotypes, will contribute to our understanding of the forces that have shaped genome architecture.

1.2 Cichlid fish

A scuba dive or snorkel in one of the Great Lakes of Africa will quickly reveal one of the most diverse set of vertebrates on earth. Various pigmentation patterns and colors of the African cichlid fish are likely the first characteristic one will notice. Perhaps the next distinction one may spot are the morphological differences between the many species of cichlids. These morphological adaptations have allowed many species to specialize into various niches spread throughout the lakes. There are also quantifiable behaviors specific to particular groups of cichlids. Many less obvious traits, such as visual sensitivity and sex determination systems, have been shown to vary greatly as well. These cichlid phenotypes have evolved not once, but multiple times and to different extents across the lakes within the African rift valley, which has its own rich geological history. More distantly related Neotropical cichlids occupy parts of central America, Madagascar and India. In total there are estimated to be 3,000 species of cichlids across the world (1). The Great Lakes of Malawi, Victoria and Tanganyika have hosted the largest radiations of cichlids. Estimates of cichlid species numbers in these lakes vary from 200-250 in Lake Tanganyika (2) to 500-1,000 in Lake Malawi (3).

1.3 The cichlid genome

These cichlid fish provide an ideal system for studying many processes underlying speciation. Here we will use comparative genomics to begin to understand some of the mechanisms of cichlid speciation and the molecular basis of their diverse phenotypes. This first requires having well assembled genome references. The complete history of the cichlid genome project is probably best told over a few drinks with my advisor, Tom Kocher. Making a long story short, I'll fast-forward to when the cichlid genome white paper was submitted to the National Human Genome Research Institute (NHGRI) in 2006. The original cichlid genome white paper proposed 5X coverage of Sanger sequencing to build a draft assembly of the riverine tilapia, *Oreochromis niloticus* and low coverage shotgun of four additional species (4). About eight years and dozens of conference calls later, we along with the cichlid community, published the first cichlid genomes in *Nature* (5). This work was based on short-read Illumina genome assemblies of five cichlids (three from the original white paper, including tilapia as the most accurate). These assemblies represented the state of the art using the latest genome sequencing technologies and remained some of the best short-read only vertebrate assemblies for some time (6).

1.4 Chromosome-scale, high-quality reference genome assemblies

In the years since, the limitations of genome assemblies based on short reads have become obvious and well cataloged. Fragmented and misassembled short read genome assemblies result in inconsistent gene predictions and for several reasons often overestimate (and sometimes underestimate) the number of genes, especially when assembling expanded gene families (7). This suggests that evolutionary histories of genes and trait mapping conclusions drawn from flawed genome assemblies could easily be wrong. Indeed, some of the most interesting and perhaps evolutionarily important regions of the genome may be the most difficult areas to assemble correctly. Recently duplicated regions are notoriously difficult to assemble due to their high sequence identity (7). Many of these problems are due to the relatively short length of these reads. PCR amplification and GC bias introduce more problems when using Illumina data (8,9).

Genome assembly problems were not limited to the short-read era. Improvements to the initial mouse draft genome assembly revealed a large amount of previously missing sequence, many missing duplicated genes, and regions containing transposable elements with importance to rodent specific biology (10). Additional examples from draft assemblies of the rhesus macaque (11) and *Bos taurus* (12) showed the importance of assembly quality on gene predictions. Another recent study has shown that an apparent loss of a critical hedgehog gene in several birds was actually due to genome assembly errors (13). These problems of poor draft assemblies have been exacerbated in the short-read era and demonstrated the need for higher quality genome assemblies (14).

There were some initial efforts made by the assembly community to improve the situation, such as the GAGE evaluation of genome assembly software by a group of assembly experts. The results of the GAGE evaluation showed that data quality plays a large factor, that assembly contiguity varies greatly between different assemblers and different genomes, and that the correctness of an assembly greatly varies and is not well correlated with assembly contiguity (15). Similarly, the Assemblathon2 competition (assemblathon.org) gathered many separate groups of experts to "compete" to produce the best possible assemblies of a bird, a snake, and a fish (the Lake Malawi *Metriaclima zebra*) using primarily short-read data. The results of the competition were useful to the genomics community and while no absolute "winners" were awarded, it showed the need to use different types of data and metrics to properly evaluate genome assemblies (6). About the same time that short read genome assembly was reaching theoretical limits on what could be assembled given the limited information of the data, long read sequencing became available. Several initial microbial long read genomes and methods were presented (16,17) and eventually vertebrate genome assembly with long read sequencing became possible (18–20). Throughout this large shift in sequencing technology, the Genome 10K project has been working (21) "to assemble a genomic zoo – a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately one for every vertebrate genus." Thankfully, the Genome 10k organizers have recognized the importance and benefits of shifting to long read based genome assemblies and have provided high standards for the genome assembly community ("N50 contig >1Mb, N50 Scaffold >10Mb, >90% of genome assembled into chromosomes, and phased as much as possible") in their recent Vertebrate Genomes Project announcement (22). While these standards were just recently set forth, they correspond with the standards that we have put on our genome assemblies of cichlids.

1.5 Genome structure of African cichlids

Accurate chromosome-level genome assemblies allow large genomic rearrangements and structural evolution to be studied. Cichlid karyotypes vary in diploid chromosome number from 32-60 (23). However, the large majority of African cichlid genomes consist of 22 chromosome pairs, including the two species that are assembled in this dissertation, *Oreochromis niloticus* and *Metriaclima zebra.* Neotropical cichlids have a mode of 24 diploid chromosomes. Among African and Neotropical cichlids there is a large amount of variation in the numbers of metacentric, submetacentric, subtelomeric and acrocentric chromosomes. Assembly and anchoring of centromere repeats on each chromosome in this work allows these differences in centromere position to be studied for the first time in cichlid genome assemblies. In *O. niloticus*, there is 1 metacentric/submetacentric chromosome pairs and 21 subtelomeric/acrocentric chromosome pairs (24). In *M. zebra*, there are 6 metacentric/submetacentric chromosome pairs and 16 subtelomeric/acrocentric chromosome pairs.

Recent work has shown there to be a large diversity in sex determination loci among various cichlid species (25–29). Sex determination loci have been identified on at least 12 different cichlid chromosomes, with several chromosomes showing convergent evolution of sex determination loci in different lineages. The number of currently identified sex determination loci is almost certainly under-sampled and further work will likely show a sex determination locus on every cichlid chromosome once more species are sampled. Many of these sex determination loci are thought to be located in relatively large inversions where recombination with a tightly-linked sexually antagonistic allele has been suppressed (30). This rapid turnover in sex determination loci has likely played an important role in cichlid speciation (31) and has shaped much of the structural evolution in the cichlid genome.

Genetic recombination maps are one method to anchor assembled contigs into linkage groups or chromosomes. These recombination maps offer the advantage of providing recombination information across each chromosome that other physical mapping techniques do not. Patterns of recombination complement the structural changes identified and can help relate previous work to our genome comparisons. This work includes the use of five genetic recombination maps first to check genome assemblies, then to anchor contigs, and finally to study genome evolution in the species used to generate the maps. Four of the recombination maps are generated from crosses of Lake Malawi cichlids. The fifth map is a high-density recombination map using a genotyping SNP array of *O. niloticus,* where we are also able to compare male and female recombination. The patterns of recombination from five different maps complement the genome assemblies and allow for an additional layer of long-range comparison to be made.

1.6 Repetitive sequences and B chromosomes pose unique challenges and questions

Since the repetitive regions of genomes are typically the hardest parts to assemble, we focused on analyzing these regions. In many cases, comparison of the new assemblies to the draft assemblies demonstrated the improvements being made and reiterated the need for high quality assemblies. The hard problem of accurately assembling short reads *de novo* can become even more difficult when other, unexpected elements of a genome are also sequenced and introduced to the read set. One such odd element that we encountered are B chromosomes (32).

B chromosomes are non-essential, supernumerary chromosomes that are present in addition to the normal ("A") karyotype of an organism. They were first identified over 100 years ago (33). B chromosomes can be regularly found in some, but not all, individuals of a given population. They are estimated to occur across 15% of all eukaryotes (34) covering a wide range of taxa from fungi to plants to animals, including mammals (35). B chromosomes have been well studied cytogenetically but are only recently beginning to be understood at the genomic level (36). Originally thought to contain mostly repetitive DNA sequence and to be completely heterochromatic, recent studies have begun to show that B chromosomes contain transcribed genic sequences (32,37).

In Lake Victoria, B chromosomes have been found in a subset of species. They were shown to play a functional role in sex determination in at least one population of cichlids, but not the majority of populations harboring B chromosomes (38). The same study also showed that the size of B chromosomes varies greatly even within the same population.

We previously sequenced an individual with two B chromosomes and an individual with zero B chromosomes of the Lake Victorian cichlid *Astatotilapia latifasciata*. We were able to characterize regions of the B chromosome (B "blocks") that were homologous to sequences along the A chromosomes that revealed insights into the origin and evolution of that B chromosome (32). When we compared those B chromosome blocks to the original draft genome assembly of the Lake Victoria cichlid, *Pundamilia nyererei*, we realized that the fish chosen for genome sequencing also carried a similar B chromosome. This contributed to errors in the *P. nyererei* assembly in regions where the B chromosome was homologous with the A genome. These results demonstrated that karyotyping is an important first step in eukaryotic genome projects, especially if B chromosomes are known to be present in closely related species. Additional analysis of the transcriptomes of this individual allowed for the identification of several B chromosome genes that are being transcribed (32).

B chromosomes have also been karyotyped in a species of the Lake Malawi cichlid, *Metriaclima lombardoi* (39). Recently, we have sequenced over 20 different populations of Lake Malawi cichlids and have identified B chromosomes present in one copy, and solely in female individuals, of at least 7 populations (40). Based on these initial findings, we have re-sequenced multiple female individuals of these populations using short read data, and a single individual using long read data. We describe the structure of the B chromosome at the sequence level for the first time.

Comparisons among the Lake Malawi B chromosomes, and between the Lake Victorian B chromosome, were made to discover what sequence content they share. The origin, evolution, maintenance and role of B chromosomes in African cichlids is presented.

1.7 Outline of dissertation chapters

Chapter 2 describes our initial work using long read sequencing to improve the original *M. zebra* genome assembly. It presents the improvements in genome assembly and downstream analysis that could be made with a moderate amount of long read data. Chapter 3 focuses on a *de novo* assembly of *O. niloticus* using high coverage long reads, the improvements made to this genome assembly and how it allowed for the characterization a sex determination system in two species. Chapter 4 refines the *O. niloticus* assembly using a new high-density map and we present a *de novo* anchored assembly of *M. zebra.* Using these two chromosome-scale genome assemblies we are able to characterize a large amount of structural variation between the two genomes, account for the karyotype differences, describe unique and interesting patterns of recombination across the genome, and relate each of these features to several known phenotypes in cichlids. Finally, chapter 5 defines the B chromosome present in Lake Malawi species using both long read sequencing of an individual and short read sequencing of many species. This B chromosome is compared to the B chromosome of Lake Victoria and that points to a possible shared but extremely diverged history of the B chromosomes in both of these lakes.

Chapter 2: An improved genome reference for the African cichlid, *Metriaclima zebra*

Previously published in: **Conte MA** and Kocher TD. An improved genome reference for the African cichlid, *Metriaclima zebra*. *BMC Genomics*. 2015;16(1):724.

2.1 Abstract

2.1.1 Background

Problems associated with using draft genome assemblies are well documented and have become more pronounced with the use of short read data for de novo genome assembly. We set out to improve the draft genome assembly of the African cichlid fish, Metriaclima zebra, using a set of Pacific Biosciences SMRT sequencing reads corresponding to 16.5x coverage of the genome. Here we characterize the improvements that these long reads allowed us to make to the state-of-the-art draft genome previously assembled from short read data.

2.1.2 Results

Our new assembly closed 68% of the existing gaps and added 90.6 Mbp of new non-gap sequence to the existing draft assembly of M. zebra. Comparison of the new assembly to the sequence of several bacterial artificial chromosome clones confirmed the accuracy of the new assembly. The closure of sequence gaps revealed thousands of new exons, allowing significant improvement in gene models. We corrected one known misassembly, and identified and fixed other likely misassemblies. 63.5Mb

(73%) of the new sequence was classified as repetitive and the new sequence allowed for the assembly of many more transposable elements.

2.1.3 Conclusions

Our improvements to the M. zebra draft genome suggests that a reasonable investment in long reads could greatly improve many other comparable vertebrate draft genome assemblies.

2.1.4 Keywords

African cichlid fish, genome assembly, Pacific Biosciences SMRT sequencing, transposable elements.

2.2 Background

Advances in high-throughput genome sequencing have allowed relatively inexpensive genome projects to be conducted for almost any organism. Projects such as the 'Genome 10K Project', which aims to sequence 10,000 vertebrate genomes (41), and the 'Bird 10K' project, which aims to sequence 10,500 bird species (42) have accelerated the production of draft genome sequences. Although attempts have been made to establish standards for declaring a genome sequence 'complete' (21), the quality of draft genomes varies dramatically. The limitations of using these draft genomes for downstream analyses has been documented (7,14). In spite of these limitations, it is clear that such draft genomes will continue to be the basis for genetic research on many species for the foreseeable future.

The use of short (up to several hundred bp) reads has been driven by the desire to reduce costs of DNA sequencing (43). Short read sequencing technologies are appealing, as the cost per base is relatively cheap. However, short reads make the *de novo* assembly process more difficult when the genome contains repeats that exceed the read length, which is typical for even relatively small genomes (44). In addition, sequencing coverage biases caused by variation in base composition and PCR amplification further complicate the task of the assembler (8,9). Many different molecular biology and computational techniques have been developed that attempt to circumvent the problems associated with short read length, while keeping the cost of genome sequencing projects low. One technique is the use of paired-end and mate-pair jumping libraries. The power of this technique was demonstrated when a usable human draft genome assembly was produced using a combination of differently sized short read jumping libraries (180bp to 40kb) with the ALLPATHS-LG assembler (45).

The Assemblathon2 contest was organized as a friendly competition to assess current methods and evaluate the state of genome assembly by providing primarily short read datasets for three different vertebrate genomes. Assemblathon2 demonstrated that there was a lot of variability between submitted assemblies, and still plenty of room for improvement (6). One of the three species used in the Assemblathon2 was the Lake Malawi cichlid fish, *Metriaclima zebra.* African cichlid fish are an ideal system for studying evolutionary mechanisms due to their phenotypic diversity and rapid speciation (31). Draft genomes of *M. zebra* and four other African cichlid fish were recently published (46). According to most assembly metrics, this

M. zebra draft assembly ('M_zebra_v0') was among the best entries submitted to Assemblathon2. However, our extensive use of this assembly has revealed problems with gene models in or near assembly gaps, misassemblies encountered during the course of chromosome walks, and spurious spikes of differentiation statistics near gap or scaffold edges. These problems are likely not unique to this genome project and complicate the use of many similar draft genomes.

To improve the *M. zebra* draft assembly, we generated a 16.5x set of Pacific Biosciences SMRT (Single Molecule, Real-Time) sequencing reads. These 'long' PacBio reads can be used to improve draft assemblies by spanning gaps around repetitive regions and joining contigs and scaffolds (47). Here we set out to improve the M_zebra_v0 genome assembly both to create a better reference assembly for the cichlid research community and to explore the improvements made possible with the addition of 16.5x of PacBio reads to even a relatively good draft vertebrate genome assembly.

2.3 Methods

2.3.1 Overview

Our new 'M_zebra_UMD1' assembly is based on the recently published M_zebra_v0 assembly (46). Misassemblies in the M_zebra_v0 assembly were identified as regions poorly supported by the existing Illumina mate-pair libraries. The assembly was 'broken' at these locations. A newly generated 16.5x coverage PacBio read set was error-corrected to improve base accuracy and identify potentially chimeric reads. These corrected PacBio reads were then used to fill in gaps and to join together scaffolds in the broken M_zebra_v0 assembly. The new M_zebra_UMD1 assembly was then evaluated by comparison to the sequence of individual bacterial artificial chromosome (BAC) clones, alignment of independently assembled transcriptomes, and calculation of assembly completeness and likelihood statistics. Figure 2.1 provides an overview of this assembly process with several assembly statistics shown at each step. Additional details of the steps in this process are discussed below.

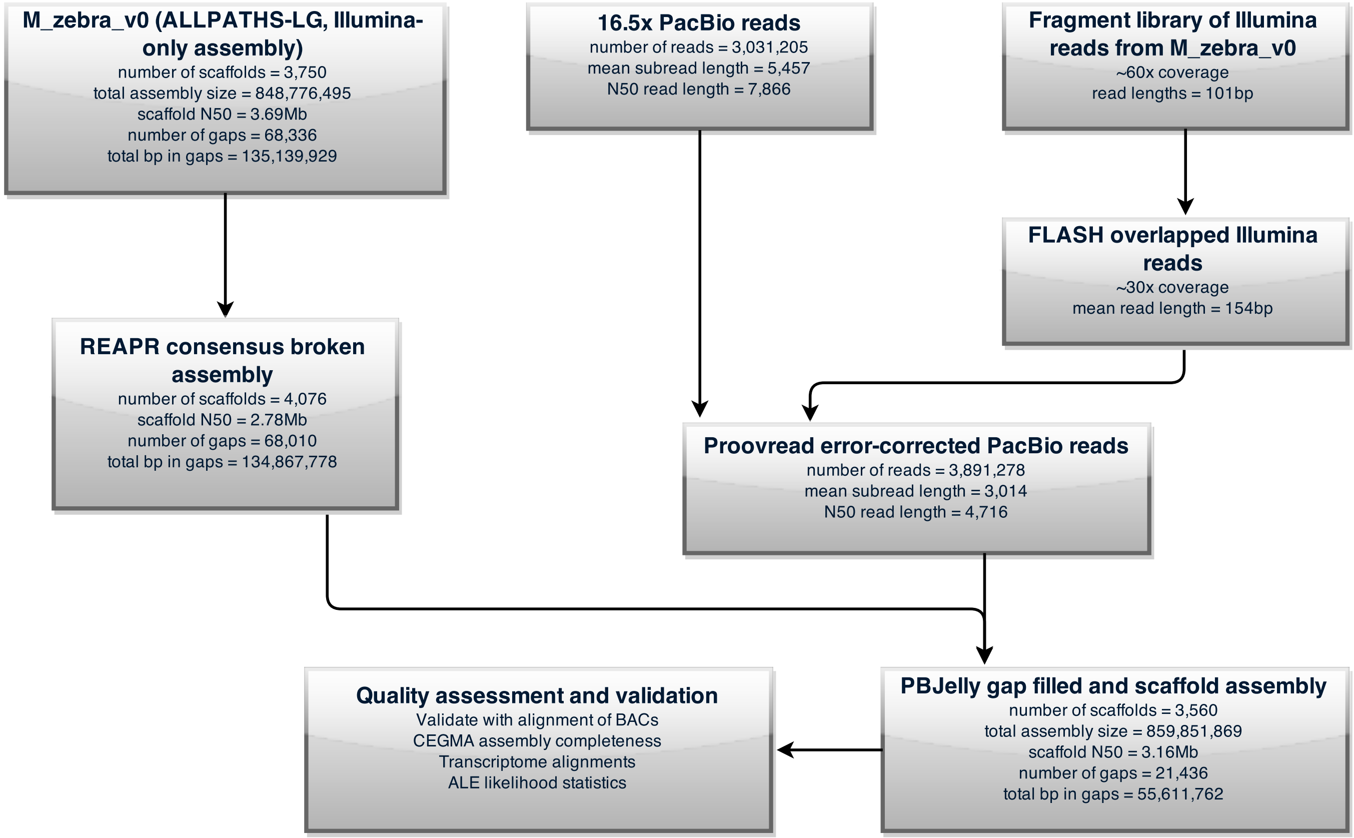

M_zebra_v0 (ALLPATHS-LG, Illumina-only assembly)
number of scaffolds = 3,750
total assembly size = 848,776,495
scaffold N50 = 3.69Mb
number of gaps = 68,336
total bp in gaps = 135,139,929
16.5x PacBio reads
number of reads = 3,031,205
mean subread length = 5,457
N50 read length = 7,866
Fragment library of Illumina reads from M_zebra_v0
~60x coverage
read lengths = 101bp
FLASH overlapped Illumina reads
~30x coverage
mean read length = 154bp
REAPR consensus broken assembly
number of scaffolds = 4,076
scaffold N50 = 2.78Mb
number of gaps = 68,010
total bp in gaps = 134,867,778
Proovread error-corrected PacBio reads
number of reads = 3,891,278
mean subread length = 3,014
N50 read length = 4,716
Quality assessment and validation
Validate with alignment of BACs
CEGMA assembly completeness
Transcriptome alignments
ALE likelihood statistics
PBJelly gap filled and scaffold assembly
number of scaffolds = 3,560
total assembly size = 859,851,869
scaffold N50 = 3.16Mb
number of gaps = 21,436
total bp in gaps = 55,611,762

Figure 2.1. Genome assembly overview. Input datasets and the various steps involved in the assembly of M_zebra_UMD1 are diagrammed along with relevant metrics provided at each step.

2.3.2 Illumina datasets

The M_zebra_v0 assembly was originally created using seven different Illumina insert size libraries (46) as input to the ALLPATHS-LG assembler (45). Table 2.1 provides details of each of the different Illumina libraries used.

Type	Library size	# of reads	# of bp	Sequence coverage
Fragment	180 +/- 15	597,610,332	60,358,643,532	60x
2-3kb jump	2,218 +/- 363	492,188,542	49,711,042,742	50x
2-3kb jump	2,738 +/- 352	217,999,666	22,017,966,266	22x
5kb jump	4,362 +/- 625	147,317,752	14,879,092,952	15x
7kb jump	6,080 +/- 759	158,260,012	15,984,261,212	16x
9kb jump	8,099 +/- 1,345	143,454,662	14,488,920,862	14x
11kb jump	9,079 +/- 2,388	114,671,088	11,581,779,888	12x
40kb jump	38,038 +/- 4,331	38,364,464	2,762,241,408	2.8x
Total		1,909,866,518	191,783,948,862	192x

Table 2.1 - Illumina insert libraries used for the original M_zebra_v0 ALLPATHS-LG assembly and here for REAPR breaking.

2.3.3 REAPR consensus breaking

Recognizing Errors in Assemblies using Paired Reads (REAPR) is a tool that uses paired-read libraries to evaluate genome assembly accuracy, flag regions with potential errors, and break incorrectly joined scaffolds (48). We ran REAPR version 1.0.17 on the M_zebra_v0 assembly using each of the libraries in Table 2.1 separately. First, the REAPR '*smaltmap'* task was run to align each of the libraries to the M_zebra_v0 assembly using SMALT version 0.7.6. SMALT is the recommended aligner for REAPR as it allows for reads in a pair to be mapped independently and not be forced to map at an expected insert distance, which is an important factor for identifying potential misassemblies. The alignments for the two separate 2-3kb libraries listed in Table1 were merged using the '*samtools merge*' command. The REAPR '*perfectfrombam*' task was run on the SMALT alignment of the short-insert fragment library to generate read-depth information and identify repetitive regions. The REAPR '*pipeline*' task was then run separately for each of the jump libraries. The high-quality short-insert alignment from the '*perfectfrombam*' task was supplied to the '*pipeline*' task for each of the jumping libraries. Aggressive breaking ('*-break a=1'*) was also performed as it breaks scaffolds at regions where the fragment coverage distribution is low and potentially misassembled. The output of the REAPR '*pipeline*' task includes the locations where REAPR broke the M_zebra_v0 assembly. Locations in the M_zebra_v0 assembly that were broken by a majority (four or more) of the insert libraries were compiled and the M_zebra_v0 assembly was broken based on this consensus. A Venn diagram of the overlap of REAPR breaks between the libraries (Figure 2.2) was created using jvenn (49).

In addition to breaking the M_zebra_v0 assembly using REAPR, we also randomly broke the assembly to evaluate how well random breaks could be put back together with the PacBio reads. The M_zebra_v0 assembly was randomly broken the same number of times as we broke the assembly according to the REAPR breaking above.

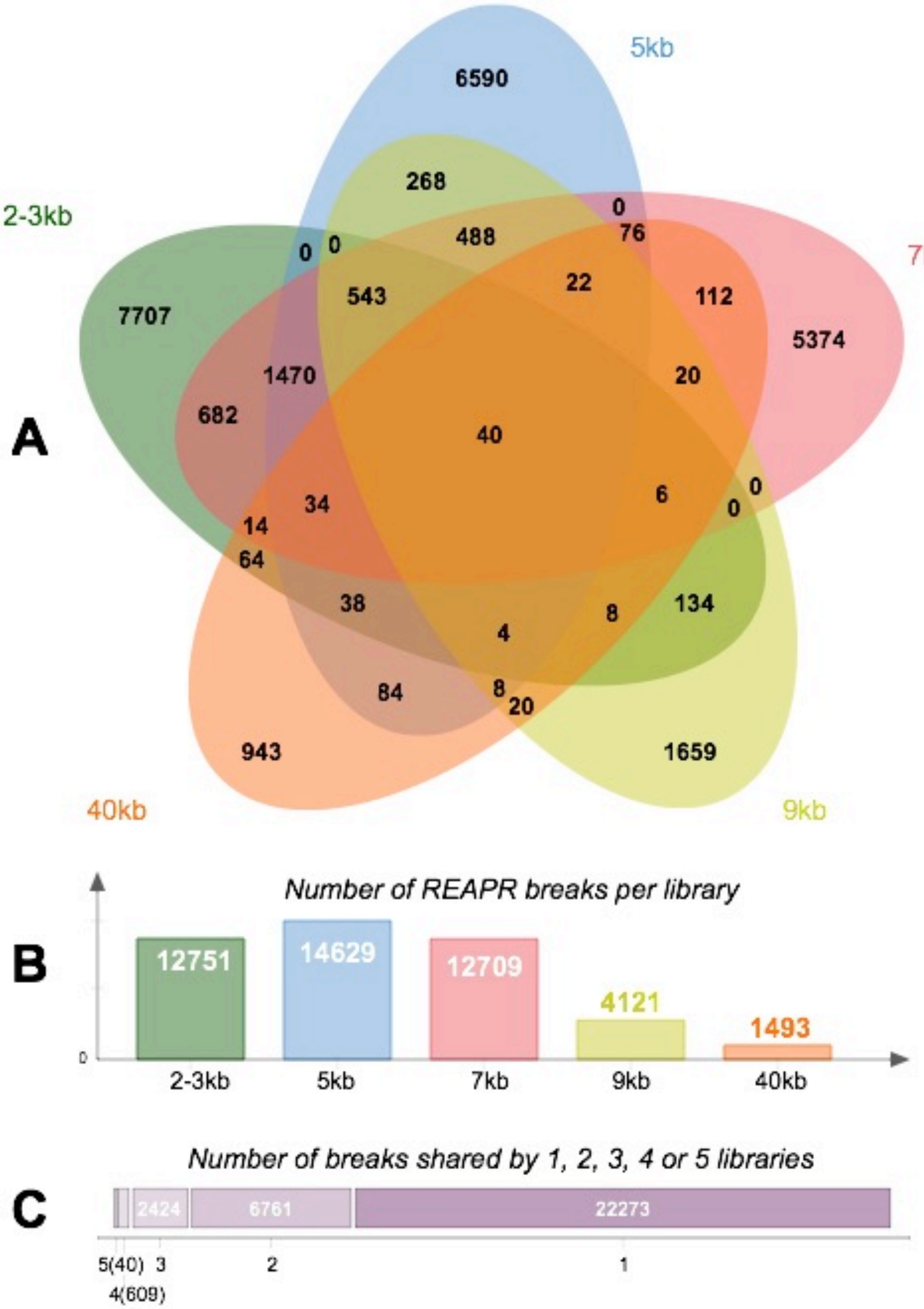

A
5kb
2-3kb
7kb
40kb
9kb
6590
268
0
488
76
0
0
543
22
112
7707
5374
1470
20
682
40
6
0
0
34
14
64
38
134
8
4
84
8
20
943
1659
B
Number of REAPR breaks per library
0
12751
14629
12709
4121
1493
2-3kb
5kb
7kb
9kb
40kb
C
Number of breaks shared by 1, 2, 3, 4 or 5 libraries
2424
6761
22273
5(40)
3
2
1
4(609)

Figure 2.2. Overlap and number of REAPR breaks with different sized Illumina insert libraries. A) Venn diagram showing the overlapping REAPR breaks generated by each of the different Illumina insert libraries provided in Table 2.1. B) Histogram showing the total number of breaks for each library. The 11kb Illumina library was omitted as it produced far more breaks (35,135) than the other libraries and was less complex overall. C) Chart showing the number of REAPR breaks shared by a particular number of libraries (40 breaks shared by all 5 libraries, 609 shared by 4 libraries, etc.)

2.3.4 Pacific Biosciences SMRT sequencing

The Qiagen MagAttract HMW DNA kit was used to extract high-molecular weight DNA from a nucleated blood cell sample from a new individual from the same population used for the Broad Institute sequencing project. Size selection was performed at the University of Maryland Genomics Resource Center using a Blue Pippin pulse-field gel electrophoresis instrument. A library was constructed and 24 SMRT cells were sequenced on their PacBio RS II using the P5-C3 chemistry.

2.3.5 Proovread error correction

Proovread is a hybrid error correction pipeline for correcting PacBio SMRT reads using short read data (50). This step is important as the raw PacBio subreads are only ~85% accurate (51) and we obtained only a modest 16.5x coverage set of reads. The PacBio subreads also contain chimeric reads at a rate of 1-2% (52). Proovread corrects PacBio reads to a high accuracy, detects and clips potentially chimeric reads, and also identifies previously undetected SMRTBell adapter sequences in the PacBio subreads (designated as “siameric” sequences within Proovread).

As shown in Figure 2.1, we used the ~60x Illumina fragment library for Proovread error correction. This Illumina library was designed so that pairs would overlap and slightly longer reads could be generated. We first trimmed and filtered these reads using Trimmomatic version 0.32 with the following settings: *ILLUMINACLIP:TruSeq2-PE.fa:2:30:10 SLIDINGWINDOW:4:20 LEADING:10 TRAILING:10 CROP:101 HEADCROP:0 MINLEN:80*. The adaptor sequences used in the *TruSeq2-PE.fa* file are provided in Additional File 1. We then used FLASH (53) version 1.2.11 with a mis-match density of 0.15 (*-x 0.15*) to overlap the trimmed reads. These trimmed, filtered and overlapped Illumina reads were used for error correction with Proovread. The Proovread ‘SeqChunker’ tool was used to split the 3,031,205 PacBio subreads into 128 similarly sized files to run Proovread on our cluster. Proovread version 2.10 was run with the following BWA mem ‘bwa-pre’ configuration settings: *-k 12 -W 20 -w 40 -r 1 -D 0 -y 20 -A 5 -B 11 -O 2,1 -E 4,3 -T 2.5 -L 30,30* and the following BWA mem ‘bwa-finish’ configuration settings: *-k 17 -W 18 -w 40 -r 1 -D 0 -y 20 -A 5 -B 11 -O 2,1 -E 4,3 -T 3.5 -L 30,30.*

2.3.6 Gap closure and scaffolding with PBJelly

PBJelly is a pipeline for upgrading genome assemblies using PacBio reads (47). PBJelly version 14.9.9 was run using the error corrected PacBio reads as described above. The initial PBJelly ‘setup’ step was run with the ‘--minGap’ parameter set to 19 to reflect the smallest gap size in the M_zebra_v0 assembly. ThePBJelly ‘mapping’ step aligned the corrected PacBio reads to the consensus REAPR broken M_zebra_v0 assembly using BLASR (54) version 1.3.1.127046 and the following parameters: *-minMatch 8 -minPctIdentity 70 -bestn 1 -nCandidates 20 -maxScore -500 –noSplitSubreads.* The PBJelly ‘assembly’ step was run with the ‘--maxWiggle' parameter set to 2000 to account for predicted gap size error in the M_zebra_v0 assembly. The other PBJelly steps (‘support’, ‘extraction’, ‘output’) were run with default parameters.

2.3.7 Quality assessment and validation

GMAP (55) version 2014-12-06 was used to align existing RNA-seq transcriptome assemblies of eleven *M. zebra* tissues. The transcriptome assemblies were created using Trinity (56) as part of the cichlid genome project (46) and made available as supplementary information (57).

Three BAC clones that were previously sequenced and assembled using Sanger technology were aligned to the existing and newly produced assemblies for validation. These published BACs correspond to several opsin gene loci: SWS2A/SWS2B/LWS (GenBank accession JF262084.1, 107.6kbp), SWS1 (GenBank accession JF262085.1, 77.6kbp), and RH2B/RH2A (GenBank accession JF262089.1, 83.5kbp) (58). The BAC sequences were aligned to the corresponding M_zebra_v0 and M_zebra_UMD1 assembly sequences using Gepard (59) version 1.30 to create dotplots for comparison.

Completeness of the intermediate and final M_zebra_UMD1 assemblies was assessed using CEGMA (60) version 2.5 optimized for vertebrate genomes (--vrt).

CEGMA relied on GeneWise version 2.4.1, HMMER version 3.1b1, and NCBI BLAST+ version 2.2.29+. The 248 mostly highly conserved core eukaryotic gene set provided by CEGMA was used.

The likelihoods of the intermediate and final M_zebra_UMD1 assemblies were evaluated using ALE (61). Each of the Illumina libraries were aligned to the assemblies using Bowtie2 (62) version 2.0.2 with the '--very-sensitive' preset parameter. The uncorrected PacBio reads were aligned to assemblies with BLASR version 1.3.1.127046 using the same parameters used above with PBJelly and the '-sam' option to produce a SAM file for input to ALE. ALE was then run on each of the respective alignment files to produce likelihood and mapping statistics for each library.

Summary statistics of the assemblies were compiled using the assemblathon_stats.pl script (63).

2.3.8 RepeatMasker comparisons

RepeatModeler (64) version open-1.0.8 was used to identify and classify *de novo* repeat families in each of the respective assemblies. To obtain a reasonable comparison, RepeatModeler was run using both the M_zebra_v0 and M_zebra_UMD1 assemblies separately. The consensus repeat sequences generated by RepeatModeler for each assembly were combined with the Repbase RepeatMasker library version 20140131. RepeatMasker (65) version open-4.0.5 was run with NCBI/RMBLAST version 2.2.27+ using the '-lib' option to specify the respective RepeatModeler and Repbase combined library so that repeats predicted for

M_zebra_v0 were modeled using the M_zebra_v0 assembly and repeats predicted for M_zebra_UMD1 were modeled using the M_zebra_UMD1 assembly.

2.4 Results and Discussion

2.4.1 REAPR consensus breaking identifies misassemblies in M_zebra_v0

A genetic linkage map of *M. zebra* consisting of 834 RAD-tag markers was previously constructed (66). Comparison of this map to the original M_zebra_v0 assembly identified a misassembly on the largest scaffold (scaffold_0). Table 2.2 shows the alignment of scaffold_0 to markers on two separate constructed linkage groups (LG7 and LG14) within this genetic map, identifying the misassembly. Based on the map data we narrowed the location of the misassembly to a 1.7Mbp region between 3,426,502 (LG14) and 5,124,400 (LG7) on scaffold_0.

Marker name	**Linkage Group**	**Map Position (cM)**	**Position on Scaffold 0**
33761	14	8.093	29,187
36558	14	7.385	169,879
12821	14	14.980	821,093
36086	14	9.480	937,855
47854	14	3.352	1,085,027
32200	14	2.455	1,988,503
55726	14	6.711	3,426,502
MZ371	7	64.131	5,124,400
Ed1012	7	58.564	13,037,865
UNH973	7	55.946	15,726,268

Table 2.2 - Genetic markers that map to scaffold 0 of the M_zebra_v0 assembly. Markers on LG7 and LG14 are ordered by their position aligned to scaffold_0 of M_zebra_v0.

Within this 1.7Mbp region there was a 19bp gap at scaffold_0:3,622,144 where REAPR also predicted a misassembly for 5 out of the 6 Illumina insert libraries listed in Table 2.1. The 40kb library was the only library where REAPR did not predict a misassembly. The 40kb library was also the only jumping library that had mate-pairs that properly spanned this gap. REAPR predicted a misassembly at this gap for the other 5 jumping libraries either because they did not have spanning mate-pairs, had mate-pairs improperly oriented, and/or had mate-pairs aligning at a distance much different than the expected insert size. This small 19bp gap also had no PacBio reads that spanned it. It is likely that this is the exact location of the misassembly identified by the genetic map data.

In addition to this known misassembly, REAPR identified many additional putative misassemblies in the M_zebra_v0 assembly. Figure 2.2 shows the number of breaks that REAPR predicted using the Illumina insert libraries listed in Table 2.1. Inspection of paired-read mappings from the 11kb library revealed that it was much less complex than any of the other libraries. Using this library, REAPR broke the M_zebra_v0 assembly 35,135 times. This was far more times than any other library and more than twice that of the 5kb library REAPR breaks (14,629 breaks). We elected to remove this 11kb library from subsequent analyses.

The number of REAPR breaks shared by 5, 4, 3, 2 or 1 libraries was 40, 649, 3073, 9835 and 32107 respectively (Figure 2.2). To begin our reassembly process we had to choose the most appropriate number of REAPR breaks of the M_zebra_v0 assembly. Breaking the assembly too few times could leave unidentified misassemblies, while breaking too many times would fragment the assembly more than necessary. PacBio provides the SMRT View tool (67) for visualizing PacBio read alignments created using their BridgeMapper SMRT Pipe module within the SMRT-Analysis software suite (68). The BridgeMapper module creates split read alignments with BLASR that can be used to identify misassemblies. Using these tools we were able to manually inspect the PacBio split read alignments and estimate that there are ~200-1000 misassemblies in the M_zebra_v0 assembly.

We also evaluated the rate of false positive breaks by quantifying the number of REAPR breaks that could be re-joined with PBJelly and the corrected PacBio reads. For the M_zebra_v0 assembly that was broken randomly, 541/649 (83.4%) of the breaks were reassembled in the original M_zebra_v0 assembly order. In contrast, only 75 (11.6%) of the 649 REAPR breaks were reassembled in the original M_zebra_v0 order. The random breaks are reassembled in the original order about

82% of the time (Table 2.3). The percentage of REAPR breaks that are reassembled increases as the number of breaks increases, but is still far from the percentage of random breaks that are rejoined. It is clear that the consensus REAPR breaks have identified regions of the M_zebra_v0 assembly that were poorly supported and often misassembled. These regions are difficult to reassemble even with the corrected PacBio reads and likely represent complex and highly repetitive regions of the genome.

Number of shared libraries	Number of breaks	REAPR breaks reassembled in M_zebra_v0 order	Random breaks reassembled in M_zebra_v0 order
5 out of 5	40	3 (7.5%)	33 (82.5%)
4 out of 5	649	75 (11.6%)	541 (83.4%)
3 out of 5	3,073	509 (16.6%)	2,530 (82.3%)
2 out of 5	9,835	2,135 (21.7%)	8,024 (81.6%)
1 out of 5	32,107	8,225 (25.6%)	25,389 (79.1%)

Table 2.3 - REAPR and random breaks reassembled.

Based on the manual inspection of split read alignments and the rate of false positive breaks that were introduced we chose to break the M_zebra_v0 assembly wherever REAPR had predicted a misassembly in 4 or more of the Illumina insert libraries. This resulted in an assembly that was broken 649 times (40 breaks found in 5 or more libraries plus 609 breaks found in 4 or more libraries, Figure 2.2).

2.4.2 Proovread error correction

We generated a 16.5x set of PacBio reads using with the P5-C3 chemistry. However, these PacBio reads are error prone (80-85% accuracy (8)) and known to contain chimeric reads at a rate higher than 1% (52). In addition, the SMRTbell adapter sequences are not always removed properly and may persist in 1% to 5% of filtered PacBio subreads (Thomas Hack, personal communication). These particular sequences are deemed "siameric" reads because they contain twin reads connected by the adapter. To detect and clip both chimeric and siameric reads as well as improve the base-level accuracy of the PacBio reads we ran Proovread (50). The ~60x short-insert Illumina library was first overlapped to produce longer reads (mean overlapped read length = 154bp, ~30x coverage) which were then used for the Proovread error-correction (Figure 2.1). Table 2.4 provides summary statistics of the PacBio reads before and after the Proovread error-correction. While the mean and N50 read length decreased, chimeric and siameric reads were detected at the expected rates and corrected by Proovread. There was a tradeoff between having longer PacBio reads with a small percentage of chimeric reads or somewhat shorter but error-corrected PacBio reads. We chose to remove the chimeric reads and use the set of shorter and error-corrected PacBio reads, especially considering the modest 16.5x coverage and the potential for chimeric/siameric introductions into the assembly.

	Uncorrected	**Proovread corrected**
Number of reads	3,031,205	3,891,278
Mean read length	5,457	3,014
N50 read length	7,866	4,716
Number of chimeric reads detected	119,924 (3.95%)	-
Number of siameric reads detected	37,836 (1.25%)	-

Table 2.4 - PacBio read statistics before and after Proovread error-correction.

2.4.3 Gap filled assembly

Once the known and putative misassemblies were broken, and the errors in the PacBio reads were corrected, the M_zebra_v0 assembly was ready to be improved using PBJelly. Table 2.5 provides summary statistics of three assemblies: 1) the original M_zebra_v0 draft assembly, 2) M_zebra_v0 after being broken 649 times by REAPR, 3) and the broken assembly after gap-filling with PBJelly using the corrected PacBio reads (M_zebra_UMD1).

Assembly	**M_zebra_v0**	**REAPR broken**	**M_zebra_UMD1**
Number of scaffolds	3,750	4,076 (+8.69%)	3,560 (-5.07%)
Total size of scaffolds	848,776,495	848,503,369 (-0.03%)	859,851,869 (+1.3%)
Longest scaffold	18,958,539	12,137,054 (-35.98%)	14,997,410 (-20.89%)
Mean scaffold size	226,340	208,171 (-8.03%)	241,531 (+6.71%)
N50 scaffold length	3,699,709	2,783,035 (-24.78%)	3,158,421 (-14.63%)
NG50* scaffold length	3,007,690	2,252,862 (-25.10%)	2,555,048 (-15.05%)
Scaffold %N	15.93	15.9 (-0.19%)	6.47 (-59.38%)
Number of gaps	68,336	68,010 (-0.48%)	21,436 (-68.63%)
Non gap bp	713,636,566	713,635,591 (~0.00%)	804,240,107 (+12.70%)
Total gap bp	135,139,929	134,867,778 (-0.2%)	55,611,762 (-58.85%)
Number of exons mapped	4,492,869	4,492,551 (-0.01%)	4,591,788 (+2.20%)

Table 2.5 - Assembly summary statistics *NG50 assumes genome size of 1.0Gb. Percentage change values in parenthesis are relative to M_zebra_v0.

Most of the 649 REAPR breaks occurred at gaps. REAPR typically broke the M_zebra_v0 assembly twice, once on each side of the gap, generating 326 more scaffolds. Many of these broken scaffolds were put back together with the corrected PacBio reads in the new M_zebra_UMD1 assembly. The new assembly has 190 (5%) fewer scaffolds relative to M_zebra_v0, and 516 (12.7%) fewer scaffolds relative to the REAPR broken assembly. These may not seem like sizeable differences, but the M_zebra_v0 assembly was scaffolded using a ~40kb jumping library, with a mean insert size (38,038bp) that is longer than the longest error-corrected PacBio read in our dataset (33,000bp). Therefore, since the M_zebra_v0 assembly was already relatively well placed into scaffolds, we did not see a large reduction in the number of scaffolds. We expect that draft assemblies that do not include mate pair libraries at this scale will experience a greater improvement using the long PacBio reads.

The total length of the M_zebra_UMD1 assembly increased by 11.1Mbp (+1.3%). However, this leaves out the fact that 79.5Mb of gaps were filled, for a total of 90.6Mb of new sequence. The total length of the assembly contained in gaps decreased from 15.93% to 6.47% of the assembly length, a 59% improvement. The number of gaps decreased by 70%, from 68,336 to 21,436. Further assembly metrics are provided in Additional File 2.

We mapped existing transcriptome assemblies from 11 tissues of *M. zebra* (46) to each of the genome assemblies using GMAP. The total number of mapped exons increased by 98,919 (+2.20%). This count includes exons that are present in multiple transcript isoforms and are thus counted multiple times.

2.4.4 Assembly completeness

To assess the completeness of the assemblies we ran CEGMA (60), which scores the presence of 248 core eukaryotic genes (CEGs) in a given assembly. Table 2.6 provides the CEGMA completeness report for both the original M_zebra_v0 and the new M_zebra_UMD1 assemblies. The total number of complete plus partial CEGs is the same in both assemblies (237). However, the new M_zebra_UMD1 assembly contains 7 (2.6%) more complete CEGs than the original M_zebra_v0 assembly. This increase in complete CEGs can be attributed to filling gaps that occur within gene models. One example of this was seen in the assembly of the predicted piwi-like protein (NCBI accession XM_004544701.1). Figure 2.3 shows this piwi-like RefSeq mRNA sequence aligned to the M_zebra_UMD1 assembly. When the transcriptome assemblies were mapped to the M_zebra_UMD1 assembly, it became evident that the gaps in the original M_zebra_v0 assembly had left out at least 10 of the exons in this piwi-like protein.

Assembly	**M_zebra_v0**	**M_zebra_UMD1**
Complete CEGs	227 (91.53%)	233 (93.95%)
% Of complete CEGs with multiple orthologs	25.55	26.61
Complete + Partial CEGs	237 (95.56%)	237 (95.56%)
% Of complete + partial CEGs with multiple orthologs	28.69	29.96
Total complete CEGs including putative orthologs	302	314
Average number of orthologs per complete CEG	1.33	1.35
Total complete + partial CEGs including putative orthologs	331	338
Average number of orthologs per complete + partial CEG	1.4	1.43

Table 2.6 - Summary of CEGMA results.

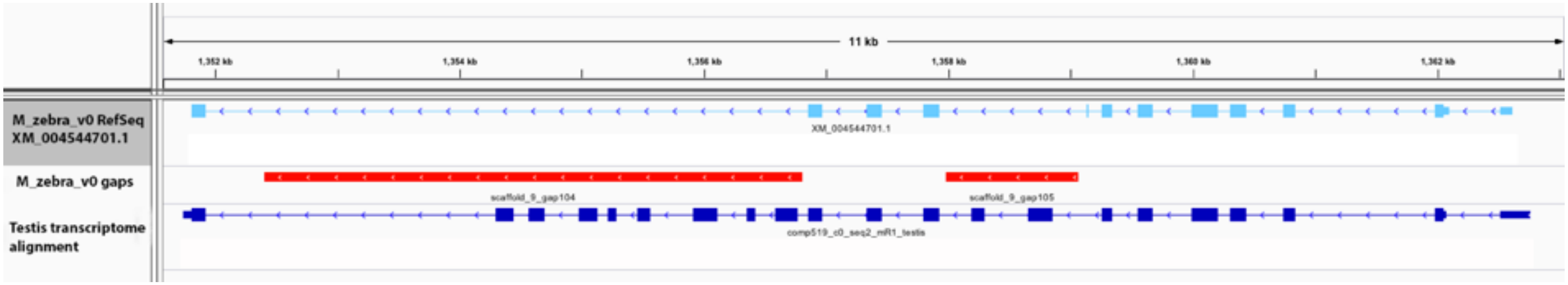


Figure 2.3 - Gap filling improves gene models. The top (light-blue) track shows the original RefSeq gene model (XM_004544701.1) based on the M_zebra_v0 assembly aligned to the M_zebra_UMD1 assembly. The middle (red) track indicates the location of the gaps (now filled) in the original M_zebra_v0 assembly. The bottom (blue) track shows the testis transcriptome assembly aligned to M_zebra_UMD1 assembly and the additional 10 exons that were originally in gaps in the assembly.

The new M_zebra_UMD1 assembly contains an increased number of CEGs that have multiple orthologs (62, increased from 58). These orthologs were collapsed in the M_zebra_v0 assembly and have been separately assembled in the M_zebra_UMD1 assembly. Extrapolated across the genome, the difference in the number of genes with multiple orthologs amounts to hundreds of new genes.

2.4.5 Comparison with BACs from opsin loci

Three *M. zebra* BAC clones previously sequenced and assembled using Sanger technology (58) were used to evaluate the accuracy of the error-correction and gap-filling procedures. Figure 2.4 shows dotplot alignments of these sequenced BACs to both the M_zebra_v0 and M_zebra_UMD1 assemblies. Most of the gaps in the M_zebra_v0 assembly have been filled in the M_zebra_UMD1 assembly. Several small gaps remain in the M_zebra_UMD1 assembly, as can be seen in Figure 2.4B and 4D. BAC clone JF262085.1 (encompassing the SWS1 opsin) was the only BAC of the three that had gaps in the original assembled BAC sequence. The incongruence in the lower left portion of the Figure 2.4D dotplot represents a difference in the size of the gap between the JF262085.1 BAC and the M_zebra_UMD1 assemblies. The abnormal alignment in the upper right portion of the dotplot in Figure 2.4D represents a small 20bp gap in the M_zebra_v0 assembly that has been “overfilled” by PBJelly with 779 bases. Both of these differences likely represent some structural sequence variation between the individual fish used for the BAC, M_zebra_v0 and M_zebra_UMD1 sequencing.

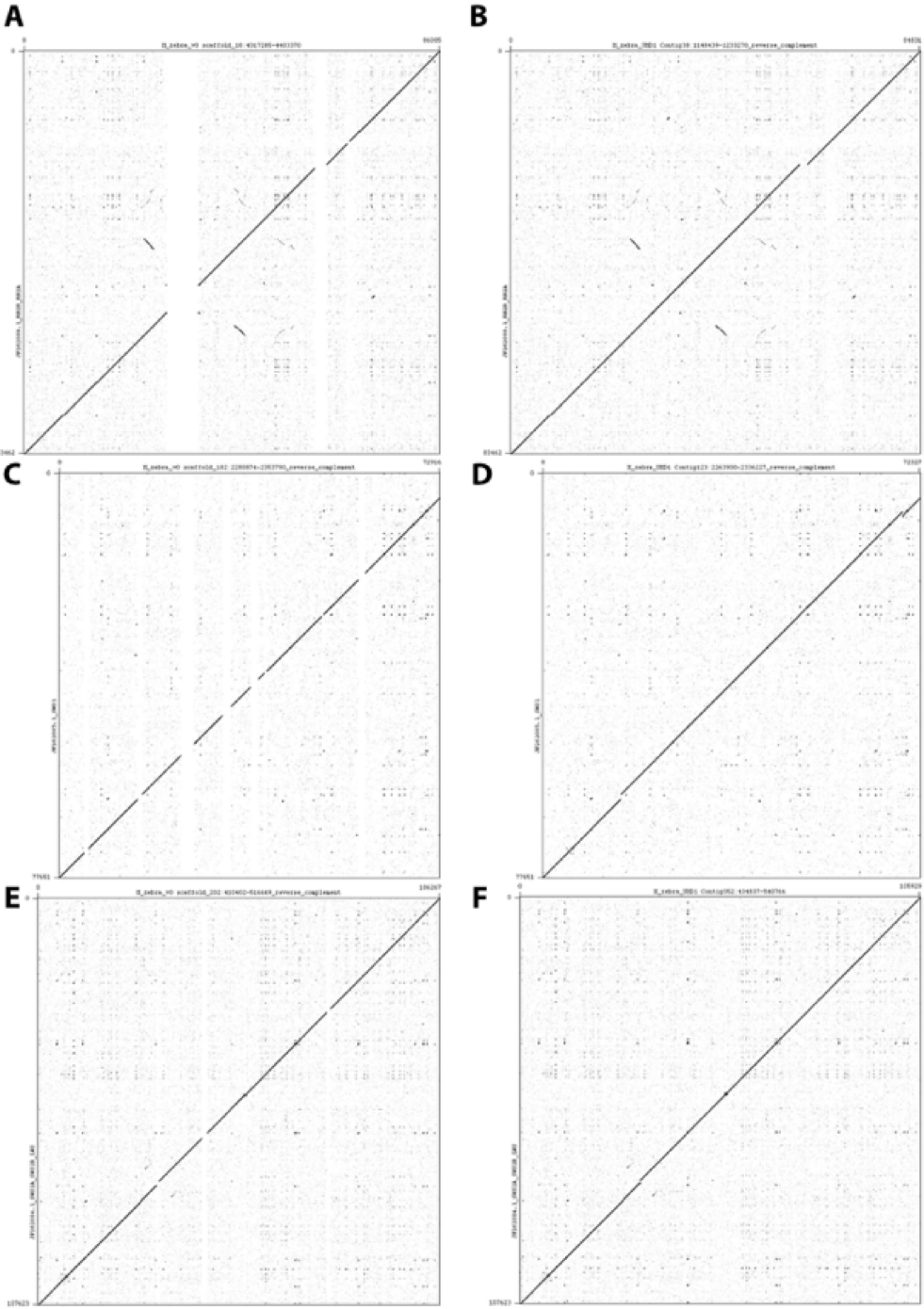

A
0
M_zebra_v0 scaffold_10:4317285-4403370
86085
0
83462
B
0
M_zebra_UMD1 Contig30:1148439-1233270_reverse_complement
84831
0
83462
C
0
M_zebra_v0 scaffold_102:2280874-2353790_reverse_complement
72916
0
77681
D
0
M_zebra_UMD1 Contig123:2263900-2336227_reverse_complement
72327
0
77681
E
0
M_zebra_v0 scaffold_202:420402-526669_reverse_complement
106267
0
107623
F
0
M_zebra_UMD1 Contig352:434837-540766
105929
0
107623

Figure 2.4 - Dotplot alignments of opsin BACs to M_zebra_v0 and M_zebra_UMD1 to validate filled gap sequence. RH2B/RH2A (JF262089.1) versus M_zebra_v0 (A) and M_zebra_UMD1 (B). SWS1 (JF262085.1) versus M_zebra_v0 (C) and M_zebra_UMD1 (D). SWS2A/SWS2B/LWS (JF262084.1) versus M_zebra_v0 (E) and M_zebra_UMD1 (F).

2.4.6 Assembly likelihood

The assembly summary metrics provided in Table 2.5 indicate the new M_zebra_UMD1 assembly is better in all respects except maximum scaffold length (-21%) and scaffold N50 (-15%). However, these decreases in continuity are accompanied by an overall improvement in accuracy and completeness of the assembly. To further quantify the accuracy of the new assembly we ran the Assembly Likelihood Evaluation (ALE) program (61). This tool integrates read quality, mate-pair orientation, insert size, coverage and k-mer frequencies to provide a statistical measurement of assembly quality. Table 2.7 provides a summary of the ALE metrics calculated for several different read sets against both the M_zebra_v0 and M_zebra_UMD1 assemblies. The overall ALE likelihood score itself is not intended to be used to compare assemblies created from different datasets as is the case for the M_zebra_v0 (Illumina only) and M_zebraUMD1 (Illumina + PacBio). However, the remaining assembly metrics provided in the ALE output are very useful for comparison. For each Illumina library, the total number of placed reads is greater, the number of unmappable bases is lower, the number of unmappable regions is lower and the number of bases with 0 coverage is less in the M_zebra_UMD1 assembly compared to the M_zebra_v0 assembly. For brevity, only 3 of the 7 Illumina libraries are shown in Table 2.7, but the remaining libraries show the same trends (Additional File 3). A surprising amount of the genome had bases with 0 coverage alignment for the Illumina libraries. Some of this can be explained by the 55.6Mbp of gaps that remain in the M_zebra_UMD1 assembly, since ALE calculates gaps as bases with 0 coverage. The remaining ~66Mbp with 0 coverage (short-insert and 2-3kb Illumina libraries in Table 2.7) is mostly covered by the PacBio library. The ~10Mbp with 0 coverage for the PacBio library reflects regions where the library either did not have any reads by chance or where only the Illumina libraries were able to sequence through. Additional PacBio coverage will help to more precisely describe such regions.

	M_zebra_v0	**M_zebra_UMD1**
Illumina short insert library		
#Total Placed Reads	384,925,943	390,482,375
# Unmappable Bases	132,637,543	57,405,631
# Unmappable Regions	57,998	14,063
Bases with 0 Coverage	139,693,095	121,246,622
Illumina 2-3kb insert library		
#Total Placed Reads	320,493,115	341,717,744
# Unmappable Bases	133,188,276	56,563,974
# Unmappable Regions	58,324	14,069
Bases with 0 Coverage	143,109,574	121,181,395
Illumina 40kb insert library		
#Total Placed Reads	20,487,153	22,971,340
# Unmappable Bases	144,670,975	60,104,659
# Unmappable Regions	73,341	25,254
Bases with 0 Coverage	518,909,366	492,713,889
16.5x PacBio library		
#Total Placed Reads	2,703,712	2,794,402
Average Read Length	3,772	4,258
Average Read Overlap	3,453	3,886
# Unmappable Bases	82,472,941	45,023,176
# Unmappable Regions	18,363	6,349
Bases with 0 Coverage	114,035,849	65,141,623

Table 2.7 - Summary of assembly likelihood (ALE) results.

2.4.7 Analysis of transposable elements and repetitive sequences

A large amount of the sequence that was added in the new M_zebra_UMD1 assembly is composed of repetitive sequences and transposable elements that were either collapsed or not assembled in the original M_zebra_v0 assembly. We analyzed the total amount of repetitive sequences in both assemblies to understand the repeat content of the sequence that was added in M_zebra_UMD1. Table 2.8 lists several of the most abundant transposable element super families in the two assemblies. For most of the transposable element super families, the number of elements increased in the M_zebra_UMD1 assembly. Those transposable elements that decreased in number still increased in total bp, which means that the sequence of individual transposable element copies were longer in the M_zebra_UMD1 assembly. The assemblies of longer repeat copies can be seen for both the DNA hAT-Ac and LINE L1 transposable elements (Figure 2.5). Additional File 4 provides a detailed list of hundreds of transposable elements and low complexity repeats that were annotated in both assemblies.

		M_zebra_v0				**M_zebra_UMD1**				**Δ from M_zebra_v0**			
order	**super-family**	**number**	**bp**	**mean size**	**median size**	**number**	**bp**	**mean size**	**median size**	**number**	**bp**	**mean size**	**median size**
DNA	TcMar-Tc1	133,563	30,394,950 (-3.58%)	227.6	152	137,896	40,100,895 (4.66%)	290.8	173	4,333	9,705,945	63.2	21
	hAT-Ac	41,018	9,251,093 (1.09%)	225.5	143	43,310	16,553,134 (1.93%)	382.2	215	2,292	7,302,041	156.7	72
LINE	L1	9,184	3,265,323 (0.38%)	355.5	190	11,186	7,488,720 (0.87%)	669.5	318.5	2,002	4,223,397	313.9	128.5
	L2	65,651	14,708,900 (1.73%)	224.0	148	62,048	18,525,102 (2.15%)	298.6	168	-3,603	3,816,202	74.5	20
	Rex-Babar	25,685	6,087,899 (0.72%)	237.0	139	30,109	14,508,668 (1.69%)	481.9	202	4,424	8,420,769	244.8	63
LTR	Gypsy	10,865	3,908,793 (0.46%)	359.8	159	14,026	6,476,548 (0.75%)	461.8	184	3,161	2,567,755	102.0	25
	Ngaro	3,955	393,178 (0.05%)	99.4	94	10,633	1,841,475 (0.21%)	173.2	157	6,678	1,448,297	73.8	63
SINE	MIR	12,756	1,741,837 (0.21%)	136.6	111	10,900	2,395,459 (0.28%)	219.8	165	-1,856	653,622	83.2	54
	tRNA-Core	7,419	953,921 (0.11%)	128.6	124	12,054	1,819,302 (0.21%)	150.9	145	4,635	865,381	22.4	21
Unknown		285,700	49,619,702 (5.85%)	173.7	126	279,557	58,688,408 (6.83%)	209.9	138	-6,143	9,068,706	36.3	12
Ancestral repeats		1,101,882	173,081,089 (20.39%)			1,153,935	234,447,039 (27.27%)				61,365,950		
Lineage specific		17,320	4,748,554 (0.56%)			15,585	6,875,733 (0.80%)				2,127,179		
Total		1,119,202	177,829,643 (20.95%)			1,169,520	241,322,772 (28.07%)				63,493,129		

Table 2.8 - Repetitive element summary.

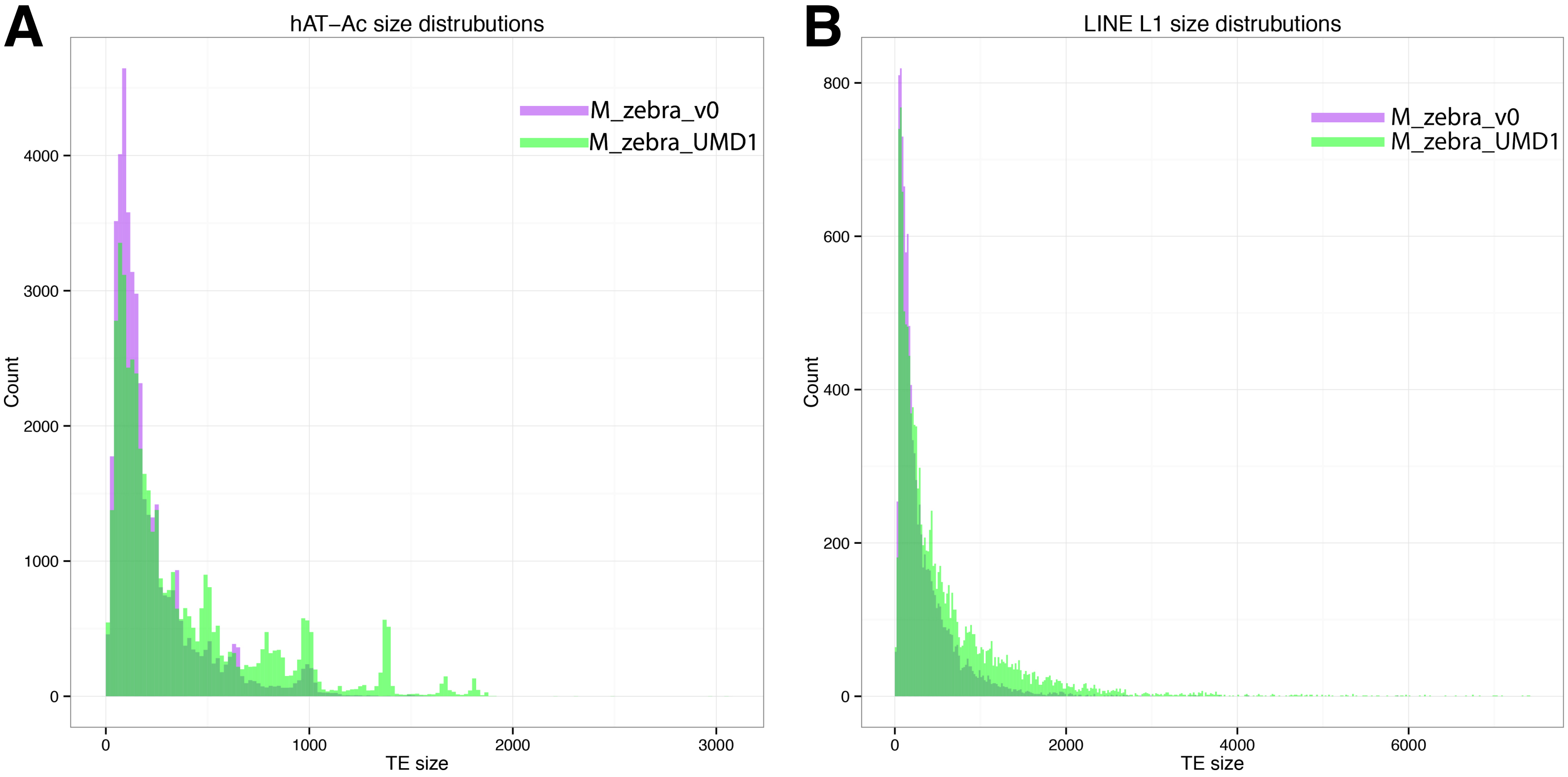


Figure 2.5 - Gap filling improves both number and length of transposable element sequences. A) Distribution of the size of DNA hAT-Ac transposable elements in the two assemblies. B) Distribution of the size of the LINE L1 transposable elements in the two assemblies.

The M_zebra_UMD1 assembly had fewer total lineage specific repeats identified (17,320 vs. 15,585), but the total amount of lineage specific repeat bases was higher compared to the M_zebra_v0 assembly (4.7Mbp vs. 6.9Mbp). Again, this shows that longer lineage specific repeats have been assembled in the M_zebra_UMD1 assembly. In terms of total repetitive sequence, the new M_zebra_UMD1 assembly contained 63.5Mbp of additional sequence that was classified as repetitive. This is consistent with the idea that most of the gaps in the original M_zebra_v0 assembly spanned sequences consisting of transposable elements and other repetitive sequences.

2.5 Conclusions

This study reports an improved assembly of the Lake Malawi African cichlid, *M. zebra*. We identified hundreds of misassemblies in the previous draft assembly (46). We then used a newly generated set of 16.5x long PacBio reads to fill in 68% of the previous assembly gaps and join together a portion of the previous scaffolds. This process added 90.6 Mbp of new sequence to the assembly. Some of the newly added sequence contained gene sequence, allowing the identification of thousands of new exons. However, the majority of the newly added sequence was annotated as repetitive (73%). The new data allowed us to assemble many more and longer copies of the transposable elements in the *M. zebra* genome. We hope this study can serve as an example of how a reasonable investment in long-read sequencing can improve even a relatively well assembled vertebrate draft genome.

2.6 Competing interests

The authors declare they have no competing interests.

2.7 Authors' contributions

MAC and TDK conceived the study. TDK extracted HMW DNA for sequencing. MAC performed the computational analysis. MAC and TDK drafted the manuscript.

2.8 Availability of supporting data

The *M. zebra* assemblies are available under NCBI BioProject 'PRJNA198780'. The raw PacBio reads are available under the NCBI SRA accession SRX985423.

2.9 Acknowledgements

We thank Thomas Hackl for the very helpful discussion and guidance in running the Proovread error correction steps. We thank Luke Tallon and Naomi Sengamalay of the Institute for Genome Sciences core facility for providing a high quality PacBio library and sequence reads. We thank the members of the Kocher and Carleton labs at UMD and the UMD Bring Your Own Bioinformatics club for providing thoughtful advice during the course of the project. We thank Karen Carleton, William Gammerdinger and Ian Misner for reading and providing thoughtful feedback on the manuscript. We acknowledge the University of Maryland supercomputing resources (www.it.umd.edu/hpcc/) made available in conducting the research reported in this paper. This work was supported by the National Science Foundation under grant DEB-1143920.

2.10 Additional files

Additional files referenced in this chapter can be found at:
https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1930-5

Chapter 3: A high quality assembly of the Nile Tilapia (*Oreochromis niloticus*) genome reveals the structure of two sex determination regions

Previously published in: **Conte MA**, Gammerdinger WJ, Bartie KL, Penman DJ, Kocher TD. A high quality assembly of the Nile Tilapia (*Oreochromis niloticus*) genome reveals the structure of two sex determination regions. *BMC Genomics*. 2017;18(1):341.

3.1 Abstract

3.1.1 Background

Tilapias are the second most farmed fishes in the world and a sustainable source of food. Like many other fish, tilapias are sexually dimorphic and sex is a commercially important trait in these fish. In this study, we developed a significantly improved assembly of the tilapia genome using the latest genome sequencing methods and show how it improves the characterization of two sex determination regions in two tilapia species.

3.1.2 Results

A homozygous clonal XX female Nile tilapia (*Oreochromis niloticus*) was sequenced to 44X coverage using Pacific Biosciences (PacBio) SMRT sequencing. Dozens of candidate *de novo* assemblies were generated and an optimal assembly (contig NG50 of 3.3Mbp) was selected using principal component analysis of likelihood scores calculated from several paired-end sequencing libraries. Comparison of the new assembly to the previous *O. niloticus* genome assembly reveals that recently

duplicated portions of the genome are now well represented. The overall number genes in the new assembly increased by 27.3%, including a 67% increase in pseudogenes. The new tilapia genome assembly correctly represents two recent *vasa* gene duplication events that have been verified with BAC sequencing. At total of 146Mbp of additional transposable element sequence are now assembled, a large proportion of which are recent insertions. Large centromeric satellite repeats are assembled and annotated in cichlid fish for the first time. Finally, the new assembly identifies the long-range structure of both an ~9Mbp XY sex determination region on LG1 in *O. niloticus*, and a ~50Mbp WZ sex determination region on LG3 in the related species *O. aureus*.

3.1.3 Conclusions

This study highlights the use of long read sequencing to correctly assemble recent duplications and to characterize repeat-filled regions of the genome. The study serves as an example of the need for high quality genome assemblies and provides a framework for identifying sex determining genes in tilapia and related fish species.

3.2 Background

Aquaculture plays an increasingly important role in providing sustainable seafood products and has significantly outpaced capture fisheries in the past several decades (69). Tilapias are among the most important farmed fishes, and tilapia production continues to expand exponentially across the globe (70). An important aspect of commercial production is the control of sexual differentiation. Males grow to marketsize earlier than females. Females also start to reproduce at a smaller size, filling production ponds with small fish (71). It is therefore advantageous to grow-out only male fish. At one time, all-male populations were produced through interspecific crosses (72), but the strains supporting this technology have been lost or contaminated. Currently, the standard way of achieving all male or nearly all male tilapia populations is via hormonal masculinization (71,73). A reliable way of producing genetically all-male tilapia would allow the replacement of hormonal masculinization, which is banned in several major producing countries (although not enforced in most cases). It is therefore important to understand the genetic basis of sex determination in current aquaculture stocks.

Sex determination in tilapias is largely genetic, although environmental factors also play a role (74–76). In Nile tilapia (*Oreochromis niloticus*), distinct XY sex determining loci have been identified on both linkage group (LG) 1 and LG23 (77,78). The closely related blue tilapia (*O. aureus*) segregates both an XY locus on LG1, and an epistatically dominant ZW locus on LG3 (79). Additional sex determining loci have been identified on LGs 5, 7, 13, 18 and 20 in closely related species of East African cichlid (26,80,81). As a group, tilapias and related species of other cichlid fishes are a promising model system for understanding the gene network controlling sex determination in vertebrates.

Work to identify the genes underlying each of these sex determiners has been hampered by the incomplete nature of previous draft genome assemblies, and by the discovery that many of these sex determiners are located in large blocks of highly differentiated, and sometimes repetitive sequence. To date, the molecular genetic basis for sex determination in cichlids has been determined for only the LG23 XY locus in *O. niloticus* (82).

Although several draft genome sequences are available for cichlids, these are mostly based on short Illumina sequencing reads (83). The previous *O. niloticus* assembly was produced using ~277X coverage of Illumina reads from several libraries including a 40kb scaffolding library. Recently duplicated and highly repetitive sequences are typically collapsed in these assemblies (84). Indeed some of the most interesting and perhaps evolutionarily important regions of the genome may be the most difficult to assemble accurately. Recently duplicated regions are notoriously difficult to assemble due to their repeat length and high sequence identity (7). The repetitive "dark-matter" part of the genome is vastly underrepresented in the majority of current genome assemblies (85). Attempts to assemble these regions using only short read sequencing are futile (14). Only long sequencing reads will produce more contiguous and complete assemblies of complex vertebrate genomes (18,86–89). The importance of such high quality assemblies for downstream applications cannot be overemphasized.

Here we report a new assembly of the tilapia genome from long PacBio sequence reads. This assembly contains much of the missing sequence from previous assemblies, and is among the most contiguous vertebrate genome assemblies to date. We use this new assembly to further characterize the tilapia sex determining loci previously identified on LGs 1 and 3.

3.3 Results

3.3.1 Assembly Overview

A homozygous clonal XX female tilapia individual (90) was chosen for genome sequencing. The individual was sequenced to 44X coverage using PacBio sequencing of 63 SMRT cells using the P6-C4 chemistry. This yielded 5,085,371 reads with a mean subread length of 8,747bp and N50 read length of 11,366bp.

An overview of the assembly process is outlined in Figure 3.1. To summarize, 37 candidate *de novo* assemblies were generated using both the FALCON (86) and Canu (91) genome assembly packages. Multiple parameters were adjusted for both algorithms to tune the assemblies. The error correction steps of both algorithms include parameters that control alignment seed length, read length, overlap length and error rates (see Methods).

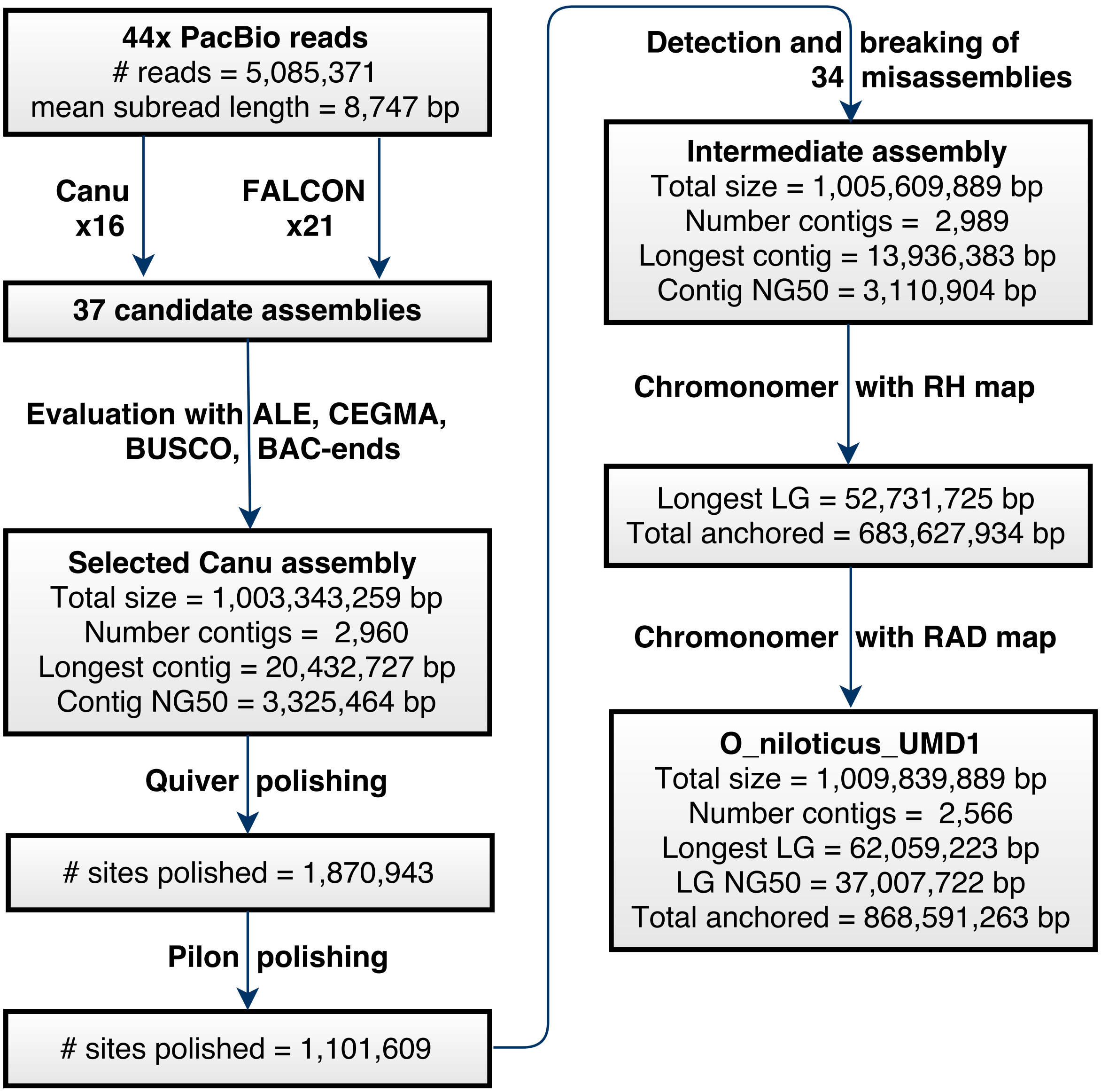

44x PacBio reads
# reads = 5,085,371
mean subread length = 8,747 bp
Canu
x16
FALCON
x21
37 candidate assemblies
Evaluation with ALE, CEGMA,
BUSCO, BAC-ends
Selected Canu assembly
Total size = 1,003,343,259 bp
Number contigs = 2,960
Longest contig = 20,432,727 bp
Contig NG50 = 3,325,464 bp
Quiver polishing
# sites polished = 1,870,943
Pilon polishing
# sites polished = 1,101,609
Detection and breaking of
34 misassemblies
Intermediate assembly
Total size = 1,005,609,889 bp
Number contigs = 2,989
Longest contig = 13,936,383 bp
Contig NG50 = 3,110,904 bp
Chromonomer with RH map
Longest LG = 52,731,725 bp
Total anchored = 683,627,934 bp
Chromonomer with RAD map
O_niloticus_UMD1
Total size = 1,009,839,889 bp
Number contigs = 2,566
Longest LG = 62,059,223 bp
LG NG50 = 37,007,722 bp
Total anchored = 868,591,263 bp

Figure 3.1. Assembly overview. Flowchart detailing the processing of the raw 44X PacBio sequencing reads, producing candidate assemblies, polishing, breaking, and final assembly anchoring. Metrics are provided at each step.

3.3.2 Evaluating Assemblies

The 37 candidate assemblies were evaluated using a number of different metrics, techniques and complementary datasets. First, each of the candidate assemblies was evaluated using ALE assembly likelihood estimates (61) (which integrated read quality, mate-pair orientation, insert size, coverage and *k*-mer frequencies) based on alignment of the reads from four separate Illumina libraries and of the 44X PacBio dataset (see Methods). Candidate assemblies were also evaluated for completeness using CEGMA (60) and BUSCO (92) core gene sets, as well as by aligning existing *O. niloticus* RefSeq (93) transcripts. A set of 193,027 BAC-end sequences (94) representing ~29X physical clone coverage were used to assess the longer range accuracy of candidate assemblies. Finally, both a physical radiation-hybrid (RH) map consisting of 1,256 markers (95) and a RAD-seq genetic map consisting of 3,802 markers (96) were used to estimate the number of misassemblies present in each of the candidate assemblies. The results of these analyses are provided in Additional File 1.

3.3.3 Ranking Assemblies

No single candidate assembly ranked the highest for all of the evaluation metrics that were computed. Principal component analysis (PCA) was used to reduce the various assembly evaluation metrics and compare the candidate assemblies. Additional File 2 shows that the Canu assemblies tend to cluster separately from the FALCON assemblies in the PCA space. The total assembly size and number of RefSeq exons mapped explained the largest amount of variance and were correlated. These two metrics did not seem like the most important metrics to base the evaluation upon since assembly parameters could be tuned to change the total size and the estimated genome size was 1.082Gbp (97).

The ALE likelihood scores explain the next largest proportion of the variance. The 37 candidate assemblies were ranked by overall ALE scores for each of the five sequencing libraries. An average of the ALE ranks was then calculated. The Canu assembly (#14) that was chosen as the best among the 37 candidate assemblies showed the best average ALE ranks. In addition, Canu assembly #14 had one of the best rates of properly mapped BAC-end sequences, and possessed among the fewest misassemblies as determined by conflicts with the RH and RAD map data (Additional File 1). These results suggest that Canu assembly #14 has the best long-range accuracy while maintaining comparable short-range accuracy.

3.3.4 Polishing

A relatively small number of sequence errors remained in the intermediate unpolished Canu #14 assembly. To correct these errors, first the raw 44X PacBio reads were aligned to the Canu assembly and Quiver was used to polish the assembly at 1,870,943 sites (see Methods). Quiver corrected 1,739,112 (92.95%) insertions, 88,037 (4.71%) substitutions and 43,794 (2.34%) deletions. Next, four Illumina libraries, totaling 277x coverage, were aligned to the intermediate Quiver-polished assembly. Based upon these alignments, Pilon polished an additional 1,101,609 sites. Pilon corrected 1,087,107 (98.68%) insertions, 12,402 (1.13%) substitutions, and 2,100 (0.19%) deletions.

3.3.5 Detection of Misassemblies

The polished intermediate assembly showed high accuracy at the level of individual bases and with respect to the placement of paired-end sequences from ~150kbp BACs (Additional File 1). However, 32 putative inter-chromosomal misassemblies were identified by alignment to the RH and RAD maps. The RH and RAD maps both identified 21 of these inter-chromosomal misassemblies. The RAD map identified an additional 8 putative misassemblies that were not identified using the RH map (the RH map had no markers aligning to these regions), while the RH map identified an additional 3 misassemblies that were not identified using the RAD map (likewise, the RAD map had no markers aligning to these regions). The regions around each putative misassembly were inspected using the genomic resources already mentioned. Each had a characteristic signature consisting of a high density of variants in the 44X PacBio read alignments, as well as low or zero physical coverage of the 40kbp insert Illumina mate-pair library. An example of these misassembly signatures is shown in Additional File 3.

Genome wide analysis of the intermediate assembly for each of these characteristic signatures detected 110 regions of high-density PacBio variants and 376 regions of low physical coverage in the 40kbp mate-pair library. 41 regions had both a high-density of PacBio variants and low physical 40kbp mate-pair coverage. Nine of these regions showed correct alignment to both maps and therefore were not included in the set of putative misassemblies. However, two of these regions were identified by the PacBio variants and low 40kbp physical coverage in regions where there were no markers in either the RH or RAD map and added to the 32 map-based misassemblies giving a total of 34 sites of likely misassembly. Table 3.1 provides a summary of the putative misassemblies that were identified by the maps and sequence alignment methods.

Evidence	Number of misassemblies detected
Both maps	21
RAD map only	8
RH map only	3
Both PacBio variants and 40kbp library	16
PacBio variants only	2
40kbp library only	16
PacBio variants and 40kbp library, but no maps	2

Table 3.1. Number of putative misassemblies identified by various methods.

Analysis of the repetitive elements within these regions revealed that misassembly locations were enriched for highly repetitive interspersed and nested repeats. We examined the region ~75kbp on both sides of the likely misassembly breakpoints and found that 94.51% of these regions were classified as repetitive (see Methods). These regions were enriched for several TE families. Table 3.2 shows the enrichment of the most common repeats and TEs within the misassembly regions. In each of these cases, the mean length of these repeats was longer within the misassembly regions. Some of the same TE families that are abundant across the whole genome (e.g. DNA-TcMar-Tc1, LINE-L2, LINE-Rex-Babar) are also present in high frequency in the misassembly regions. However, some TE families that occurred in relatively low frequency across the whole genome (e.g. DNA-Sola, LTR-ERV1, RC-Helitron, and satellite repeats) are highly enriched in the misassembly regions.

Repetitive element		O_niloticus_UMD1 genome wide				Within 34 misassembly regions				Enrichment	
Order	**Family**	**#**	**Total bp**	**%**	**Mean length**	**#**	**Total bp**	**%**	**Mean length**	**Ratio of genome/ mis-assembly regions**	**Δ mean length**
DNA	Sola	7,007	1,536,337	0.15%	219.3	142	53,045	1.37%	373.6	913.3%	154.3
	TcMar-Tc1	156,588	46,394,192	4.60%	296.3	846	354,606	9.18%	419.2	199.6%	122.9
	hAT	36,441	10,103,158	1.00%	277.2	289	99,016	2.56%	342.6	256.0%	65.4
	hAT-Ac	49,528	17,626,929	1.75%	355.9	445	191,054	4.95%	429.3	282.9%	73.4
	hAT-Charlie	37,049	13,709,558	1.36%	370.0	236	99,913	2.59%	423.4	190.4%	53.3
LINE	L1	10,712	8,879,041	0.88%	828.9	63	79,410	2.06%	1,261	234.1%	431.6
	L2	76,937	29,334,193	2.91%	381.3	539	269,645	6.98%	500.3	239.9%	119.0
	Penelope	28,509	7,214,522	0.71%	253.1	258	70,627	1.83%	273.7	257.7%	20.7
	Rex-Babar	38,996	19,208,630	1.90%	492.6	386	199,730	5.17%	517.4	272.1%	24.9
LTR	ERV1	10,756	6,450,995	0.64%	599.8	112	97,055	2.51%	866.6	392.2%	266.8
	Gypsy	29,201	13,615,743	1.35%	466.3	274	172,231	4.46%	628.6	330.4%	162.3
RC	Helitron	3,882	2,685,111	0.27%	691.7	216	187,805	4.86%	869.5	1800.0%	177.8
Unk.	Unknown	350,007	97,456,302	9.66%	278.4	2,188	841,264	21.78%	384.5	225.5%	106.0
Satellite	Satellite	10,061	7,662,597	0.76%	761.6	93	115,309	2.99%	1,240	393.4%	478.3
	Simple	322,732	15,543,664	1.54%	48.2	1,733	173,824	4.50%	100.3	292.2%	52.1

Table 3.2. Repeats in putative misassembly regions compared to the whole genome.

3.3.6 Anchoring

Table 3.3 provides the anchored size of each LG, including gaps. The new O_niloticus_UMD1 assembly anchored 868.6Mbp of the total genome (86.9%), which is 211Mbp (32%) more than was anchored in the previous "Orenil1.1" assembly (657Mbp) (83). When gaps are not counted, the amount of anchored, non-gap, sequence is 864Mbp (86.4%) compared to 606Mbp (60.6%) in the previous Orenil1.1 assembly. LG3 is the largest anchored LG (68.6Mbp), which agrees with cytogenetic studies that show LG3 as the largest and most repetitive chromosome in the *O. niloticus* genome (24,95,98). Cytogenetic studies also indicate that LG7 is the second largest chromosome in the *O. niloticus* genome, and LG7 is the second largest anchored LG in the new "O_niloticus_UMD1" assembly.

Linkage Group	Orenil1.1	O_niloticus_UMD1	Difference (%)
LG1	31,194,787	38,372,991	7,178,204 (23.0%)
LG2	25,048,291	35,256,741	10,208,450 (40.8%)
LG3	19,325,363	68,550,753	49,225,390 (254.7%)
LG4	28,679,955	38,038,224	9,358,269 (32.6%)
LG5	37,389,089	34,628,617	-2,760,472 (-7.4%)
LG6	36,725,243	44,571,662	7,846,419 (21.4%)
LG7	51,042,256	62,059,223	11,016,967 (21.6%)
LG8	29,447,820	30,802,437	1,354,617 (4.6%)
LG9	20,956,653	27,519,051	6,562,398 (31.3%)
LG10	17,092,887	32,426,571	15,333,684 (89.7%)
LG11	33,447,472	36,466,354	3,018,882 (9.0%)
LG12	34,679,706	41,232,431	6,552,725 (18.9%)
LG13	32,787,261	32,337,344	-449,917 (-1.4%)
LG14	34,191,023	39,264,731	5,073,708 (14.8%)
LG15	26,684,556	36,154,882	9,470,326 (35.5%)
LG16	34,890,008	43,860,769	8,970,761 (25.7%)
LG17	31,749,960	40,919,683	9,169,723 (28.9%)
LG18	26,198,306	37,007,722	10,809,416 (41.3%)
LG19	27,159,252	31,245,232	4,085,980 (15.0%)
LG20	31,470,686	36,767,035	5,296,349 (16.8%)
LG22	26,410,405	37,011,614	10,601,209 (40.1%)
LG23	20,779,993	44,097,196	23,317,203 (112.2%)
Total	**657,350,972**	**868,591,263**	**211,240,291 (32.1%)**
Total (minus gaps%)	**606,480,097**	**864,361,263**	**257,881,166 (42.5%)**

Table 3.3 – Size of each anchored linkage group for both the previous assembly, Orenil1.1 (83) and the new assembly (O_niloticus_UMD1).

3.3.7 Assembly Completeness

To determine the completeness of the new O_niloticus_UMD1 assembly, the assembly was compared against two established sets of core vertebrate gene sets. Table 3.4 shows the number of the 248 CEGMA and the 3,023 BUSCO conserved vertebrate genes that were identified in the new assembly. The number of conserved genes identified increased for both the CEGMA and BUSCO gene sets. The number of complete single-copy BUSCOs increased by 223 (10%), while the number of complete duplicated BUSCOs increased by 26 (59%). The number of missing BUSCOs decreased by 288 (67%) in the O_niloticus_UMD1 assembly compared to the Orenil1.1 assembly.

	Orenil1.1 (LGs)	**O_niloticus_UMD1 (LGs)**
Complete CEGs	244 (98.39%)	245 (98.79%)
Complete + partial CEGs	247 (99.61%)	248 (100%)
Total complete CEGs including putative orthologs	333	342
Complete BUSCOs	2185 (72.28%)	2408 (79.66%)
Complete and single-copy BUSCOs	2141 (70.82%)	2338 (77.34%)
Complete and duplicated BUSCOs	44	70
Fragmented BUSCOs	411 (13.60%)	476 (15.75%)
Missing BUSCOs	427 (14.13%)	139 (4.60%)

Table 3.4 – Genome completeness as measured by CEGMA and BUSCO.

3.3.8 Annotation

The O_niloticus_UMD1 assembly was annotated using the NCBI RefSeq automated eukaryotic genome annotation pipeline. This same pipeline was previously used to annotate the Orenil1.1 assembly. Several additional, new transcriptome datasets (particularly gill tissues, see Methods) were available to annotate the O_niloticus_UMD1 assembly that were not available during the Orenil1.1 annotation process. A comparison of both genome assembly annotations is provided in Table 3.5. The O_niloticus_UMD1 assembly contains 8,238 more gene and pseudogene annotations than the Orenil1.1 assembly (27.3% increase). Similarly, the number of mRNA annotations increased markedly by 10,374 (21.7% increase). The number of partial mRNA annotations decreased from 3,050 to 393 (87.1% decrease). CDS annotations also increased overall (21.9%). The RefSeq annotation pipeline makes corrections to CDS annotations that contain premature stop-codons, frameshifts and internal gaps that would disrupt protein sequence coding. These corrections are based on transcriptome data and corrected 743 CDSs in O_niloticus_UMD1 compared to 817 previously for Orenil1.1 (9.1% decrease). The number of non-coding RNAs more than doubled in the O_niloticus_UMD1 assembly (115.5% increase).

Feature	Orenil1.1	O_niloticus_UMD1	Difference (%)
Genes and pseudogenes	30,174	38,412	8,238 (27.3%)
protein-coding	26,329	29,249	2,920 (11.1%)
non-coding	3,508	8,599	5,091 (145.1%)
pseudogenes	337	564	227 (67.4%)
mRNAs	47,700	58,074	10,374 (21.7%)
fully-supported	45,245	55,760	10,515 (23.2%)
partial	3,050	393	-2,657 (-87.1%)
with filled gap(s)	2,480	67	-2,413 (-97.3%)
known RefSeq (NM_)	145	178	33 (22.8%)
model RefSeq (XM_)	47,555	57,896	10,341 (21.7%)
Other RNAs	5,694	12,899	7,205 (126.5%)
fully-supported	5,071	10,881	5,810 (114.6%)
model RefSeq (XR_)	5,071	10,929	5,858 (115.5%)
CDSs	47,892	58,398	10,506 (21.9%)
fully-supported	45,245	55,760	10,515 (23.2%)
partial	2,467	401	-2,066 (-83.7%)
with major correction(s)	817	743	-74 (-9.1%)
known RefSeq (NP_)	145	178	33 (22.8%)
model RefSeq (XP_)	47,555	57,896	10,341 (21.7%)

Table 3.5 – RefSeq annotation summary.

3.3.9 O_niloticus_UMD1 Assembly Summary

Table 3.6 provides summary statistics for the previous *O. niloticus* assembly (Orenil1.1), each intermediate of the new assembly, and our new final assembly (O_niloticus_UMD1). The O_niloticus_UMD1 assembly is more contiguous, with 45% fewer contigs than the number of scaffolds in Orenil1.1. The overall size of the O_niloticus_UMD1 assembly is 1.01Gbp compared to 927Mbp of Orenil1.1. The O_niloticus_UMD1 contains only 424 gaps that were introduced in the anchoring step. These anchoring gaps amount to 4.2Mbp (0.42%) due entirely to the arbitrary

10kbp gaps placed between anchored contigs. This compares to 111.5Mbp (12.04%) of gaps in Orenil1.1. Overall, 189.5Mbp of new sequence has been assembled in O_niloticus_UMD1 that was either previously in gaps or not assembled at all in Orenil1.1.

Assembly	O_niloticus_wgs_v1 (scaffolds)	Orenil1.1 (anchored to LGs)	*O. niloticus* Canu (contigs)	*O. niloticus* Canu broken (contigs)	O_niloticus_UMD1 (anchored to LGs)
Number of contigs/scaffolds	5,900	5,677	2,960	2,989	2,566
Total size	927,725,912	927,679,487	1,003,343,259	1,005,609,889	1,009,839,889
Longest contig/scaffold/LG	13,623,339	51,042,256	20,432,727	13,936,383	62,059,223
Mean contig/scaffold/LG size	157,242	29,879,589.6	338,967	336,437	37,672,228.8
NG50 contig/scaffold/LG	2,629,658	26,684,556	3,325,464	3,110,904	37,007,722
% N	12.04	12.03	0	0	0.42
Number of gaps	71,854	72,077	0	0	424
Non gap bp	816,139,901	816,140,124	1,003,343,259	1,005,609,889	1,005,610,312
Total gap bp	111,586,011	111,539,363	0	0	4,240,000

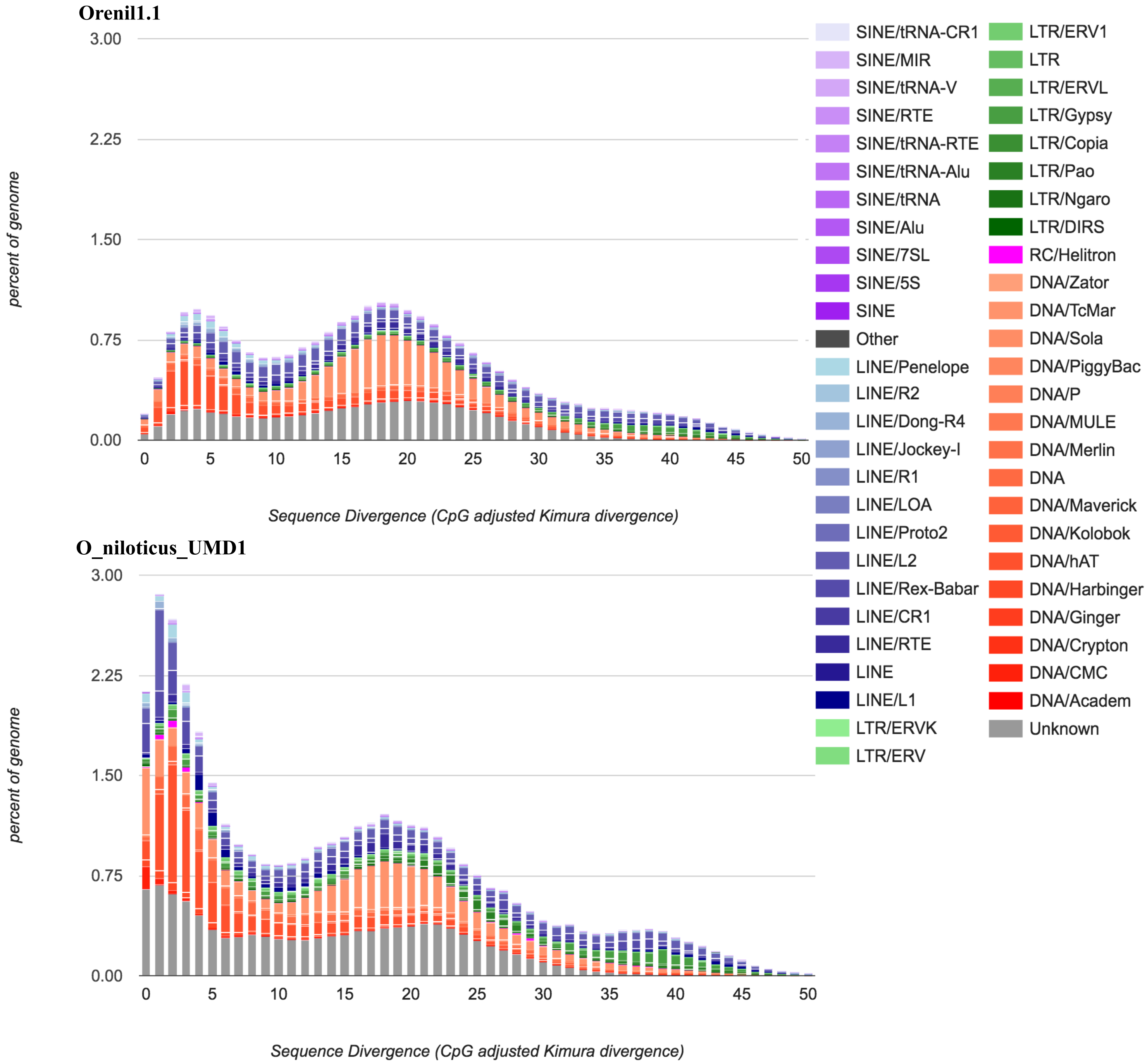
Table 3.6 – Summary statistics for the various assemblies.

3.3.10 Repeat Content

The TE and repeat portion of the genome is vastly under underrepresented in most genome assemblies (85). The use of long PacBio reads allowed for the assembly of more of the repetitive regions of the *O. niloticus* genome. 379Mbp (37.6%) of the total O_niloticus_UMD1 assembly was annotated as repetitive. Table 3.7 provides a summary of the repeat and TE families that were most abundant in the assembly. The new assembly includes an additional 146Mbp (14.6%) of repetitive sequence that was either hidden in gaps or not present at all in the previous assembly. The entire repeat catalog is provided in Additional File 4.

Repetitive element		**Orenil1.1**			**O_niloticus_UMD1**			**Δ from Orenil1.1**	
Order	**Family**	**Total bp**	**%**	**Mean length (bp)**	**Total bp**	**%**	**Mean length (bp)**	**Total bp**	**Mean length (bp)**
DNA	TcMar-Tc1	39,070,443	4.21%	252.0	46,394,192	4.60%	296.3	7,323,749	44.3
	hAT	3,502,443	0.38%	210.2	10,103,158	1.00%	277.2	6,600,715	67.1
	hAT-Ac	11,264,479	1.21%	259.5	17,626,929	1.75%	355.9	6,362,450	96.4
	hAT-Charlie	8,266,601	0.89%	218.1	13,709,558	1.36%	370.0	5,442,957	152.0
LINE	L1	4,469,636	0.48%	389.9	8,879,041	0.88%	828.9	4,409,405	439.0
	L1-1_AFC	1,277,360	0.14%	671.6	3,197,003	0.32%	1,686.2	1,919,643	1,014.6
	L2	20,015,588	2.16%	248.8	29,334,193	2.91%	381.3	9,318,605	132.4
	Rex-Babar	9,422,494	1.02%	276.1	19,208,630	1.90%	492.6	9,786,136	216.5
LTR	ERV1	1,872,564	0.20%	302.2	6,450,995	0.64%	599.8	4,578,431	297.6
	Gypsy	6,734,826	0.73%	415.0	13,615,743	1.35%	466.3	6,880,917	51.3
	Pao	1,745,361	0.19%	686.9	4,892,623	0.48%	833.9	3,147,262	147.0
Unknown	**Unknown**	58,952,108	6.36%	215.4	97,456,302	9.66%	278.4	38,504,194	63.0
Low complexity	**Satellite**	1,110,151	0.12%	268.4	7,662,597	0.76%	761.6	6,552,446	493.2
	Simple	11,784,382	1.27%	42.0	15,543,664	1.54%	48.2	3,759,282	6.2
TOTAL	-	232,691,524	25.10%	183.5	379,017,551	37.56%	254.9	146,326,027	71.4

Table 3.7 – Summary of repeat families in the new assembly.



Orenil1.1

O_niloticus_UMD1

Figure 3.2. Repeat Landscape comparison. The percentage of both the Orenil1.1 and O_niloticus_UMD1 and assemblies that each TE family is represented at in particular substitution levels analogous to the age of TEs (Kimura substitution level – CpG adjusted).

Figure 3.2 provides a comparison of the repeat landscape of the Orenil1.1 and O_niloticus_UMD1 assemblies. Most notably, recently inserted (~ < 5% Kimura divergence) TEs have been assembled in far greater number in the new O_niloticus_UMD1 assembly. The overall number of repetitive elements increased at all divergence levels (218,992 more elements, Additional File 4), with most at lower divergences (165,607 additional elements at < 5% Kimura divergence). The graph suggests that TE insertions less than 1% diverged are still underrepresented in the assembly.

Satellite regions represent one of the most highly repetitive regions of the genome and are often associated with centromeric and heterochromatic regions. Two tilapia-specific satellite repeats have been previously described. ONSATA is a 209bp repeat unit and shows variability between related tilapiine species (99). Only 29 copies of ONSATA (comprising a total of 2,917bp) were assembled and annotated in the original Orenil1.1 assembly. In the new O_niloticus_UMD1 assembly, 226 regions of ONSATA comprising a total of 1,386,985bp were assembled and annotated. Many of the ONSATA regions, the longest of which was 43,805bp on the unanchored contig908, were composed of multiple, nested ONSATA copies. ONSATB is a 1,904bp repeat unit that is organized in tandem arrays and appears to be more conserved and perhaps under selective constraint (100). 48 copies of ONSATB (comprising a total of 11,036bp) were assembled and annotated in the original Orenil1.1 assembly. In the new O_niloticus_UMD1 assembly, 1,481 copies of ONSATB (comprising a total of 2,889,496bp) were assembled and annotated. Again, many of the ONSATB regions were composed of multiple ONSATB copies, the longest of which was 11,210bp located near the beginning of LG12 (607,345-618,555).

TEs specific to African cichlid species have been previously sequenced and used as molecular markers to study evolutionary history and phylogenetics of African cichlids (101,102). Some of these African cichlid specific or “AFC” LINEs and SINEs had been previously assembled and annotated in the Orenil1.1 assembly. An additional 2.3Mbp of AFC-specific TE sequence was annotated in the new O_niloticus_UMD1 assembly. This 2.3Mbp increase was assembled across 55 fewer AFC TE copies, which resulted in longer mean length AFC TE copies. This suggests that the previous assembly contained many fragmented AFC specific TE copies.

3.3.11 Recently Duplicated Regions

Recently duplicated genes are notoriously difficult to assemble due to their high sequence identity (7). Using short Illumina reads to assemble these regions is a difficult task even with mate-pair sequence data across multiple spatial scales. In a previous study of the tilapia *vasa* gene, we identified three partial gene sequences in the Orenil1.1 assembly (103). We then screened a tilapia BAC library for *vasa* gene sequences and identified three BAC clones containing *vasa* sequences. The three clones came from separate restriction fingerprint contigs (104), and represent duplications of the ancestral *vasa* gene. Sanger sequencing identified a full-length *vasa* gene in each of these BAC clones. Figure 3.3a shows how the previous Orenil1.1 assembly failed to correctly assemble any of the three *vasa* gene copies. Figure 3.3b indicates how these genes were assembled from each of the BAC clones.

Figure 3.3c details how the new O_niloticus_UMD1 assembly correctly assembles the three copies of the *vasa* gene corresponding to the three BAC clones.

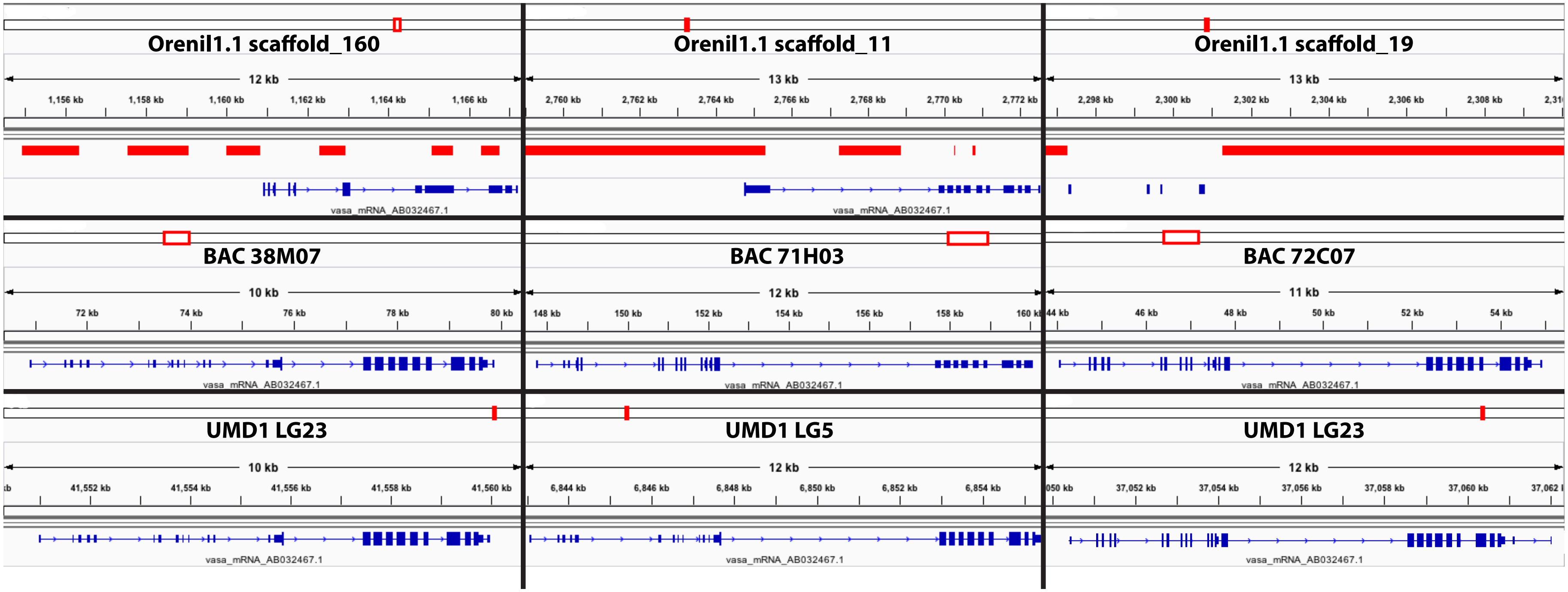


Figure 3.3. *Vasa* gene duplication. a) The top row shows the *vasa* transcript (NCBI accession number AB032467.1) aligned to Orenil1.1 assembly scaffolds with gaps shown in solid red. b) The middle row shows this same *vasa* transcript aligned to the separate BAC assemblies (NCBI accession numbers AB649031-AB649033). c) The bottom row shows the *vasa* transcript aligned to O_niloticus_UMD1 LGs. For each row there are three alignments corresponding to the three copies of each *vasa* transcript.

3.3.12 Sex Determination Regions

The new O_niloticus_UMD1 assembly was used to study sequence differentiation across two sex-determining regions in tilapias. The first region is an XX/XY sex-determination region on LG1 found in many strains of tilapia (77,96,105–108). We previously characterized this region by whole genome Illumina re-sequencing of pooled DNA from males and females (27). We realigned these sequences to the new O_niloticus_UMD1 assembly and searched for variants that were fixed in the XX female pool and polymorphic in the XY male pool. Figure 3.4 shows the F_{ST} and the sex-patterned variant allele frequencies for XX/XY *O. niloticus* comparison across the complete Orenil1.1 and O_niloticus_UMD1 assemblies, while Figure 3.5 focuses on the highly differentiated ~9Mbp region on LG1 with a substantial number of sex-patterned variants, indicative of a reduction in recombination in a sex determination region that has existed for some time (27).

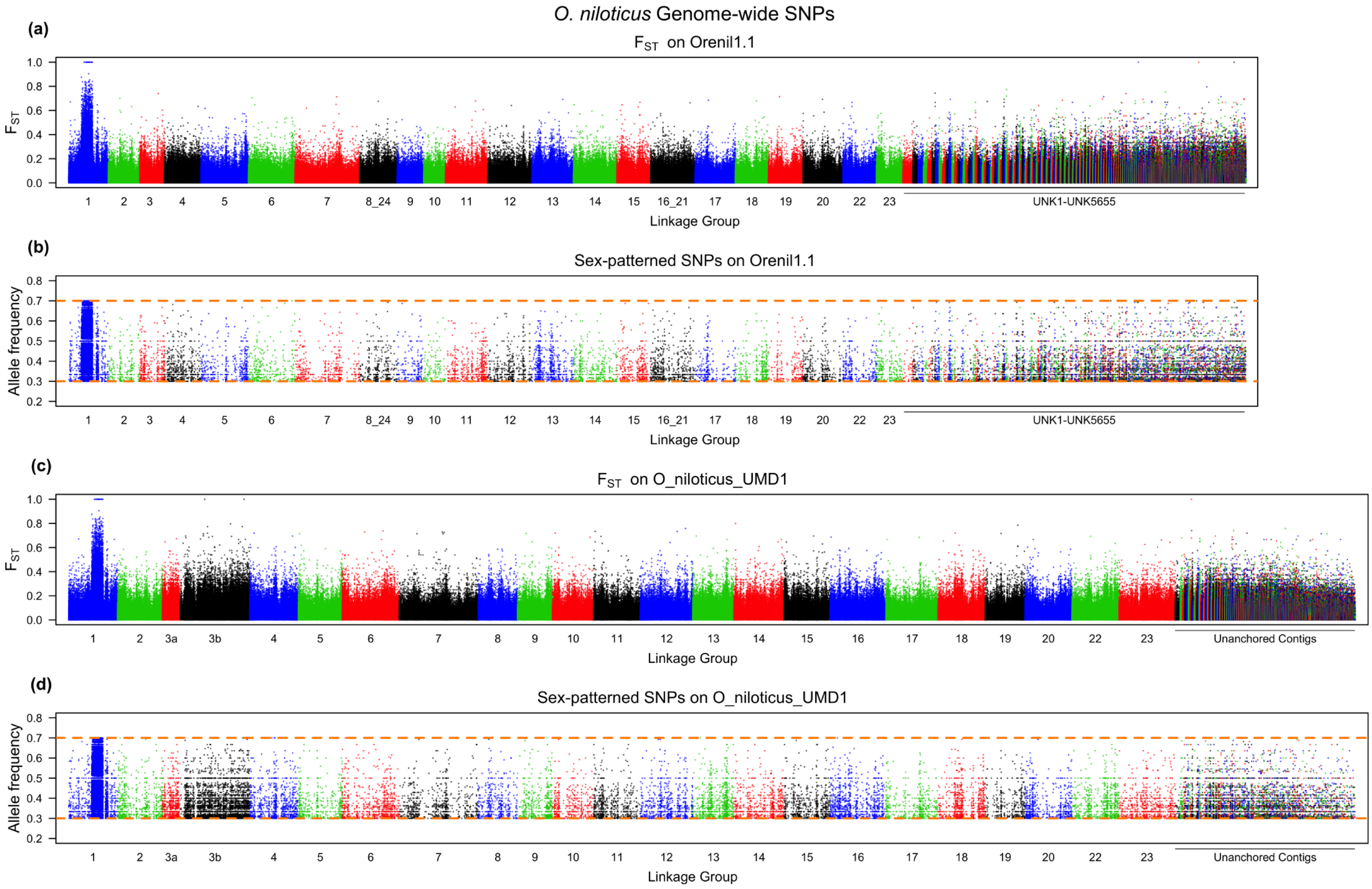


Figure 3.4. Whole genome *O. niloticus* sex comparison. a) F_{ST} comparison of XX female pool versus XY male pool on Orenil1.1. b) Sex-patterned variants across Orenil1.1. c) F_{ST} comparison of XX female pool versus XY male pool on O_niloticus_UMD1. d) Sex-patterned variants across O_niloticus_UMD1.

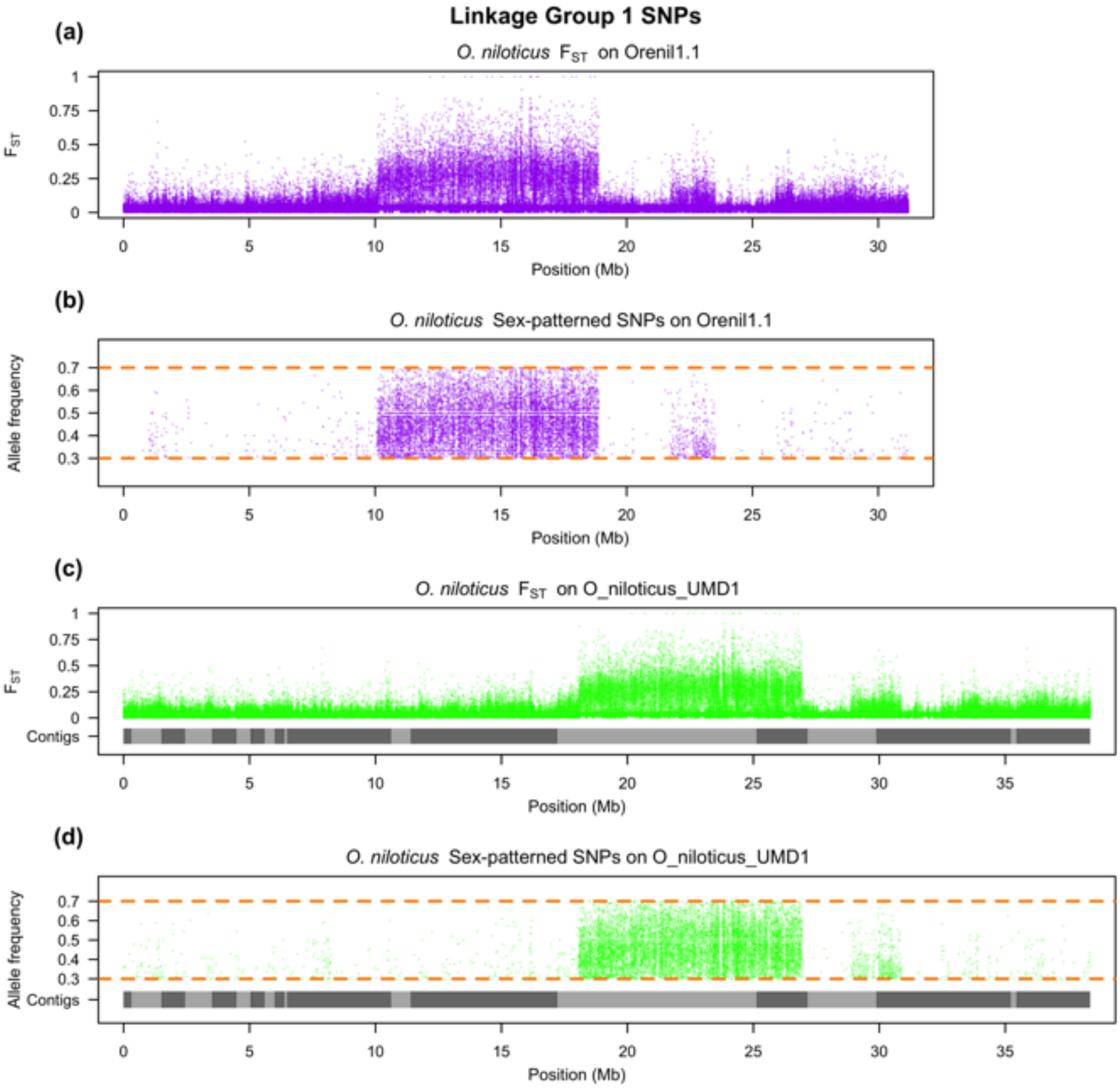


Figure 3.5. LG1 *O. niloticus* sex comparison. a) F_{ST} comparison of XX female pool versus XY male pool on LG1 of Orenil1.1. b) Sex-patterned variants on LG1 of Orenil1.1. c) F_{ST} comparison of XX female pool versus XY male pool on LG1 of O_niloticus_UMD1. Anchored contig boundaries are depicted with grey bars. d) Sex-patterned variants on LG1 of O_niloticus_UMD1.

The second sex comparison is for an ZZ/WZ sex-determination region on LG3 in a strain of *O. aureus* (79,109). This region has not previously been characterized using whole genome sequencing. For this comparison we identified variants alleles fixed in the ZZ male pool and polymorphic in the WZ female pool. Figure 3.6 shows the F_{ST} and the sex-patterned variant allele frequencies for this comparison across the whole O_niloticus_UMD1 assembly, while Figure 3.7 focuses on the differentiated region on LG3. *O. aureus* LG3 contains a large ~50Mbp region of differentiated sex-patterned variants, also indicative of a reduction in recombination in the sex determination region. Figure 3.6 also shows this differentiation pattern on several other LGs (LG7, LG9, LG14, LG16, LG18, LG22 and LG23). It is possible that these smaller regions of sex-patterned differentiation are actually translocations in *O. aureus* relative to the *O. niloticus* genome assembly.

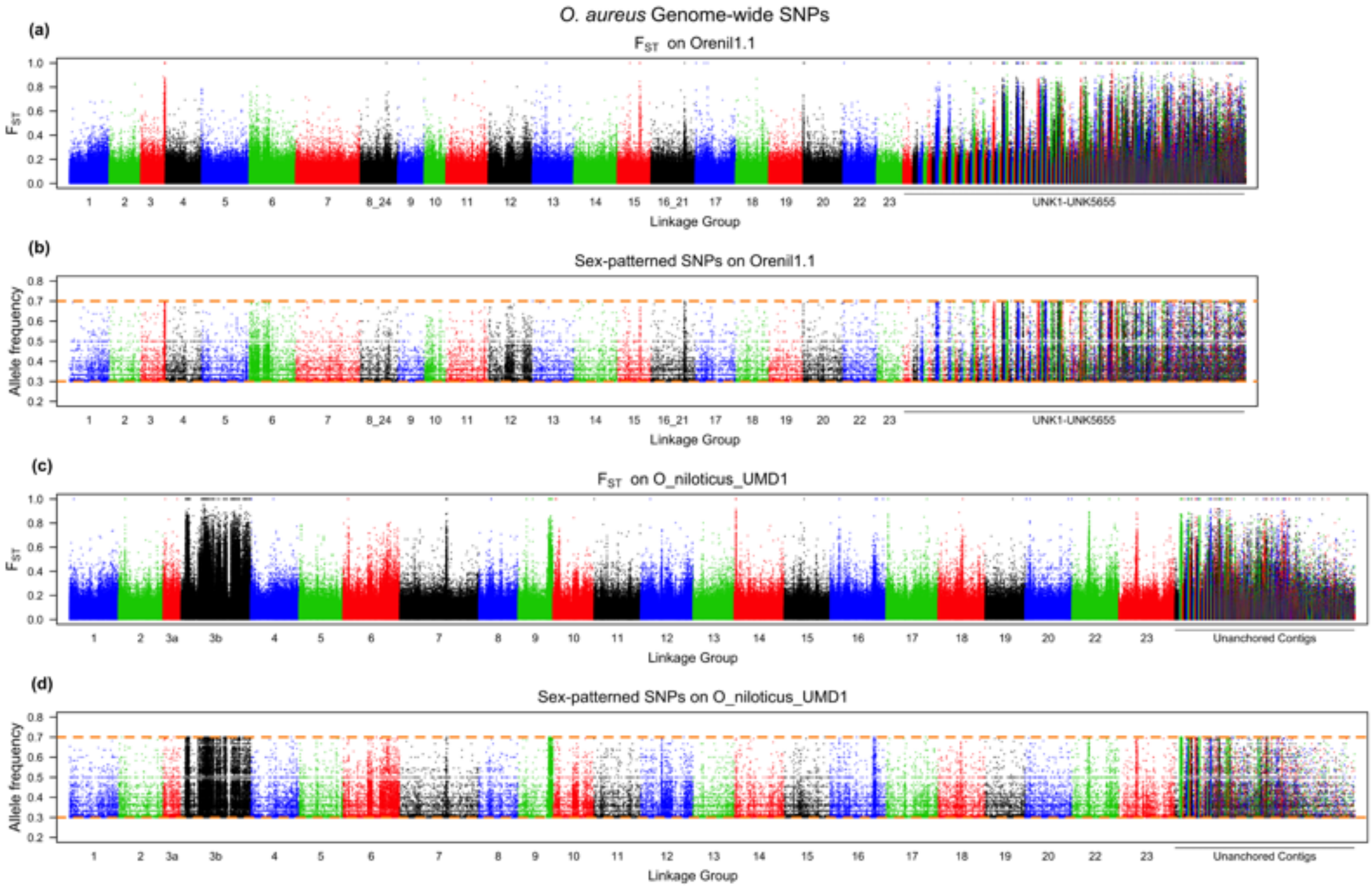


Figure 3.6. Whole genome *O. aureus* sex comparison. a) F_{ST} comparison of ZW female pool versus ZZ male pool on Orenil1.1. b) Sex-patterned variants across Orenil1.1. c) F_{ST} comparison of ZW female pool versus ZZ male pool on O_niloticus_UMD1. d) Sex-patterned variants across O_niloticus_UMD1.

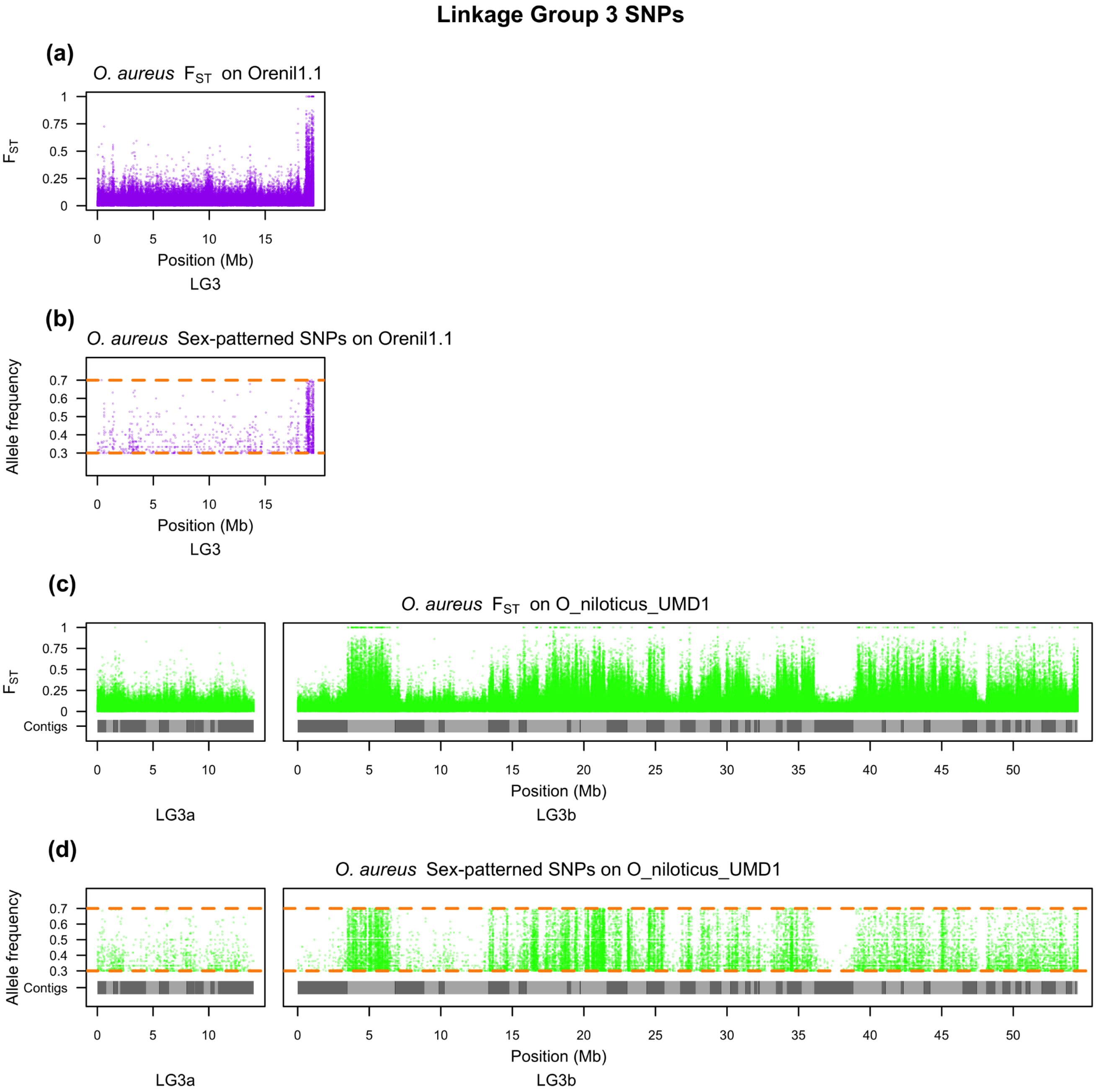


Figure 3.7. LG3 *O. aureus* sex comparison. a) F_{ST} comparison of ZW female pool versus ZZ male pool on LG3 of Orenil1.1. b) Sex-patterned variants on LG3 of

Orenil1.1. c) F_{ST} comparison of ZW female pool versus ZZ male pool on LG3 of O_niloticus_UMD1. Anchored contig boundaries are depicted with grey bars. d) Sex-patterned variants on LG3 of O_niloticus_UMD1.

The overall number of sex-patterned variants was markedly increased for both sex comparisons using the new assembly. Table 3.8 indicates this and provides the number of sex-patterned variants in each comparison across the whole genome as well as on the respective sex-determination LG. LG3 saw the largest gain of sex-patterned variants (1,445 to 24,983 variants) due to the fact that the LG3 assembly now includes 49.3Mbp of new sequence (Table 3.3).

	Orenil1.1	**O_niloticus_UMD1**
O. niloticus		
LG1 sex-patterned variants	11,894	12,225
Non-LG1 sex-patterned variants	17,579	26,493
Total sex-patterned variants	29,473	38,718
O. aureus		
LG3 sex-patterned variants	1,445	24,983
Non-LG3 sex-patterned variants	79,936	78,423
Total sex-patterned variants	81,381	103,406

Table 3.8 – LG1 and LG3 sex-patterned variants using both assemblies.

3.4 Discussion

3.4.1 Genome Assembly

We explored the parameter space of both the FALCON and Canu genome assembly packages and produced 37 candidate assemblies (Additional File 1). Since the true sequence is not known, we had to deduce which of the candidate assemblies best represented the true sequence of the homozygous clone. We elected to assess the assemblies with a variety of metrics, and to select the assembly that scored well across all of the most important metrics.

The first metric is the overall size of the assembly, which should closely match the estimated size of the genome. The size of the *O. niloticus* genome has been measured by both Feulgen densitometry and bulk fluorometric assay. Five separate measurements range between 0.95-1.20 picograms or ~0.929-1.174Gbp (97). The average genome size of these five estimates is 1.082Gbp. The various assemblies ranged in size from 975.1Mbp to 1.07Gbp. The assembly that was chosen (#14) has a length of 1.01Gbp, which corresponds to 93.3% of the estimated size of the genome.

The second set of metrics we considered were the standard measures of assembly contiguity such as NG50, number of contigs, longest contig and mean contig size. The third set of metrics consisted of assembly likelihood (ALE) scores, which were calculated by aligning four Illumina libraries (fragment, 3kbp, 6-7kbp, and 40kbp – Table 3.9, Methods) as well as the 44X PacBio library against each candidate assembly. The fourth metric measured the accuracy of the assemblies at larger scales by aligning the contigs to a ~29X clone coverage library of ~150kbp BAC-end sequences (94) and to existing genetic and physical maps of *O. niloticus* (95,96). Alignment of the RH and RAD maps to the candidate assemblies indicated that every assembly had a relatively low and consistent number of misassemblies (Additional File 1). The fifth metric assessed the completeness of each candidate genome assembly by looking for two core eukaryotic gene sets, CEGMA (60) and BUSCO (92).

No candidate assembly ranked the best for all of these different metrics. In order to choose a preferred assembly, we used principal component analysis to organize the several scores for each assembly. The PCA analysis showed a noticeable difference between the Canu assemblies and the FALCON assemblies (Additional File 2). All of the Canu assemblies clustered together in PCA space. The FALCON assemblies fell into two separate clusters because five of the FALCON assemblies (#17, 32, 34, 35, and 36, Additional File 1) had low ALE scores and NG50s. The other FALCON assemblies tended to show overall better ALE scores for the 44X PacBio library than did the Canu assemblies. This is due to differences in the consensus accuracy between Canu and FALCON assemblies. The 44X PacBio ALE placement and insert scores were virtually the same across all candidate assemblies, but the 44X PacBio ALE k-mer scores were lower for the Canu assemblies. This suggests a slight difference in consensus between Canu and FALCON, although it is probably not noticeable after the polishing steps.

Leaving aside the five low quality FALCON assemblies, a major tradeoff in the PCA is between size of the assembly and the PacBio ALE score. The FALCON assemblies are all smaller than the Canu assemblies, and for the reasons discussed above, have higher ALE scores for the PacBio library. We elected to focus on the Canu assemblies, where the major tradeoff is between the quality of the assembly (ALE scores, NG50, completeness) and size of the assembly (Total size, exon bp mapped). Ultimately, we chose the assembly (Canu #14) with the best overall ALE average rank. This assembly was 28.8Mbp shorter than the longest Canu assembly (#15).

Alignment of the RH and RAD maps to the candidate assemblies indicated that every assembly had a relatively similar and low number of misassemblies (Additional File 1). To correct these misassemblies in the polished version of assembly #14, the locations of misassemblies were first narrowed using the RH and RAD map data together. This typically narrowed the location of a misassembly to a region of less than 1Mbp. From there, the region around each misassembly breakpoint was inspected using alignments of the PacBio data, Illumina data, RefSeq gene set, BAC-end sequences as well as the RepeatMasker annotations. A characteristic signal of high variation in the PacBio alignments, low physical coverage in the Illumina libraries (best characterized with the largest 40kbp Illumina library), and a high density of large and nested repeats was seen in each region of misassembly. Regions of high variation in the PacBio alignments and low 40kbp physical coverage were then calculated genome-wide to investigate whether additional misassemblies might be hidden in the assembly. When considering the PacBio highly variant regions and the low physical 40kbp coverage regions individually, both sets over-estimated the number of misassembly regions. These false-positive potential misassemblies occurred in regions where there was support for correct and continuous assembly based on both RH and RAD map alignments, which together lend stronger support. Only in two cases were there regions that had high PacBio variation, low physical 40kbp coverage and no alignment of RH or RAD map data. We decided to break the assembly at these two locations as well.

3.4.2 Anchoring

A total of 868.6Mbp of the assembled contigs were anchored to the 22 LGs in O_niloticus_UMD1. Overall, 258Mbp of additional (non-gap) sequence has been anchored in the O_niloticus_UMD1 assembly (Table 3.3). All but two of the O_niloticus_UMD1 LGs (LG5 and LG13) are larger in size than in the previous Orenil1.1 assembly. LG5 is 2.7Mbp smaller and LG13 is 0.4Mbp smaller. It is possible that the Orenil1.1 assembly correctly assembled more of these LG5 and LG13. Alternatively, the size difference could be due to overestimates of gap sizes in the Orenil1.1 assembly and/or incorrect assignment of contigs/scaffolds to the wrong LG, which have now been correctly assigned.

It should be noted that although two markers were required to anchor and orient any contig to a particular LG, not all of the markers in the RAD map were located at distinct map positions (i.e. the map has multiple markers at the same genetic position). Therefore, in some cases (particularly involving many of the smaller and repetitive contigs that were anchored to LG3b), the orientation of contigs on LGs is ambiguous. We chose to allow anchoring of these contigs to maximize the anchoring of the many small repetitive contigs that make up LG3.

3.4.3 Annotation

Table 3.5 provides the RefSeq annotation summary of both the Orenil1.1 and new O_niloticus_UMD1 assemblies. The increase in gene and pseudogene annotations is at least partly due to the fact that the O_niloticus_UMD1 assembly contains an additional 189.5Mbp of sequence that was not present in Orenil1.1 as well as the fact that additional transcriptome reads were available for RefSeq annotation of O_niloticus_UMD1. These additional annotations include protein-coding genes (2,920, 11.1% increase), non-coding RNAs (5,091, 145.1% increase) and pseudogenes (227, 67.4% increase). At the same time, there was a decrease in the number of partial mRNA (2,657, 87.1% decrease) and partial CDS (2,066, 83.7% decrease) annotations. This is most likely due to the fact that O_niloticus_UMD1 gene annotations are not disrupted by assembly gaps. The remaining partial annotations may represent recent pseudogenes that the annotation pipeline has little way of differentiating.

The NCBI RefSeq annotation pipeline corrects CDS annotations that have premature stop-codons, frameshifts and internal gaps that would disrupt protein-coding sequence. The RefSeq annotation pipeline corrected 743 CDSs in O_niloticus_UMD1 compared to 817 previously for Orenil1.1. These remaining 743 CDS annotations that required corrections may be due to incomplete polishing in the final O_niloticus_UMD1 assembly, but this number is less than the amount of corrected CDSs annotated in the smaller Orenil1.1 assembly.

3.4.4 Repeats

The vast majority of TE families are represented by more sequence in the new assembly (Table 3.7 and Additional File 4). It is likely that the fragmented Orenil1.1 assembly caused there to be an inflated count of annotated TE copies in places where gaps were inserted within TE copies. The O_niloticus_UMD1 assembly has assembled TE families in longer overall copies than in Orenil1.1 It is also likely that having longer repeat copies and overall 146Mbp more repeat sequence allowed for more accurate annotation of all repeat sequences. In turn, several TE families (such as SINE tRNA-V and LINE Dong-R4, Additional File 4) have decreased in overall number in the O_niloticus_UMD1 assembly, which is likely due to these TEs being more accurately annotated as different, but related TEs. The most recent and less diverged TE copies have been assembled in far greater number in the new O_niloticus_UMD1 assembly (Figure 3.2).

The two tilapia-specific satellite repeats, ONSATA (99) and ONSATB (100), have been shown to be present in high copy number. Both of these satellite repeats have previously been physically mapped using fluorescent *in situ* hybridization (FISH) in *O. niloticus* (110). ONSATA was found almost solely in the centromeres, while ONSATB was also scattered throughout the length of each chromosome arm. Consistent with this, we found nested ONSATA repeat segments assembled near the very ends of several anchored chromosomes (LG3b, LG4, LG8, LG14, and LG17). ONSATB nested repeat segments were found near one or both ends of several anchored chromosomes (LG2, LG3a, LG3b, LG4, LG6, LG11, LG12, LG14, LG16, LG17, LG18, LG19, LG20, and LG23). These data suggest that our assembly of these chromosomes extend into the centromeres. These satellite nested repeats were also abundant in several of the misassembled regions (Table 3.2) suggesting that they remain an obstacle to complete assembly of the genome.

3.4.5 Recently Duplicated Regions

As the recent *vasa* gene duplication in *O. niloticus* (Figure 3.3) shows, the use of long reads has enabled the assembly of such recently duplicated regions. It is likely that there are many other recently duplicated regions that have now been assembled. This is supported by the genome completeness analysis with BUSCO that showed there were 26 additional duplicated BUSCOs out of 3023 searched (Table 3.4). Even though this is a small percentage of the genes analyzed (0.86%), when extrapolated over all the genes in the genome this would amount to hundreds of recently duplicated genes being assembled for the first time. The RefSeq annotation shows that the O_niloticus_UMD1 assembly contained 227 additional pseudogenes (67.4% increase from the Orenil1.1 assembly), which also supports this notion.

3.4.6 Sex Determination Regions

Manipulation of sex-determination in tilapia has important economic implications. The O_niloticus_UMD1 assembly was used to confirm the known and previously described *O. niloticus* ~9Mbp sex-determination region on LG1 (27). The size and pattern of sex differentiation on LG1 and across the genome is similar in both the Orenil1.1 and O_niloticus_UMD1 assemblies (Figure 3.4 and Figure 3.5). A total of 331 additional LG1 sex-patterned variants are identified in the O_niloticus_UMD1 assembly.

The sex-determination region in *O. aureus* is located on the large and highly repetitive LG3. Due to the fact that LG3 is highly repetitive, it was poorly assembled in Orenil1.1 and the vast amount of sex-patterned variants were previously found on unanchored contigs and scaffolds (Figure 3.6a and 3.6b). An additional 23,538 LG3-specific *O. aureus* sex-patterned variants are identified in the O_niloticus_UMD1 assembly. Now that LG3 has been assembled and anchored into a much larger LG (68.5Mbp versus 19.3Mbp, Table 3.3), many of these sex-patterned variants are confirmed on LG3 (Figure 3.6c and 3.6d). There still exist a substantial number of sex-patterned variants on unanchored contigs in the new assembly. The overall pattern of *O. aureus* sex differentiation on LG3 is characterized by several sharp transitions between low and high differentiation (e.g. ~5Mbp and ~37Mbp, Figure 3.7c and 3.7d). These sharp transitions may be explained by either errors in the anchoring process or structural differences between the reference species (*O. niloticus*) and *O. aureus*. Indeed, there are several peaks of differentiation on other LGs (LG7, LG9, LG14, LG16, LG18, LG22 and LG23, Figure 3.6). These may also be chromosomal translocation differences between the two species that will need to be investigated further with FISH.

3.5 Conclusions

This study provides a new assembly and annotation of the Nile tilapia *O. niloticus* (O_niloticus_UMD1), which provides a high-quality reference for the cichlid research community as well as one for studying the evolution of vertebrate genomes. The study also serves as a template for vertebrate genome assembly with current technology and describes many of genomic features that can now be represented correctly. Generation of O_niloticus_UMD1 began by comparing candidate *de novo* assemblies systematically comparing them to select a single best assembly. A small number of misassemblies present in this candidate assembly were identified using several different datasets and subsequently corrected. The final anchored O_niloticus_UMD1 assembly remained very contiguous with a contig NG50 of 3.1Mbp and 86% of contigs anchored to LGs. The number of annotated genes increased 27.3% from the previous assembly of *O. niloticus.* Additionally, a vast amount of repetitive sequences (~146Mbp) were added in the O_niloticus_UMD1 assembly, many of which represent very recent TEs. Finally, the O_niloticus_UMD1 assembly was used to better characterize two large sex-determination regions. The first is a ~9MBp region in *O. niloticus* and the second is a ~50Mbp region in the related species *O. aureus*. Further characterization of these sex-determination regions will have important economic implications for farmed tilapia.

3.6 Methods

3.6.1 PacBio Sequencing

PacBio sequencing was performed on a new individual from the same XX homozygous clonal line used for the previous whole genome sequencing of *O. niloticus* (83). This mitogynogenetic line was developed and maintained at the University of Stirling, UK (90). All working procedures complied with the UK Animals (Scientific Procedures) Act (111).

The Qiagen MagAttract HMW DNA kit was used to extract high-molecular weight DNA from a nucleated blood cell sample of the female “F11D_XX” individual. Size selection was performed at the Genomics Resource Center, Institute for Genome Sciences using a Blue Pippin pulse-field gel electrophoresis instrument. A library was constructed and 63 SMRT cells were sequenced on their PacBio RS II instrument using the P6-C4 chemistry.

3.6.2 Assembly

Both Canu (91) (*version 1.0*) and FALCON (86) (*versions 0.3.0 and 0.5.0*) were run to generate candidate *de novo* genome assemblies. The wide range of parameters tested for both algorithms are provided in Additional File 5. The final assembly (#14), chosen based on the evaluation and likelihood calculations (see below), was run using Canu with the following relevant parameters: '*minReadLength=7000 minOverlapLength=2000 MhapSensitivity=high genomeSize=1g errorRate=0.025 -pacbio-raw*'.

3.6.3 Assembly Accuracy Measurements

Assembly summary metrics were calculated using the assemblathon_stats.pl script (63). Illumina libraries generated previously (83) were aligned to each candidate *de novo* assembly using Bowtie2 (*version 2.2.5* in '*--very-sensitive*' mode). The four different insert size Illumina libraries used are presented in Table 3.9.

Insert size	**NCBI SRA accession IDs**	**Combined coverage**	**Platform(s)**
Fragment	SRR071589, SRR071593, SRR071594, SRR071601, SRR071604, SRR071605, SRR071610, SRR071619	51.6x	Illumina Genome Analyzer II
3kbp	SRR071588, SRR071591, SRR071597, SRR071599, SRR071603, SRR071612, SRR071614	196.3x	Illumina HiSeq 2000
6-7kbp	SRR071602, SRR071607, SRR071615, SRR071616, SRR071617, SRR071620, SRR071622,	24.6x	Illumina Genome Analyzer II
40kbp	SRR071595, SRR071598, SRR071611	4.8x	Illumina Genome Analyzer II

Table 3.9 – *O. niloticus* Illumina libraries used for ALE calculations and Pilon polishing.

For each SRA run, raw reads were downloaded from NCBI using the '*fastq-dump*' program from the SRA Toolkit (112) (*version 2.5.2*). Raw fastq files were combined for each insert size group and Trimmomatic (113) (*version 0.32*) was run on the combined fastq files. The 101bp fragment library and 3kbp library reads were each trimmed with the following Trimmomatic settings:

'*ILLUMINACLIP:TruSeq2-PE.fa:2:30:10 SLIDINGWINDOW:4:20 LEADING:10 TRAILING:10 CROP:101 HEADCROP:0 MINLEN:80*'

The 36bp 6-7kbp library reads were trimmed with the following settings:

'*ILLUMINACLIP:TruSeq2-PE.fa:2:30:10 SLIDINGWINDOW:4:20 LEADING:10 TRAILING:10 CROP:36 HEADCROP:0 MINLEN:31*'

The 76bp 40kbp library reads were trimmed with the following settings:

'*ILLUMINACLIP:TruSeq2-PE.fa:1:30:10 SLIDINGWINDOW:10:20 LEADING:5 TRAILING:10 CROP:76 HEADCROP:4 MINLEN:70*'

For the fragment library, the trimmed and filtered reads were next overlapped with FLASH (53) (*version 1.2.11*) using the following parameters: '*-m 20 -x 0.15 -z*'. The samtools (114) (*version 1.1*) '*view*' and '*sort*' commands were used to convert the Bowtie2 SAM outputs to BAM format. The Picard (115) (*version 2.1.0*) '*MarkDuplicates*' program was run on each of these Bowtie2 alignments with '*REMOVE_DUPLICATES=true*'.

The Assembly Likelihood Estimator (ALE) (61) was then run on each of these filtered BAM files to generate likelihood statistics for each candidate Canu and FALCON *de novo* assembly for each Illumina library. Additionally, to generate ALE scores for the raw PacBio data aligned to each assembly, the 44X raw PacBio reads were aligned using BLASR (54) (version *1.3.1.127046*) with the following parameters: '*-minMatch 8 -minPctIdentity 70 -bestn 1 -nCandidates 10 -maxScore -500 -nproc 40 -noSplitSubreads –sam'.* ALE was then run on these BLASR alignments as well.

A set of *O. niloticus* paired BAC-end sequences (94) were aligned against each candidate assembly using BLAST (116–118) (*version 2.3.0+*). The top hit with an *E-value* less than 1e-150 were kept and then assigned a category of alignment relative to the candidate assemblies according to the details described previously (94) and briefly explained for Additional File 1.

To evaluate the completeness of the candidate assemblies, BUSCO (92) (*version 1.22*) was run (in '*-m genome*' mode) using the '*vertebrata*' lineage-specific profile library. CEGMA (60) (*version 2.5*) was also run on each of the candidate assemblies. CEGMA was run optimized for vertebrate genomes (option ‘*--vrt*') and relied on GeneWise (*version 2.4.1)*, HMMER (*version 3.1b1*), and NCBI BLAST+ (*version 2.3.0+*) using the provided set of 248 CEGs.

3.6.4 Principal Component Analysis

The following metrics were calculated and culled for each of the 37 candidate assemblies: Total ALE score for the aligned Illumina fragment, 3kbp, 6-7kbp, and 40kbp libraries; Total ALE score for the aligned PacBio library; Total number of complete CEGs as defined by CEGMA; Longest contig; NG50; Total assembly size (bp); Total number of RefSeq exon bp mapped. *O. niloticus* RefSeq transcripts (93) (*release 70*) were aligned to each of the candidate assemblies using GMAP (55) (*version 2015-07-23*) and exon bp mapped were calculated from the output GFF3 file. R version 3.2.3 was used to perform the PCA analysis using the ‘*prcomp*’ function with ‘*center=TRUE, scale=TRUE*’ and to create plots with the ‘*biplot*’ function.

3.6.5 Polishing the Assembly

SMRT-Analysis (68) (*version 2.3.0.140936*) was used for polishing the Canu #14 assembly using the 44X raw PacBio reads. First, each SMRT cell was separately aligned to the unpolished Canu assembly using pbalign (*version 0.2.0.138342*) with the ‘*--forQuiver*' flag. Next, cmph5tools.py (*version 0.8.0*) was used to merge and sort (with the ‘*--deep*’ flag) the pbalign .h5 output files for each SMRT cell. Finally, Quiver (*GenomicConsensus version 0.9.2* and *ConsensusCore version 0.8.8*) was run on the merged and sorted pbalign output to produce an initial polished assembly.

Pilon (119) (*version 1.18*) was run on the intermediate Quiver-polished assembly produced above. Again, Bowtie2 (*version 2.2.5* in '*--very-sensitive*' mode) was used to align Illumina reads to this intermediate assembly for Pilon polishing. The fragment library alignment was supplied to Pilon with '*--unpaired*' while the other 3 insert library alignments were specified with '*--jumps*'. Additionally, Pilon was run with the following parameters: '*--changes --vcf --chunksize 40000000 --fix all*'.

3.6.6 Detecting Misassemblies

The 44X coverage raw PacBio reads were aligned to the Quiver- and Pilon-polished Canu #14 assembly using BLASR (54) (version *1.3.1.127046*) with the same parameters as mentioned above. Variants were called using FreeBayes (120) (version *v1.0.2-33-gdbb6160-dirty*). To facilitate FreeBayes processing, regions of the polished assembly were broken into 500kbp chunks using the FreeBayes "fasta_generate_regions.py" script. The separate VCF output files were then concatenated using the VCFtools (121) '*vcf-concat*' program. The FreeBayes utility '*vcffilter*' was used to filter these variants for quality greater than 10 ('*-f "QUAL > 10"* '). VCFtools was then used to compute variant density by specifying '*--SNPdensity 10000*' to calculate variant density in 10kbp windows. Highly variant regions were flagged if there were more than 1 variant per 1kbp over a 10kbp window.

The 40kbp mate-pair Illumina reads of the same homozygous inbred *O. niloticus* line (83) were downloaded from the NCBI SRA (SRR071595, SRR071598, and SRR071611). Trimmomatic (113) (*version 0.32*) was run to remove adaptor sequences and to trim/quality filter these reads. The relevant parameters for Trimmomatic were '*PE -phred33 ILLUMINACLIP:TruSeq2-PE.fa:1:30:10 SLIDINGWINDOW:10:20 LEADING:5 TRAILING:10 CROP:76 HEADCROP:4 MINLEN:70*'. The trimmed and filtered reads were combined and aligned to the polished assembly using BWA mem (122) (*version 0.7.12-r1044*) with the '*-M*' flag. The Picard (115) (*version 2.1.0*) '*SortSam*' program was used to convert the SAM output to BAM (*'SORT_ORDER=coordinate'*) and the Picard '*MarkDuplicates*' program was used to identify duplicate reads. The physical coverage of the 40kbp mate-pairs was calculated on a per-contig basis using a series of piped samtools (114) (*version 1.1*) and bedtools (*version v2.26.0*) commands using the following template, where '*contig*' and '*contig_size*' are the specific contig and its respective size: '*samtools sort -no <(samtools view -bh -F 2 -q 1 40kb.bam contig) tmp | bamToBed -i stdin -bedpe | cut -f 1,2,6 | sort -k 1,1 | bedtools genomecov -i stdin -g <(echo -e "contig\tcontig_size\n") -bga -pc | grep ^contig > output*'. Regions within 200kbp of the start or end of a contig were then excluded from this analysis. Regions below 20x physical coverage of 40kbp mate-pair reads were flagged.

Regions of high variant density within 20kbp of each other, based on raw PacBio alignments, were merged using the bedtools '*merge*' program (*'-d 20000*'). The same merging of windows was performed for regions of low physical coverage based on the 40kbp mate-pair library. The bedtools '*intersect*' program was then used to determine regions of high-density PacBio variants and low 40kbp mate-pair physical coverage that overlapped by at least 80% in the high-density PacBio variants merged windows (*'-f 0.8'*).

Regions of both high-density PacBio variants and low 40kbp mate-pair physical coverage were compared to the alignments of the RH map and RAD map to confirm or contradict the putative misassemblies. Putative misassembled regions were manually inspected using the BLASR and BWA alignments using IGV (123). In addition to these tracks, both RefSeq (93) (*release 70*) *O. niloticus* transcripts aligned to the polished Canu assembly using GMAP (55) (*version 2015-07-23*) and RepeatMasker (65) repeat annotations were considered when defining the exact location of a misassembly. Break locations were chosen so that they did not occur within RefSeq transcripts or within single repeat annotations. The REAPR (48) (*version 1.0.18*) '*break*' program was used to break and fix the polished Canu assembly by providing the determined break locations.

3.6.7 Anchoring with Chromonomer

Chromonomer (124) (*version 1.05*) was first used to anchor the polished and misassembly-corrected assembly using the RH map for *O. niloticus* (95). This initial anchored assembly was then subsequently anchored again with a RAD map for *O. niloticus* (96). BWA mem (*version 0.7.12-r1044)* was used in both Chromonomer runs to create the input SAM file by aligning respective map marker sequences to the appropriate intermediate assembly. A minimum of two markers were required to anchor a contig to a particular LG. Gaps of 10kbp were placed between anchored contigs using '*--join_gap_size 10000*' in Chromonomer. Several RH linkage groups required manual placement were fixed by replacing their entries in the SAM file used by Chromonomer. The RH map LGs that were not anchored using the RAD map ("LOD4.9-RH10- LG10", "LOD6.5-RH17- LG15", and "LOD5.7-RH31- LG3") were manually placed onto the final LGs by using the additional mapping data provided in the previous publication, '*Additional file 4. Data S4*' of (95) which integrated FISH mapping of BAC markers and an previous genetic map (125). Three RH LGs also had to be fixed as they contained a number of repetitive markers, which was causing them to be anchored to incorrect linkage groups in the RAD map ("LOD4.5-RH5- LG9," LOD6.9-RH6-LG5.rev", and "LOD5.1-RH8-LG13").

To further evaluate the candidate assemblies described above, the Chromonomer output file '*problem_scaffolds.tsv*' was used to count the number of contigs in each assembly that had multiple markers that mapped to two or more separate linkage groups.

3.6.8 RefSeq Annotation

The O_niloticus_UMD1 assembly was submitted to NCBI to perform the Eukaryotic Genome Annotation Pipeline (126). This automated pipeline masks the assembly, and aligns existing transcript, protein, RNA-seq, and curated RefSeq sequences to it. Gene prediction based on these alignments is performed and the best gene models are selected among the RefSeq and predicted models which are then made available as the annotation release. The O_niloticus_UMD1 assembly was annotated as "annotation release 103" (127) using software version 7.2 on December 5 2016, while the previous "annotation release 102" (128) consisted of the Orenil1.1 annotation using software version 6.4 on July 30 2015. The Orenil1.1 annotation used 1,319,429,488 reads available and the O_niloticus_UMD1 used 2,295,445,708 reads available at the times of the respective annotations. The newer transcriptome datasets were derived from testis, ovary, liver, and gill tissues. Only the gill tissue was not present in the annotation of Orenil1.1. The numbers in Table 3.5 were extracted from these summaries.

3.6.9 Repeat Annotation

The annotation of repetitive elements was run on several of the intermediate assemblies as well as the final O_niloticus_UMD1 assembly. For each of these assemblies, RepeatModeler (64) (*version open-1.0.8*) was first used to identify and classify *de novo* repeat families present in each assembly. These *de novo* repeats were then combined (separately for each assembly) with the RepBase-derived RepeatMasker libraries (129). RepeatMasker (65) (*version open-4.0.5*) was then run on each of these assemblies using NCBI BLAST+ (*version 2.3.0+*) as the engine ('*-e ncbi*') and specifying the combined repeat library ('*-lib*'). The more sensitive slow search mode ('*-s*') was used.

3.6.10 Analysis of Duplicated *Vasa* Regions

The *vasa* transcript (NCBI accession AB032467.1) was aligned to three assembled BAC clones (NCBI accessions AB649031-AB649033) corresponding to the three copies of *vasa* present in the *O. niloticus* genome (103) using GMAP (55) (*version*

2015-07-23). The *vasa* transcript was also aligned to the scaffolds of the Orenil1.1 assembly and the final anchored O_niloticus_UMD1 assembly. IGV was used to generate images displaying the transcript alignments of the duplicated *vasa* genes.

3.6.11 Sex Comparisons

Sex comparisons were run on the O_niloticus_UMD1 assembly for two species of tilapia, *O. niloticus* and *O. aureus*. The *O. niloticus* sequence data used in this study was previously described (27). The *O. aureus* individuals used were F1 individuals derived from a stock originally provided by Dr. Gideon Hulata (Institute of Animal Science, Agricultural Research Organization, The Volcani Center, Bet Dagan, Israel) and maintained at University of Maryland. These animal procedures were conducted in accordance with University of Maryland IACUC Protocol #R-10-74. A total of 58 *O. niloticus* XY males, 33 *O. niloticus* XX females, 22 *O. aureus* ZZ males and 22 *O. aureus* WZ females were pooled separately, sheared to ~500bp on a Covaris shearer, and sequenced on an Illumina HiSeq 2000. The reads from each pool were separately mapped to O_niloticus_UMD1 using BWA mem (*v0.7.12*). The alignments were sorted and duplicates were marked with Picard (*v2.1.0*). Alignments were converted into an mpileup file using Samtools (*v0.1.18*) and subsequently into a sync file using Popoolation2 (*v1201*) (130). Estimates of F_{ST} and analyses of sex-patterned variants (SNPs and short deletions that are fixed or nearly fixed in the homogametic sex and in intermediate frequency in the heterogametic sex) were carried out using *Sex_SNP_finder_GA.pl* (https://github.com/Gammerdinger/sex-SNP-finder). For the

O. niloticus sex comparison, the XX females were set to be the homogametic sex. For the *O. aureus* comparison, the ZZ males were set to be the homogametic sex.

3.7 Ethics approval and consent to participate

All working animal procedures complied with the UK Animals (Scientific Procedures) Act (111) under project license number PPL 60/4397 and were performed at the University of Stirling, UK. The animal used for this study was developed and maintained at the University of Stirling, UK.

3.8 Availability of data and material

Sequencing data is available via NCBI using the accessions provided below.

Female *O. niloticus* pool: SRR1606304
Male *O. niloticus* pool: SRR1606298
Female *O. aureus* pool: SRR5121055
Male *O. aureus* pool: SRR5121056
44X *O. niloticus* PacBio reads: SRP093160
O_niloticus_UMD1 Assembly: MKQE00000000

3.9 Funding

This work was supported by the United States Department of Agriculture under project number MD.W-2014-05906 to improve and close the tilapia genome via PacBio sequencing. The work was also funded by National Institutes of Health project R01-EY024639 providing partial salary support for MAC.

3.10 Authors' contributions

MAC and TDK conceived and designed study. DJP provided the blood sample from the homozygous clonal line of *O. niloticus* and KLB optimized and performed the HMW DNA extraction. MAC performed the genome assembly, evaluation and analyses. WJG and TDK raised specimens, performed DNA extraction and prepared sequencing library construction of the *O. aureus* samples. WJG organized map data and performed the computational analysis of the sex comparisons. MAC and TDK wrote the paper. All authors read and approved the manuscript.

3.11 Acknowledgements

We thank Luke Tallon and Naomi Sengamalay at the Genomics Resource Center, Institute for Genome Sciences for providing a high quality PacBio library and sequence reads. We thank Gideon Hulata of the Agricultural Research Organization in Israel for providing the original stocks of *O. aureus* that we maintained for this study. We thank Christopher Hill at University of Washington for helpful guidance on assembly parameters. We thank the members of the Kocher and Carleton labs at UMD for providing thoughtful advice during the course of the project. We thank Karen Carleton for reading and providing thoughtful feedback on the manuscript. We thank Christos Palaiokostas (Roslin Institute) and Michaël Bekaert (University of Stirling) for providing the marker sequences of the RAD map. We thank Lucile Soler (Bioinformatics Infrastructure for Life Sciences) for BAC-end analysis scripts. We thank the NCBI RefSeq curators for annotation of the O_niloticus_UMD1 genome assembly. We acknowledge the University of Maryland supercomputing resources (www.it.umd.edu/hpcc/) made available in conducting the research reported in this paper.

Chapter 4: Chromosome-scale assemblies reveal the structural evolution of African cichlid genomes

Authors

Matthew A. Conte[1], Rajesh Joshi[2], Emily C. Moore[3], Sri Pratima Nandamuri[1], William J. Gammerdinger[1], Frances E. Clark[1], Reade B. Roberts[3], Cesar Martins[4], Karen L. Carleton[1], Sigbjørn Lien[2], Thomas D. Kocher[1*].

Affiliations

1 Department of Biology, University of Maryland, College Park, MD 20742, USA
2 Norwegian University of Life Sciences (NMBU), Norway
3 Department of Biological Sciences and W. M. Keck Center for Behavioral Biology, North Carolina State University, Raleigh, USA
4 Department of Morphology, Institute of Biosciences, UNESP - São Paulo State University, Brazil

*** -** Corresponding author orcid.org/0000-0002-7547-0133

4.1 Abstract

4.1.1 Background

African cichlid fishes are well known for their rapid radiations and provide a good model system for studying evolutionary processes. In particular, we do not have a good understanding of how genome structure evolves in rapidly radiating lineages. Here we compare multiple, high-quality, chromosome-scale genome assemblies to understand the mechanisms of cichlid diversification at a genomic level.

4.1.2 Results

We re-anchored our recent assembly of the Nile tilapia (*Oreochromis niloticus*) genome using a new high-density genetic map. We also developed a new *de*

novo genome assembly of the Lake Malawi cichlid, *Metriaclima zebra*, using high-coverage PacBio sequencing and anchor contigs to linkage groups (LGs) using four different genetic maps. These new anchored assemblies allow for the first chromosome length comparisons of African cichlid genomes.

Large (~2-28Mbp) intra-chromosomal structural differences among species are common. However, there are relatively few inter-chromosomal differences (< 10Mbp total). By assembling and placing centromere arrays in the genome assemblies, we show that the large structural differences account for many of the karyotype differences among species. Most chromosomes share a characteristic pattern of recombination along their length. The exceptions involve regions of large structural change associated with sex-determination chromosomes. Structural differences on LG9, LG11 and LG20 are associated with reductions in recombination, and suggest the presence of inversions unique to the sand-dwelling clade of Lake Malawi cichlids. *M. zebra* has a larger number of recent transposable element (TE) insertions compared to *O. niloticus*, indicating that several TE families have a higher rate of insertion in the haplochromine cichlid lineage.

4.1.3 Conclusion

This study provides a new set of genomic resources that sets the stage for future cichlid research to elucidate the mechanisms driving African cichlid speciation.

4.2 Background

African cichlid fishes, due to their phenotypic diversity and rapid speciation over the last several million years, are a model system for studying the mechanisms of evolution (31). The utility of the system has been enhanced by increasingly complete genome assemblies. Draft genomes of five African cichlid species were previously generated using Illumina short-read sequencing and used in an initial analysis exploring some of the forces at play in African cichlid speciation (5). The draft genome assembly of the Lake Malawi cichlid, *Metriaclima zebra*, was one of the most cutting edge short-read only genomes, as revealed in the Assemblathon 2 competition (6). However, these five draft genome assemblies contain a large number of gaps, and only the assembly of the Nile tilapia, *Oreochromis niloticus*, has been anchored to linkage groups (LGs), making it difficult to compare the structure of cichlid genomes at chromosomal scales.

To improve these cichlid genome resources, we have employed long-read Pacific Bioscience SMRT sequencing. Long-read DNA sequencing technology has made it much easier to create accurate and contiguous genome assemblies (18,86,131–133). In particular, long-read technologies have allowed assembly of repetitive sequences, and the identification of structural variants. We previously improved the genome assembly for the Lake Malawi cichlid, *M. zebra*, by sequencing 16.5X coverage of PacBio reads to fill in gaps and characterize repetitive sequences (84). We also produced a new high-quality genome assembly of *O. niloticus,* using 44X coverage PacBio sequencing. We were able to anchor 86.9% of the assembly to linkage groups, which allowed us to characterize the structure of two sex determination regions in tilapias (29).

Cichlid karyotypes are fundamentally similar. Nevertheless, the diploid number varies from 32-60 (23), and the proportion of metacentric chromosomes varies among cichlid species (24,98). These karyotype changes may have played an important role in the evolution and speciation of African cichlids. Classical cytogenetic studies are able to characterize differences in chromosome number, as well as large fusion or translocation events, that are easily seen under the microscope. They are less suited to studying smaller genome rearrangements, including inversions smaller than a few megabases. Comparisons of chromosome scale assemblies in other vertebrate groups have begun to identify extensive structural differences at both the cytogenetic and the sequence assembly level (134,135), but the role of chromosome rearrangements in recent adaptive radiations has not been well studied.

Chromosome scale assemblies can be achieved either by physical mapping techniques (136), or by anchoring the contigs of the sequence assembly to genetic linkage maps. Genetic maps have the advantage of reflecting another important feature of genomes, namely variation in recombination rate, which has manifold impacts on the levels of genetic polymorphism (137). Recombination rate also has practical effects on the efficiency of genome scans (138).

Here we describe chromosome-scale assemblies of two cichlid genomes. First, we re-anchor our previously published PacBio assembly of the Nile tilapia (*Oreochromis niloticus*) genome (29) using a new high-density genetic map. Second, we present a new assembly of *Metriaclima zebra* based on long PacBio sequence reads. Finally, we anchor the *M. zebra* assembly with several recombination maps produced from hybrid crosses among closely related species from Lake Malawi. The anchored genome assemblies of these two species allow for the first chromosome-scale comparison of African cichlid genomes.

In this paper, we integrate the whole genome alignments of these two species with the location of centromeres and the patterns of recombination, to arrive at a panoramic view of African cichlid genome structure. This new view allows us to see how the structure of cichlid genomes has evolved over the last several million years. This perspective raises new questions about the evolution of cichlid genomes, and how it relates to the evolution of sex chromosomes and the adaptive radiation of East African cichlids. We focus our analyses on three aspects of genome evolution that are revealed by these new chromosome-scale assemblies.

First, we describe the pattern of recombination along each chromosome. Spatial variation in recombination rate has implications for patterns of genetic variation (139,140), the evolution of sex chromosomes (141), and the analysis of genome-wide associations between phenotypes and genotypes (138). Despite the importance of recombination in shaping genome architecture (142), it is only beginning to be studied in cichlids (143). A great diversity of sex chromosomes have evolved in East African cichlids, likely the result of sexual genetic conflict (144). Rapid changes in sex determination mechanism, which are frequently variable even within species, may play an important role in cichlid speciation (31). The evolution of new sex chromosomes often involves chromosomal inversions, which change the pattern of recombination (26,27,80,108,145). Studies of these changing patterns of recombination, and their effects on genetic variation, have been hampered by the incomplete nature of the previous draft genome assemblies.

Second, we characterize the patterns of chromosome rearrangement among species. It has been suggested that teleost karyotypes have remained largely stable since the fish-specific whole genome duplication more than 300 million years ago (146). This is in contrast to recent reports of chromosomal fusions among closely related cichlid species (147,148), and a large number of putative inversions associated with the evolution of sex chromosomes in various species (27–29). Chromosome-scale assemblies of cichlids will allow us to quantify the levels of synteny among teleost lineages, and the rate of intra-chromosomal rearrangement among cichlid lineages in East Africa.

To further explore these distinct patterns of recombination and structural changes in cichlids, an older evolutionary comparison with the detailed genomic history of the medaka (*Oryzias latipes*) was utilized. Previous studies in medaka have shown that subsequent to the teleost-specific whole-genome duplication 320-350 million years ago, a subset of medaka chromosomes remained stable while another subset underwent more dramatic fusion and translocation events (146,149). Related comparisons using additional teleost species have shown that the diploid number of chromosome are relatively stable (24-25 diploid chromosomes in 58% of teleosts) and that when the chromosome number is lower in a particular species or group that it is due to chromosome fusion events (150).

Finally, we quantify the abundance and distribution of various transposable element families in each genome. Several studies have documented the expansion of particular transposon families in East African cichlids (151,152). Transposable elements (TEs) may play important roles in shaping genome architecture, particularly the divergence of sex chromosomes. Transposable elements also may be an important source of regulatory mutations (153). Insertion of an AFC-SINE into a gene promoter is associated with the evolution of a novel egg-spot coloration pattern in haplochromine cichlids (154). Similar promoter element re-wiring events have been shown to control cichlid opsin visual sensitivity (155). Since transposons may have been involved in the evolution of many other phenotypes, it is important that these sequences be well-represented in genome assemblies. Unfortunately, transposable elements are not well-represented in genome assemblies based on short Illumina sequence reads. Our previous work has shown how long-read sequencing greatly improves both the amount and length of TE repeats in cichlid genome assemblies (29,84). A comparative analysis of transposable elements will improve our understanding of the patterns of transposon insertion and deletion during the radiation of East African cichlids.

4.3 Analyses

4.3.1 Anchoring the *O. niloticus* assembly to a high-density linkage map

The recently assembled *O. niloticus* genome (29) was re-anchored in this study using a new high-density map (40,190 SNP markers, see Methods). This new map identified 22 additional misassemblies. Table 4.1 provides a comparison of the previous O_niloticus_UMD1 assembly with this newly anchored O_niloticus_UMD_NMBU assembly.

Linkage group	O_niloticus_UMD1	O_niloticus_UMD_NMBU	Change
LG1	38,372,991	40,673,430	2,300,439
LG2	35,256,741	36,523,203	1,266,462
LG3	68,550,753	87,567,345	19,016,592
LG4	38,038,224	35,549,522	-2,488,702
LG5	34,628,617	39,714,817	5,086,200
LG6	44,571,662	42,433,576	-2,138,086
LG7	62,059,223	64,772,279	2,713,056
LG8	30,802,437	30,527,416	-275,021
LG9	27,519,051	35,850,837	8,331,786
LG10	32,426,571	34,704,454	2,277,883
LG11	36,466,354	39,275,952	2,809,598
LG12	41,232,431	38,600,464	-2,631,967
LG13	32,337,344	34,734,273	2,396,929
LG14	39,264,731	40,509,636	1,244,905
LG15	36,154,882	39,688,505	3,533,623
LG16	43,860,769	36,041,493	-7,819,276
LG17	40,919,683	38,839,487	-2,080,196
LG18	37,007,722	38,636,442	1,628,720
LG19	31,245,232	30,963,196	-282,036
LG20	36,767,035	37,140,374	373,339
LG22	37,011,614	39,199,643	2,188,029
LG23	44,097,196	45,655,644	1,558,448
Total anchored (%)	868,591,263 (86.0%)	907,601,988 (90.2%)	39,010,725 (4.2%)

Table 4.1 – Anchoring comparison of O_niloticus_UMD1 and O_niloticus_UMD_NMBU.

The previous O_niloticus_UMD1 assembly anchored a total of 868.6Mbp and the new O_niloticus_UMD_NMBU assembly anchored a total of 907.6Mbp (90.2%), The majority of the newly anchored sequence is on LG3, which increased 19Mbp from 68.6Mbp to 87.6Mbp. The new O_niloticus_UMD_NMBU assembly also joined LG3 into a single LG, whereas previously LG3 was broken into LG3a and

LG3b. LG3 is known to be the largest and most repetitive chromosome in *O. niloticus* (98), as well as being a known sex determination chromosome (79). The repetitive nature of *O. niloticus* LG3 is highlighted by the fact that it required this new dense map to anchor these smaller contigs. Several LGs (e.g. LG16) have fewer total bp anchored in the O_niloticus_UMD_NMBU compared to the previous O_niloticus_UMD1 assembly. This is due to the fact that misassembled contigs that have been broken by the new map are now assigned to their correct LG.

4.3.2 Diploid sequence assembly of *Metriaclima zebra*

The 65X PacBio reads were assembled using FALCON/FALCON-unzip (86) to generate the new diploid *M. zebra* assembly, "M_zebra_UMD2". FALCON first assembles the PacBio reads into primary contigs (p-contigs) and associate contigs (a-contigs) that correspond to alternate alleles. During the FALCON-unzip step, reads are assigned to haplotypes by phasing of heterozygous SNPs and then a final set of p-contigs and haplotigs are produced. Table 4.2 provides the assembly summary statistics for each of these assembly parts. Measuring the completeness of the p-contigs is simple since the total size of p-contigs (957Mb) closely matches the estimated cichlid genome size of 1Gbp (97). To measure the completeness of the haplotigs, the theoretical sizes of heterozygous regions under null expectations of recombination rates and effective population sizes were compared to the size distribution of the haplotigs. Appendix A shows the size distribution of the assembled haplotigs and how it relates to the theoretical recombination rate for several different effective population sizes. The shape of this haplotig size distribution is closest to the curves representing effective population sizes of 1,000-2,500 which closely matches a recent estimate of the effective population size in *M. zebra* (156).

Assembly fraction	Assembly size (Mbp)	Number of contigs	NG50 N50 (Mbp)	LG50 L50	Mean contig size (kbp)	Max contig size (Mbp)
FALCON p-contigs	986.67	3931	1.38	200	251.00	10.04
FALCON a-contigs	261.12	5625	0.054	1615	46.42	0.381
FALCON-unzip p-contigs	957.01	2313	1.42	186	413.75	10.01
FALCON-unzip haplotigs	642.33	6367	0.214	891	100.89	1.17

Table 4.2. FALCON assembly results for *M. zebra*. NG50 and LG50 are based on estimated genome size of 1Gbp. N50 and L50 sizes provided for a-contigs and haplotigs since there is no known size for the alternate haplotype.

4.3.3 Anchoring of the *M. zebra* genome assembly

Four genetic recombination maps generated from RAD-seq studies of F_2 crosses of six Lake Malawi cichlid species were first used to detect misassemblies, then to anchor the contigs to LGs, and finally to compare species level structural differences. Two previously generated maps consisted of crosses of *Metriaclima zebra* x *Metriaclima mbenjii* with 160 F_2 (66) and *Labeotropheus fuelleborni* and *Tropheops 'red cheek'* with 262 F_2 (157). Two new maps consisted of crosses of *M. mbenjii* x *A. koningsi* (331 F_2) (*in preparation*) and *M. mbenjii* x *A. baenschi* (161 F_2) (*submitted*). Table 4.3 provides the total bp that was anchored to each LG for each of the four maps and the final M_zebra_UMD2 assembly (760.7Mbp).

Linkage group	*M. zebra x M. mbenjii* (160 F2)	*L. fuelleborni x Tropheops 'red cheek'* (262 F2)	*M. mbenjii x A. koningsi* (331 F2)	*M. mbenjii x A. baenschi* (161 F2)	M_zebra_UMD2
LG1	31,191,433	32,150,205	**38,662,702**	36,192,366	38,662,702
LG2	25,783,542	28,952,651	**32,647,892**	33,362,328	32,647,892
LG3	18,498,838	14,707,016	**37,717,145**	24,847,713	37,309,556
LG4	28,418,370	24,424,243	**29,889,472**	23,743,562	30,507,480
LG5	29,725,229	34,008,850	**36,154,892**	30,984,548	36,154,892
LG6	15,868,181	32,717,361	**39,879,506**	32,438,073	39,760,669
LG7	29,333,014	57,016,972	**64,381,187**	50,973,986	64,889,811
LG8	19,307,854	16,999,744	**24,280,574**	18,082,738	23,959,896
LG9	**21,018,370**	22,620,859	18,771,712	24,011,483	21,018,370
LG10	25,942,318	26,176,893	**32,583,833**	25,149,136	32,346,187
LG11	**32,253,887**	30,903,800	34,404,464	31,577,152	32,434,411
LG12	23,231,402	31,401,442	**34,043,602**	31,595,605	34,077,077
LG13	25,893,161	24,034,634	**31,886,878**	28,831,406	32,061,881
LG14	32,750,971	32,025,991	**37,909,455**	30,978,148	37,855,742
LG15	28,015,059	28,462,857	**34,537,245**	28,405,563	34,537,245
LG16	24,665,172	26,935,058	**34,727,877**	29,158,962	34,727,877
LG17	28,473,329	31,631,813	**35,766,785**	31,607,415	35,766,785
LG18	19,927,984	23,757,304	**29,457,134**	30,047,761	29,494,144
LG19	24,076,222	19,992,035	**25,739,093**	22,726,673	25,955,740
LG20	28,281,247	30,800,769	24,975,175	**29,774,176**	29,774,176
LG22	27,460,019	31,372,369	**34,717,234**	30,512,954	34,717,234
LG23	27,069,552	27,967,022	**42,736,004**	37,848,175	42,076,657
Total anchored (%)	567,185,154 (59.3%)	629,059,888 (65.7%)	755,869,861 (79.0%)	662,849,923 (69.3%)	760,736,424 (79.5%)
Total including unanchored	957,158,042	957,163,242	957,185,442	957,167,042	957,200,631

Table 4.3. Anchoring of the *M. zebra* assembly to four different linkage maps. The FALCON assembly was anchored to each map separately, and the total bases anchored shown for each LG and map. The anchored map LGs that were used for the M_zebra_UMD2 anchoring are indicated in **bold**. The *L. fuelleborni x Tropheops 'red cheek'* map had four LGs that were combined into two (LG10a/LG10b and

LG13a/LG13b). Usage of particular LGs in the final anchoring is based on accuracy and not necessarily overall length.

Prior to the final anchoring, these four maps were also used to detect and confirm potential misassemblies in the FALCON assembly. Appendix B provides the list of FALCON p-contigs where markers from two or more different LGs maps aligned, indicating a potential inter-LG misassembly. Each of these potential misassemblies were further inspected using alignments of a 40kb Illumina mate-pair library (158), RefSeq gene annotations (93), and repeat annotations (see Methods). In some cases, it was determined that some map marker sequences were repetitive and giving a false misassembly signal. A total of 33 potential misassemblies were inspected and 16 likely misassemblies were identified and broken. An example view of one of these misassemblies is provided (Appendix C). Whole genome alignment comparisons (see section below) detected one additional intra-chromosomal misassembly brining the final total to 17 breaks.

The *M. mbenjii* x *A. koningsi* map typically anchored more of the *M. zebra* assembly contigs and in a more accurate order (relative to *O. niloticus*) than did the other three maps. This is likely due to the fact that the *M. mbenjii* x *A. koningsi* map had both more F_2 Individuals and more map markers than the other three maps, giving it the highest resolution. However, for several LGs (LG2, LG9, LG18, LG20, see Table 4.3), one of the other three maps anchored more contigs. However, the map that produced the longest anchored LG did always appear to be the most accurate. To determine this accuracy, each *M. zebra* LG (anchored with each of the four maps) was aligned to the anchored *O. niloticus* assembly and compared (Appendix D). In the final assembly, the *M. zebra* x *M. mbenjii* map was used to anchor LG9 and LG11 and for LG20 the *M. mbenjii* x *A. baenschi* map was used. The anchoring of LG9 using *M. mbenjii* x *A. koningsi* map was very short compared to the other LGs and may be indicative of a hybrid incompatibility on LG9 in that cross. The other three maps anchored significantly more of LG9. The *M. zebra* x *M. mbenjii* map was chosen to anchor LG9 as it showed the closest ordering relative to the *O. niloticus* assembly (Appendix D). The *M. zebra* x *M. mbenjii* map was also chosen to anchor LG11 as the other three maps showed large structural differences (Appendix D and also seen in the recombination maps, presented below). LG20 was best represented by the *M. mbenjii* x *A. baenschi* map based on alignment to *O. niloticus*, overall size and by ordering of markers in the recombination maps. The final M_zebra_UMD2 anchoring used three of the four maps to assign, order and orient contigs. The *L. fuelleborni x Tropheops 'red cheek'* map was not used in the final anchoring but helped confirm many misassemblies and informed structural similarities and differences. Several LGs have slightly different overall sizes than when the assembly was anchored with just a single map (e.g. LG3 changed from 37,717,154bp to 37,309,556bp, Table2). This is due to the fact that several small contigs are assigned to different LGs by the four different maps.

4.3.4 Structural differences among Lake Malawi cichlid genomes

The process of anchoring the M_zebra_UMD2 assembly to the four maps also allowed for a comparison of the six species used to generate the maps to see if there were any large structural differences between species. Since a large number of the same p-contigs were assigned to LGs by each map separately, we could look for contigs that were assigned to different LGs in any of the four maps. Table 4.4 provides the list of the 9 contigs that were assigned to different LGs by at least two maps and represents putative inter-chromosomal differences.

	contig size	*Mz.* x *Mb.* map LG	*Lf.* x *Tr.* map LG	*Mb.* x *Ak.* map LG	*Mb.* x *Ab.* map LG	Notes
000084F_pilon|quiver	2,383,905	**LG1 (1)**	LG3 (3)	LG3 (6)	LG3 (3)	
000105F_pilon|quiver_1_1312536	1,312,536	NA	**LG10a (1)**	LG2 (1)	LG2 (3)	
000201F_pilon|quiver	1,489,552	LG3 (1)	*LG1 (3)*	LG3 (3)	LG3 (1)	
000223F_pilon|quiver	1,452,516	LG8 (4)	LG8 (8)	**LG3 (2)**	LG8	repetitive markers on LG3
000256F_pilon|quiver	1,241,607	LG20 (1)	LG20 (1)	NA	**LG9 (1)**	
000414F_pilon|quiver	805,874	LG5 (1)	LG5 (1)	NA	**LG3 (1)**	
000521F_pilon|quiver	566,343	LG15 (2)	NA	LG17 (1)	NA	repetitive marker on LG17
000541F_pilon|quiver	515,490	NA	LG2 (1)	LG3 (1)	NA	
000671F_pilon|quiver	374,096	LG23 (1)	NA	LG23 (1)	***LG22 (1)***	

Table 4.4. Putative inter-chromosomal differences as identified by map anchoring comparison. The number of markers aligned to each contig for each LG is indicated in (*N*). 'NA' indicates that a particular map had no markers aligned to that contig. Potential species-specific inter-chromosomal differences are indicated in **bold,** where possible.

Seven of these nine contigs are anchored by only a single marker on a different LGs than the three other maps and so it is difficult to determine if there is a true inter-chromosomal difference with such little evidence. Even when all nine contig anchoring differences are considered, it amounts to only 10.1Mbp of total inter-chromosomal differences between the species used to generate the maps. This estimates at most 1% of these Lake Malawi cichlid genomes are different at the inter-chromosomal level.

760.7Mbp of the 957.2Mbp total assembly was anchored (Table 4.3). In this 80% of the genome, we detected only 10.1Mbp of potential inter-chromosomal differences. It is possible, that there are some other significant inter-chromosomal differences that we did not detect in the unanchored portion of the genome. If they do exist, they are likely to be highly repetitive portions of these genomes that could not be assembled into long contigs and/or reliable map markers.

4.3.5 Localization of centromeric repeats

Figure 4.1 shows the karyotype of *O. niloticus* and *Metriaclima lombardoi,* two species that diverged 17-28 million years ago (159). *M. lombardoi* is a sister species to *M. zebra* and very closely related. O. *niloticus* was chosen for this comparison since it is the closest relative to *M. zebra* to have an anchored assembly available. The *O. niloticus* SATA repeat (160) is mapped and counter stained in the *O. niloticus* karyotype and maintained in African cichlid centromeres (98). The SATA consensus repeat also closely matches most cichlid satellite repeats assembled from a multitude of datasets in a recent analysis of centromeres across many taxa (161).

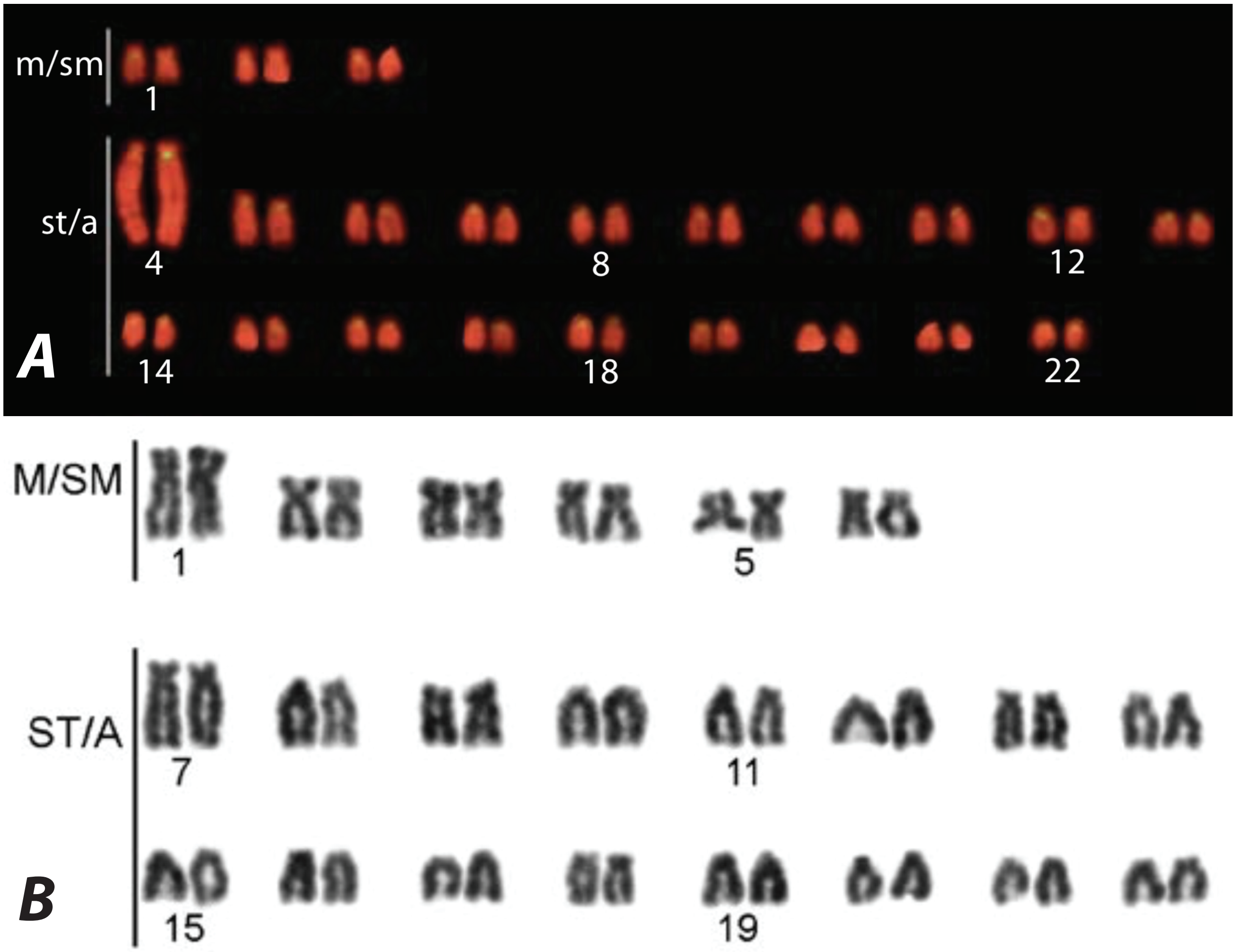


Figure 4.1. A) Chromosome mapping of SATA satellite DNA in *O. niloticus* reproduced with permission from (98). B) Giemsa-stained karyograms of the Lake Malawi *Metriaclima lombardoi* reproduced with permission from (40).

The karyotypes of *M. zebra* and *O. niloticus* each have 22 chromosome pairs, as do the majority of African cichlids. *O. niloticus* has only one meta-submetacentric and 21 subtelo-acrocentric chromosomes whereas *M. zebra* has six meta-submetacentric and 16 subtelo-acrocentric chromosomes. The chromosomes in Figure 4.1 have been ordered by type and then by size but have not been assigned to LGs via genetic maps, previously. Whole genome alignments of M_zebra_UMD2 and O_niloticus_UMD_NMBU were performed and visualized. Appendix D contains images of these whole genome alignments for each LG. Figure 4.2 shows the LG23 alignment of *M. zebra* and *O. niloticus*. Placement of centromere repeats and a large structural rearrangement on LG23 indicates that it is a chromosome that is subtelo-acrocentric in *O. niloticus*, but meta-submetacentric in *M. zebra.* Perhaps the most diverged LG is LG3. Appendix E shows an F_{ST} comparison of the *Oreochromis aureus* male versus female pools described in (29). There is a very wide region of sex-patterned differentiation on LG3 from ~40Mbp to 85Mbp.

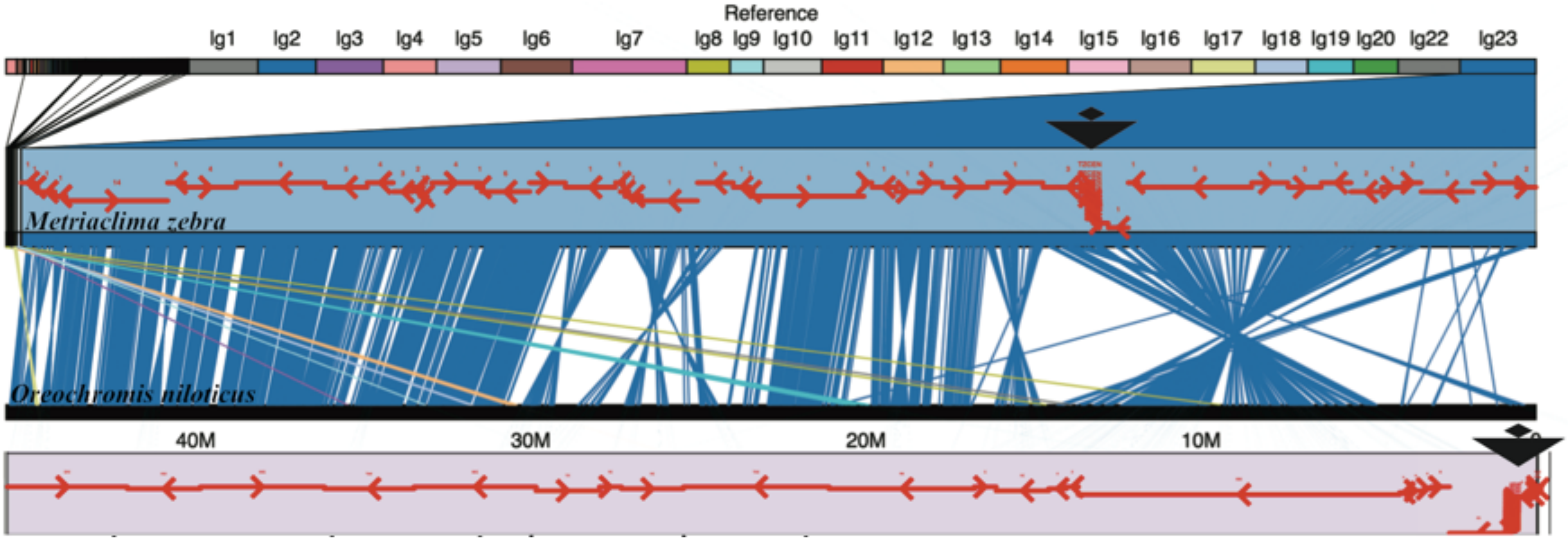


Figure 4.2 – Alignment comparison of LG23 in *M. zebra* and *O. niloticus*. Centromere repeats in each assembly are indicated by large black triangles. Anchored contigs in each assembly are show as red arrows indicating the orientation of each contig.

Centromere repeats were not assembled on every single LG in both *M. zebra* and *O. niloticus.* However, on LGs where centromere repeats were placed in both assemblies and a large structural difference was observed, we were able to identify a large centromere repositioning event indicating acrocentric/metacentric changes (LG4, LG7, LG16, LG17). Although centromeres were not identified in both genome assemblies LG2, LG6, LG20, and LG22 show similar rearrangement events at the ends of chromosomes that may indicate acrocentric/metacentric changes as well (Appendix D).

In addition to identifying and assigning LGs to the karyotype changes between *M. zebra* and *O. niloticus*, the whole genome alignment comparisons have also identified a number of large intra-chromosomal structural rearrangements. On LG2 there are two large rearrangements of ~15Mbp and ~20Mbp (Appendix D). The largest single structural change appears on LG19 where there is a ~23Mbp rearrangement between *M. zebra* and *O. niloticus.* A similar ~20Mbp rearrangement is present on LG20. There is an ~11Mbp rearrangement at one end of LG22 that may be associated with another centromere location change, although the centromere was not localized on LG22 in either assembly.

4.3.6 Variation in recombination rate among species

The four Lake Malawi genetic recombination maps were also used to compare differences in rates and patterns of recombination across LGs and to detect any noticeable differences between the crosses. To do this, each set of map markers were aligned to the final M_zebra_UMD2 assembly and plotted against their recombination map positions. Figure 4.3 shows the comparison of the four maps relative to M_zebra_UMD2 on LG23. Each of the four maps shows high recombination from 0-15Mbp and then much lower recombination to the end of the LG23. The centromere is placed at 30.1Mbp on LG23 and is in the middle of the region of low recombination. This region of low recombination also corresponds to the large (~15Mb) structural rearrangement relative to *O. niloticus* (Figure 4.2). Appendix F contains plots of each of the four maps relative to M_zebra_UMD2 for each LG. During this process, one additional misassembly was detected at 6,922,000 on contig 000000F on LG12 and subsequently broken for the final anchoring (included in Appendix F).

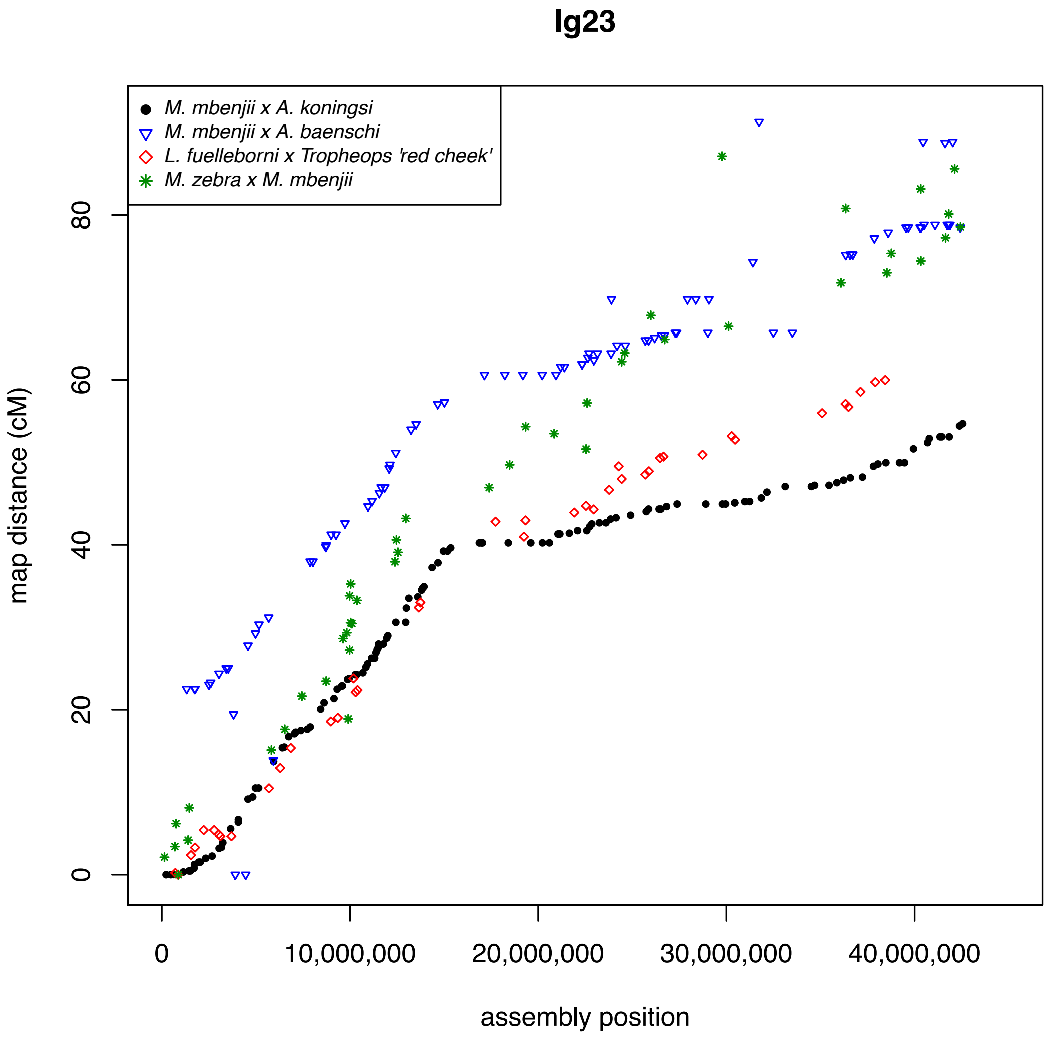


Figure 4.3 - Comparison of the four maps relative to M_zebra_UMD2 on LG23.

The male and female *O. niloticus* recombination curves are plotted against the O_niloticus_UMD_NMBU assembly and provided in Appendix G. Overall, both the *O. niloticus* and the Lake Malawi LGs are characterized by low recombination on the ends of LGs and higher recombination in the middle of LGs. There were several notable exceptions to this pattern though. In Lake Malawi on LG2 there is a region of low recombination for the first ~15Mb that also corresponds with a large structural rearrangement relative to *O. niloticus* (Appendix D), where recombination is not as suppressed. LG7 maintains the overall pattern of low recombination at the ends, but also has a region of low recombination in the middle (at ~30Mbp in M_zebra_UMD2) near several smaller scale rearrangements relative to *O. niloticus*. LG7 is a known sex determination chromosome in Lake Malawi (80), and this odd recombination pattern may represent multiple strata or independent sex inversion(s) on LG7. LG9 appears to be experiencing a large amount of structural rearrangement within all four crosses as seen by both the recombination map and whole genome alignment comparisons. However, in *O. niloticus*, LG9 does not have any abnormal recombination patterns. There appears to be a ~2Mbp inversion on LG10 (relative to *O. niloticus*) that is associated with lowered recombination near the position at 20Mbp in M_zebra_UMD2. LG11 follows the typical recombination pattern for the *M. zebra* x *M. mbenjii* map, but there appears to be a large 15Mbp inversion in the *Aulonocara* genus as seen by both the *M. mbenjii* x *A. koningsi* and *M. mbenjii* x *A. baenschi* maps. The *L. fuelleborni* x *Tropheops 'red cheek'* map also shows a large, but different rearrangement on LG11 when compared to *O. niloticus* (which does not have abnormal recombination on LG11). LG15 has a region of lower recombination in the middle that also associated with structural rearrangements relative to *O. niloticus.* There is a large structural rearrangement on LG20 present in each of the four map anchored assemblies that is also associated with a large (~15Mbp) region of low recombination.

Each of the *O. niloticus* LGs show a difference in recombination between males and females. The typical pattern is higher recombination in the females than the males. However, LG6 and parts of LG4, LG9, LG16, LG20, and LG22 show higher recombination in males than females. LG3 and LG23 are both known sex determination chromosomes in tilapias (78,79), and each deviates from the normal recombination patterns. On LG3, the largest chromosome in *O. niloticus* (Figure 4.1), there is very low recombination for ~70Mbp. On LG23 there is a ~28Mbp region of greatly reduced recombination.

4.3.7 Major structural rearrangements of ancient cichlid chromosomes

We aligned the O_niloticus_UMD_NMBU assembly to the recently published "HSOK" *O. latipes* medaka assembly (149). *O. niloticus* has 22 chromosome pairs, while the medaka HSOK genome has 24 chromosome pairs. Table 4.5 shows the correspondence of cichlid LGs with the medaka HSOK chromosome numbers.

O_niloticus_UMD_NMBU linkage group	Primary medaka HSOK chromosome (alignment length)	Secondary medaka HSOK chromosome (alignment length)
LG1	3	
LG2	10	
LG3	18	
LG4	8	
LG5	5	
LG6	1	
LG7	6 (32Mbp)	12 (31Mbp)
LG8	19	
LG9	20	
LG10	14	
LG11	16	
LG12	9	
LG13	15	
LG14	13	
LG15	24 (31Mbp)	4 (5Mbp)
LG16	21	
LG17	23 (23Mbp)	4 (12Mbp)
LG18	17	
LG19	22	
LG20	7	
LG22	11	
LG23	2 (23Mbp)	4 (17Mbp)

Table 4.5. Corresponding *O. niloticus* and *O. latipes* LG and chromosomes. Chromosomes with large fusion/translocation events have alignment lengths provided.

We identify several large chromosome rearrangement events that happened in the cichlid ancestor. Tilapia LG7, the second largest LG (Table 4.1), is comprised of medaka chromosomes 6 and 12 in their entirety (Figure 4.4). This suggests a chromosome fusion that of these ancestral chromosomes that have remained in cichlids. Tilapia LG23, the third largest LG (Table 4.1), is comprised of medaka chromosome 2 in its entirety and roughly 17Mbp or roughly half of medaka chromosome 4 (Figure 4.5). The other half of medaka chromosome 4 was likely translocated onto LG15 and LG17. The remaining 18 chromosomes have undergone extensive intra-chromosomal rearrangements in many cases yet largely correspond to the same chromosomes having evolved over the course of the 120 million years of evolution since the divergence of the common ancestor of medaka and tilapia. LG3 is the largest tilapia LG (Table 4.1), but surprisingly does not show any evidence of a chromosomal fusion or translocation event. Tilapia LG3 aligns well to medaka chromosome 18 along the first ~30Mbp of LG3, and the remainder of LG3 aligns to medaka chromosome 18 in a much more sporadically and with much less contiguity. This divergent region of LG3 corresponds to the large ~70Mbp of low recombination.

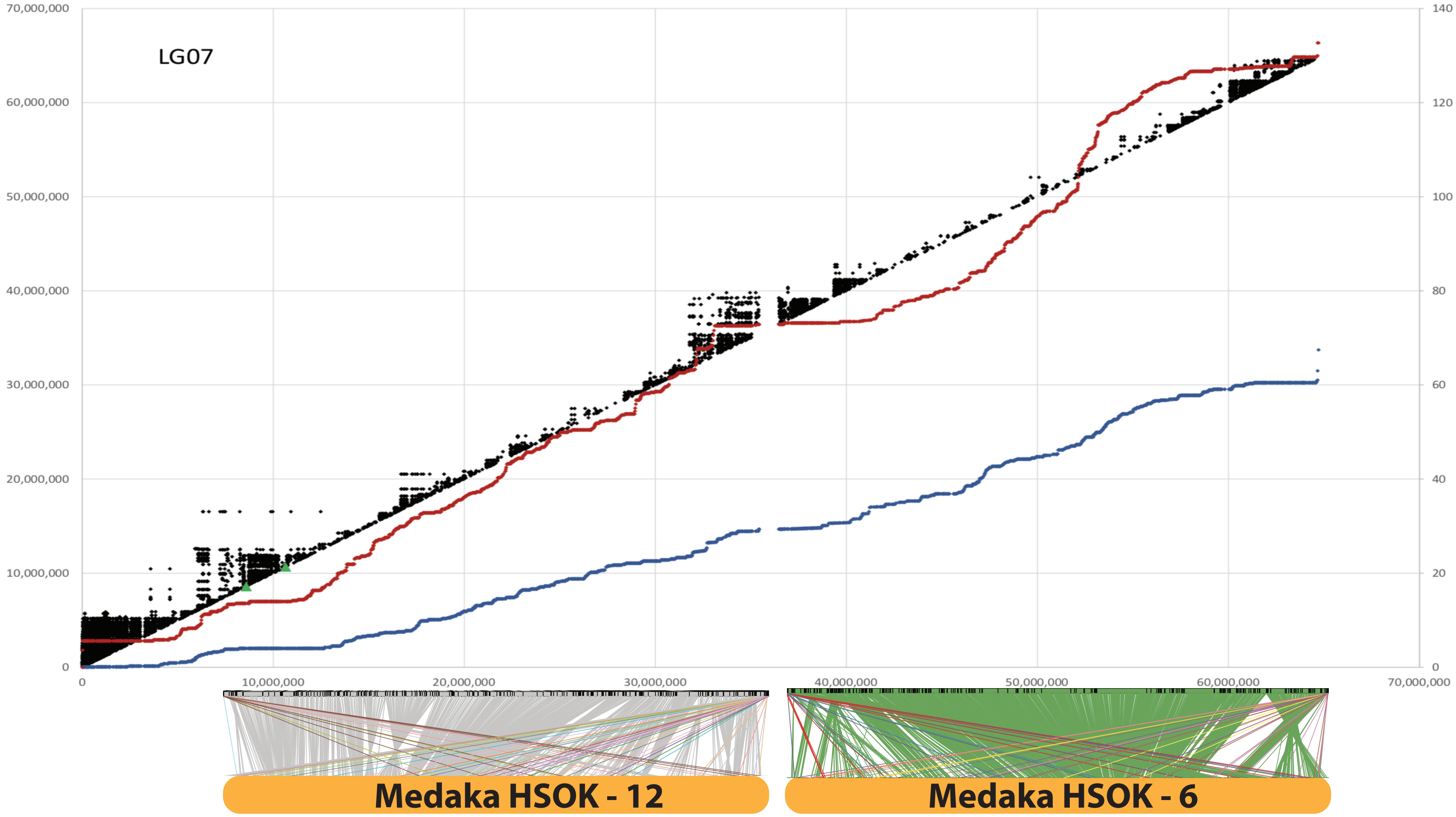


Figure 4.4. O_niloticus_UMD_NMBU LG07 is an ancient fusion of medaka HSOK 12 and 6.

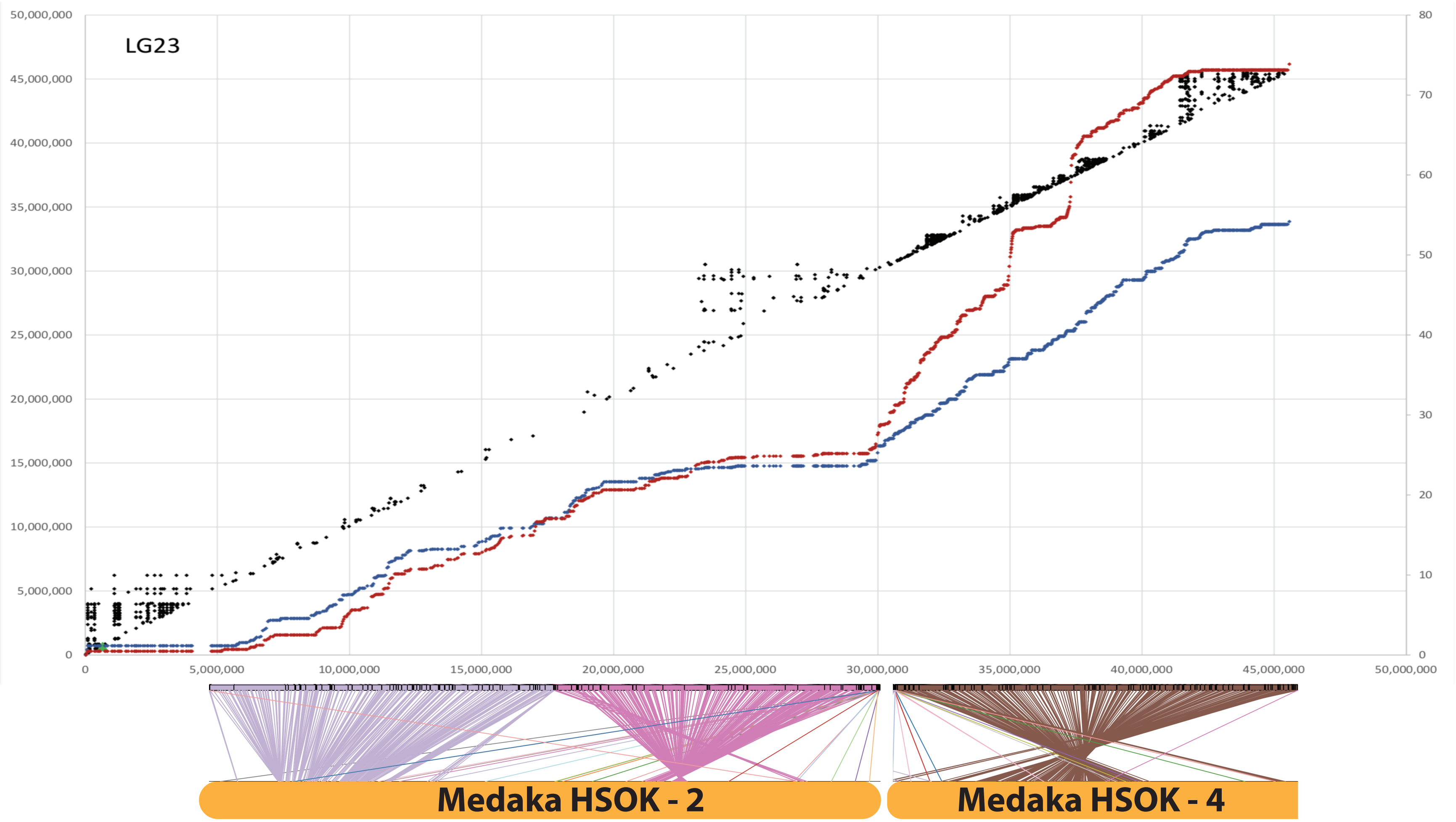


Figure 4.5. O_niloticus_UMD_NMBU LG23 is an ancient fusion of medaka HSOK 2 and part of medaka HSOK 4.

4.3.8 Repeat landscape of the *Metriaclima zebra* assembly

Similar to the O_niloticus_UMD1 assembly which is 37% repetitive (29), the M_zebra_UMD2 assembly is 35% repetitive. Figure 4.6 shows the repeat landscape for the *M. zebra* and *O. niloticus* assemblies. While the *O. niloticus* genome assembly does have a slightly larger total amount of repeats, the *M. zebra* genome assembly has a noticeably larger amount of recent TE insertions (sequence divergence < 2%). To test that this observation was not an artifact of differences between the two assembly processes, we assembled the *M. zebra* PacBio reads at the same coverage, with the same parameters, using the same software version and on the same compute cluster as was performed for the O_niloticus_UMD1 assembly. RepeatMasker was subsequently run on this assembly and the pattern of more recent insertion became more pronounced (Appendix H).

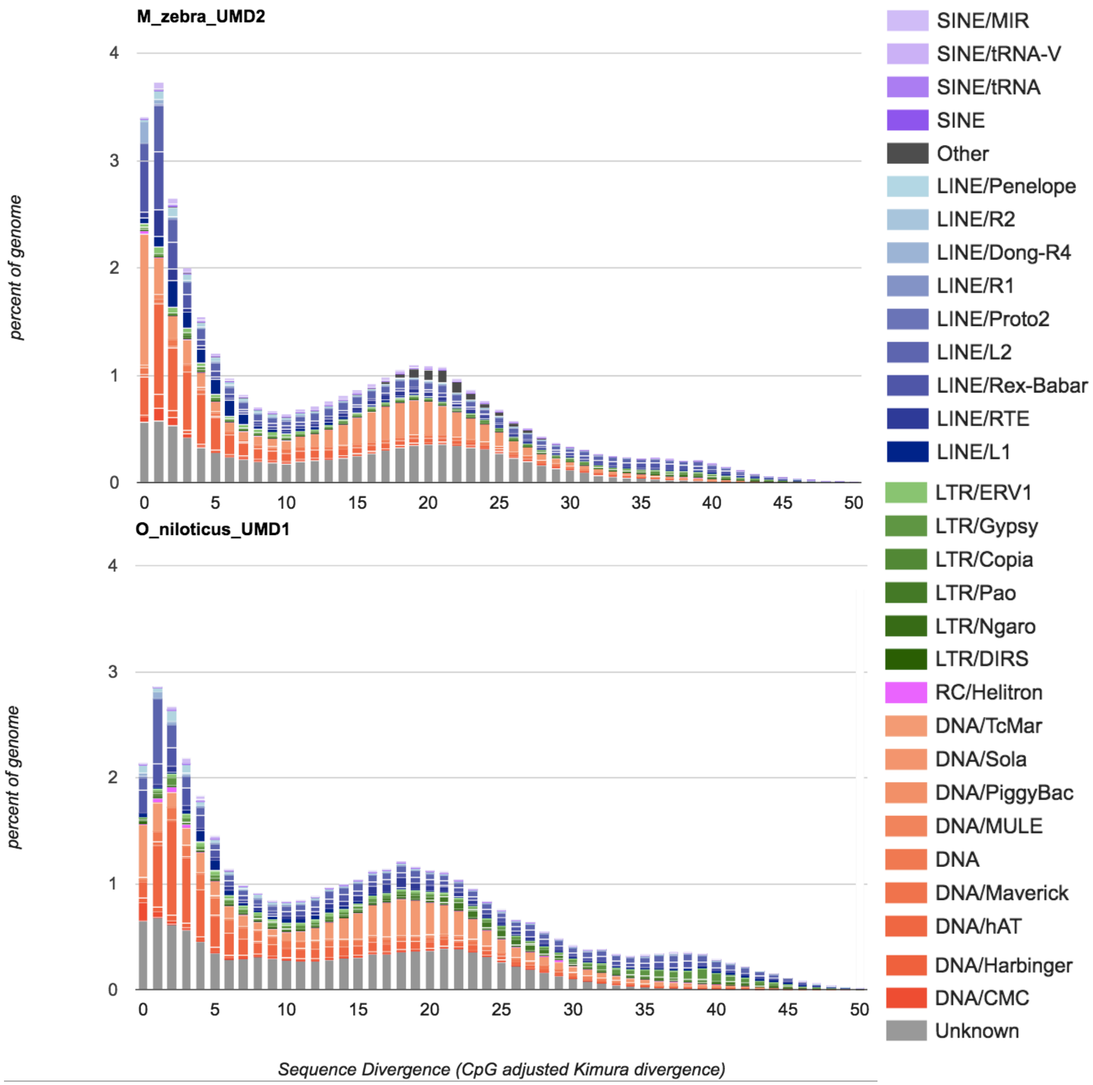


Figure 4.6 - Comparison of the repeat landscape in the *M. zebra* and *O. niloticus* genome assemblies.

Three TE families account for the largest differences in the recent TE activity difference seen between the two species. The class II DNA transposon super family, Tc1-Mariner, makes up 0.5% of the total O_niloticus_UMD1 assembly, whereas it makes up 1.3% of the *M. zebra* assembly for recent insertions with 0-1% sequence divergence. Another class II DNA transposon super family, hAT, is present at 0.15% in O_niloticus_UMD1, but present at 0.45% in the equivalent *M. zebra* assembly for recent insertions with 0-1% sequence divergence. The class I retrotransposon super family, LINE-Rex-Babar, is present 0.2% in the O_niloticus_UMD1 assembly, but present at 0.6% in the equivalent *M. zebra* assembly for recent insertions with 0-1% sequence divergence. Other TE super families show smaller increases in *M. zebra* as well. This indicates that *M. zebra*, and perhaps Lake Malawi cichlids in general, have experienced more recent TE expansion than the riverine counter-part, *O. niloticus*.

Overall, the amount of TEs assembled has increased from the original Illumina-only based *M. zebra* assembly(158), to the moderate PacBio coverage gap-filled M_zebra_UMD1 assembly (84), and now with the M_zebra_UMD2 assembly. Appendix I provides a comparison of the repeat landscape for each of these three *M. zebra* assemblies. The overall number of TEs and particularly, the most recently inserted TEs are better represented as the assemblies improve. The African Cichlid-specific AFC-SINEs and AFC-LINEs (162), have been assembled in greater length as well. For example, the “L1-1_AFC” LINE was assembled into 2,874 copies (across 1.29Mbp) in the original M_zebra_v0 assembly, 1,350 copies (across 1.66Mbp) in the M_zebra_UMD1 assembly and 2,295 copies (across 4.77Mbp) in the new M_zebra_UMD2 assembly.

4.3.9 Genome completeness

Benchmarking Universal Single-Copy Orthologs (BUSCO) (92,163) was used to assess the completeness of the new *M. zebra* genome assembly. 2,586 complete vertebrate BUSCOs were searched and 2,465 (95.3%) complete BUSCOs were found, 71 of which duplicated (2.7%) and 2,394 that were single-copy. Only 39 (1.5%) BUSCOs were reported as missing and 82 (3.2%) were reported as fragmented.

4.4 Discussion

4.4.1 Anchoring to produce chromosome-scale assemblies

The genetic maps and whole genome alignment comparison to the *O. niloticus* assembly were very useful in identifying large and mostly inter-chromosomal misassemblies in the new *M. zebra* assembly. A 40kb Illumina jumping library was also used in this process to determine if disagreements between the maps and the assembly were true misassemblies, errors in the maps, or structural differences between samples. It is likely that several misassemblies still remain in the final M_zebra_UMD2 anchoring. However, these potential misassemblies are probably only present on smaller contigs where there were not enough markers to detect misassembly events. An anchoring analysis that combined the anchored assemblies from all four maps resulted in a slightly more anchored assembly (833Mbp total compared to 760Mbp for M_zebra_UMD2). However, the ordering of contigs in this combined anchored assembly was far less accurate (when aligned to *O. niloticus*) and so it was not used. However, if the four maps were able to anchor 833Mbp combined, then this portion (87%) of the assembly was also checked for misassemblies. There was only a single contig longer than 1Mbp (“000254F”) that was not anchored by at least one map. Therefore, any possible remaining misassemblies are likely to occur on these smaller contigs.

4.4.2 Patterns of continuity in genome assemblies

The longest contigs tend to be anchored in the middle of LGs and in regions where there is greater recombination. The ends of LGs, typically in regions of lower recombination, tend to have smaller contigs. Perhaps the clearest example of this is on LG13 (Appendix D and Appendix F). On LG7, smaller contigs appear in the middle of the LG where there is also a reduction in recombination uncharacteristic of most other LGs. Regions abundant with smaller contigs are likely the result of large repetitive regions that could not be assembled completely and caused a more fragmented assembly. These regions have likely accumulated large TE arrays, unable to be spanned by even the longest of the reads in our datasets. It is known that TEs accumulate in regions of suppressed recombination, but it is still unclear if this is due to relaxed ectopic recombination or a reduction in the efficacy of selection to remove these insertions (164). These regions abundant with smaller contigs also tend to have more structural rearrangements relative to *O. niloticus*. This pattern could also be caused by ambiguities in the maps due to there being fewer recombination events and therefore less map resolution in these regions. There are also fewer markers used to anchor smaller contigs that may also contribute to this pattern. Orthogonal mapping technologies, such as optical mapping, that do not rely on recombination will be needed to attempt to resolve the structure of these regions in finer detail in future studies.

4.4.3 Diploid assembly

We present the new *M. zebra* assembly in both haploid and diploid representations. The majority of current genomics tools assume a haploid reference assembly and all subsequent analysis is based on the initial use of this haploid representation. The use of multiple diploid assemblies will be required to capture population level patterns of heterozygosity and complex structural variation. This genome assembly should be the beginning of a larger effort to properly represent cichlid genomes. A study of *Arabidopsis thaliana* and *Vitis vinifera* (Cabernet Sauvignon) showed that a phased diploid assembly produced by FALCON-unzip improved identification of haplotype structure and heterozygous structural variation (165). Sequencing and assembly of F1 in cattle has also been shown to recover these complex regions better and may be the way forward for assembly of diploid genomes (166). Additional diploid long-read assemblies will be able to better describe the variation particularly in regions of complex variation where current long read assemblies are beginning to span such regions (167). Moving beyond a haploid reference has begun and the advantage of using graph genome representations (168,169) has been shown to improve variant calling in these complex regions such as the human leukocyte antigen (HLA) (170), major histocompatibility complex (MHC) (171) and centromeres (172).

4.4.4 Patterns of recombination in *O. niloticus*

Several patterns are evident in the recombination maps for *O. niloticus*. First, the level of recombination in females is generally higher than in males. The total female map length is 1,641 cM, while the male map is only 1,321 cM. The sex differences in recombination rate are smaller than observed in salmonids (173,174), and the pattern of recombination is generally similar in males and females. Second, the pattern of recombination on each chromosome is usually sigmoidal, with relatively little recombination over about 5Mb at each end of the chromosome. The highest levels of recombination are found in the middle of each chromosome. This pattern is exactly opposite the pattern observed in stickleback and catfish, where recombination is highest at the ends of the chromosomes (175,176).

These patterns of recombination have implications for the pattern of linkage disequilibrium along each chromosome. Linkage disequilibrium will be more extensive in regions of low recombination near the ends of each chromosome. Regions of high linkage disequilibrium are likely to accumulate repetitive elements. Regions of high linkage disequilibrium are also likely to experience episodes of genetic hitchhiking, which will alter the pattern of genetic variation across the genome. The pattern of linkage disequilibrium also has implications for the probability of fixation of adaptive variants and may affect the probability that a given chromosomal segment can evolve into a new sex chromosome. The patterns of recombination and LD should be carefully considered when interpreting results from genome-wide association studies.

4.4.5 Patterns of recombination in Lake Malawi cichlids

The four genetic maps of Lake Malawi cichlids show the same general pattern of recombination as *O. niloticus*. Again, the pattern of recombination on most Lake Malawi LGs is characterized by low recombination at the ends of the LGs and high recombination in the middle of the LGs. Several exceptions all indicate intra-chromosomal rearrangements among the Lake Malawi species, or between the Malawi species and *O. niloticus*.

Perhaps the most striking difference between these four maps is a large (~19Mbp) putative inversion on LG11 in *Aulonocara,* as evidenced by the lack of recombination in the *M.*

mbenjii x *A. koningsi* and *M. mbenjii* x *A. baenschi* maps (Appendix F). The *M. zebra* x *M. mbenjii* map does not contain this putative inversion and shows the normal recombination pattern across LG11. It is possible that this large region of no recombination is associated with a sex-determination region on LG11 in *Aulonocara.* The *L. fuelleborni x Tropheops 'red cheek'* cross appears to show a different structural arrangement on LG11 but shows no evidence of suppressed recombination. There also appears to be a large inversion on LG20 in *Aulonocara*. Both of the *Aulonocara* maps show highly reduced recombination across a 15Mb region of this chromosome (Appendix D).

All four crosses showed a reduction in recombination for the first ~15Mbp on LG2 (Appendix F) that corresponds exactly with a structural rearrangement relative to *O. niloticus* (Appendix D). There is no such reduction in recombination on LG2 in *O. niloticus* (Appendix G).

Recombination is also reduced in the middle of *M. zebra* LG7, centered at ~32Mbp, that is not associated with the centromere (located at 61Mbp). It should also be noted that there is a single marker in this region that appears out of order in the *M. zebra* x *M. mbenjii* map, perhaps indicating a structural difference relative to *O. niloticus* (Appendix D and Appendix F). This region is near a previously identified sex determination locus on LG7 (177).

4.4.6 Patterns of recombination on sex chromosomes

Sex chromosomes typically accumulate inversions that reduce recombination between the sex determining gene and linked sexually antagonistic alleles (178). In the strain of *O. niloticus* studied here, sex is determined by an XY locus on LG23 (Li et al. 2015), and we observed reduced recombination in males relative to females adjacent to the sex locus at 34.5Mbp on *O.*

niloticus LG23 (Appendix G). We also observed significant differences in recombination between the sexes on LG7, LG11, LG14 and LG15. An XY sex locus has been identified on LG14 in *O. mossambicus* (Gammerdinger, *in submission*), and XY sex loci have been identified on LG7 (25) and LG11(unpublished) in Lake Malawi cichlids. The current sex-specific patterns of recombination in *O. niloticus* might represent more ancient sex chromosomes, or these particular chromosomes might be predisposed to become sex chromosomes because of inherent sex-specific differences in recombination.

The new anchoring provides the most complete assembly to date of LG3, the largest chromosome in the karyotype. This chromosome carries a ZW sex locus in several species of *Oreochromis* (29,179). The first 30 Mbp of LG3 shows a standard rate and pattern of recombination. However, the remaining 60Mbp exhibits almost no recombination in either males or females. This region of highly reduced recombination contains the ZW sex locus, and a very high density of repetitive elements. It is not clear whether the low recombination rates in this region are a consequence of the ZW sex locus, or whether an inherently low rate of recombination predisposed this region to become a sex chromosome.

4.4.7 Conservation of ancient synteny

Synteny is remarkably conserved among even distantly related teleosts (150,180). Medaka show few inter-chromosomal rearrangements since shortly after the fish-specific whole genome duplication more than 300 MY ago (146). Our whole genome alignment of tilapia to medaka supports the previously reported findings that the syntenic organization of teleost genomes is largely stable. The ancestral teleost chromosome number was 24, and contraction of diploid chromosome numbers are usually the result of chromosome fusion and/or translocation events

(150). In cichlids, where the most common chromosome number is 22 (24), we find evidence for two large fusion events on LG7 and LG23 and additional translocations on LG15 and LG17. Cichlid LG7 corresponds to a fusion of medaka chromosomes 6 and 12, while cichlid LG23 is a fusion of medaka chromosomes 2 and 4. Clearly, the variation in diploid number observed in other cichlid species implies there have been additional more inter-chromosomal rearrangements, but we predict these will be simple fission/fusion events and not the result of homogenization of these ancient syntenic relationships.

The patterns of recombination across these particular LGs provide additional evidence of fusion and translocation events (Appendix F and Appendix G). There are large deviations from the slope of the recombination curves located precisely where we suggest that these fusion and translocation events have occurred. This also suggests that the pattern of recombination evolves slowly, as these oddly shaped recombination patterns have persisted for at least ~15 million years since the divergence of the common ancestor of *O. niloticus* and the Lake Malawi species. Interestingly, the odd pattern of recombination on LG3 does not seem to be the result of a chromosome fusion event. This lends support to the hypothesis that LG3 has been accumulating repetitive sequences after it became a sex chromosome.

There are many examples of large-scale (>2Mbp) intra-chromosomal rearrangements between *O. niloticus* and Lake Malawi cichlids, as well as rearrangements evident between the Lake Malawi species. In some cases, the anchoring of the *M. zebra* assembly using each map showed the same large structural rearrangement relative to *O. niloticus* for each map (see LG2, LG19, LG20 in Appendix D). This suggests that these rearrangements happened prior to the Lake Malawi radiation. In other cases, there are large structural differences relative to *O. niloticus* for each map, but these are different between the four maps (LG12, Appendix D). This suggests that these rearrangements occurred during the radiation in Lake Malawi. For example, on LG11, the *M. zebra* x *M. mbenjii* map is mostly collinear with *O. niloticus*, the other three maps show a large rearrangement. The other three maps also show some differences in the order of this rearrangement. LG9 of *M. zebra* was particularly difficult to anchor with the *M. mbenjii* x *A. koningsi* map (Table 4.3). We believe this may indicate a hybrid incompatibility locus on LG9 in this cross. Additional work is needed to better define the structure of these chromosomes in each lineage.

4.4.8 Evolution of centromere position and sequence

Long-read sequencing has made it possible to assemble centromere repeats (167,181,182). A recent study of centromere evolution in medaka provides an example of the role of centromere evolution in speciation (149). The study showed that the centromere position of a certain set of medaka chromosomes has remained unchanged in both acro-centric and non-acro-centric chromosomes. In other chromosomes, the position of centromeres did change and involve chromosomes that have undergone other major structural rearrangements. Alignment of the O_niloticus_UMD_NMBU assembly to these new medaka assemblies showed that this pattern was not the same in cichlids where different chromosomes have remained relatively static and others have evolved more structural changes. Additionally, the medaka study showed that centromere sequence repeats were more conserved in the chromosomes that remained acro-centric than in chromosomes that switched between acro- and non-acro-centric or that were non-acro-centric. Assembly and placement of cichlid centromere repeats in multiple species will allow for both refining previous karyotype studies in the context of whole genome assembly comparisons, but also centromere evolution at the sequence level. Are there differences in centromere sequence/rate of evolution between non-acro-centric and acro-centric chromosomes? Are these differences great enough to create meiotic incompatibilities in hybrids? Are the positions of centromeres conserved across man species? This study provides a starting point to begin to answer these questions.

4.4.9 Evolutionary patterns of African cichlid chromosomes via karyotyping and genome assembly

The karyotypes of *O. niloticus* and *M. zebra* in Figure 4.1 show that there have been at least 5 or 6 changes from subtelo-acrocentric to meta-submetacentric chromosomes. The clearest example of this in the new genome assemblies is the 15Mbp rearrangement on LG23 (Figure 4.2). Additionally, three similar centromere location changes have happened on LG3, LG4 and LG16 (Appendix D). We were able to identify centromere-containing repeats on both the *M. zebra* and *O. niloticus* assemblies in just over half of the LGs (LG3, LG4, LG5, LG7, LG8, LG9, LG11, LG13, LG14, LG16, LG17, LG19, LG23). The ONSATA and TZSAT satellite sequences (99) have not explicitly been shown as the centromeric binding sequences, but rather highly associated in the centromeres via *in situ* staining (98). It is possible that these ONSATA and TZSAT repeat sequences may be present in other portions of the chromosome, or that some of them have been assembled incorrectly. Indeed, there are several LGs where the ONSATA and TZSAT repeats were identified in multiple distant locations along the chromosome in one or both assemblies (LG6, LG16, LG17, LG19). On LG6 a centromere was not identified in the *M. zebra* assembly, but it does appear to have undergone a centromere location change.

Two of the LGs where we have identified karyotype changes, have also been shown to harbor sex-determining loci in African cichlids. The first was the previously mentioned XY sex determination region in *O. niloticus* on LG23 (82). On LG3, a WZ sex determination region has been previously identified (179) and characterized (29) and reanalyzed on the O_niloticus_UMD_NMBU assembly (Appendix E). There is a very wide region of sex-patterned differentiation from LG3 at ~40Mbp to 85Mbp. This same region corresponds with the low recombination in male and female *O. niloticus*. The largest LG in the *O. niloticus* karyotype is LG3, although LG7 is the largest *M. zebra* chromosome. The assembled and anchored LGs support these karyotypes (Figure 4.1, Table 4.3 and Table 4.1). We suggest that LG3 has expanded from the ancestral state in the *O. niloticus* lineage, by accumulation of a large amount of TEs and segmental duplications and is likely involved in the sex determination region on *O. niloticus* LG3 (29). It is difficult to determine if this apparent runaway elongation of LG3 in *O. niloticus* is due to the sex-determination locus or if recombination was suppressed first due to some other process. Additional genome-assemblies of similar quality in related *Oreochromis* species that also harbor the LG3 sex-determination system should allow for further refinement of the evolutionary history of this large tilapia sex chromosome.

Similar to LG3, there is a large (~28Mbp) region of greatly reduced recombination on LG23 in each of the four Lake Malawi maps as well as the *O. niloticus* map. LG23 is also the second largest anchored LG in the *M. zebra* assembly and third largest LG in the *O. niloticus* assembly. It is possible that this arm of LG23 is accumulating TEs similar to LG3. There is an XY sex determination locus on LG23 (78,82) which may be driving or contributing to the expanding effect that is seen. However, while LG23 has been shown to be a sex determination chromosome in *O. niloticus,* LG23 has not been shown to harbor a sex locus in Lake Malawi. Three scenarios may explain these observations: 1) That LG23 is an older sex chromosome still sorting in *Oreochromis* genus, but has maintained the recombination pattern in riverine and Lake

Malawi cichlids; 2) That there is a LG23 sex determination locus sorting in Lake Malawi that has yet to be identified and described; 3) The recombination pattern on LG23 is not necessarily involved in the sex determination pattern on LG23 and has been maintained for some other unknown reason in both lineages.

Many LGs have shown extensive rearrangement, but it should also be noted that several LGs have undergone very little change since the divergence of *M. zebra* and *O. niloticus*. Other than relatively small structural changes at the ends of LGs, conserved synteny seems to have been maintained across the entire length of LG13, LG14, LG17 and LG18. It is possible that selective pressures have acted to maintain the synteny of these LGs or that synteny has been maintained by chance. Since 20% of the *M. zebra* and 10% of the *O. niloticus* genome assemblies remain unanchored, future studies may provide additional structural insights. For example, LG9 in *M. zebra* remains under-anchored. Future *in situ* studies should confirm these results in *O. niloticus* and *M. zebra*. Moreover, our work will greatly inform more fine-scale cytogenetic studies to be performed by providing many starting points for intra-chromosomal differences in cichlids to be studied.

4.4.10 Recent transposable element expansion in M. zebra

Recent evidence has shown that cis-regulatory AFC-SINE insertions are highly associated with innovative cichlid phenotypes such as egg-spots (154) and a deletion that may be TE-mediated responsible for controlling the expression of *SWS2A* opsin (155). It is likely that other AFC-specific and other TE-mediated mutations have also contributed to the diverse phenotypes of African cichlids. Therefore, it is important that these TE insertion events are well represented in the genome assemblies.

Figure 4.6 shows a comparison of the repeat landscapes for *M. zebra* and *O. niloticus* assemblies. *M. zebra* has a higher amount of recent TE insertions (sequence divergence < 2%) than *O. niloticus*. Since the *O. niloticus* assembly is 43.4Mbp longer than the *M. zebra* assembly, it is possible that the difference in recent TE insertions is even greater than what we see. Each new version of the *M. zebra* genome assembly has improved upon the AFC-specific and TE super families in general.

We present this finding with several caveats. It is possible that the two species have divergent patterns of insertions across the genome. We suggested *O. niloticus* contains larger clusters of repeat arrays that are experiencing recent insertions (29). These arrays do not seem to be present in the *M. zebra* genome. It is possible that many of the recent TE insertions in *O. niloticus* were not assembled and remain hidden in these large arrays. DNA of the two samples were extracted and sequenced at different times and the *M. zebra* dataset included 16.5X coverage using a different PacBio chemistry (P5-C3). Other unknown technical factors may also have contributed to the difference that we have described. It is also possible that *O. niloticus* may have a different but active TE superfamily that is too long to be assembled with our current read lengths. Future comparisons of additional samples and species assembled using the same sequencing coverage and assembly software/parameters will be useful in more accurately quantifying the recent TE expansion in African Great Lake cichlids.

4.5 Potential implications

This study highlights evolutionary insight that can be gained using a comparison of high-quality chromosome-scale genome assemblies, genetic recombination maps and cytogenetics across multiple related and, in this case, rapidly evolving species. It further illustrates the necessity of high-quality, chromosome scale genome assemblies for answering many basic biologically relevant questions. The study will serve as a unique example of the structural changes that have happened in the genomes of rapidly evolving clade and should prove interesting to compare to other radiations in the tree of life, both large and small. This study provides a wide-angle view of the African cichlid genome history by demonstrating how these high-quality resources can be used for many different types of evolutionary genomic analyses going forward. As additional high-quality cichlid genomes are generated, this study provides the groundwork for comparisons of structural, recombination, cytogenetic and repetitive sequences across the cichlid phylogeny. Many new questions have been generated here. How do the structural changes of African cichlid genomes compare to other groups? Is the pattern of few inter-chromosomal, but many intra-chromosomal differences seen here in Lake Malawi cichlids similar in additional Malawi genera as well as other radiations in Lake Tanganyika and Lake Victoria? Are these patterns of recombination observed across the majority of cichlids? Are any deviations from these typical recombination patterns related to specific phenotypic patterns and sex chromosome history and how have they evolved structurally? We look forward to the renaissance in cichlid genomics that is coming.

4.6 Methods

4.6.1 *O. niloticus* SNP array map, misassembly detection and new anchoring

Offspring (n=689) and parents from 41 full-sib families belonging to the 20th, 24th and 25th generations of the GST® strain were analyzed using a custom 57K SNP Axiom® Nile Tilapia Genotyping Array (*in preparation*). SNPs classified as "PolyHighRes" or "No-MinorHom" by Axiom Analysis Suite (Affymetrix, Santa Clara, USA), and having a minor-allele frequency ≥ 0.05, and call rate ≥ 0.85 were used in genetic map construction (n= 40,548). Lep-MAP2 (183) was used to order these SNPs into linkage groups in a stepwise process beginning with SNPs being assigned to linkage groups using the 'SeparateChromosomes' command. LOD thresholds were adjusted until 22 linkage groups, which correspond with the *O. niloticus* karyotype. Unassigned SNPs were subsequently added to linkage groups using the 'JoinSingles' command and a more relaxed LOD threshold, and ordered within each linkage group using the 'OrderMarkers' command.

Sequence flanking each SNP (2 x 35nt) was used to precisely position 40,190 SNPs to the O_niloticus_UMD1 assembly (MKQE00000000) and thereby integrate the linkage and physical maps. This revealed 22 additional contig misassemblies (i.e. contigs containing SNPs from different LGs) that were not detected in the original anchoring for O_niloticus_UMD1. These contigs that were subsequently broken. Linkage information was subsequently used to order and orientate contigs and build sequences for 22 Nile tilapia LGs in the new O_niloticus_UMD_NMBU assembly following the previous cichlid nomenclature (5,29,95,125) .

4.6.2 PacBio Sequencing of *M. zebra*

The previous version of the *M. zebra* assembly, M_zebra_UMD1 (84), included 16.5X PacBio sequencing (25 SMRT cells using the P5-C3 chemistry) on an PacBio RS II machine (84). An additional library was prepared using the same Qiagen MagAttract HMW DNA extraction and Blue Pippin pulse-field gel electrophoresis size selection that was previously sequenced. An additional 60 SMRT cells (using the P6-C4 chemistry) were sequenced on the same PacBio RS II at the University of Maryland Genomics Resource Center as the previous 16.5X P5-C3 data. These P6-C4 SMRT cells comprised 50X coverage to bring combined total to ~65X coverage.

4.6.3 *M. zebra* diploid genome assembly

The 65X PacBio reads were assembled using FALCON-integrate/FALCON_unzip (*version 0.4.0*) (86). The following parameters were used for the '*fc_run.py*' assembly step:

length_cutoff = 9000

length_cutoff_pr = 9000

pa_HPCdaligner_option = -v -dal128 -H10000 -M60 -t16 -e.70 -l2000 -s100 -k14 -h480 -w8

ovlp_HPCdaligner_option = -v -dal128 -H10000 -M60 -t32 -h1024 -e.96 -l1000 -s100 -k24

falcon_sense_option = --output_multi --min_idt 0.70 --min_cov 4 --max_n_read 350 --n_core 5

overlap_filtering_setting = --max_diff 100 --max_cov 150 --min_cov 0 --bestn 10 --n_core 18

This was followed by the unzip step (‘*fc_unzip.py*’) and quiver polishing of the diploid assembly with the ‘*fc_quiver.py*’ assembly step.

4.6.4 Polishing of the *M. zebra* diploid genome assembly

The diploid assembly described above includes a PacBio polishing (quiver) step. However, there were also Illumina reads available to for *M. zebra* from the first version of the assembly (5). Trimming and filtering of the raw *M. zebra* Illumina reads are described for the previous version of the assembly (84). The trimmed and filtered fragment library corresponded to 30.1X coverage and the trimmed and filtered 2-3kb library corresponded to 32.6X coverage for a total of 62.7X Illumina coverage. These Illumina reads were aligned to the diploid assembly with BWA mem (122) (*version 0.7.12-r1044*). Pilon (119) (*version 1.22*) was run supplying the fragment library with the ‘*--frags*' option, the 2-3kb library with the ‘*--jumps*' option and the following options: ‘*--diploid --fix bases --mindepth 10 --minmq 1 --minqual 1 --nostrays*'.

This intermediate, Illumina-polished assembly was then polished again with the PacBio reads using SMRT-Analysis (68) (*version 2.3.0.140936*) using the 65X raw PacBio reads. First, each SMRT cell was separately aligned to the intermediate polished assembly using pbalign (*version 0.2.0.138342*) with the ‘*--forQuiver*' flag. Next, cmph5tools.py (*version 0.8.0*) was used to merge and sort (with the ‘*--deep*’ flag) the pbalign .h5 output files for each SMRT cell. Finally, Quiver (*GenomicConsensus version 0.9.2* and *ConsensusCore version 0.8.8*) was run on the merged and sorted pbalign output to produce an initial polished assembly.

4.6.5 Detecting misassemblies in *M. zebra*

To detect misassemblies present in the intermediate polished assemble, several datasets were analyzed and compared. This included four genetic maps: A genetic map with 834 markers generated from RAD genotyping of 160 F_2 individuals from a cross of *M. zebra* and *M. mbenjii* (66); a genetic map with 946 markers generated from RAD genotyping of 262 F_2 individuals from a cross of *Labeotropheus fuelleborni* and *Tropheops 'red cheek'* (157); a genetic map of 2,553 markers generated from RAD genotyping of 331 F_2 individuals from a cross of *M. mbenjii* and *Aulonocara koningsi* (cross and map construction details in separate Methods section); a genetic map of 1,217 markers generated from RAD genotyping of 161 F_2 individuals from a cross of *M. mbenjii* and *A. baenschi* (cross and map construction details in separate Methods section).

The markers for each of the four maps were aligned to the intermediate polished assembly using BWA mem (122) (*version 0.7.12-r1044*) and a separate SAM file was generated. Chromonomer (124) (*version 1.05*) was run for each map using these respective SAM files and map information as input. Chromonomer detected contigs in the intermediate assembly that were mapped to multiple linkage groups.

To narrow the location of these identified misassemblies, the Illumina 40kb mate-pair library from the first *M. zebra* assembly (5) was aligned to the intermediate assembly. The raw PacBio reads were aligned using BLASR (54) (version 1.3.1.127046) with the following parameters: *'-minMatch 8 -minPctI- dentity 70 -bestn 1 -nCandidates 10 -maxScore -500 -nproc 40 -noSplitSubreads –sam*'. Regions of abnormal coverage in the PacBio read alignments as well as abnormal clone coverage in the 40kb mate-pair were identified for most potential misassemblies identified by the genetic maps. These misassembly regions were manually inspected using these alignments in IGV (123). Additionally, RefSeq (93) (*release 76*) *M. zebra* transcripts were aligned to the intermediate assembly using GMAP (55) (*version 2015-07-23*) and RepeatMasker (65) repeat annotations were considered when defining the exact location of a misassembly break.

One additional misassembly was identified during the comparison of linkage maps (next section) and was subsequently broken using the same process as above.

4.6.6 *M. zebra* assembly anchoring

The same four genetics maps used above for misassembly detection were also used for anchoring the assembly contigs (after breaking) into the final set of linkage groups. Chromonomer (124) (*version 1.05*) was run on each of these four genetic maps to anchor the polished and misassembly corrected contigs. BWA mem (*version 0.7.12-r1044)* was used to create the input SAM file by aligning respective map marker sequences to these contigs. Gaps of 100bp were placed between anchored contigs. The final M_zebra_UMD2 anchoring was generated by anchoring LG9 and LG11 with the *M. zebra* and *M. mbenjii* map (66), LG20 with the *M. mbenjii* and *A. baenschi* map and the remaining 19 LGs with the *M. mbenjii* and *A. koningsi* map. To accomplish this anchoring, the markers for each of those respective maps and LGs were used with Chromonomer as described above.

4.6.7 *M. zebra* repeat annotation

RepeatModeler (64) (*version open-1.0.8*) was first used to identify and classify *de novo* repeat families present in the final anchored assembly. These *de novo* repeats were combined with the RepBase-derived RepeatMasker libraries (129). RepeatMasker (65) (*version open-4.0.5*) was run on the final anchored assembly using NCBI BLAST+ (*version 2.3.0+*) as the engine ('*-e ncbi*') and specifying the combined repeat library ('*-lib*'). The more sensitive slow search mode ('*-s*') was used. The repeat landscape was generated with the RepeatMasker '*calcDivergenceFromAlign.pl*' and '*createRepeatLandscape.pl*' utility scripts.

4.6.8 *M. zebra* BUSCO genome-completeness analysis

BUSCO (*version 3.0.2*) was run on the M_zebra_UMD2 anchored assembly in the genome mode (*-m geno*) and compared against the vertebrate BUSCO set ('vertebrata_odb9').

4.6.9 Whole genome alignment of *M. zebra* to *O. niloticus*

The final anchored M_zebra_UMD2 assembly was aligned to the O_niloticus_UMD2 assembly using the '*nucmer*' program of the MUMmer package (184) (*version 3.1*). The default *nucmer* parameters were used and the raw *nucmer* alignments were filtered using the '*delta-filter*' program with the following options: '*-o 50 -l 50 -1 -i 10 -u 10*'. These filtered alignments were converted to a tab-delimited set of coordinates using the '*show-coords*' program with the following options: '*-I 10 -L 5000 -l -T -H*'. This set of coordinates was then visualized using Ribbon (185).

4.6.10 Whole genome alignment of *M. zebra* to medaka

The HSOK medaka genome assembly version 2.2.4 was downloaded from http://utgenome.org/medaka_v2/#!Assembly.md and corresponds to NCBI accession (GCA_002234695.1). Similar to the M_zebra_UMD2 comparison, O_niloticus_UMD_NMBU was aligned to the medaka HSOK genome with *nucmer*. The '*delta-filter'* settings were adjusted to *‘-1 -l 50 -i 50 -u 50*’ to account for the increased divergence between the two more distantly related species. The ‘*show-coords*’ settings were also adjusted to ‘*-I 50 -L 50 -l -T -H*’. Alignments were again viewed with Ribbon to identify putative chromosome fusion and translocation events.

4.7 Declarations

4.7.1 Funding

This work was supported by the United States Department of Agriculture under project number MD.W-2014-05906 to TDK and National Institutes of Health project R01-EY024639 to KLC.

4.7.2 Author's contributions

MAC, TDK, and KLC conceived the study. TDK carried out HMW DNA extraction. MAC carried out computational analyses. ECM, SPN, and RBR performed genetic map construction. FEC and CM performed karyotype experiments. WJG organized map data for anchoring. MAC and TDK wrote the manuscript. All authors read and approved the manuscript.

4.7.3 Acknowledgements

We acknowledge all members of the UMD cichlid labs for their countless discussions on this project. We thank Luke Tallon and Naomi Sengamalay at the Genomics Resource Center, Institute for Genome Sciences for providing a high quality PacBio library and sequence reads. We acknowledge the University of Maryland supercomputing resources (www.it.umd.edu/hpcc/)

made available in conducting the research reported in this paper. We thank many individuals in the cichlid community for their patience while we developed these resources.

Chapter 5: Origin, evolution and history of B chromosomes in African cichlids.

Authors

Matthew A. Conte, Frances E. Clark, Karen L. Carleton, Cesar Martins and Thomas D. Kocher.

5.1 Abstract

B chromosomes have proven an enigmatic genomic compartment present in some, but not all, individuals of a population and found across roughly 15% of eukaryotes. B chromosomes have long been characterized via cytogenetic methods. Recently, new information about B chromosome content and organization has been discovered through genome sequencing. B chromosomes have been identified in multiple cichlid species of Lake Malawi and Lake Victoria. In all of the Lake Malawi B chromosome carrying species, B chromosomes are found solely in females. However, in Lake Victoria, B chromosomes are also found in males, and are female limited in only one of 12 species previously studied. This study compares the B chromosomes of Lake Malawi and Lake Victoria cichlids using whole genome sequencing. We find several relatively short regions (totaling 149kb) of shared ancestry of the B chromosomes in the two lakes. A large amount of the B chromosome sequence is unique to the cichlids of each lake, which indicates very rapid evolution of B chromosomes. The rapid evolution of B chromosomes is further supported by a comparison of six species across three genera within Lake Malawi. Additional comparisons within and between the B chromosomes in the two lakes may suggest an introduction of a Lake Victorian B chromosome into Lake Malawi. Additionally, long read

sequencing and *de novo* assembly of a single female B chromosome shows the dynamic DNA sequence structure of the B chromosome for the first time. This *de novo* B chromosome assembly also revealed that the transposable element activity of this B chromosome differs greatly from the A genome. Several genes identified on the B chromosome are possible candidates for sex-determination and B chromosome drive functions.

5.2 Background

B chromosomes were first identified over 100 years ago in the insect genus *Metapodius* (33). B chromosomes are non-essential, supernumerary chromosomes that are present in addition to the normal ("A") karyotype of an organism. B chromosomes can be regularly found in some, but not all, individuals of a given population. They are estimated to occur across 15% of all eukaryotes (34) covering a wide range of taxa from fungi to plants to animals, including mammals (35). B chromosomes have been well studied cytogenetically but are only recently beginning to be better understood at the genomic level (36). Originally thought to contain mostly repetitive DNA sequence and to be completely heterochromatic, recent studies have begun to show that B chromosomes do indeed contain transcribed genic sequences (32,37). The role of B chromosomes as selfish genetic elements has been described in many taxonomic groups (186).

B chromosomes are prevalent in African cichlids. In Lake Victoria, B chromosomes have been found in a subset of species and shown to play a functional role in sex determination in at least one population of cichlids, but not the majority of populations harboring B chromosomes (38). The same study also showed that the size of B chromosomes varies greatly even within the same population. We previously sequenced an individual with two B chromosomes (2B) and an individual with zero B (0B) chromosomes of the Lake Victorian cichlid *Astatotilapia latifasciata*. We were able to characterize regions of the B chromosome (B "blocks") that were homologous to sequences along the A chromosomes, which provided insights into the origin and evolution of that B chromosome (32). We also compared these B chromosome blocks to the Illumina-based genome assembly of the Lake Victorian, *Pundamilia nyererei* (158), and realized that the individual sample for this genome also carried similar B chromosome. This caused misassemblies in this *P. nyererei* assembly in the regions where the B chromosome was homologous with the A genome. This demonstrated that karyotyping is an important first step in eukaryotic genome projects, especially if B chromosomes are known to be present in closely related species. Additional analysis of the transcriptomes of this *P. nyererei* individual showed transcription of several genes from this B chromosome (32). Overall, the study described a B chromosome at the genomic sequence level for the first time, showed that most genes present on the B are fragmented, and that the genes that appear intact are transcriptionally active. We put forth a model of this B chromosome originating as a proto-B fragment from one autosome that expanded by the insertion of fragments from many chromosomes in the rest of the genome.

B chromosomes have also been karyotyped and described in a species of Lake Malawi cichlid, *Metriaclima lombardoi* (39). Recently, we sequenced over 20 different populations of Lake Malawi cichlids and have identified B chromosomes present solely in female individuals of at least 7 populations, including the previously identified *M. lombardoi* (40). We found Lake Malawi B chromosomes to be present in 13% of females and 0% of males across 323 and 317 samples, respectively.

These previous studies of B chromosomes in Lake Victoria and Lake Malawi generated many interesting evolutionary and biological questions that we address in the present study. Is there a shared evolutionary origin of the B chromosomes in Lake Victoria and Lake Malawi? If two B chromosomes share the same origin, what genomic parts do the two distinct B chromosomes share and in what ways have the B chromosomes diverged? What mechanism restricts the Lake Malawi B chromosomes to female individuals? Are the evolutionary trajectories of B chromosomes in the two African Great Lakes similar? Do the B chromosomes in the two lakes contain similar active TE sequences contributing to their repetitive content? Does B chromosome structure in both lakes show similar or different patterns? Are there detectable gene fusion events? Are the patterns of genome structure and recombination in Lake Malawi cichlids associated with sequence content and divergence patterns on these B chromosomes? This study begins to answer many of these questions while also generating many new interesting questions about the history, evolution and function of African cichlid B chromosomes.

5.3 Results

5.3.1 Comparison of B chromosome blocks to the M_zebra_UMD2 reference

A total of 2,528,172 PacBio reads totaling 20.97Gbp (~20X coverage) were obtained from an individual *M. lombardoi* female who had a B chromosome. These reads were aligned to the M_zebra_UMD2 assembly and scored for high coverage regions similar to our previously study (32) (see Methods). The total size of the B chromosome as estimated by this set of B blocks is 16.9Mbp. This represents a conservative calling of B blocks and probably underestimates the size of the B chromosome.

The distribution of B chromosome blocks across the genome can be seen in Table 5.1. There is a fairly even distribution of B chromosome content and this is similar to what was seen on the Lake Victorian *A. latifasciata* B chromosome (32). Appendix F provides plots of the distribution of B blocks along the lengths of each M_zebra_UMD2 LG. These plots also show that the distribution of B blocks is rather uniform and there does not appear to be a propensity of B content derived from any particular LG or LGs. Likewise, these plots show that there does not appear to be any discernable pattern associated with recombination or genome structure that was described in chapter 4.

M_zebra_UMD2 LG	B block span in M_zebra_UMD2 (bp)	Average B block copy number per LG	Total estimated size of B per LG
LG1	31,500	4.32	136,050
LG2	38,392	3.94	151,341
LG3	73,900	3.58	264,429
LG4	72,100	6.43	463,589
LG5	30,500	8.99	274,086
LG6	100,700	8.51	857,119
LG7	60,800	8.26	502,444
LG8	14,400	11.59	166,958
LG9	194,300	8.41	1,634,780
LG10	52,400	5.07	265,411
LG11	93,600	8.17	764,726
LG12	57,400	10.84	622,281
LG13	65,600	8.27	542,206
LG14	119,600	5.13	613,156
LG15	52,700	4.68	246,617
LG16	66,000	6.56	433,244
LG17	104,500	3.51	366,432
LG18	24,000	6.18	148,262
LG19	59,600	7.10	423,064
LG20	20,500	3.37	69,070
LG22	57,800	3.99	230,347
LG23	238,000	8.25	1,964,160

Table 5.1. Distribution and size of B blocks.

There were two B chromosome blocks that showed highest copy number on lg7:14,298,900-14,299,000 (~72 copies) and lg6:33,687,100-33,687,200 (~66 copies). On lg7 this B chromosome block includes a large portion of "inner centromere protein A" (LOC101482374), including the BED zinc finger DNA binding domain of this protein. The high copy B block on lg6 has a portion of an unannotated gene (LOC101463671) and a portion of the "catenin delta-1" gene (LOC101487010). The longest continuous Lake Malawi B block along the M_zebra_UMD2 assembly is a 172kb block on lg9: 15,993,700-16,165,800. This block contains "cadherin 18, type 2" (cdh18) and is near a region of high structural variation described in chapter 4 and again here (also see lg9 image in Appendix D).

5.3.2 Conserved and dynamic content of the Lake Malawi B chromosome blocks in Lake Malawi

To determine the extent of shared and variable regions of Lake Malawi B chromosomes, we re-sequenced 12 individual females from six species across three Malawi genera that were genotyped as having B chromosomes using our techniques described previously (40). Table 5.2 provides the sequencing results and B block sizes computed for of these individuals.

Species	Identifier	Coverage	B block bp
Metriaclima greshakei	M_greshakei_2012_3493	14.7	9,664,979
Labeotropheus trewavasae	L_trewavasae_2005_1306	15.2	13,195,343
Melanochromis auratus	M_auratus_2008_1601	14.7	7,241,178
Metriaclima zebra (Nkhata Bay)	M_zebra_NkhBay_2012_5347	16.4	12,138,194
Metriaclima zebra (Nkhata Bay)	M_zebra_NkhBay_2012_5340	13.4	9,112,260
Metriaclima mbenji	M_mbenji_2012_3997	14.7	6,632,183
Metriaclima lombardoi	M_lombardoi_2014_1108	11.9	17,900,132
Metriaclima lombardoi	M_lombardoi_2014_1021	17.3	14,873,538
Metriaclima lombardoi	M_lombardoi_2014_1018	16.4	17,324,878
Metriaclima zebra (Boadzulu island)	M_zeb_boadzulu_2005_0986	12.6	10,283,813
Metriaclima zebra (Boadzulu island)	M_zeb_boadzulu_2005_0976	15.4	11,560,422
Metriaclima zebra (Boadzulu island)	M_zeb_boadzulu_2005_0983	14.9	10,533,188

Table 5.2. Individual samples containing B chromosomes that were re-sequenced.

B chromosome blocks for each of these individuals were computed and blocks that were present in every sample were considered the "core" content of the Lake Malawi B chromosome. The core content of the Lake Malawi B chromosome spans 1.28Mbp of A chromosome space in the M_zebra_UMD2 assembly.

However, the B chromosome blocks vary considerably between individuals and species. B chromosome blocks that are not part of the core blocks, "variable blocks", are typically limited to a single individual or a particular species. The total size of the Lake Malawi variable blocks was 64.3Mbp of A chromosome space across the M_zebra_UMD2 assembly. B chromosomes in Lake Malawi are composed of a core and conserved set of blocks derived from a common ancestor. Lake Malawi B chromosomes are also composed of a variable set of blocks that are dynamic and can be added or lost in different species and individuals.

5.3.3 History of African cichlid B chromosomes

The same 0B and 2B Lake Victorian *A. latifasciata* samples from our previous study (32) were aligned to the M_zebra_UMD2 assembly and B blocks were called. Several regions were determined as shared B blocks in both Lake Victoria and Lake Malawi. A total of 149.7kbp is shared between the Lake Victoria 2B blocks and Lake Malawi core B blocks. Of note, the longest shared region between the two B chromosomes were two blocks on lg23 at 15.7Mbp that are 7.8kbp and 3.9kb in length. This region of lg23 is also at the exact breakpoint of the ancient cichlid fusion of this chromosome (Figure 4.5 depicts this in the *O. niloticus* assembly), and is discussed further below. An inspection of the Lake Malawi B chromosome-specific alleles in these regions revealed that they largely correspond with Lake Victorian alleles, but not necessarily Lake Victoria B-specific alleles (Figure 5.1). These Lake Victorian alleles present on the Lake Malawi B chromosome do not appear to be present in Lake Malawi samples without a B chromosome. An inspection of Lake Victorian specific B blocks does not show allele sharing with Lake Malawi samples.

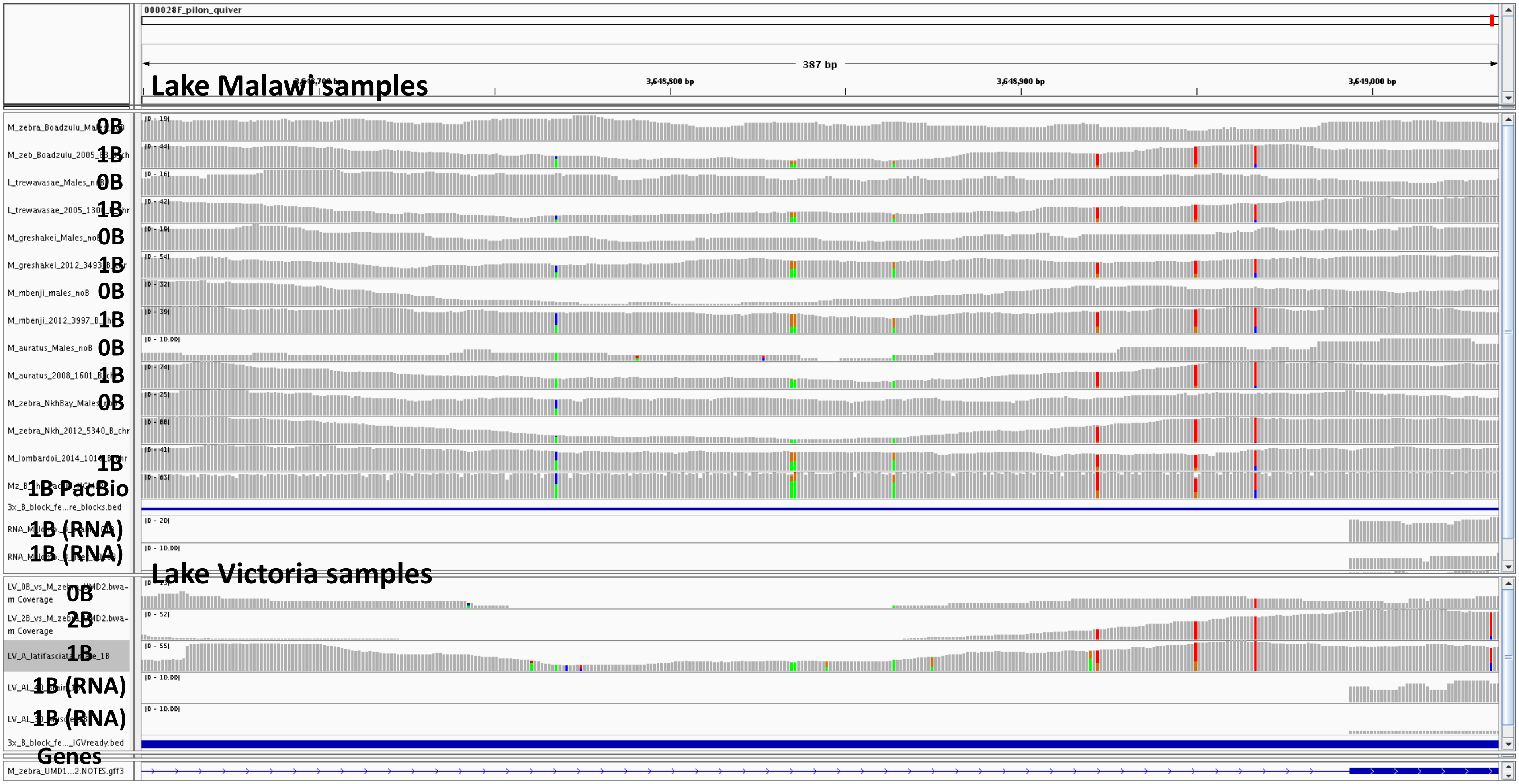


Figure 5.1. Lake Malawi B chromosomes show shared alleles with Lake Victoria. Samples with B chromosomes in Lake Malawi show the same alleles as samples with and without B chromosomes in Lake Victoria.

Phylogenetic trees of B and no B carrying samples from Lake Malawi and Lake Victoria as well as an outgroup Lake Tanganyika cichlid were constructed to further refine this initial finding of shared allelism. A phylogeny of the whole genome is shown in Figure 5.2. The samples cluster according to the species tree of African cichlids with the Lake Victorian *A. latifasciata* samples clustering together, the Lake Malawi B and no B samples clustering by species, and the *Neolamprologus brichardi* sample from Lake Tanganyika clustering as the outgroup. Several additional phylogenies were generated from subsets of the genome reference based on the presence of core, variable and shared B blocks. First, a phylogeny was generated only for regions where core Lake Malawi B chromosome blocks were called. This includes many regions where the Lake Malawi B chromosome shares alleles with Lake Victoria and the phylogenetic tree in Figure 5.3 depicts this pattern. Second, to test the opposite case, a phylogeny was generated only in regions where Lake Victorian B chromosome blocks were called. In this case, the Lake Victorian samples do not cluster with anything from Lake Malawi (Figure 5.4). Next, a phylogeny was generated only within the relatively short content of regions that were shared B blocks in both Lake Malawi and Lake Victoria (Figure 5.5). Figure 5.5 shares a very similar topology to Figure 5.3, meaning that the core Malawi B blocks and the B blocks shared with Lake Victoria have a shared history. Finally, a phylogeny was generated from the variable Lake Malawi B chromosome blocks and these blocks largely match the species tree (Figure 5.6).

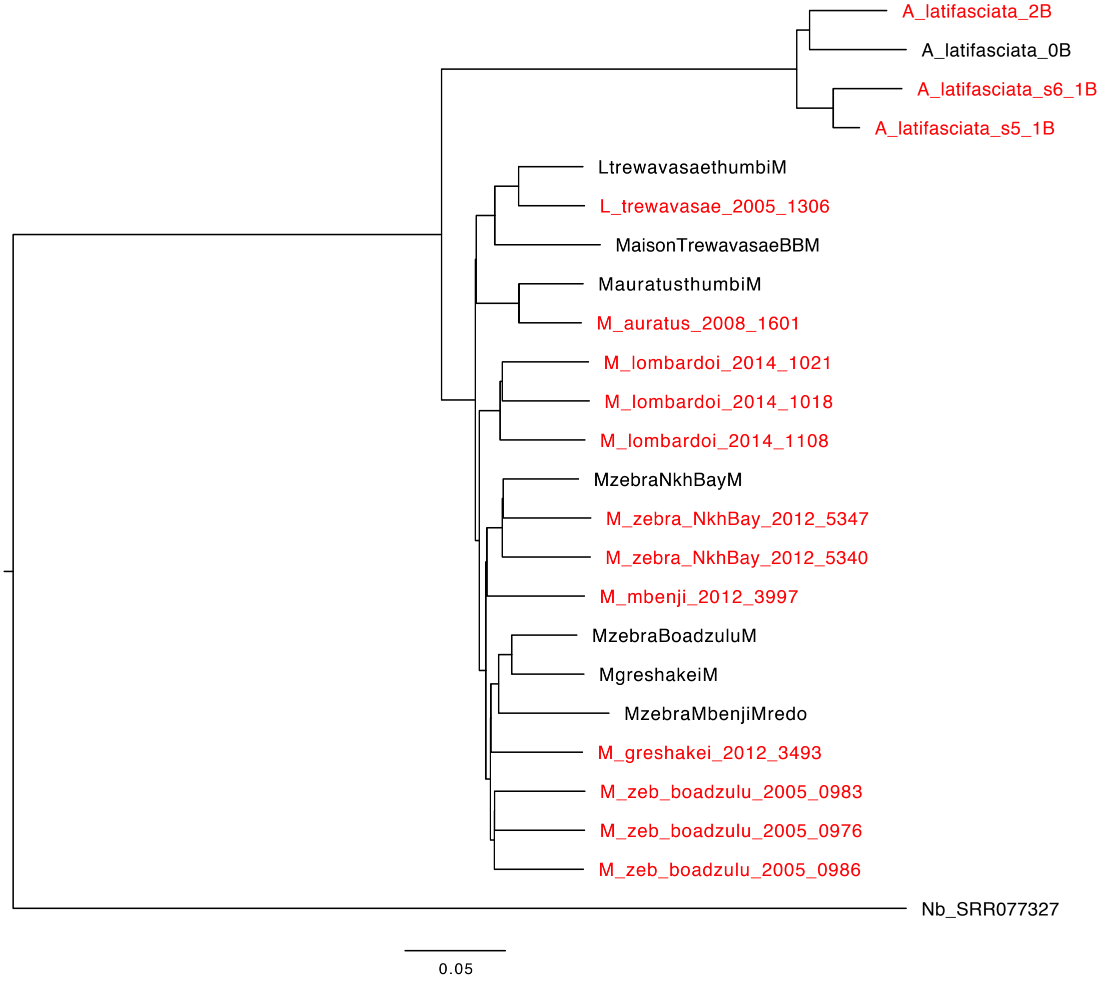


Figure 5.2 Whole genome phylogeny of Lake Malawi and Lake Victoria B and noB samples. Samples labeled in red indicate the presence of B chromosomes. Samples labeled black are samples without B chromosomes. *Neolamprologus brichardi* (Lake Tanganyika) is used as an outgroup.

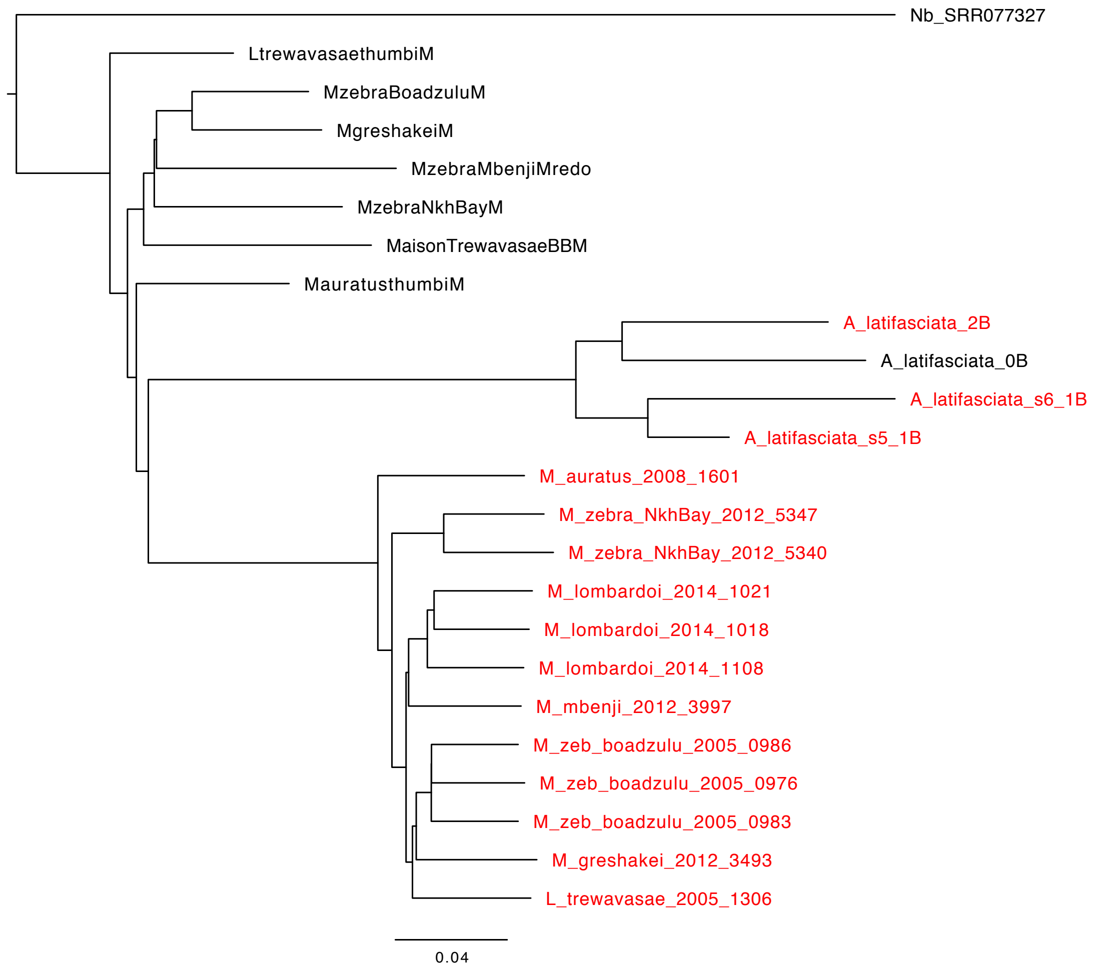


Figure 5.3 Phylogeny of 1.28Mbp “core” Lake Malawi B blocks regions. Samples labeled in red indicate the presence of B chromosomes. Samples labeled black are samples without B chromosomes. *Neolamprologus brichardi* (Lake Tanganyika) is used as an outgroup.

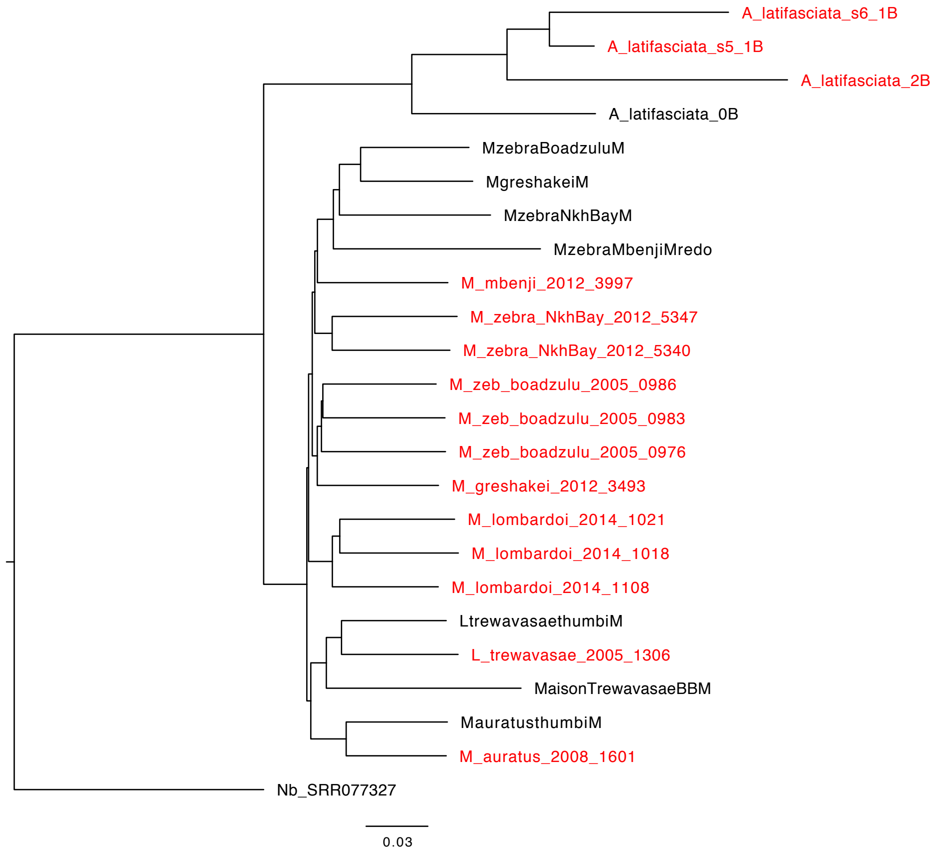


Figure 5.4 Phylogeny of Lake Victoria B chromosome blocks regions. Samples labeled in red indicate the presence of B chromosomes. Samples labeled black are samples without B chromosomes. *Neolamprologus brichardi* (Lake Tanganyika) is used as an outgroup.

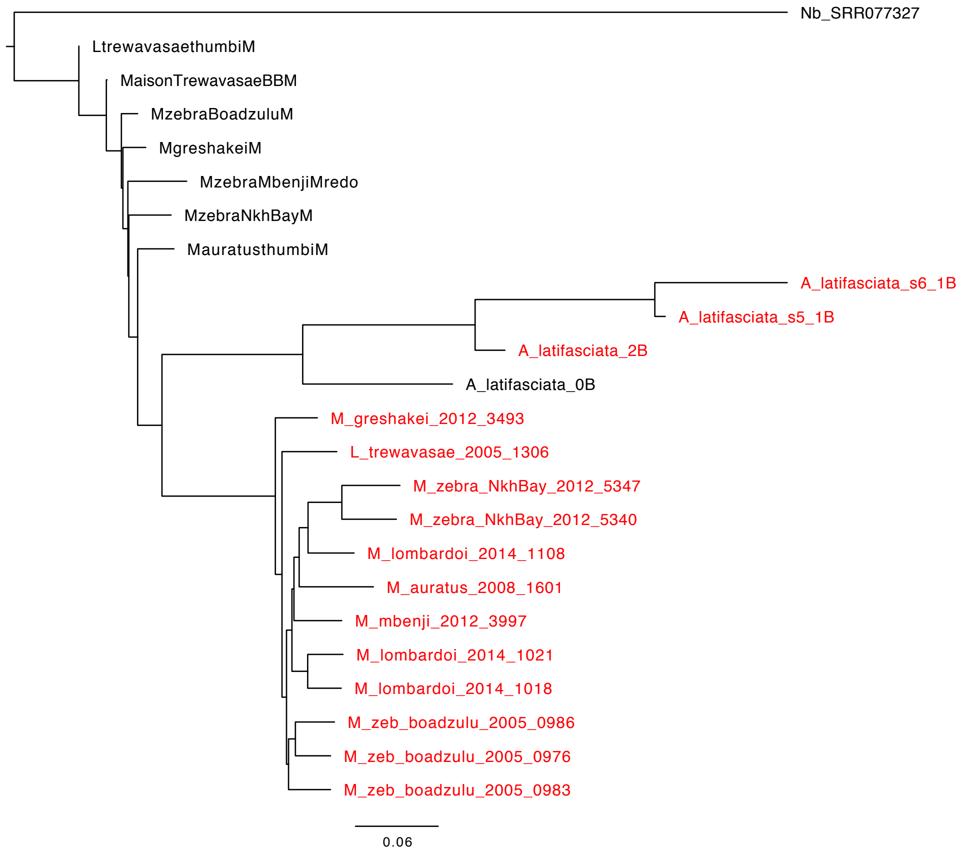


Figure 5.5 Phylogeny of shared B blocks regions between Lake Malawi and Lake Victoria. Samples labeled in red indicate the presence of B chromosomes. Samples labeled black are samples without B chromosomes. *Neolamprologus brichardi* (Lake Tanganyika) is used as an outgroup.

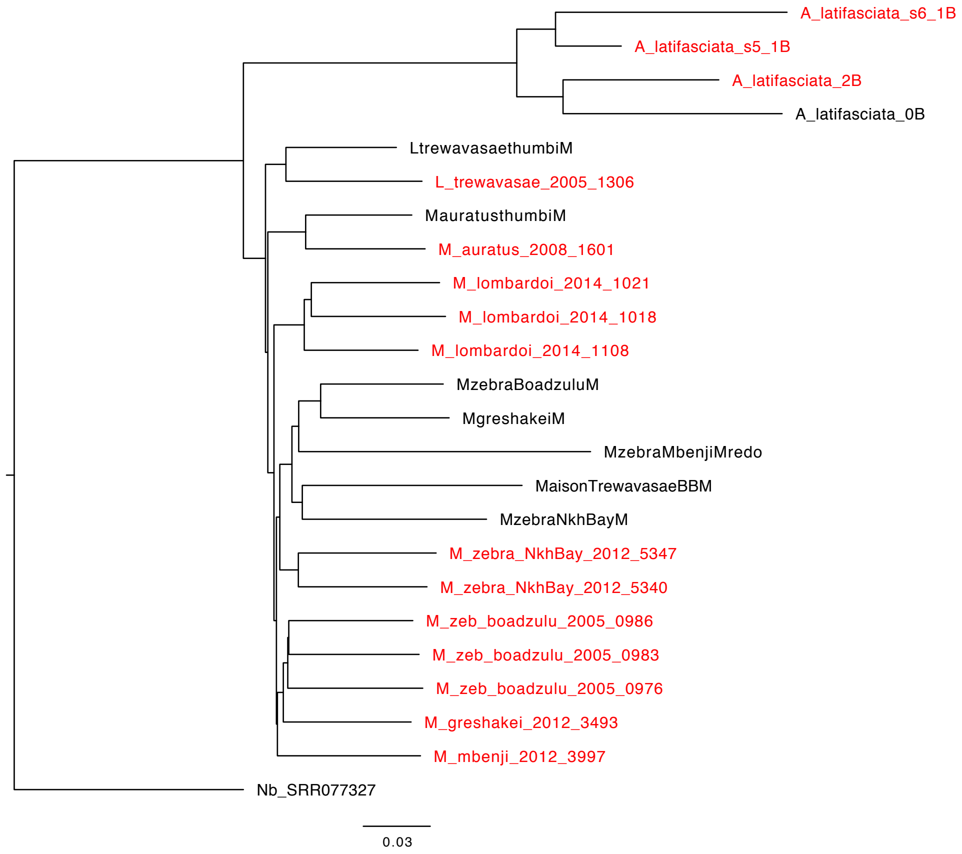


Figure 5.6 Phylogeny of regions corresponding to the variable B chromosome blocks in Lake Malawi. Samples labeled in red indicate the presence of B chromosomes. Samples labeled black are samples without B chromosomes. *Neolamprologus brichardi* (Lake Tanganyika) is used as an outgroup.

These findings show a shared ancestry of the African cichlid B chromosome in Lake Victoria and Lake Malawi. One possible scenario is that an ancestral B chromosome evolved first in Lake Victoria and later spread to Lake Malawi via hybridization. Since that introduction into Lake Malawi, the B chromosomes in each lake have diverged significantly in overall content and at present day share only 149kb of A chromosome content. The 149kb of shared B chromosome blocks on lg9 (11.1kb) and lg23 (11.7kb) show strong allele sharing as do several of the other longer blocks (contig 000028F, 8.3kb). There are several other smaller blocks that overlap in both lakes but appear to do so by chance in most cases as there is little allele sharing in these regions. Another possible scenario is that the ancestral cichlid B chromosome arose in the riverine species outside of Lake Victoria and Lake Malawi and colonized each lake separately. A strongly driving B chromosome could pass through species boundaries relatively easily, similar to the P-element in *Drosophila* (187,188). A less likely scenario is that the B chromosomes of Lake Malawi do not share a common origin and that the B block overlap and allele sharing is due by chance or produced by some process not related to B chromosomes that shows both high coverage sequence and sharing of alleles.

5.3.4 Structure of the B chromosome

An initial alignment of the B chromosome PacBio reads to the M_zebra_UMD2 reference revealed patterns of large structural differences on the B chromosome. An example of this can be seen in Figure 5.7. In this example there are portions of B chromosome specific reads that map to this short ~5.5kbp region on lg7 and other portions that map to at least 3 other places in the genome.

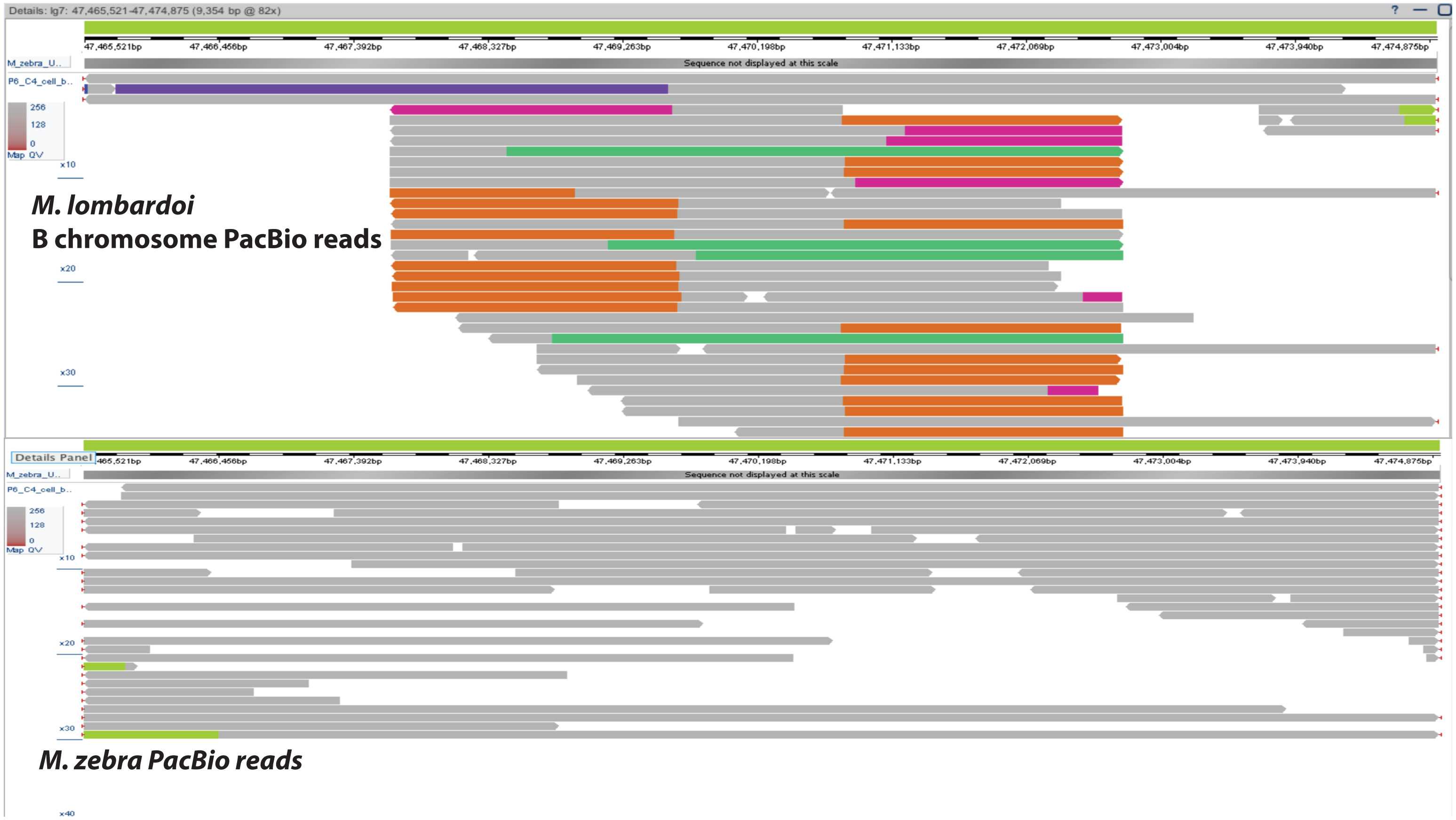


Figure 5.7. The top track shows that reads derived from the B chromosome match for part of LG7 (grey color) but have parts that match to other LGs (other colors, folded back). The bottom track shows alignment of the original *M. zebra* reads used to generate the genome assembly that have normal alignment (grey color).

Our previous work has shown that much of the B chromosome is composed of sequence that is in multiple copies and that the B chromosome blocks derived from the A chromosomes have diverged rapidly (32,40). Given this fact, we decided to assemble the B chromosome PacBio reads even though it contains a mixture of both A and B chromosomes that can cause mis-assembly errors that we described in the original *Pundamilia nyererei* reference assembly (5,32). We isolated the B chromosome specific reads by adjusting our *de novo* assembly parameters to only assemble parts that were in much higher coverage than the 20X of our sample (see Methods). The resulting B chromosome *de novo* assembly produced by miniasm consists of 650 contigs, is 22.8Mbp in total size with a contig N50 of 42.8kbp. Mapping these B chromosome contigs back to M_zebra_UMD2 agreed with our B chromosome block analysis described above. In other words, assembled B chromosome contigs aligned to the same regions as the B chromosome blocks identified by coverage.

Several of the B chromosome contigs were long enough to explore the structure of by aligning to M_zebra_UMD2. Figure 5.8 shows an alignment of the longest B chromosome contig (684kb) to M_zebra_UMD2. This contig is primarily composed of parts of lg9, lg10, lg11, lg23, lg17, lg18, lg19, lg22, and lg23 (alignments longer than 5kb). ~183kb of this contig is composed of a long B chromosome block on lg9. This is the same lg9 block that was the longest B block identified. This suggests that this is one of the longest regions on the Lake Malawi B chromosome syntenic with its A chromosome derivative.

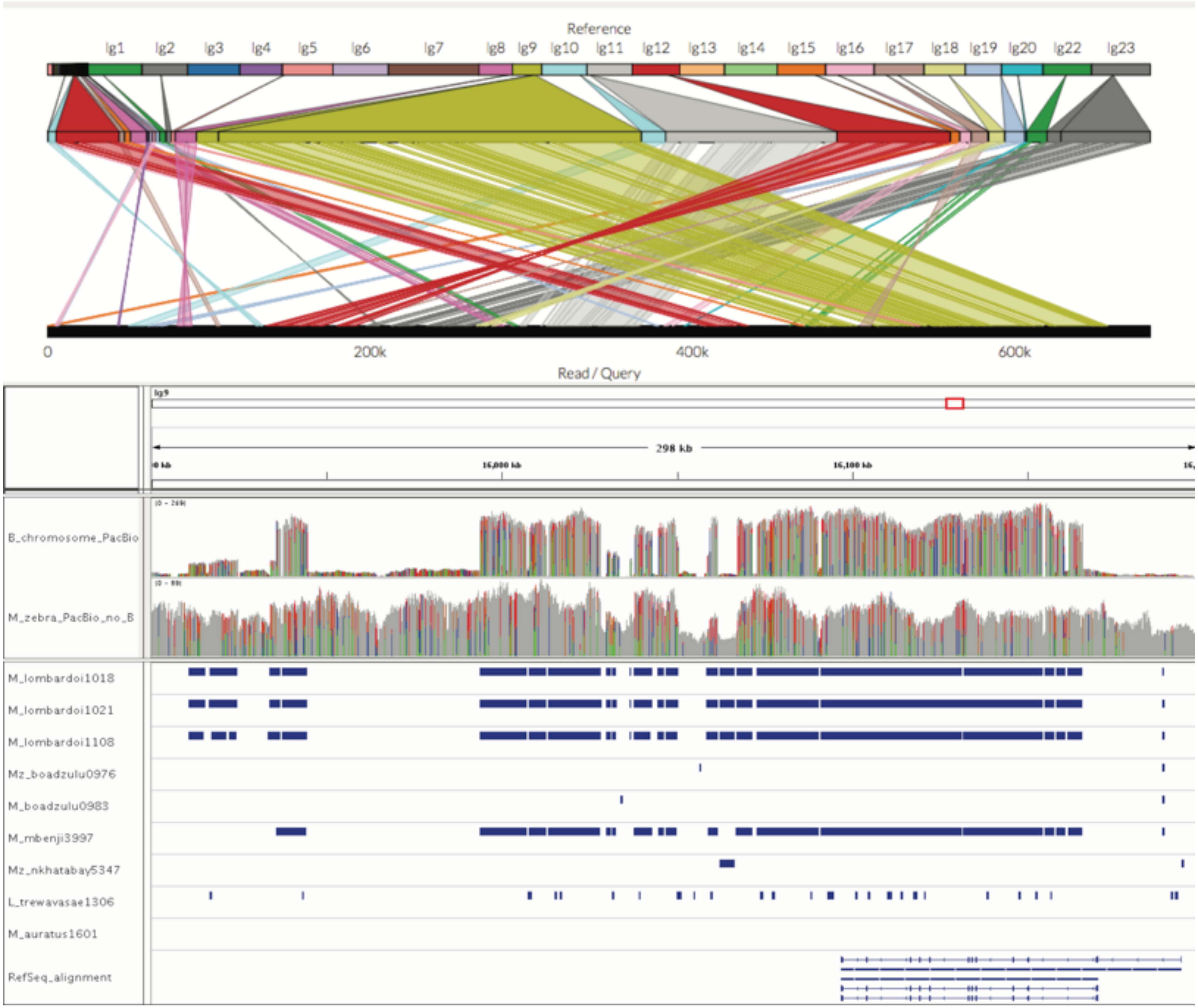

Reference
lg1
lg2
lg3
lg4
lg5
lg6
lg7
lg8
lg9
lg10
lg11
lg12
lg13
lg14
lg15
lg16
lg17
lg18
lg19
lg20
lg22
lg23
0
200k
400k
600k
Read / Query
lg9
298 kb
16,000 kb
16,100 kb
B_chromosome_PacBio
M_zebra_PacBio_no_B
M_lombardoi1018
M_lombardoi1021
M_lombardoi1108
Mz_boadzulu0976
M_boadzulu0983
M_mbenji3997
Mz_nkhatabay5347
L_trewavasae1306
M_auratus1601
RefSeq_alignment

Figure 5.8. On top is the alignment of the longest B chromosome contig to M_zebra_UMD2. On bottom are coverage plots on the corresponding lg9 region for the B chromosome PacBio sample (max coverage shown = 269x) and no-B chromosome PacBio sample (max coverage shown = 89x). Below the coverage plot are B block calls for a subset of the samples listed in Table 5.2.

5.3.5 Transposable element activity and repetitive sequences on the B chromosome

45.98% of the B chromosome *de novo* assembly was annotated as repetitive in contrast to the M_zebra_UMD2 assembly which was only 33.95% repetitive. The repeat landscape on the B chromosome is provided in Figure 5.9. There is a lack of recent TE insertions (0-2% divergence) due to the fact that this B chromosome assembly was not polished and miniasm does not attempt to correct the raw PacBio reads. Divergence levels of 3-7% are associated with sequences still containing the raw PacBio error rates. The peak around 5% divergence likely represents recent TE insertion on the B chromosome. The B chromosome has a large amount of recent TE insertion. The Tc1-Mariner element is well represented, similar to the activity seen in the M_zebra_UMD2 assembly. However, LINE/L2, LTR/ERV1 and LTR/Gypsy are present in far greater amount and copies on the B chromosome compared to the M_zebra_UMD2 genome (Figure 4.6).

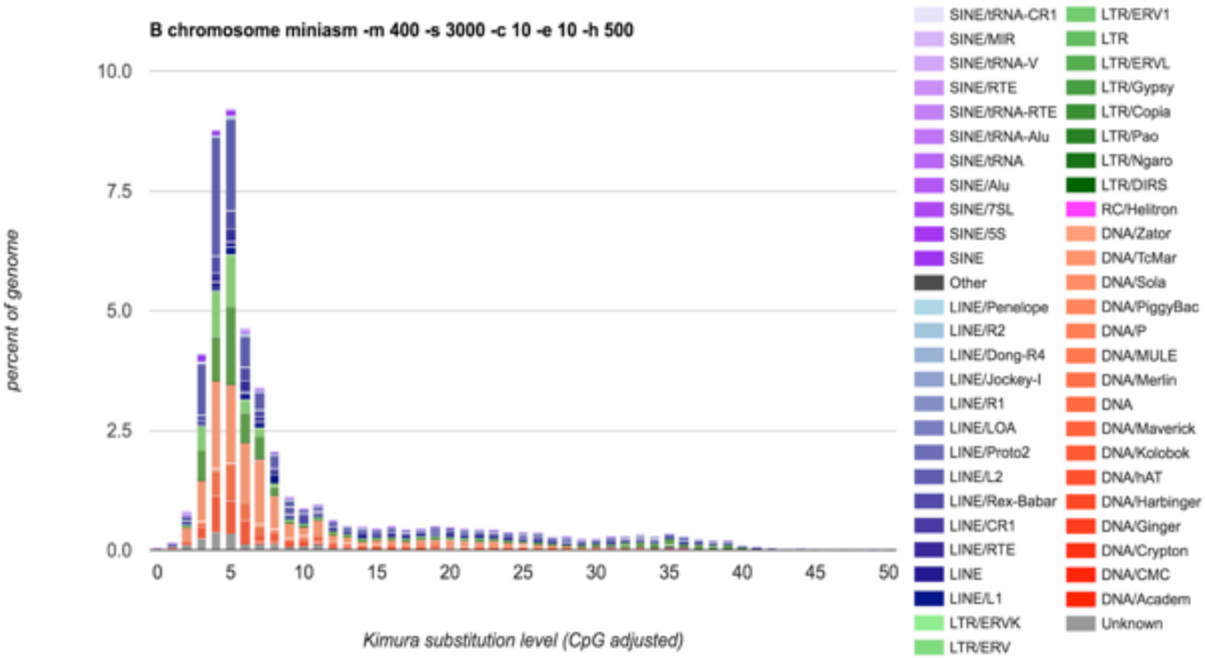


Figure 5.9. Repeat landscape of transposable elements on the B chromosome.

Six copies of the tandemly repeated telomere motif (TTAGGG) (189) were detected on the second longest contig (322kb) of the B chromosome assembly. This contig aligned to parts of many chromosomes (similar to the contig in Figure 5.8). However, the telomere repeat was annotated near a part of the contig that aligned to an unanchored contig in the M_zebra_UMD2 assembly. Therefore, it is difficult to identify where this telomere repeat may have derived from originally. There were no centromere specific repeats, ONSATA or TZSAT (99), annotated in the B chromosome assembly.

5.4 Discussion

5.4.1 Shared origin and divergent history of B chromosomes in African cichlids

The results of this study show that the B chromosomes of Lake Victoria and Lake Malawi cichlids share a common origin and have diverged dramatically within each of the lakes. Comparison of B chromosome blocks in both lakes revealed that only 149kb of blocks are common between the B chromosomes in both lakes. Variant detection and phylogenetic analysis show that the core blocks on the Lake Malawi B, and the regions shared in both lakes, are of shared origin (Figure 5.3 and Figure 5.5). Counter to this result, a phylogeny of the Lake Victorian specific blocks showed no shared origin as each sample clustered according to the species tree (Figure 5.4). Additionally, a phylogeny of the variable B chromosome blocks in Lake Malawi also followed the species tree (Figure 5.6). This evidence lends support of a scenario where a B chromosome arose in Lake Victoria and was subsequently spread to Lake Malawi, where it then diverged greatly in both lakes. Another possible scenario may be that a B chromosome arose in riverine species and spread into Lake Victoria and Lake Malawi separately. One factor that may be affecting our current analysis is the amount of ancestral polymorphism sorting that has happened in the samples that we have examined. It may be difficult to determine if the alleles specific to the four Lake Victorian samples (and the Lake Malawi B chromosomes samples) are also present at low frequency in any non-B chromosome carrying fish in Lake Malawi. Whole genome sequencing of additional species from Lake Victoria (with and without B chromosomes), riverine species surrounding Lake Victoria and surrounding Lake

Malawi may provide additional evidence as to the amount of ancestral polymorphism that has sorted on these B chromosomes. It would also likely provide additional clues as to the origin and history of African cichlid B chromosomes. Another factor that might have affected our analysis is the slight possibility of gene conversion acting on B chromosomes. The process of gene conversion from the B with the A genome is unlikely as cytogenetic work has shown that most B chromosomes are not homologous to A chromosomes and therefore do not pair with A chromosomes (190). However, it is possible that a small proto-B chromosome may have paired with an A chromosome and gene conversion may have happened in either direction (from the A to the B or from the B to the A). Tracks of gene conversion (if gene conversion has happened or is happening) may explain some of the allele sharing that is seen, particularly on the shared blocks on lg23 and lg9. Gene conversion could be acting on this particularly odd portion of lg23 that is the site of an ancient chromosome fusion.

5.4.2 B chromosome function and maintenance

In addition to sequence content, the functions of B chromosomes in Lake Victoria and Lake Malawi have also diverged. B chromosomes are present in male and female fish of at least 12 of species in Lake Victoria cichlids (32,38,98,191). In one of these species, *Lithochromis rubripinnis,* the B chromosome was shown to be female-specific and crosses showed that presence of this B chromosome led to a female-biased sex ratio of offspring (38). However, in Lake Malawi we have found B chromosomes present only in females from six species across 3 genera (Table 5.2)

(40). We have not yet found a Lake Malawi male carrying a B chromosome, but it is possible that males in Lake Malawi do have B chromosomes at very low frequencies. It is unclear if the female-biased B chromosome was a function that evolved in an ancestral B chromosome and has continued functioning as such in Lake Malawi but has lost this function in most species in Lake Victoria. It is also possible that the female-biased function evolved independently twice, once in each lake and has not yet spread in Lake Victoria. One gene of interest found on the Lake Malawi B chromosome is a potential candidate gene for the sex-bias. A portion of the know medaka sex-determination gene, *gsdf* (192) (LOC101465072) on lg7, was present on the Lake Malawi B chromosome. It is possible that this gene is being expressed on the B chromosome, although there are no identifiable B-specific variants in this block. Long-read sequencing of B chromosome transcriptomes may help in learning more about the structure of this important candidate and other genes on the B. Likewise, sequencing of female-biased Lake Victorian *L. rubripinnis* individuals with and without B chromosomes and comparing to the Lake Malawi B chromosome genomes would likely help to answer questions about the gene(s) involved in B chromosome-induced female-bias and the history of how this important function has evolved in both lakes and B chromosomes.

It is interesting that one of the longest B chromosome blocks in Lake Malawi is also shared with B chromosome blocks in Lake Victoria on lg23. The annotated gene present within these shared B blocks does not suggest an immediate role for B chromosome maintenance and/or a drive mechanism. The shared B block on lg23 is within the neuroligin-1 gene (*nlgn1*). Neuroligin-1 is cell surface protein involved in cell to cell interactions and also plays a role in synapse function and synaptic signal transmission (193,194). It should also be noted that these shared blocks are fragments of these genes. It remains unclear if this portion of the B chromosome is being expressed and has a functional role.

The location of the shared block on lg23 is precisely at the breakpoint of our previously identified ancient cichlid chromosome fusion (shown in Figure 4.5 for *O. niloticus* assembly). Since we know the history of this region, it may say something about the early history of the proto-B chromosome. A recent review argues that B chromosomes are likely to originate from genomic locations where there have been evolutionary breakpoints or regions that have a higher frequency of non-homologous recombination (195). It is possible that this region of lg23 may be prone to breaking and a small part of this region on lg23 could have potentially formed the proto-B chromosome that is still present and conserved in both Lake Victoria and Lake Malawi. This proto-B chromosome would have accumulated additional sequence from the A genome and perhaps other gene(s)/sequence(s) that altered the function(s) of the B chromosomes in each lake.

There are several genes of interest identified on the core Lake Malawi B chromosome that may be important for the drive mechanism and maintenance of this B chromosome. Regulator of telomere elongation helicase 1, RTEL1 (LOC101471057) appears to be present in multiple copies in both no-B and B-carrying individuals (Appendix J). However, there are additional copies in the core B blocks that appear to have decayed, suggesting a pseudogenization event specific to the B chromosome. The divergence of this potential RTEL pseudogene is so severe that there are regions where Illumina reads do not align and which only PacBio reads can span. It is unclear if the duplicate copy(ies) of RTEL1 on the B chromosome is(are) still functioning. RTEL1 functions as a helicase with important telomere functions (196) that could play a major role on the B chromosome. A study in humans of individuals with an autosomal recessive mutation in RTEL1 showed patients had evidence of telomere dysfunction, shortened length, and extra-chromosomal circular telomeric DNA (197). It is possible that the additional RTEL1 copy or pseudogene may be playing an important role during bouquet formation and attachment to the nuclear membrane during meiosis (198). The "non-structural maintenance of chromosome element 4" (nsmce4a or NSE4A) is another gene of interest on the Lake Malawi B. The NSE4A gene spans 5.7kb and a 15kb core B chromosome block entirely contains this gene, 4kb of its promotor, and 5.3kb of downstream sequence in all Lake Malawi B chromosomes. NSE4A plays an important role in meiosis by functioning in homologous recombination repair of double strand breaks and recovery of stalled replication forks (193,194). Impairments in the NSE family of proteins have been shown to lead to chromosome breakage disorders (199). Finally, a 2.3kb- and 4kb- core Lake Malawi B chromosome block on lg7 the encompasses the promotor and exons 1 and exons 3-11 (of 19 total) within the 12.2kb "Inner centromere protein antigens 135/155kDa" (INCENP) gene. One of the two functional domains of INCENP is retained in these core blocks. INCENP is important in regulation of mitosis and functions at the centromere for alignment and chromosome segregation (193,194). No centromere repeat was identified on the B chromosome. It is possible that the B chromosome centromere was simply not assembled. It is also possible, though unlikely, that the B chromosome centromere has diverged sufficiently that it was not detectable.

5.4.3 B chromosome structure

To our knowledge, this study is the first to produce and analyze a *de novo* B chromosome assembly via long read sequencing. The results of this *de novo* B chromosome assembly and alignment to the M_zebra_UMD2 genome reference revealed a very striking picture structure of the B chromosome organization. Figure 5.8 provides an example of how parts of the A genome are spread throughout the contemporary B chromosome. In this example, the block is variable (not core) and present in some *Metriaclima* samples, but not others. The *Labeotropheus trewavasae* sample appears as it is being lost from the B chromosome. Seemingly every large B chromosome contig contains parts both large and small from various LGs along neighboring regions of the B chromosome similar to the depiction in Figure 5.8.

The *de novo* B chromosome assembly that has been produced here is likely an under representation of the total B chromosome, since the total size was 22.8Mbp and the average *M. zebra* LG is 45.5Mbp. Karyotypes of *M. lombardoi* have shown this B chromosomes to be one of the three largest chromosomes (40) which would put it on the higher end of this average 45.5Mbp size. Likely missing from this B chromosome *de novo* assembly are B chromosome regions that are present in low copy, as the coverage settings would preclude most of these regions from being assembled.

This view of the structure of the Lake Malawi B chromosome provides new insight into how the B chromosome may be gaining and losing parts of the A genome, but additional data is needed to determine exactly how this has happened. Flow sorting and longer read sequencing of B chromosomes may prove to be a useful technique to further refine the structure of these complex B chromosomes.

5.4.4 Transposable element dynamics differ greatly between A and B genomes

Transposable elements may also be playing a large role in shaping the structure of the B chromosome. TEs are abundant in many B chromosomes (32,36,190,200–203) and can play large roles in the structure of genomes (187,204–206), especially when selection against TE activity is relaxed, as is likely the case on B chromosomes. The overall amount of transposable element sequence annotated on our *de novo* B chromosome assembly was 12% higher than the M_zebra_UMD2 assembly. This likely represents an underestimate of the total amount of TEs on the B chromosome both due to lack of polishing as previously mentioned and since some TEs are likely not fully assembled due to the highly repetitive nature of the B chromosome. Nevertheless, the pattern of TE activity on the B chromosome is much different than the A genome. LINE/L2, LTR/ERV1 and LTR/Gypsy elements are present in much higher numbers on the B chromosome than on the A genome. These particular elements are probably active in Lake Malawi cichlids, but are able insert into the B chromosome more often due to relaxed selective pressures of the B.

5.5 Methods

5.5.1 DNA extraction and PacBio and Illumina Sequencing

Female *M. lombardoi* individuals were genotyped using previously published markers to identify fish sith a B chromosome (40). A single female was then sacrificed to obtain a blood sample, using animal procedures that were conducted in accordance with University of Maryland IACUC Protocol #R-10-74. The Qiagen MagAttract HMW DNA kit was used to extract high-molecular weight DNA from nucleated blood cells from the sample. Size selection was performed at the University of Maryland Genomics Resource Center using a Blue Pippin pulse-field gel electrophoresis instrument. There, a library was also constructed and 9 SMRT cells were sequenced on their PacBio RS II using the P6-C4 chemistry. An additional 9 SMRT cells were sequenced on their PacBio Sequel instrument.

DNA from the female individuals listed in Table 5.2 were extracted from fin clips by phenol-chloroform extraction. DNA concentrations were measured via fluorescence spectroscopy and individual libraries were generated using the Illumina TruSeq DNA PCR-Free LT kit. Each library was then run on one lane of 101bp paired-end sequencing on an Illumina HiSeq 2000 at the University of Maryland IBBR genomics facility.

5.5.2 Alignment of B chromosome reads and B block detection in the PacBio sample

PacBio B chromosome reads were aligned to the M_zebra_UMD2 assembly using NGMLR *version 0.2.6* (207) with the '*-x pacbio*' setting. Copycat (208) was used to bin read coverage in 100bp bins. Bins with coverage higher than 50x (2.5 times the mean coverage of the read set) were kept. Bedtools (209) *version 2.26.0* was run with the following command ‘*bedtools merge -d 10000 -c 1 -o count -i*’ to merge the high coverage bins.

5.5.3 B block detection with Illumina data

The Illumina reads for each sample listed in Table 5.2 were aligned with BWA *mem* (122) (*version 0.7.12-r1044*) to the M_zebra_UMD2 assembly. The samtools (114) utility, mpileup, was run to quantify Illumina read coverage across the genome at the base pair level. The ‘-A’ flag to ‘do not discard anomalous read pairs’ was turned on since many B chromosome reads map as anomalous read pairs. In order to compare coverage across samples, coverage at each bp was normalized by the average coverage across the genome for that sample, resulting in the “scaled coverage” value. Reads from B sequences that are still highly homologous to their A genome counter parts were expected to align to the A genome, resulting in higher coverage proportional to the number of copies of that sequence on both the A and B chromosomes. To identify these regions of higher coverage, the scaled coverage of B samples was compared to the scaled coverage of NoB samples, resulting in a ratio of scaled coverages, or the scaled coverage ratio (SCR) value. Base pairs with a SCR of 3 or higher (meaning the scaled coverage was 4+ times higher in the B sample than the NoB sample) were identified as possible B sequence and provided to bedtools merge in order to identify consecutive regions of high SCR. The bedtools *version 2.26.0* (209) ‘merge’ was run to connect any bp meeting the SCR threshold within

300 bp of each other, resulting in regions of sequence identified as B blocks as opposed to individual bp. These blocks were further refined by removing any block with less than 10% of its bp meeting the SCR threshold value or any block less than 500 bp in length. This refinement is intended to eliminate regions misidentified as B block sequence due to variable Illumina coverage. In addition to the identification of B blocks in each B individual sample, blocks shared across B samples were also identified. The B blocks BED files were consecutively fed to the bedtools 'intersect' command to identify regions shared in at least 12 of the 13 B samples. These shared regions are referred to as the Lake Malawi "core" blocks. Likewise, the bedtools 'intersect' command was used to identify regions shared between Lake Malawi and Lake Victoria blocks.

5.5.4 Variant calling and phylogenetic analysis of Illumina samples

The Picard *version 2.1.0* (115) 'SortSam', 'MarkDuplicates', and 'BuildBamIndex' programs were run on each subsequent SAM file produced by BWA. The Lake Tanganyika *Neolamprologus brichardi* NCBI SRA run 'SRR077327' was aligned and processed the same way as the samples listed above and is used as an outgroup for the phylogenetic analysis. Variants across the entire genome were called using FreeBayes (120) (version *v1.0.2-33-gdbb6160-dirty*). VCFtools *version 0.1.13* (121) was used to merge each of these VCF files into a single VCF file. For the whole genome phylogeny, a neighbor-joining tree was generated on this VCF file using the VCF-kit 'phylo tree nj' *version 0.1.6* (210). To compute the phylogeny of only the Lake Malawi B chromosome blocks, the Bedtools ‘intersect’ command was used to extract variants of the B block regions from the whole genome VCF file. Again, VCF-kit ‘phylo tree nj’ was run on this subset of variants to generate the phylogeny of Lake Malawi B block regions.

5.5.5 *De novo* assembly of a B chromosome

The B chromosome PacBio reads were assembled with minimap and miniasm (211). First, minimap *version 0.2-r123* was run with the following parameters on the FASTA file of PacBio reads: ‘*minimap -w5 -L1000 -m70 -f 0.00001 -c 6 -g 500 -t40 -I6G*’. These approximate read mappings were then assembled with miniasm *version 0.2-r128* with the following parameters: ‘*miniasm -m 400 -s 3000 -c 10 -e 10 -h 500*’. The resulting gfa output assembly graph was converted to FASTA with a simple awk command (*awk '/^S/{print ">"$2"\n"$3}'*). The assembled B chromosome contigs were confirmed by comparison to the called B chromosome blocks by alignment of the contigs to the M_zebra_UMD2 assembly using ‘*nucmer*’ program of the MUMmer package (184) (*version 4.0.0beta2*). This allowed for refinement of the final assembly parameters used about as confirmation that the B blocks were indeed the only portion being assembled. This also allowed for visual inspection of the structure of the B chromosome.

5.5.6 Repeat annotation of the de novo B chromosome assembly

To identify and classify repeats on the B chromosome contigs produced by miniasm (above), we again used RepeatModeler (64) (*version open-1.0.8*). These *de novo* repeats specific to the B chromosome were combined with the RepBase-derived

RepeatMasker libraries (129). RepeatMasker (65) (*version open-4.0.5*) was run on the contigs using NCBI BLAST+ (*version 2.3.0+*) as the engine (‘*-e ncbi*’) and specifying the combined repeat library (‘*-lib*’). The more sensitive slow search mode (‘*-s*’) was used. The repeat landscape was generated with the RepeatMasker ‘*calcDivergenceFromAlign.pl*’ and ‘*createRepeatLandscape.pl*’ utility scripts.

Chapter 6: Conclusions and future directions

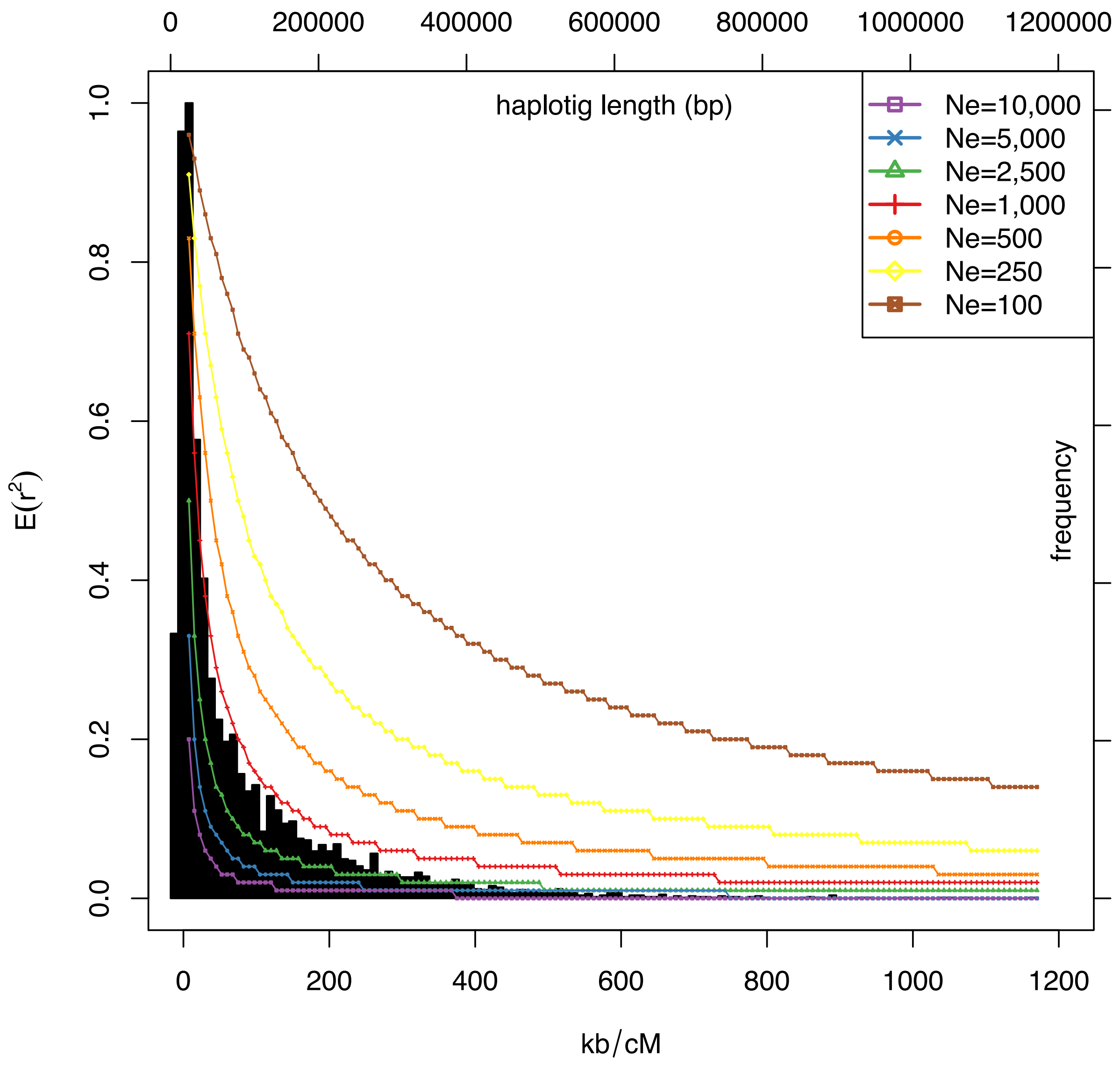
This work provides chromosome-scale, high-quality genomic resources for the cichlid and genomics communities while integrating a variety of datasets to achieve these end products. Through this process, we were able to understand many unique characteristics of the genomes of the diverse African cichlids. Aspects of African cichlid genomes such as the large structural changes, sex chromosome evolution, patterns of recombination across chromosomes, ancient chromosome fusion events, variation in transposable element activity, and the role of B chromosomes all provide a rich context to the genome assemblies that have been produced.

These genome assemblies are a significant step forward in terms of completeness and accuracy, but remain incomplete as do most other vertebrate genomes, including the human genome. While this work demonstrates the ability to assemble large recently duplicated genes and regions, we know that others remain poorly assembled. For example, the duplication of the 417kb "CUB and Sushi multiple domains" (csmd1) gene within the *Oreochromis niloticus* sex-determination region on LG1 remains collapsed. We have been able to partially assemble a subset of centromeres in cichlids for the first time and this has allowed us to place them in the context of evolving karyotypes. However, some centromeres remain unplaced. Full length assembly of all centromeres will allow for many new questions related to their functional and sequence evolution to be studied. The representation of recent transposable element insertions, that are likely rewiring key divergent gene regulator networks, are now well represented for the first time. However, these recent TEs have only been assembled in two genomes and many additional high-quality assemblies will be required for assessing the role of recent TE activity across African cichlid species. The dizzying array of duplications, deletions and rearrangements on the rapidly diverging B chromosomes were assembled *de novo* for the first time. A comparison across Lake Malawi and Lake Victoria points to a shared and interesting history of B chromosome evolution in African cichlids. Additional genomes of B chromosome carrying species will help fully describe the evolution that has happened within them, their functional effects, and the role of B chromosomes in the evolution of cichlid genomes.

Future advances in genome sequencing technologies will result in longer, more accurate and less expensive reads. It is my expectation that the resources produced in this dissertation will become stepping stones for newer, more complete genomes of many African cichlid species in the near future. Comparison of telomere-to-telomere complete genome assemblies of a wide variety of African cichlid species is something that is feasable and exciting. This work has provided a blueprint for the present and future, while also generating many new and interesting questions that should be asked as these new resources are generated, analyzed and interpreted.

Appendix A

Size distribution of the M_zebra_UMD2 assembled haplotigs and theoretical recombination rate for several different effective population sizes



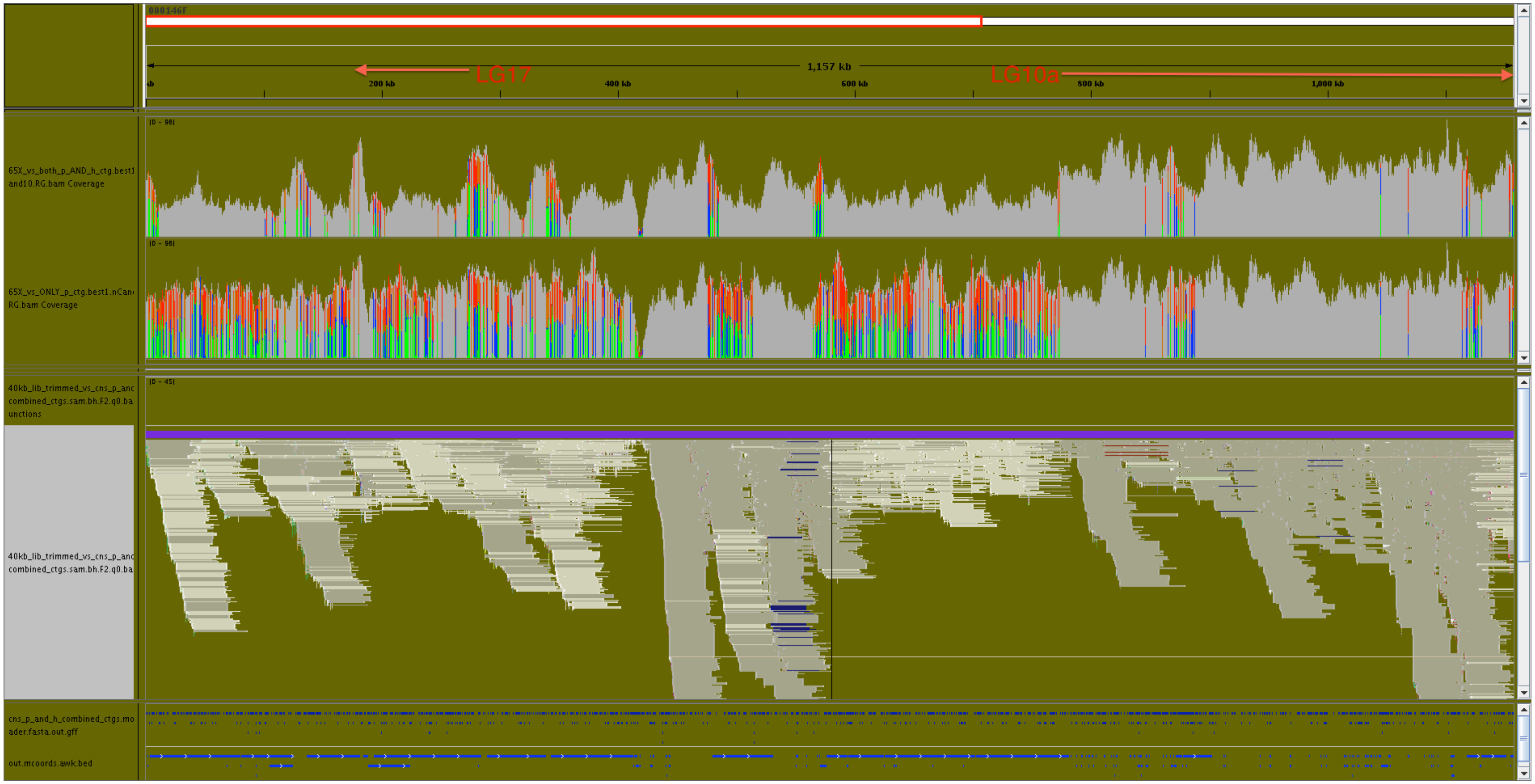
Appendix B

M_zebra_UMD2 FALCON p-contigs where markers from two or more different LGs maps aligned, indicating a potential inter-LG misassembly.

key:			
detected in 1 of the 4 maps			
detected in 2 of the 4 maps			
detected in 3 of the 4 maps			
detected in all 4 maps			
likely not a misassembly			
		inspect this location	Break?
M. zebra x M. mbenjii (160 F2):			
000006F_pilon	lg20, lg5		no
000075F_pilon	lg1, lg23	000075F_pilon|quiver:2233963-2424134	yes
000432F_pilon	lg13, lg2	000432F_pilon|quiver:568720-710844	yes
L. fuelleborni x Tropheops ‘red cheek’ (262 F2):			
000034F_pilon	lg18, lg6		yes
000036F_pilon	lg13a, lg13b		no
000075F_pilon	lg1, lg21	000075F_pilon|quiver:2233963-2424134	yes
000105F_pilon	lg10a, lg2	000105F_pilon|quiver	yes
000117F_pilon	lg10b, lg20		yes
000146F_pilon	lg10a, lg17	000146F_pilon|quiver:216380-308914	yes
000197F_pilon	lg3, lg4	000197F_pilon|quiver	no
000201F_pilon	lg1, lg3	000201F_pilon|quiver:738288-1391900	no

000455F_pilon	lg14, lg8	000455F_pilon|quiver:108290-402879	yes
M. mbenjii x A. baenschi (161 F2):			
000034F_pilon	lg17, lg8		yes
000075F_pilon	lg1, lg9	000075F_pilon|quiver:2233963-2424134	yes
000117F_pilon	lg13, lg7		yes
000258F_pilon	lg20, lg22	000258F_pilon|quiver:150109-582071	yes
M. mbenjii x A. koningsi (331 F2):			
000002F_pilon	lg17, lg22	000002F_pilon|quiver:94485-465360	yes
000007F_pilon	lg11, lg7	000007F_pilon|quiver:1-477055	yes
000034F_pilon	lg18, lg6		yes
000045F_pilon	lg17, lg6	000045F_pilon|quiver:128119-312105	yes
000075F_pilon	lg1, lg23	000075F_pilon|quiver:2233963-2424134	yes
000117F_pilon	lg10, lg20		yes
000146F_pilon	lg10, lg17	000146F_pilon|quiver:216380-308914	yes
000149F_pilon	lg12, lg16	000149F_pilon|quiver:1063847-1399960	no
000216F_pilon	lg17, lg3	000216F_pilon|quiver:633635-885577	no
000223F_pilon	lg3, lg8	000223F_pilon|quiver:245368-788334	no
000245F_pilon	lg11, lg19	000245F_pilon|quiver:32581-349492	yes
000261F_pilon	lg1, lg23	000261F_pilon|quiver:489307-963566	yes
000369F_pilon	lg23, lg3	000369F_pilon|quiver:557152-578801	no
000404F_pilon	lg3, lg7	000404F_pilon|quiver:612567-697734	yes
000415F_pilon	lg1, lg18	000415F_pilon|quiver	yes
000521F_pilon	lg15, lg17	000521F_pilon|quiver	no
000580F_pilon	lg4, lg9	000580F_pilon|quiver	no

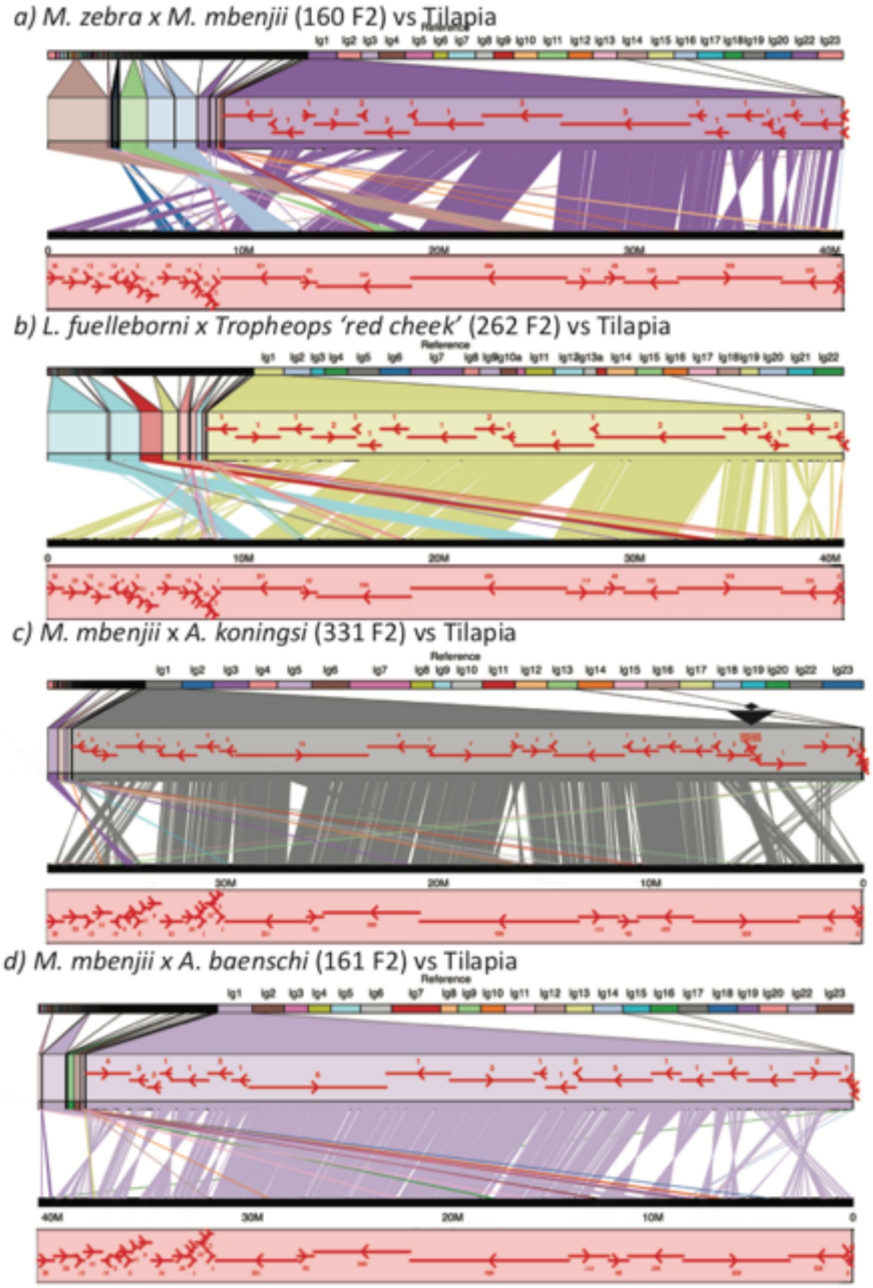
Appendix C


000146F
1,157 kb
LG17
LG10a
kb
200 kb
400 kb
600 kb
800 kb
1,000 kb
[0 - 98]
65X_vs_both_p_AND_h_ctg.best1
and10.RG.bam Coverage
[0 - 98]
65X_vs_ONLY_p_ctg.best1.nCan
RG.bam Coverage
[0 - 45]
40kb_lib_trimmed_vs_cns_p_and
combined_ctgs.sam.bh.F2.q0.ba
unctions
40kb_lib_trimmed_vs_cns_p_and
combined_ctgs.sam.bh.F2.q0.ba
cns_p_and_h_combined_ctgs.mo
ader.fasta.out.gff
out.mcoords.awk.bed

Screenshot of IGV view to inspect potential misassemblies. In this example, a misassembly on this contig was confirmed at position 420,665. The top two tracks are the read coverage plots for the PacBio read alignments and show a sharp decrease in coverage at the misassembly location. The track below that show 40kb mate-pair alignments and also show no coverage at the location of the misassembly.

<u>*Appendix D*</u>

M. zebra assembly contigs anchored with each of the 4 maps and aligned to O_niloticus_UMD_NMBU. Centromeres indicated with black triangles. Contigs are represented as red lines above each respective assembly. LG1:



a) M. zebra x M. mbenjii (160 F2) vs Tilapia

Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
0
10M
20M
30M
40M

b) L. fuelleborni x Tropheops 'red cheek' (262 F2) vs Tilapia

Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10a lg11 lg12 lg13a lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg21 lg22
0
10M
20M
30M
40M

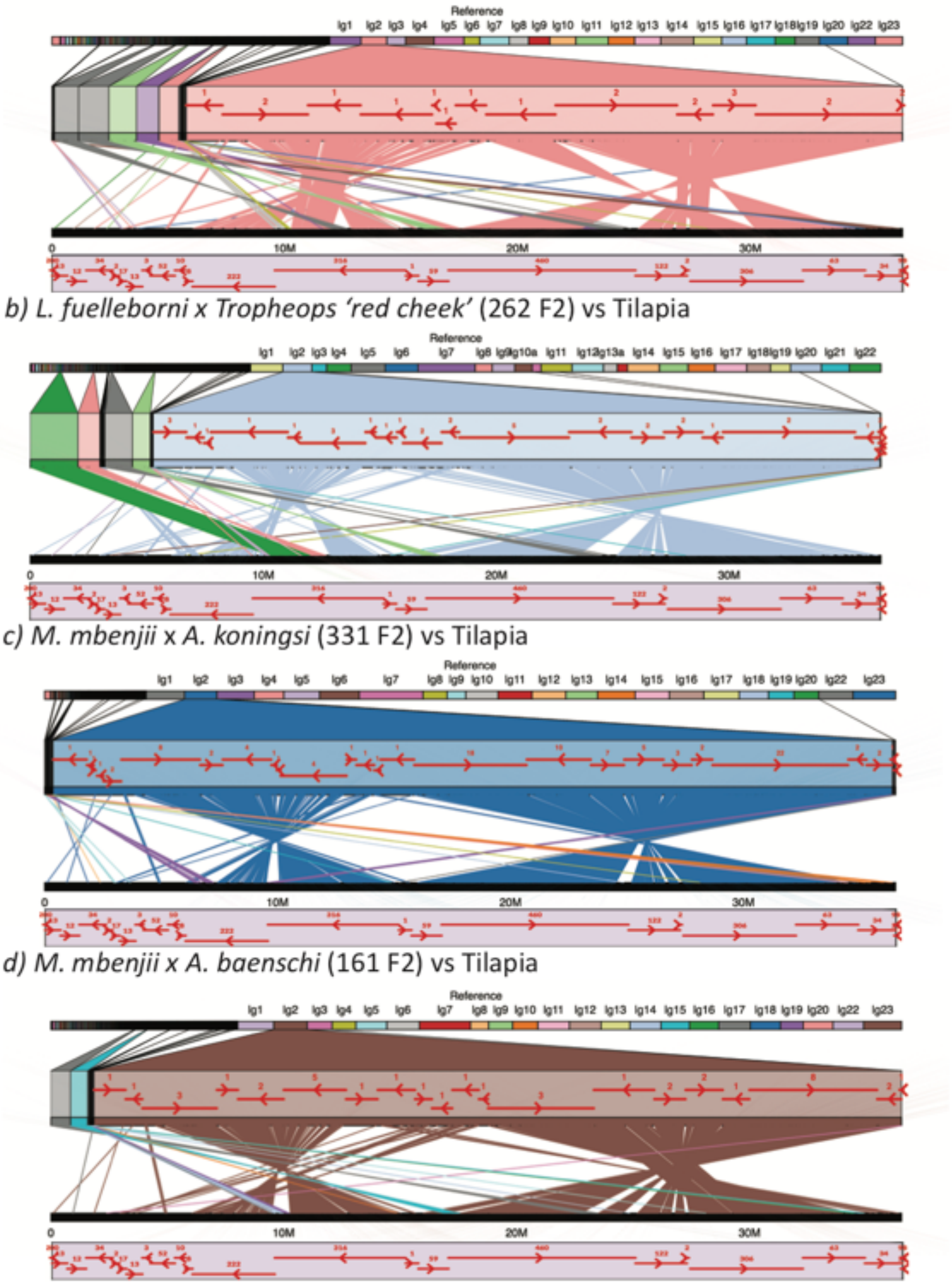
c) M. mbenjii x A. koningsi (331 F2) vs Tilapia

Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
30M
20M
10M
0

d) M. mbenjii x A. baenschi (161 F2) vs Tilapia

Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
40M
30M
20M
10M
0

LG2

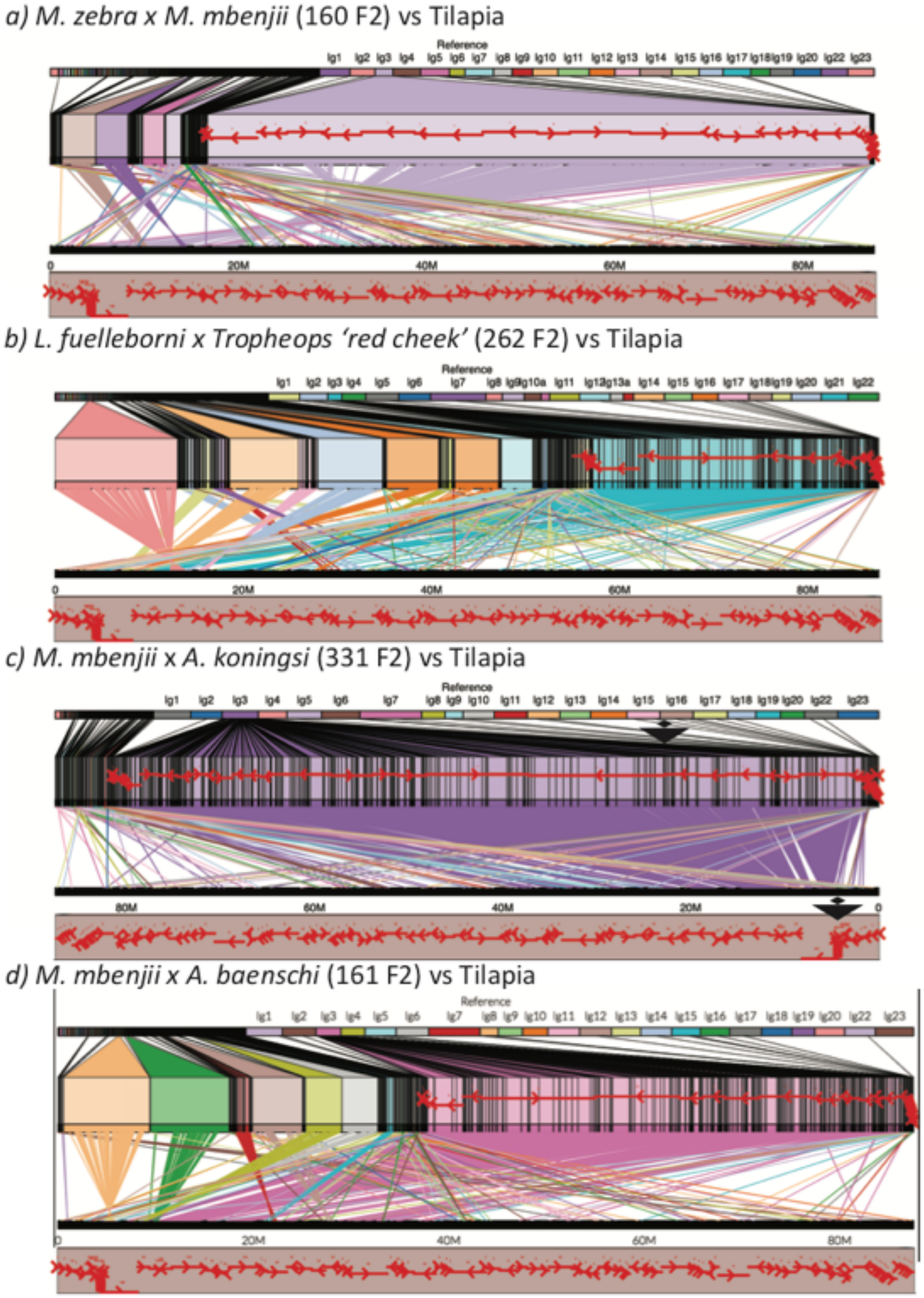

a) *M. zebra* x *M. mbenjii* (160 F2) vs Tilapia

b) *L. fuelleborni* x *Tropheops 'red cheek'* (262 F2) vs Tilapia

c) *M. mbenjii* x *A. koningsi* (331 F2) vs Tilapia

d) *M. mbenjii* x *A. baenschi* (161 F2) vs Tilapia

LG3



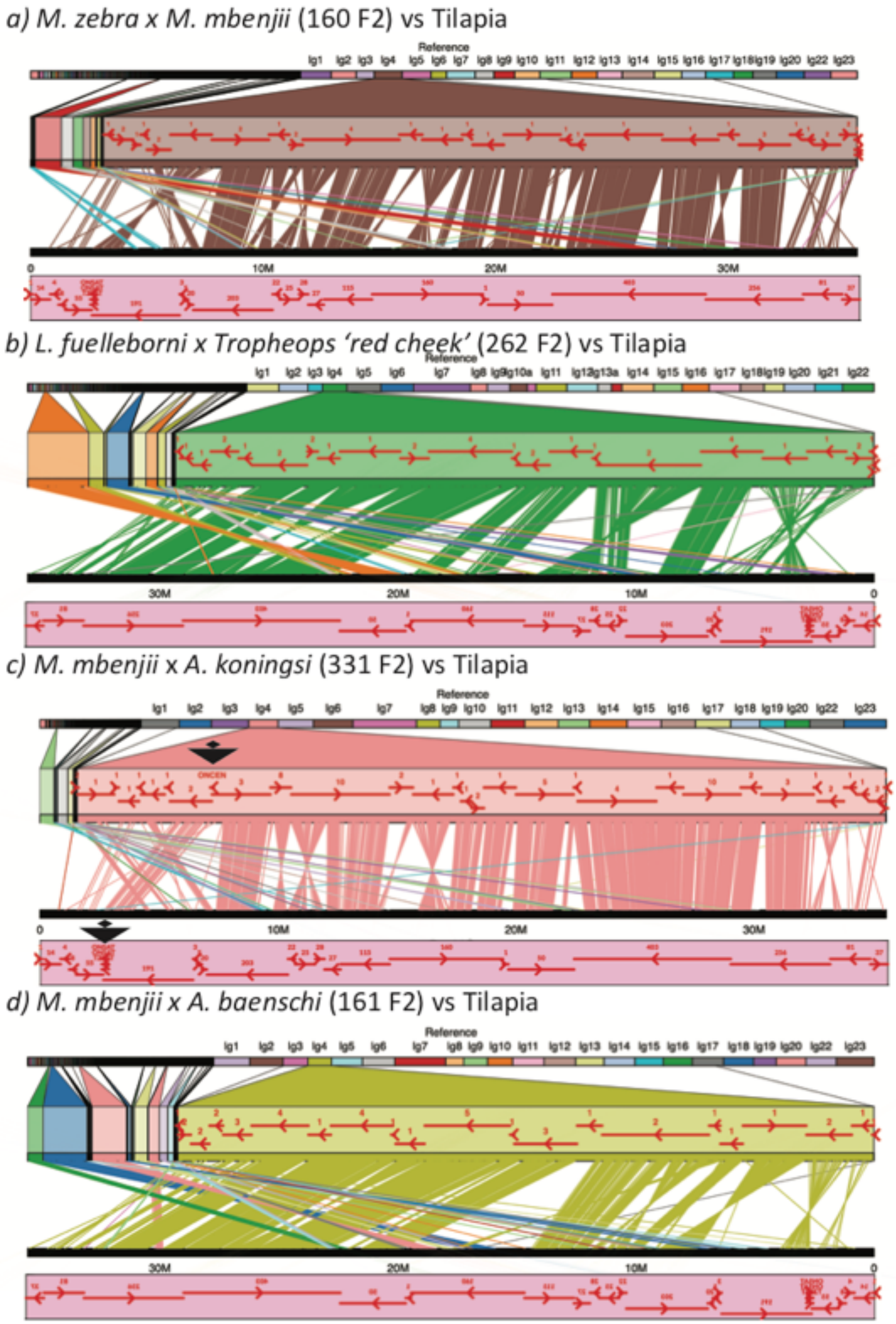
a) *M. zebra x M. mbenjii* (160 F2) vs Tilapia

b) *L. fuelleborni x Tropheops 'red cheek'* (262 F2) vs Tilapia

c) *M. mbenjii* x *A. koningsi* (331 F2) vs Tilapia

d) *M. mbenjii* x *A. baenschi* (161 F2) vs Tilapia

LG4



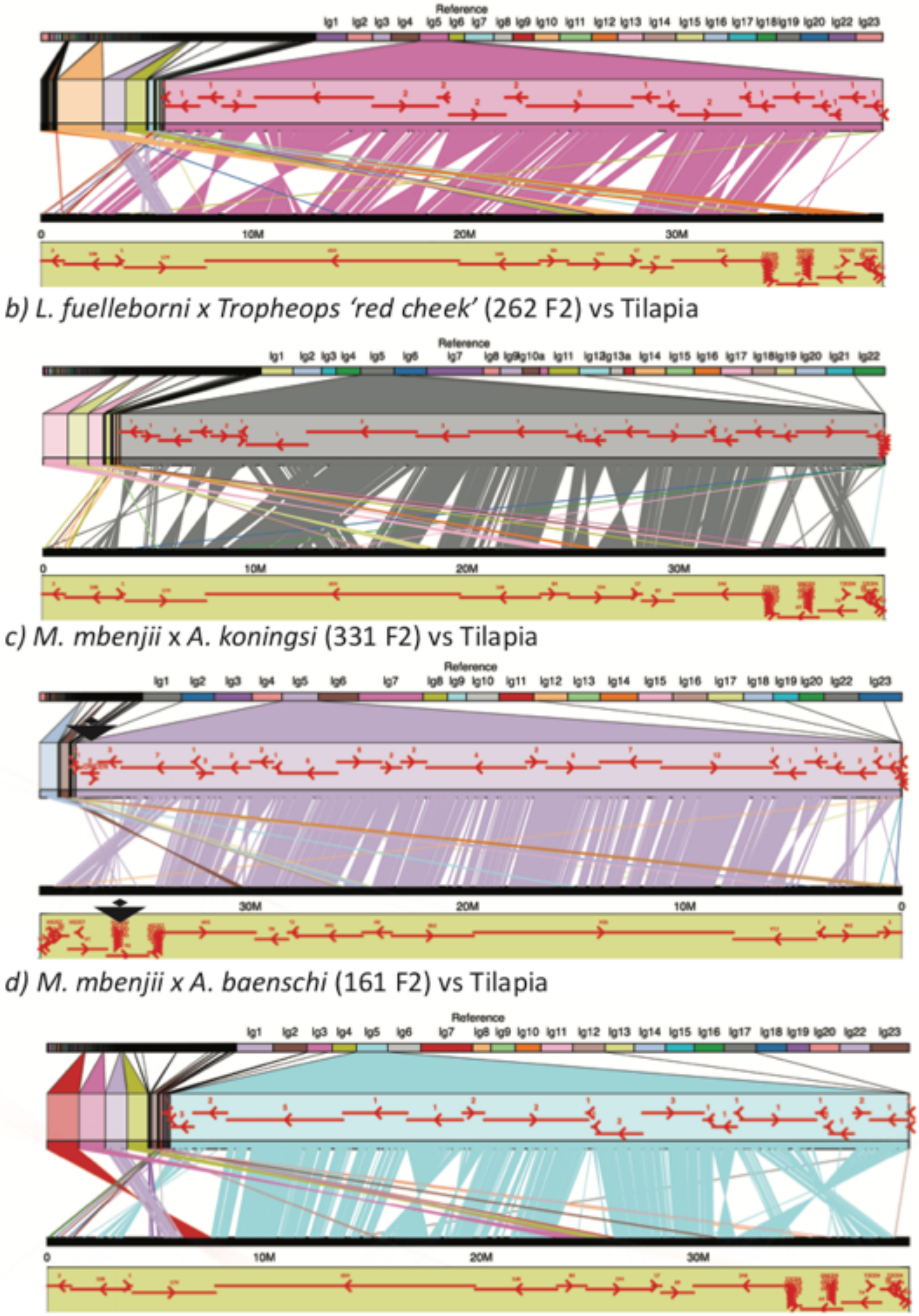
a) *M. zebra x M. mbenjii* (160 F2) vs Tilapia

b) *L. fuelleborni x Tropheops 'red cheek'* (262 F2) vs Tilapia

c) *M. mbenjii x A. koningsi* (331 F2) vs Tilapia

d) *M. mbenjii x A. baenschi* (161 F2) vs Tilapia

LG5



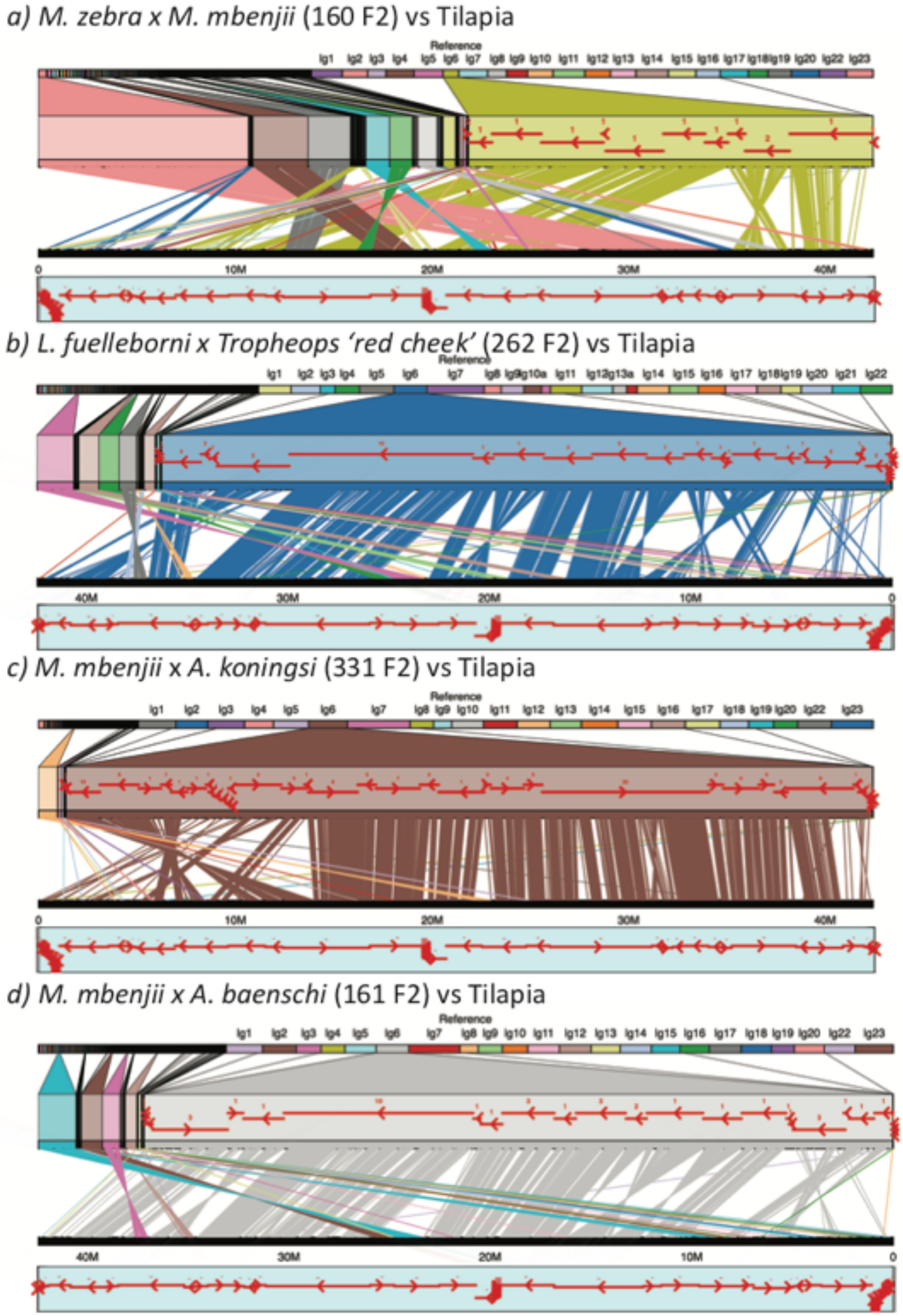
a) M. zebra x M. mbenjii (160 F2) vs Tilapia

b) L. fuelleborni x Tropheops 'red cheek' (262 F2) vs Tilapia

c) M. mbenjii x *A. koningsi* (331 F2) vs Tilapia

d) M. mbenjii x A. baenschi (161 F2) vs Tilapia

LG6



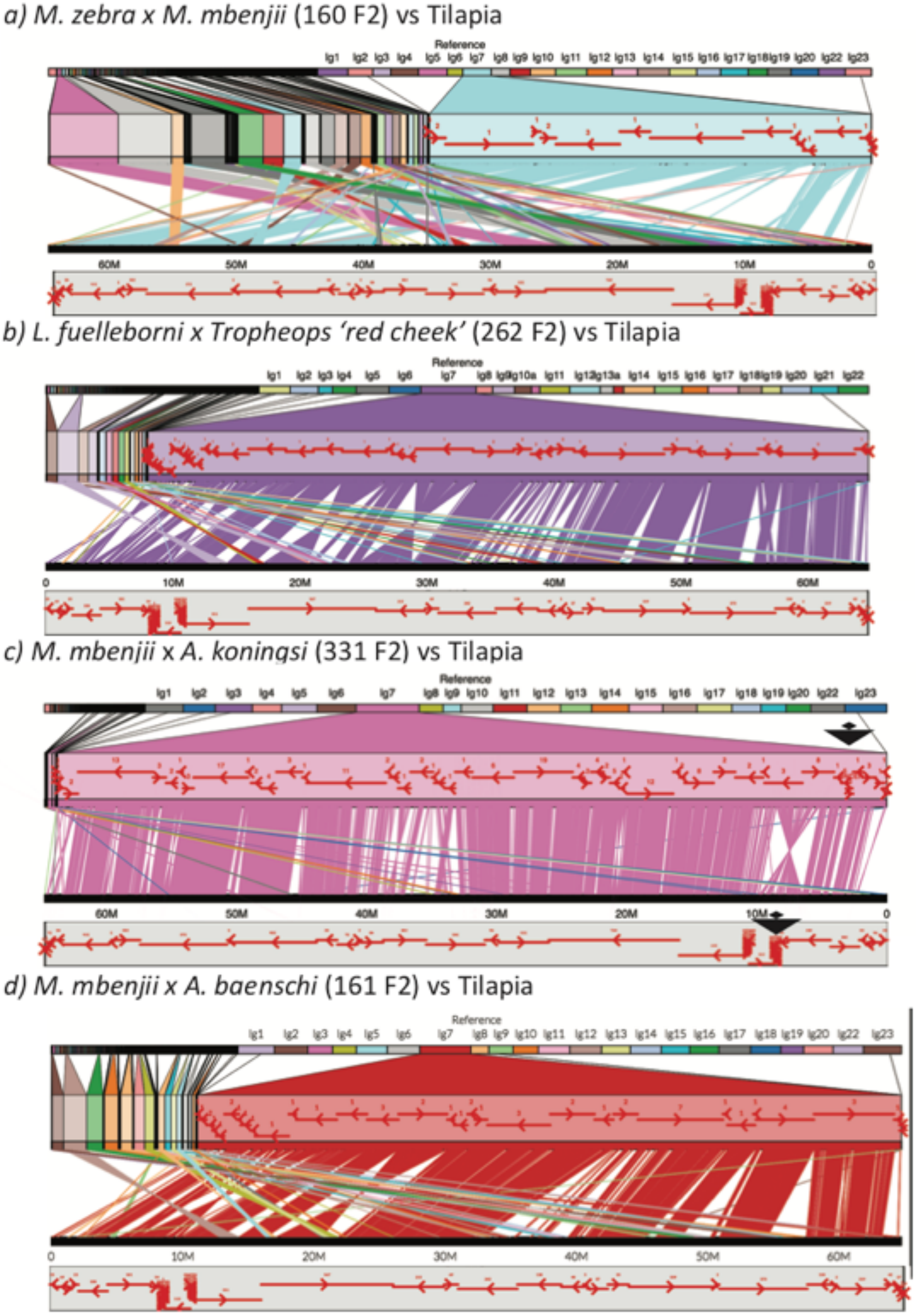
a) *M. zebra x M. mbenjii* (160 F2) vs Tilapia

b) *L. fuelleborni x Tropheops 'red cheek'* (262 F2) vs Tilapia

c) *M. mbenjii* x *A. koningsi* (331 F2) vs Tilapia

d) *M. mbenjii x A. baenschi* (161 F2) vs Tilapia

LG7



a) *M. zebra x M. mbenjii* (160 F2) vs Tilapia

b) *L. fuelleborni x Tropheops 'red cheek'* (262 F2) vs Tilapia

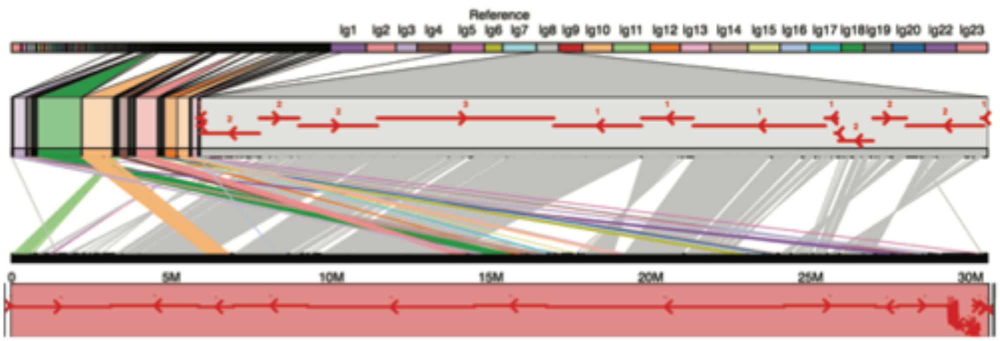
c) *M. mbenjii x A. koningsi* (331 F2) vs Tilapia

d) *M. mbenjii x A. baenschi* (161 F2) vs Tilapia

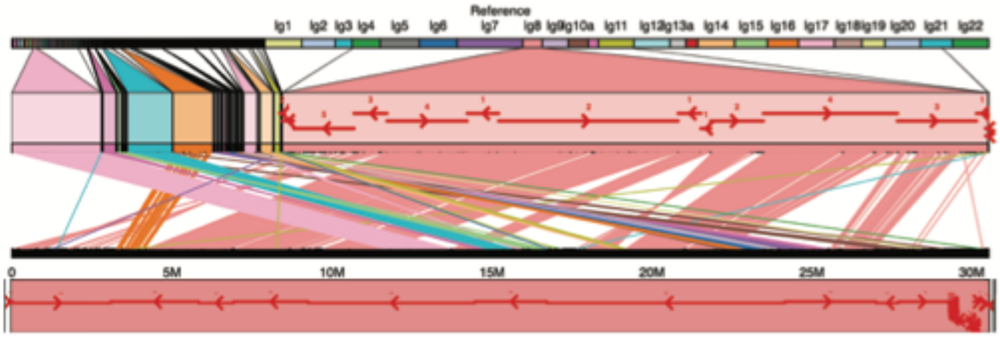
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
60M 50M 40M 30M 20M 10M 0
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10a lg11 lg12 lg13a lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg21 lg22
0 10M 20M 30M 40M 50M 60M
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
60M 50M 40M 30M 20M 10M 0
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
0 10M 20M 30M 40M 50M 60M

LG8

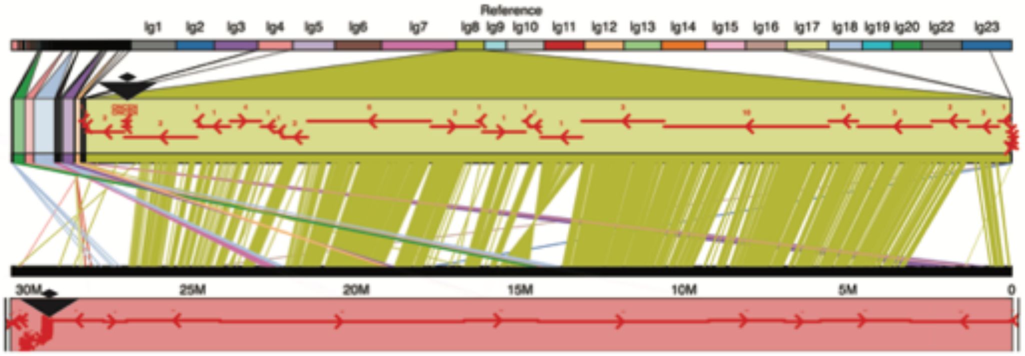
a) M. zebra x M. mbenjii (160 F2) vs Tilapia


Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
0
5M
10M
15M
20M
25M
30M

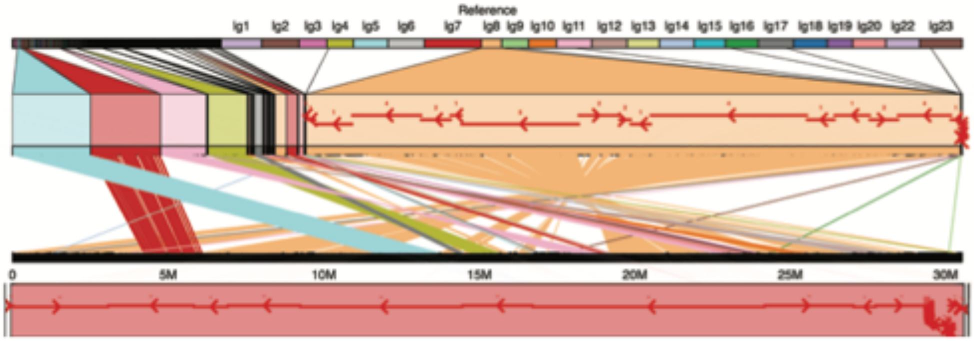
b) L. fuelleborni x Tropheops 'red cheek' (262 F2) vs Tilapia


Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10a lg11 lg12 lg13a lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg21 lg22
0
5M
10M
15M
20M
25M
30M

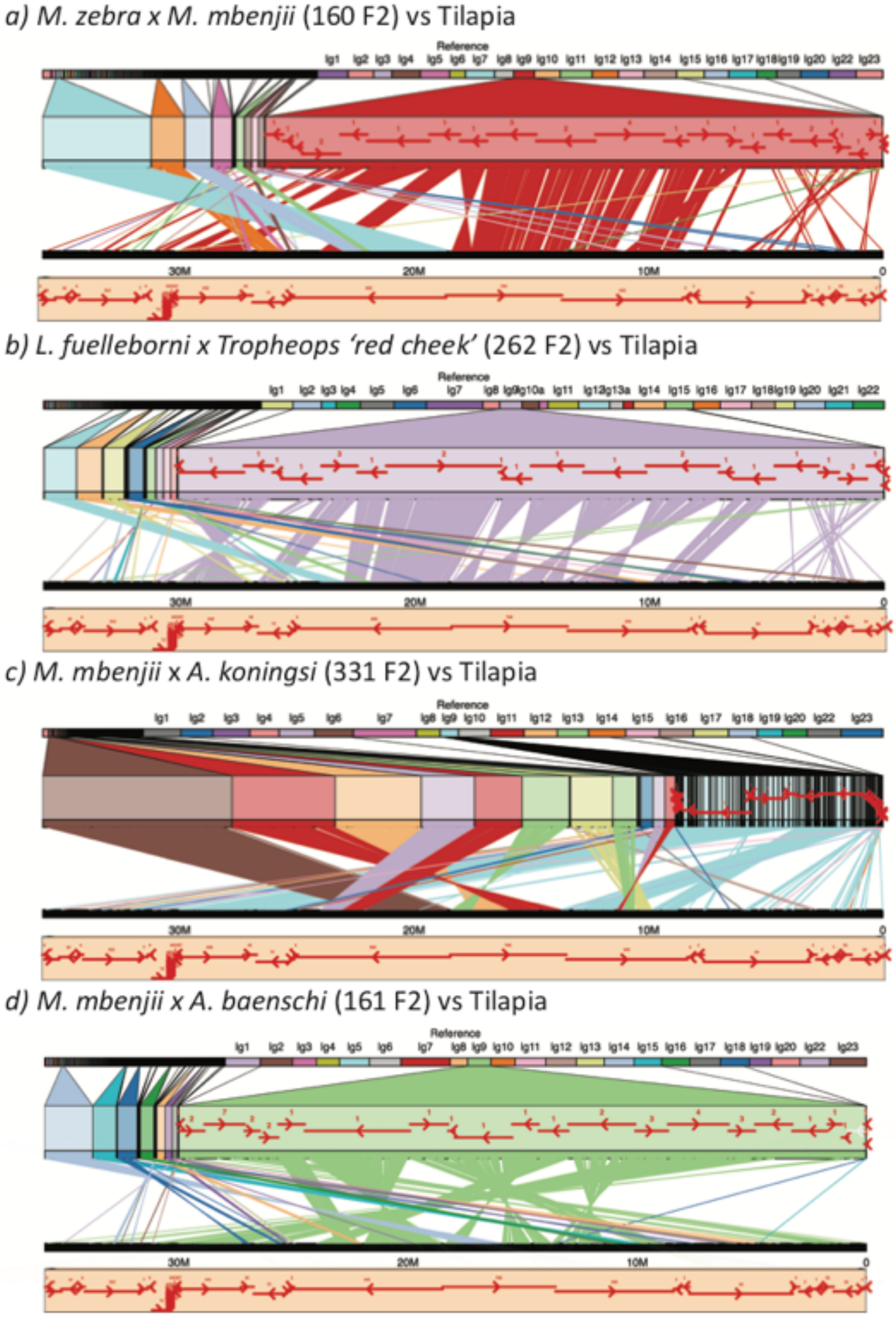
c) M. mbenjii x A. koningsi (331 F2) vs Tilapia


Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
30M
25M
20M
15M
10M
5M
0

d) M. mbenjii x A. baenschi (161 F2) vs Tilapia


Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
0
5M
10M
15M
20M
25M
30M

LG9



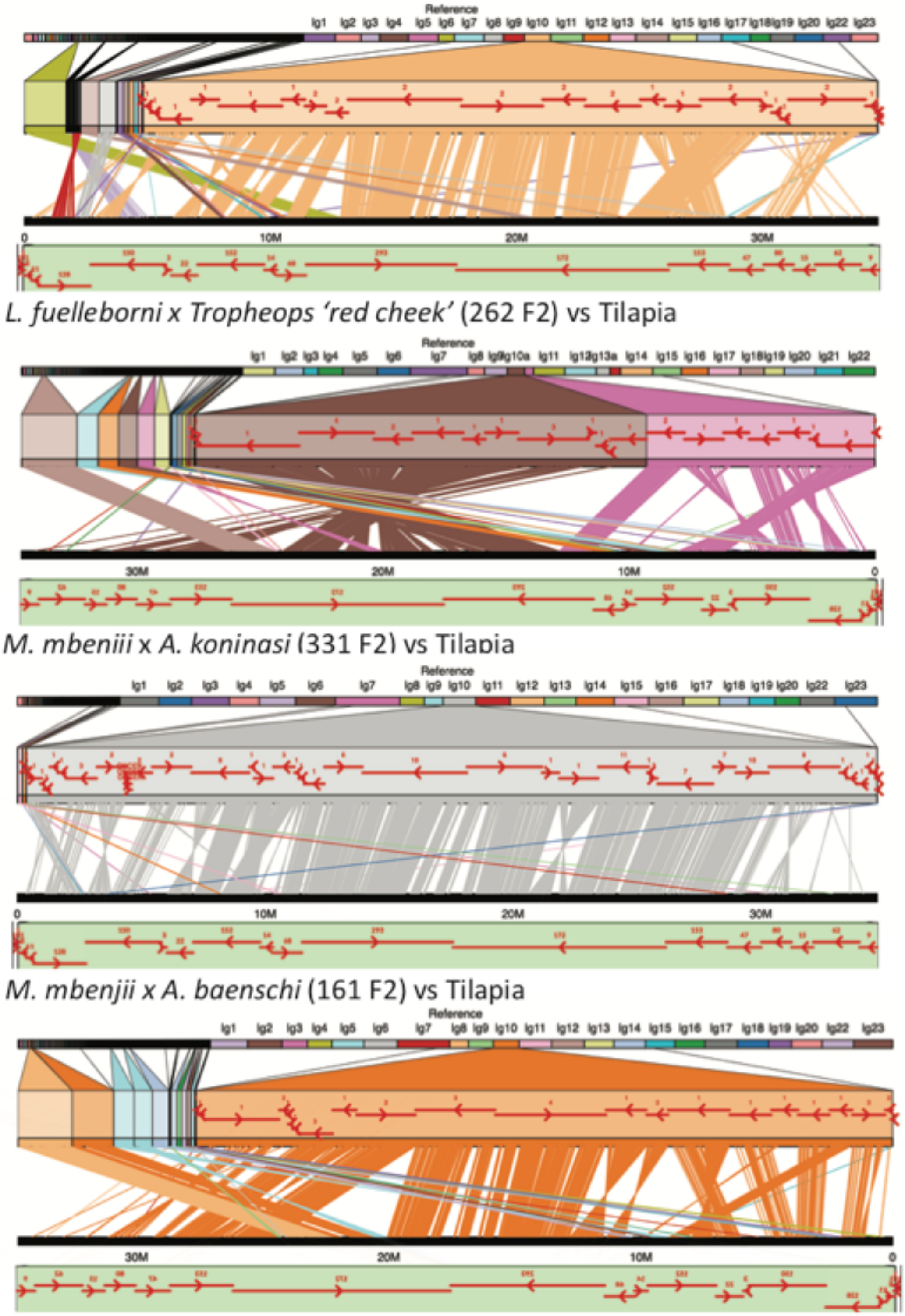
a) *M. zebra x M. mbenjii* (160 F2) vs Tilapia

b) *L. fuelleborni x Tropheops 'red cheek'* (262 F2) vs Tilapia

c) *M. mbenjii x A. koningsi* (331 F2) vs Tilapia

d) *M. mbenjii x A. baenschi* (161 F2) vs Tilapia

LG10



a) *M. zebra x M. mbenjii* (160 F2) vs Tilapia

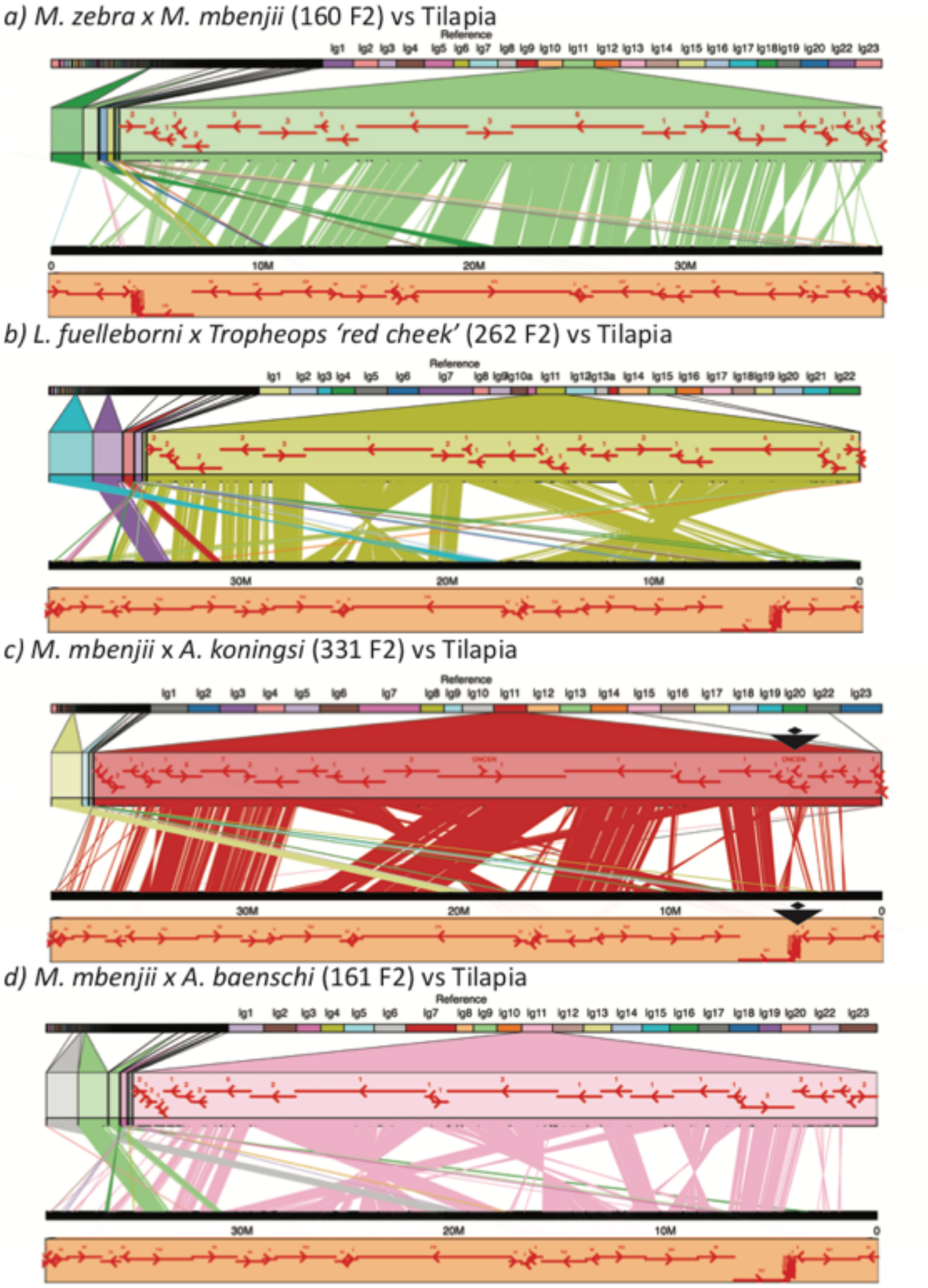
b) *L. fuelleborni x Tropheops 'red cheek'* (262 F2) vs Tilapia

c) *M. mbenjii* x *A. koningsi* (331 F2) vs Tilapia

d) *M. mbenjii x A. baenschi* (161 F2) vs Tilapia

Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
0
10M
20M
30M
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10a lg11 lg12 lg13a lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg21 lg22
30M
20M
10M
0
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
0
10M
20M
30M
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
30M
20M
10M
0

LG11



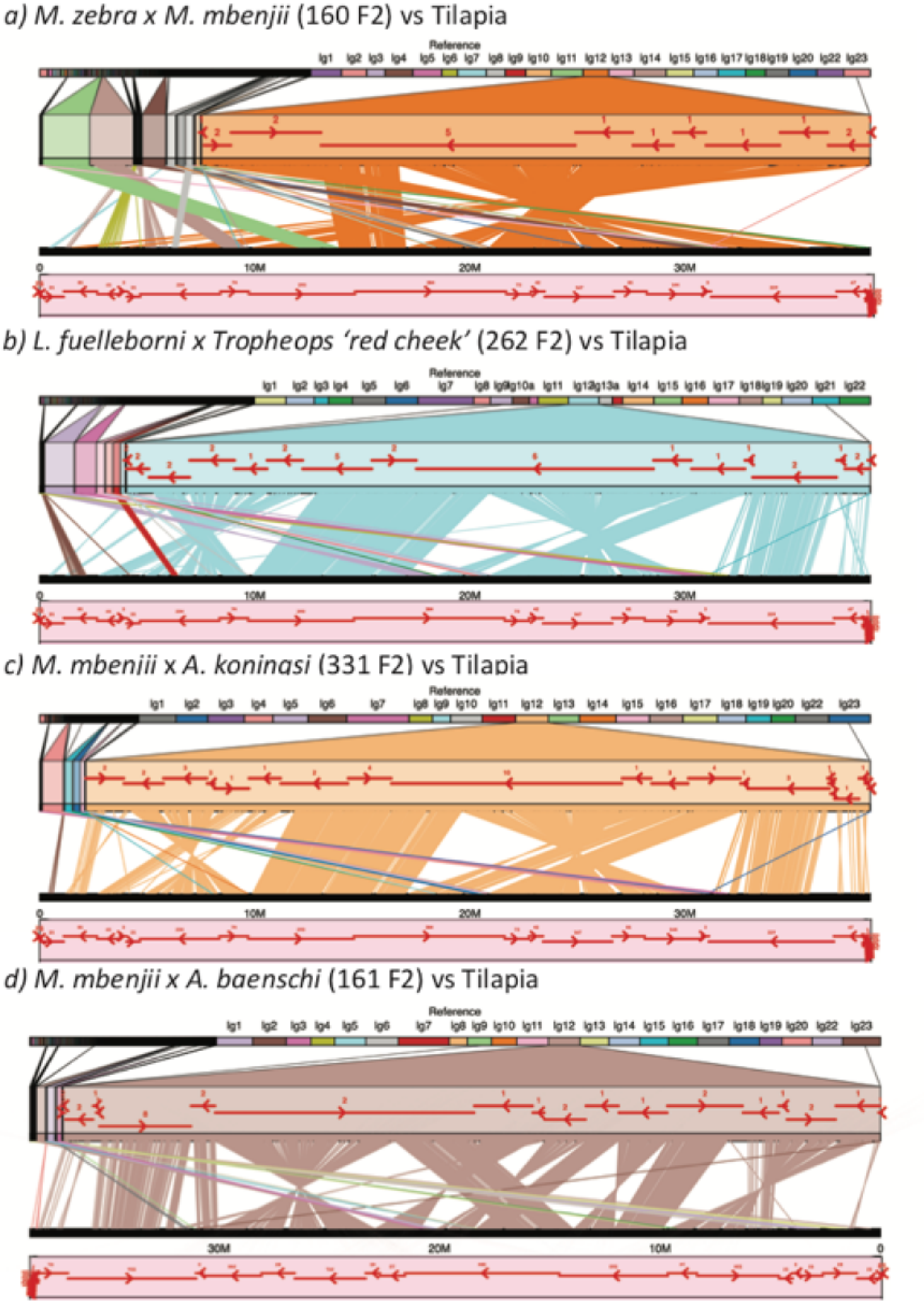
a) *M. zebra x M. mbenjii* (160 F2) vs Tilapia

b) *L. fuelleborni x Tropheops 'red cheek'* (262 F2) vs Tilapia

c) *M. mbenjii x A. koningsi* (331 F2) vs Tilapia

d) *M. mbenjii x A. baenschi* (161 F2) vs Tilapia

LG12

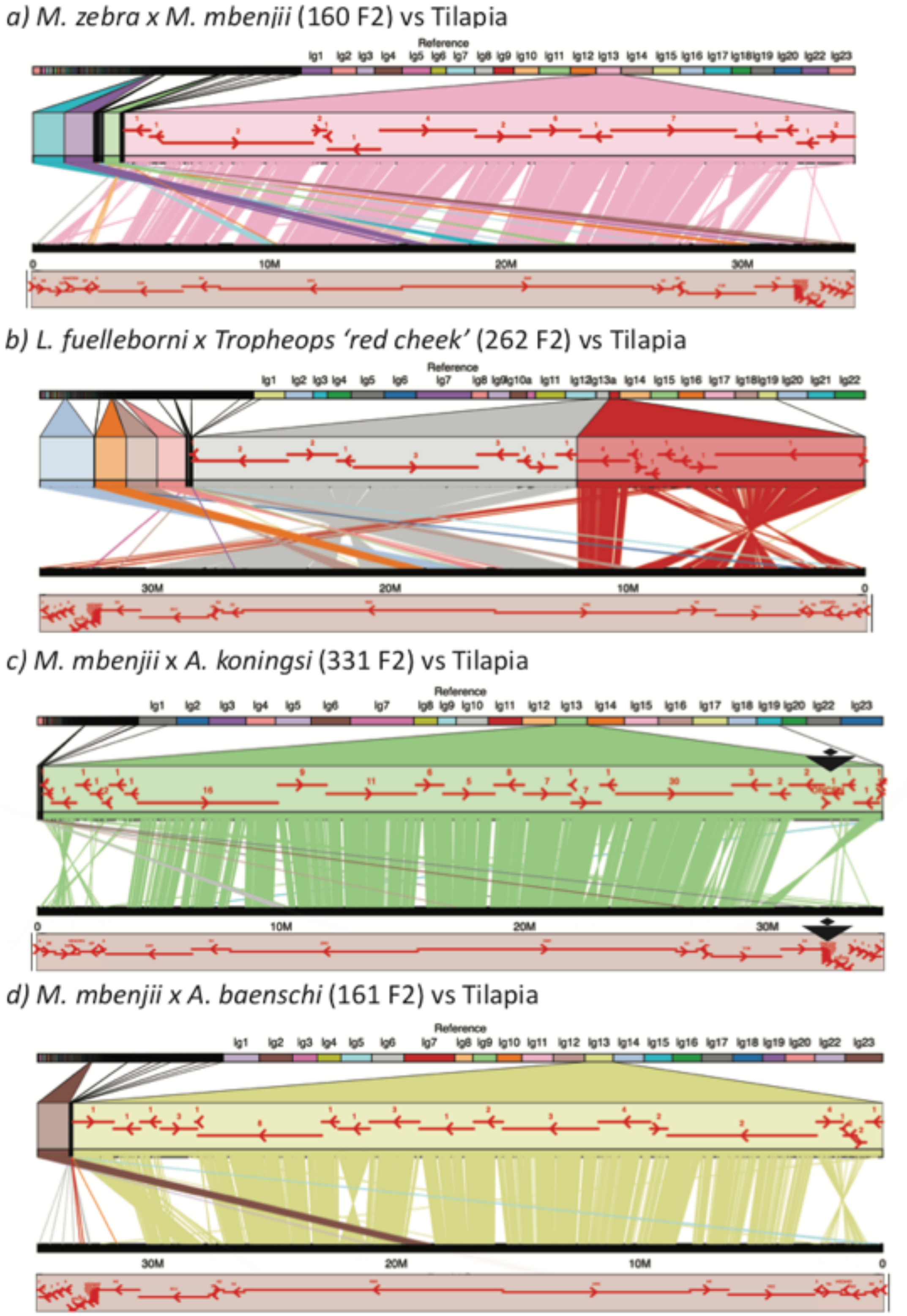

a) *M. zebra x M. mbenjii* (160 F2) vs Tilapia

b) *L. fuelleborni x Tropheops 'red cheek'* (262 F2) vs Tilapia

c) *M. mbenjii x A. koningsi* (331 F2) vs Tilapia

d) *M. mbenjii x A. baenschi* (161 F2) vs Tilapia

LG13



a) *M. zebra x M. mbenjii* (160 F2) vs Tilapia

Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
0
10M
20M
30M

b) *L. fuelleborni x Tropheops 'red cheek'* (262 F2) vs Tilapia

Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10a lg11 lg12 lg13a lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg21 lg22
30M
20M
10M
0

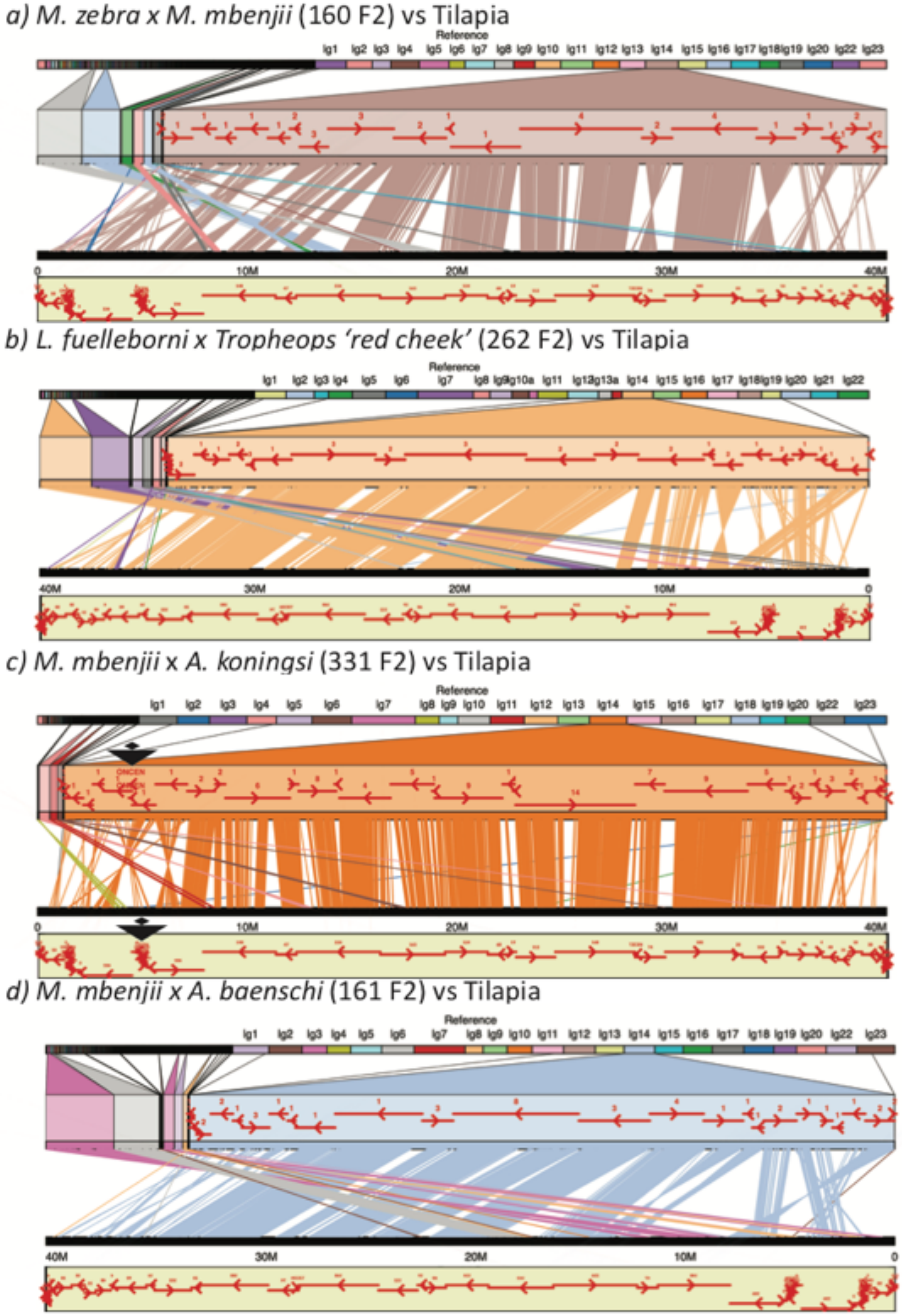
c) *M. mbenjii* x *A. koningsi* (331 F2) vs Tilapia

Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
0
10M
20M
30M

d) *M. mbenjii x A. baenschi* (161 F2) vs Tilapia

Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
30M
20M
10M
0

LG14



a) *M. zebra* x *M. mbenjii* (160 F2) vs Tilapia

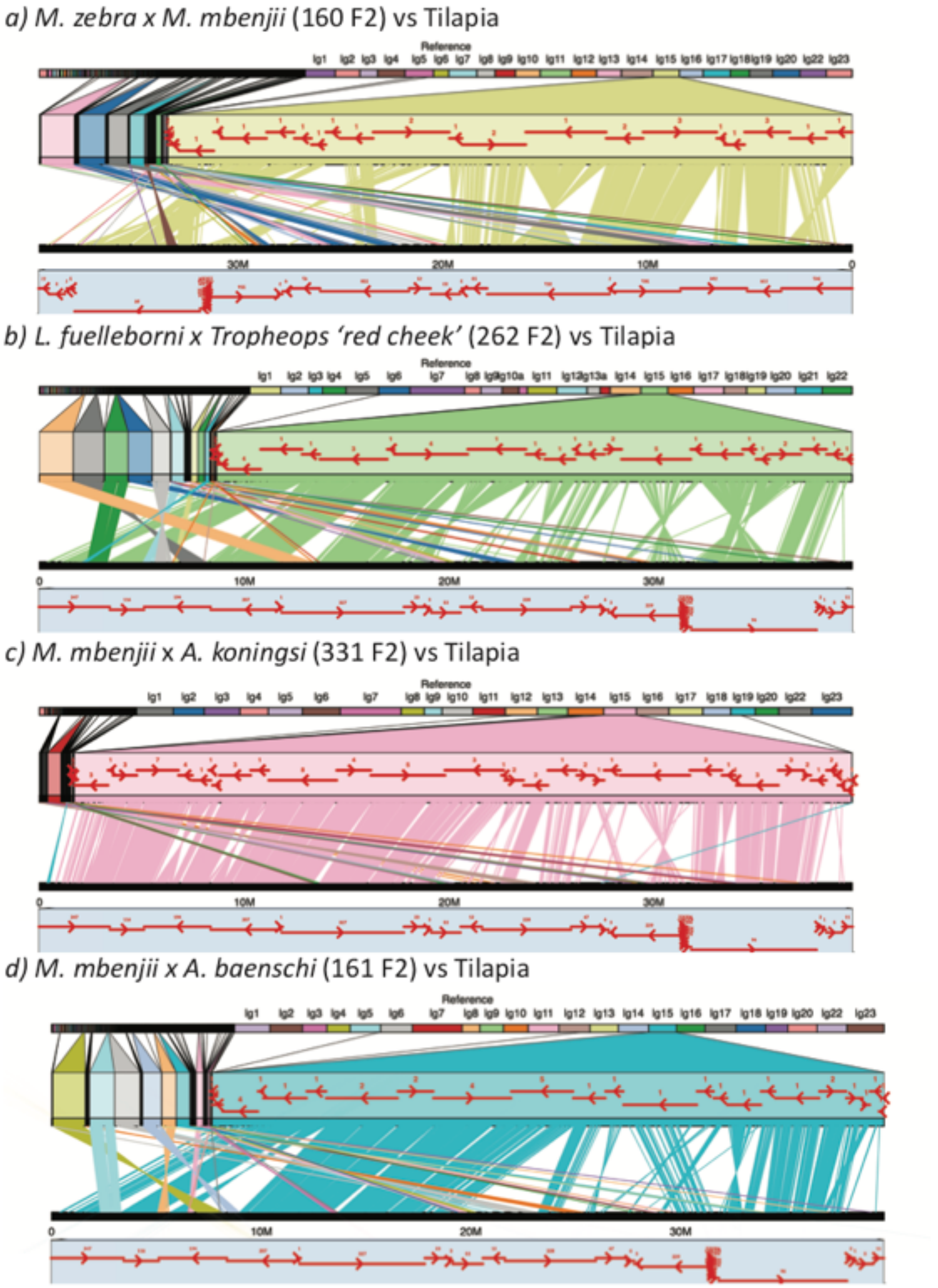
b) *L. fuelleborni* x *Tropheops 'red cheek'* (262 F2) vs Tilapia

c) *M. mbenjii* x *A. koningsi* (331 F2) vs Tilapia

d) *M. mbenjii* x *A. baenschi* (161 F2) vs Tilapia

Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
0
10M
20M
30M
40M
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10a lg11 lg12 lg13a lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg21 lg22
40M
30M
20M
10M
0
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
ONCEN
0
10M
20M
30M
40M
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
40M
30M
20M
10M
0

LG15



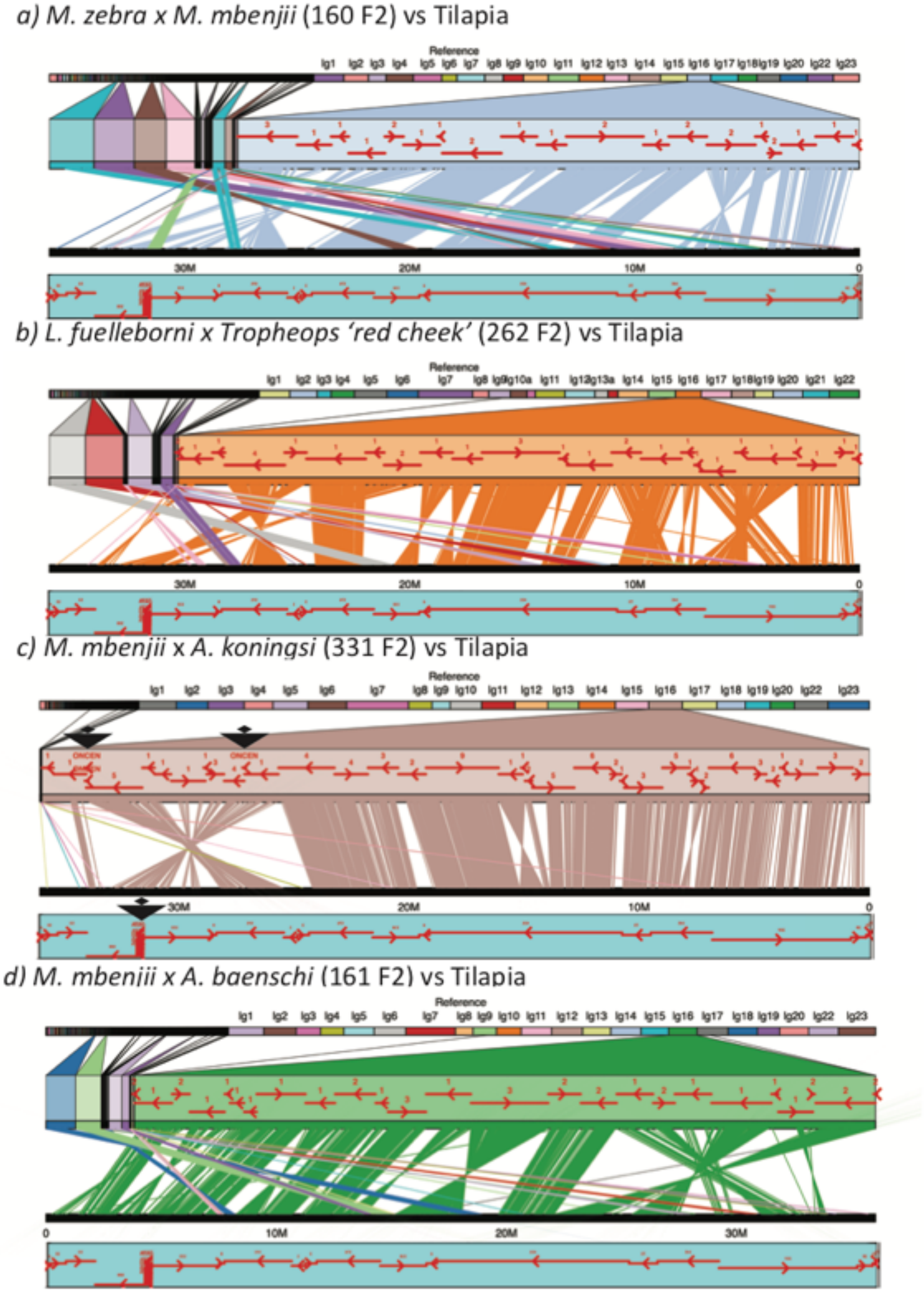
a) *M. zebra x M. mbenjii* (160 F2) vs Tilapia

b) *L. fuelleborni x Tropheops 'red cheek'* (262 F2) vs Tilapia

c) *M. mbenjii x A. koningsi* (331 F2) vs Tilapia

d) *M. mbenjii x A. baenschi* (161 F2) vs Tilapia

LG16



a) *M. zebra x M. mbenjii* (160 F2) vs Tilapia

b) *L. fuelleborni x Tropheops 'red cheek'* (262 F2) vs Tilapia

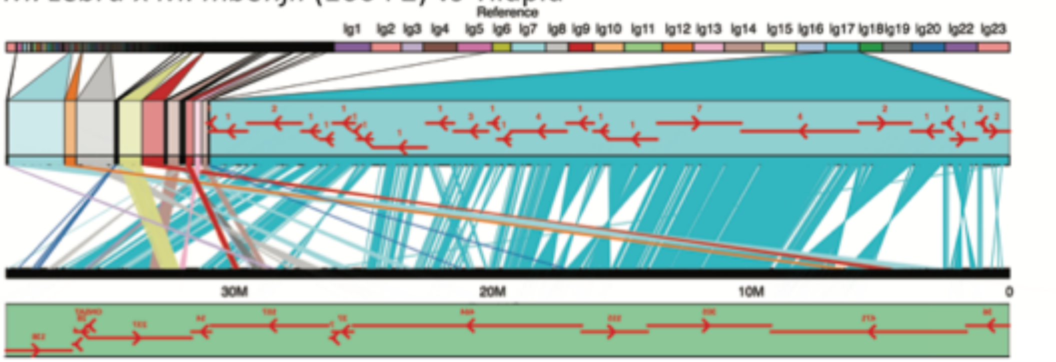
c) *M. mbenjii x A. koningsi* (331 F2) vs Tilapia

d) *M. mbenjii x A. baenschi* (161 F2) vs Tilapia

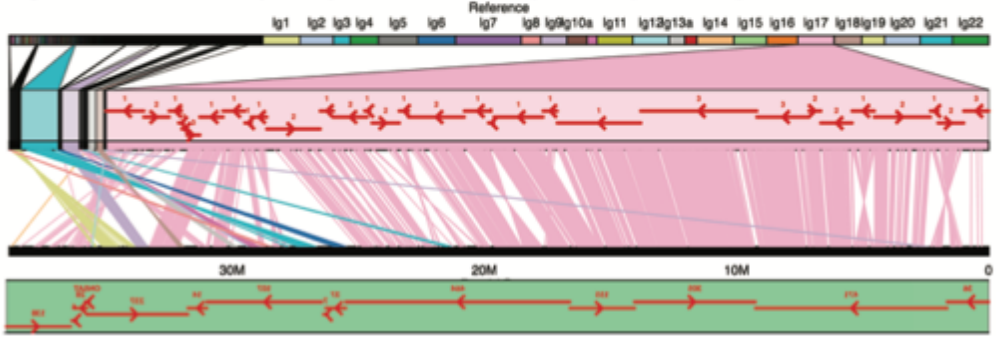
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
30M
20M
10M
0
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10a lg11 lg12 lg13a lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg21 lg22
30M
20M
10M
0
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
ONCEN
ONCEN
30M
20M
10M
0
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
0
10M
20M
30M

LG17

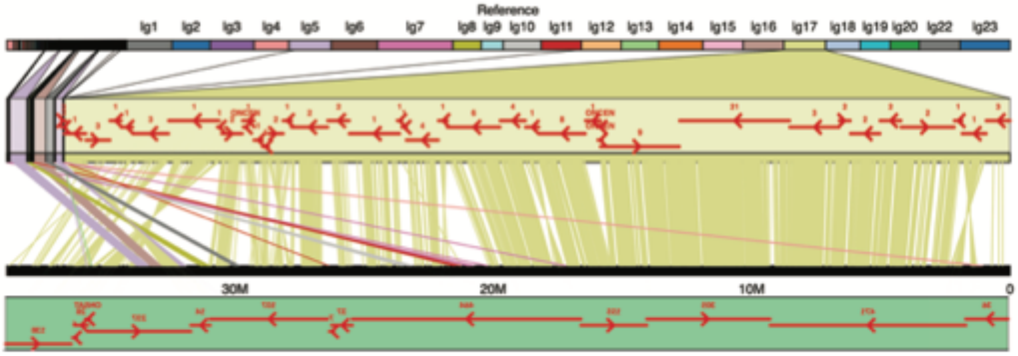
a) M. zebra x M. mbenjii (160 F2) vs Tilapia


Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
30M
20M
10M
0

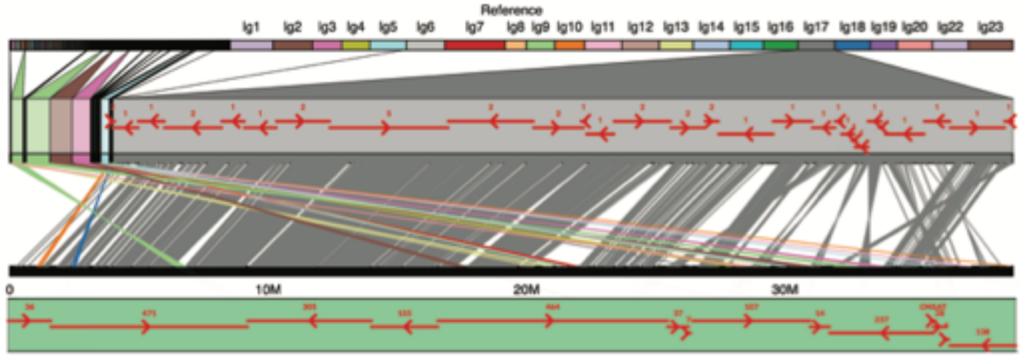
b) L. fuelleborni x Tropheops 'red cheek' (262 F2) vs Tilapia


Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10a lg11 lg12 lg13a lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg21 lg22
30M
20M
10M
0

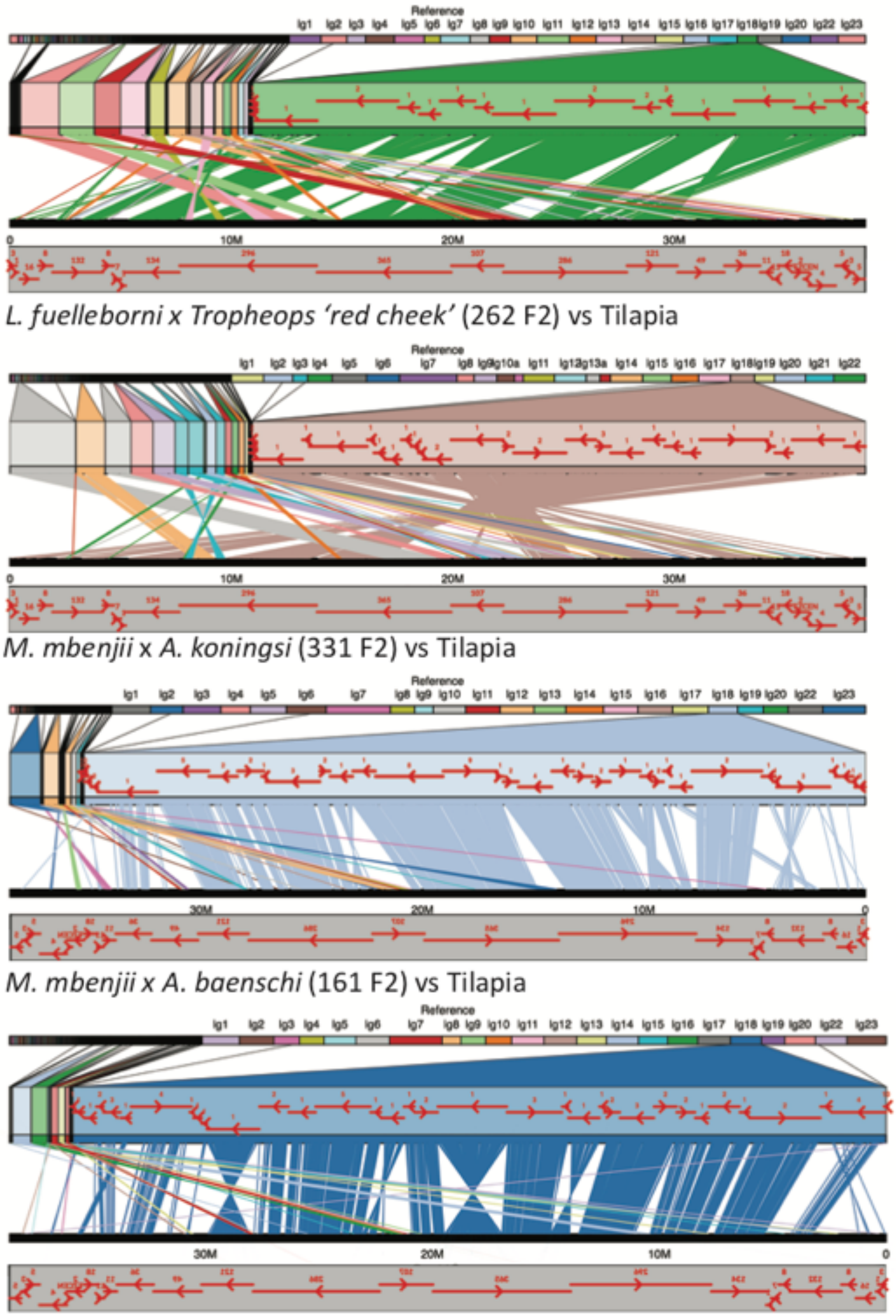
c) M. mbenjii x *A. koningsi* (331 F2) vs Tilapia


Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
30M
20M
10M
0

d) M. mbenjii x A. baenschi (161 F2) vs Tilapia


Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
0
10M
20M
30M

LG18



a) *M. zebra x M. mbenjii* (160 F2) vs Tilapia

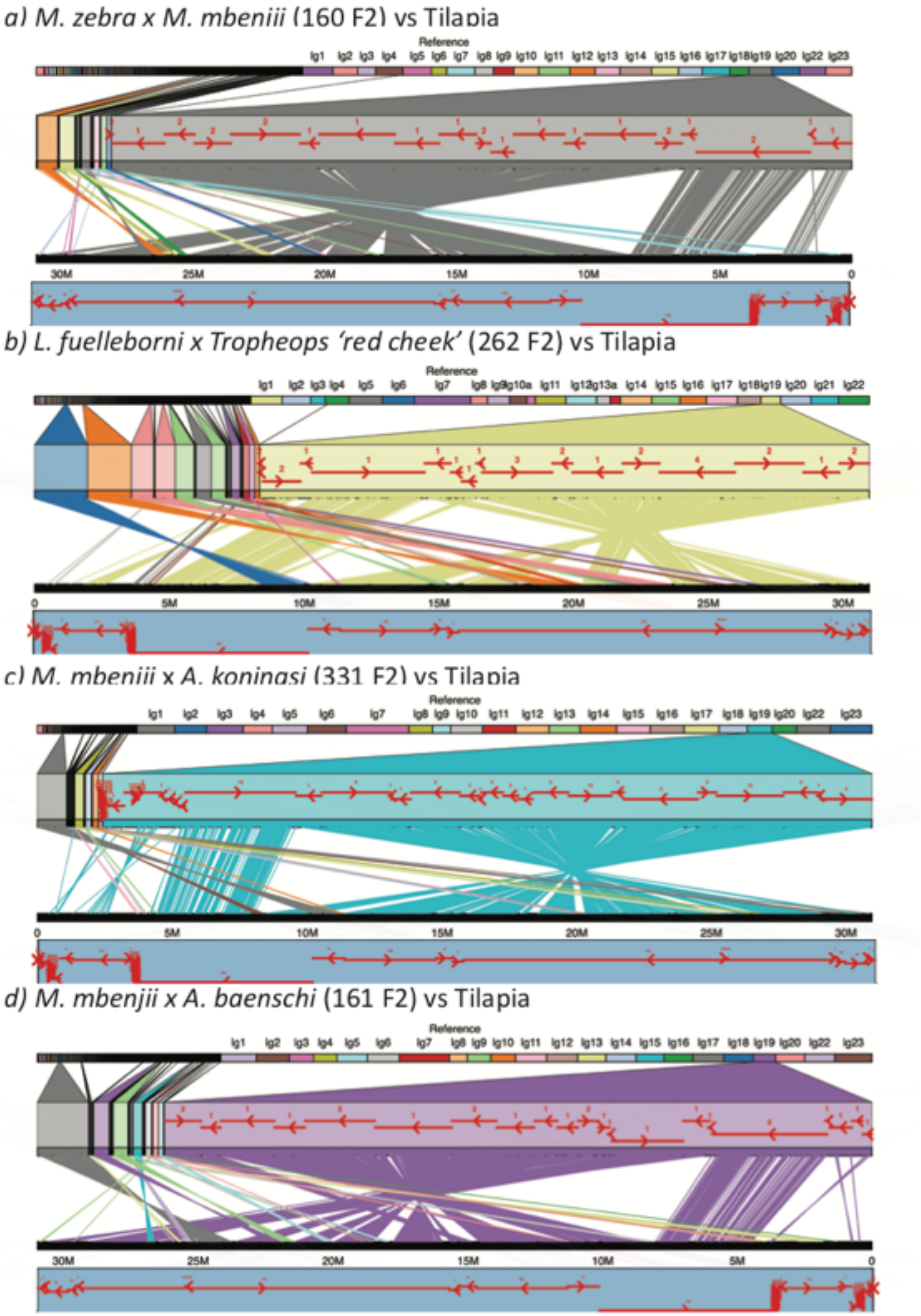
b) *L. fuelleborni x Tropheops 'red cheek'* (262 F2) vs Tilapia

c) *M. mbenjii x A. koningsi* (331 F2) vs Tilapia

d) *M. mbenjii x A. baenschi* (161 F2) vs Tilapia

Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
0
10M
20M
30M
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10a lg11 lg12 lg13a lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg21 lg22
0
10M
20M
30M
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
30M
20M
10M
0
Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
30M
20M
10M
0

LG19



a) *M. zebra x M. mbeniii* (160 F2) vs Tilapia

Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
30M
25M
20M
15M
10M
5M
0

b) *L. fuelleborni x Tropheops 'red cheek'* (262 F2) vs Tilapia

Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10a lg11 lg12 lg13a lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg21 lg22
0
5M
10M
15M
20M
25M
30M

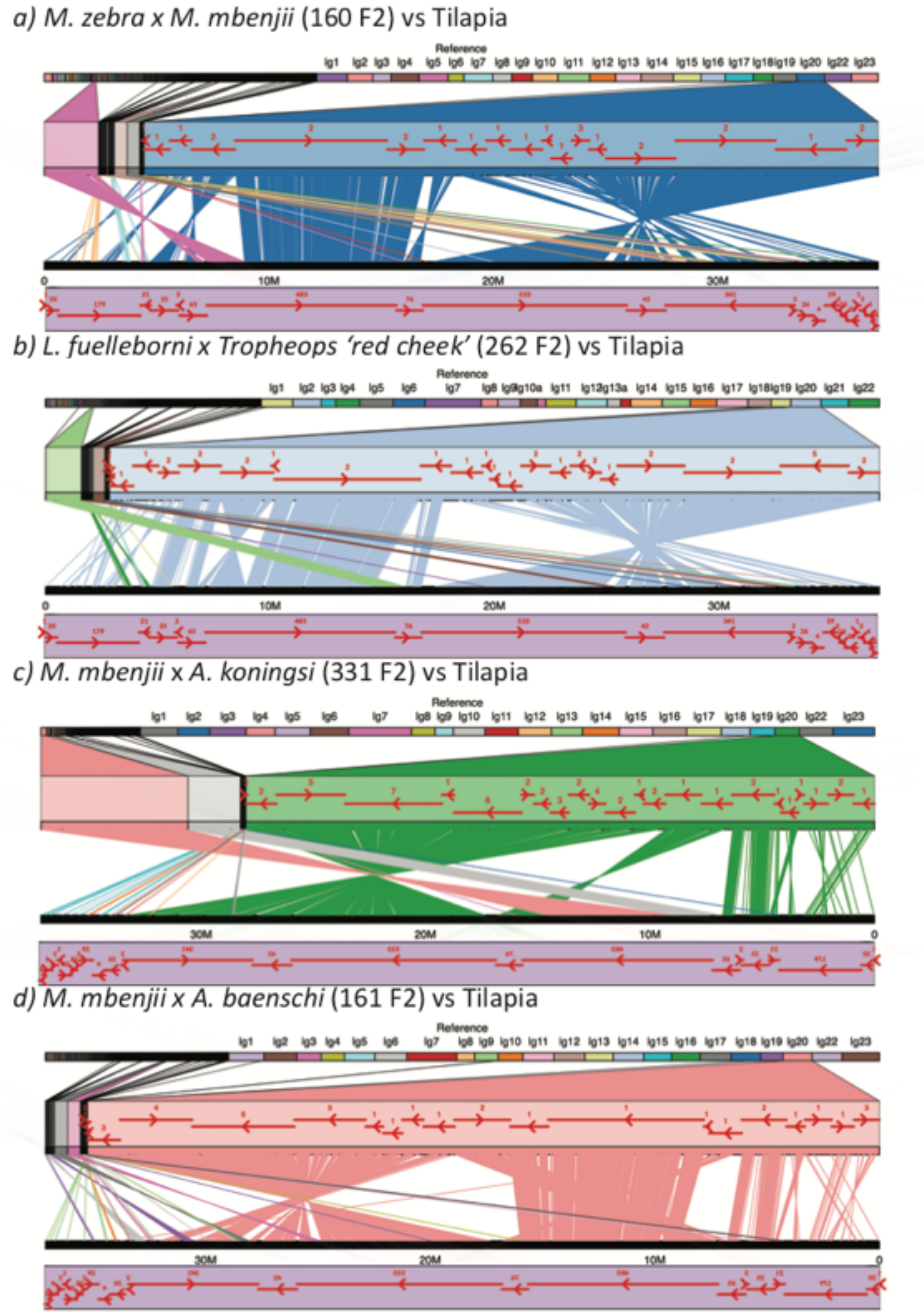
c) *M. mbeniii x A. koningsi* (331 F2) vs Tilapia

Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
0
5M
10M
15M
20M
25M
30M

d) *M. mbenjii x A. baenschi* (161 F2) vs Tilapia

Reference
lg1 lg2 lg3 lg4 lg5 lg6 lg7 lg8 lg9 lg10 lg11 lg12 lg13 lg14 lg15 lg16 lg17 lg18 lg19 lg20 lg22 lg23
30M
25M
20M
15M
10M
5M
0

LG20



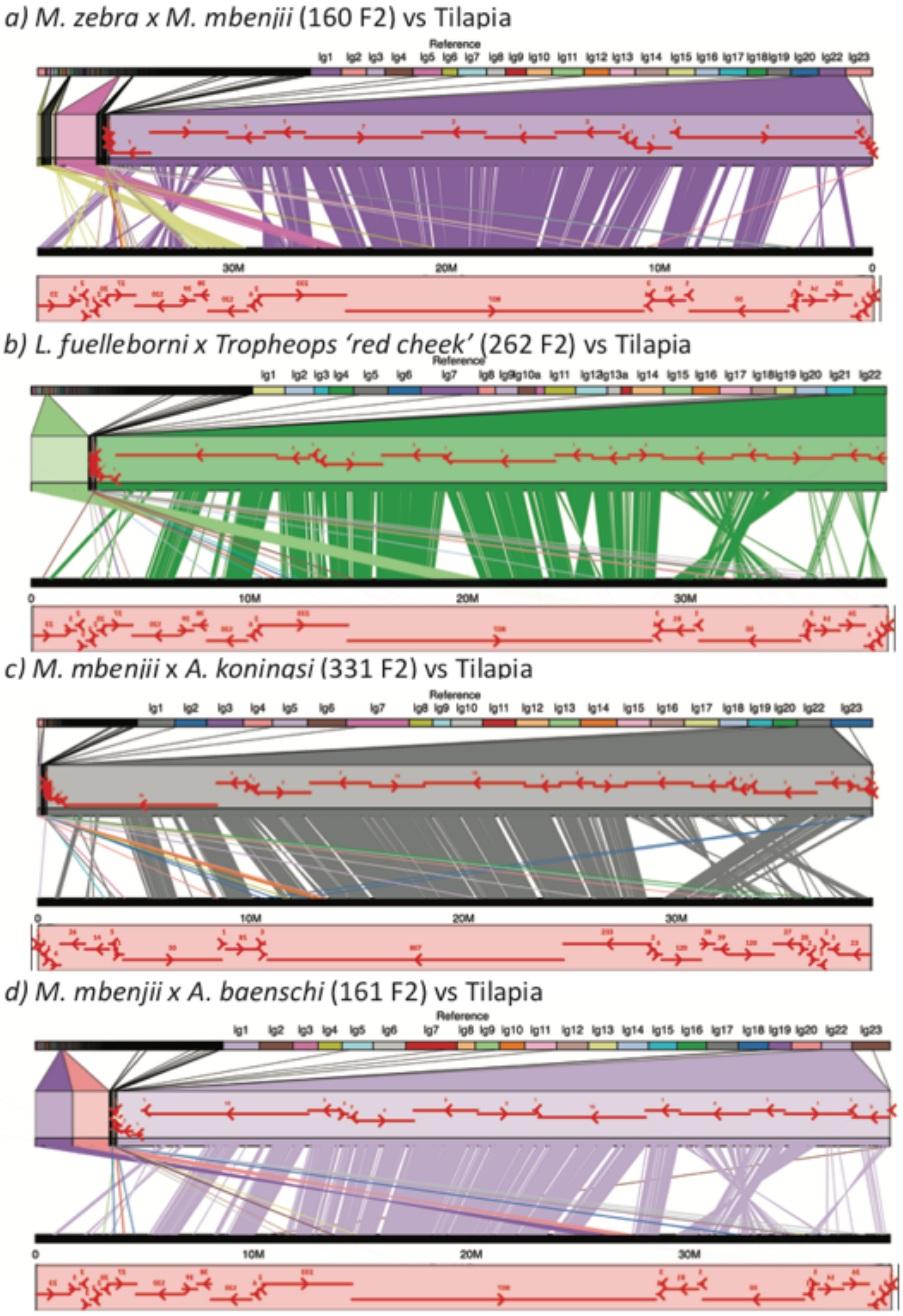
a) *M. zebra x M. mbenjii* (160 F2) vs Tilapia

b) *L. fuelleborni x Tropheops 'red cheek'* (262 F2) vs Tilapia

c) *M. mbenjii x A. koningsi* (331 F2) vs Tilapia

d) *M. mbenjii x A. baenschi* (161 F2) vs Tilapia

LG22



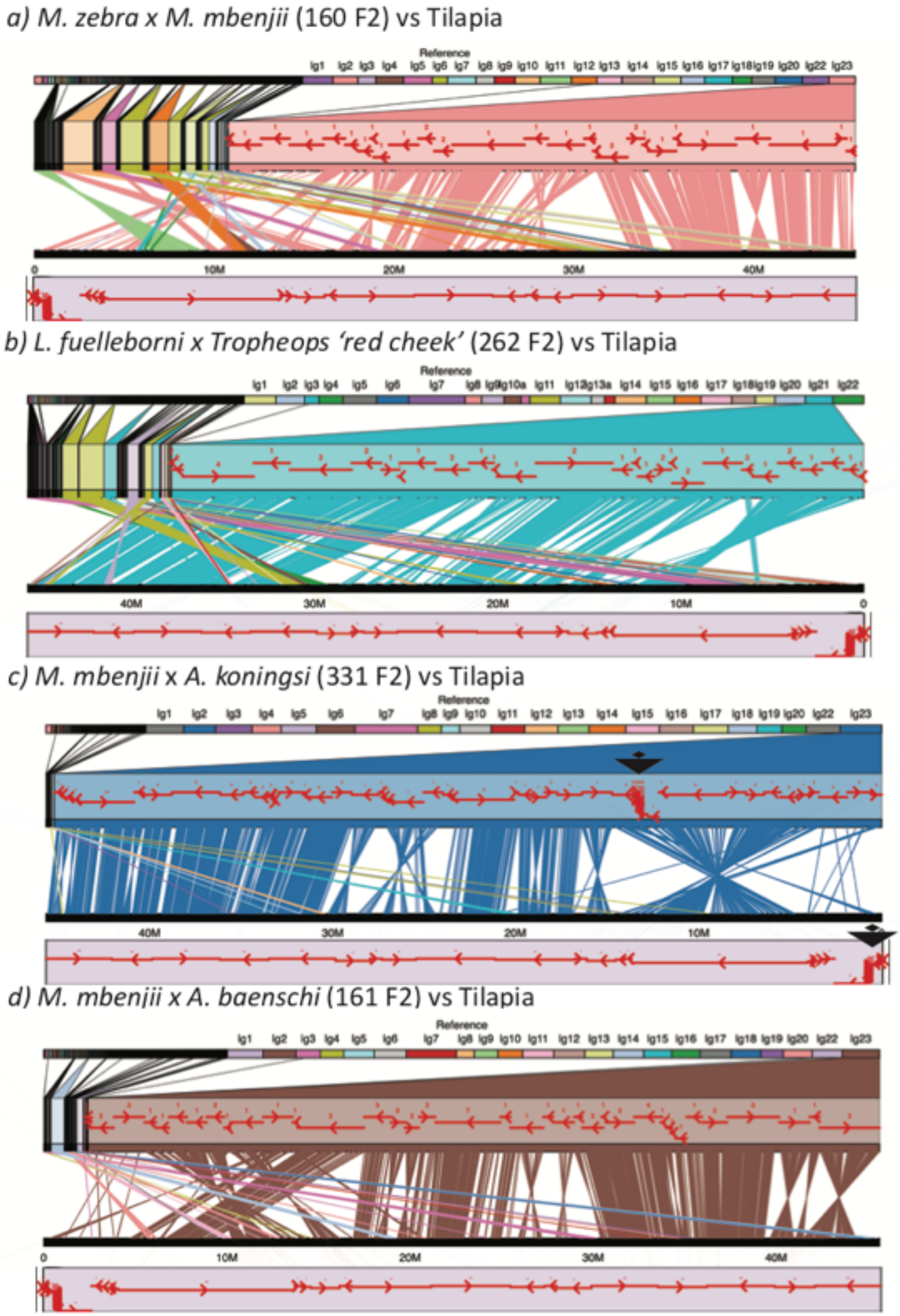
a) *M. zebra x M. mbenjii* (160 F2) vs Tilapia

b) *L. fuelleborni x Tropheops 'red cheek'* (262 F2) vs Tilapia

c) *M. mbenjii x A. koningsi* (331 F2) vs Tilapia

d) *M. mbenjii x A. baenschi* (161 F2) vs Tilapia

LG23



a) *M. zebra x M. mbenjii* (160 F2) vs Tilapia

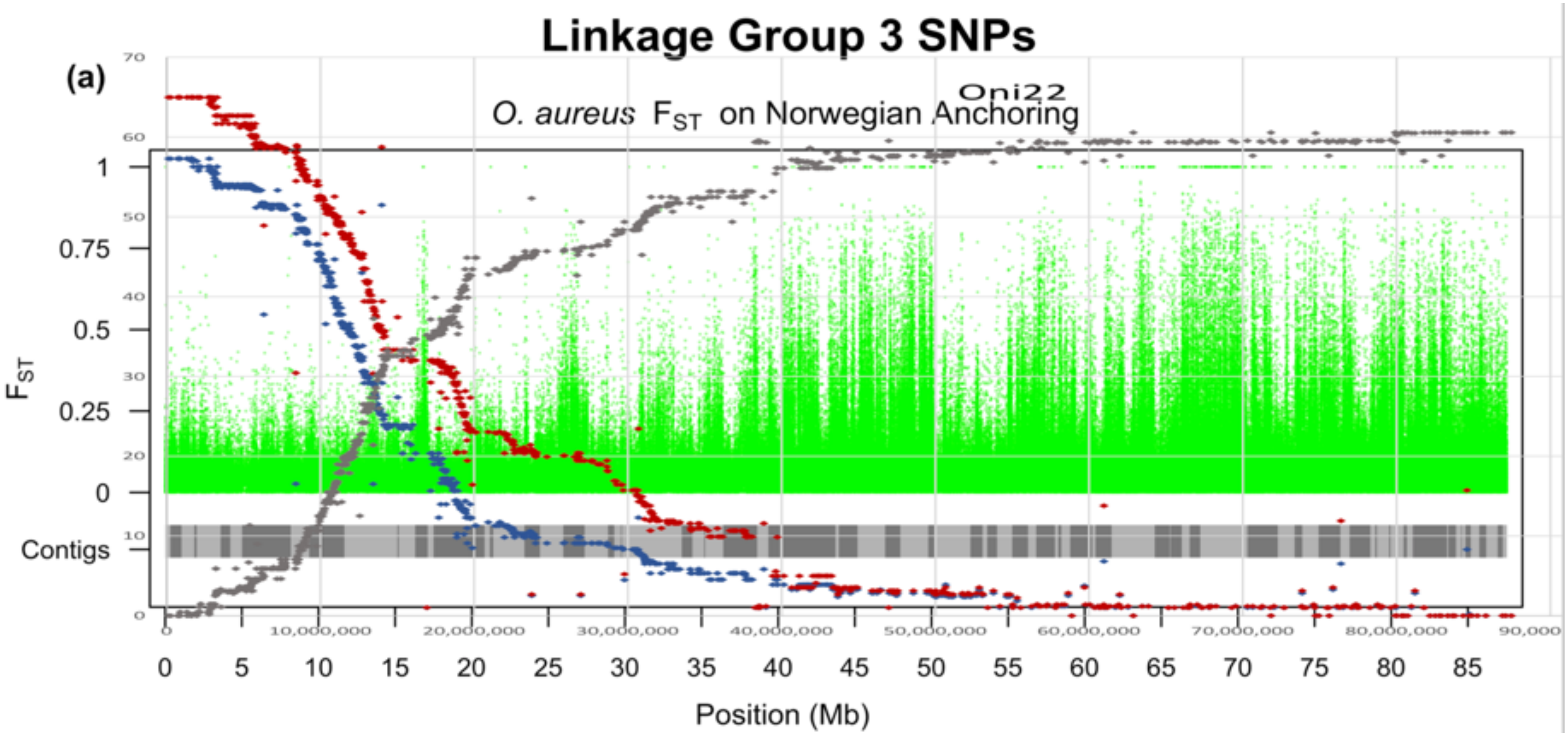
b) *L. fuelleborni x Tropheops 'red cheek'* (262 F2) vs Tilapia

c) *M. mbenjii x A. koningsi* (331 F2) vs Tilapia

d) *M. mbenjii x A. baenschi* (161 F2) vs Tilapia

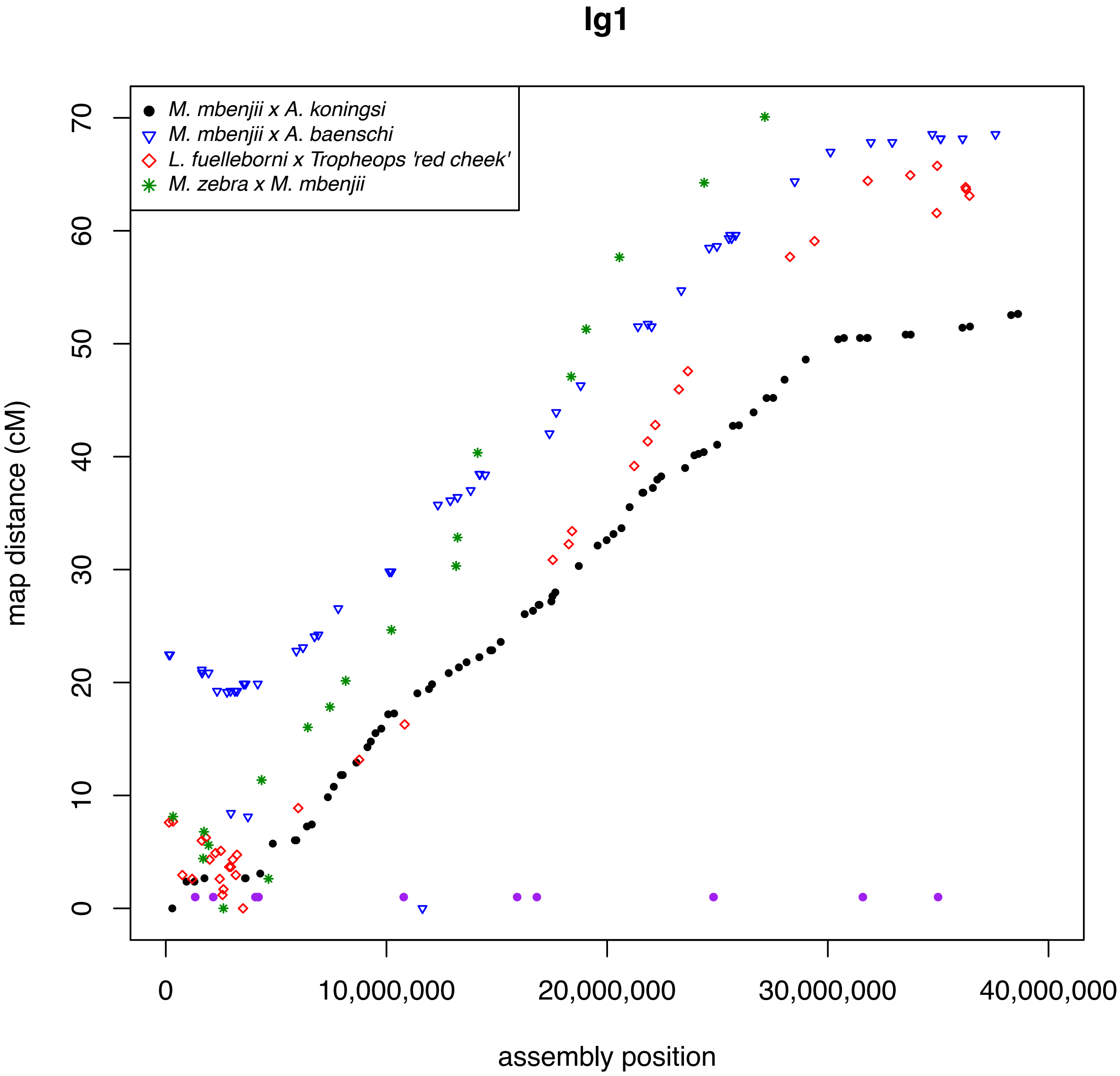
Appendix E

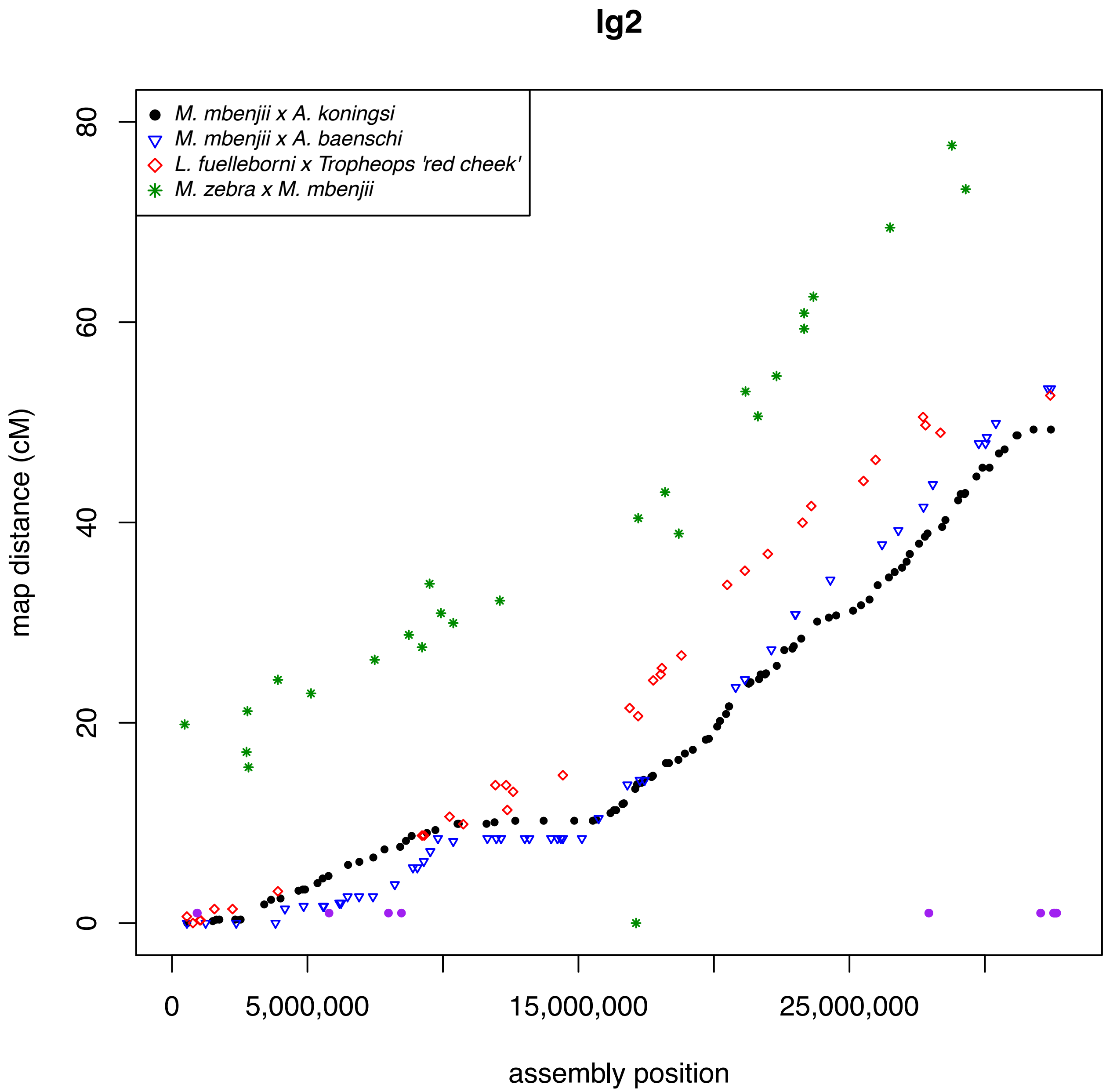
Comparison of male and female *O. aureus* LG3 WZ with an overlay of the *O. niloticus* recombination curves in Appendix G.

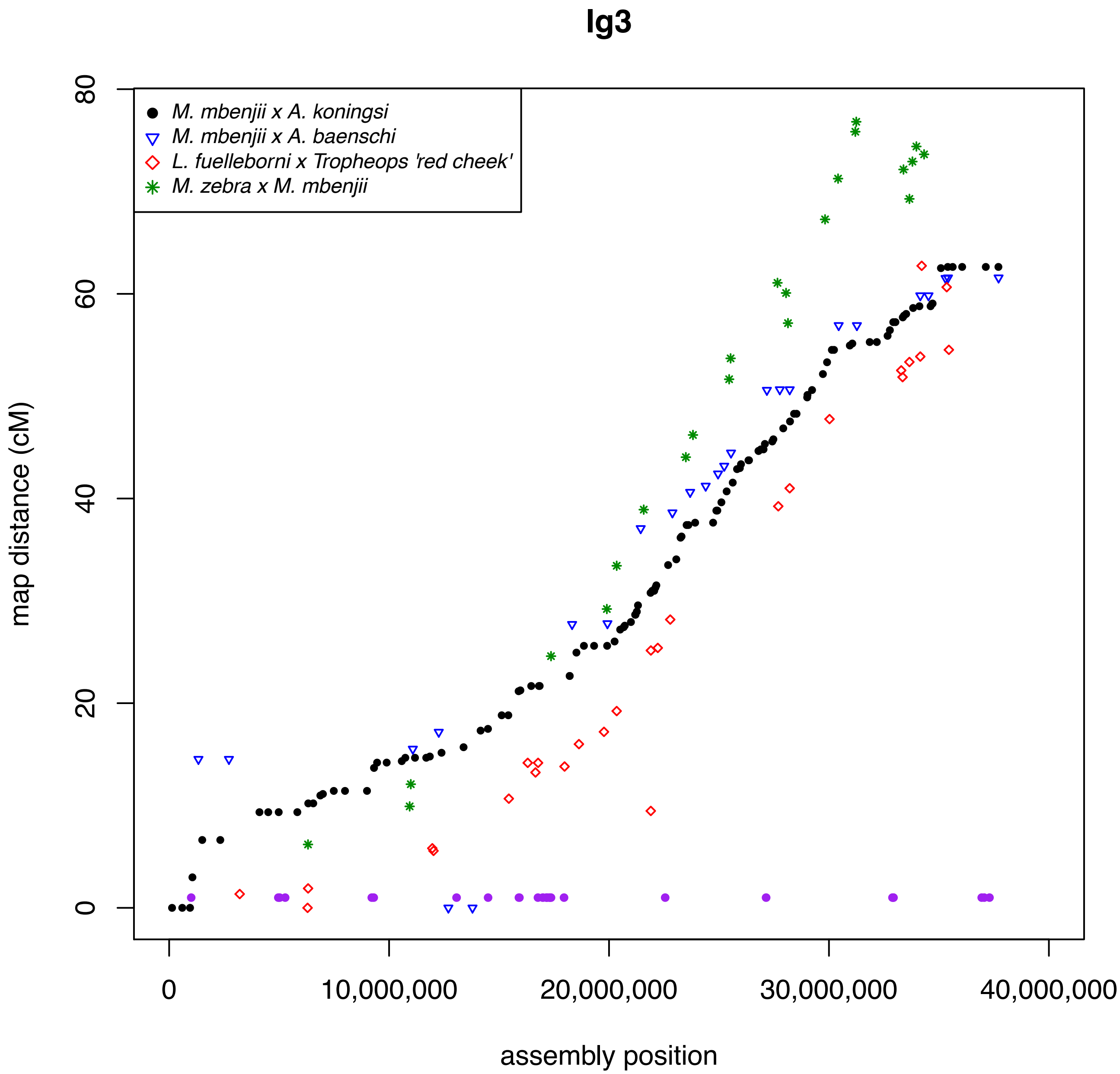


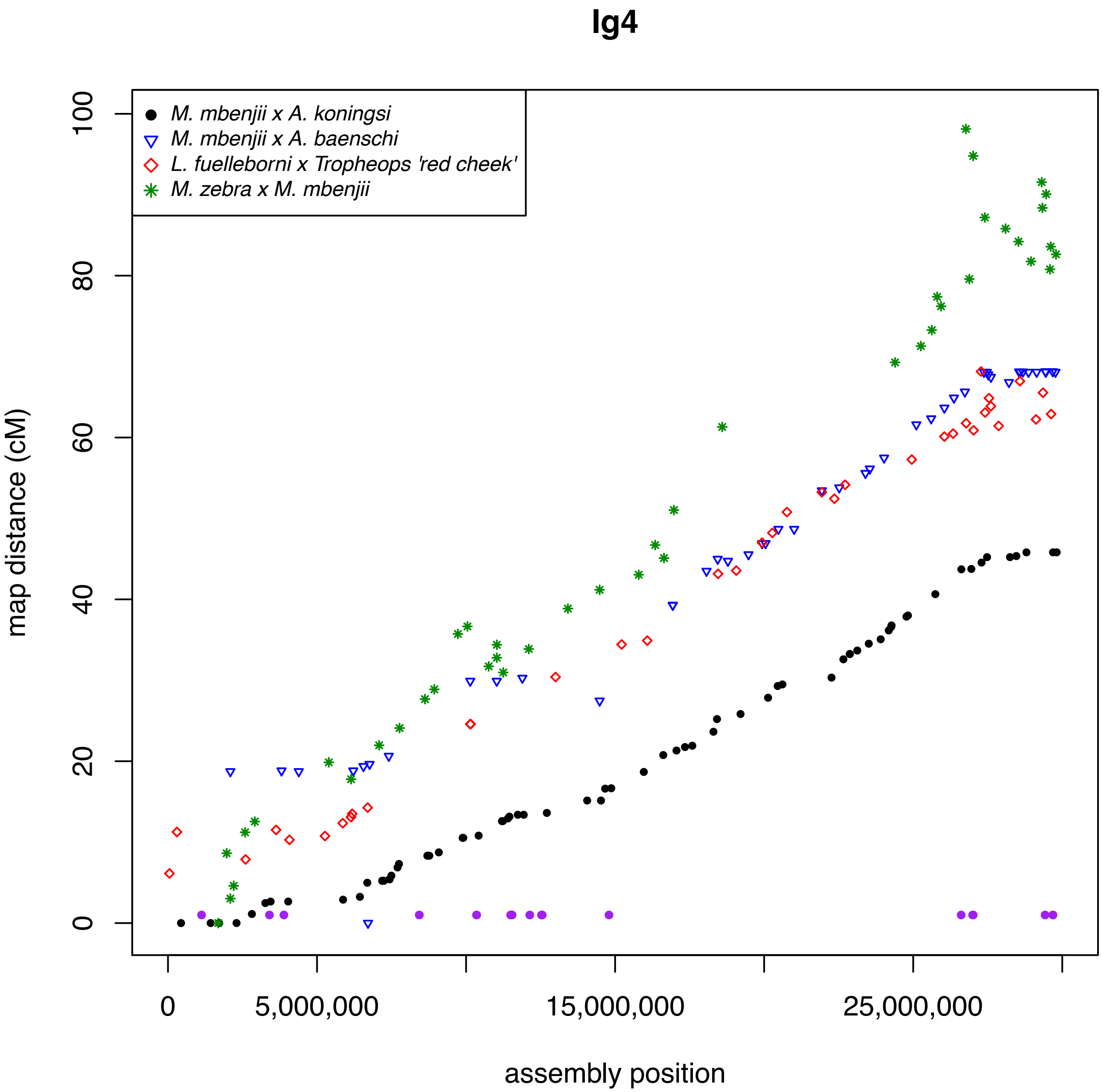
<u>*Appendix F*</u>

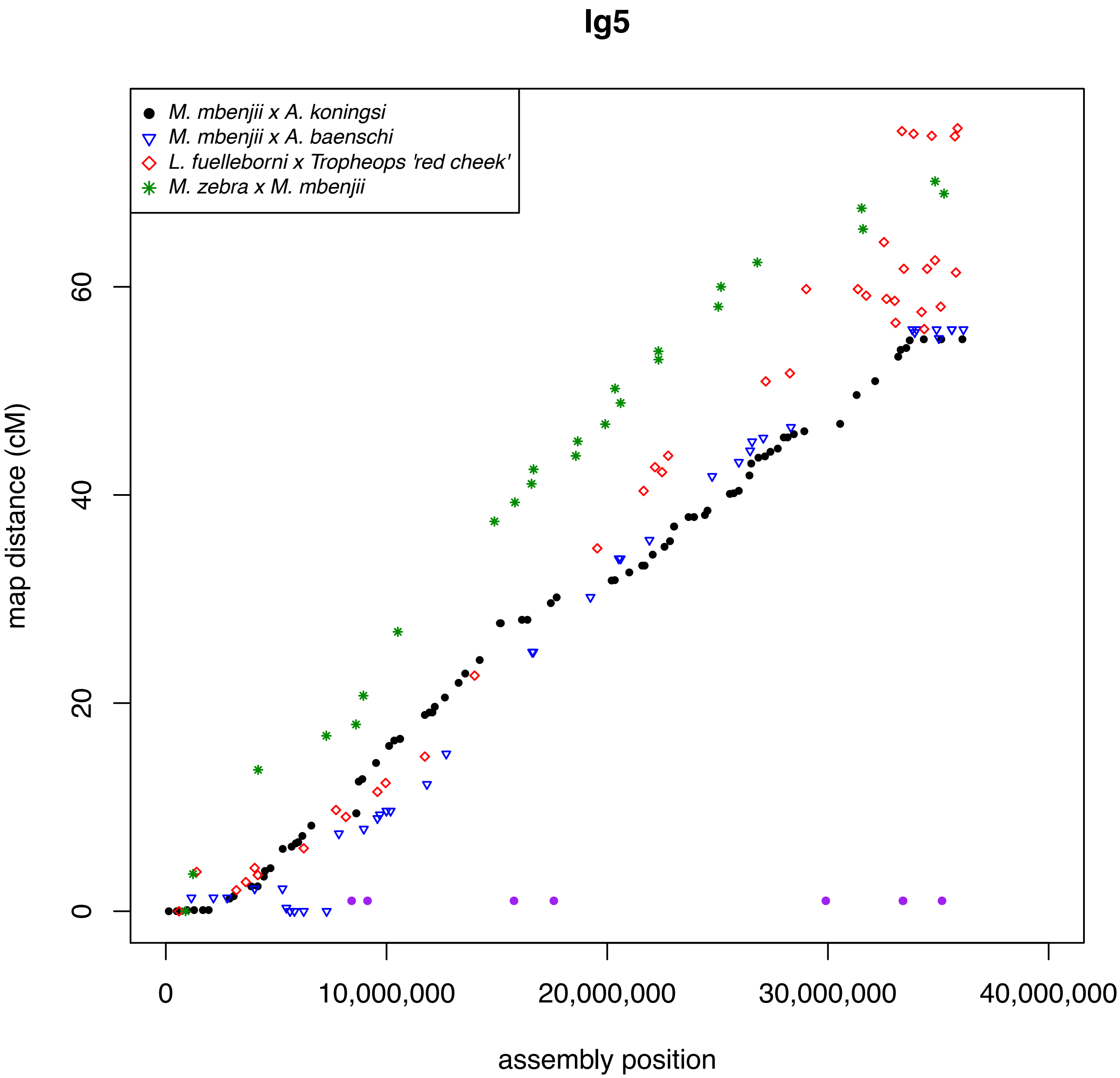
Comparison of recombination in the four genetic maps. LGs from maps that needed to be reversed from their original published order are indicated in the legend. The detected misassembly is included as "LG12 misassembly". B chromosome "blocks" (Chapter 5) are shown in purple.

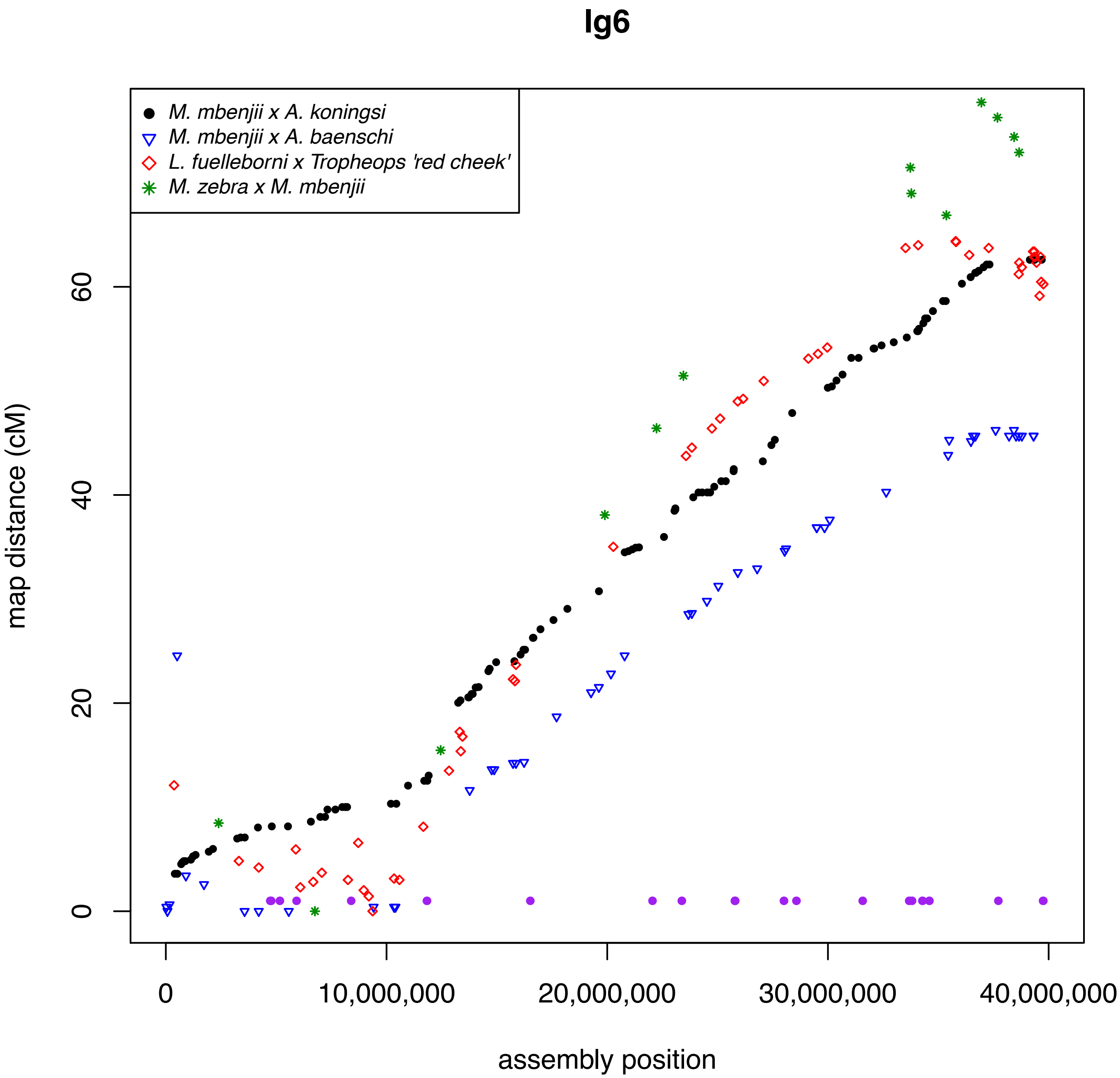


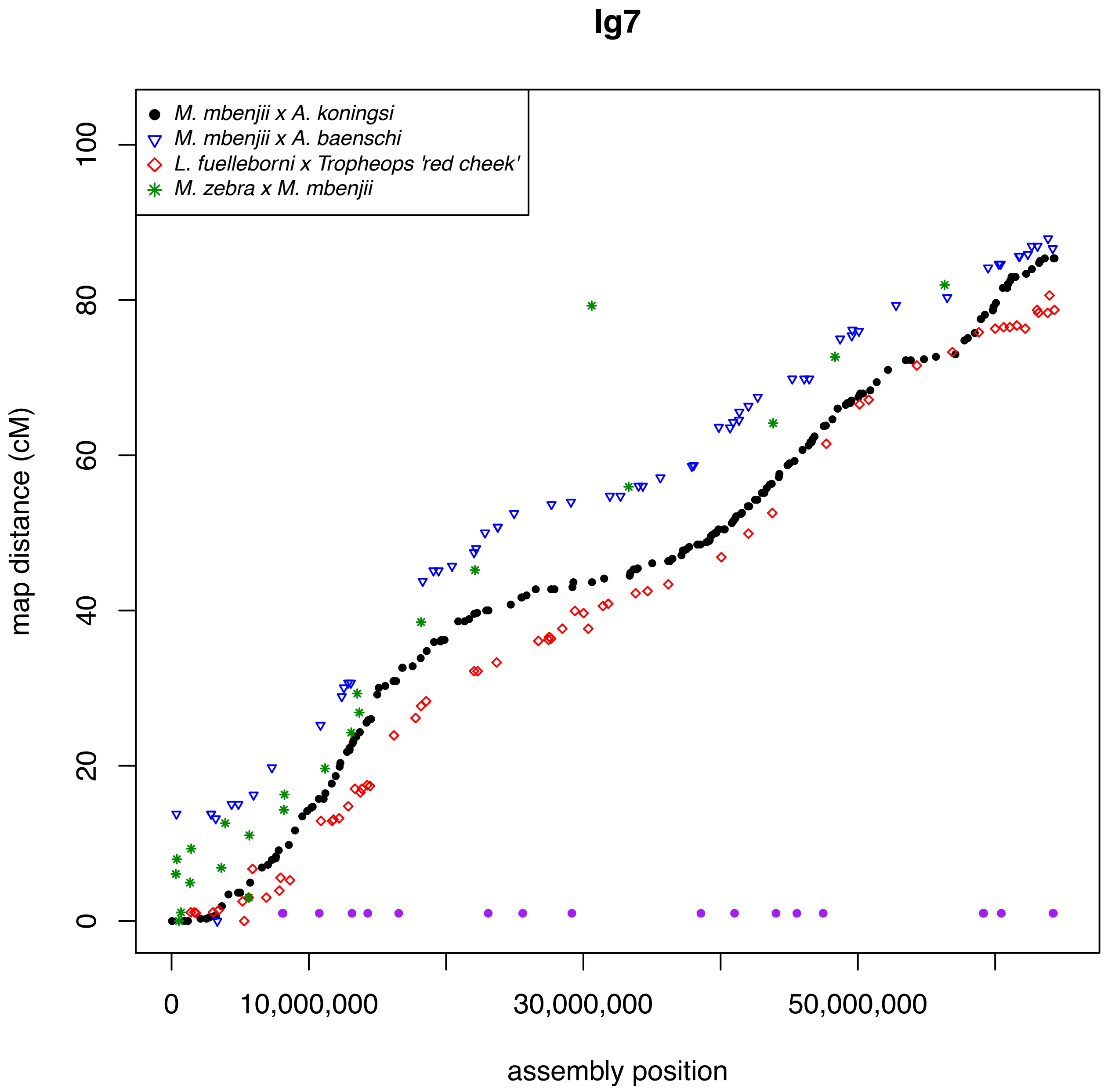

lg2
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
0
5,000,000
15,000,000
25,000,000
assembly position

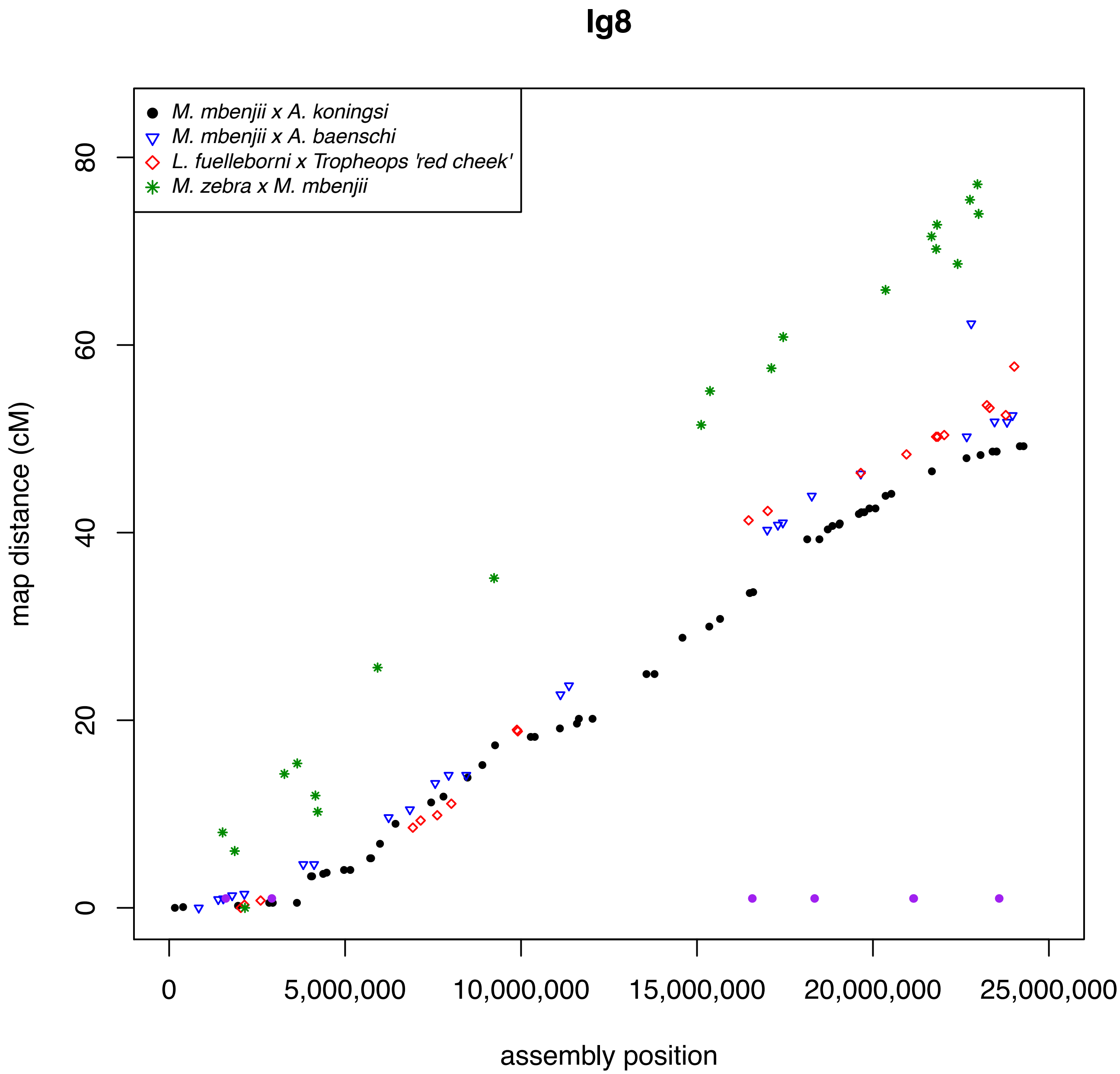

lg3
M. mbenjii x A. koningsi
M. mbenjii x A. baenschi
L. fuelleborni x Tropheops 'red cheek'
M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
assembly position
0
10,000,000
20,000,000
30,000,000
40,000,000

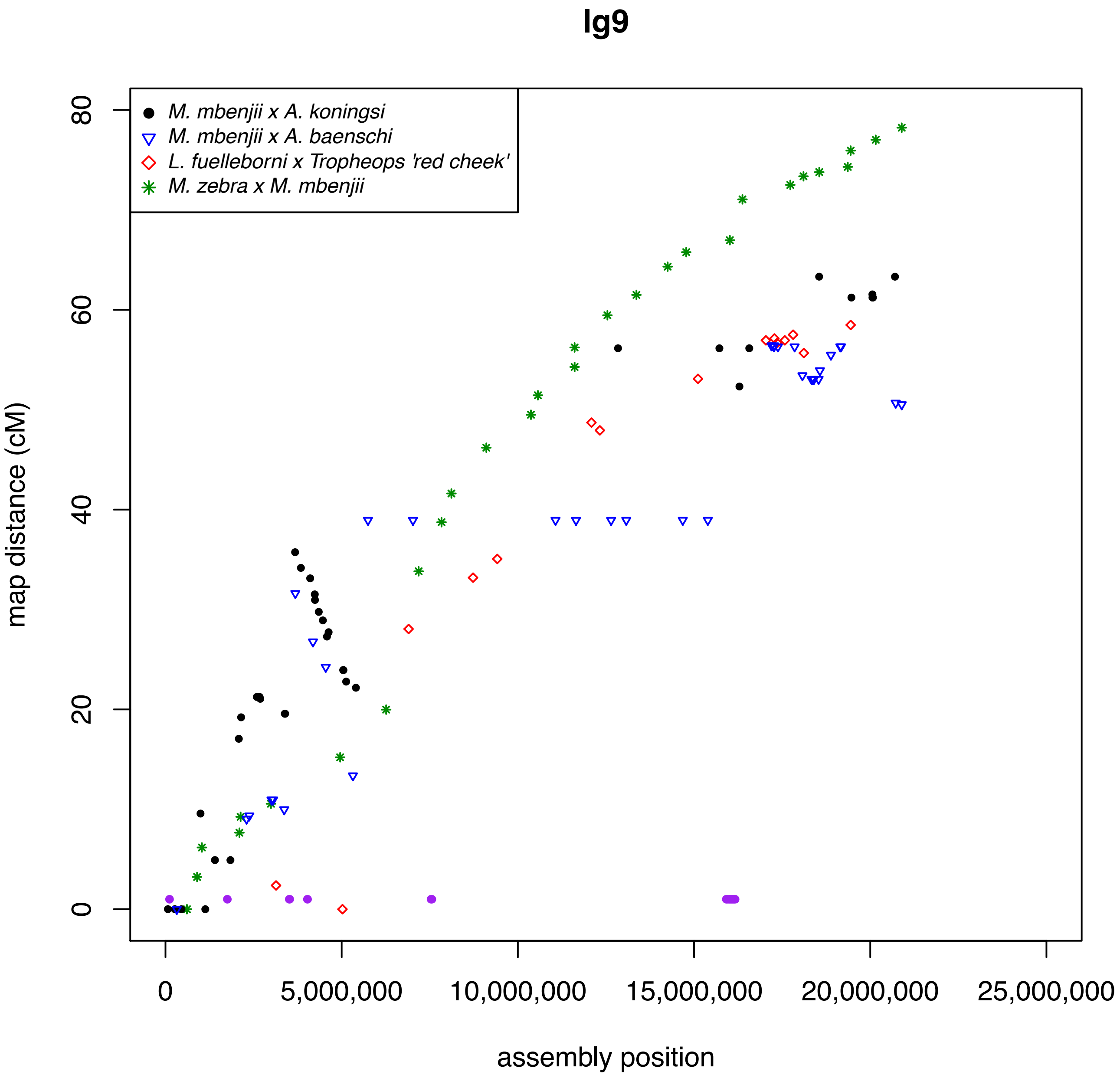

lg4
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
100
0
5,000,000
15,000,000
25,000,000
assembly position

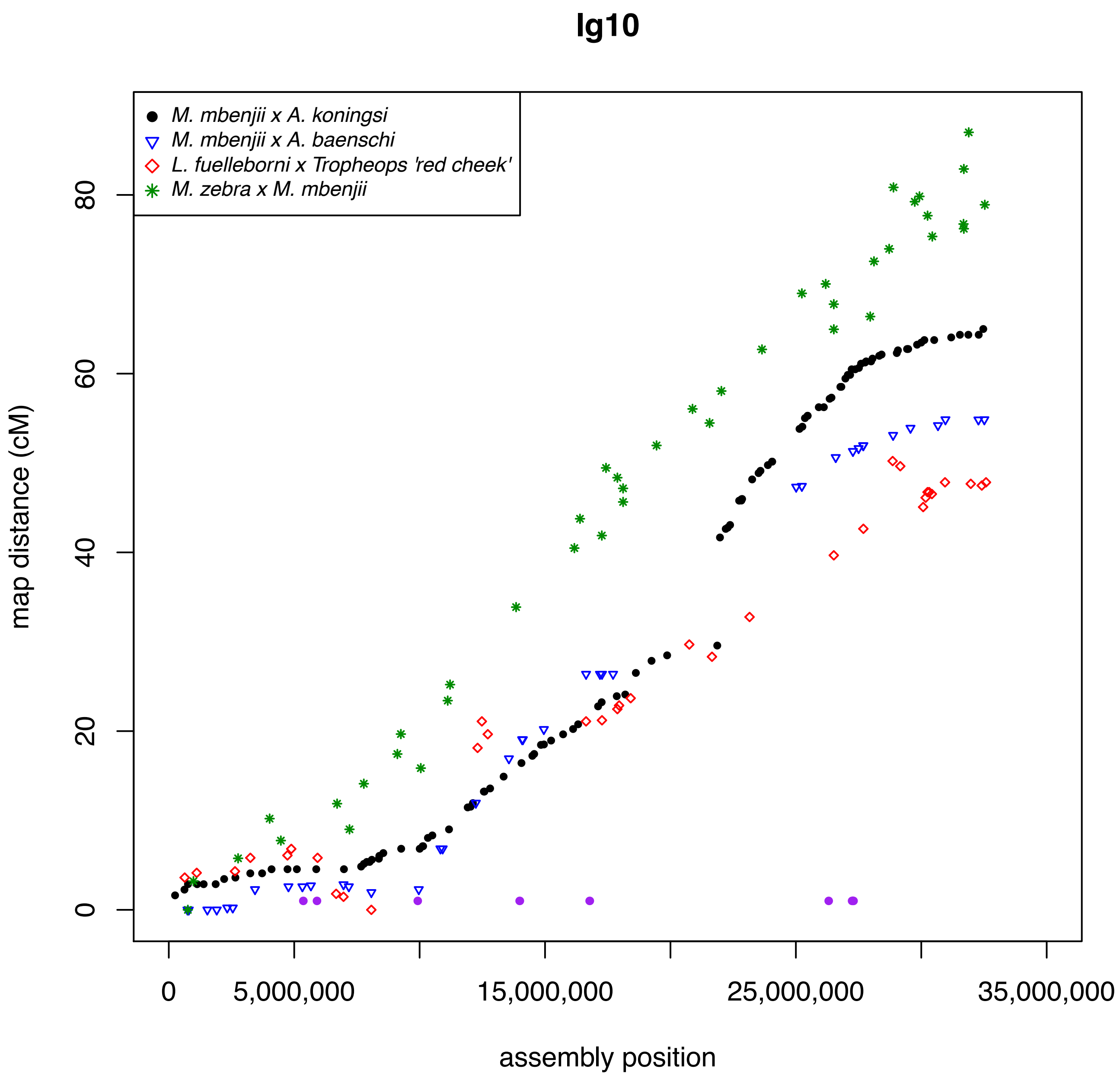

lg5
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
0
10,000,000
20,000,000
30,000,000
40,000,000
assembly position

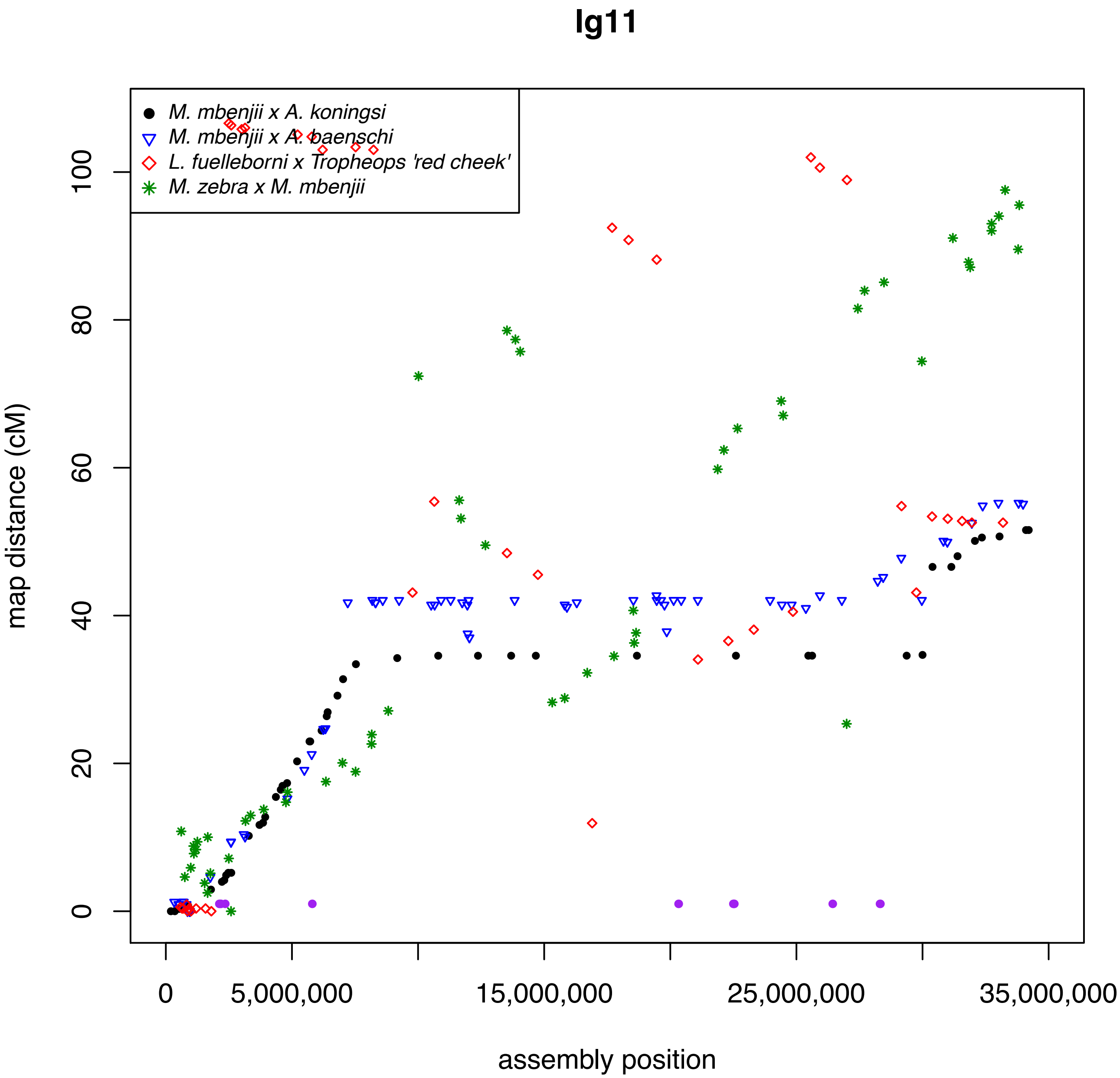

lg6
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
0
10,000,000
20,000,000
30,000,000
40,000,000
assembly position

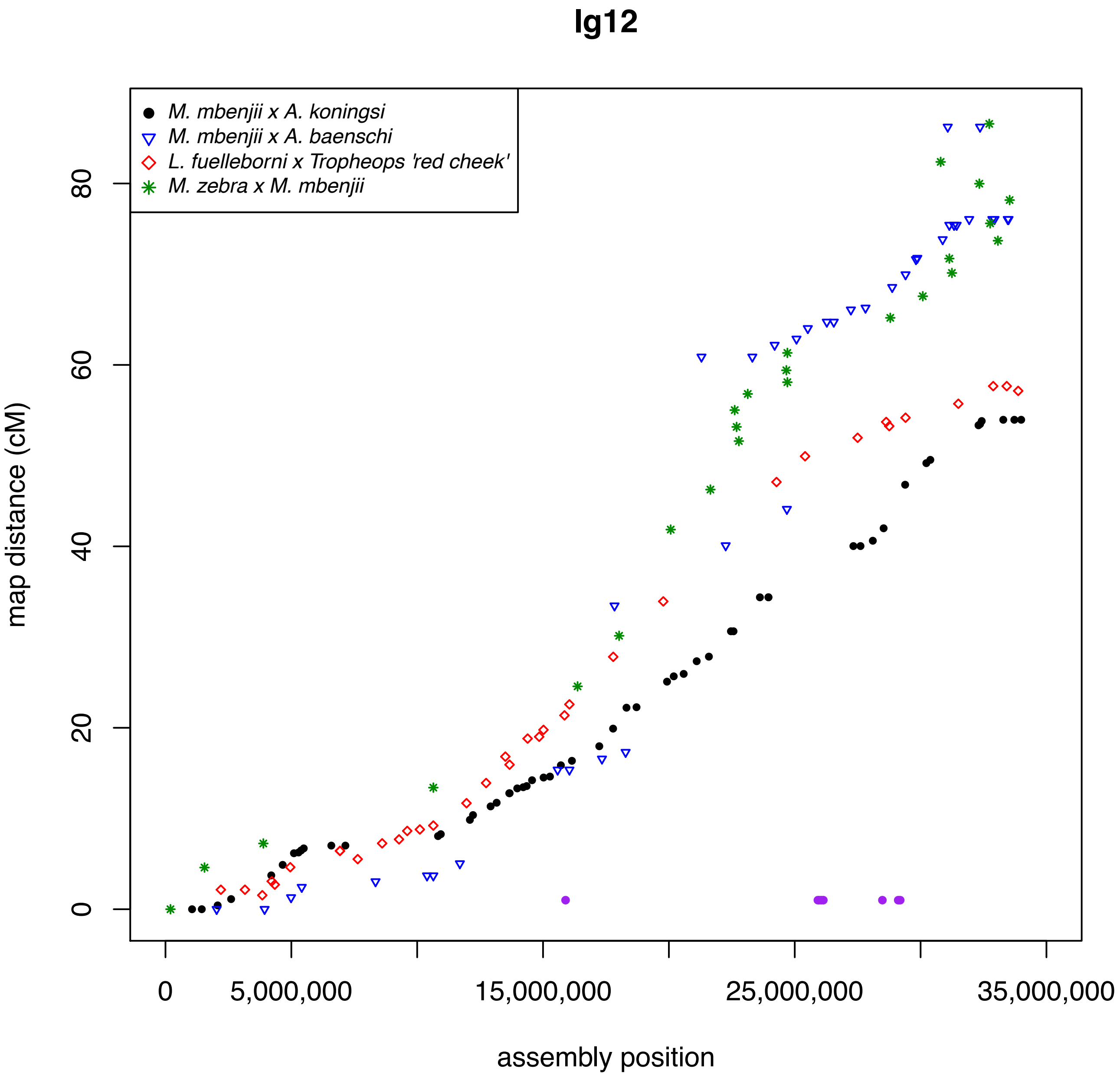

lg7
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
100
0
10,000,000
30,000,000
50,000,000
assembly position


lg8
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
assembly position
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000

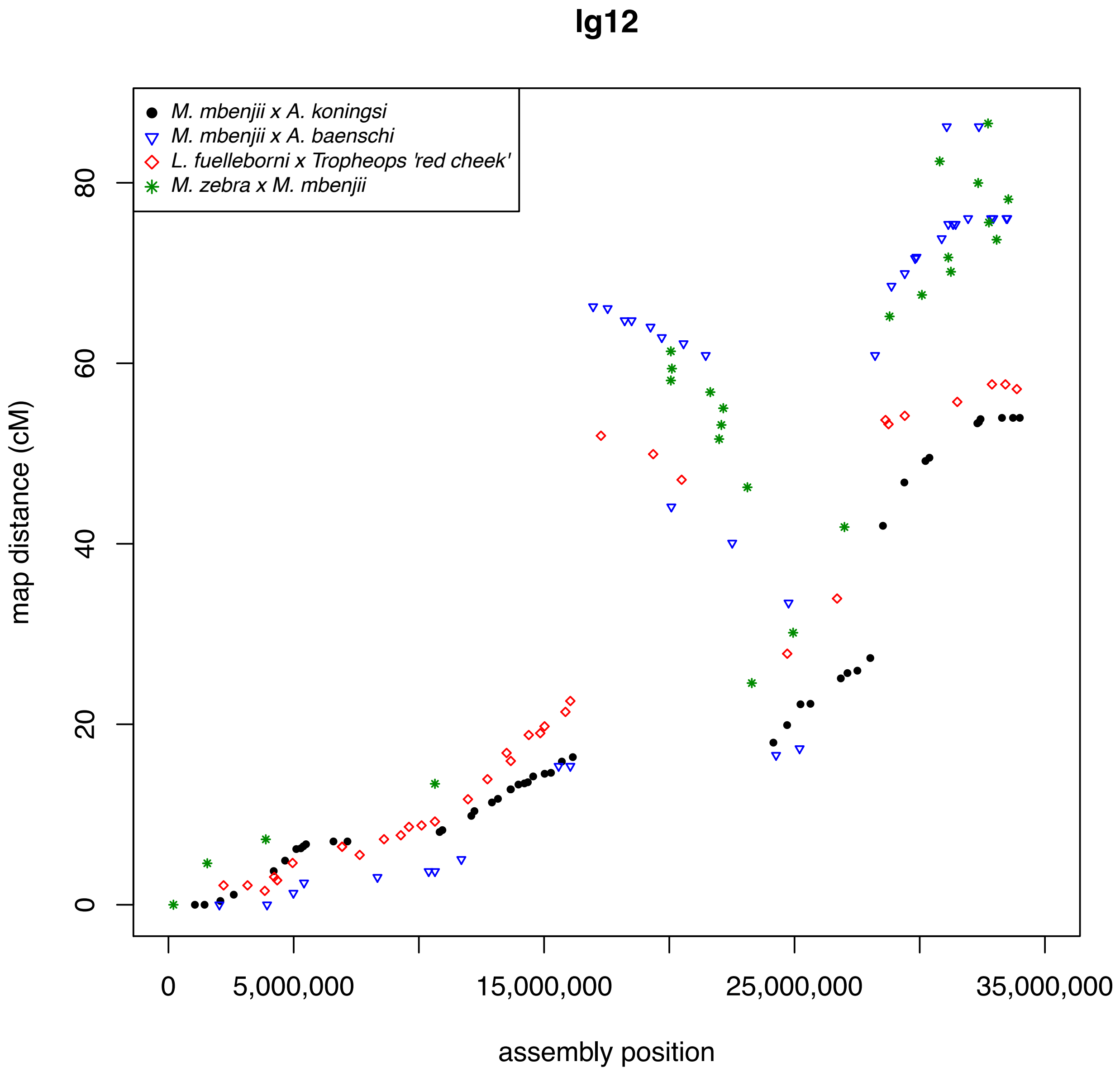

lg9
map distance (cM)
0
20
40
60
80
assembly position
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii

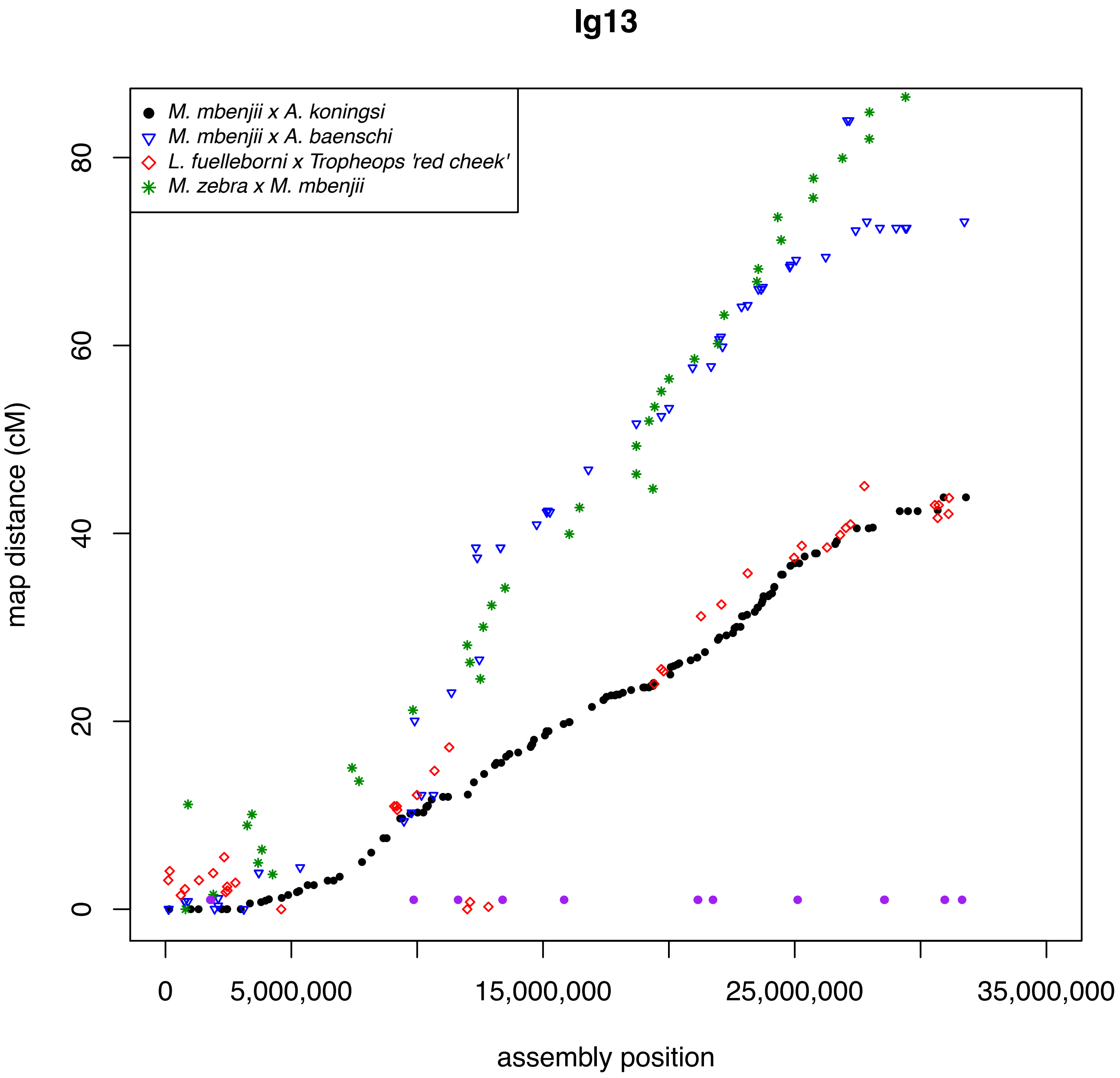

lg10
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
assembly position
0
5,000,000
15,000,000
25,000,000
35,000,000

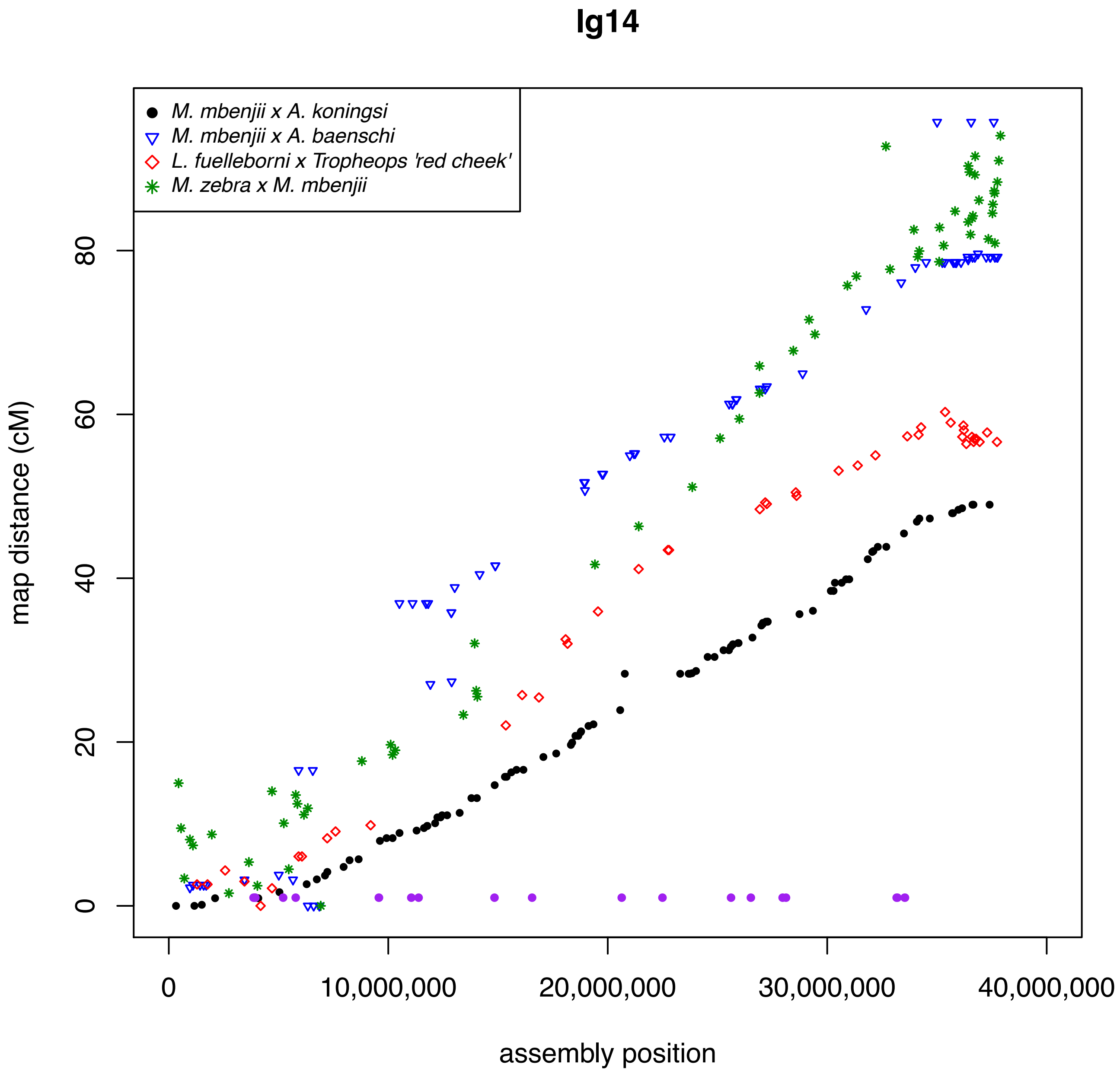

lg11
M. mbenjii x A. koningsi
M. mbenjii x A. baenschi
L. fuelleborni x Tropheops 'red cheek'
M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
100
assembly position
0
5,000,000
15,000,000
25,000,000
35,000,000

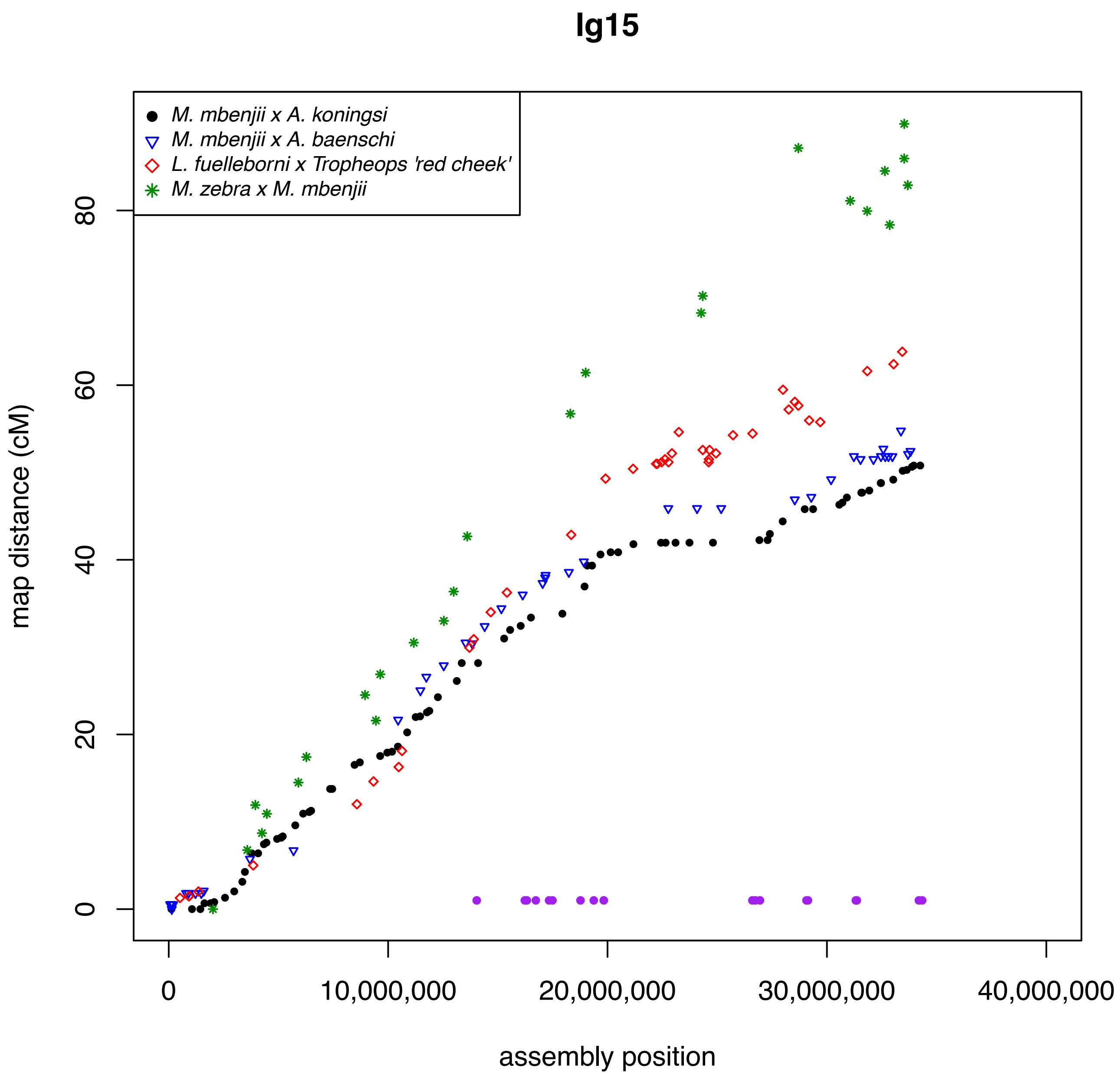

lg12
M. mbenjii x A. koningsi
M. mbenjii x A. baenschi
L. fuelleborni x Tropheops 'red cheek'
M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
assembly position
0
5,000,000
15,000,000
25,000,000
35,000,000

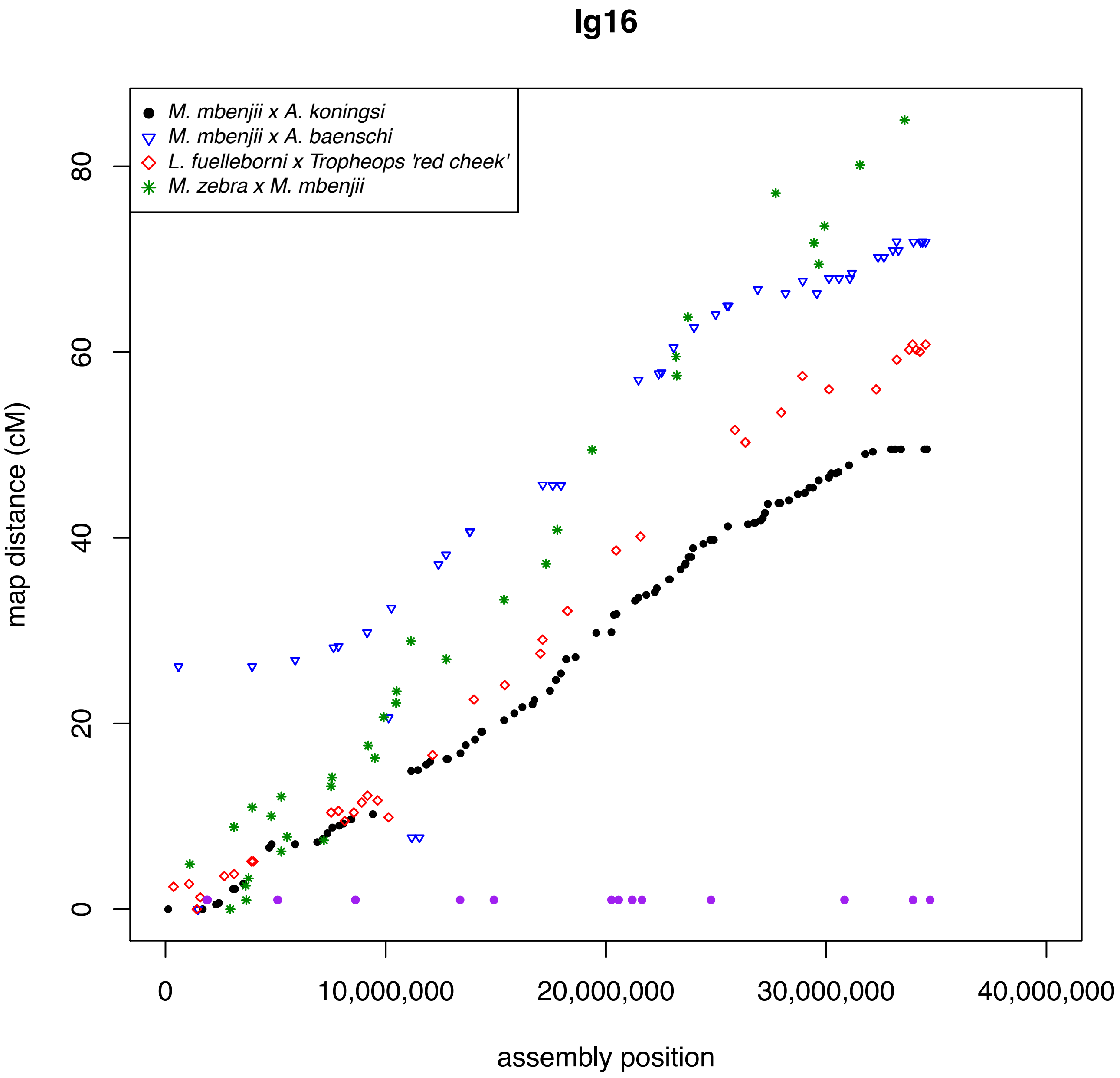
LG12 misassembly

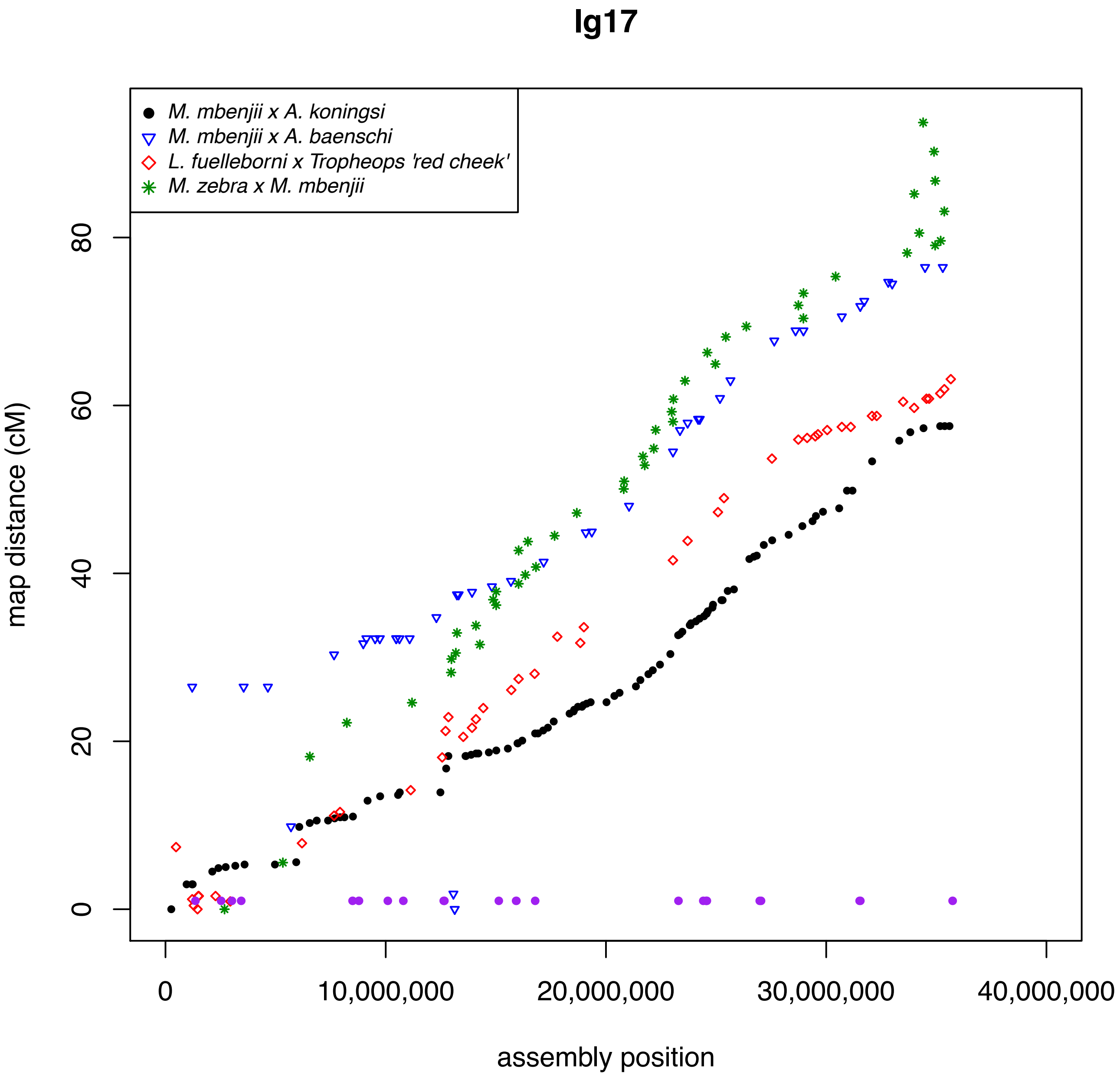


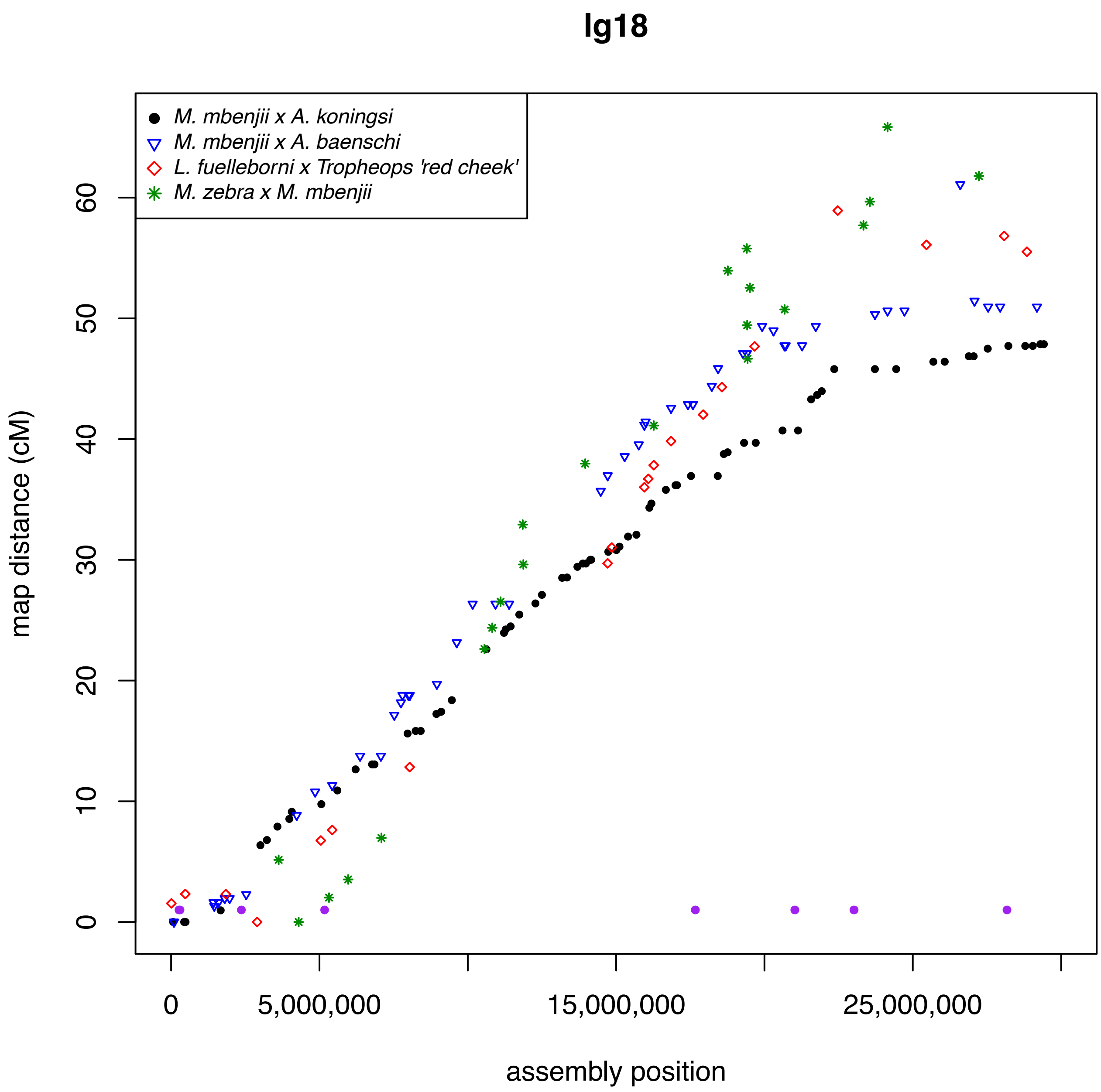

lg13
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
assembly position
0
5,000,000
15,000,000
25,000,000
35,000,000

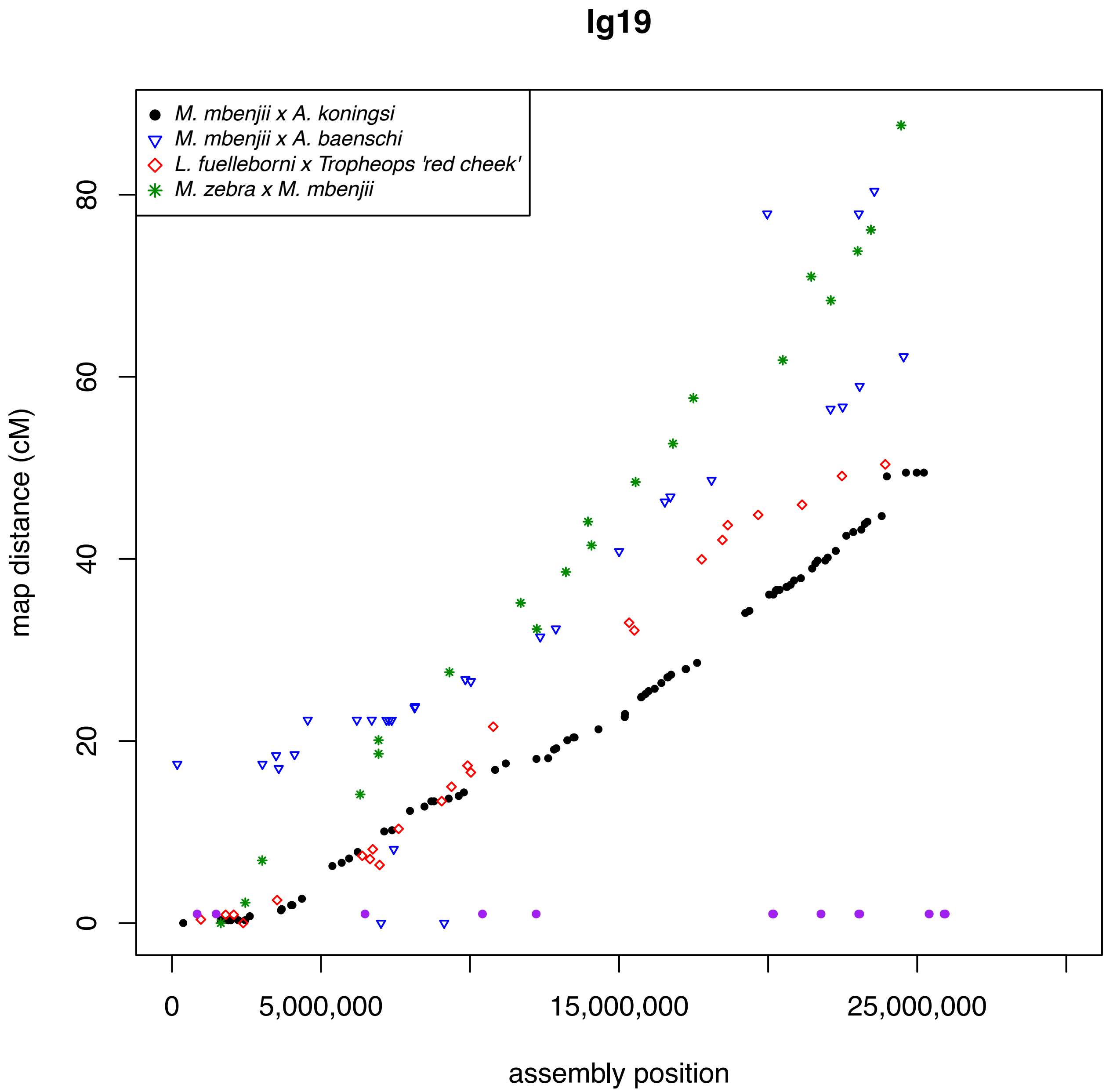

lg14
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
0
10,000,000
20,000,000
30,000,000
40,000,000
assembly position

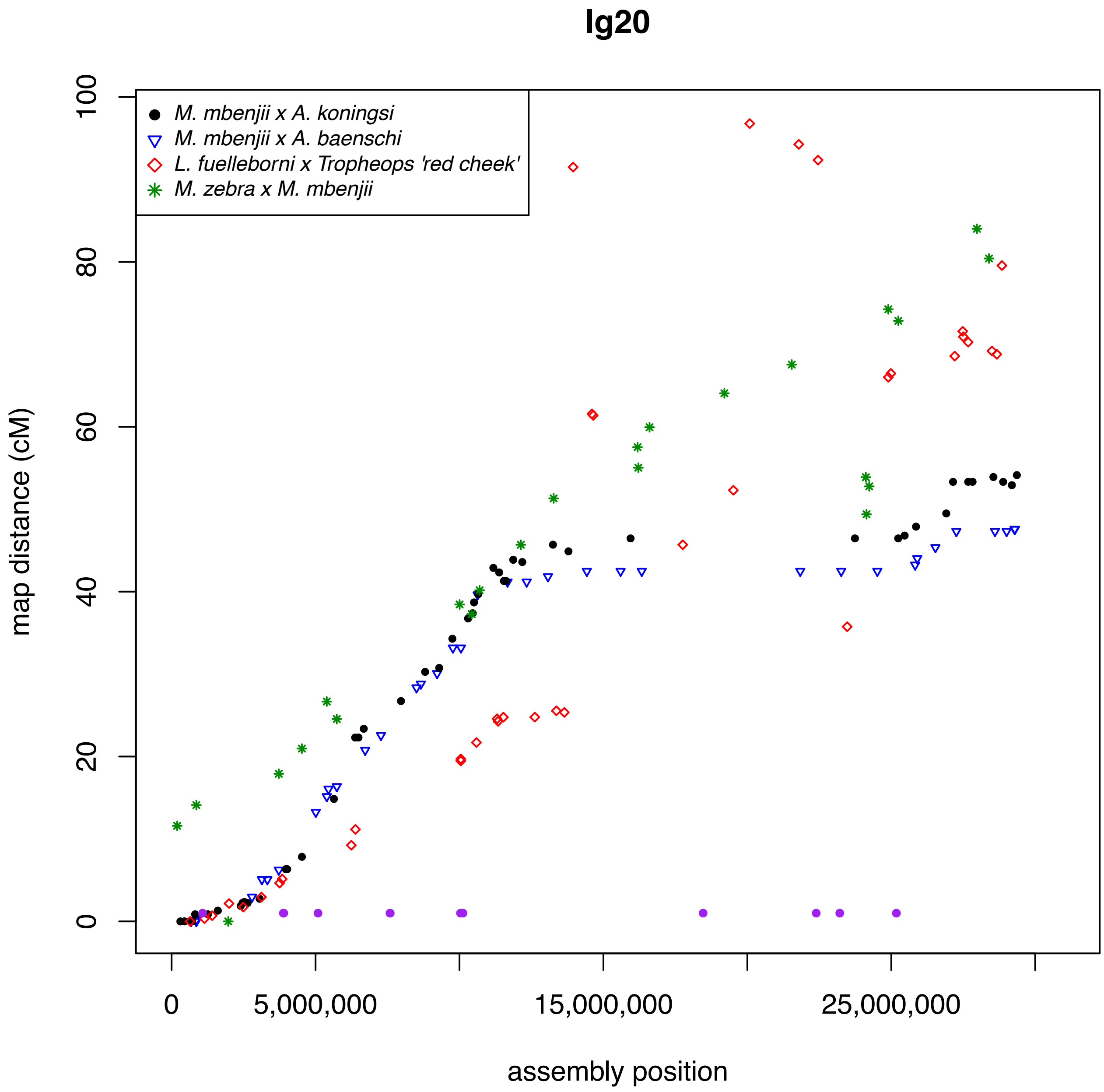

lg15
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
0
10,000,000
20,000,000
30,000,000
40,000,000
assembly position

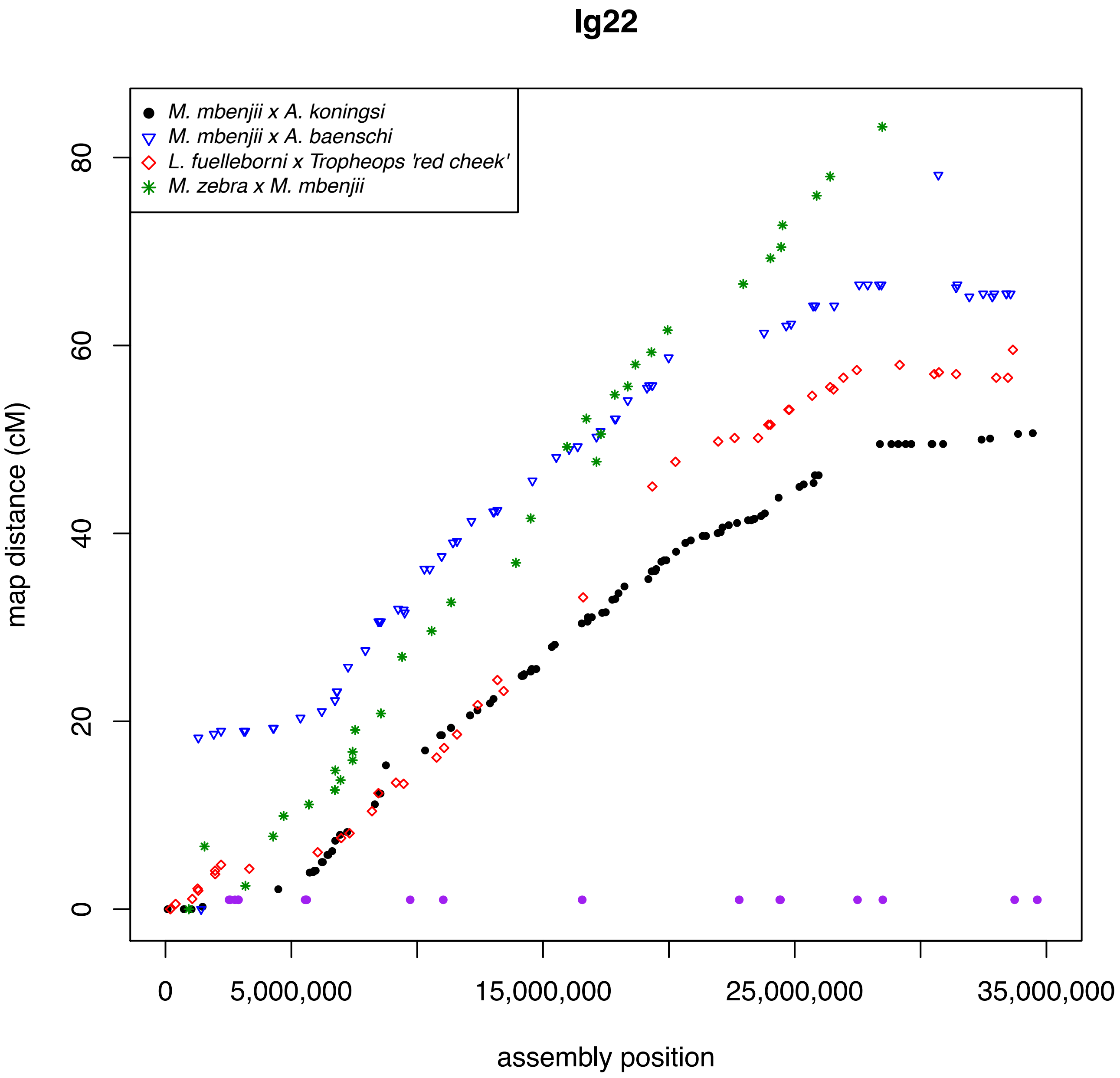

lg16
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
0
10,000,000
20,000,000
30,000,000
40,000,000
assembly position

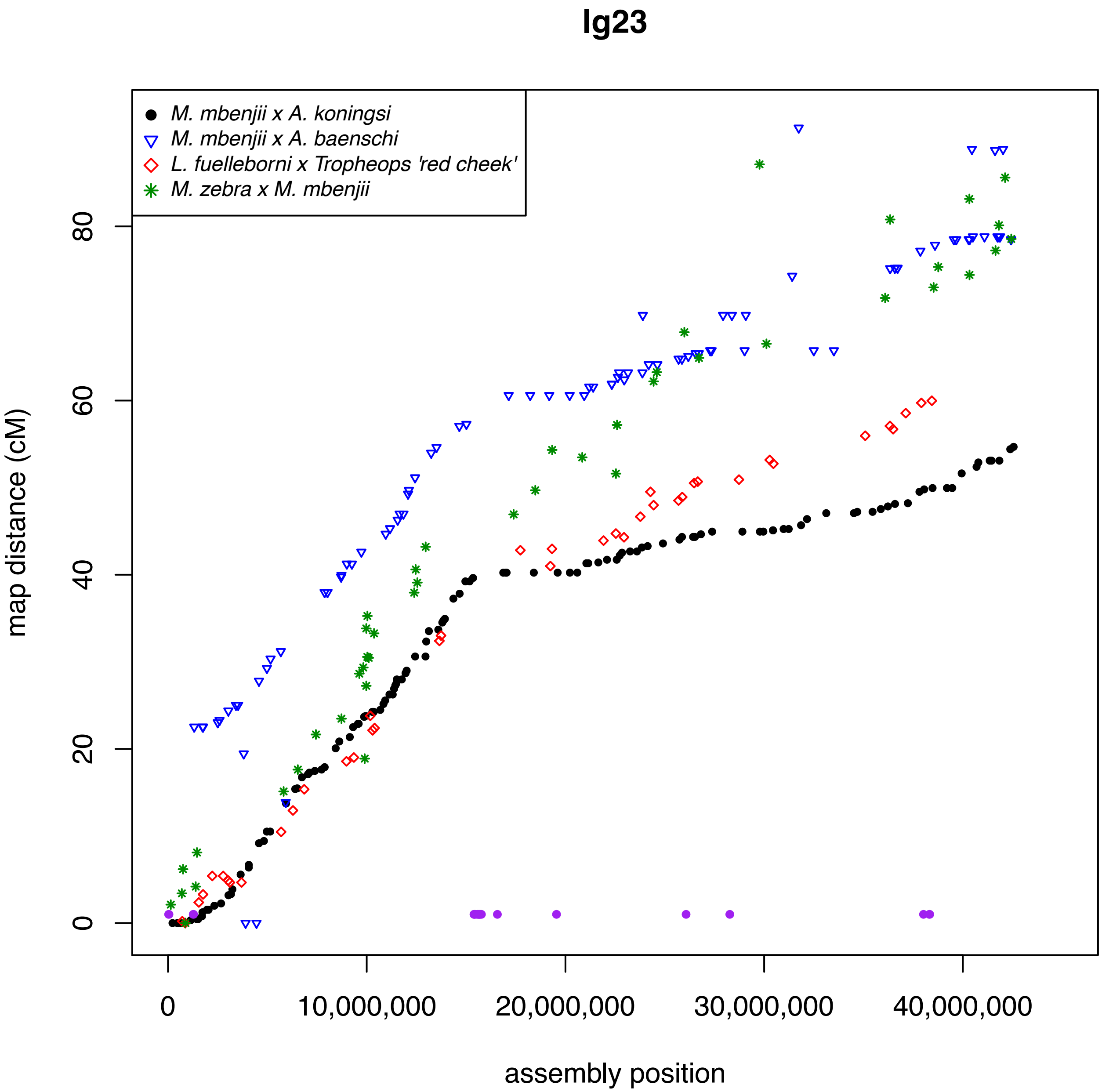

lg17
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
0
10,000,000
20,000,000
30,000,000
40,000,000
assembly position


lg18
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii
map distance (cM)
0
10
20
30
40
50
60
assembly position
0
5,000,000
15,000,000
25,000,000


lg19
M. mbenjii x A. koningsi
M. mbenjii x A. baenschi
L. fuelleborni x Tropheops 'red cheek'
M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
0
5,000,000
15,000,000
25,000,000
assembly position

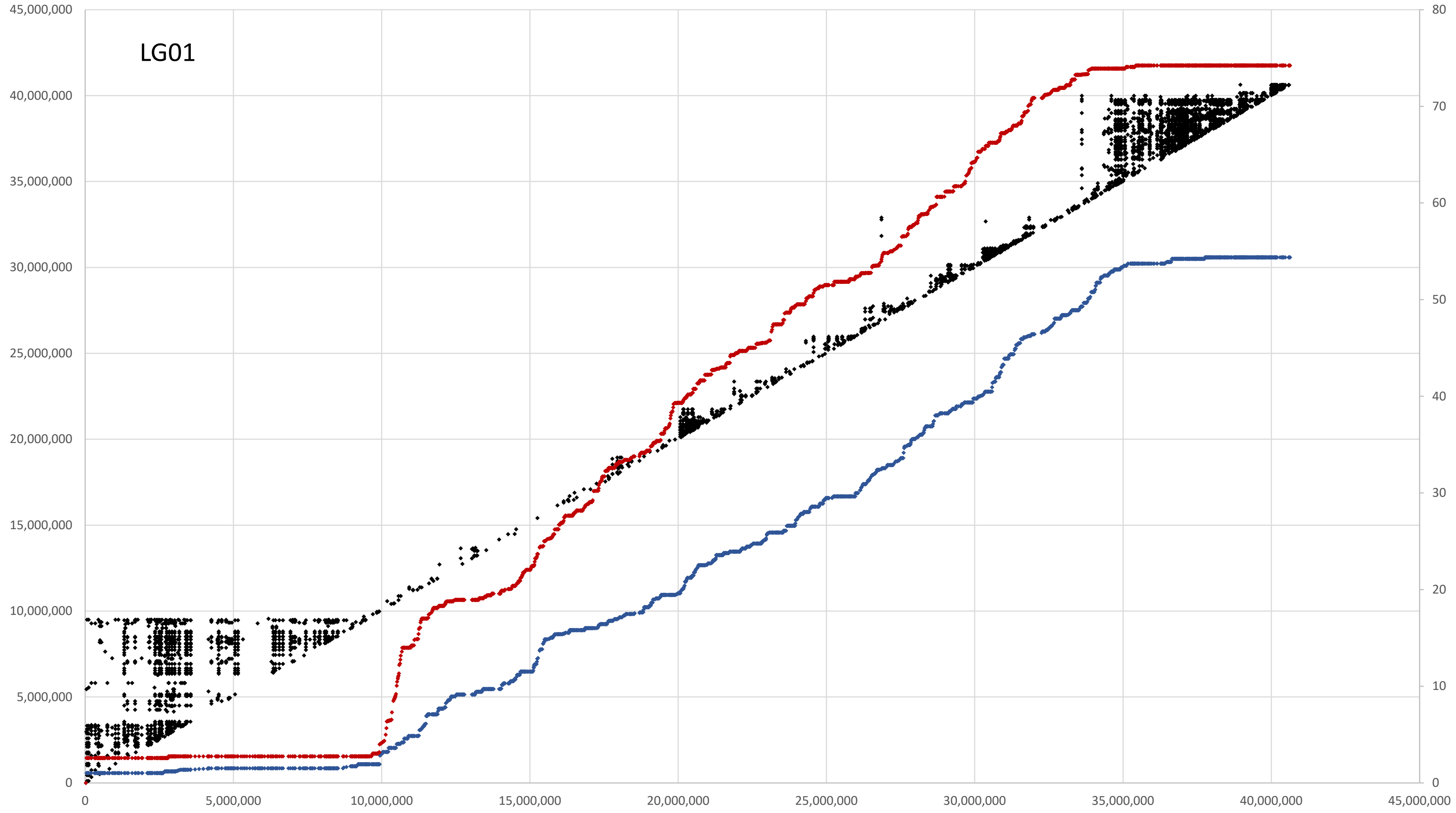

lg20
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
100
0
5,000,000
15,000,000
25,000,000
assembly position

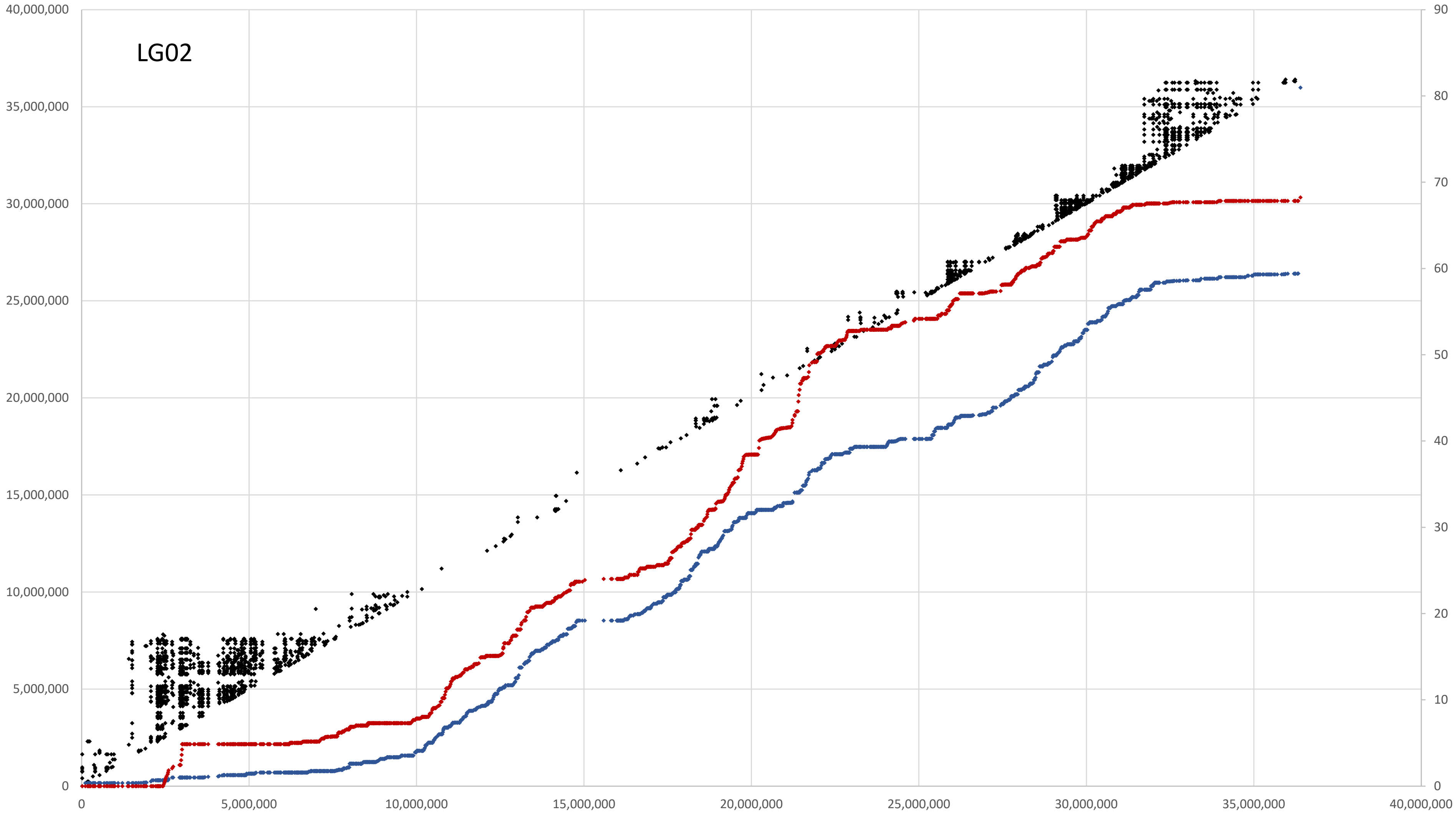

lg22
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
assembly position
0
5,000,000
15,000,000
25,000,000
35,000,000

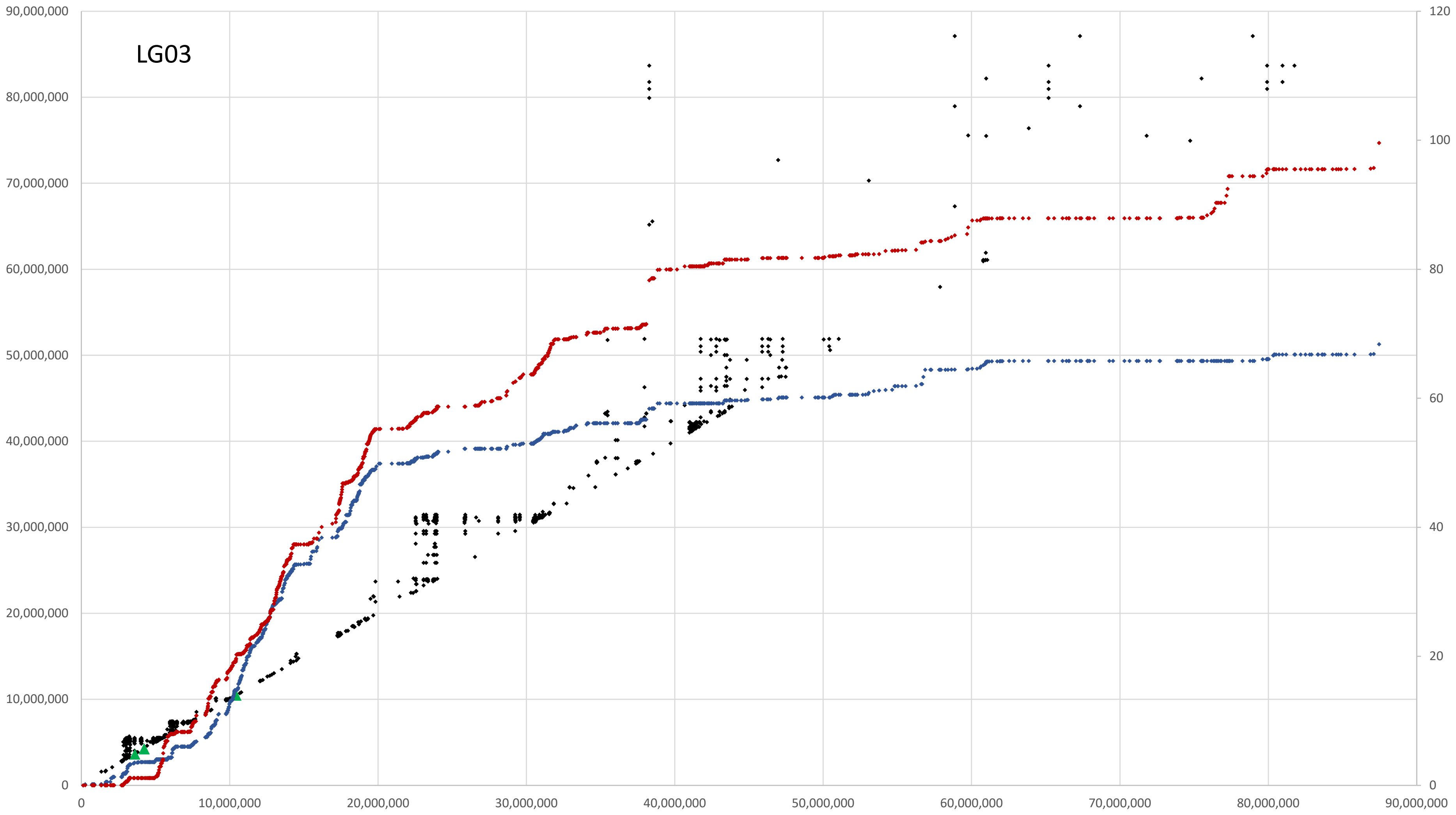

lg23
● M. mbenjii x A. koningsi
▽ M. mbenjii x A. baenschi
◇ L. fuelleborni x Tropheops 'red cheek'
✳ M. zebra x M. mbenjii
map distance (cM)
0
20
40
60
80
0
10,000,000
20,000,000
30,000,000
40,000,000
assembly position

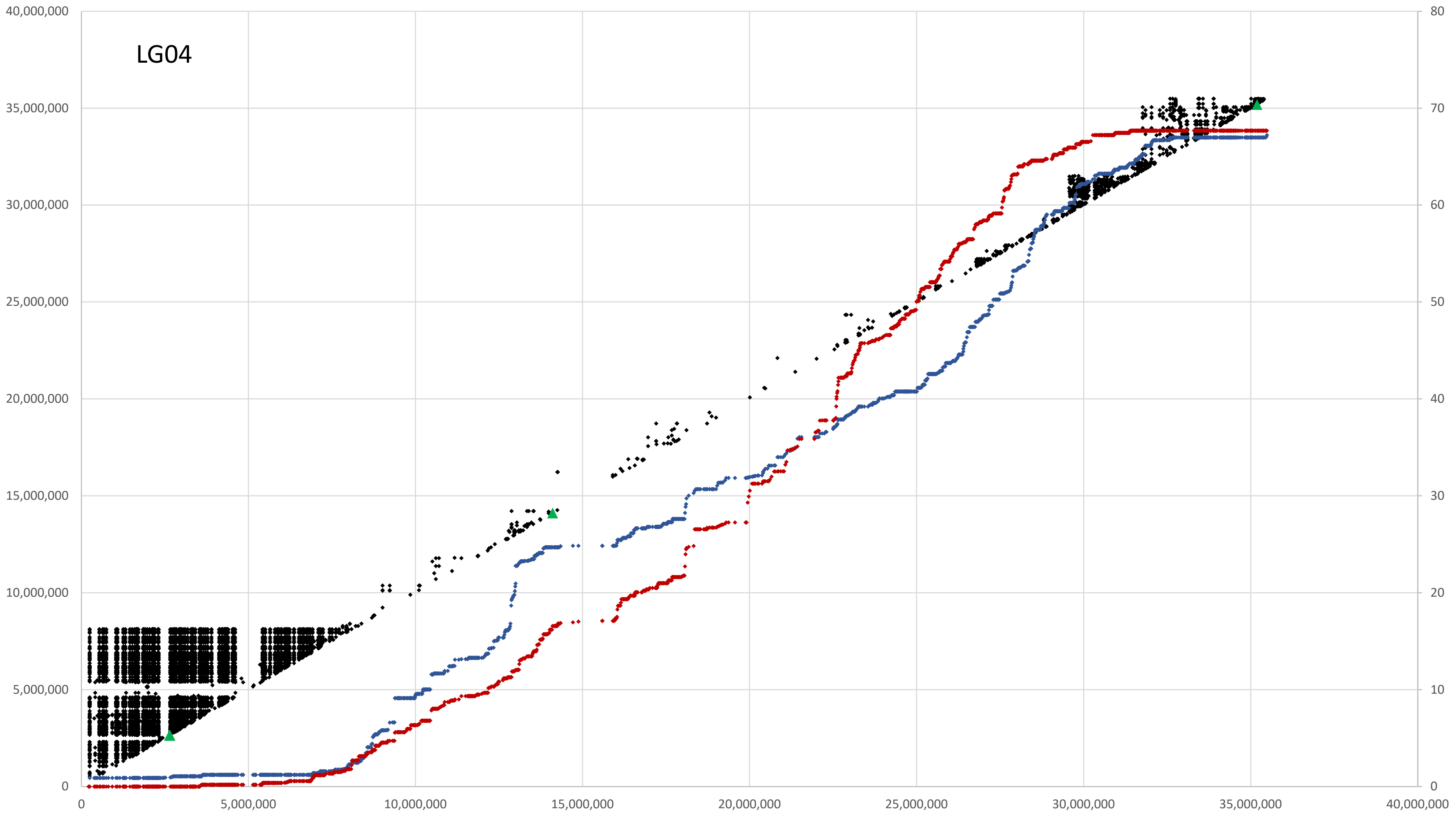
Appendix G

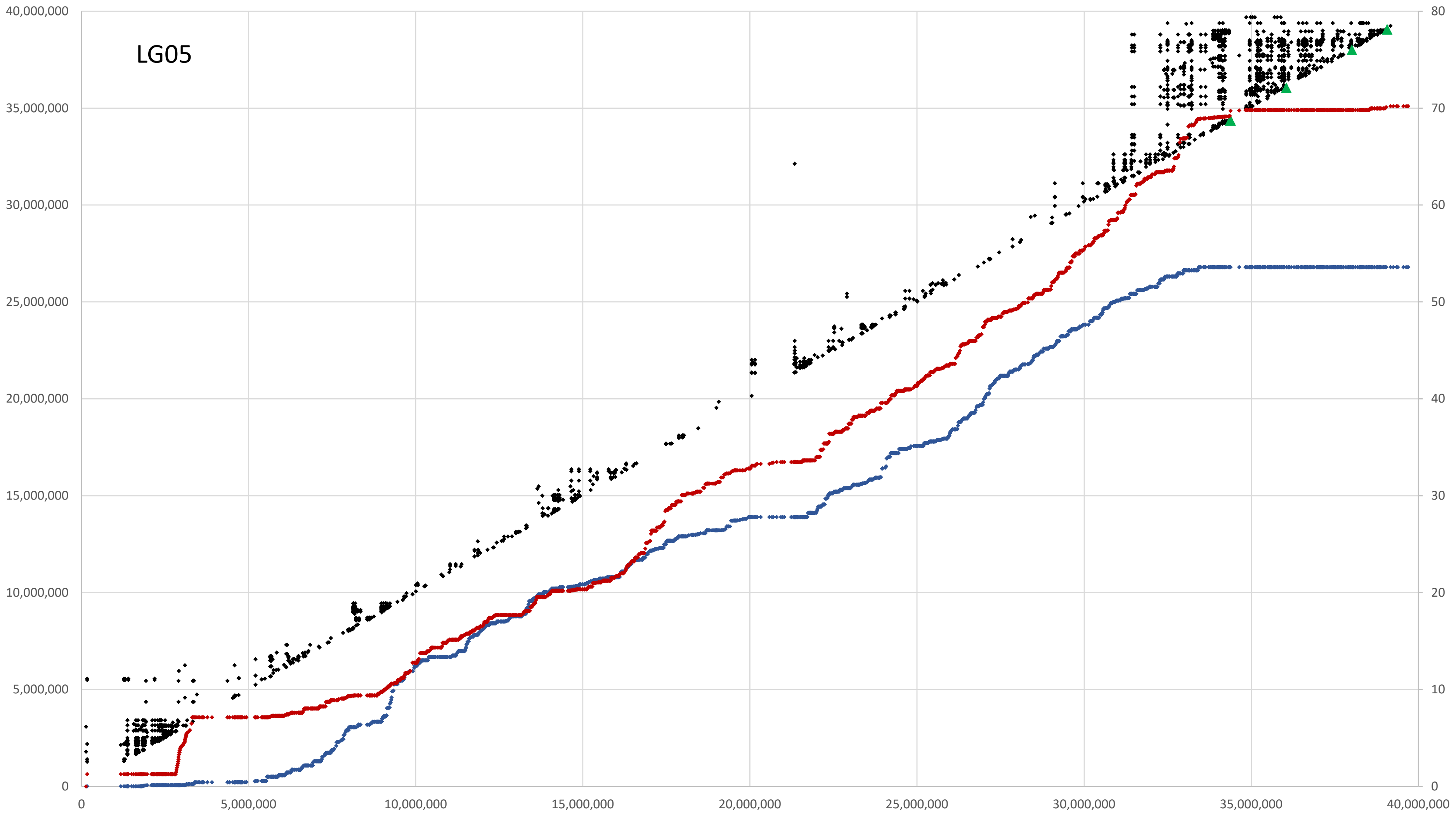
O. niloticus recombination curves for females (red) and males (blue). Centromere repeats are displayed as green triangles where applicable. X-axis represent the location along the anchored LG. Left Y-axis represents linkage disequilibrium and right Y-axis shows the map location for each marker.

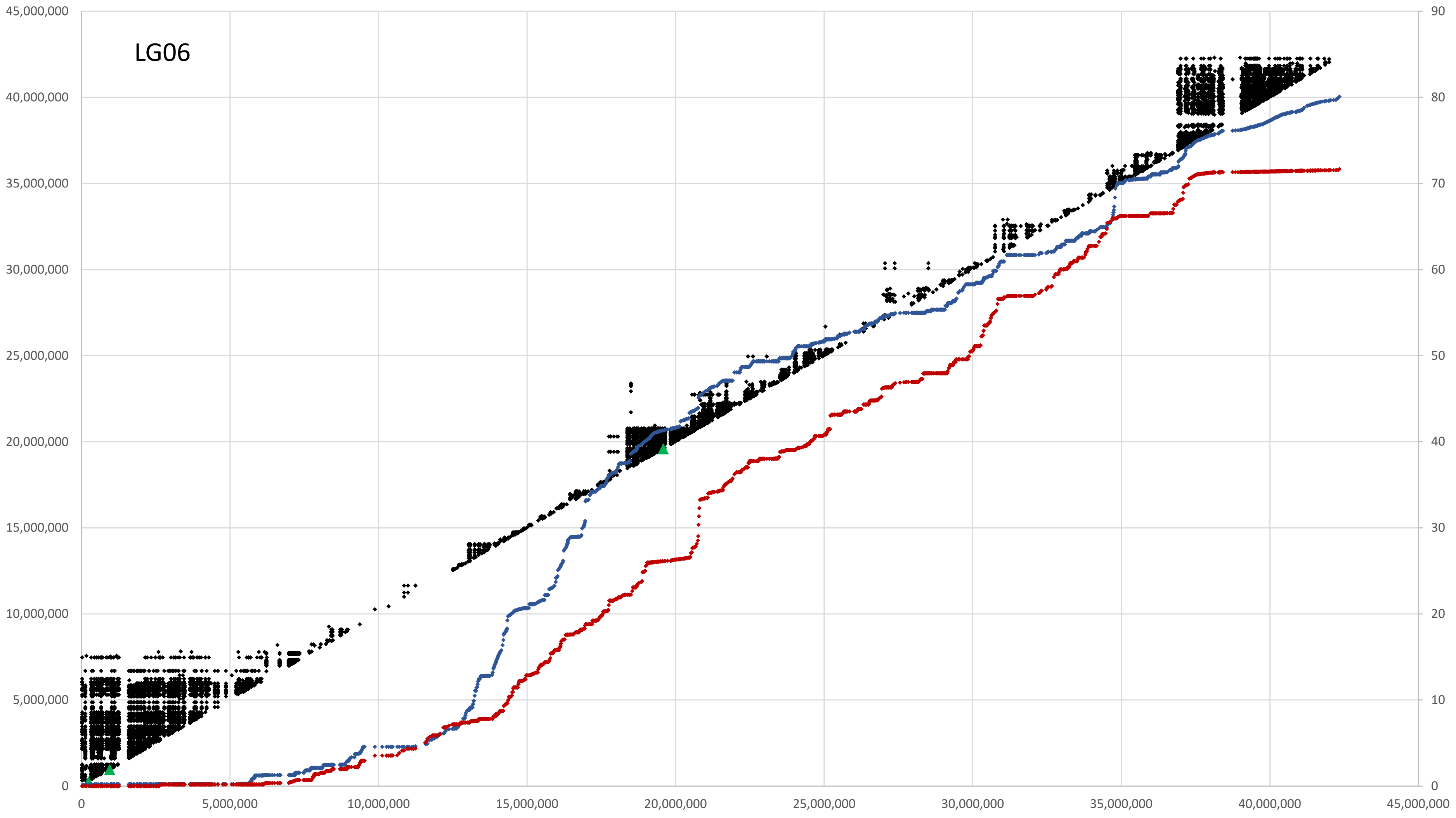



LG02
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
90
80
70
60
50
40
30
20
10
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000

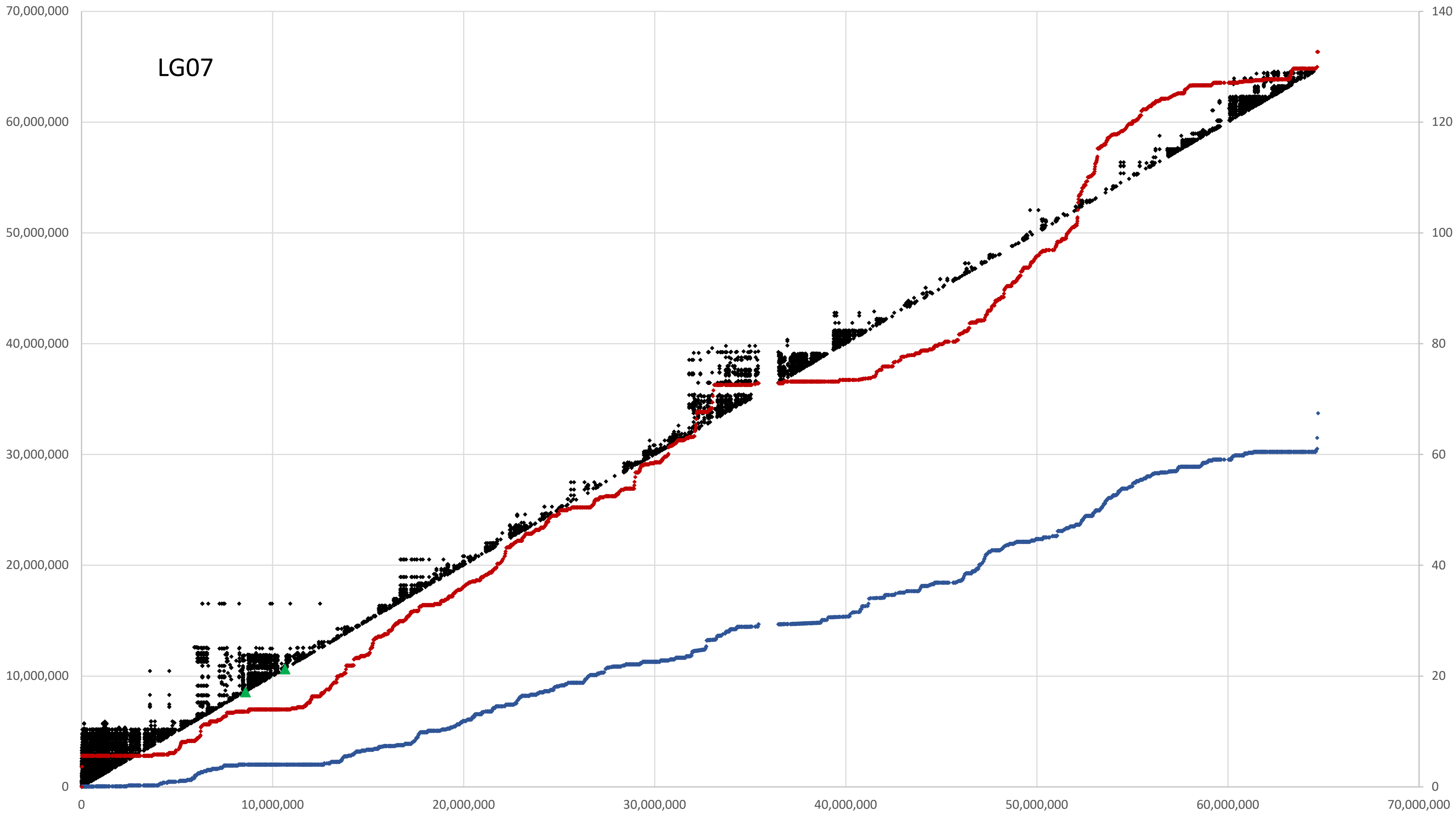

LG03
90,000,000
80,000,000
70,000,000
60,000,000
50,000,000
40,000,000
30,000,000
20,000,000
10,000,000
0
120
100
80
60
40
20
0
0
10,000,000
20,000,000
30,000,000
40,000,000
50,000,000
60,000,000
70,000,000
80,000,000
90,000,000

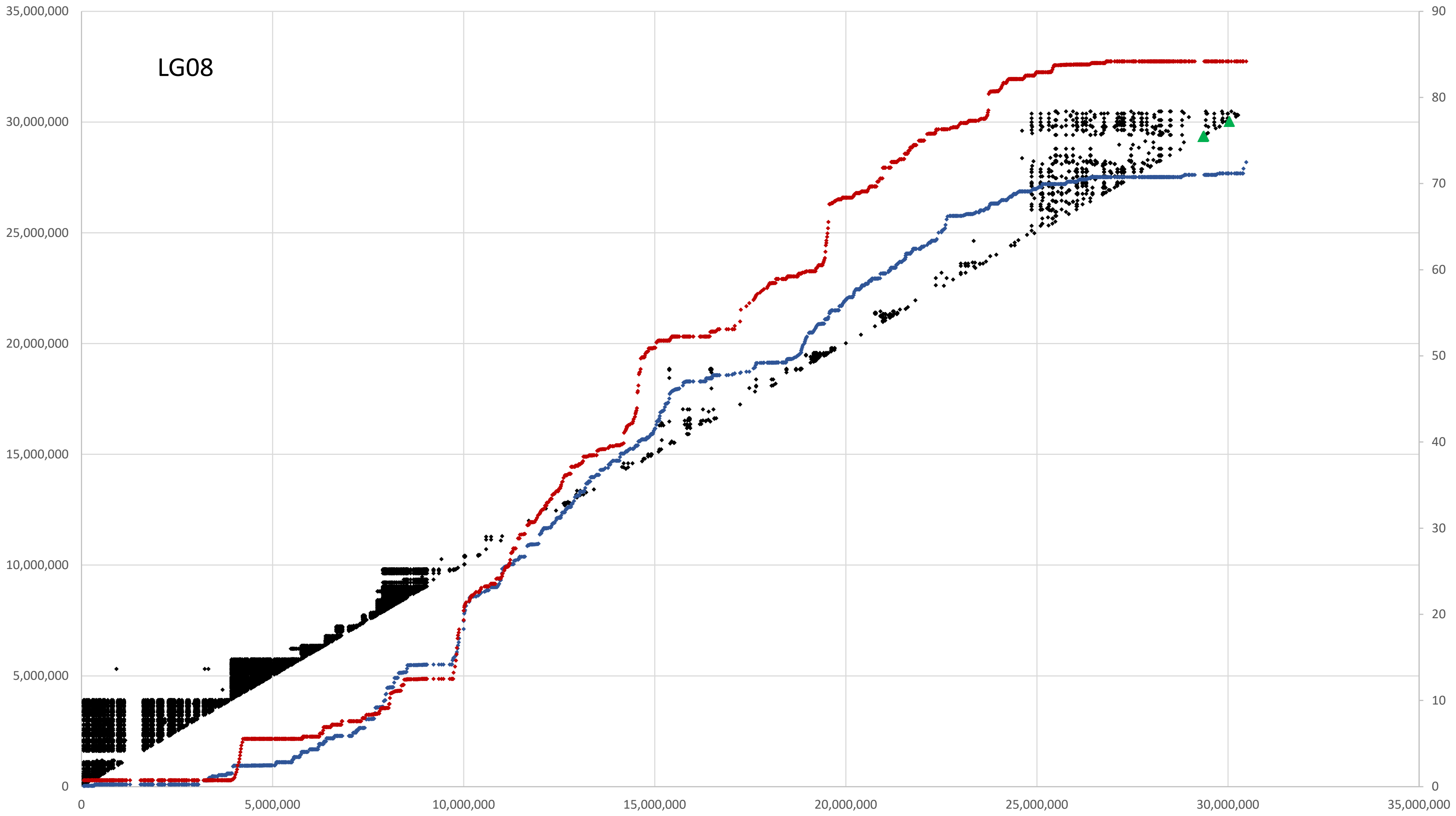

LG04
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000
80
70
60
50
40
30
20
10
0

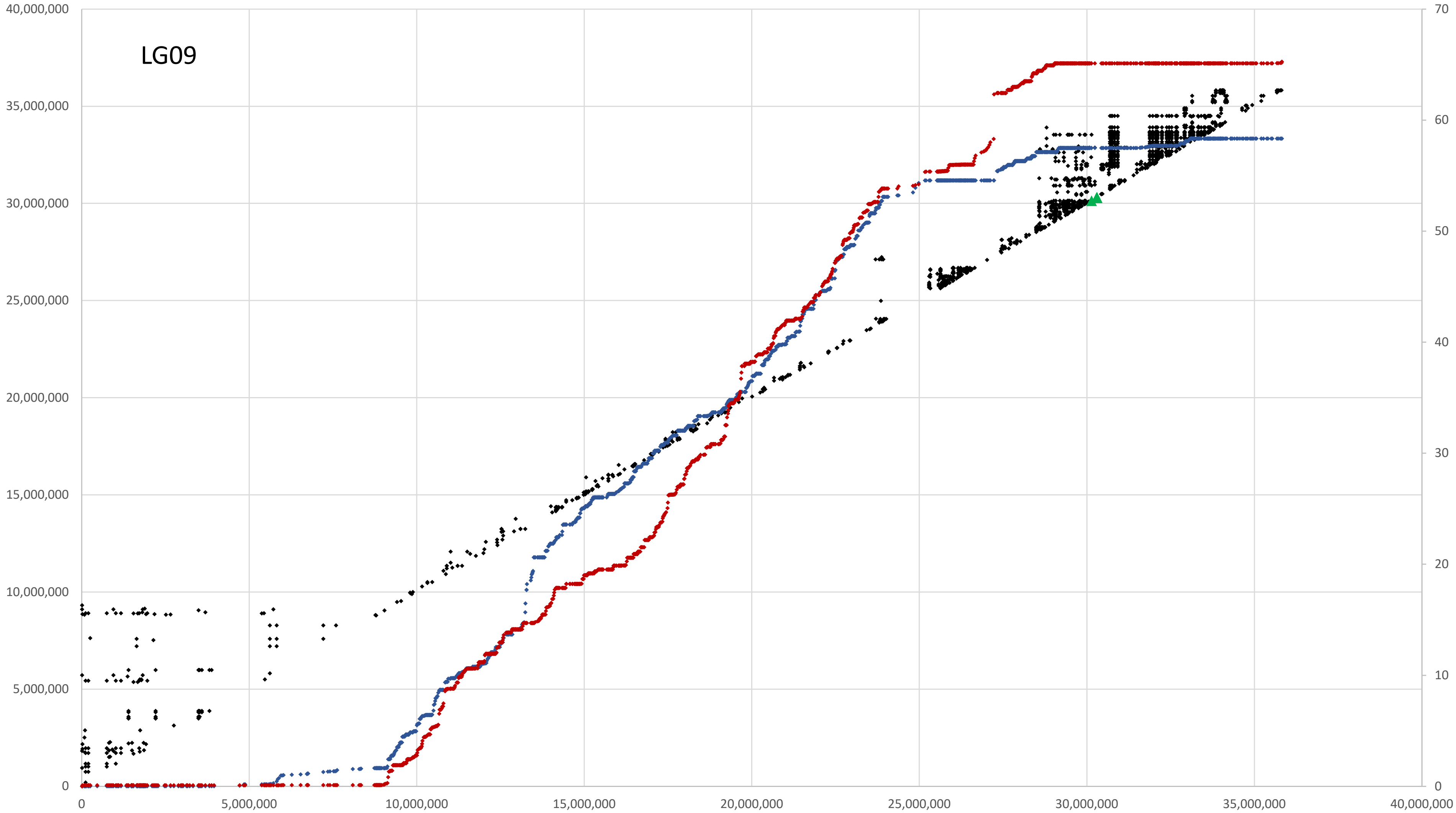

LG05
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000
80
70
60
50
40
30
20
10
0

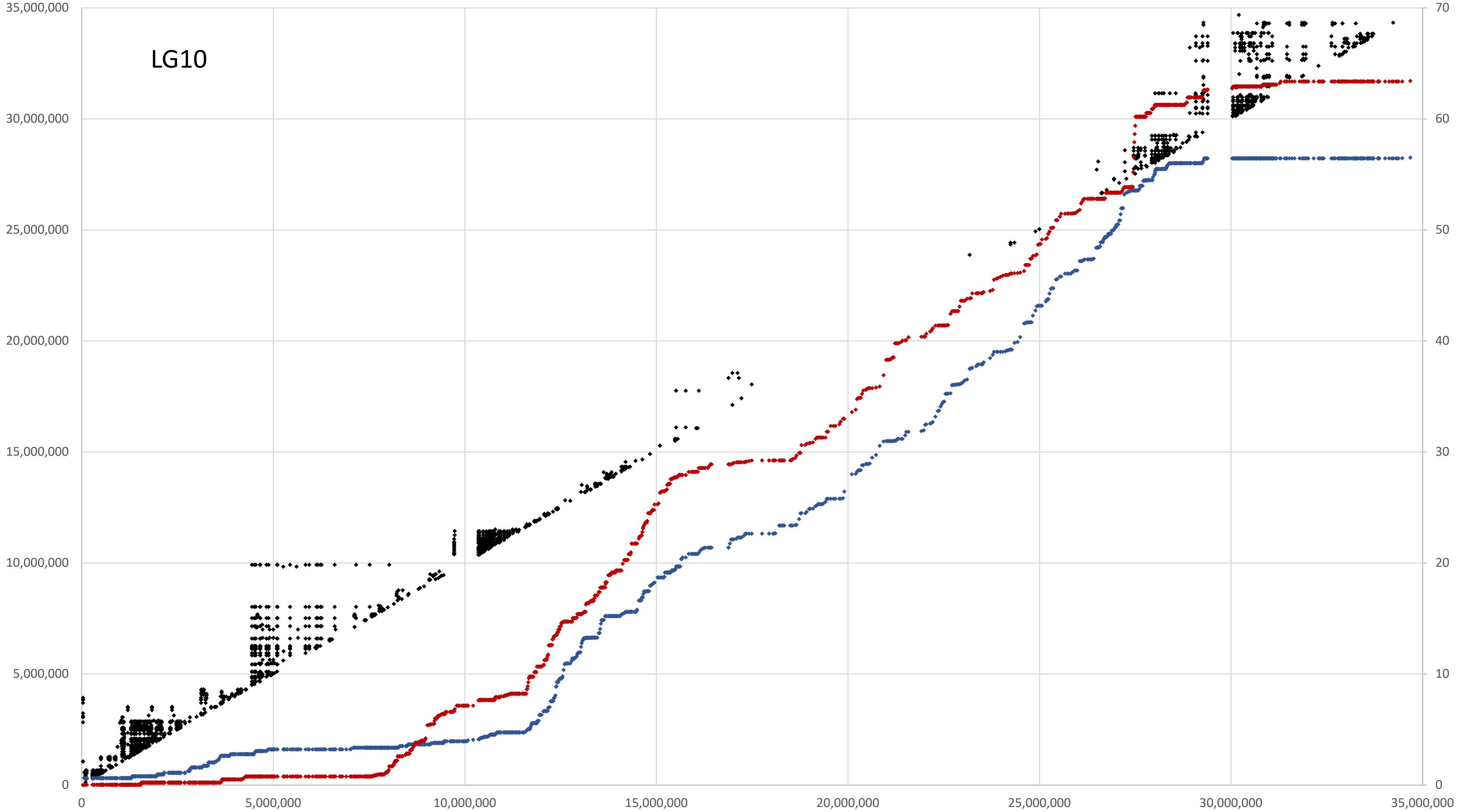


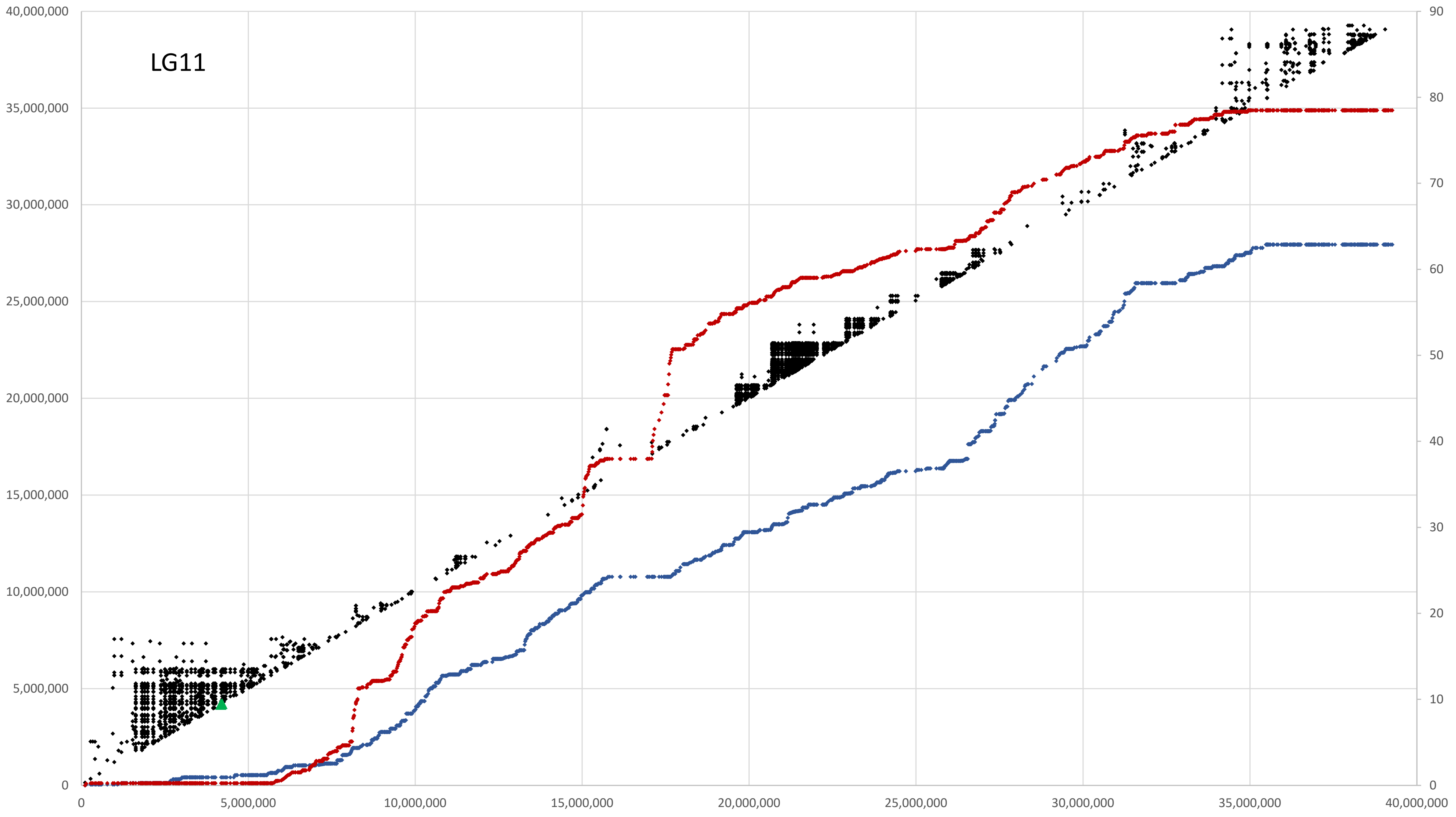
LG06
45,000,000
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
90
80
70
60
50
40
30
20
10
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000
45,000,000

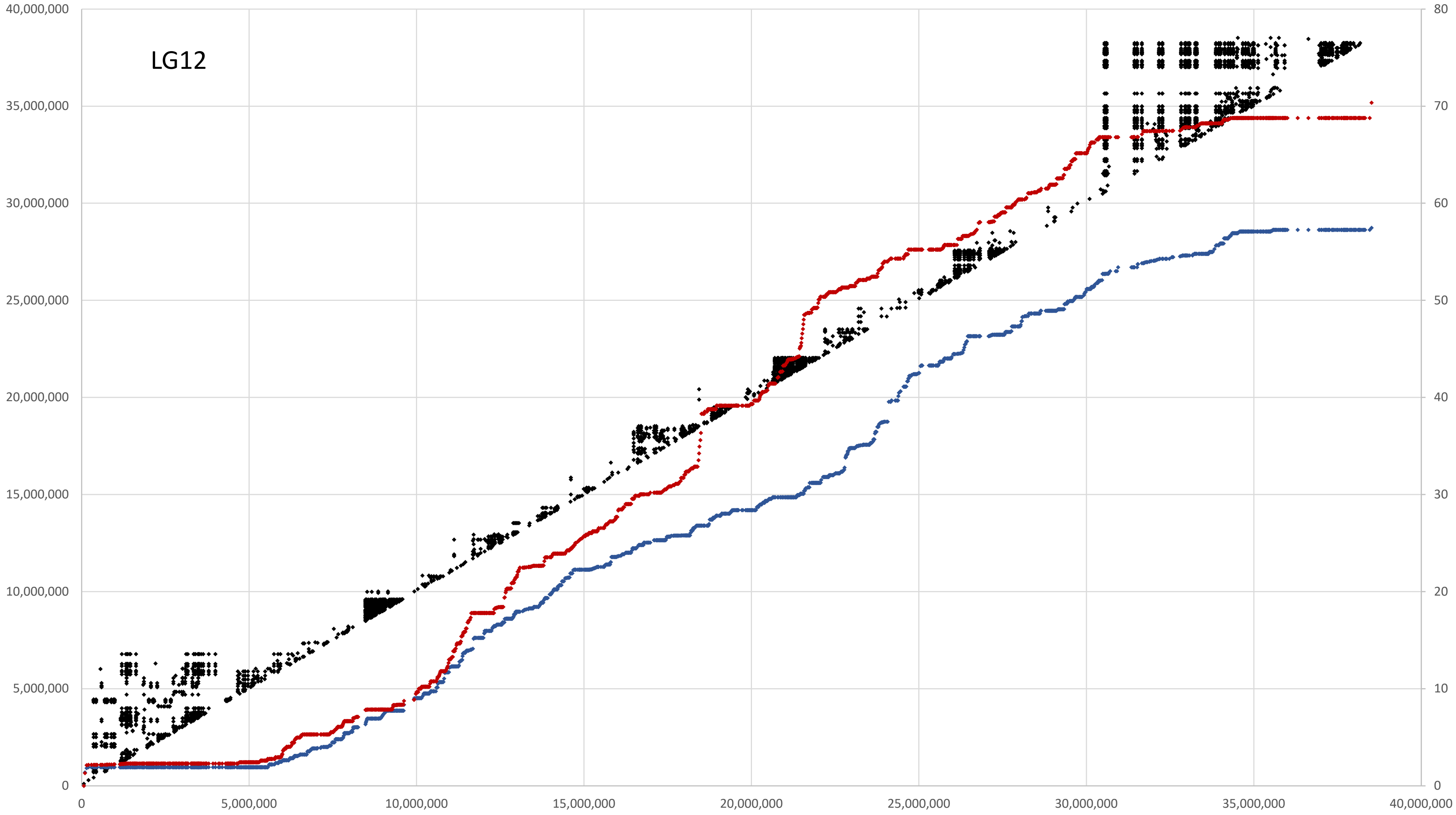

LG07
70,000,000
60,000,000
50,000,000
40,000,000
30,000,000
20,000,000
10,000,000
0
140
120
100
80
60
40
20
0
0
10,000,000
20,000,000
30,000,000
40,000,000
50,000,000
60,000,000
70,000,000

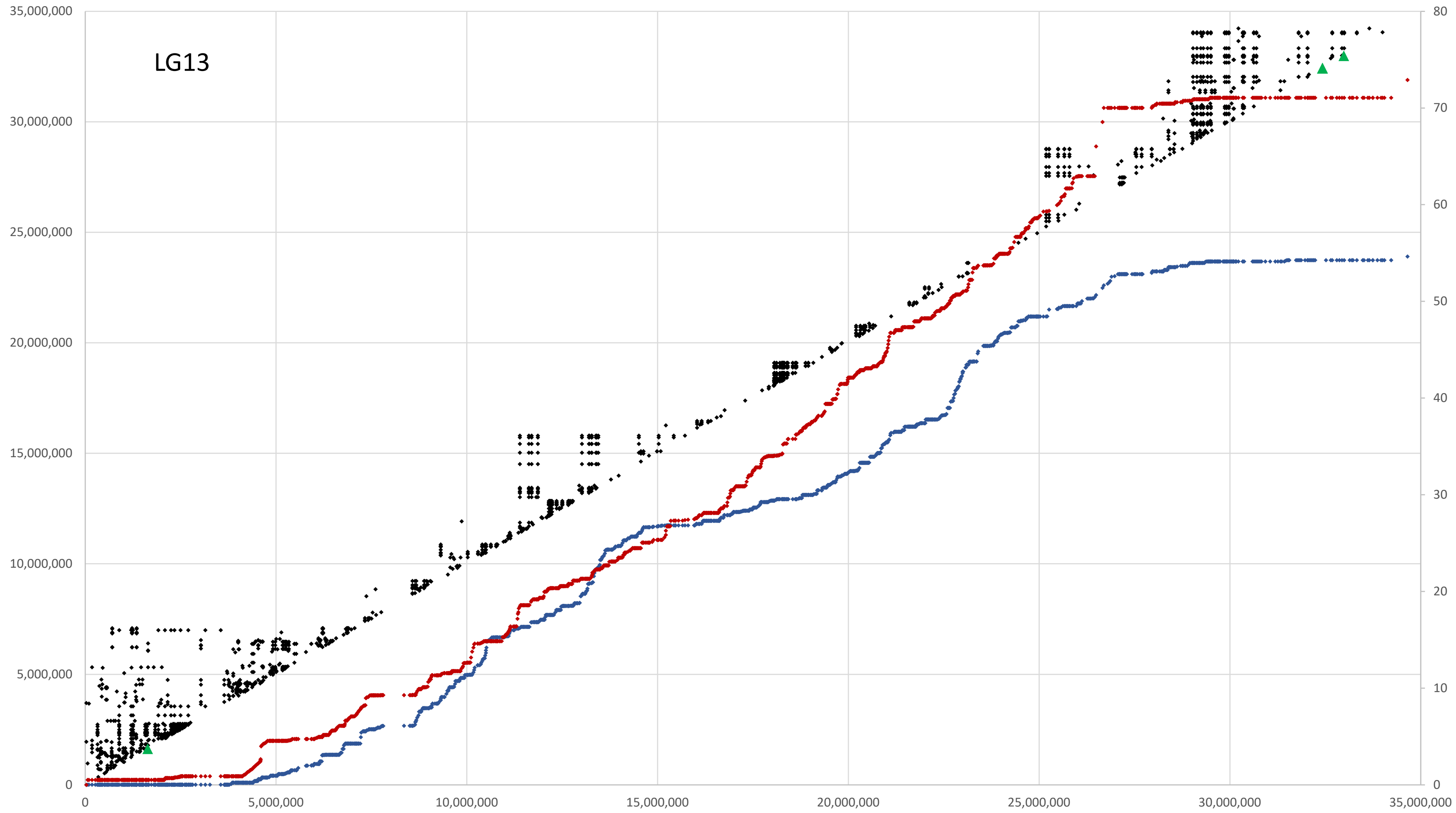

LG08
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
90
80
70
60
50
40
30
20
10
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000

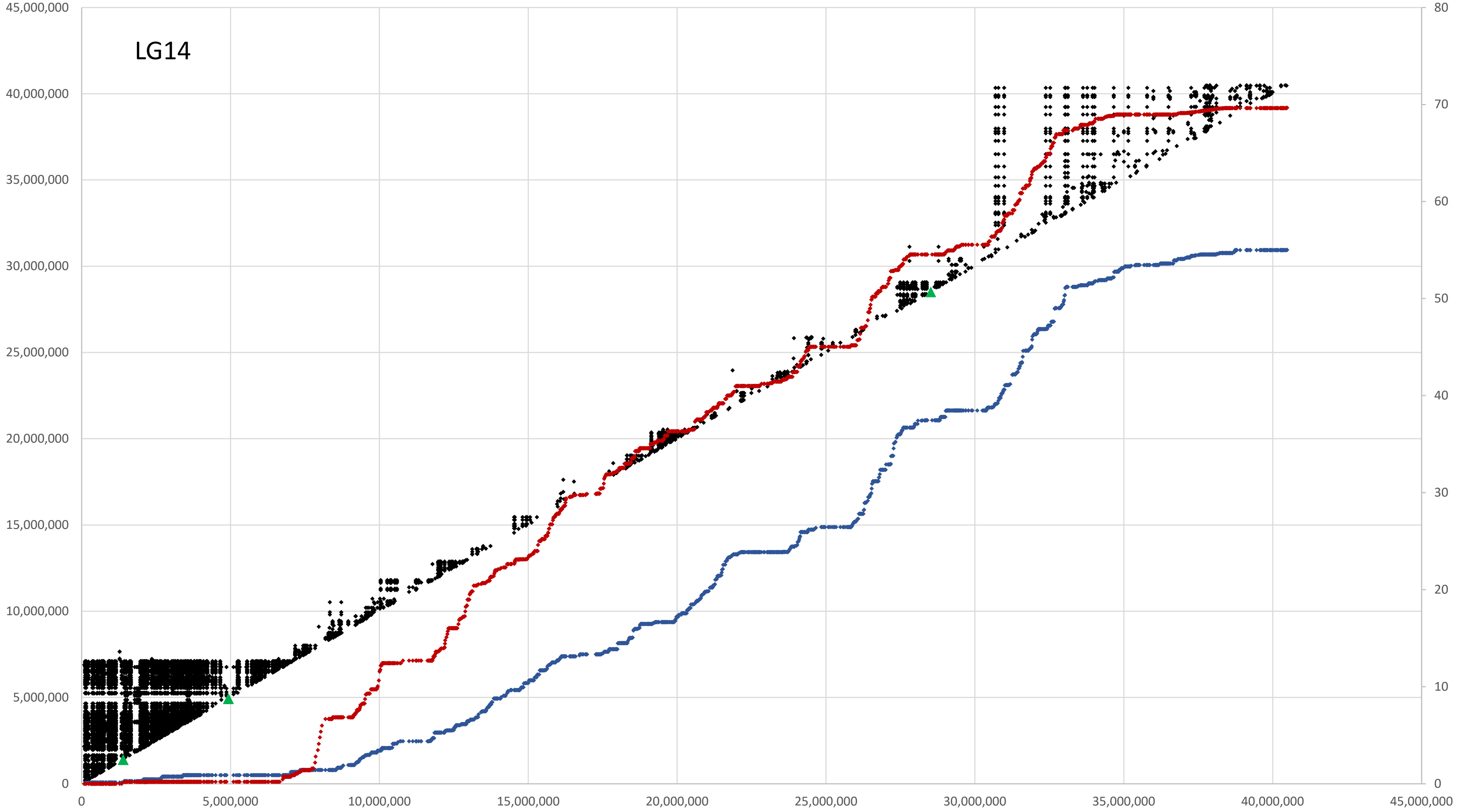

LG09
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000
70
60
50
40
30
20
10
0

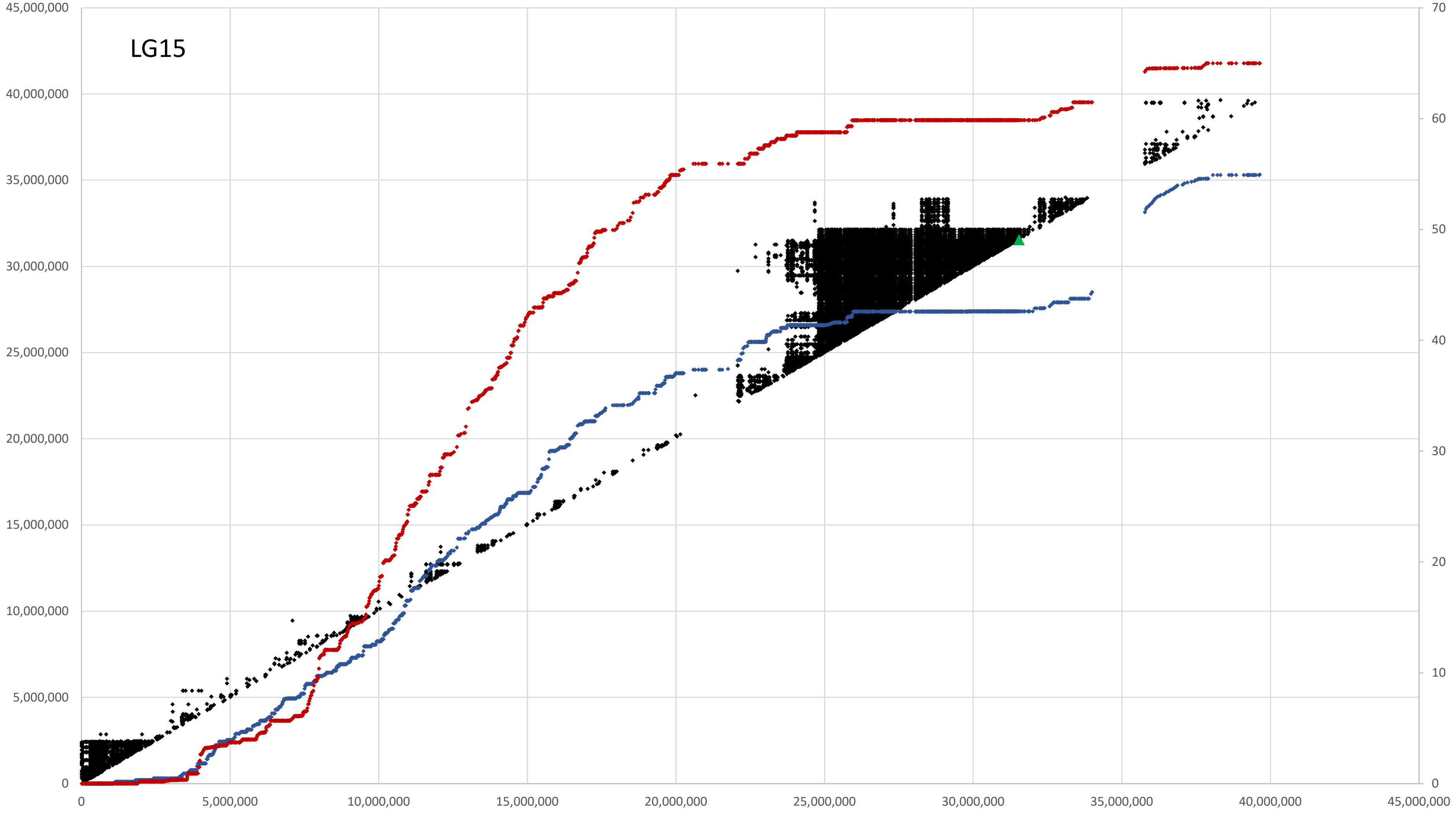

LG10
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
70
60
50
40
30
20
10
0

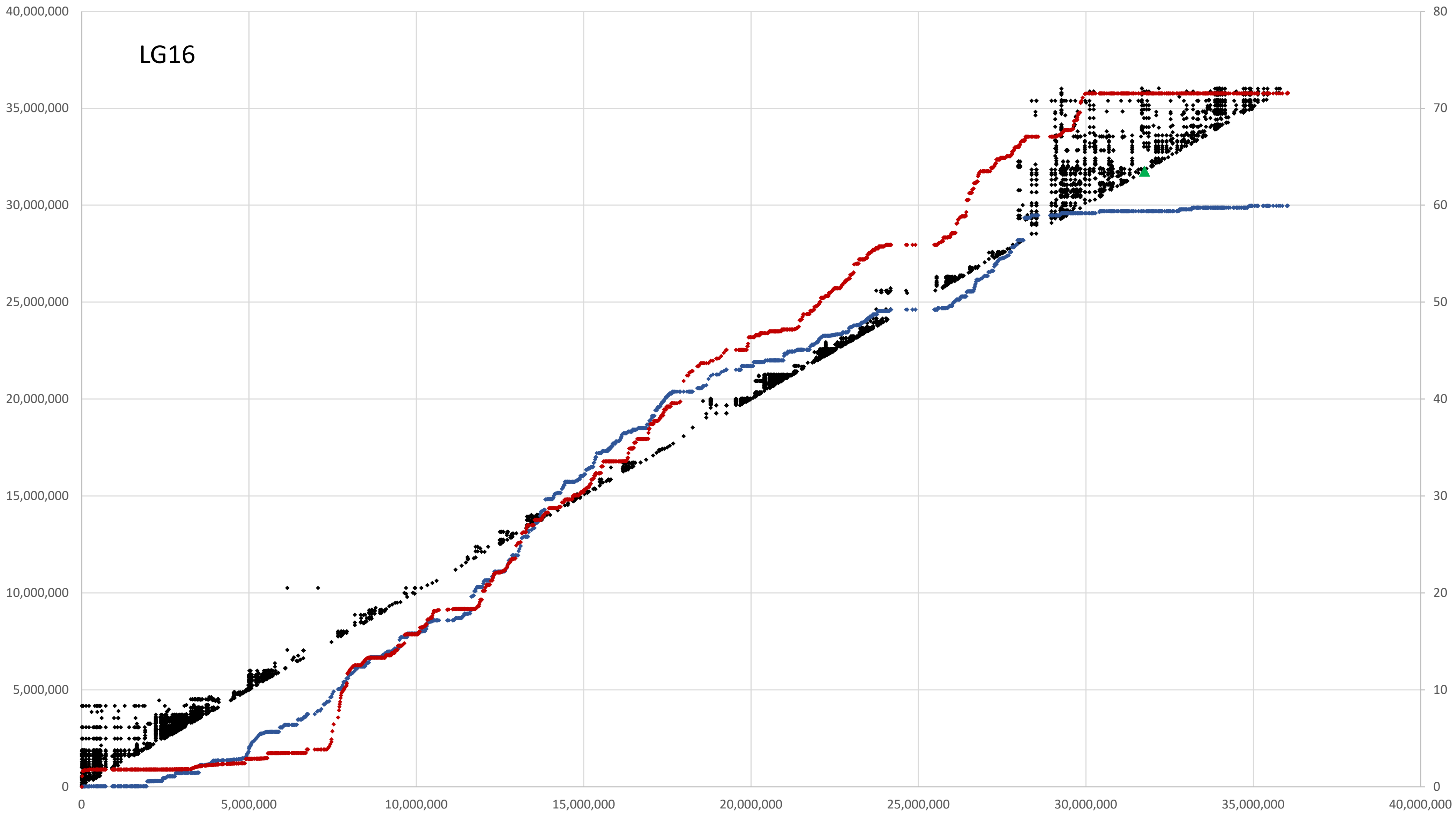

LG11
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000
90
80
70
60
50
40
30
20
10
0

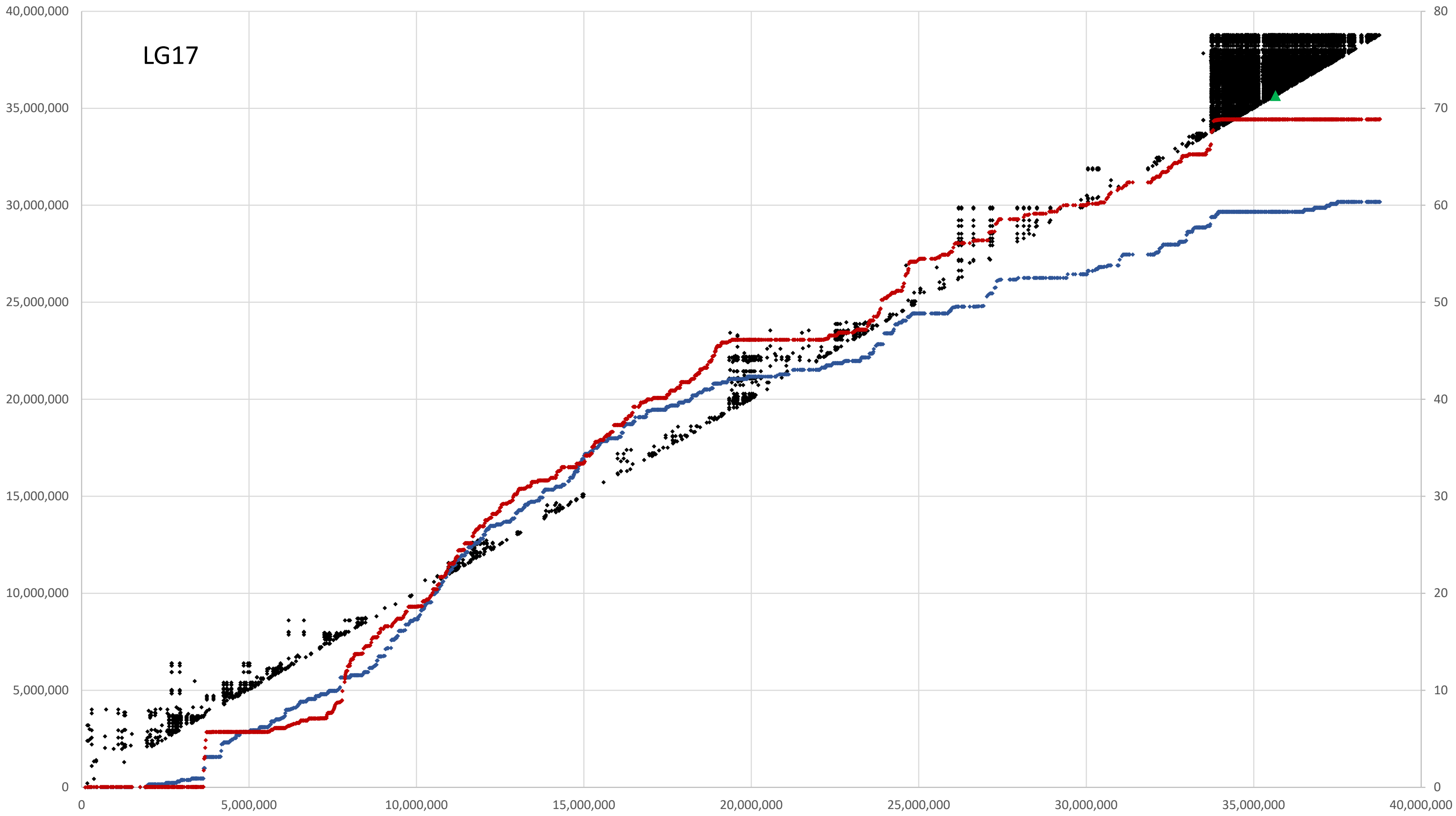

LG12
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
80
70
60
50
40
30
20
10
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000

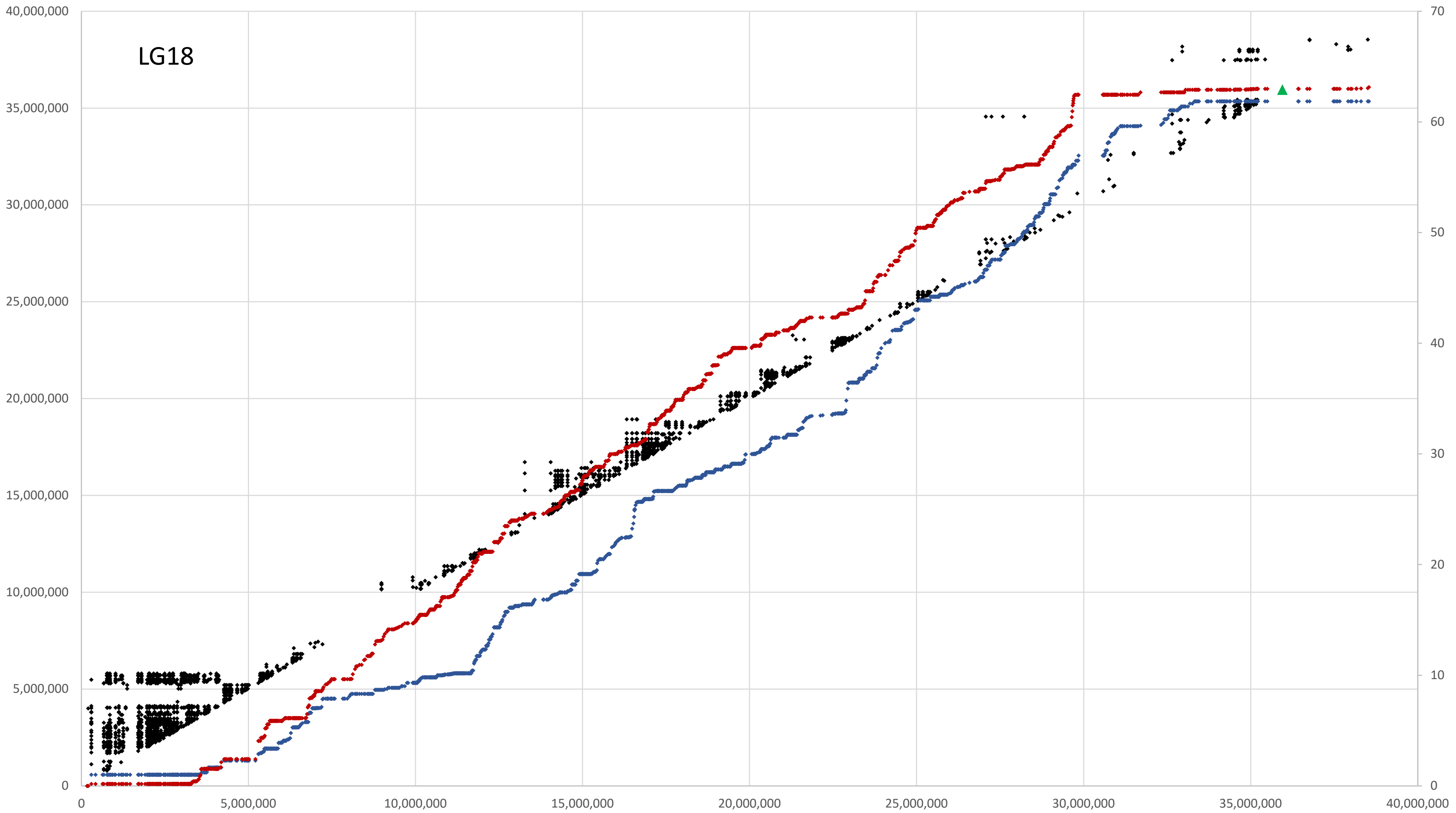

LG13
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
80
70
60
50
40
30
20
10
0

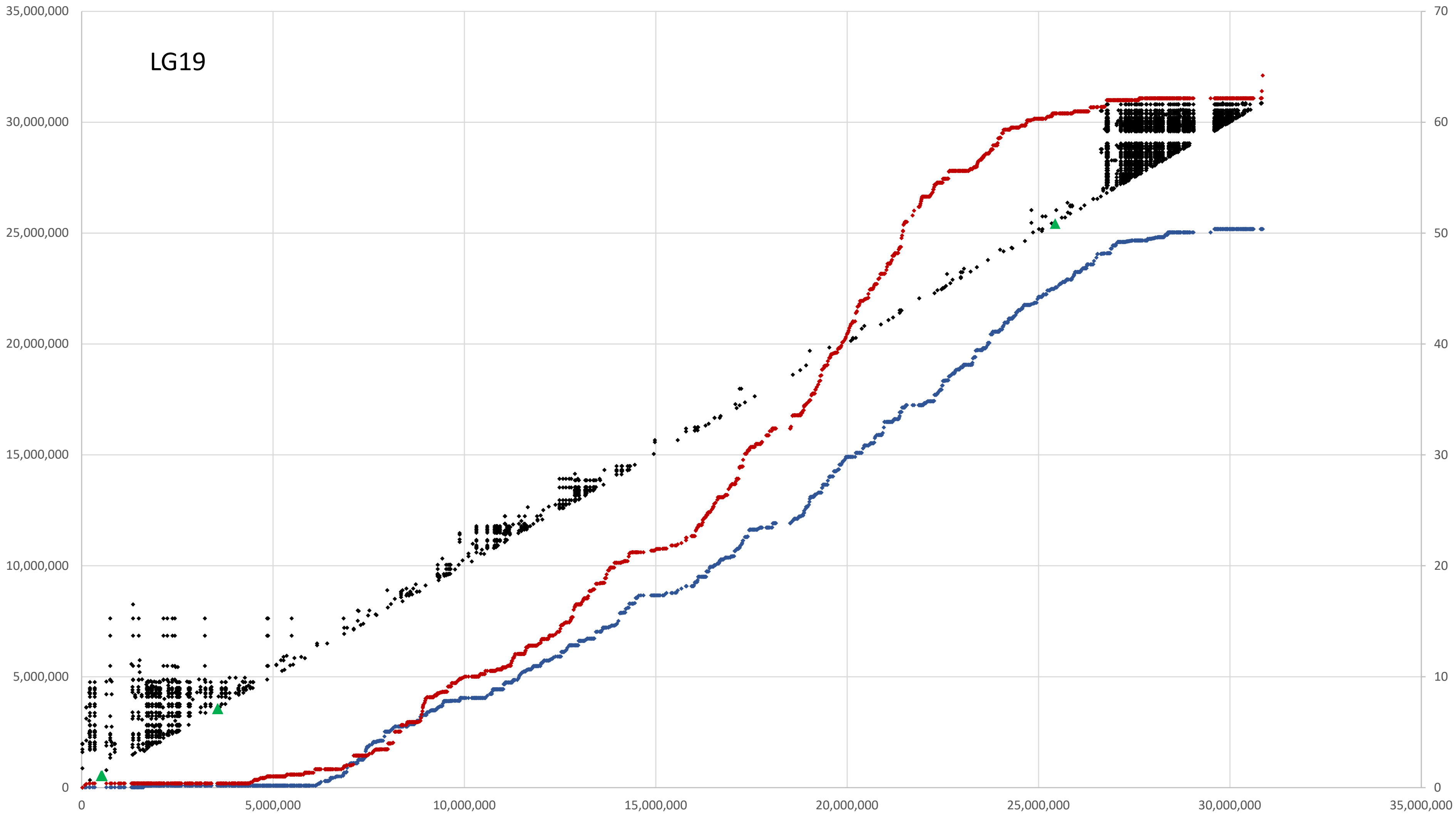

LG14
45,000,000
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000
45,000,000
80
70
60
50
40
30
20
10
0

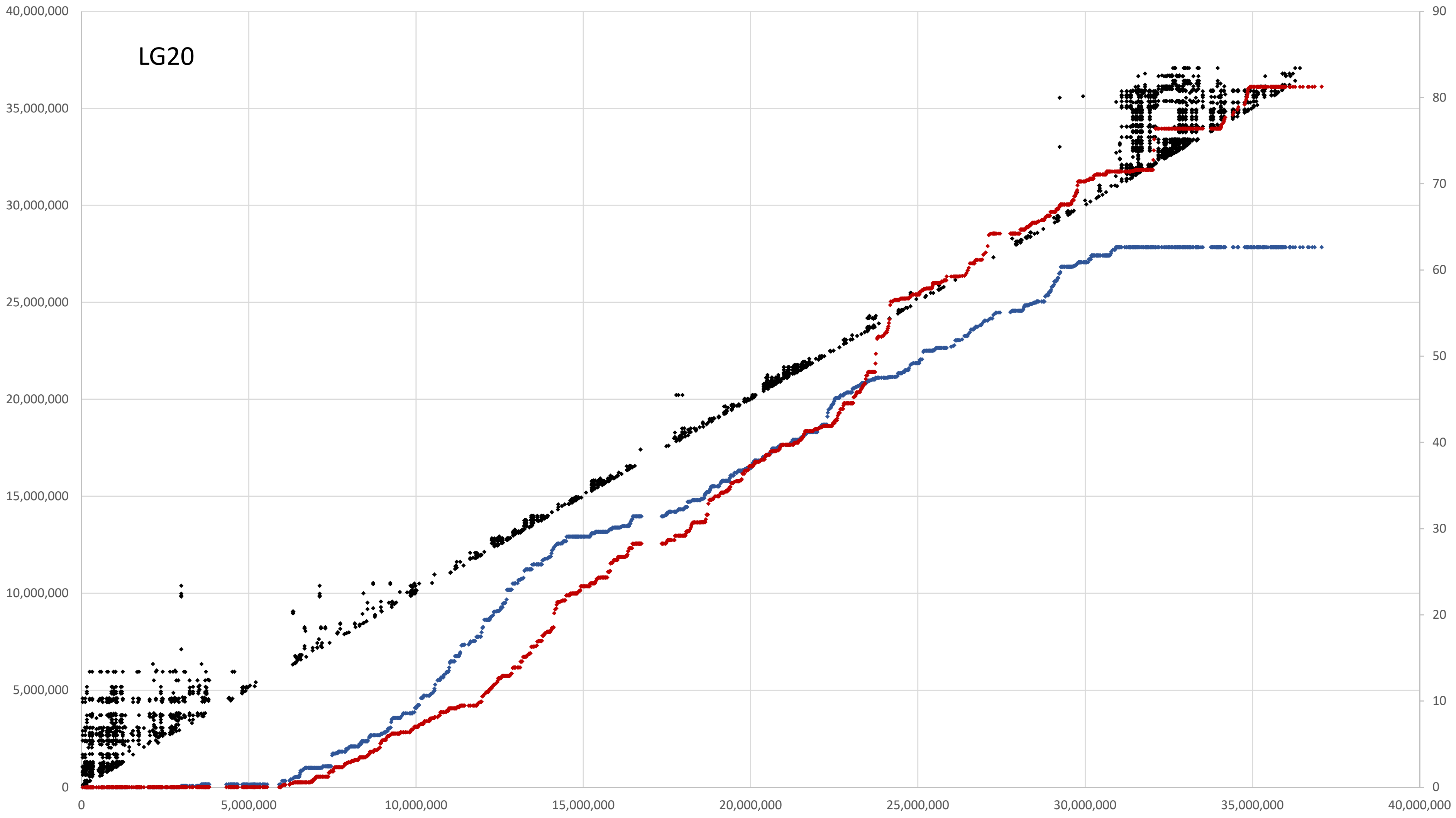

LG15
45,000,000
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
70
60
50
40
30
20
10
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000
45,000,000

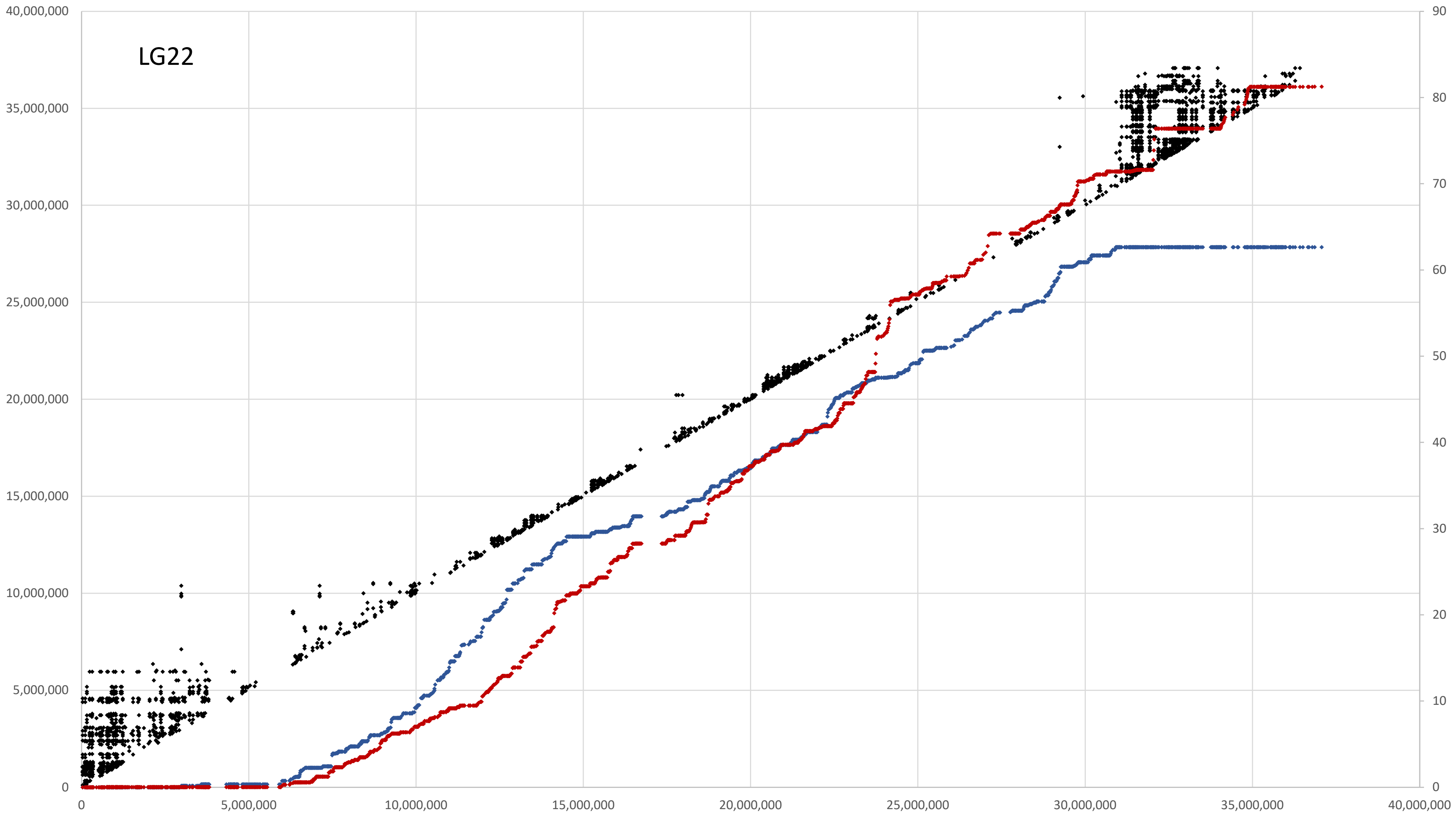

LG16
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000
80
70
60
50
40
30
20
10
0

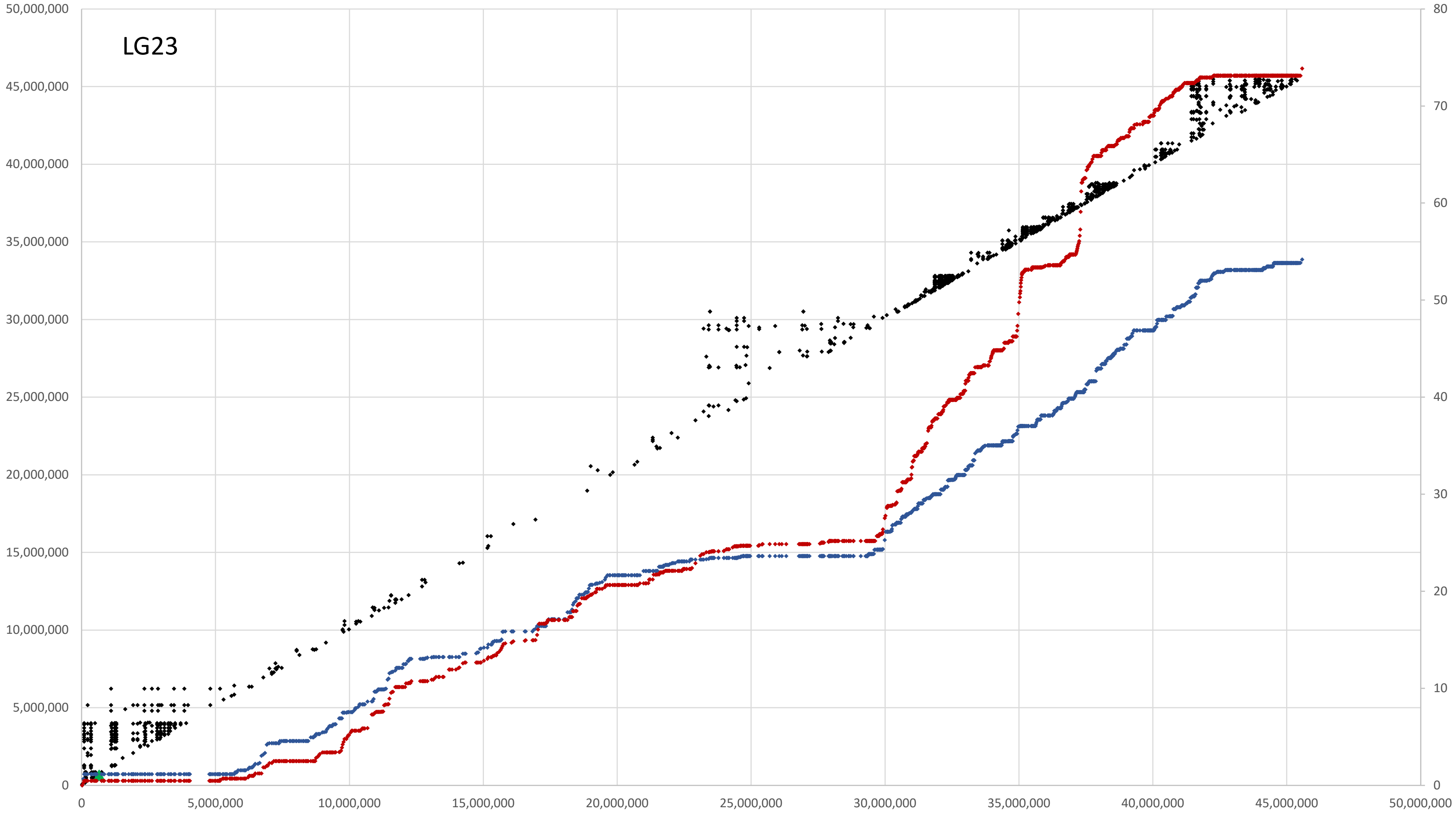

LG17
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
80
70
60
50
40
30
20
10
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000


LG18
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000
70
60
50
40
30
20
10
0


LG19
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
70
60
50
40
30
20
10
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000

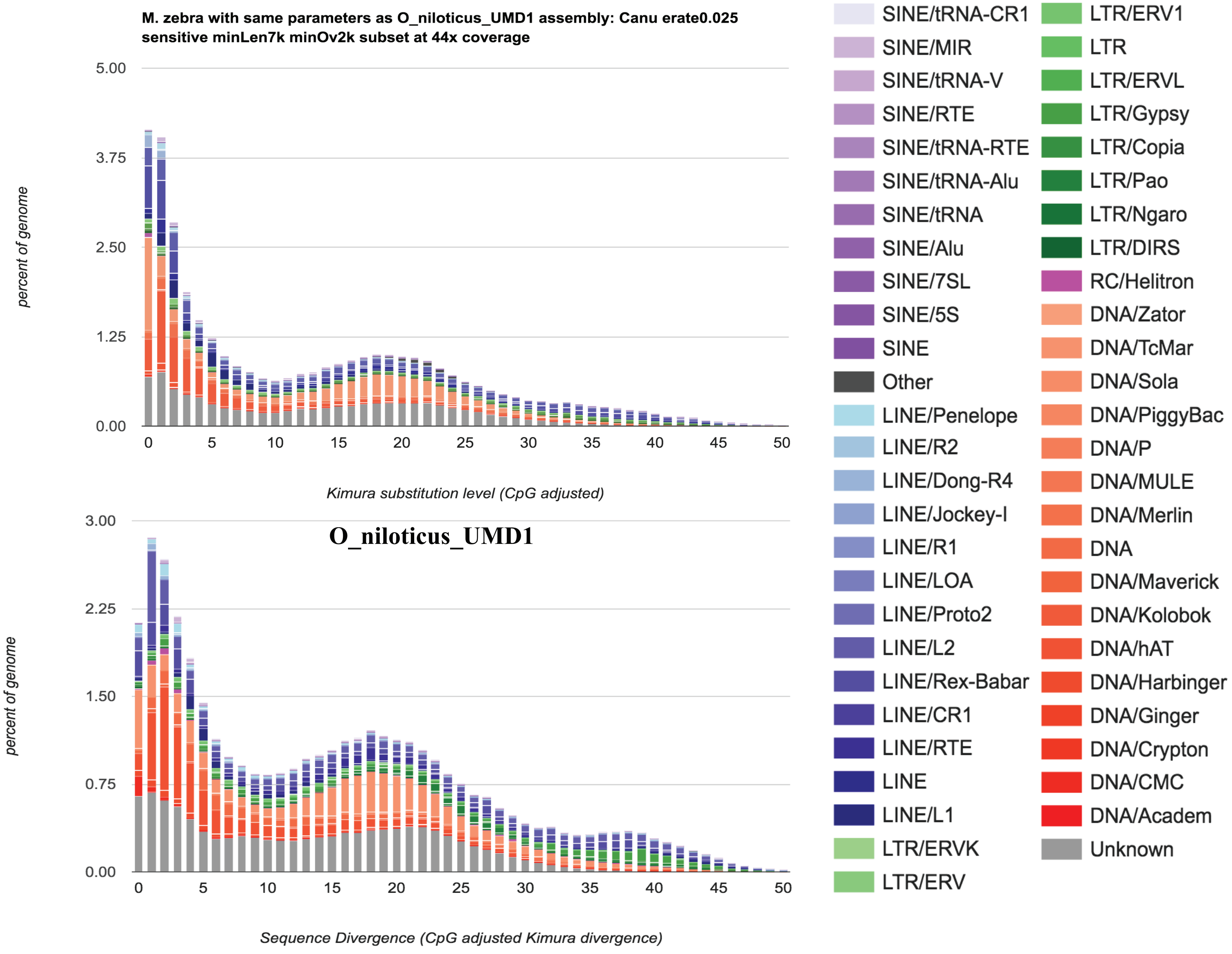

LG20
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000
90
80
70
60
50
40
30
20
10
0


LG22
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000
90
80
70
60
50
40
30
20
10
0


LG23
50,000,000
45,000,000
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
80
70
60
50
40
30
20
10
0
0
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
30,000,000
35,000,000
40,000,000
45,000,000
50,000,000

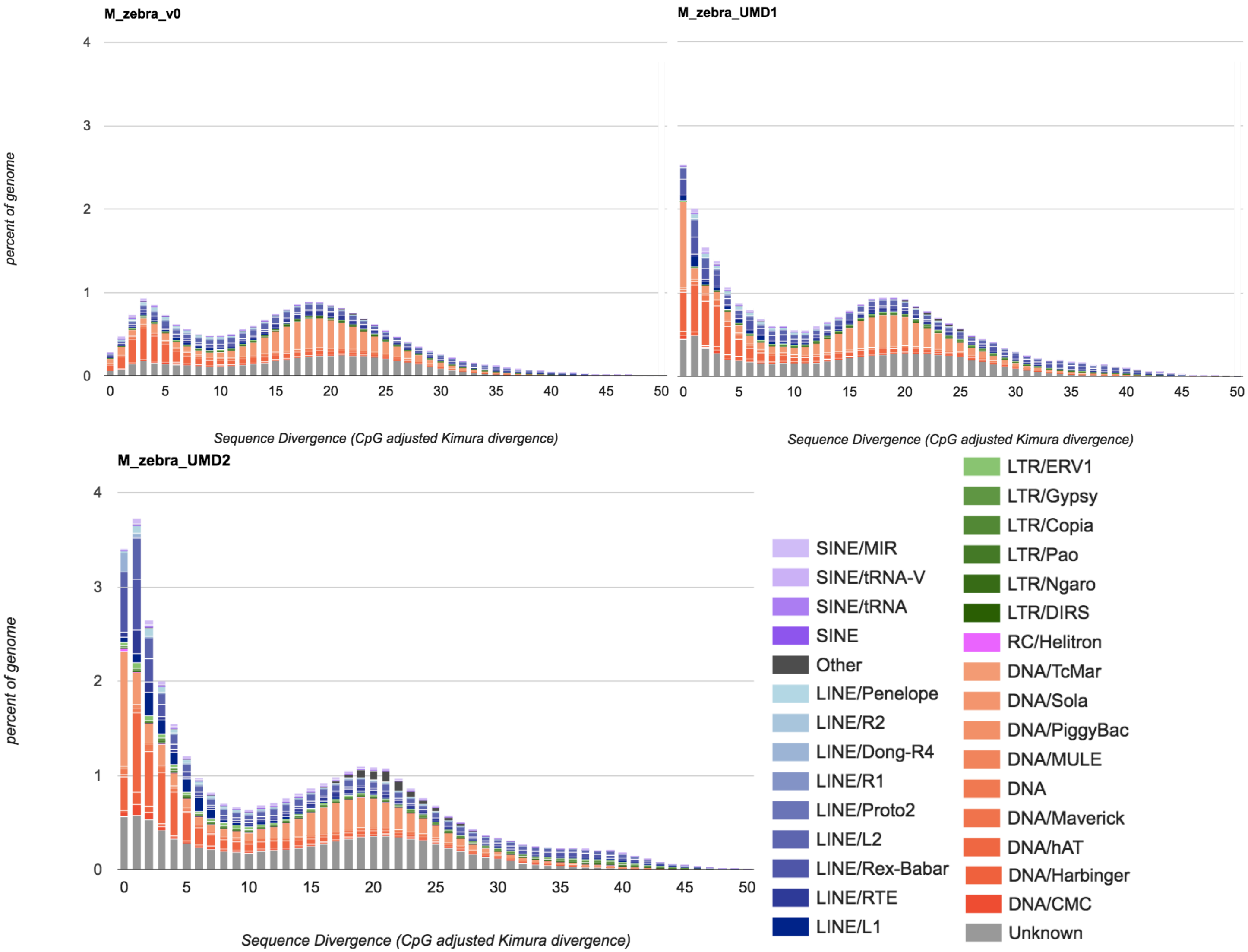
Appendix H

Comparison of the repeat landscape in the *M. zebra* and *O. niloticus* genome assemblies using same assembly parameters.



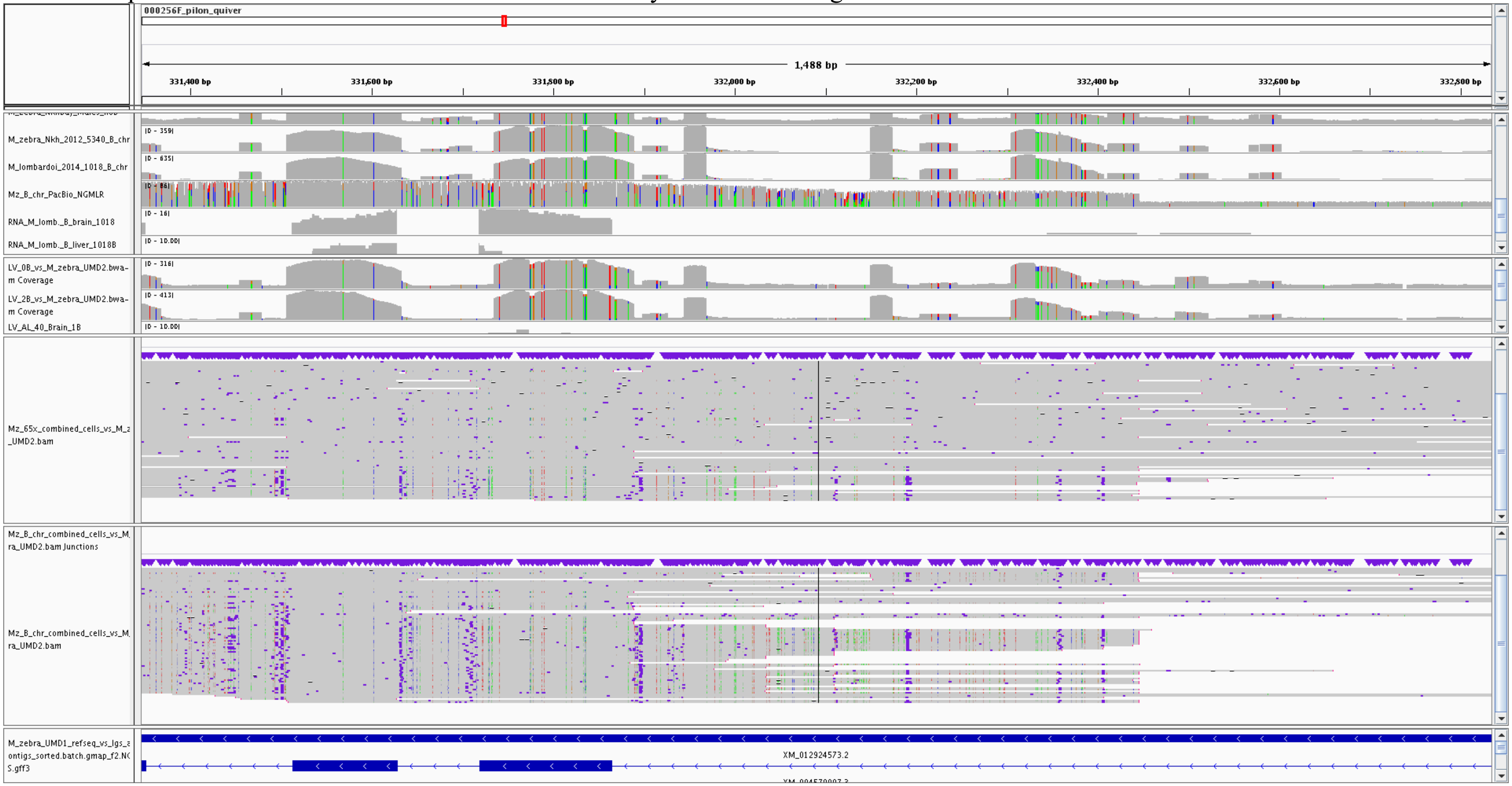
Appendix I

Comparison of the repeat landscape in the three M. zebra assembly versions.



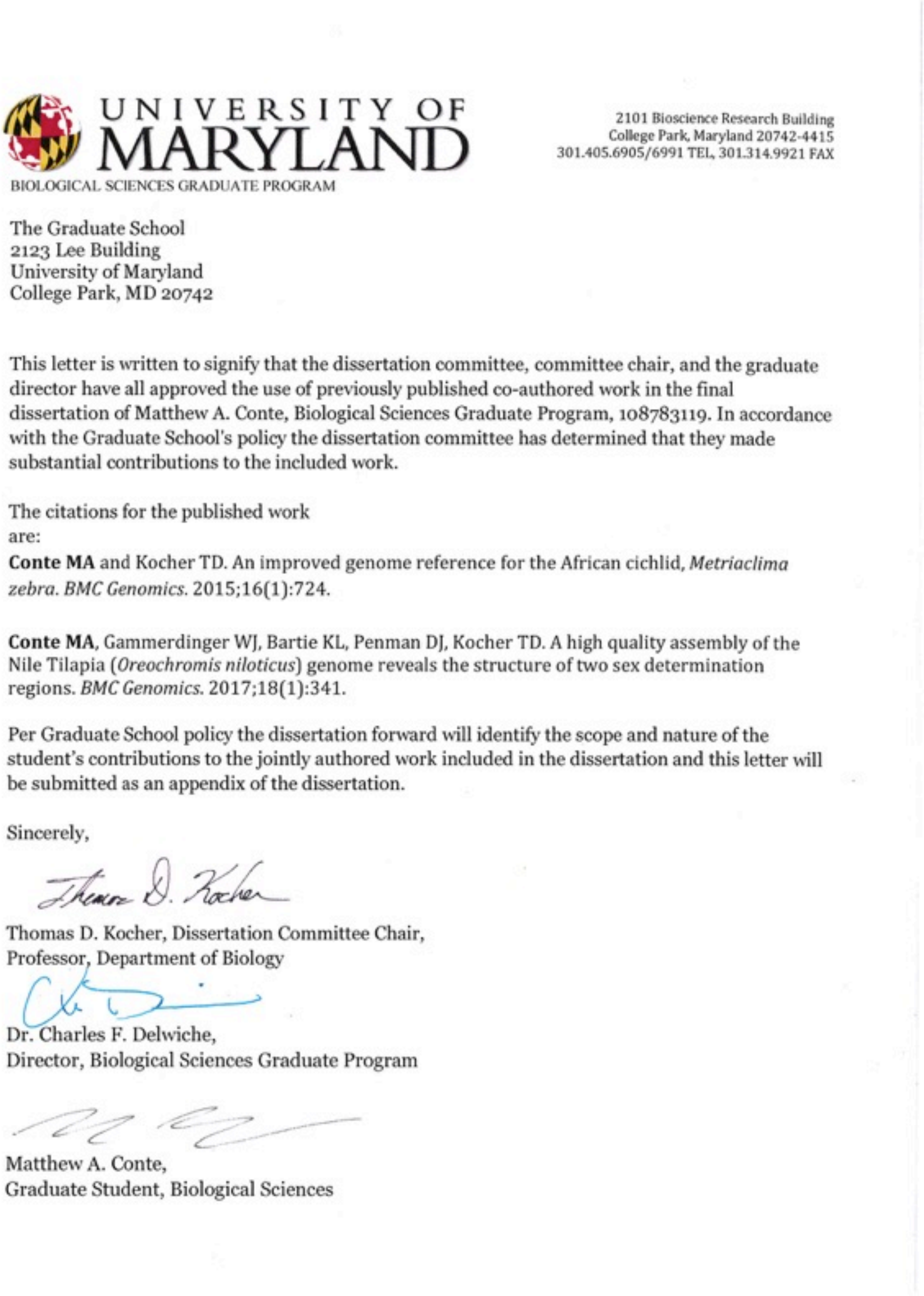
Appendix J

Additional copies of RTEL1 on the B chromosome identified by PacBio read alignment.


000256F_pilon_quiver
1,488 bp
331,400 bp
331,600 bp
331,800 bp
332,000 bp
332,200 bp
332,400 bp
332,600 bp
332,800 bp
M_zebra_Nkh_2012_5340_B_chr
[0 - 359]
M_lombardoi_2014_1018_B_chr
[0 - 635]
Mz_B_chr_PacBio_NGMLR
[0 - 86]
RNA_M_lomb._B_brain_1018
[0 - 16]
RNA_M_lomb._B_liver_1018B
[0 - 10.00]
LV_0B_vs_M_zebra_UMD2.bwa-m Coverage
[0 - 316]
LV_2B_vs_M_zebra_UMD2.bwa-m Coverage
[0 - 413]
LV_AL_40_Brain_1B
[0 - 10.00]
Mz_65x_combined_cells_vs_M_z_UMD2.bam
Mz_B_chr_combined_cells_vs_M_ra_UMD2.bam Junctions
Mz_B_chr_combined_cells_vs_M_ra_UMD2.bam
M_zebra_UMD1_refseq_vs_lgs_a ontigs_sorted.batch.gmap_f2.NC S.gff3
XM_012924573.2

Appendix K

Previously co-authored work approval letter.



UNIVERSITY OF
MARYLAND
BIOLOGICAL SCIENCES GRADUATE PROGRAM

2101 Bioscience Research Building
College Park, Maryland 20742-4415
301.405.6905/6991 TEL, 301.314.9921 FAX

The Graduate School
2123 Lee Building
University of Maryland
College Park, MD 20742

This letter is written to signify that the dissertation committee, committee chair, and the graduate director have all approved the use of previously published co-authored work in the final dissertation of Matthew A. Conte, Biological Sciences Graduate Program, 108783119. In accordance with the Graduate School's policy the dissertation committee has determined that they made substantial contributions to the included work.

The citations for the published work are:

Conte MA and Kocher TD. An improved genome reference for the African cichlid, *Metriaclima zebra. BMC Genomics.* 2015;16(1):724.

Conte MA, Gammerdinger WJ, Bartie KL, Penman DJ, Kocher TD. A high quality assembly of the Nile Tilapia (*Oreochromis niloticus*) genome reveals the structure of two sex determination regions. *BMC Genomics.* 2017;18(1):341.

Per Graduate School policy the dissertation forward will identify the scope and nature of the student's contributions to the jointly authored work included in the dissertation and this letter will be submitted as an appendix of the dissertation.

Sincerely,

Thomas D. Kocher

Thomas D. Kocher, Dissertation Committee Chair,
Professor, Department of Biology

Dr. Charles F. Delwiche,
Director, Biological Sciences Graduate Program

Matthew A. Conte,
Graduate Student, Biological Sciences

Bibliography

1. Salzburger W, Meyer A. The species flocks of East African cichlid fishes: recent advances in molecular phylogenetics and population genetics. Naturwissenschaften [Internet]. 2004;91(6). Available from: http://link.springer.com/10.1007/s00114-004-0528-6
2. Martens K. Speciation in ancient lakes. Vol. 12, Trends in Ecology and Evolution. 1997. p. 177–82.
3. Turner GF, Seehausen O, Knight ME, Allender CJ, Robinson RL. How many species of cichlid fishes are there in African lakes? Mol Ecol. 2001;10(3):793–806.
4. Kocher TD, Baroiller J-F, Fernald RD, Hey J, Hofman H, Meyer A, et al. Genetic Basis of Vertebrate Diversity: the Cichlid Fish Model [Internet]. 2006. Available from: https://www.genome.gov/pages/research/sequencing/seqproposals/cichlidgenomeseq.pdf
5. Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, et al. The genomic substrate for adaptive radiation in African cichlid fish. Nature. 2014;513(18 September 2014):375–81.
6. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol İ, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience. 2013;2(10).
7. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. PLoS Comput Biol. 2014 Dec;10(12):e1003998.
8. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol [Internet]. 2013 May 29 [cited 2014 Jul 10];14(5):R51. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4053816&tool=pmcentrez&rendertype=abstract
9. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol [Internet]. 2011 Jan [cited 2014 Jul 10];12(2):R18. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3188800&tool=pmcentrez&rendertype=abstract
10. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol [Internet]. 2009 May 5 [cited 2014 Dec 1];7(5):e1000112. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2680341&tool=pmcentrez&rendertype=abstract
11. Zhang X, Goodsell J, Norgren RB. Limitations of the rhesus macaque draft genome assembly and annotation. BMC Genomics [Internet]. 2012 Jan [cited

2015 Jan 7];13(1):206. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3426473&tool=pmcentrez&rendertype=abstract

12. Florea L, Souvorov A, Kalbfleisch TS, Salzberg SL. Genome assembly has a major impact on gene content: a comparison of annotation in two Bos taurus assemblies. PLoS One [Internet]. 2011 Jan [cited 2015 Jan 7];6(6):e21400. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3120881&tool=pmcentrez&rendertype=abstract

13. Pereira J, Johnson WE, O'Brien SJ, Jarvis ED, Zhang G, Gilbert MTP, et al. Evolutionary genomics and adaptive evolution of the hedgehog gene family (shh, ihh and dhh) in vertebrates. PLoS One [Internet]. 2014 Jan [cited 2015 Jan 10];9(12):e74132. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25549322

14. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2011 Jan;8(1):61–5.

15. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, et al. GAGE : A critical evaluation of genome assemblies and assembly algorithms. 2012;

16. Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, McVey SD, et al. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol [Internet]. 2013 Sep 13 [cited 2013 Nov 13];14(9):R101. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24034426

17. Chin C, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid , finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10(6).

18. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2014 Nov 10;517:608–11.

19. Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, et al. Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res. 2014;1–9.

20. Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol [Internet]. 2015;33(6):623–30. Available from: http://dx.doi.org/10.1038/nbt.3238

21. Koepfli K-P, Paten B, O'Brien SJ. The Genome 10K Project: A Way Forward. Annu Rev Anim Biosci [Internet]. 2015;3(1):57–111. Available from: http://www.annualreviews.org/doi/abs/10.1146/annurev-animal-090414-014900

22. Jarvis E. Reference Vertebrate Genomes Project (VGP) [Internet]. 2017. Available from: https://genome10k.soe.ucsc.edu/projects/Reference_VGP

23. Feldberg E, Ivan J, Porto R, Antonio L, Bertollo C. Chromosomal Changes and Adaptation Of Cichlid Fishes During Evolution. AL Val BG Kapoor (eds),

Fish Adapt. 2003;285–309.

24. Poletto AB, Ferreira I a, Cabral-de-Mello DC, Nakajima RT, Mazzuchelli J, Ribeiro HB, et al. Chromosome differentiation patterns during cichlid fish evolution. BMC Genet [Internet]. 2010 Jan;11(1):50. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2896337&tool=pmcentrez&rendertype=abstract

25. Ser JR, Roberts RB, Kocher TD. Multiple interacting loci control sex determination in Lake Malawi cichlid fish. Evolution (N Y). 2010;64(2):486–501.

26. Roberts RB, Ser JR, Kocher TD. Sexual conflict resolved by invasion of a novel sex determiner in Lake Malawi cichlid fishes. Science. 2009 Nov 13;326(5955):998–1001.

27. Gammerdinger WJ, Conte MA, Acquah EA, Roberts RB, Kocher TD. Structure and decay of a proto-Y region in Tilapia, Oreochromis niloticus. BMC Genomics [Internet]. 2014 Jan [cited 2015 Jan 7];15(1):975. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4251933&tool=pmcentrez&rendertype=abstract

28. Gammerdinger WJ, Conte MA, Baroiller J, Cotta HD, Kocher TD. Comparative analysis of a sex chromosome from the blackchin tilapia , Sarotherodon melanotheron. BMC Genomics [Internet]. 2016;1–10. Available from: http://dx.doi.org/10.1186/s12864-016-3163-7

29. Conte MA, Gammerdinger WJ, Bartie KL, Penman DJ, Kocher TD. A high quality assembly of the Nile Tilapia (Oreochromis niloticus) genome reveals the structure of two sex determination regions. BMC Genomics [Internet]. 2017;18(1):341. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3723-5

30. Bachtrog D. Y-chromosome evolution: Emerging insights into processes of Y-chromosome degeneration. Vol. 14, Nature Reviews Genetics. 2013. p. 113–24.

31. Kocher TD. Adaptive evolution and explosive speciation: the cichlid fish model. Nat Rev Genet [Internet]. 2004 Apr [cited 2014 Jul 15];5(4):288–98. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15131652

32. Valente GT, Conte M a, Fantinatti BE a, Cabral-de-Mello DC, Carvalho RF, Vicari MR, et al. Origin and evolution of B chromosomes in the Cichlid Fish Astatotilapia latifasciata based on integrated genomic analyses. Mol Biol Evol [Internet]. 2014 Aug [cited 2015 Jan 13];31(8):2061–72. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24770715

33. Wilson E. The supernumerary chromosomes of Hemiptera. Science (80-). 1907;(26):870–1.

34. Beukeboom LEOW. Review article Bewildering Bs : an impression of the 1st B-Chromosome Conference. 1994;73(March).

35. Jones RN, Rees H. B chromosomes. 1982.

36. Banaei-Moghaddam AM, Martis MM, Macas J, Gundlach H, Himmelbach A, Altschmied L, et al. Genes on B chromosomes: Old questions revisited with

new tools. Biochim Biophys Acta [Internet]. 2014 Dec 3 [cited 2014 Dec 9];1849(1):64–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25481283

37. Banaei-Moghaddam AM, Meier K, Karimi-Ashtiyani R, Houben A. Formation and expression of pseudogenes on the B chromosome of rye. Plant Cell [Internet]. 2013 Jul [cited 2014 Dec 9];25(7):2536–44. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3753381&tool=pmcentrez&rendertype=abstract

38. Yoshida K, Terai Y, Mizoiri S, Aibara M, Nishihara H, Watanabe M, et al. B chromosomes have a functional effect on female sex determination in Lake Victoria cichlid fishes. PLoS Genet [Internet]. 2011 Aug [cited 2013 Mar 25];7(8):e1002203. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3158035&tool=pmcentrez&rendertype=abstract

39. Poletto AB, Ferreira I a, Cabral-de-Mello DC, Nakajima RT, Mazzuchelli J, Ribeiro HB, et al. Chromosome differentiation patterns during cichlid fish evolution. BMC Genet [Internet]. 2010 Jan;11:50. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2896337&tool=pmcentrez&rendertype=abstract

40. Clark FE, Conte MA, Ferreira-Bravo IA, Poletto AB, Martins C, Kocher TD. Dynamic sequence evolution of a sex-Associated b chromosome in lake Malawi cichlid fish. J Hered. 2017;108(1):53–62.

41. Haussler D, O'Brien SJ, Ryder OA, Keith Barker F, Clamp M, Crawford AJ, et al. Genome 10K: A proposal to obtain whole-genome sequence for 10000 vertebrate species. J Hered. 2009;100(6):659–74.

42. Zhang G. Bird sequencing project takes off. Nature. 2015;522:34–34.

43. Mardis ER. A decade's perspective on DNA sequencing technology. Nature [Internet]. 2011;470(7333):198–203. Available from: http://dx.doi.org/10.1038/nature09796

44. Schatz M, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. Genome Res [Internet]. 2010;20(9):1165–73. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.176.246&rep=rep1&type=pdf

45. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A [Internet]. 2011 Jan 25 [cited 2014 Jul 13];108(4):1513–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3029755&tool=pmcentrez&rendertype=abstract

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

47. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing

technology. PLoS One [Internet]. 2012 Jan [cited 2013 Nov 7];7(11):e47768. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3504050&tool=pmcentrez&rendertype=abstract

48. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome assembly evaluation. Genome Biol [Internet]. 2013 Jan;14(5):R47. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3798757&tool=pmcentrez&rendertype=abstract

49. Bardou P, Mariette J, Escudié F, Djemiel C, Klopp C. jvenn : an interactive Venn diagram viewer. BMC Bioinformatics. 2014;15(293).

50. Hackl T, Hedrich R, Schultz J, Förster F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics [Internet]. 2014 Jul 10 [cited 2014 Oct 18];30(21):3004–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25015988

51. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol [Internet]. 2012 Jul [cited 2013 Mar 2];30(7):693–700. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22750884

52. Fichot EB, Norman RS. Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. Microbiome [Internet]. 2013 Jan [cited 2014 Nov 20];1(1):10. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3971627&tool=pmcentrez&rendertype=abstract

53. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics [Internet]. 2011 Nov 1 [cited 2014 Jul 13];27(21):2957–63. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3198573&tool=pmcentrez&rendertype=abstract

54. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. 2012;13(238).

55. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21(9):1859–75.

56. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52.

57. Cichlid RNASeq_assemblies [Internet]. [cited 2012 Feb 6]. Available from: ftp://ftp.broadinstitute.org/pub/vgb/cichlids/RNASeq_Assemblies/M_zebra.transcripts.tgz

58. O'Quin KE, Smith D, Naseer Z, Schulte J, Engel SD, Loh Y-HE, et al. Divergence in cis-regulatory sequences surrounding the opsin gene arrays of African cichlid fishes. BMC Evol Biol [Internet]. 2011;11(1):120. Available from: http://www.biomedcentral.com/1471-2148/11/120

59. Krumsiek J, Arnold R, Rattei T. Gepard: A rapid and sensitive tool for creating

dotplots on genome scale. Bioinformatics. 2007;23(8):1026–8.

60. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics [Internet]. 2007 May 1 [cited 2014 Jul 11];23(9):1061–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17332020

61. Clark SC, Egan R, Frazier PI, Wang Z. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. Bioinformatics [Internet]. 2013 Mar 15 [cited 2014 Jul 29];29(4):435–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23303509

62. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods [Internet]. 2012 Apr [cited 2013 Feb 28];9(4):357–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22388286

63. Bradnam K. assemblathon2-analysis/assemblathon_stats.pl [Internet]. [cited 2015 Jan 15]. Available from: https://github.com/ucdavis-bioinformatics/assemblathon2-analysis/blob/master/assemblathon_stats.pl

64. Smit, AFA, Hubley R. RepeatModeler Open-1.0 [Internet]. 2010. Available from: www.repeatmasker.org

65. Smit, AFA, Hubley, R & Green P. RepeatMasker Open-4.0 [Internet]. 2010. Available from: www.repeatmasker.org

66. O'Quin CT, Drilea AC, Conte MA, Kocher TD. Mapping of pigmentation QTL on an anchored genome assembly of the cichlid fish , Metriaclima zebra. BMC Genomics [Internet]. 2013;14(1):1. Available from: BMC Genomics

67. SMRT View · PacificBiosciences/DevNet Wiki [Internet]. [cited 2014 May 2]. Available from: https://github.com/PacificBiosciences/DevNet/wiki/SMRT-View

68. PacificBiosciences/SMRT-Analysis [Internet]. [cited 2014 May 5]. Available from: https://github.com/PacificBiosciences/SMRT-Analysis

69. FAO. World Review of Fisheries and Aquaculture [Internet]. 2012 [cited 2016 Dec 21]. Available from: http://www.fao.org/docrep/016/i2727e/i2727e01.pdf

70. FAO. Cultured Aquatic Species Information Programme. In: FAO Fisheries and Aquaculture Department [Internet]. Available from: http://www.fao.org/fishery/culturedspecies/Oreochromis_niloticus/en

71. Mair GC, Abucay JS, Skibinski DOF, Abella TA, Beardmore JA. Genetic manipulation of sex ratio for the large-scale production of all-male tilapia *Oreochromis niloticus*. Can J Fish Aquat Sci. 1997;54(2):396–404.

72. Hickling C. The Malacca tilapia hybrids. J Genet. 1960;57:1–10.

73. Little D, Hulata G. Strategies for tilapia seed production. In: Beveridge M, McAndrew B, editors. Tilapias: Biology and Exploitation. 2000. p. 267–326.

74. Baroiller JF, D'Cotta H, Bezault E, Wessels S, Hoerstgen-Schwark G. Tilapia sex determination: Where temperature and genetics meet. Vol. 153, Comparative Biochemistry and Physiology - A Molecular and Integrative Physiology. 2009. p. 30–8.

75. Wessels S, Sharifi RA, Luehmann LM, Rueangsri S, Krause I, Pach S, et al. Allelic variant in the anti-müllerian hormone gene leads to autosomal and

temperature-dependent sex reversal in a selected nile tilapia line. PLoS One [Internet]. 2014 Jan [cited 2014 Sep 5];9(8):e104795. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4144872&tool=pmcentrez&rendertype=abstract

76. Palaiokostas C, Bekaert M, Khan MG, Taggart JB, Gharbi K, McAndrew BJ, et al. A novel sex-determining QTL in Nile tilapia (Oreochromis niloticus). BMC Genomics [Internet]. 2015;16(1):1–10. Available from: http://www.biomedcentral.com/1471-2164/16/171

77. Lee BY, Coutanceau JP, Ozouf-Costaz C, D'Cotta H, Baroiller JF, Kocher TD. Genetic and physical mapping of sex-linked AFLP markers in Nile tilapia (Oreochromis niloticus). Mar Biotechnol (NY). 2011;13(3):557–62.

78. Eshel O, Shirak A, Weller JI, Hulata G, Ron M. Linkage and Physical Mapping of Sex Region on LG23 of Nile Tilapia (Oreochromis niloticus). G3 Genes, Genomes, Genet. 2012 Jan;2(1):35–42.

79. Lee B-YB, Hulata G, Kocher TD. Two unlinked loci controlling the sex of blue tilapia (Oreochromis aureus). Heredity (Edinb). 2004 Jun;92(6):543–9.

80. Ser JR, Roberts RB, Kocher TD. Multiple interacting loci control sex determination in lake Malawi cichlid fish. Evolution. 2010 Feb 1;64(2):486–501.

81. Parnell NF, Streelman JT. Genetic interactions controlling sex and color establish the potential for sexual conflict in Lake Malawi cichlid fishes. Heredity (Edinb). 2013 Mar;110(3):239–46.

82. Li M, Sun Y, Zhao J, Shi H, Zeng S, Ye K, et al. A Tandem Duplicate of Anti-Müllerian Hormone with a Missense SNP on the Y Chromosome Is Essential for Male Sex Determination in Nile Tilapia, *Oreochromis niloticus*. PLoS Genet. 2015;11(11):1–23.

83. Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, et al. The genomic substrate for adaptive radiation in African cichlid fish. Nature. 2014;513(18 September 2014):375–81.

84. Conte MA, Kocher TD. An improved genome reference for the African cichlid, Metriaclima zebra. BMC Genomics. 2015;16(1):724.

85. Copetti D, Wing RA. The dark side of the genome: revealing the native transposable element/repeat content of eukaryotic genomes. Mol Plant. 2016;9(12):1664–1666.

86. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing. Nat Methods. 2016;13:1050–4.

87. Seo J-S, Rhie A, Kim J, Lee S, Sohn M, Kim C-U, et al. De novo assembly and phasing of a Korean human genome. Nature [Internet]. 2016;538(7624):243–7. Available from: http://www.nature.com/doifinder/10.1038/nature20098

88. Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, et al. Long-read sequencing and de novo assembly of a Chinese genome. Nat Commun. 2016;7(1):12065.

89. Vij S, Kuhl H, Kuznetsova IS, Komissarov A, Yurchenko AA, Van Heusden P, et al. Chromosomal-Level Assembly of the Asian Seabass Genome Using

Long Sequence Reads and Multi-layered Scaffolding. PLOS Genet [Internet]. 2016;12(4):e1005954. Available from: http://dx.plos.org/10.1371/journal.pgen.1005954

90. Sarder MRI, Penman DJ, Myers JM, McAndrew BJ. Production and propogation of fully inbred Clonal lines in the Nile Tilapia (Oreochromis niloticus L.). J Exp Zool. 1999;284(6):675–85.

91. Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. bioRxiv. 2016;71282.

92. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.

93. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45.

94. Soler L, Conte MA, Katagiri T, Howe AE, Lee B-Y, Amemiya C, et al. Comparative physical maps derived from BAC end sequences of tilapia (Oreochromis niloticus). BMC Genomics [Internet]. 2010;11(1):636. Available from: http://www.biomedcentral.com/1471-2164/11/636

95. Guyon R, Rakotomanga M, Azzouzi N, Coutanceau JP, Bonillo C, D'Cotta H, et al. A high-resolution map of the Nile tilapia genome: a resource for studying cichlids and other percomorphs. BMC Genomics [Internet]. 2012 Jan;13:222. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3441813&tool=pmcentrez&rendertype=abstract

96. Palaiokostas C, Bekaert M, Khan MGQ, Taggart JB, Gharbi K, McAndrew BJ, et al. Mapping and Validation of the Major Sex-Determining Region in Nile Tilapia (Oreochromis niloticus L.) Using RAD Sequencing. PLoS One. 2013;8(7):1–9.

97. Gregory TR. Animal Genome Size Database [Internet]. 2016. Available from: http://www.genomesize.com

98. Ferreira I a, Poletto a B, Kocher TD, Mota-Velasco JC, Penman DJ, Martins C. Chromosome evolution in African cichlid fish: contributions from the physical mapping of repeated DNAs. Cytogenet Genome Res [Internet]. 2010 Jan [cited 2014 Dec 27];129(4):314–22. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3202915&tool=pmcentrez&rendertype=abstract

99. Franck J, Wright J, McAndrew B. Genetic variability in a family of satellite DNAs from tilapia (Pisces: Cichlidae). Genome [Internet]. 1992;35(5):719–25. Available from: http://dx.doi.org/10.1139/g92-111%255Cnhttp://www.nrcresearchpress.com/doi/abs/10.1139/g92-111

100. Franck JPC, Wright JM. Conservation of a satellite DNA sequence (SATB) in the tilapiine and haplochromine genome (Pisces: Cichlidae). Genome [Internet]. 1993 Feb;36(1):187–94. Available from:

http://www.nrcresearchpress.com/doi/abs/10.1139/g93-025

101. Takahashi K, Terai Y, Nishida M, Okada N. A novel family of short interspersed repetitive elements (SINEs) from cichlids: the patterns of insertion of SINEs at orthologous loci support the proposed monophyly of four major groups of cichlid fishes in Lake Tanganyika. Mol Biol Evol. 1998;15(4):391–407.

102. Terai Y, Takahashi K, Nishida M, Sato T, Okada N. Using SINEs to probe ancient explosive speciation: "Hidden" radiation of African cichlids? Mol Biol Evol [Internet]. 2003 Jun [cited 2015 Jan 18];20(6):924–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12716991

103. Fujimura K, Conte MA, Kocher TD. Circular DNA intermediate in the duplication of Nile tilapia vasa genes. PLoS One [Internet]. 2011;6(12):e29477. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245284&tool=pmcentrez&rendertype=abstract

104. Katagiri T, Kidd C, Tomasino E, Davis JT, Wishon C, Stern JE, et al. A BAC-based physical map of the Nile tilapia genome. BMC Genomics [Internet]. 2005 Jan [cited 2015 Jan 9];6:89. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1180826&tool=pmcentrez&rendertype=abstract

105. Cnaani A, Lee BY, Zilberman N, Ozouf-Costaz C, Hulata G, Ron M, et al. Genetics of sex determination in tilapiine species. Sex Dev. 2008;2(1):43–54.

106. Lee BY, Penman DJ, Kocher TD. Identification of a sex-determining region in Nile tilapia (Oreochromis niloticus) using bulked segregant analysis. Anim Genet. 2003;34(5):379–83.

107. Ezaz MT, Harvey SC, Boonphakdee C, Teale AJ, McAndrew BJ, Penman DJ. Isolation and physical mapping of sex-linked AFLP markers in Nile tilapia (Oreochromis niloticus L.). Mar Biotechnol [Internet]. 2004;6(5):435–45. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15791488

108. Gammerdinger WJ, Conte MA, Baroiller J-F, D'Cotta H, Kocher TD. Comparative analysis of a sex chromosome from the blackchin tilapia, Sarotherodon melanotheron. BMC Genomics [Internet]. 2016;17(1):808. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3163-7

109. Campos-Ramos R, Harvey SC, Masabanda JS, Carrasco LAP, Griffin DK, McAndrew BJ, et al. Identification of putative sex chromosomes in the blue tilapia, Oreochromis aureus, through synaptonemal complex and fish analysis. Genetica. 2001;111(1–3):143–53.

110. Ferreira IA, Martins C. Physical chromosome mapping of repetitive DNA sequences in Nile tilapia Oreochromis niloticus: Evidences for a differential distribution of repetitive elements in the sex chromosomes. Micron. 2008;39(4):411–8.

111. Parliament of the United Kingdom. The Animals (Scientific Procedures) Act 1986, revised. [Internet]. Available from: http://www.legislation.gov.uk/ukpga/1986/14

112. NCBI SRA Toolkit [Internet]. Available from: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software#header-global
113. Bolger AM, Lohse M, Usadel B, Planck M, Plant M, Mühlenberg A. Trimmomatic : A flexible trimmer for Illumina Sequence Data. Bioinformatics. 2014;1–7.
114. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics [Internet]. 2009 Aug 15 [cited 2014 Jul 9];25(16):2078–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002&tool=pmcentrez&rendertype=abstract
115. Picard Tools - By Broad Institute [Internet]. [cited 2016 Feb 6]. Available from: http://broadinstitute.github.io/picard
116. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol [Internet]. 1990;215(3):403–10. Available from: http://www.sciencedirect.com/science/article/pii/S0022283605803602
117. Altschul SF, Madden TL, Schaffer A, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PS I-BLAST: a new generation of protein database search programs. Nucleic Acids Res [Internet]. 1997;25(17):3389–402. Available from: http://dx.doi.org/10.1093/nar/25.17.3389%255Cnhttp://nar.oxfordjournals.org/content/25/17/3389.short
118. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST plus: architecture and applications. BMC Bioinformatics. 2009;10(421):1.
119. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11).
120. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv Prepr arXiv12073907 [Internet]. 2012;9. Available from: http://arxiv.org/abs/1207.3907
121. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo M a., et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
122. Li H. Aligning sequence reads , clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;0(0):1–2.
123. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178–92.
124. Catchen J, Amores A. Chromonomer [Internet]. Available from: http://catchenlab.life.illinois.edu/chromonomer/
125. Lee B-Y, Lee W-J, Streelman JT, Carleton KL, Howe AE, Hulata G, et al. A second-generation genetic linkage map of tilapia (Oreochromis spp.). Genetics [Internet]. 2005 May [cited 2014 Dec 11];170(1):237–44. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1449707&tool=pmcentrez&rendertype=abstract
126. NCBI. The NCBI Eukaryotic Genome Annotation Pipeline [Internet].

Available from: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/

127. NCBI. NCBI Oreochromis niloticus Annotation Release 103 [Internet]. Available from: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Oreochromis_niloticus/103/

128. NCBI. NCBI Oreochromis niloticus Annotation Release 102 [Internet]. Available from: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Oreochromis_niloticus/102/

129. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA [Internet]. 2015;6(1):11. Available from: http://www.mobilednajournal.com/content/6/1/11

130. Kofler R, Pandey RV, Schlötterer C. PoPoolation2 : Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics. 2011;27(24):3435–6.

131. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36.

132. Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, et al. Long-read sequence assembly of the Gorilla Genome. Science (80-). 2016;352(6281).

133. Zimin A V., Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL. The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. Gigascience [Internet]. 2017; Available from: http://academic.oup.com/gigascience/article/doi/10.1093/gigascience/gix097/4561661

134. Carbone L, Alan Harris R, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, et al. Gibbon genome and the fast karyotype evolution of small apes. Nature [Internet]. 2014;513(7517):195–201. Available from: http://dx.doi.org/10.1038/nature13679

135. Damas J, O'Connor R, Farré M, Lenis VPE, Martell HJ, Mandawala A, et al. Upgrading short-read animal genome assemblies to chromosome level using comparative genomics and a universal probe set. Genome Res. 2017;27(5):875–84.

136. Lewin HA, Larkin DM, Pontius J, O'Brien SJ. Every genome sequence needs a good map. Genome Res. 2009;19(11):1925–8.

137. Hellmann I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M. A Neutral Explanation for the Correlation of Diversity with Recombination Rates in Humans. Am J Hum Genet [Internet]. 2003;72(6):1527–35. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0002929707604510

138. Wolf JBW, Ellegren H. Making sense of genomic islands of differentiation in light of speciation. Nat Rev Genet [Internet]. 2017;18(2):87–100. Available from: http://dx.doi.org/10.1038/nrg.2016.133

139. Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism

correlate with recombination rates in D. melanogaster. Nature [Internet]. 1992;356(6369):519–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1560824%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/1560824?ordinalpos=1&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_DefaultReportPanel.Pubmed_RVDocSum

140. Kulathinal RJ, Bennett SM, Fitzpatrick CL, Noor MAF. Fine-scale mapping of recombination rate in Drosophila refines its correlation to diversity and divergence. Proc Natl Acad Sci [Internet]. 2008;105(29):10051–6. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0801848105

141. Charlesworth D. Evolution of recombination rates between sex chromosomes. Phil Trans R Soc B [Internet]. 2017;372(1736):20160456. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29109220%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5698619%0Ahttp://dx.doi.org/10.1098/rstb.2016.0456

142. Stapley J, Feulner PGD, Johnston SE, Santure AW, Smadja CM. Variation in recombination frequency and distribution across Eukaryotes: patterns and processes. Phil Trans R Soc BPhil Trans R Soc B. 2017;In press.

143. Gante HF, Matschiner M, Malmstr??m M, Jakobsen KS, Jentoft S, Salzburger W. Genomics of speciation and introgression in Princess cichlid fishes from Lake Tanganyika. Mol Ecol. 2016;

144. Werren JH. Selfish genetic elements, genetic conflict, and evolutionary innovation. Proc Natl Acad Sci U S A [Internet]. 2011 Jun 28 [cited 2013 Oct 22];108 Suppl:10863–70. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3131821&tool=pmcentrez&rendertype=abstract

145. Frances E Clark, Conte M a., Ferreira-Bravo I a, Poletto AB, Martins C, Thomas D Kocher. Dynamic Sequence Evolution of a Sex-Associated B Chromosome in Lake Malawi Cichlid Fish. 2016;1–10.

146. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, et al. The medaka draft genome and insights into vertebrate genome evolution. Nature. 2007;447(7145):714–9.

147. Roberts NB, Juntti SA, Coyle KP, Dumont BL, Stanley MK, Ryan AQ, et al. Polygenic sex determination in the cichlid fish. BMC Genomics [Internet]. 2016;1–13. Available from: http://dx.doi.org/10.1186/s12864-016-3177-1

148. Mazzuchelli J, Kocher TD, Yang F, Martins C. Integrating cytogenetics and genomics in comparative evolutionary studies of cichlid fish. BMC Genomics. 2012;13(1).

149. Ichikawa K, Tomioka S, Suzuki Y, Nakamura R, Doi K, Yoshimura J, et al. Centromere evolution and CpG methylation during vertebrate speciation. Nat Commun [Internet]. 2017;8(1):1833. Available from: http://www.nature.com/articles/s41467-017-01982-7

150. Amores A, Catchen J, Nanda I, Warren W, Walter R, Schartl M, et al. A RAD-tag genetic map for the platyfish (Xiphophorus maculatus) reveals mechanisms of karyotype evolution among teleost fish. Genetics. 2014;197(2):625–41.

151. Takahashi K, Terai Y, Nishida M, Okada N. A Novel Family of Short

Interspersed Repetitive Elements (SINEs) from Cichlids : The Patterns of Insertion of SINEs at Orthologous Loci Support the Proposed Monophyly of Four Major Groups of Cichlid Fishes in Lake Tanganyika. Mol Biol Evol. 1998;15(4):391–407.

152. Takahashi K, Okada N. Mosaic structure and retropositional dynamics during evolution of subfamilies of short interspersed elements in African cichlids. Mol Biol Evol. 2002;19(8):1303–12.

153. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: From conflicts to benefits. Nat Rev Genet [Internet]. 2017;18(2):71–86. Available from: http://dx.doi.org/10.1038/nrg.2016.139

154. Santos ME, Braasch I, Boileau N, Meyer BS, Sauteur L, Böhne A, et al. The evolution of cichlid fish egg-spots is linked with a cis-regulatory change. Nat Commun [Internet]. 2014 Oct 9 [cited 2014 Oct 9];5:5149. Available from: http://www.nature.com/doifinder/10.1038/ncomms6149

155. Schulte JE, O'Brien CS, Conte M a, O'Quin KE, Carleton KL. Interspecific variation in Rx1 expression controls opsin expression and causes visual system diversity in African cichlid fishes. Mol Biol Evol [Internet]. 2014 Sep [cited 2015 Jan 4];31(9):2297–308. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24859246

156. Husemann M, Nguyen R, Ding B, Danley PD. A genetic demographic analysis of Lake Malawi rock-dwelling cichlids using spatio-temporal sampling. Mol Ecol. 2015;24(11):2686–701.

157. Albertson RC, Powder KE, Hu Y, Coyle KP, Roberts RB, Parsons KJ, et al. Genetic basis of continuous variation in the levels and modular inheritance of pigmentation in cichlid fishes. Mol Ecol. 2014;23(21):5135–50.

158. Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, et al. The genomic substrate for adaptive radiation in African cichlid fish. Nature. 2014;513(18 September 2014):375-.

159. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol. 2017;34(7):1812–9.

160. Franck JPC, Kornfield I, Wright JM. The utility of sata satellite dna sequences for inferring phylogenetic relationships among the three major genera of tilapiine cichlid fishes. Mol Phylogenet Evol. 1994;3(1):10–6.

161. Melters DP, Bradnam KR, Young H a, Telis N, May MR, Ruby JG, et al. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol [Internet]. 2013 Jan 30 [cited 2014 Jul 11];14(1):R10. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4053949&tool=pmcentrez&rendertype=abstract

162. Terai Y, Takahashi K, Okada N. SINE cousins: the 3'-end tails of the two oldest and distantly related families of SINEs are descended from the 3' ends of LINEs with the same genealogical origin. Mol Biol Evol. 1998;15:1460–71.

163. Waterhouse RM, Seppey M, Simao Neto FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene

prediction and phylogenomics. bioRxiv [Internet]. 2017;1–9. Available from: http://www.biorxiv.org/content/early/2017/08/17/177485

164. Dolgin ES, Charlesworth B. The effects of recombination rate on the distribution and abundance of transposable elements. Genetics. 2008;178(4):2169–77.

165. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods [Internet]. 2016;13(12):1050–4. Available from: http://dx.doi.org/10.1038/nmeth.4035

166. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. Complete assembly of parental haplotypes with trio binning. bioRxiv [Internet]. 2018;271486. Available from: https://www.biorxiv.org/content/early/2018/02/26/271486

167. Jain M, Koren S, Quick J, Rand AC, Sasani TA, Tyson JR, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol [Internet]. 2018;(January 2018). Available from: https://www.biorxiv.org/content/early/2017/04/20/128835

168. Novak AM, Hickey G, Garrison E, Blum S, Connelly A, Dilthey A, et al. Genome Graphs. doi.org [Internet]. 2017;101378. Available from: https://www.biorxiv.org/content/early/2017/01/18/101378

169. Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. Genome Res. 2017;27(5):665–76.

170. Eggertsson HP, Jonsson H, Kristmundsdottir S, Hjartarson E, Kehr B, Masson G, et al. Graphtyper enables population-scale genotyping using pangenome graphs. Nat Genet [Internet]. 2017;49(11):1654–60. Available from: http://dx.doi.org/10.1038/ng.3964

171. Dilthey A, Cox C, Iqbal Z, Nelson MR, McVean G. Improved genome inference in the MHC using a population reference graph. Nat Genet [Internet]. 2015;47(6):682–8. Available from: http://dx.doi.org/10.1038/ng.3257

172. Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. Centromere reference models for human chromosomes X and Y satellite arrays. 2014;697–707.

173. Sakamoto T, Danzmann RG, Gharbi K, Howard P, Ozaki A, Khoo SK, et al. A microsatellite linkage map of rainbow trout (Oncorhynchus mykiss) characterized by large sex-specific differences in recombination rates. Genetics. 2000;155(3):1331–45.

174. Moen T, Hoyheim B, Munck H, Gomez-Raya L. A linkage map of Atlantic salmon (Salmo salar) reveals an uncommonly large difference in recombination rate between the sexes. Anim Genet. 2004;35(2):81–92.

175. Roesti M, Moser D, Berner D. Recombination in the threespine stickleback genome - Patterns and consequences. Mol Ecol. 2013;22(11):3014–27.

176. Zeng Q, Fu Q, Li Y, Waldbieser G, Bosworth B, Liu S, et al. Development of a 690 K SNP array in catfish and its application for genetic mapping and validation of the reference genome sequence. Sci Rep [Internet]. 2017;7(October 2016):1–14. Available from:

http://dx.doi.org/10.1038/srep40347

177. O'Quin CT. Thesis Dissertation. 2014;

178. Charlesworth B. The evolution of sex chromosomes. Science (80-) [Internet]. 1991;251(4997):1030–3. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1998119

179. Lee B, Hulata G, Kocher TD. Two unlinked loci controlling the sex of blue tilapia (Oreochromis aureus). Heredity (Edinb). 2004 Jun;92(6):543–9.

180. Guyomard R, Boussaha M, Krieg F, Hervet C, Quillet E. A synthetic rainbow trout linkage map provides new insights into the salmonid whole genome duplication and the conservation of synteny among teleosts. BMC Genet. 2012;13:1–12.

181. Sevim V, Bashir A, Chin CS, Miga KH. Alpha-CENTAURI: Assessing novel centromeric repeat sequence variation with long read sequencing. Bioinformatics. 2016;32(13):1921–4.

182. Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel K V., Paten B, et al. Linear Assembly of a Human Y Centromere using Nanopore Long Reads. DoiOrg [Internet]. 2017;170373. Available from: https://www.biorxiv.org/content/early/2017/07/31/170373

183. Rastas P, Calboli FCF, Guo B, Shikano T, Merilä J. Construction of Ultradense Linkage Maps with Lep-MAP2: Stickleback F2 Recombinant Crosses as an Example. Genome Biol Evol. 2015;8(1):78–93.

184. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol [Internet]. 2004;5(2):R12. Available from: http://genomebiology.com/2004/5/2/R12

185. Nattestad M, Chin C-S, Schatz MC. Ribbon: Visualizing complex genome alignments and structural variation. bioRxiv. 2016;344:82123.

186. Houben A, Banaei-Moghaddam AM, Klemme S, Timmis JN. Evolution and biology of supernumerary B chromosomes. Cell Mol Life Sci [Internet]. 2013 Aug 3 [cited 2013 Nov 23]; Available from: http://www.ncbi.nlm.nih.gov/pubmed/23912901

187. Lee YCG, Langley CH. Transposable elements in natural populations of Drosophila melanogaster. Philos Trans R Soc Lond B Biol Sci [Internet]. 2010;365(1544):1219–28. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2871824&tool=pmcentrez&rendertype=abstract

188. Kofler R, Hill T, Nolte V, Betancourt AJ, Schlötterer C. The recent invasion of natural *Drosophila simulans* populations by the P-element. Proc Natl Acad Sci [Internet]. 2015;112(21):6659–63. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1500758112

189. Ocalewicz K. Telomeres in fishes. Cytogenet Genome Res. 2013;141(2–3):114–25.

190. Douglas RN, Birchler JA. B chromosomes. In: Chromosome Structure and Aberrations. 2017. p. 13–39.

191. Fantinatti BE a, Mazzuchelli J, Valente GT, Cabral-de-Mello DC, Martins C.

Genomic content and new insights on the origin of the B chromosome of the cichlid fish Astatotilapia latifasciata. Genetica [Internet]. 2011 Oct [cited 2013 Mar 26];139(10):1273–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22286964

192. Myosho T, Otake H, Masuyama H, Matsuda M, Kuroki Y, Fujiyama A, et al. Tracing the emergence of a novel sex-determining gene in medaka, Oryzias luzonensis. Genetics [Internet]. 2012 May [cited 2013 Apr 2];191(1):163–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22367037

193. EMBL, SIB Swiss Institute of Bioinformatics, Protein Information Resource (PIR). UniProt. In: Nucleic acids research. 2013. p. 41: D43-D47.

194. The UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res [Internet]. 2015;43(Database issue):D204-12. Available from: http://nar.oxfordjournals.org/cgi/content/long/43/D1/D204

195. Makunin AI, Dementyeva P V., Graphodatsky AS, Volobouev VT, Kukekova A V., Trifonov VA. Genes on B chromosomes of vertebrates. Mol Cytogenet. 2014;7(1):1–10.

196. Vannier J-B, Sandhu S, Petalcorin MIR, Wu X, Nabi Z, Ding H, et al. RTEL1 is a replisome-associated helicase that promotes telomere and genome-wide replication. Science [Internet]. 2013 Oct 11 [cited 2014 Jul 15];342(6155):239–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24115439

197. Ballew BJ, Joseph V, De S, Sarek G, Vannier JB, Stracker T, et al. A Recessive Founder Mutation in Regulator of Telomere Elongation Helicase 1, RTEL1, Underlies Severe Immunodeficiency and Features of Hoyeraal Hreidarsson Syndrome. PLoS Genet. 2013;9(8).

198. Siderakis M, Tarsounas M. Telomere regulation and function during meiosis. Chromosom Res. 2007;15(5):667–79.

199. Crabben SN Van Der, Hennus MP, Mcgregor GA, Ritter DI, Nagamani SCS, Wells OS, et al. Destabilized SMC5 / 6 complex leads to chromosome breakage syndrome with severe lung disease. J Clin Investig. 2016;126(8):2881–92.

200. Poletto AB, Ferreira I a, Martins C. The B chromosomes of the African cichlid fish Haplochromis obliquidens harbour 18S rRNA gene copies. BMC Genet. 2010;11:1.

201. Martis MM, Klemme S, Banaei-Moghaddam AM, Blattner FR, Macas J, Schmutzer T, et al. Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. Proc Natl Acad Sci U S A [Internet]. 2012 Aug 14 [cited 2013 Apr 2];109(33):13343–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3421217&tool=pmcentrez&rendertype=abstract

202. Klemme S, Banaei-Moghaddam AM, Macas J, Wicker T, Novák P, Houben A. High-copy sequences reveal distinct evolution of the rye B chromosome. New Phytol [Internet]. 2013 Jul;199(2):550–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23614816

203. Valente GT, Mazzuchelli J, Ferreira I a, Poletto a B, Fantinatti BE a, Martins

C. Cytogenetic mapping of the retroelements Rex1, Rex3 and Rex6 among cichlid fish: new insights on the chromosomal distribution of transposable elements. Cytogenet Genome Res [Internet]. 2011 Jan [cited 2013 Apr 2];133(1):34–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21196713

204. Böhne A, Brunet F, Galiana-Arnoux D, Schultheis C, Volff JN. Transposable elements as drivers of genomic and biological diversity in vertebrates. Vol. 16, Chromosome Research. 2008. p. 203–15.

205. Zeh DW, Zeh JA, Ishida Y. Transposable elements and an epigenetic basis for punctuated equilibria. BioEssays. 2009;31(7):715–26.

206. Wilkins AS. The enemy within: An epigenetic role of retrotransposons in cancer initiation. Vol. 32, BioEssays. 2010. p. 856–65.

207. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M. Accurate detection of complex structural variations using single molecule sequencing. bioRxiv. 2017;1–24.

208. Nattestad M. copycat [Internet]. 2016. Available from: https://github.com/MariaNattestad/copycat

209. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.

210. Cook DE, Andersen EC. VCF-kit: Assorted utilities for the variant call format. Bioinformatics. 2017;33(10):1581–2.

211. Li H. Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016;32(14):2103–10.