Broad characterization of structural variation and genetic differentiation in two hybridizing macaque species

by

Chase Arey Kirkland Rushton

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
August 4th, 2018

Keywords: *Hybridization, Speciation, Whole Genome Sequencing, Population Genetics, Structural Variation*

Approved by

Laurie Stevison, Chair, Assistant Professor of Biological Sciences
Leslie Goertzen, Associate Professor of Biological Sciences
Tonia Schwartz, Assistant Professor of Biological Sciences

Thesis Overview

Species are the fundamental units of biological diversity; speciation itself is the engine of biodiversity. Therefore, understanding speciation patterns and mechanisms is paramount to evolutionary biology. Evolutionary genetics is a subset of biology spurred on by the modern synthesis movement in the 1930s that focuses on what evolutionary factors produce the diversity that we see on earth; three of these important facets are speciation (the formation of new species), natural selection (pioneered by Charles Darwin, the process where organisms adapt to their environment) and common decent (starting from a single common ancestor). Current research has incorporated not only the fundamental pillars of biology, evolution, developmental biology and systematics, it has also incorporated molecular genetics and computer science to increase our scientific gains. Here I seek to better understand the underlying processes and genomic information potentially important to speciation in a group of macaques broadly distributed across Southeast Asia and known to hybridize in an overlapping geographic range. Speciation generates all the biodiversity on the planet, but what keeps them from merging all back together? Defining what a species is and what allows species to persist is a controversial topic. Commonly, a species is split into two when some geographic barrier presents itself, such as a mountain range, a river formation, or even the building of a railway. This new barrier causes reduction of interbreeding between the two populations and over time two distinct groups develop. If the two populations are unable to interbreed when later brought together, this is referred to as allopatric speciation. However, what occurs when species that are not bound by

geographically isolating barriers and share overlapping ranges? The conundrum of what maintains a species once they form is a popular research topic. Here, we provide mix of traditional wet lab work with an *in-silica* approach to provide a broad overview of genomic structural variation between *Macaca mulatta* and *Macaca fascicularis*, combining five structural variant programs in an attempt gain confidence in variants calls that are shared or unique to each respective species. In conjunction with the structural variation analysis, a small-scale population genomics analysis was performed to identify areas of high divergence in the genome between these two species. Finally, unique structural variants and areas of high divergence driven by intraspecies not interspecies differences were compared to see if regions of high divergence underlie areas of structural variation. The results of this study can provide a basis for further research into the genomic structure of macaques, helping to understand primate evolution and genomic structure.

Acknowledgments

First, I would like to thank my advisor, Dr. Laurie Stevison, for her support, guidance, and patience throughout my studies. Next I would like to thank my committee Dr. Tonia Schwartz and Dr. Leslie Goertzen for their guidance, input, support and tempering of my wild ambitions throughout my project. I would also like to thank my lab mates who have helped with brainstorming ideas, providing feedback and edits on my thesis, and support during my time at Auburn University. Thank you to all of my friends, both within graduate school and outside of graduate school for your constant support and understanding, you all kept me grounded during my graduate studies, and I honestly couldn’t have done it without all of you! Last but not least, thank you to my parents Dr. Lyn Kirkland Rushton and Ian Rushton, for giving me the foundation and support that allow me to succeed. Without you I could have never made it this far. Finally, I would like to thank Auburn University for funding my research.

Table of contents

Thesis Overview ii

Acknowledgments iv

Abstract 1

Background 2

1) Identification and comparison of shared and unique structural variants 9

2) Identify areas of high divergence between *M. mulatta* and *M. fascicularis.* 10

3) Do areas of high divergence match areas of areas of structural variation? 10

Materials and Methods 12

Identification and comparison of shared and unique Structural variants 12

Chromosome 5 analysis 17

Identify areas of high divergence between *M . mulatta* and *M . fascicularis* 18

Do areas of high divergence match areas of areas of structural variation? 23

Results 24

What type of structural variants occur between and within species? 24

What regions of the genome harbor high levels of diversity and divergence. 27

Do areas of high divergence match areas of SV calls? 30

Discussion 32

Identification and comparison of shared and unique Structural variants 32

Lack of empirical validation of chromosome 5 putative inversions 35

Areas of high divergence between *M. mulatta* and *M. fascicularis* highest on the X chromosome 36

Areas of high divergence overlap SVs more often than expected by chance. 39

Summary 40

Figures 42

References 70

Appendix 75

M. mulatta BAMQC 75

M.fascicularis BAMQC 81

Table of figures

Figure 1: Geographic ranges for M. mulatta and M. fascicularis. Range of M. mulatta can be seen in green spanning India and China, M. fascicularis in teal, and overlapping zone of hybrid alleles in purple. (Adapted from Fooden 1980). 42

Figure 2:Diagram of potential structural variant signatures analyzed via computational methods: reference and query sequences can be seen with respective variants in red. 43

Figure 3:Overview of genome preparation pipeline from raw genome file download with Sratoolkit to merged structural variant analysis. Pipeline primarily followed GATK best practices with minor revisions. 46

Figure 4::Comparison of reference genomes between rheMac8 and macFas5 created with lastZ, alignments were filtered for 95% mapping quality and plotted using ggplot2. 49

Figure 5 :UCSC genome browser custom track of putative 15mb inversion from 42mb-57mb not found in M. mulatta. Whole inversion can be viewed in(top) and the left and right breakpoints respectively below. Breakdancer can be seen in blue, Delly in green and Lumpy in red 50

Figure 6:Weir and Cockrams Fst calculated between M. mulatta and M. fascicularis calculated at varying bin sizes(50kb,250kb,500kb,1mb & 1.5mb). 50kb bin size was chosen for further analysis to reduce confounding effect of F_{ST} outliers on the X chromosome. A 1% cutoff was added to filter only the regions with the greatest outliers. A- 50kb window , B – 250kb window , C- 500kb window, D – 1Mb window , E- 1.5mb window... 52

Figure 7: (A) Pairwise nucleotide divergence was calculated for M . fascicularis .Areas of low PI were extracted to be used alongside areas of high FST. PI was plotted along the genome using R and is bound from 0<->1. (B) Pairwise nucleotide divergence for M. mulatta . (C) Tajimas D calculated in 50kb bin size using Vcftools for M . fascicularis , visualized using R, Abline at 0 as negative values were ones of interest. (D) : Tajimas D calculated in 50kb bin size using Vcftools for M. mulatta , visualized using R, Abline at 0 as negative values were ones of interest.......... 53

Figure 8:: Karyoplot showing where areas of High FST , low PI and negative Tajima's D lie across the genome. The massive amount of overlap on the X chromosome may hint at a faster X effect in this syste.. 54

Figure 9:dN/dS between M. mulatta and M. fascicularis, comparisons of Autosomes, PAR1, PAR2 and X. Sample size and statistical grouping can be seen above, with autosome and PAR1 grouping out. 61

Figure 10: M. fascicularis large structural variants were plotted with KaryoploteR in R , SVs along the genomes in the large bin size(10kb-1mb) can be seen respectively. Additionally, F_{st} outliers were plotted on top of the SV dataset for an overall view of overlap. 62

Figure 11: M. fascicularis Extra-large variants filtered for 3/5 match overlaid with areas of high FST.. 63

Figure 12: M. fascicularis Medium variants filtered for 3/5 match overlaid with areas of high FST.. 64

Figure 13: M. fascicularis small variants filtered for 3/5 match overlaid with areas of high FST 65

Figure 14:M. mulatta Extra-large variants filtered for 3/5 match overlaid with areas of high FST 66
Figure 15: M. mulatta Large variants filtered for 3/5 match overlaid with areas of high FST..... 67
Figure 16: M. mulatta Medium variants filtered for 3/5 match overlaid with areas of high FST 68
Figure 17:M. mulatta small variants filtered for 3/5 match overlaid with areas of high FST 69

Table of Tables

Table 1:Publicly available whole genome data used in this study. In total, 12 samples were used from a broad geographic range, including six from the extremely isolated Republic of Mauritius. 44
Table 2: Raw structural variant calls before intersection and filtering from Breakdancer, Lumpy, Delly, Pindel, and CNVnator for both M. mulatta and M. fascicularis. 47
Table 3: Merged structural variant calls after filtering. Variants were filtered for size into bins of small (<1kb), medium (1 kb - 10 kb), large (10 kb - 1 Mb) and extra-large (>1Mb) and at least three program hits. A drastic reduction in SV calls was noted after intersection. M. mulatta and M. fascicularis were used respectively. 48
Table 4: Comparison of Structural variation call set between the aligners BWA and Bowtie2 provides an insight into differing algorithms and potential downstream bias. 51
Table 5: Ten top and bottom gene family functions reported from DAVID, Top 1% FST outliers can be seen on the left, and the bottom 1% outliers can be seen on the right. Inputs were region matched and controlled against the whole genome. 55
Table 6:Table 6: Top 1 percent gene ontology outputs from DAVID gene ontology.................. 56
Table 7: Bottom one percent gene families from DAVID gene ontology................................... 58
Table 8: Gene ontology analysis of unique structural variant regions of M. mulatta and M. fascicularis conducted with DAVID gene ontology. 61

Abstract

Speciation and hybridization drive all of the biodiversity seen on the plant. Understanding underlying mechanisms of these cornerstones of evolutionary genetics if of vital importance. Here we seek to apply a combination of *in silica* methods to characterize at the genomic level a group of hybridizing macaques broadly distributed across southeast Asia. Using a mixture of structural variation programs and population genetics we aimed to better understand this system. Here we establish the need to run multiple variant detection programs and intersect the outputs to attain a high quality variant list. Additionally, evidence was found supporting faster X evolution in these two macaques. Finally, rates of variation and divergence were found to overlap more often than by chance at a rate of 81%. This research can provide a vast foundation for further research into primate evolution and macaque genomics and genetics.

Background

Genomes are a mosaic of hybridization, speciation, selection and mutation. By examining an organism's genome, we can infer what contributed to making it unique in the grand scheme of the tree of life. With the advent of genome wide data sets it is now possible to look at variation among and between species not only on a massive scale, but allowed for sequencing faster and to more depth, but also with increased resolution. The advancement in sequencing technologies continues to provide higher resolution pictures of genomes with increased accuracy in the base calls. This is of vital importance to a researcher interested at the genomic level to have not only the whole picture, or as much as possible of the genome, but also confidence that the base calls are correct, as sometimes the difference could be a single base pair you are looking for. The genomic era has allowed us to start getting a basis of understanding on all the genetic diversity on the earth. One of the central goals of evolutionary biology research is to elucidate these underlying mechanisms (evolution, selection, drift) in an attempt to understand how the known world functions. Additionally, the more we understand about genomics in well studied systems, will lay the foundations for best practices that can then be applied to organisms that are not well defined.

Species concepts were largely championed by Ernst Mayr. He described species as groups of breeding natural populations which are reproductively isolated from other such groups. This concept is now commonly known as the biological species concept (BSC) and is one of the leading views of species concepts out there (Aldhebiani, 2018). It is most easily defined as a group of reproducing natural populations incapable of effectively mating or breeding with other such groups, and which inhabits a particular niche in nature (Bisby, Bisby, & Coddington; Mayr, 1982) . The BLC however, does not apply to either asexual organisms or those that are in allopatry (isolated by some geographic barrier).

At the genetic level, speciation is the genome diversifying via selection, mutation, and drift while at the same time homogenizing through migration and recombination elsewhere along the genome. Understanding the origin of species and what allows them to persist requires an understanding of what caused the barriers between species to arise. One method of breaking down species barriers is the occurrence of hybridization via the migration of new alleles into a population. Hybridization is defined as the mating between genetically or phenotypically distinguishable species. This can have a large variety of evolutionary outcomes: species could merge into one population; genomic information could be asymmetrically shared or symmetrically shared. Additionally hybrid inviability, has been witnessed in tetrapods, mammals and reptiles (Prager & Wilson, 1975; Wilson, Maxson, & Sarich, 1974), and hybrid sterility such as Mules or Ligers.

Hybrid inviability occurs after pre-zygotic barriers are overcome, or when no pre-zygotic barriers exist, and a zygote is produced from two mating species. More often than not, the embryo dies before birth due to conflicting genes from the parents. Mammal species have been shown to only produce hybrids that remain reproductively viable for 2-3 million years after a speciation event (Wilson et al., 1974). This phenomenon is much larger than other species ability to remain reproductively viable intraspecies after a hybridization event. Two hypotheses were proposed to potentially explain this nuance of mammalian life history and why there is a faster rate of hybrid in-viability compared to other groups. These hypotheses are the (1) regulatory and (2) immunological hypothesizes, both of which have found support in the scientific community as viable options to explain the propensity for mammals to have faster rate of inviability. Additionally, hybridization has other effects on a system such as increasing or decreasing the diversity within the populations. However, hybridization can rescue small populations of inbred species by giving an influx of non-deleterious alleles (Frankham, 2015). Finally, hybridization can reduce diversity by breaking through the reproductively isolating blocks of the genome potentially allowing for the merging of two previously independent lineages. Hybridization between species is common in plants, however in animals it was previously thought to be unusual and rare. Hybridization and introgression are the opposite of reproductive isolation, which challenges the reality of the biological species complex. Here I present work on a system that has secondary gene flow and asymmetric hybridization using next generation sequencing data.

Here, two macaque species will be the focal groups in this thesis, *Macaca mulatta* otherwise known as the rhesus macaque and *Macaca fascicularis* known as the crab eating macaque. Old world monkeys including the rhesus and the crab eating macaque are the most widespread radiation of primates besides humans themselves. *Macaca* comprises more than 20 species comprising four groups (*Sylvanus* , *Silenus* , *Sinica* and *Fascicularis*) that diverged 5-6 million years ago (J Fooden, 1983). Macaques share a last common ancestor (LCA) with humans about 25 million years ago (Fleagle 2013). This makes them crucial to both biomedical and primate evolution studies as they provide not only vast biomedical functionality, but through their diversity with 19 extant living species are ecologically significant as well. *M. mulatta* have a lifespan of around 25 years with a population estimate of 6/km and a broad geographic range

(Figure 1). They have an average height of 500mm and weight of 7.7kg (Cawthorn 2005). Rhesus macaques are divided according to their country of origin, they are often referred to as Chinese or Indian derived. In primate centers in the United States, macaques are sometimes bred according to their country of origin, however crossbreeding has occurred and this can cause some confusion in the taxonomic separation of individuals (Smith & Mcdonough, 2005) .They have a range of colours from dusty brown to auburn, and are sexually dimorphic, males range from 531.8mm and weigh on average 7.7kg where females average 468mm and 5.34kg (Jack Fooden & Fooden, 2000).

In contrast, *M. fascicularis* has an average life span of 31 years, also an unknown population with a broad geographic range (Figure 1). They have an average height of 400-600mm and a weight of 4.7-8kg (Jack Fooden & Fooden, 2000). Captive stocks are known to hybridize and can lead to erroneous estimations of gene flow. Since these captive stocks are also used for biomedical research, understanding these systems is an important facet for biology. Genetic background in these model organisms could potentially be mixed depending on how they are kept. Previous work in these species has shown asymmetrical hybridization from *M. mulatta* into *M. fascicularis* (Singh & Sinha, 2004). These two species have an estimated divergence time of 1-2.5 million years ago (Stevison & Kohn, 2009).The geographic ranges of *M. mulatta* and *M. fascicularis* adjoin in Indochina where they have a zone of hybridization. Gene flow is restricted to mainland Indochina (Figure 1).

In most plants and animals, hybrid sterility acts as a post fertilization barrier to hybridization (Kirkpatrick & Barton, 2006). The genetic background of *M. fascicularis* is made complex by the high genetic diversity and population structure. In combination with gene introgression from the *M. mulatta* , *M. fascicularis* has a unique genomic background (Higashino et al., 2012).

Previous karyotype evidence has shown that among 20 papionini species (macaques, baboon, mandrill), encompassing 62.4 million years of evolution, there is remarkable conservation of karyotype evolution (Stanyon, Fantini, Camperio-Ciani, Chiarelli, & Ardito, 1988). However, this study found a difference in *M. fascicularis* compared to the other 19 species. It was proposed to possibly have arisen from either a packaging difference (tightness of chromatin binding) or a chromosomal structural variant (SV). Structural variants are generally defined as variation in the structure of an organism's genome or chromosome. These variations can consist of deletions, duplications, copy number variants, inversions, insertions and translocations. The variants affect a sequence area larger than a single nucleotide polymorphism, but smaller than a chromosomal abnormality (Feuk 2006). Structural variants, like SNPS can be associated with genetic disorders and are more difficult to accurately detect than SNPs (Sudmant 2015). Structural variants of deletions (DEL), insertions (INS), duplications (DUP) and inversions (INV) and translocation (CTX/ITX) were the focus of this study. A deletion is an event where the DNA is removed, either a single base pair or a series of nucleotides, and the two ends flanking this excision fuse back together. In contrast to deletions, insertions occur when one or more nucleotides are added into the genomic sequence a special kind of insertion is a duplication, where a segment of the genome is duplicated and either resides next to the original copy (tandem duplication) or elsewhere in the genome. Translocations come in two types intra-chromosomal and inter-chromosomal. In the former, a section of the genome breaks off and is relocated onto the same chromosome, while the latter has that break relocate to a region of the genome on a different chromosome. Finally, inversions occur when a sequence of DNA breaks, inverts its orientation and lands back in the same position. Any combination of these variants can occur in the genome, and lead to complex genomic shuffling (Holland & Cleveland, 2012).

SV’s, such as an insertion, deletion, inversion, or translocation, are common in genomes, and like single nucleotide polymorphisms can act as genetic markers. While serving as genetic markers, SV’s have also been implicated as contributing factors to speciation (Feulner & De-Kayne, 2017). Insertions or deletions may occur near promoters/enhancers/repressors facilitating speciation, or a translocation may occur causing genes to shuffle to another chromosome and be under a different set of genomic controlling factors. Next generation sequencing (NGS), has become the primary way of identifying SVs from the genome and evaluating their effect on the individual or populations. However our current methodology of sequencing short (~100bp) reads in parallel from randomly fragmented copies of a genome with NGS technology is limited in large scale structural variant detection (Mardis, 2008; Metzker, 2010). Thus, sophisticated computational methods are needed to accurately depict if a variant is real or not. Other methods of detecting structural variation can include PCR validation of either breakpoints or small (>1kb) variants that could then be sent off for sequencing. Additionally, visualization of structural variants can be done using FISH (fluorescent in situ hybridization) and numerous other methods.

Differing structural variant algorithms utilize different features of the sequence set correlating to each SV type, and each has their own strengths and weaknesses. Typically, structural variant programs identify breaks in the alignment. When the millions of NGS reads are aligned to a reference genome, signatures of misalignment are used to infer structural differences between the sample and the reference. Each type of SV has its own signature, which are summarized in (Figure 2) (Liu et al., 2015).

While there are numerous methods available all using complex computational algorithms to detect variants from genomic datasets, five programs were selected in this study. While next generation short read sequencing has generated a magnitude of data, and in turn, a plethora of variant detection programs exist, none of them are comprehensive or accurate enough to run alone. Therefore we used five in this study (Breakdancer (Chen et al., 2009), Lumpy (Layer, Chiang, Quinlan, & Hall, 2014), Delly (Rausch et al., 2012), CNVnator (Abyzov, Urban, Snyder, & Gerstein, 2011), and Pindel (Ye, Schulz, Long, Apweiler, & Ning, 2009)). This combined approach allowed for a variety of detection methodologies and detection bias (depending on the algorithms, some programs are more stringent than others) were taken into account.

Some structural variants can potentially be linked back to speciation and hybridization by capturing locally adapted alleles when two populations are hybridizing. These could be genes such as ones related to phenotypic characteristics, immunological differences (such as MHC or RH factors), it has been shown that immune response genes might cluster in areas of genomic rearrangement in the macaque lineage. Additionally, some pathway underlying ecological speciation causing reproductive isolation could be implicated.

Potentially *M. mulatta* and *M. fascicularis* have existed as two different gene pools and species because SVs are driving linkage disequilibrium or some other factor between them, allowing recombination to homogenize the genome outside these blocks, however keeping the species individuality intact. Another factor to be considered is that genomic variants of all sizes and types can contribute to genetic disease or a deleterious phenotype, providing potential substrates for selection to act upon, resulting in a phenotypic difference between individuals or populations.

While all structural variants will be analyzed, previous work in the literature has highlighted the importance of inversions in speciation (Kirkpatrick & Barton, 2006; McGaugh & Noor, 2012; Stevison et al., 2011), as such, they were the initial focus of the project. Inversions have been shown to capture locally adapted alleles when two populations are hybridizing. It is possible that *M. mulatta* and *M. fascicularis* might exist as two distinct gene pools because inversions have been the driving force of LD (linkage disequilibrium) between them, allowing recombination to homogenize these two species outside inversions. There is complete recombination suppression up to 2.5mb outside inversion breakpoints shown in *Drosophila* (Laurie S. Stevison et al., 2011). As such, regions of high divergence could match with regions of SVs, hinting at variant regions of the genome and vice versa. Inversions may facilitate speciation by protecting locally adapted alleles inside the inversions from gene flow with other populations, allowing further divergence between species. Creating LD in ways besides reproductive isolations (inversions reducing recombination) facilitates species persistence. However, inversions are not an absolute block to recombination, so selection/local adaptation should also be implicated.

When analyzing structural variation between *M. mulatta* and *M. fascicularis,* it is important to understand the population level differences between them. As such, measures of F_{ST} (Weir & Cockerham, 1984), Tajima's D (Tajima, 1989) and π (Nei & Li, 1979), and dN/dS were analyzed. The genetic measures of both divergence and diversity will help characterize this system. In combination with these statistical measures of divergence, diversity, and selective pressures, gene categories were also assayed. Measuring the underlying gene categories is often beneficial. It's one thing to know what regions of the genome are under selection or are differentiated but knowing what lies in those regions can also be of vital importance.

Here I have three key research objectives

1) <u>Identification and comparison of shared and unique structural variants</u>

Where do structural variants occur along the genome in both *M. mulatta* and *M. fascicularis*?

To achieve this, I used a combination of five structural variant programs: Breakdancer, Lumpy, Delly, CNVnator and Pindel and publicly available whole genome sequences of *M. mulatta* and *M. fascicularis*. These programs were used in conjunction with a custom script to merge structural variant calls providing a high confidence list of variation across both species genomes. The expectation is that there would be more shared variants than unique variants, due to the recent divergence of these two-macaque species. Additionally, I expected that the number of SV calls would reduce as the stringency of the approaches increased. Unique structural variants were the target of a gene ontology analysis to gather information on the gene families within the variants.

2) Identify areas of high divergence between *M. mulatta* and *M. fascicularis.*

Next, areas of divergence between *M. mulatta* and *M. fascicularis* were analyzed using a population genetics approach. F_{ST}, Tajima's D, pairwise nucleotide divergence(π), and D_nD_s was calculated. Combining these measurements of differentiation further elucidated underlying genetic differences of this unique system allowing for the attribution of genetic divergence to be given one species or the other respectively. Here regions were attributed to *M. fascicularis* only, however addition of *M. mulatta* filtering could further refine the system. Regions that are highly divergent between *M. mulatta* and *M. fascicularis* were the target of another gene ontology analysis to understand the underlying gene categories in these divergent regions. The expectations from the GO analysis were follows: if sex or immunological related gene families are found, then these regions are important to reproductive isolation or creating LD blocks in speciation. However, if other gene families are found, relating to geographic or differences in diet/habitat one could attribute these factors to being important in speciation.

3) Do areas of high divergence match areas of areas of structural variation?

Finally, structural variation along the genome will be compared to areas of high divergence. While it will not be possible to say if the structural variation is causing the high divergence or the divergence is causing the structural variation, it will be possible to report if they are overlapping more often than expected by chance. Additionally, the results from the structural variation gene ontology analysis will be compared to that of the divergence gene ontology analysis. This will elucidate any common gene families present in both sets of data. We expect to find SVs correlating with divergent regions and potentially underlying important GO terms for speciation. Using both *in silica* and laboratory work we hope to elucidate some driving factors of speciation and the maintenance of species.

Materials and Methods

Sample Information

All genomic samples were publicly available and found on NCBI/EBI databases. Raw reads were downloaded, analyzed independently from the original work (Table 1).

Sample Verification

Mitochondrial sequences were extracted from each genomic sample to verify its geographic origin. Following protocol previously outlined in (L.S. Stevison & Kohn, 2008).

DNA samples for PCR analysis

DNA samples were acquired from Oregon National Primate Research Center. Samples ranged in origin from Cambodia, Philippines, Vietnam, Indonesia, Mauritius, China and India. Overall 20 samples were procured 10 *M. mulatta* and 10 *M. fascicularis* (Table 2).

Identification and comparison of shared and unique Structural variants

Genome analysis

Genomic samples raw data were downloaded from NCBI using SRA-toolkit version v2.8 (Leinonen, Sugawara, Shumway, & International Nucleotide Sequence Database Collaboration, 2011). Files were then aligned using BWA mem version 0.7.12 (H. Li & Durbin, 2009) to the rhemac3 reference genome (UCSC edition) with options M -v 2 -t 4. Newly aligned sam (sequence alignment map file) files were sorted with samtools v1.2 (H. Li et al., 2009) and converted to bam files (binary sam files). Next, read group information was added with Picard tools v1.79 (DePristo et al., 2011) to standardize statistical error groups for downstream analysis. As files are often uploaded as multiple sequence files, after alignment individual bam files were merged into a complete bam file using Picardtools and sorted with Picard tools. Next duplicates were marked with Picardtools. GATK version 3.6 (DePristo et al., 2011) was used for local indel realignment. Finally, GATK base quality score realignment (BQSR) was performed using GATK. First variants were called with GATK haplotype caller. Basic variant stats were calculated with VCF tools version 1.14-14 (Danecek et al., 2011) - site quality, depth and site depth - to filter variants. Next VCF tools was used to filter the variants for quality, min and max depth, min and max quality 2 standard deviations past the mean respectively. The next step was to create the pre- and post- recalibration table using GATK base recalibrator. Plots of the recalibrations were created using GATK analyze covariates. Filtered variants were applied to the bam file using GATK print reads. Haplotype caller was run and VCF compare was used to calculate the percent of sites called in both datasets, variants were called, filtered and applied until VCF compare revealed that less than 1% new variants were called to ensure confidence in the recalibration. For my analysis BQSR was run 3 times for each sample to converge at this 99% similarity cutoff. A sample pipeline is pictured in (Figure 3).

Structural Variation detection and intersection

M. mulatta (SRA023856) and *M. fascicularis* (SRA023855) were used with five structural variant programs. Initially, these were the only samples in the project, after the addition of new publicly available sequences, the assumption was made that SVs are relatively conserved within species and as such, one representative from each species could be used to reduce computational run time and rerunning all analysis pipelines. Breakdancer (v1.3.6), Lumpy (v1.0), Delly (v0.7.7), CNVnator (v0.3.2) and Pindel (v0.2.5b9) were chosen for this analysis. Breakdancer is a genome wide structural variant detector using paired end NGS reads. It predicts five types of SVs: Insertions, Deletions, Inversions, Inter and Intra chromosomal translocations. Breakdancer works by analyzing short paired-end sequencing reads using read pairs that are mapped with unexpected separation distances or orientation. As such insert size of your library can play a role in breakdancers efficiency. Thus, large insert size genomic samples were preferred for structural variation analysis. The output from breakdancers (ctx file) was converted into an extended bed file format for intersection with the other programs. The next program, Delly is an integrated structural variant prediction method that can discover, genotype and visualize deletions, tandem duplications, inversions and translocations at single-nucleotide resolution in short-read NGS data .Delly uses paired-ends, split-reads and read-depth to detect structural variants along the genome. Delly output, like that of Breakdancer was converted from VCF to an extended bed format for intersection. The third program, Lumpy integrates multiple signals to give higher accuracy on low coverage samples. It can use information generated from read alignments or prior evidence, and that can readily adapt to any additional source of evidence that may become available with future technological advances. Lumpy output was converted into an extended bed format for intersection from its source VCF output. Initially, these three programs were used. However, after the failure to empirically validate a putative chromosome 5 variant, two additional programs were added. The first of these was CNVnator, which uses a combination of split-read and read-pair approaches using mapping density. CNVnator is able to discover CNVs in a vast range of sizes, from a few hundred bases to megabases in length, to the whole genome. CNVnator output was converted into an extended bed format for intersection. Finally, Pindel was used, which can detect breakpoints of deletions, insertions, inversions and duplications at the single base pair resolution from NGS datasets. It uses a pattern growth approach to identify breakpoints from paired end reads. Pindels output was converted into an extended bed format for intersection. A diagram on structural variant discovery can be seen in (Figure 2).

To get a high quality variant list, a custom python script (https://github.com/aubcar/Masters_Work) was written to parse each unique SV output into a BEDPE file. The script uses object-oriented programming to take SV outputs from any of the five programs, and potentially other SV programs. It parses the output into a uniform usable format and outputs both a bed and extended bed format for each SV type. Length is recalculated or verified, as not all programs accurately calculate length. Additionally, a column is added into the output to tell which program was used. Next, the tool multiIntersect bed from bedtools version 2.27 (Quinlan & Hall, 2010) was used to take the intersection of all programs. These five program sets were comprehensively filtered into one "high" quality SV set which have higher confidence in the putative variant calls being true positives than a single program.

It is important in SV analysis to use large insert size paired end libraries, as these give you the highest power to detect SV's. However, while the power of current NGS technologies is the billions of short reads generated, part of the issue is the promiscuity of mapping. Reads in repetitive regions are not often mapped correctly, as such, a masked (repetitive and low-quality regions removed). While paired end sequencing has its advantages, the accuracy in SV detection is highly correlated to the insert size of the library or read length, small insertion or deletion events are often missed (Liu et al., 2015). Even if larger insert sizes and multiple sequencing libraries are used, the issue of multiple mappings remains for longer repeat elements that number in the millions in primate genomes (Liu et al., 2015).

Variants validated in 3/5, 4/5 and 5/5 of the programs were analyzed. Outputs from the intersection were binned by size, type and number of structural variant program calls in accordance. Shared and unique variants between *M. mulatta* and *M. fascicularis* were determined by using bedtools intersect and filtering for ones that overlapped in both call sets (shared) and ones that did not (unique). To be considered shared, windows of reported SV's had to significantly overlap.

Alignment Bias

Different alignment tools or different parameter settings of the same tool will result in different alignment results (Heng Li & Homer, 2010; Ruffalo, Laframboise, & Koyutürk, 2011), which will impact the performance of SV detections. This difference in alignment algorithms was taken into account via a test of BWA (H. Li & Durbin, 2009) and Bowtie2 (Langmead & Salzberg, 2012). Raw reads for both species were aligned to rhemac3 and processed via BWA and Bowtie2 as described in section "Genome analysis" above through the step of indel realignment. After this, SVs were called with the quickest of the programs to run Breakdancer. Number of variants called, average number based on size category, quality, and supporting reads were also analyzed.

Reverse complement SV analysis using macFas5

Genomic samples were aligned to the macFas5 (UCSC) reference genome using BWA mem. As described in "Genome analysis". Breakdancer was run to see if the indels recovered corresponded to deletions in the mapping of reads to the rhemac3 reference. Additionally, comparison of reference genomes (rheMac3, rhemac8, macFas5) was performed.

Chromosome 5 analysis

Breakpoint refinement

PCR primers were designed using Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/) with the right breakpoint of the putative 15mb inversion on chr5. Overall eight pairs of primers were developed for PCR validation of the chr5 inversion identified.

De novo assembly to refine breakpoints

Three structural variant programs predicted a putative 15Mb inversion from 42Mb to 57Mb on chromosome 5 of *M. fascicularis* that was not present in *M. mulatta.* The region was expanded by 2.5mb proximal to each side and a fasta sequence of the region was then extracted from the aligned BAM file. Reads within this 20mb region were then extracted using Samtools version 1.3.1 for a *de novo* assembly. Extracted reads were split into forward and reverse complement reads using a combination of grep and awk. The extracted reads were then assembled with Ray (Boisvert, Raymond, Godzaridis, Laviolette, & Corbeil, 2012) at four Kmer values (k=22,25,28,31).

Breakpoint refinement utilizing blast

Contigs from the Ray *de novo* assembly (696k total) were blasted against a custom blast database of *M.mulatta* rhemac3 chromosome 5 in an attempt to refine the breakpoint of the 15mb putative inversion using NCBI BLAST (Altschul et al., 1997). A custom python script was written to filter blast hits with more than one hit in more than one direction. A positive control training dataset was generated to validate the program and its ability to recover a putative breakpoint region. All 696k contigs generated from the *de novo* assembly were blasted against the reference library. We set the blast to output format 6 and parsed it for a sequence that aligned in both positive and negative orientation across two scaffolds.

Syntenic region plots

Synteny plots were created using LastZ (Harris, 2007) The Rhesus reference genome rheMac8 was aligned against MacFas5 the *M. fascicularis* reference genome. Dot plots were created from the LastZ output for synteny analysis and large variant analysis. Dot plots were filtered for 95% mapping quality before being plotted using ggplot2 (Wickham, 2011) in R.

PCR Validation of LastZ breakpoints

PCR validation on large scale synteny variants from chromosome 5 LastZ alignment were performed. PCR primers were designed using Geneious Primer3 (Untergasser et al., 2012). PCR validation was carried out in 20ul reactions using Promega kit with a touchdown PCR procedure: 0.2mM dNTP, 1X buffer, 2.5x $MgCl_2$, 0.5uM Forward primer, 0.5 uM Reverse Primer, 0.5 units of Taq.

Thermocycler conditions were:
Lid-105°C
95°C for 3:00
94 for 0:30,
65°C for 0:30
-0.5°C per cycle
72°C for 0:45
25X
72°C for 5:00
Permanent hold at 4.

Samples were then examined using gel electrophoresis on a 2% gel using a Biorad gel imager.

Identify areas of high divergence between *M . mulatta* and *M . fascicularis*

SNP calling and filtering:

SNPS were called using GATK haplotype caller using the ALLSITES option in reference confidence mode to output variant and invariant sites from each samples BAM file. The output was filtered with VCFtools and GATK and combined with using the following filters: QD < 2.0 (Quality by depth), MQ <40.0 (Root mean square mapping quality), FS>60(Fisher strand), SOR

>3.0 (Strandsoddratio), MQRanksum < -12.5, (Mapping quality rank sum test) ReadPosRankSum <- 8.0 (Read position rank sum test), DP <5, DP >120 (Depth). Indels were filtered using VCFtools with the following parameters: FS > 200(Fisher strand), ReadPosRankSum <-20.0 (Read position rank sum test), SOR> 10.0 (Strandsoddratio , DP <5 ,DP >120 (Depth).

F_{ST}

F_{st} is often referred to as the fixation index (Nei, 1973). It is a measure of population differentiation, often using SNPS in its estimates of genetic differentiation between populations. F_{ST} is a special case of Wright’s F statistic and is and commonly used statistic in population genetics. F_{ST} measures the amount of genetic variation along the genome that can be explained by populations structure. It is bound from 0 to 1 and is the probability that two randomly selected alleles are from different source populations. Weir and Cockrams F_{ST} was calculated using VCFtools version 1.14-14 between *M. mulatta* and *M. fascicularis* sample sets. Wrights original formula assumes precise allele frequency per population and in a reality, this is impossible. Wier and Cockrams formula calculates F_{ST} with multi allelic loci, first pioneered using gel electrophoresis data. Here F_{ST} was calculated from filtered SNPs in 50 kilobase sliding windows across the genome.

Tajima’s D was also calculated. It is another population genetic test that computes the difference between the number of pairwise differences and the number of segregating sites (Tajima, 1989). Each of these values is scaled to assume a neutrally evolving population that is not changing in size. Tajima’s D can be used to find out if a population is not evolving under random chance but is under some sort of selection such as balancing, negative or directional selection. Tajima's statistic computes a standardized measure of the total number of SNPs in the sampled DNA and the average number of mutations between pairs in the sample. Tajima’s D was calculated using VCFtools version 1.14-14 for both *M. mulatta* and *M. fascicularis* populations respectively, in 50kb sliding windows across the genome. Tajima’s D was calculated using an infinite sites model where each new mutation affects a new site in the sequence.

*Nucleotide diversity (*π*)* – is used to measure the degree of polymorphisms within a population (Nei & Li, 1979). Nei and Li first introduced a commonly used measure of nucleotide diversity in 1979. Their measure is defined as the average number of nucleotide differences per site between two DNA sequences. Nucleotide diversity can be calculated by examining the DNA sequences directly or may be estimated from molecular marker data was calculated using VCFtools version 1.14-14 for both *M. mulatta* and *M. fascicularis* populations in 50kb sliding windows across the genome respectively.

D_N /D_S is a ratio of nonsynonymous (point mutations that results in an amino acid change) to synonymous (point mutations that do not change the amino acid) per nonsynonymous site in a set of coding sequences. The theoretical results of D_N /D_S can be interpreted as followed:
Positive selection- A theoretical D_N /D_S of greater than 1 implies N>S and there has been positive selective pressure on the system. Positive selection often occurs in immune system geneses.
Neutral evolution- a D_N /D_S ratio of one implies there have been equal number of changes N=S but this is relative to the number of sites sampled.

Negative selection- A D_N /D_S of less than one implies S>N and there was evolutionary pressure to maintain an ancestral state over the new mutations that arose. However, theoretical and empirical estimations vary wildly (W.-H. Li, 1993; W. H. Li, Wu, & Luo, 1985; Nei & Gojobori, 1986). Depending on your system D_N /D_S ratios of ~0.3-0.5 can be interpreted as positive selection as long as proper control genes are used.

In calculating D_N /D_S there is a systematic bias that programs must take into consideration; the rate at which the various nucleotides ATCG are swapped, as certain swaps are more probable than other. Because transitions (t<->c and a<->g) are favoured over transversions (the substitution of a purine for a pyrimidine) (W. H. Li et al., 1985), computational models must take this into account. Programs take accuracy over speed or speed over accuracy, so understanding your program and its potential shortfalls is crucial to interpretation. For example Nei & Gojoboris algorithm neglects the transition/transversion bias for the sake of computational time, but at the expense of accuracy as this method will overestimate N and underestimate S (Pamilo & Bianchi, 1993). Here D_N /D_S was calculated in 50kb bin sizes across the coding region of the genome. SNPs were applied to the reference genome rheMac3 using VCF tools consensus to create a new fasta file for input. Coding regions were extracted using the rhesus GTF file and stop codons were masked using a custom script. D_N /D_S was calculated using MEGA7 (Kumar, Stecher, & Tamura, 2016) for autosomes, chrX, PAR1 region and PAR2 region to test for faster chrX evolution. Transition/transversion rates were estimated and counted for each codon. Correction for multiple hits was applied and D_N /D_S ratios were fit. Additionally, verification or rejection of the PAR 2 region in macaques was analyzed using the known genome location in humans.

Gene Ontology for divergent and unique structural variant regions

A useful tool to understand underlying gene families is a Gene Ontology analysis (GO) (Ashburner et al., 2000). Most GO enrichment tools deal with statistics methods (p-value, FDR, Bonferroni etc). DAVID is a well maintained and curated database that uses a variant of the Fisher exact statistic for p value calculation called EASE (Hosack, Dennis, Sherman, Lane, & Lempicki, 2003) which is more conservative than the standard way of calculating P values in GO analysis. Here the GO analysis was used to identify enriched gene families in highly divergent regions. Regions of high divergence – the top 1 percent of F_{ST} outliers, combined with a low Pi and negative Tajima's D in *M. fascicularis* were used for the gene ontology analysis. This intersection between various diversity/divergence statistics ensures that the population differentiation indicated by the high F_{ST} estimate was caused by reduced population diversity in the *M. fascicularis* population and is consistent with positive selection as indicated by the Tajima's D result. Both top and bottom 1 percent F_{ST} regions were extracted. Genes within these matched regions were input into DAVID Bioinformatics toolkit for a GO analysis. After filtering ambiguous gene families, gene families with an enrichment score of >1.2 were analyzed.

Unique SV regions binned from 10kb-100kb were used for a separate GO analysis. The smaller bin sizes were chosen based on the infeasibility of using all 1Mb+ variants and the vast number of genes in them, which could have a large bias on potentially false positives and false gene categories in the gene ontology analysis.

These regions were used as the targets for gene extraction. The output list of genes were and input into DAVID Bioinformatics toolkit for a GO analysis (Huang, Sherman, & Lempicki, 2009). Gene IDs were extracted using BEDOPS and GTF2BED using the list of regions and annotation files. ID's were then filtered against the database so only ID's DAVID recognizes were used. Gene families with an enrichment score of >1.4 were used. Gene families underlying structural variants were analyzed for relevance to the scientific question.

Do areas of high divergence match areas of areas of structural variation?

F_{ST} & SV overlap

First F_{ST} in 50kb bin sizes were compared to SV call sets in small(1kb), medium(1-10kb), large(10kb-1mb) bin sizes to get the actual amount of overlap between the two datasets. Then a custom python script was written to simulate the occurrence of SV and F_{ST} region overlap. This was used to calculate whether empirical SV and F_{ST} overlap was more often than expected by chance. Regions of the genome space were randomly sampled in 50kb sections for the "FST" dataset and in the varying bin sizes of 1kb,10kb,50kb,100kb and 1mb for the "SV" dataset. Ten thousand iterations of ten thousand samples were run of the overlap to get a distribution of the percentage overlap.

Overlap of FST&SV Gene ontology

Gene families from both gene ontology analyses were compared to better understand what categories are shared and unique between these datasets.

Results

What type of structural variants occur between and within species?

Whole genome Structural Variant analysis

After alignment of *M. mulatta* (SRA023856) and *M. fascicularis* (SRA023855) to the rhemac3 reference genome, 34x and 33x coverage were achieved respectively, with an average insert size of 1000bp: range (500-15000bp). Larger insert sizes aide structural variant prediction (see Materials and Methods). Qualimap (Okonechnikov, Conesa, & García-Alcalde, 2015) summary statistics on each bam file can be seen in (Appendix 2&3). Whole genome analysis of SV raw counts before filtering and merging yielded a total of 3,726,607 structural variants across five programs for *M. mulatta* and 576,220 for *M. fascicularis* respectively (Table 3).

Insertions were the majority of the call set in *M. mulatta*, however these SVs were all 1bp in size. Alternatively, no insertions were recovered from the *M. fascicularis* dataset. 6.4X more variants were recovered in *M. mulatta* than *M. fascicularis*, however, after removing insertions, this reduces to a ratio of 1.2x more in *M. mulatta* as compared to *M. fascicularis.*

After merging with bedtools *multiIntersect bed* (Quinlan & Hall, 2010) SVs were analyzed at intersections of 3/5, 4/5, and 5/5 programs. Variants were then binned into four categories, small (1kb), medium (1kb-10kb), large (10kb-1mb) and extra-large (1mb+). Results of 3 structural variant programs making the same call for a particular variant can be seen in (Table 4). The 3/5 program consensus was selected because there were no variants that were called by either 4 or 5 programs. Despite the reduction in the total call set after requiring three programs to predict the same SV, *M. mulatta* to had ~3x more SV than *M. fascicularis* (N=15,541 vs N= 4,073). As expected the majority of structural variant calls between the two species were shared (N= 3,960). There were 11,581 SVs unique to *M. mulatta* and 113 unique to *M. fascicularis* shared which is a ~5% of the total. To get an understanding as to why SV algorithms called variants based on the alignments, 10 variants of each type were visualized and can be seen in (Appendix 1). From this pairwise alignment analysis comparing the two published genomes (rheMac8, macFas5), several potential large-scale structural differences are present on eight chromosomes: 2, 3, 5, 12, 13, 16, 19, and 20.

Chromosome five putative inversion

Consistent with previous karyotype results from (Stanyon et al., 1988), *in silica* methods here predicted a large variant present in *M. fascicularis* but not in *M. mulatta.* This putative 15 Mb inversion was supported by three separate SV callers, Lumpy, Breakdancer and Delly, and spans 42 Mb-57 Mb of *M. fascicularis* chromosome 5. Among the three programs there was slight variation of the precise breakpoints (Figure 5), nonetheless, they were all within PCR distance from each other. Further analysis of this putative inversion via PCR was in conjunction with breakpoint refinement via *de novo* assembly using the raw *M. fascicularis* reads (see Materials and Methods). Of the ~694k contigs assembled no contig matched inversion criteria of two different contigs with differing oreintations.

Breakpoints refinement via PCR was also conducted using eight sets of primers spanning the potential right breakpoint regions of each software. The reported left breakpoint regions were not easily within PCR range of eachother, therefore the right breakpoint was selected for PCR validation. If a primer pair spanned a breakpoint region, it would be too large to amplify via PCR in *M. fascicularis*, however, *M. mulatta* samples would be expected to have sucessful amplification due to a lack of the inversion variation. In our results, every primer set had sucessful amplification of PCR product in all samples, thus not validating the inversion variant predicted in *M. Fascicularis.* These results were interpereted as the inversion variant being a false positive. Alternativly the breakpoint could be located outside the specific regions predicted by the three programs, lumpy delly and breakdancer. While we did not validate inversion variant at this location, it does not rule out the potential karyotype difference as being due to a structural variant.

LastZ comparison of reference genomes

In addition to PCR and *de novo* validation attempts, I attempted validation of a similar putative inversion identified through the synteny plots of the two genome assemblies rheMac8 and macFas5. Specifically, an inverted region between rhemac8 and macFas5 located on chromosome 5 from the LastZ alignment was noted for analysis (Figure 4). As done above, primers were designed flanking the putative breakpoint regions and touchdown PCR was performed. As the primers were designed with the macFas5 reference genome, expected results would be amplification in *M. fascicularis* and no amplification present in *M. mulatta* samples for confirmation of the inversion variant. Similar to the other predicted chr5 inversion variant, none of the PCRs validated the breakpoints with amplification across all samples. Therefore, among the small set of inversion variants selected we were unable to empirically validate any detected structural variant in this dataset.

Potential structural variation caused by alignment bias.

To test whether the aligner of choice has an effect on structural variant calls reads were aligned using both BWA and Bowtie2. Both aligners are commonly used in next generation sequencing analysis, but they use slightly different methods. Because softwares to detect SVs use the mapping information for SV detection, slight variation in the alignment algorithm could possibly cause variation in the SV call sets produced. Both BWA and Bowtie 2 are written in C++, take in fasta or fastq input files, output a sequence alignment file, work with paired end reads and have the ability to be multithreaded (Ruffalo et al., 2011). However only BWA can do gapped alignment or trimming of reads. There appears to be a tradeoff between speed and accuracy, BWA is not as accurate in mapping but is significantly faster (Ziemann, 2016). BWA has a higher tolerance for mismatches than bowtie (Ziemann, 2016). The primary speed hindrance is indexing in bowtie takes is indexing which takes ~100% longer in Bowtie2 than BWA. Indexing with BWA uses an fm index and backtracking for inexact match sets. As for SV calls, using only breakdancer, no major difference between Bowtie2/BWA on high/low stringencies, or defaults were apparent in our dataset. Additionally, literature search could not find any information on the choice of aligner and its settings creating false positives. It was found previously that Bowtie2 best overall but BWA is better for longer reads (Hatem, Bozdağ, Toland, & Çatalyürek, 2013). However, there was slight variation between the two aligners on the number of SV calls, their quality and the average length. Average score was 68.926 for BWA and 61.9 for Bowtie2 respectively. 312,402 variants were called in the BWA dataset as compared to 228,032 for Bowtie2. Average size of SVs between aligners was 1.43925e+06 for BWA vs 9.33101e+06 for Bowtie2. A summary of SV call rates can be seen in between the two aligners can be found in (Table 6).

What regions of the genome harbor high levels of diversity and divergence.

Diversity and Divergence Analysis

Calculation and analysis of F_{ST} between *M. mulatta* and *M. fascicularis.*

A combined VCF file with samples from both *M. mulatta* and *M. fascicularis* was used to calculate divergence statistics. In total 12 samples were used, 8 from *M. fascicularis* and 4 from

M. mulatta. Weir and Cockrams F_{ST} was calculated to identify areas of high divergence between these two populations using VCFtools in bin sizes of 50Kb,250Kb,500Kb,1Mb and 1.5Mb (Figure 6). After initial analysis, in an attempt to reduce the confounding effect of all of the areas of high divergence being on the X chromosome. And to better facilitate the GO analysis and in accordance with the average gene size in macaques, the 50kb bin size was chosen for downstream analysis (Figure 6). The top one percent of F_{ST} outliers were used in a GO analysis and compared, to SV regions. The F_{ST} cutoff for the top 1% was to 0.831. Overall 54,951 regions were sampled for F_{st} between *M. mulatta* and *M. fascicularis* within the 50kb dataset. After filtering for only the top 1% of F_{st}, 862 regions remained. Overall 45% of the total 1% highest F_{ST} regions were located on the X chromosome.

Pairwise nucleotide divergence and Tajimas D was also calculated in varying bin sizes but ultimately analyzed at the 50kb bin size. Pi was calculated as a measure to control for within species diversity of the samples. This intersection between various diversity/divergence statistics ensures that the population differentiation indicated by the high F_{ST} estimate was caused by reduced population diversity in the *M. fascicularis* population and is consistent with positive selection in *M. fascicularis* populations as indicated by the Tajimas D result Overall Pi ranged from 0.0-0.3 +/- 0.1 in both *M. mulatta* and *M. fascicularis* (Figure 7)

Regions of high F_{ST} , low Pi and negative Tajimas D within *M. fascicularis* were plotted across the genome to further visualize their location and confirm a potential large X effect (Figure 8). A chrX effect can be seen where areas of high F_{ST} that are also associated with low PI and negative Tajimas D primarily occur on the X chromosome. This propensity for outliers to be concentrated on the X chromosome led to further research into potential Large X or faster X effects in this macaque system (see Discussion).

GO from regions of high divergence

Areas of high between species divergence were the target of a subsequence GO analysis. Genes in these regions were extracted and uploaded into DAVID bioinformatics to get an idea of the gene families within high divergence regions between these two species. Results can be seen in (Table 7). Overall 74 regions were input into David, corresponding to the top F_{ST} outliers. The top 10 set of GO categories from each top and bottom F_{ST} outlier regions can be seen in (Table 8).

The majority of the bottom gene families are highly conserved cellular functions. Metal binding pathways, cell recognition and cellular proteins are all highly conserved. Inversely, the top 1% gene families were highly differential gene categories such as sexual/immunological and metabolic pathways.

GO analysis of structural variant regions

Structural variant gene ontology was also conducted with DAVID bioinformatics tool. Unique structural variant regions for both *M. mulatta* and *M. fascicularis* were inputs for the program. In total 831 unique regions for *M. mulatta* and 113 unique to *M. fascicularis* were used. This led to 51 genes for *M. fascicularis* and 109 for *M. mulatta.* David outputs can be seen in (Table 9,10) for each respective species. *M. fascicularis* primarily returned gene categories pertaining to myosin and motor proteins with addition to the SNARE complex, which mediates vesicle binding in the neurons. *M. mulatta* returned numerous gene categories dealing with regulation of apoptotic processes and apoptotic chromosome condensation. Additionally, histone exchanges and steroid receptor regulation was recovered. The final two gene families recovered were the regulation of circadian rhythm and that of protein heterodimerization.

Could a Faster X effect explain higher differentiation on the X between M. mulatta and M. fascicularis?

As stated above the majority of regions of high differentiation were clustered on the X chromosome. To test for a faster X effect, we performed a variety of subsequent analysis. First, Dn/Ds was calculated in coding regions to test if recent X polymorphisms are under stronger selection in macaques. Additionally, the presence of pseudo autosomal regions PAR1 and PAR2 was determined based on comparison to the autosomes and the X chromosome. D_ND_S was calculated in 50k bin sizes resulting in a sample size of (N=54,624), (N=55), (N=7) and (N=3.046) respectively (Figure 18). First our results showed higher D_ND_S on the X chromosome as compared to the autosomes. While the autosomes were statistically different than the X, a posthoc Tukey's test did not support a statistical difference between the autosomes and PAR1. Nor did it support a statistical difference between the X chromosome and PAR2. This validates PAR1 in macaques, as it supports its similarity to autosomes. Meanwhile PAR2 being similar to the X chromosome does not support its existence in macaques. Therefore, only PAR1 will be considered for further analysis.

Do areas of high divergence match areas of SV calls?

For better data visualization structural variants and F_{ST} outliers were plotted using KaryoPloteR (Gel & Serra, 2017) for visualization of their overlapping regions along the genome. Additionally, a two-pronged analysis was performed to ascertain if areas of divergence matched areas of structural variation more often than expected by chance. Computational random sampling found that 1.7-3.4% of the time randomly sampled regions of the F_{ST} and SV bin sizes overlapped. However, the overlap of actual F_{ST} and SV datasets was (81%) In total 21,544/26,391 or 0.8115621 structural variants hit an F_{ST} outlier. This is a significant amount of overlap as compared to simulations, indicating that F_{ST} outliers and SVs co-occur more often than expected. Distributions of overlap can be seen in (Figure 19-24).

Comparison of SV GO and F_{ST} GO

There was only one significant overlap between the structural variant and F_{ST} GO. This gene category had to deal with mitochondrial transport of cytochrome C oxidase. Cytochrome C oxidase is the terminal electron receptor in the electron transport chain and is an area of intensive study (Castresana, Lübben, Saraste, & Higgins, 1994).

Discussion

Here, we provide a mix of traditional wet lab work with an *in-silica* approach to provide a broad overview of structural variation between *M. mulatta* and *M. fascicularis*. Understanding the factors that drive biological diversity is a key concept in evolutionary biology. Species represent the base unit of biological diversity; as such, understanding potential mechanisms that drive speciation could help shape our understanding of evolutionary biology. The three objectives were as followed:

1) Identification and comparison of shared and unique Structural variants
2) Identify areas of high divergence between *M. mulatta* and *M. fascicularis*.
3) Do areas of high divergence match areas of areas of structural variation?

we took a computational approach using publicly available genomic data. The goal was to provide a broad characterization of structural variation and genetic differentiation between two closely related hybridizing species. In conjunction with this, underlying gene families from both of these analyses were analyzed via a gene ontology analysis. Finally, a statistical analysis was used to determine if these reported areas of high nucleotide divergence and structural variation overlapped more often than expected by chance.

Identification and comparison of shared and unique Structural variants

First, five structural variant programs were used and two genomic samples to analyze where structural variants occur along the genome, and how many are shared or unique between the two species. The whole genome structural variant analysis of *M. mulatta* (SRA023856) and *M. fascicularis* (SRA023855) raw variant calls, as compared to the filtered calls, provides an insight into the large call rates of SV programs and SV call sets. These results highlight the importance of both using multiple programs and filtering. Next generation sequencing, specifically Illumina datasets, like those used in this study, provide us with an interesting perspective on structural variation due to their millions of short reads. The variation of read size has increased the difficulty to detect SVs with modern day next generation sequencing data, given its short-read size (Mohiyuddin 2015). However, these short mate paired reads can also lead to aberrant variant calls during alignment to a reference. No one structural variant program can detect all variants with the same accuracy. Depending on the program, it may have more power to detect specific SV types and limitations in detecting others. Currently, there are a plethora of SV detection programs available with large variation in their overall accuracy for specific SV calls based on the type of data used to call variants.

In my analysis, variants were binned by type and then by size. Pindel was the least conservative of the SV programs, accounting for nearly half of all variants called, with the vast majority (~3 million) being single base pair insertions in the *M. mulatta* call set.Intersecting the variants between softwares drastically cut down on the number of putative calls, from ~4.4 million to 19,214 (0.44%). This large reduction in SV calls after intersection points to a very high rate of false positives in each program, which is a known limitation of these types of softwares.

The majority of the SV dataset was shared between the two species, likely due to a lower mutation rate (and subsequent fixation rate) of SVs than SNPs (Scally, 2016; Ségurel, Wyman, & Przeworski, 2014). Structural variation is relatively conserved and slow evolving in systems. The ability to use multiple structural variant programs and differing detection methods can provide researchers with two vital functions – (a) the potential to recover putative structural variations that would not have been called based on another program used, and (b) each SV program has its own strengths and weaknesses based on the algorithm it uses to detect variation. Secondly, using multiple programs and limiting analysis to their intersection allows a drastic reduction in putative false positives calls to give higher confidence a variant is real. Conversely, one could filter out a high-quality variant that did not meet whatever threshold the user provides, yielding increased false negatives. This is one of the shortfalls in current structural variant technologies and methodologies, and it a vital reason why always adapting methods with new algorithms and methods important, such as Breakseq (Lam et al., 2010), Hydra (Lindberg, Hall, & Quinlan, 2015), (Wijaya, Shimizu, Asai, & Hamada, 2014) Genomevip (Mashl et al., 2017), Crest (Wang et al., 2011) and Novobreak (Chong et al., 2017) has the potential to vastly improve this type of research in the near future.

Were this analysis to be redone, new softwares such as those listed above would be incorporated which have been shown to have greater power to detect SVs. Additionally, more sophisticated programs to intersect structural variation programs have come out SVmerge (Wong, Keane, Stalker, & Adams, 2010) and NextSV (https://github.com/Nextomics/NextSV) and potentially have a higher recovery rate and more optimization than the custom python script used here. Additionally, datasets with 4/5 and 5/5 program calls were not recovered. Future studies should incorporate more programs and more refined accumulation methods. However, in this study all variant types studied were recovered after merging, except insertions. Inter-chromosomal translocations could not be recovered due to the nature of our genome analysis pipeline. Specifically, datasets were split up by chromosomes to speed up computational time. However, *M. mulatta* had the majority of the SV calls as compared to *M. fascicularis* which is interesting, considering it is the reference. Typically, one would expect the sample not matching the reference genome to have more differences. In addition, a manual inspection of structural variants called by 1/5, 2/5, 3/5 programs were performed using the genome browser IGV to ascertain on how many appear to be false positives. With a false positive rate of ~33%.

Lack of empirical validation of chromosome 5 putative inversions

Based on a karyotype study in 1988, we sought to focus our efforts for empirical validation on inversions on chromosome 5. Initially, 3 structural variant programs were selected to analyze the genomic data: Breakdancer, Lumpy and Delly. Results from these softwares were filtered and all three were in high accordance of a putative chromosomal inversion located from 42mb-57mb on chromosome 5. While the breakpoints were not precise by the three programs, they were within 1kb distances allowing for an attempt at PCR validation

Unfortunately, all 8 sets of PCR primers and their subsets run did not recover an inversion breakpoint in *M. fascicularis, all PCRs* yielded a product when the expectation was failure of a PCR product. Additionally, since all contigs created from the Ray de novo assembly and subsequent blast and dot plot analysis did not recover any inversion breakpoints. The combination of these two empirical results led to the conclusion that this putative 15mb is a false positive or not present at the reported breakpoints via the three programs.

Additional chr5 SVs were predicted on chromosome 5 through a LastZ alignment of the two reference genomes, macFas5 and rheMac8. These two inversions were again the targets of PCR, however, like the inversion found with the SV programs, this also failed to validate. Unfortunately, this means that no variant call computationally found was empirically validated. While this does not entirely rule out the potential variation proposed by Stanyon, it does give insight into the potential for algorithms to inaccurately call structural variants in a genome. Additionally, this gives more credence for the need to refine these methods and add in more approaches to further validate any potential variant detected. While it is an imperfect system, it is still potentially far more accurate and efficient than trying to detect variants in other ways such as inspection of karyotypes and backcrosses (Escaramís, Docampo, & Rabionet, 2015).

Potential downstream effects of choice of aligner

To ascertain if a researcher's choice of aligner has any downstream effect on analysis, two popular aligners BWA and Bowtie2 were chosen. Often a researcher picks analysis programs for their pipeline unbeknownst to potential downstream effects. To investigate this in the context of structural variation, we aligned one sample from both *M. mulatta* and *M. fascicularis*. Then basic statistics and information about the structural variants recovered were calculated. There does exist not only call type (one program calling more of a variant than another) but also size difference between BWA and Bowtie2, as such this could lead to unknown bias in downstream analysis of structural variation experiments by skewing your data. Other research on the comparison between BWA and bowtie has found BWA had the highest tolerance for mistakes (Ziemann, 2016). However, it was also reported that Bowtie2 was the best overall, with BWA being better for long reads (Hatem et al., 2013).While BWA and Bowtie2 seem to be very comparable initially, with preference coming down to just runtime and type of input data (Hatem et al., 2013; Ruffalo et al., 2011; Shang et al., 2014; Yu et al., 2012; Ziemann, 2016), further analysis of downstream effects needs to be conducted.

Areas of high divergence between *M. mulatta* and *M. fascicularis* highest on the X chromosome

Identifying areas of divergence between completely or partially isolated populations within the genome can be an important factor in understanding species. F_{ST}, Tajima's D, pairwise nucleotide divergence, and $D_{n/}D_s$ were all calculated to help elucidate population differences between *M. mulatta* and *M. fascicularis.* These differences can accumulate due to various evolutionary forces, such as natural selection, genetic drift, or mutation. Demographic differences can also impact several of these statistics, such as population substructure or differences in effective population sizes (Nei, 1973; Tajima, 1989). Previous work had shown that structural variation can cordon off species-specific adaptations or isolating factors (Feulner & De-Kayne, 2017). These variants can drastically reduce or eliminate introgression. Because *M. mulatta* and *M. fascicularis* are so closely related and known to hybridize, relative genomic differentiation should be low. Previous work has shown that genomic differentiation is around ~ 20million SNPs or 0.6% of the total genome (Yan et al., 2011). Sliding window scans of F_{ST}, PI and Tajima's D in 50kb regions along the genome identified regions with high divergence between the two species that are driven by between species differences not within species differences. This was achieved by combining F_{ST} a between species measurement with Pi and Tajimas D, two within population measures of divergence. This approach would attribute the high F_{ST} outliers solely to between population differences. All divergence and diversity statistics were calculated in multiple bin sizes to select a specific bin size for further analysis, and the 50kb bin size was chosen based on the average macaque gene size of 10-15kb. Over 33% of the regions of high differentiation were on the X chromosome. Next, these regions of high F_{ST} were analyzed with a gene ontology analysis. The GO analysis looked for enriched gene families in these highly divergent regions. The majority of gene families recovered over the 1.2 enrichment score cutoff had to deal with sexual or metabolic/immunological gene families supporting the potential for immunological genes being an importance of species maintenance between these two species. While gene ontology analyses are not an endpoint analysis and are primarily used to survey what is in the region, it is interesting to find these specific gene families. A dividing factor between new species can be the evolution or differentiation of not only sexual characteristics but also those of immunological changes. Often when a species reaches a new environment, its immune system is one of the first things to change, as it has more of an effect than the baseline genomic sequence (Brodin et al., 2015).

The combination of the majority of the gene families being involved in sex/immune/metabolism in conjunction with the F_{ST} outliers being on the X chromosome led to the incorporation of some fast X hypothesis and testing. The two hypotheses that were tested were (1) new variations on the X would be under higher selection (Garrigan, Kingan, Geneva, Vedanayagam, & Presgraves, 2014; Ge, Kwok, & Shieh, 2015; Presgraves, 2008) and (2) that there should be less divergence but more diversity between *M. mulatta* and *M. fascicularis* than between the nearest outgroup, the baboon (Elango, Lee, Peng, Loh, & Yi, n.d.).

To test hypothesis 1, in addition to looking at new variations under higher selection on the X, the relative rates of the pseudo autosomal region (PAR) one and two were assayed against autosomes and the X chromosome. While it is not known if macaques have a PAR2 region the region identified as PAR2 in humans was used. Results did not support the hypothesis, where the pseudo autosomal region acted like an autosome in PAR2 but did support it for PAR1, grouping with the autosomes during a statistical post hoc Tukey test and was generally lower than the X region. This led us to assume that macaques either do not possess a PAR2 region or that the region is in a different location than humans. For the remainder of the study, only PAR1 and X were considered separately from the autosomes. It is possible that *M. mulatta* and *M. fascicularis* exist as two distinct gene pools because structural variants have been the driving force between them, allowing recombination to homogenize these two species outside SVs (only in inversions however) while remaining strikingly different within them. However, this study did not look into the rates of LD or reduction in recombination from structural variation between the two species and if they lined up with either areas of high F_{ST} or areas of structural variation.

Areas of high divergence overlap SVs more often than expected by chance.

While it was not possible to ascertain if structural variation was causing the high divergence or if the high divergence was causing the structural variation, it was possible to determined how often these two sets of data overlap. To assay whether or not these areas of high F_{ST} and areas of structural variation overlapped more often than by chance, a simulation approach was taken. First, empirical data was compared against simulated data. Empirical data showed an overlap between high divergence and SV regions by 81%. Simulated data overlapped with a range of 1.7% - 3.7% after 10,000 simulations. Thus, structural variant regions overlap with regions of high nucleotide divergence more often than by chance together. The comparison of gene ontology analysis did not bear any significant similarities. The only major gene family in both GO analysis was the mitochondrial transport of cytochrome C oxidase. All other gene families present in both analysis did not overlap.

Summary

Overall, a drastic amount of structural variation can be detected in *M. mulatta* and *M. fascicularis* with the majority share between these two closely related species. This variation can be validated using additional programs and more sophisticated methods of intersection. The majority of the variants detected drastically reduced as stringency increased. The number of variants called between BWA and Bowtie2 was not significantly different, however quality differed from 61 to an average of 68. Indels between rhemac3 and macFas5 were inversely recovered, adding an additional layer to confirm structural variants between two species if you possess a reference genome for both taxa, which is likely very unique to our system. Genome quality went up respectively, with macfas5 being less powerful as compared to rheMac3 and finally rhemac8. Genome quality increased with improved reference genome versions, as one would expect. Comparisons of the reference genomes with LastZ led to the identification of additional putative structural variants. In addition, several structural variants were recovered in the dataset of the 5 programs that were corroborated by the synteny plots. These variants could be targets of validation or inquiry for future projects. In addition to these structural variants, regions of high divergence were also inspected. With the majority of these regions clustering on the X chromosome, a range of questions were asked relating to faster X evolution. Many of which were corroborated, suggesting that new SNPs on the X chromosome are under higher selection than SNPS on the autosomes or the PAR1 region. This effect can be driven by the genome not having two copies to protect itself from innately deleterious mutations. Thus, these regions that are exposed must be under a stronger pressure of maintenance. These hypotheses were supported by the data, as more variation was found on the X chromosome than on the autosomes and the pseudo autosomal region of the X chromosome. Finally, the data showed that regions of structural variation lined up with regions of high divergence more often than expected by chance. With a maximum of 3.4% simulated overlap chance compared to the 81% observed, we are led to believe that in this system, SVs and F_{ST} outliers are correlated. While our SV detection methods were not perfect, and further refinement of not only detection, but validation is needed, it is probable this trend would still persist. However, it is still unknown which is causal, the SVs or the genetic divergence.

Figures

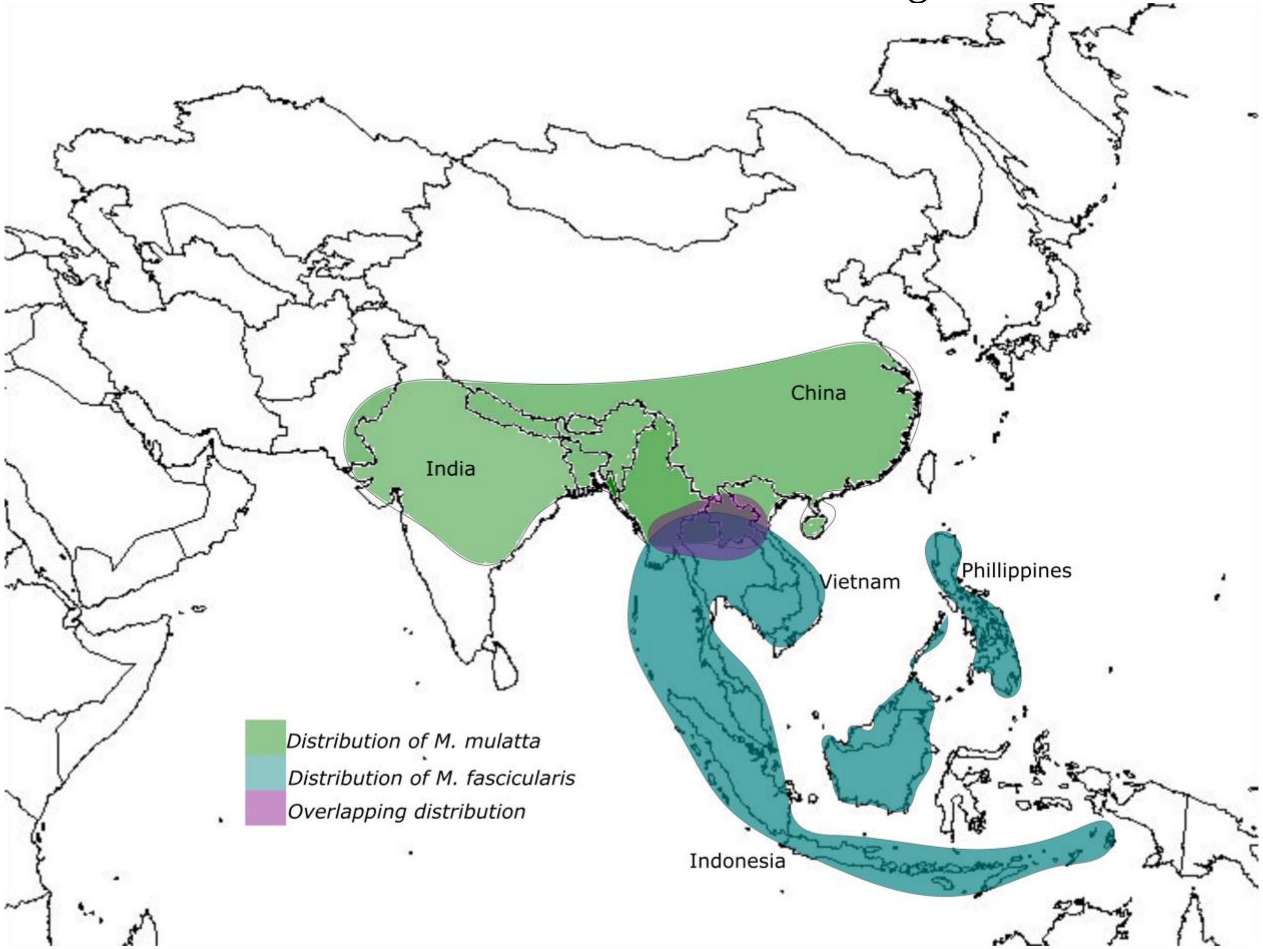


Figure 1: Geographic ranges for M. mulatta and M. fascicularis. Range of M. mulatta can be seen in green spanning India and China, M. fascicularis in teal, and overlapping zone of hybrid alleles in purple. (Adapted from Fooden 1980).

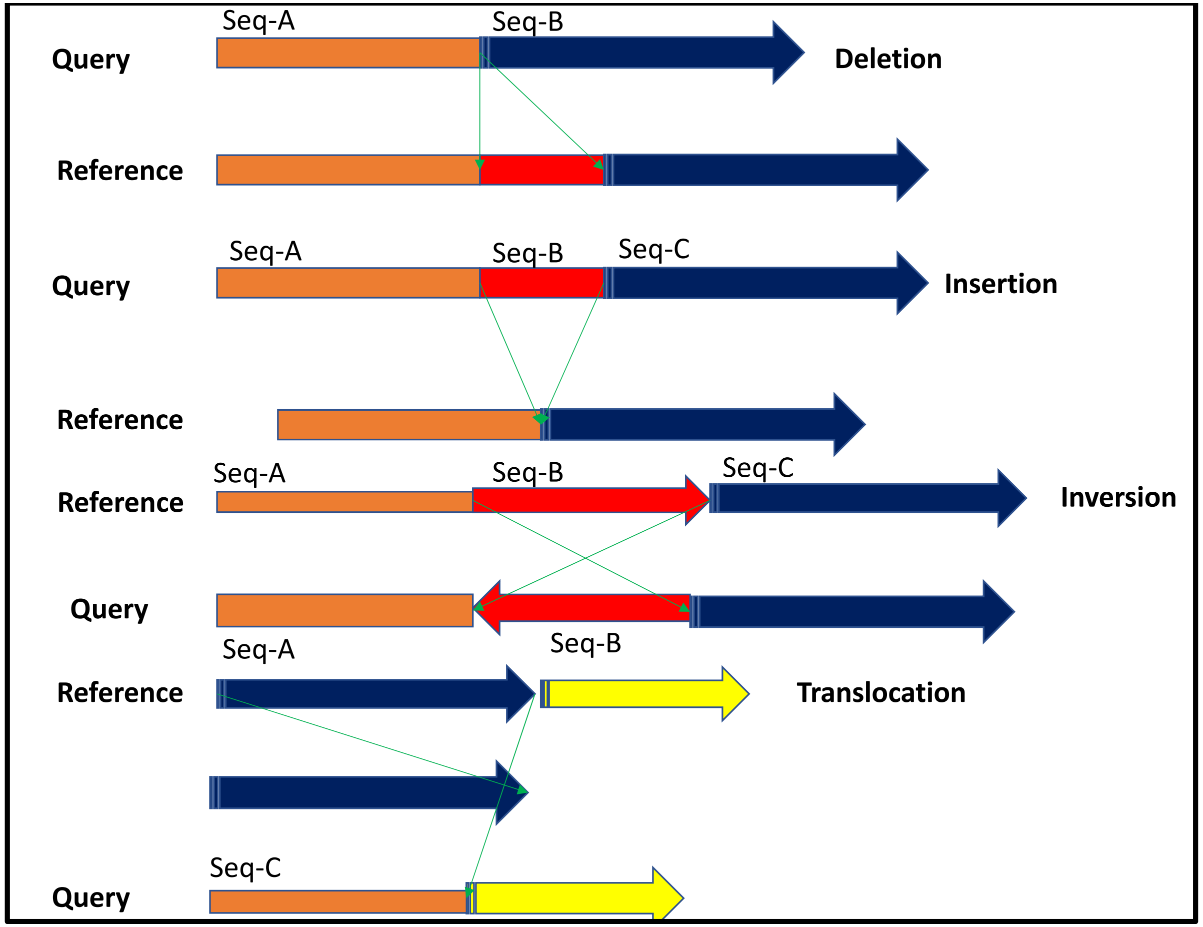


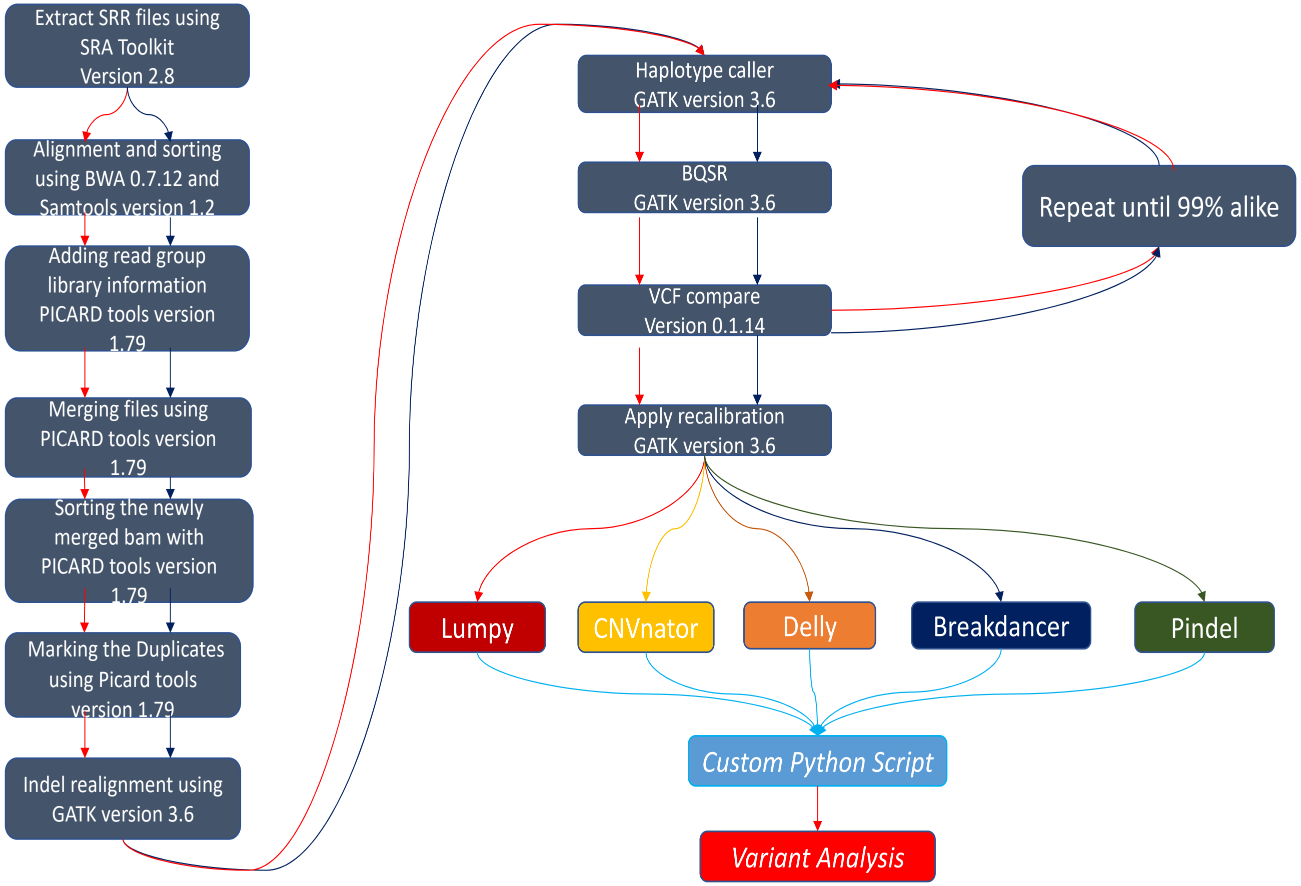
Figure 2:Diagram of potential structural variant signatures analyzed via computational methods: reference and query sequences can be seen with respective variants in red.

Table 1:Publicly available whole genome data used in this study. In total, 12 samples were used from a broad geographic range, including six from the extremely isolated Republic of Mauritius.

Sample	Location	GenBank Accession	Platform
M. mulatta	India	SRR278739	Illumina
M. mulatta	South China	SRP003590	Illumina
M. mulatta	China	SRA023856	Illumina
M. mulatta	India	PRJNA382404	Illumina
fascicularis	Vietnam	SRP045755	Illumina
***M. fascicularis* (6)**	Mauritius	PRJEB7871	Illumina
M. fascicularis	Indonesia	SRA023856	Illumina

Table 2:DNA samples obtained from Oregon Primate center that were used for PCR validation. DNA samples came from a range of geographic origins to represent a broad representation of M. mulatta and M. fascicularis.

Animal ID	Species	Geographic Origin	Gender
OR981	*M. fascicularis*	Cambodia	Male
OR272	*M. fascicularis*	Cambodia	Male
OR840	*M. fascicularis*	Philippines	Female
OR854	*M. fascicularis*	Cambodia	Female
OR362	*M. fascicularis*	Indonesia	Male
OR571	*M. fascicularis*	Indonesia	Male
OR005	*M. fascicularis*	Vietnam	Male
OR560	*M. fascicularis*	Vietnam	Male
OR652	*M. fascicularis*	Mauritius	Female
OR066	*M. fascicularis*	Mauritius	Female
OR414	*M. mulatta*	China	Female
OR163	*M. mulatta*	China	Female
OR551	*M. mulatta*	China	Female
OR864	*M. mulatta*	China	Female
OR638	*M. mulatta*	China	Male
OR451	*M. mulatta*	India	Female
OR140	*M. mulatta*	India	Female
OR800	*M. mulatta*	India	Male
OR273	*M. mulatta*	India	Male



Extract SRR files using SRA Toolkit Version 2.8
Alignment and sorting using BWA 0.7.12 and Samtools version 1.2
Adding read group library information PICARD tools version 1.79
Merging files using PICARD tools version 1.79
Sorting the newly merged bam with PICARD tools version 1.79
Marking the Duplicates using Picard tools version 1.79
Indel realignment using GATK version 3.6
Haplotype caller GATK version 3.6
BQSR GATK version 3.6
VCF compare Version 0.1.14
Repeat until 99% alike
Apply recalibration GATK version 3.6
Lumpy
CNVnator
Delly
Breakdancer
Pindel
Custom Python Script
Variant Analysis

Figure 3:Overview of genome preparation pipeline from raw genome file download with Sratoolkit to merged structural variant analysis. Pipeline primarily followed GATK best practices with minor revisions.

Table 2: Raw structural variant calls before intersection and filtering from Breakdancer, Lumpy, Delly, Pindel, and CNVnator for both M. mulatta and M. fascicularis.

	Inversion	Deletion	Duplication	Insertion	Total
M. mulatta	21,528	677,526	86,282	3,018,925	3,726,607
M. fascicularis	5,880	498,719	71,621	1	576,220

Table 3: Merged structural variant calls after filtering. Variants were filtered for size into bins of small (<1kb), medium (1 kb - 10 kb), large (10 kb - 1 Mb) and extra-large (>1Mb) and at least three program hits. A drastic reduction in SV calls was noted after intersection. *M. mulatta* and *M. fascicularis* were used respectively.

M. mulatta	Inversion	Deletion	Insertion	Duplication	Total
Small (1kb)	113	160	0	0	273
Medium (1-10kb)	2,674	3,947	0	0	6,621
Large (10kb-1mb)	5,304	3,340	0	0	8,604
Extra Large (1mb+)	21	22	0	0	43
M. fascicularis	Inversion	Deletion	Insertion	Duplication	Total
Small (1kb)	640	0	0	0	640
Medium (1-10kb)	18	2	0	1,532	1,552
Large (10kb-1mb)	497	306	0	431	1,234
Extra Large (1mb+)	647	0	0	0	647

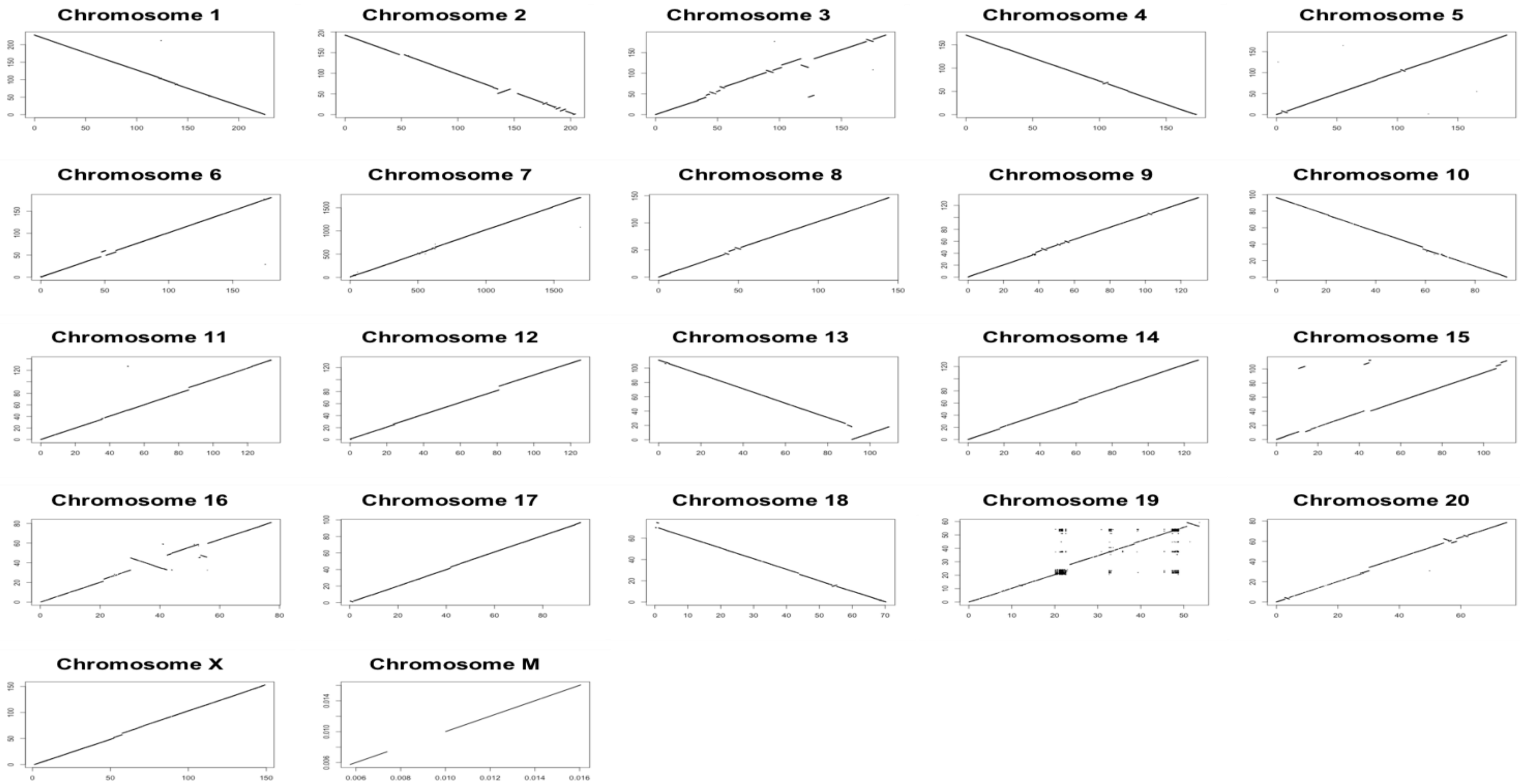


Figure 4::Comparison of reference genomes between rheMac8 and macFas5 created with lastZ, alignments were filtered for 95% mapping quality and plotted using ggplot2.

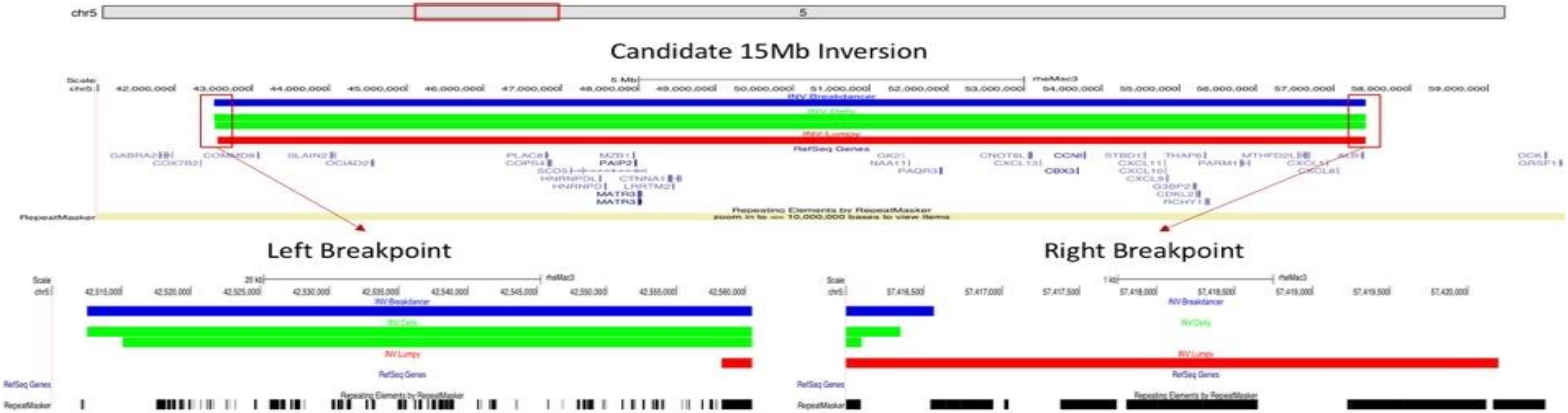


Figure 5 :UCSC genome browser custom track of putative 15mb inversion from 42mb-57mb not found in M. mulatta. Whole inversion can be viewed in(top) and the left and right breakpoints respectively below. Breakdancer can be seen in blue, Delly in green and Lumpy in red

Table 4: Comparison of Structural variation call set between the aligners BWA and Bowtie2 provides an insight into differing algorithms and potential downstream bias.

	BWA	Bowtie2
Type		
Inversion	***16,079***	***30,137***
Deletion	***84,301***	***104,703***
Insertion	***119,767***	***36,466***
Duplication	0	***0***

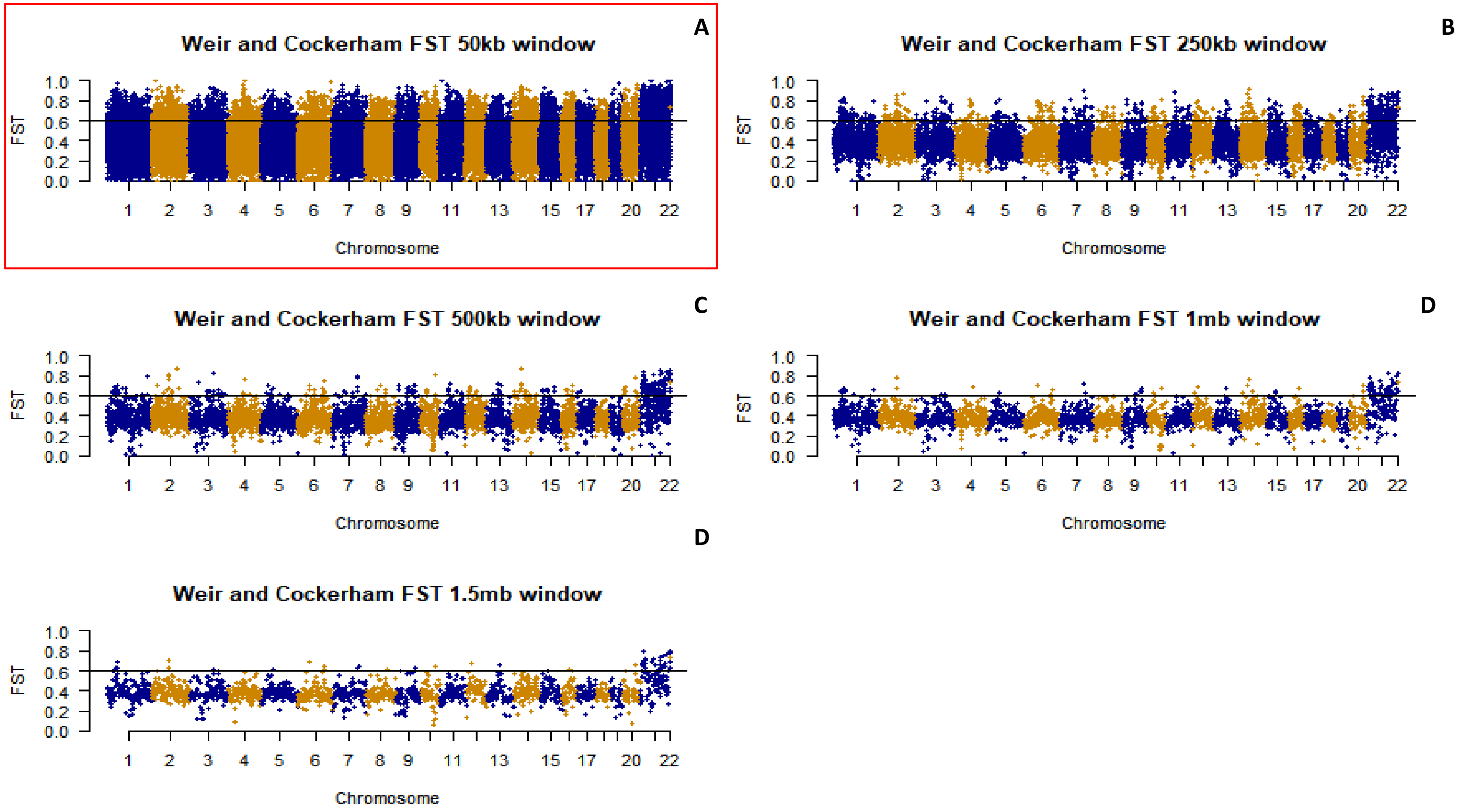
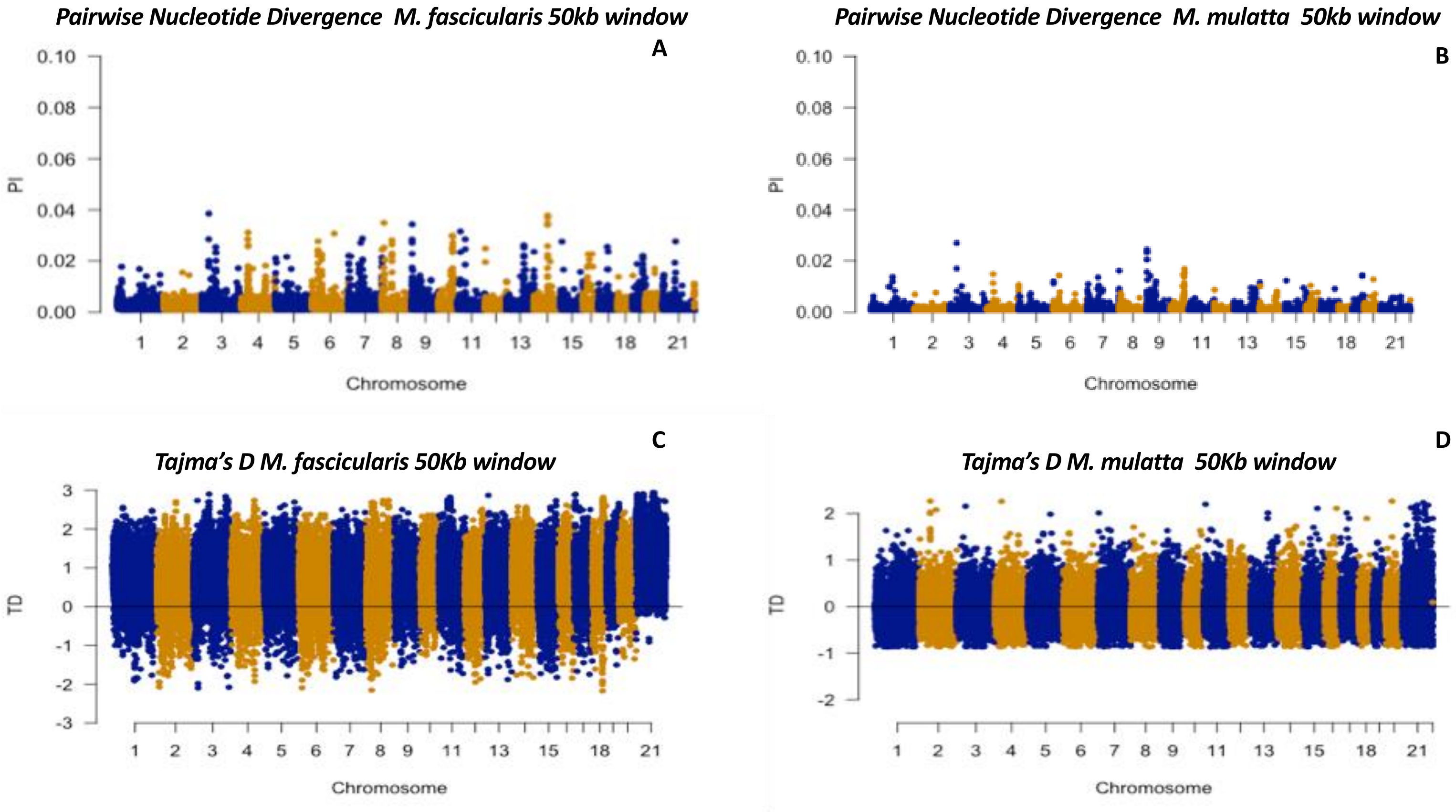


Figure 6:Weir and Cockrams Fst calculated between M. mulatta and M. fascicularis calculated at varying bin sizes(50kb,250kb,500kb,1mb & 1.5mb). 50kb bin size was chosen for further analysis to reduce confounding effect of F_{ST} outliers on the X chromosome. A 1% cutoff was added to filter only the regions with the greatest outliers. A- 50kb window , B – 250kb window , C- 500kb window, D – 1Mb window , E- 1.5mb window.



Pairwise Nucleotide Divergence M. fascicularis 50kb window

Pairwise Nucleotide Divergence M. mulatta 50kb window

Tajma's D M. fascicularis 50Kb window

Tajma's D M. mulatta 50Kb window

Figure 7: (A) Pairwise nucleotide divergence was calculated for M . fascicularis .Areas of low PI were extracted to be used alongside areas of high FST. PI was plotted along the genome using R and is bound from 0<->1. (B) Pairwise nucleotide divergence for M. mulatta . (C) Tajimas D calculated in 50kb bin size using Vcftools for M . fascicularis , visualized using R, Abline at 0 as negative values were ones of interest. (D) : Tajimas D calculated in 50kb bin size using Vcftools for M. mulatta , visualized using R, Abline at 0 as negative values were ones of interest.

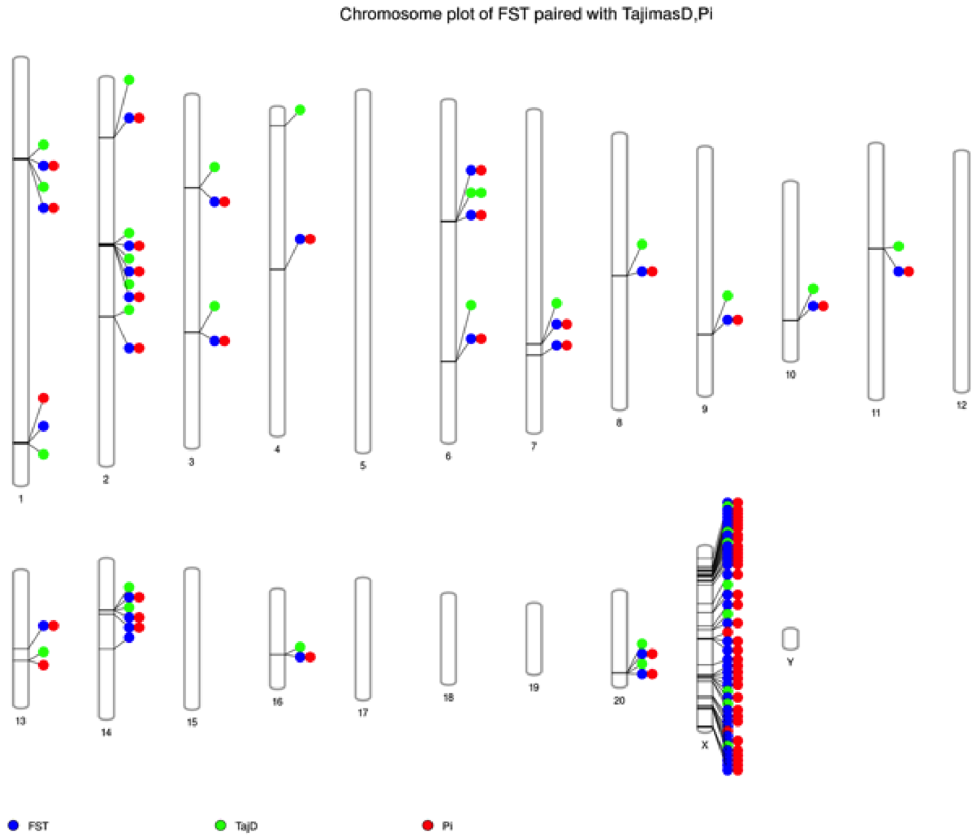


Figure 8:: Karyoplot showing where areas of High FST , low PI and negative Tajima's D lie across the genome. The massive amount of overlap on the X chromosome may hint at a faster X effect in this syste

Table 5: Ten top and bottom gene family functions reported from DAVID, Top 1% FST outliers can be seen on the left, and the bottom 1% outliers can be seen on the right. Inputs were region matched and controlled against the whole genome.

Top 1% gene families	**Bottom 1% gene families**
Sperm	Metal Binding
Fertilization	Cell recognition
Reproductive processes	Cytokines
Immunoglobin	K/Na pumps
Metabolism	Transmembrane protein
Biosynthetic Pathways	Cellular composition
Metabolic Pathways	Transmembrane signal transduction
RNA Regulation	Zinc Finger Motif
Gene Expression Regulation	Gland Development
Mitochondrial transport	Cell Development

Table 6:Table 6: Top 1 percent gene ontology outputs from DAVID gene ontology

Term	Fold Enrichment	P Value	Bonferroni	FDR	Fisher exact
DNA-binding region:Nuclear receptor	75.2	1.30E-02	2.00E-01	9.20E+00	1.30E-04
region of interest:Ligand-binding	75.2	1.30E-02	2.00E-01	9.20E+00	1.30E-04
organelle part	1.4	2.30E-02	2.80E-01	1.50E+01	1.60E-02
intracellular organelle part	1.4	3.10E-02	8.10E-01	2.60E+01	2.20E-02
cytoplasm	1.7	1.20E-02	6.20E-01	1.20E+01	6.50E-03
sperm part	7.3	6.20E-02	9.70E-01	4.60E+01	8.00E-03
organelle part	1.4	2.30E-02	7.10E-01	2.00E+01	1.60E-02
intracellular organelle	1.2	9.90E-02	1.00E+00	6.40E+01	8.40E-02
microtubule cytoskeleton	2.5	3.80E-02	1.00E+00	3.90E+01	1.50E-02
intracellular organelle part	1.4	3.10E-02	1.00E+00	3.30E+01	2.20E-02
sperm part	7.3	6.20E-02	1.00E+00	5.60E+01	8.00E-03
sperm part	7.3	6.20E-02	1.00E+00	5.20E+01	8.10E-03
organelle part	1.4	2.30E-02	1.00E+00	2.60E+01	1.60E-02
cytoskeletal part	2.1	6.00E-02	1.00E+00	5.50E+01	2.70E-02
cytoskeleton	1.8	9.40E-02	1.00E+00	7.20E+01	4.90E-02
secretory vesicle	4.5	2.40E-02	9.80E-01	2.50E+01	5.20E-03
cytoskeletal part	2.1	6.00E-02	1.00E+00	5.10E+01	2.70E-02
intracellular organelle part	1.4	3.10E-02	9.80E-01	3.00E+01	2.20E-02
secretory vesicle	4.7	2.10E-02	9.80E-01	2.30E+01	4.20E-03
nucleoplasm	1.8	9.30E-02	1.00E+00	6.50E+01	4.80E-02
secretory vesicle	4.9	1.70E-02	9.80E-01	2.00E+01	3.40E-03
sperm-egg recognition	19.2	9.70E-02	1.00E+00	7.50E+01	4.80E-03
sperm part	6.9	6.80E-02	1.00E+00	5.80E+01	9.40E-03
regulation of cellular localization	2.5	8.40E-02	1.00E+00	7.00E+01	3.10E-02
sperm-egg recognition	19.6	9.60E-02	1.00E+00	7.70E+01	4.60E-03
macroautophagy	3.6	9.70E-02	1.00E+00	7.50E+01	2.50E-02
microtubule cytoskeleton	2.3	4.80E-02	1.00E+00	4.60E+01	2.00E-02
macromolecule biosynthetic process	1.4	9.20E-02	1.00E+00	7.60E+01	6.30E-02
microtubule cytoskeleton	2.2	6.00E-02	1.00E+00	5.20E+01	2.50E-02
positive regulation of cellular protein localization	3.8	8.40E-02	1.00E+00	6.30E+01	2.10E-02
establishment of protein localization to organelle	2.9	8.80E-02	1.00E+00	7.40E+01	2.80E-02
cellular macromolecule biosynthetic process	1.4	7.50E-02	1.00E+00	6.80E+01	5.00E-02
cytoskeletal part	1.9	9.50E-02	1.00E+00	7.00E+01	4.70E-02
single-organism organelle organization	1.8	8.20E-02	1.00E+00	6.90E+01	4.20E-02
binding of sperm to zona pellucida	21.4	8.80E-02	1.00E+00	7.40E+01	3.90E-03
single-organism organelle organization	1.9	7.40E-02	1.00E+00	6.80E+01	3.70E-02
regulation of intracellular protein transport	3.5	1.00E-01	1.00E+00	7.90E+01	2.70E-02
positive regulation of cellular protein localization	3.9	8.10E-02	1.00E+00	6.90E+01	2.00E-02
regulation of establishment of protein localization	2.6	7.20E-02	1.00E+00	6.70E+01	2.60E-02
cytoskeletal part	2	8.00E-02	1.00E+00	6.30E+01	3.80E-02
sperm-egg recognition	19	9.80E-02	1.00E+00	7.90E+01	4.90E-03
pallium development	6.8	7.00E-02	1.00E+00	6.30E+01	9.80E-03

pallium development	6.9	6.80E-02	1.00E+00	6.40E+01	9.30E-03
cellular macromolecule biosynthetic process	1.4	9.60E-02	1.00E+00	7.80E+01	6.60E-02
protein autophosphorylation	6.1	8.30E-02	1.00E+00	6.70E+01	1.30E-02
cerebral cortex development	9.5	3.90E-02	1.00E+00	4.20E+01	3.90E-03
autophagy	3.2	6.40E-02	1.00E+00	5.30E+01	1.90E-02
regulation of intracellular protein transport	3.5	9.90E-02	1.00E+00	8.20E+01	2.60E-02
fertilization	7.2	6.30E-02	1.00E+00	6.20E+01	8.30E-03
macroautophagy	3.6	9.80E-02	1.00E+00	8.20E+01	2.60E-02
establishment of protein localization to organelle	2.8	9.60E-02	1.00E+00	7.80E+01	3.20E-02
IQ motif, EF-hand binding site	9.3	4.10E-02	1.00E+00	4.00E+01	4.20E-03
olymerase II transcription factor activity, ligand-activated sequence-specific DNA	20.7	9.00E-02	1.00E+00	6.40E+01	4.10E-03
sperm-egg recognition	19.1	9.80E-02	1.00E+00	8.20E+01	4.80E-03
fertilization	7.1	6.50E-02	1.00E+00	6.00E+01	8.80E-03
binding of sperm to zona pellucida	20.7	9.00E-02	1.00E+00	7.60E+01	4.10E-03
cellular macromolecule biosynthetic process	1.4	9.50E-02	1.00E+00	8.10E+01	6.50E-02
positive regulation of cellular protein localization	3.8	8.30E-02	1.00E+00	7.30E+01	2.00E-02
sperm-egg recognition	18.8	1.00E-01	1.00E+00	8.20E+01	5.00E-03
establishment of protein localization to organelle	2.8	9.50E-02	1.00E+00	8.00E+01	3.10E-02
protein stabilization	7.5	5.90E-02	1.00E+00	5.90E+01	7.40E-03
regulation of protein export from nucleus	20.8	9.00E-02	1.00E+00	7.90E+01	4.10E-03
binding of sperm to zona pellucida	20.8	9.00E-02	1.00E+00	7.90E+01	4.10E-03
regulation of cellular localization	2.5	8.50E-02	1.00E+00	7.70E+01	3.20E-02
single-organism organelle organization	1.8	9.30E-02	1.00E+00	7.90E+01	4.80E-02
establishment of protein localization to organelle	2.8	1.00E-01	1.00E+00	8.20E+01	3.30E-02
regulation of establishment of protein localization	2.6	7.90E-02	1.00E+00	7.10E+01	2.90E-02
binding of sperm to zona pellucida	24.1	7.80E-02	1.00E+00	6.50E+01	3.00E-03
single-organism organelle organization	1.8	8.50E-02	1.00E+00	7.70E+01	4.40E-02
binding of sperm to zona pellucida	20.5	9.20E-02	1.00E+00	7.90E+01	4.20E-03
regulation of protein export from nucleus	20.5	9.20E-02	1.00E+00	7.90E+01	4.20E-03
mitochondrion organization	2.9	5.00E-02	1.00E+00	5.30E+01	1.60E-02
positive regulation of cellular protein localization	3.8	8.30E-02	1.00E+00	7.60E+01	2.00E-02
intracellular organelle part	1.3	6.80E-02	1.00E+00	5.70E+01	5.10E-02
regulation of cellular localization	2.5	9.10E-02	1.00E+00	7.90E+01	3.40E-02
regulation of establishment of protein localization	2.6	7.80E-02	1.00E+00	7.40E+01	2.90E-02
positive regulation of mitochondrion organization	6.4	7.70E-02	1.00E+00	7.00E+01	1.10E-02
single fertilization	9.7	3.70E-02	1.00E+00	4.30E+01	3.70E-03
cerebral cortex development	9.7	3.70E-02	1.00E+00	4.30E+01	3.70E-03
positive regulation of cellular protein localization	3.8	8.60E-02	1.00E+00	7.70E+01	2.10E-02
organelle organization	1.6	3.10E-02	1.00E+00	3.50E+01	1.80E-02
positive regulation of mitochondrion organization	6.4	7.70E-02	1.00E+00	7.30E+01	1.10E-02
regulation of establishment of protein localization	2.5	8.30E-02	1.00E+00	7.60E+01	3.10E-02
pallium development	6.7	7.10E-02	1.00E+00	7.00E+01	9.90E-03
regulation of cellular protein localization	3.7	2.10E-02	1.00E+00	2.60E+01	5.30E-03
pallium development	6.7	7.10E-02	1.00E+00	6.70E+01	1.00E-02
positive regulation of mitochondrion organization	6.3	7.90E-02	1.00E+00	7.40E+01	1.20E-02
fertilization	7	6.60E-02	1.00E+00	6.70E+01	8.90E-03
pallium development	6.6	7.30E-02	1.00E+00	7.10E+01	1.00E-02
cerebral cortex development	9.4	3.90E-02	1.00E+00	4.50E+01	4.00E-03
single fertilization	9.4	3.90E-02	1.00E+00	4.50E+01	4.00E-03
autophagy	3.3	6.30E-02	1.00E+00	6.60E+01	1.80E-02
fertilization	6.9	6.80E-02	1.00E+00	6.80E+01	9.40E-03
protein stabilization	7.3	6.10E-02	1.00E+00	5.60E+01	8.00E-03
fertilization	7	6.60E-02	1.00E+00	6.40E+01	9.00E-03
regulation of cellular protein localization	3.6	2.30E-02	1.00E+00	3.00E+01	6.10E-03
protein stabilization	7.3	6.10E-02	1.00E+00	6.40E+01	8.00E-03
autophagy	3.2	6.70E-02	1.00E+00	6.80E+01	2.00E-02
cerebral cortex development	9.4	3.90E-02	1.00E+00	4.80E+01	3.90E-03
single fertilization	9.4	3.90E-02	1.00E+00	4.80E+01	3.90E-03
mitochondrion organization	2.8	5.50E-02	1.00E+00	6.10E+01	1.80E-02
protein stabilization	7.2	6.30E-02	1.00E+00	6.50E+01	8.30E-03
organelle organization	1.6	3.30E-02	1.00E+00	4.30E+01	2.00E-02
mitochondrion organization	2.8	5.90E-02	1.00E+00	6.30E+01	2.00E-02
single fertilization	9.3	4.00E-02	1.00E+00	4.90E+01	4.10E-03
cerebral cortex development	9.3	4.00E-02	1.00E+00	4.90E+01	4.10E-03
regulation of cellular protein localization	3.6	2.30E-02	1.00E+00	3.20E+01	6.00E-03
regulation of cellular protein localization	3.6	2.40E-02	1.00E+00	3.30E+01	6.50E-03

Table 7: Bottom one percent gene families from DAVID gene ontology

Term	Fold Enrichment	P Value	Bonferroni	FDR	Fisher exact
Potassium channel, voltage dependent, Kv1	95.3	4.00E-04	8.00E-02	5.10E-01	3.00E-06
axon	7.3	1.30E-03	1.80E-01	1.50E+00	1.70E-04
axon	7.2	1.30E-03	2.80E-01	1.70E+00	1.80E-04
axon	6.6	1.90E-03	3.10E-01	2.30E+00	2.80E-04
axon	6.6	1.80E-03	2.60E-01	2.30E+00	2.70E-04
Metal-binding	2.5	3.30E-03	1.90E-01	3.30E+00	1.30E-03
negative regulation of multicellular organismal process	3.1	4.40E-03	9.90E-01	4.80E+00	1.40E-03
negative regulation of multicellular organismal process	3.1	3.70E-03	1.00E+00	5.90E+00	1.10E-03
Potassium channel, voltage dependent, Kv	27.8	5.00E-03	3.80E-01	6.20E+00	1.70E-04
negative regulation of multicellular organismal process	3	4.60E-03	8.30E-01	6.20E+00	1.50E-03
Zinc	2.7	6.30E-03	6.50E-01	6.30E+00	2.20E-03
negative regulation of multicellular organismal process	3.1	4.20E-03	3.40E-01	6.80E+00	1.30E-03
Potassium channel	22.3	7.70E-03	6.50E-01	7.60E+00	3.20E-04
main axon	24	6.70E-03	8.10E-01	7.70E+00	2.60E-04
main axon	23.9	6.70E-03	4.00E-01	8.40E+00	2.60E-04
main axon	21.9	7.90E-03	8.00E-01	9.20E+00	3.40E-04
main axon	21.9	7.80E-03	7.10E-01	9.40E+00	3.30E-04
neuron projection	3.8	9.40E-03	7.00E-01	1.00E+01	2.40E-03
neuron projection	3.7	9.80E-03	7.80E-01	1.10E+01	2.50E-03
neuron projection	3.7	1.00E-02	9.20E-01	1.20E+01	2.60E-03
protein homodimerization activity	14.8	1.70E-02	9.50E-01	1.60E+01	1.10E-03
neuron projection	3.4	1.40E-02	9.80E-01	1.70E+01	4.00E-03
protein homodimerization activity	14	1.90E-02	7.40E-01	1.70E+01	1.30E-03
protein dimerization activity	13.1	2.10E-02	9.80E-01	1.90E+01	1.50E-03
Potassium transport	13.3	2.10E-02	9.80E-01	1.90E+01	1.50E-03
protein homodimerization activity	14.8	1.70E-02	7.00E-01	2.00E+01	1.10E-03
neuron part	2.8	2.20E-02	9.90E-01	2.00E+01	7.50E-03
protein homodimerization activity	14.6	1.70E-02	9.90E-01	2.00E+01	1.10E-03
Potassium channel tetramerisation-type BTB domain	14.5	1.80E-02	7.50E-01	2.00E+01	1.20E-03
protein dimerization activity	14	1.90E-02	6.90E-01	2.20E+01	1.30E-03
protein dimerization activity	13.8	1.90E-02	9.60E-01	2.20E+01	1.30E-03
nucleus	1.4	2.10E-02	9.90E-01	2.30E+01	1.50E-02
neuron part	2.8	2.20E-02	7.10E-01	2.30E+01	7.60E-03
Voltage-dependent potassium channel, four helix bundle domain	13.1	2.20E-02	9.40E-01	2.40E+01	1.60E-03
Potassium	11.5	2.70E-02	1.00E+00	2.50E+01	2.20E-03
voltage-gated potassium channel complex	12	2.50E-02	1.00E+00	2.50E+01	2.00E-03
identical protein binding	11.2	2.80E-02	9.20E-01	2.50E+01	2.40E-03
axolemma	74.4	2.60E-02	1.00E+00	2.50E+01	2.90E-04
binding	1.2	4.90E-02	9.80E-01	2.60E+01	4.00E-02
Salmonella infection	10.8	2.80E-02	1.00E+00	2.60E+01	2.50E-03
neuron part	2.7	2.40E-02	9.80E-01	2.70E+01	8.20E-03
voltage-gated potassium channel complex	11.8	2.60E-02	1.00E+00	2.70E+01	2.10E-03
Zinc-finger	2.6	3.00E-02	1.00E+00	2.70E+01	1.10E-02
axolemma	74.6	2.60E-02	9.20E-01	2.70E+01	2.90E-04
identical protein binding	11.9	2.50E-02	8.40E-01	2.80E+01	2.00E-03
nucleus	1.4	2.50E-02	1.00E+00	2.80E+01	1.80E-02
identical protein binding	11.7	2.60E-02	7.90E-01	2.90E+01	2.10E-03
voltage-gated potassium channel complex	11.7	2.60E-02	1.00E+00	2.90E+01	2.10E-03
axolemma	74.4	2.60E-02	8.90E-01	2.90E+01	2.90E-04
axolemma	68	2.90E-02	9.90E-01	2.90E+01	3.50E-04
axolemma	68.3	2.80E-02	1.00E+00	3.00E+01	3.50E-04
potassium channel complex	11.3	2.80E-02	8.70E-01	3.10E+01	2.30E-03
voltage-gated potassium channel complex	10.7	3.10E-02	9.90E-01	3.10E+01	2.70E-03
Voltage-gated channel	9.9	3.60E-02	1.00E+00	3.10E+01	3.50E-03
voltage-gated potassium channel complex	10.8	3.00E-02	1.00E+00	3.20E+01	2.70E-03
potassium channel complex	10.4	3.20E-02	1.00E+00	3.40E+01	2.90E-03
integral component of plasma membrane	2.3	3.60E-02	1.00E+00	3.50E+01	1.50E-02
neuron part	2.5	3.40E-02	1.00E+00	3.50E+01	1.30E-02
domain:B30.2/SPRY	28.7	5.90E-02	9.90E-01	3.60E+01	1.40E-03
membrane-enclosed lumen	1.5	6.10E-02	1.00E+00	3.60E+01	3.80E-02
integral component of plasma membrane	2.3	3.70E-02	9.10E-01	3.70E+01	1.50E-02
negative regulation of axon extension	54.3	3.60E-02	1.00E+00	3.80E+01	5.70E-04

BTB/POZ-like	5.4	3.70E-02	1.00E+00	3.80E+01	6.50E-03
cell projection	2.2	4.80E-02	1.00E+00	3.80E+01	2.10E-02
organelle lumen	1.6	4.80E-02	1.00E+00	3.80E+01	2.90E-02
BTB	4.8	4.80E-02	1.00E+00	3.90E+01	9.60E-03
membrane-bounded organelle	1.2	4.90E-02	1.00E+00	3.90E+01	4.10E-02
ion channel complex	5.1	4.10E-02	1.00E+00	3.90E+01	7.50E-03
neuron projection membrane	45.2	4.30E-02	1.00E+00	4.00E+01	8.50E-04
integral component of plasma membrane	2.3	3.90E-02	1.00E+00	4.00E+01	1.60E-02
nuclear part	1.6	4.40E-02	1.00E+00	4.10E+01	2.70E-02
neuron projection membrane	44.8	4.30E-02	1.00E+00	4.10E+01	8.60E-04
negative regulation of developmental growth	9.6	3.80E-02	1.00E+00	4.10E+01	3.70E-03
regulation of cytokine production	3.2	3.80E-02	1.00E+00	4.10E+01	1.10E-02
Ion channel	4.7	5.20E-02	1.00E+00	4.20E+01	1.00E-02
voltage-gated potassium channel activity	9.1	4.10E-02	1.00E+00	4.20E+01	4.30E-03
ion channel complex	5.1	4.10E-02	1.00E+00	4.20E+01	7.70E-03
intrinsic component of plasma membrane	2.2	4.70E-02	1.00E+00	4.30E+01	2.00E-02
poly(A) RNA binding	2.3	5.10E-02	1.00E+00	4.30E+01	2.10E-02
BTB/POZ fold	5.1	4.30E-02	1.00E+00	4.30E+01	8.10E-03
regulation of cytokine production	3.3	3.40E-02	1.00E+00	4.30E+01	9.70E-03
voltage-gated potassium channel activity	8.9	4.30E-02	1.00E+00	4.30E+01	4.50E-03
cell projection	2.2	4.80E-02	1.00E+00	4.30E+01	2.10E-02
intracellular organelle lumen	1.6	4.80E-02	1.00E+00	4.30E+01	2.90E-02
organelle lumen	1.6	4.80E-02	1.00E+00	4.30E+01	2.90E-02
nuclear part	1.5	4.70E-02	1.00E+00	4.40E+01	2.90E-02
neuron projection membrane	44.6	4.30E-02	1.00E+00	4.40E+01	8.70E-04
negative regulation of developmental growth	9.6	3.80E-02	1.00E+00	4.40E+01	3.70E-03
negative regulation of developmental growth	9.7	3.70E-02	1.00E+00	4.40E+01	3.70E-03
neuron projection membrane	40.8	4.70E-02	1.00E+00	4.40E+01	1.00E-03
morphogenesis of a branching structure	5.1	4.20E-02	9.40E-01	4.40E+01	7.70E-03
negative regulation of developmental growth	9.9	3.60E-02	9.40E-01	4.50E+01	3.40E-03
intrinsic component of plasma membrane	2.2	4.90E-02	1.00E+00	4.50E+01	2.10E-02
neuron projection membrane	41	4.70E-02	1.00E+00	4.50E+01	1.00E-03
regulation of cytokine production	3.2	3.60E-02	1.00E+00	4.50E+01	1.10E-02
transmembrane transporter complex	4.7	5.20E-02	1.00E+00	4.60E+01	1.00E-02
negative regulation of developmental growth	9.7	3.70E-02	9.40E-01	4.60E+01	3.60E-03
intracellular organelle lumen	1.6	5.10E-02	1.00E+00	4.70E+01	3.10E-02
axon part	8	5.30E-02	3.90E-01	4.70E+01	6.30E-03
ion channel complex	4.7	5.10E-02	1.00E+00	4.70E+01	1.00E-02
morphogenesis of a branching structure	5.1	4.20E-02	9.40E-01	4.70E+01	7.70E-03
cell recognition	8.7	4.50E-02	1.00E+00	4.70E+01	4.90E-03
morphogenesis of a branching structure	5.3	3.80E-02	1.00E+00	4.70E+01	6.90E-03
binding	1.2	4.90E-02	1.00E+00	4.80E+01	4.00E-02
ion channel complex	4.7	5.00E-02	1.00E+00	4.80E+01	1.00E-02
transmembrane transporter complex	4.6	5.30E-02	1.00E+00	4.80E+01	1.10E-02
nuclear lumen	1.6	5.30E-02	9.90E-01	4.80E+01	3.20E-02
transporter complex	4.6	5.50E-02	1.00E+00	4.80E+01	1.10E-02
regulation of biological process	1.2	8.20E-02	1.00E+00	4.80E+01	6.80E-02
axon part	7.9	5.40E-02	9.70E-01	4.80E+01	6.50E-03
nuclear part	1.5	5.00E-02	1.00E+00	4.90E+01	3.10E-02
intrinsic component of plasma membrane	2.2	5.00E-02	1.00E+00	4.90E+01	2.20E-02
morphogenesis of a branching structure	5.2	4.00E-02	1.00E+00	4.90E+01	7.40E-03
cell projection	2.2	5.10E-02	1.00E+00	5.00E+01	2.20E-02
axonal fasciculation	36.2	5.30E-02	1.00E+00	5.10E+01	1.30E-03
branching involved in mammary gland duct morphogenesis	36.2	5.30E-02	1.00E+00	5.10E+01	1.30E-03
transmembrane transporter complex	4.6	5.30E-02	1.00E+00	5.10E+01	1.10E-02
cell recognition	9	4.30E-02	1.00E+00	5.10E+01	4.50E-03
cytokine production	2.9	5.00E-02	1.00E+00	5.10E+01	1.60E-02
axon part	7.9	5.40E-02	1.00E+00	5.10E+01	6.50E-03
integral component of plasma membrane	2.1	5.80E-02	1.00E+00	5.10E+01	2.60E-02
plasma membrane part	1.7	7.10E-02	1.00E+00	5.20E+01	3.80E-02
integral component of plasma membrane	2.1	5.60E-02	1.00E+00	5.20E+01	2.50E-02
organelle lumen	1.5	5.50E-02	1.00E+00	5.20E+01	3.30E-02
intracellular organelle lumen	1.5	5.50E-02	1.00E+00	5.20E+01	3.30E-02
cell recognition	8.8	4.40E-02	1.00E+00	5.20E+01	4.80E-03
zinc finger region:RING-type	17.2	9.60E-02	1.00E+00	5.30E+01	4.50E-03
nuclear lumen	1.6	5.60E-02	1.00E+00	5.30E+01	3.40E-02
cytokine production	3	4.50E-02	1.00E+00	5.30E+01	1.40E-02
transporter complex	4.5	5.70E-02	6.60E-01	5.30E+01	1.20E-02
axon	7	6.70E-02	1.00E+00	5.40E+01	9.10E-03
axon part	7.2	6.30E-02	6.40E-01	5.40E+01	8.30E-03
regulation of multicellular organismal process	1.6	6.70E-02	1.00E+00	5.40E+01	3.90E-02
protein binding	1.9	8.80E-02	1.00E+00	5.50E+01	4.30E-02
axon part	7.2	6.20E-02	1.00E+00	5.60E+01	8.20E-03

cytokine production	3	4.80E-02	1.00E+00	5.60E+01	1.50E-02
transmembrane transporter complex	4.2	6.60E-02	1.00E+00	5.60E+01	1.50E-02
membrane-enclosed lumen	1.5	6.10E-02	1.00E+00	5.60E+01	3.80E-02
nucleus	1.3	6.70E-02	1.00E+00	5.70E+01	5.00E-02
Ion transport domain	7.2	6.40E-02	1.00E+00	5.70E+01	8.50E-03
transmembrane transporter complex	4.2	6.50E-02	1.00E+00	5.70E+01	1.40E-02
plasma membrane part	1.7	7.10E-02	1.00E+00	5.70E+01	3.80E-02
interleukin-6 production	7.7	5.70E-02	1.00E+00	5.80E+01	7.00E-03
transporter complex	4.1	6.90E-02	1.00E+00	5.90E+01	1.60E-02
interleukin-6 production	7.9	5.40E-02	1.00E+00	5.90E+01	6.40E-03
plasma membrane part	1.7	7.50E-02	1.00E+00	6.10E+01	4.10E-02
potassium channel activity	6.8	6.90E-02	1.00E+00	6.10E+01	9.70E-03
intrinsic component of plasma membrane	2	7.50E-02	1.00E+00	6.10E+01	3.50E-02
interleukin-6 production	7.8	5.50E-02	1.00E+00	6.10E+01	6.80E-03
intrinsic component of plasma membrane	2	7.20E-02	1.00E+00	6.10E+01	3.40E-02
plasma membrane protein complex	3	7.90E-02	1.00E+00	6.10E+01	2.40E-02
potassium channel activity	6.7	7.10E-02	1.00E+00	6.10E+01	1.00E-02
membrane-bounded organelle	1.2	7.10E-02	1.00E+00	6.20E+01	5.90E-02
axonal fasciculation	30.8	6.20E-02	1.00E+00	6.20E+01	1.90E-03
gated channel activity	3.6	9.50E-02	1.00E+00	6.30E+01	2.50E-02
plasma membrane protein complex	3	8.10E-02	1.00E+00	6.40E+01	2.50E-02
axonal fasciculation	31.6	6.00E-02	1.00E+00	6.40E+01	1.80E-03
regulation of catecholamine secretion	28.6	6.70E-02	1.00E+00	6.40E+01	2.20E-03
poly(A) RNA binding	1.9	9.90E-02	1.00E+00	6.40E+01	4.70E-02
cell part	1.1	1.00E-01	1.00E+00	6.50E+01	9.20E-02
telencephalon glial cell migration	28.8	6.60E-02	1.00E+00	6.50E+01	2.10E-03
regulation of catecholamine secretion	28.8	6.60E-02	9.80E-01	6.50E+01	2.10E-03
plasma membrane part	1.7	7.70E-02	1.00E+00	6.50E+01	4.20E-02
axonal fasciculation	31.1	6.10E-02	1.00E+00	6.50E+01	1.80E-03
negative regulation of developmental process	2.5	8.90E-02	1.00E+00	6.50E+01	3.30E-02
cation channel activity	3.9	7.90E-02	1.00E+00	6.50E+01	1.90E-02
voltage-gated cation channel activity	6.3	7.90E-02	1.00E+00	6.50E+01	1.20E-02
gland morphogenesis	6.8	6.90E-02	1.00E+00	6.60E+01	9.60E-03
gland morphogenesis	6.9	6.80E-02	1.00E+00	6.60E+01	9.40E-03
protein binding	1.9	8.00E-02	1.00E+00	6.60E+01	3.80E-02
Calcium signaling pathway	5.4	9.80E-02	1.00E+00	6.60E+01	1.70E-02
voltage-gated cation channel activity	6.2	8.10E-02	1.00E+00	6.60E+01	1.20E-02
nucleolus	2.4	9.40E-02	1.00E+00	6.60E+01	3.60E-02
cation channel activity	3.8	8.20E-02	1.00E+00	6.60E+01	2.00E-02
telencephalon glial cell migration	29.5	6.50E-02	1.00E+00	6.70E+01	2.00E-03
cerebral cortex radial glia guided migration	29.5	6.50E-02	1.00E+00	6.70E+01	2.00E-03
regulation of catecholamine secretion	29.5	6.50E-02	1.00E+00	6.70E+01	2.00E-03
catecholamine secretion	26.8	7.10E-02	1.00E+00	6.70E+01	2.50E-03
gland morphogenesis	7.1	6.50E-02	1.00E+00	6.70E+01	8.80E-03
regulation of multicellular organismal process	1.6	6.50E-02	1.00E+00	6.70E+01	3.70E-02
plasma membrane protein complex	3	8.20E-02	1.00E+00	6.70E+01	2.60E-02
telencephalon glial cell migration	29	6.60E-02	1.00E+00	6.70E+01	2.10E-03
cerebral cortex radial glia guided migration	29	6.60E-02	1.00E+00	6.70E+01	2.10E-03
regulation of catecholamine secretion	29	6.60E-02	8.30E-01	6.70E+01	2.10E-03
gland morphogenesis	6.9	6.70E-02	1.00E+00	6.80E+01	9.20E-03
catecholamine secretion	27.6	6.90E-02	1.00E+00	6.90E+01	2.30E-03
catecholamine secretion	27.2	7.00E-02	1.00E+00	7.00E+01	2.40E-03
intracellular membrane-bounded organelle	1.2	9.90E-02	1.00E+00	7.00E+01	8.10E-02
gated channel activity	3.7	9.00E-02	1.00E+00	7.10E+01	2.30E-02
nuclear part	1.4	9.70E-02	1.00E+00	7.10E+01	6.50E-02
poly(A) RNA binding	2	9.00E-02	1.00E+00	7.10E+01	4.20E-02
Regulator of G-protein signaling, domain 1	20.2	9.30E-02	9.40E-01	7.10E+01	4.30E-03
neuron recognition	23.8	7.90E-02	1.00E+00	7.10E+01	3.10E-03
cation channel complex	5.8	9.10E-02	1.00E+00	7.10E+01	1.50E-02
neuron recognition	24	7.90E-02	1.00E+00	7.10E+01	3.10E-03
gated channel activity	3.6	9.30E-02	1.00E+00	7.20E+01	2.40E-02
potassium ion transmembrane transporter activity	5.7	9.30E-02	1.00E+00	7.20E+01	1.50E-02
potassium ion transmembrane transporter activity	5.6	9.60E-02	1.00E+00	7.30E+01	1.60E-02
poly(A) RNA binding	2	9.70E-02	1.00E+00	7.30E+01	4.50E-02
neuron recognition	24.6	7.70E-02	1.00E+00	7.30E+01	2.90E-03
negative regulation of histone methylation	24.6	7.70E-02	1.00E+00	7.30E+01	2.90E-03
negative regulation of developmental process	2.5	9.10E-02	1.00E+00	7.30E+01	3.40E-02
neuron recognition	24.2	7.80E-02	1.00E+00	7.40E+01	3.00E-03
negative regulation of histone methylation	24.2	7.80E-02	1.00E+00	7.40E+01	3.00E-03
branching involved in mammary gland duct morphogenesis	21.6	8.70E-02	1.00E+00	7.50E+01	3.80E-03
negative regulation of developmental process	2.5	8.20E-02	1.00E+00	7.50E+01	3.00E-02
regulation of biological process	1.2	8.20E-02	1.00E+00	7.60E+01	6.80E-02
regulation of secretion by cell	2.8	9.80E-02	1.00E+00	7.60E+01	3.30E-02
cerebral cortex radially oriented cell migration	20.5	9.10E-02	8.40E-01	7.70E+01	4.20E-03
branching involved in mammary gland duct morphogenesis	22.1	8.50E-02	1.00E+00	7.70E+01	3.60E-03
branching involved in mammary gland duct morphogenesis	21.7	8.70E-02	1.00E+00	7.80E+01	3.70E-03
negative regulation of developmental process	2.5	8.70E-02	1.00E+00	7.80E+01	3.30E-02
cerebral cortex radially oriented cell migration	21.1	8.90E-02	1.00E+00	7.80E+01	4.00E-03
regulation of secretion by cell	2.8	9.70E-02	1.00E+00	7.90E+01	3.20E-02
regulation of secretion by cell	2.9	9.00E-02	1.00E+00	7.90E+01	2.90E-02
cerebral cortex radially oriented cell migration	20.7	9.10E-02	1.00E+00	7.90E+01	4.10E-03
regulation of secretion by cell	2.8	9.50E-02	1.00E+00	8.10E+01	3.10E-02

Table 8: Gene ontology analysis of unique structural variant regions of M. mulatta and M. fascicularis conducted with DAVID gene ontology.

M.fascicularis	Term	P-value	Fold Enrichm	Bonferroni	*FDR*	Fisher Exact
	Motor Protein	6.20E-02	29.6	4.90E-01	3.90E+01	2.10E-03
	Myosin	3.70E-02	51.1	5.40E-01	2.60E+01	6.90E-04
	Myosin Complex	3.60E-02	51.5	6.10E-01	3.20E+01	6.70E-04
	SNARE Complex	4.50E-02	40.5	7.00E-01	3.20E+01	1.00E-03
	Motor Activity	3.60E-02	50.9	7.20E-01	2.80E+01	6.80E-04
M. mulatta	Term	P-value	Fold Enrichm	Bonferroni	*FDR*	Fisher Exact
	Positive regulation of apoptotic processe	1.00E-02	8.80	9.10E-01	1.20E+01	1.10E-03
	Apoptotic chromosome condensation	1.10E-02	175.10	9.30E-01	1.30E+01	4.20E-05
	Histone exchange	1.90E-02	105.00	9.90E-01	2.10E+01	1.40E-04
	Intracellular steroid hormone receptor si	2.20E-02	87.60	9.90E-01	2.50E+01	2.10E-04
	Positive regulation of release of cytochrc	7.50E-02	25.00	1.10E+00	6.40E+01	2.80E-03
	Regulation of circadian rhythm	8.90E-02	21.00	1.00E+00	7.00E+01	4.00E-03
	Regulation of apoptotic process	9.90E-02	5.50	1.00E+00	7.40E+01	1.70E-02
	Protein heterodimerization activity	4.20E-02	44.90	1	5.80E+01	8.80E-04

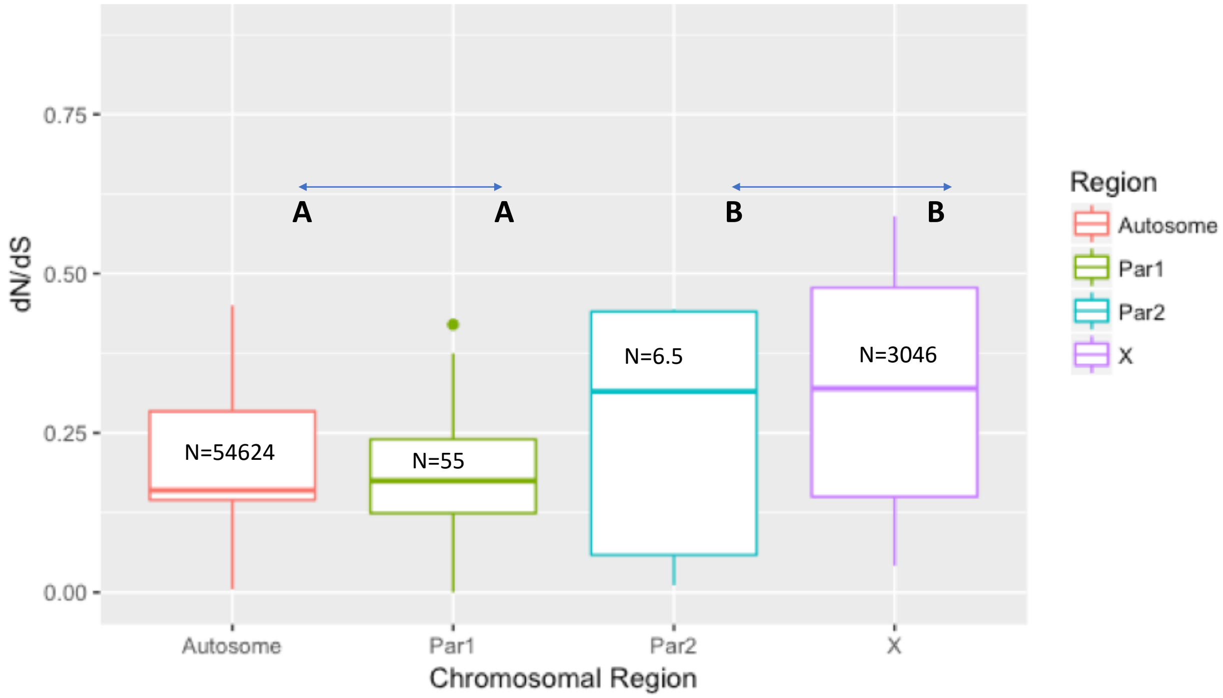


Figure 9:dN/dS between M. mulatta and M. fascicularis, comparisons of Autosomes, PAR1, PAR2 and X. Sample size and statistical grouping can be seen above, with autosome and PAR1 grouping out.

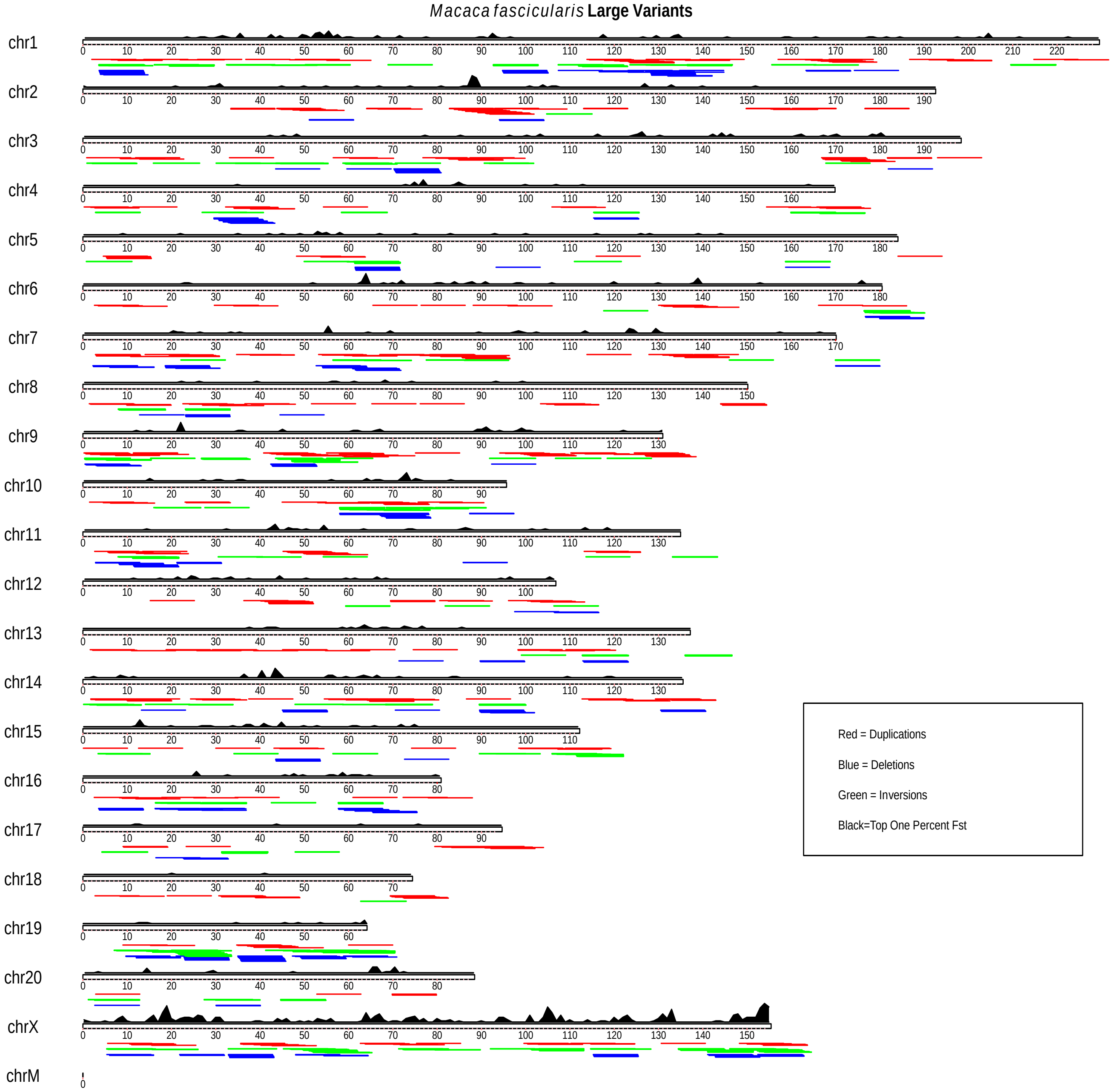


Figure 10: M. fascicularis large structural variants were plotted with KaryoploteR in R , SVs along the genomes in the large bin size(10kb-1mb) can be seen respectively. Additionally, F_{st} outliers were plotted on top of the SV dataset for an overall view of overlap.

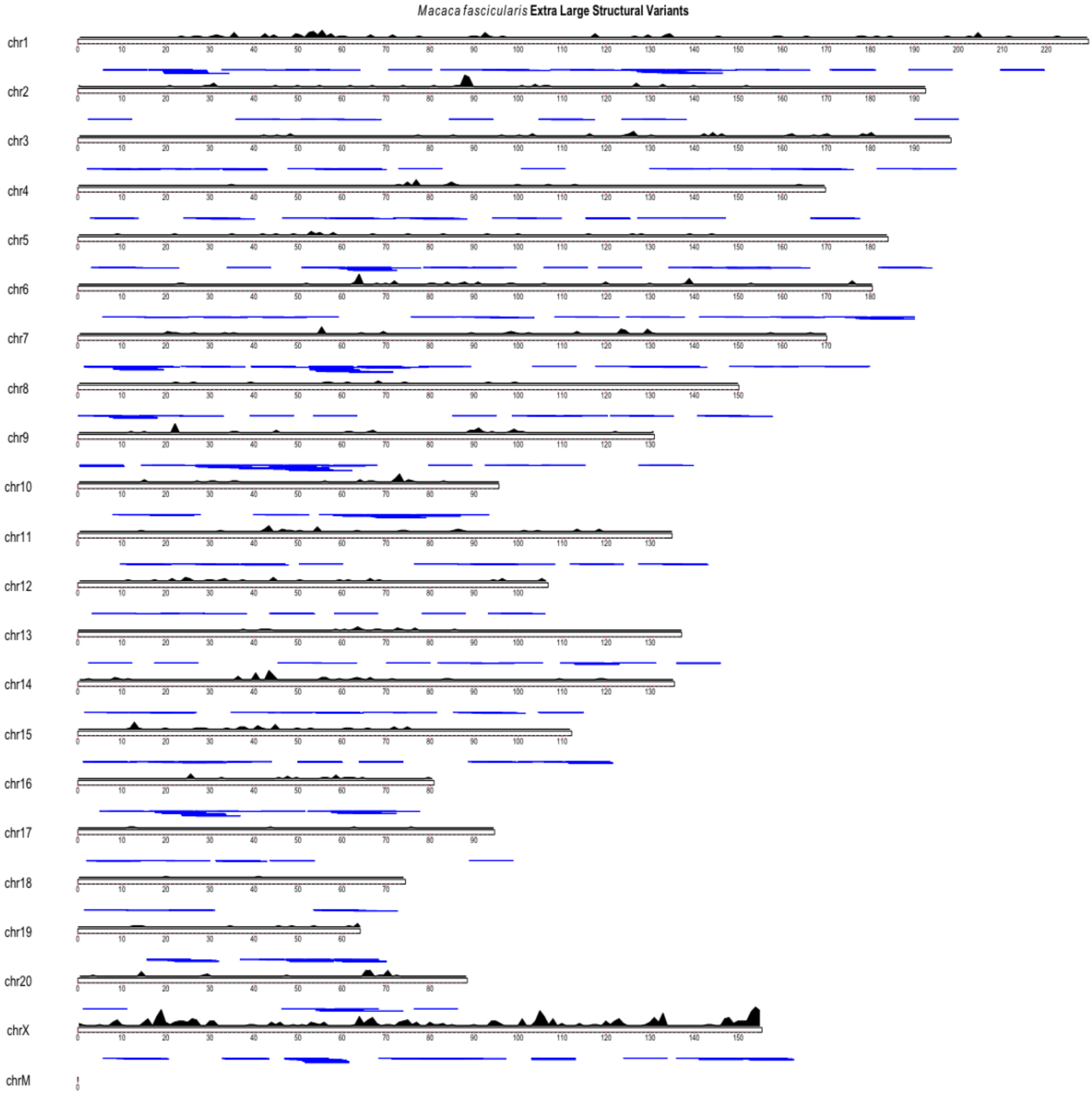


Figure 11: M. fascicularis Extra-large variants filtered for 3/5 match overlaid with areas of high FST

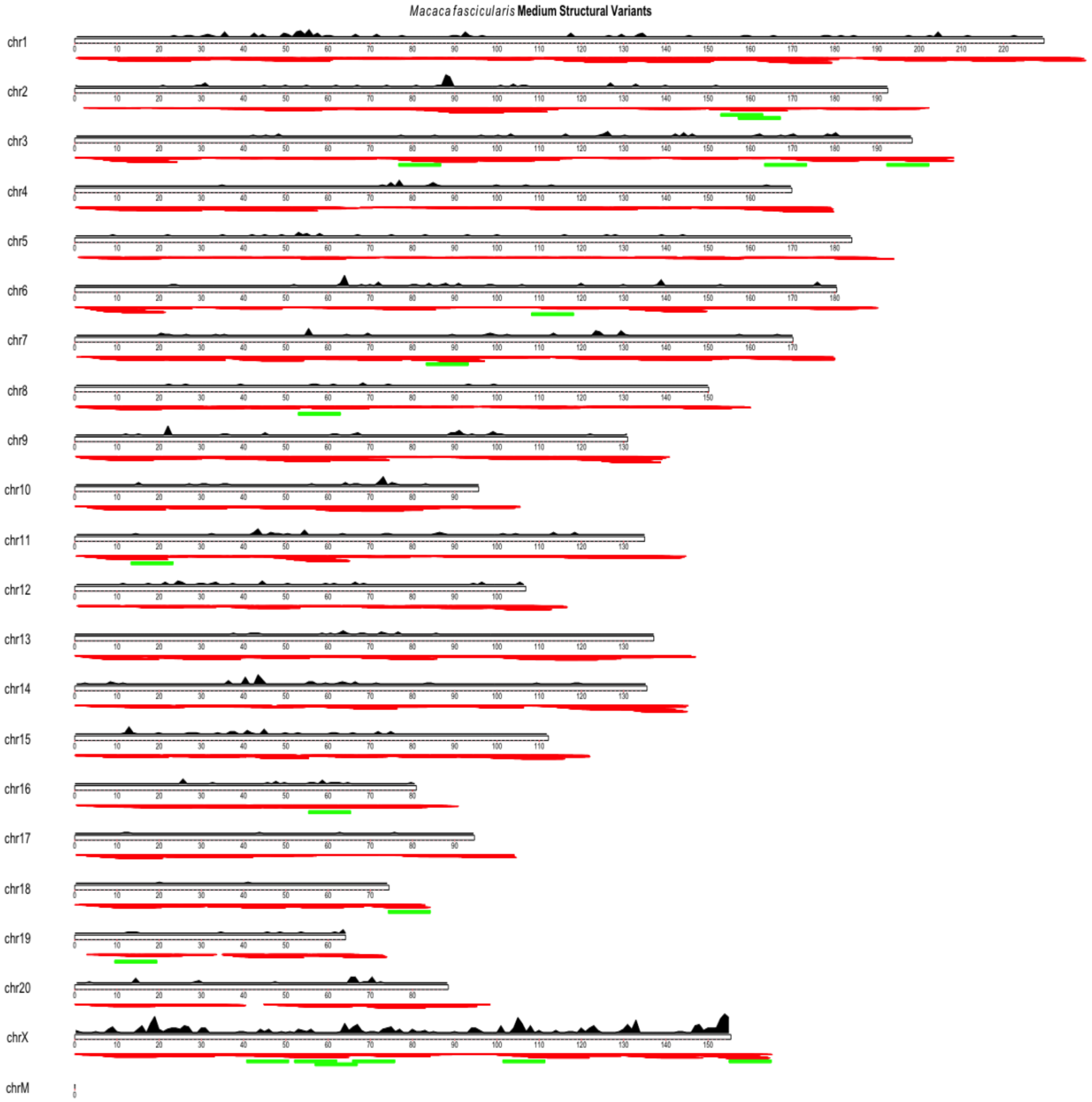


Figure 12: M. fascicularis Medium variants filtered for 3/5 match overlaid with areas of high FST

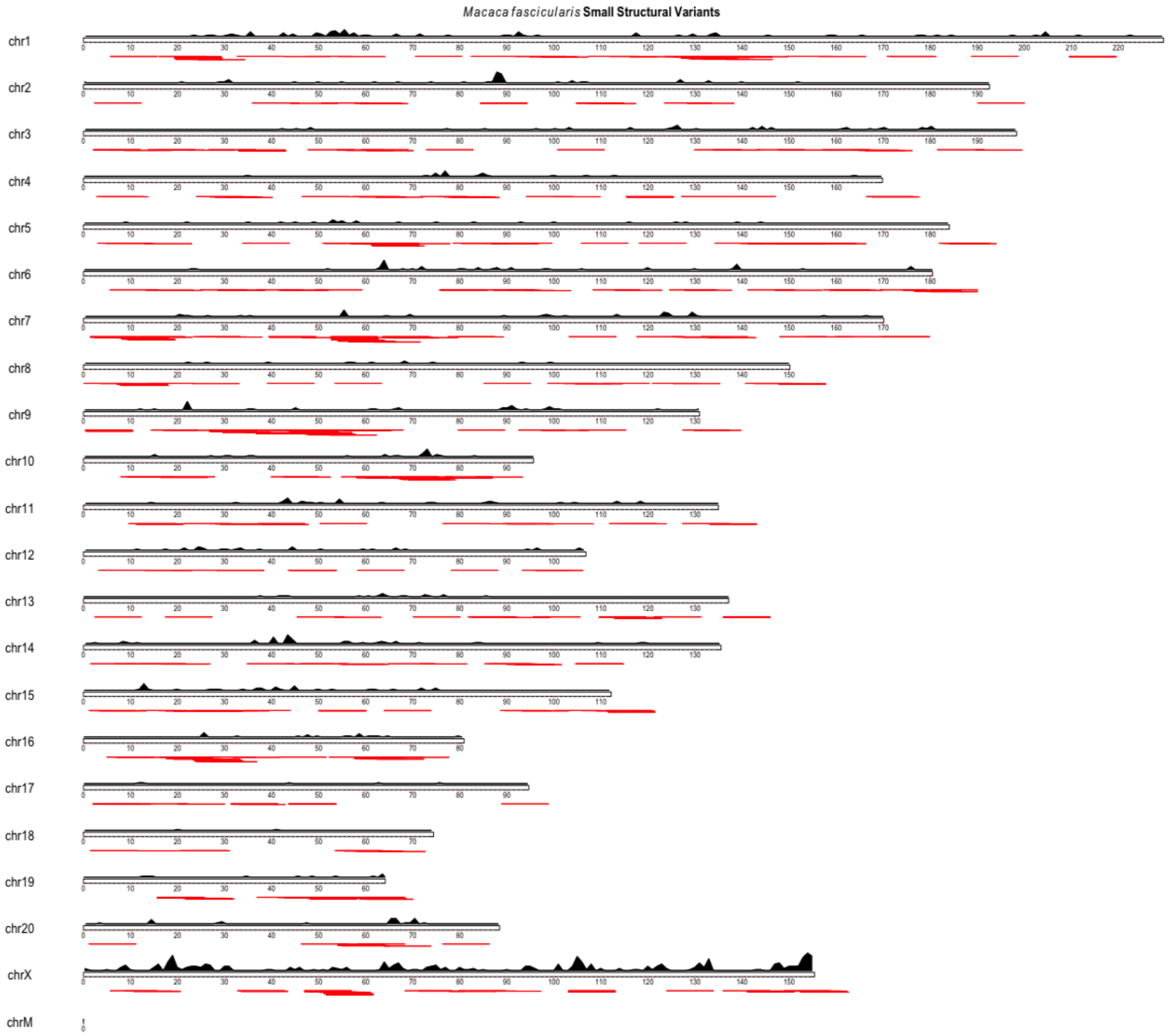


Figure 13: M. fascicularis small variants filtered for 3/5 match overlaid with areas of high FST

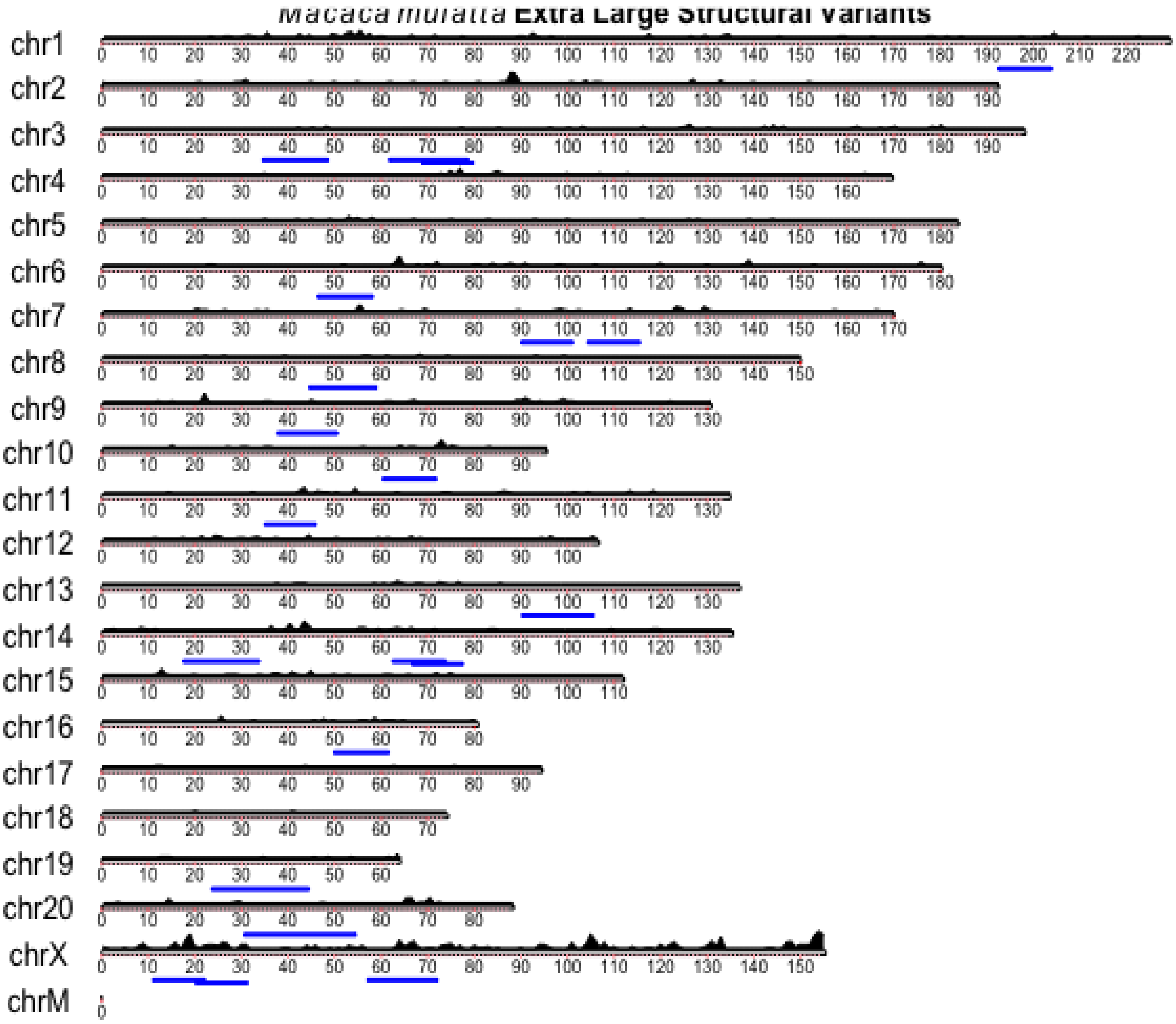


Figure 14:M. mulatta Extra-large variants filtered for 3/5 match overlaid with areas of high FST

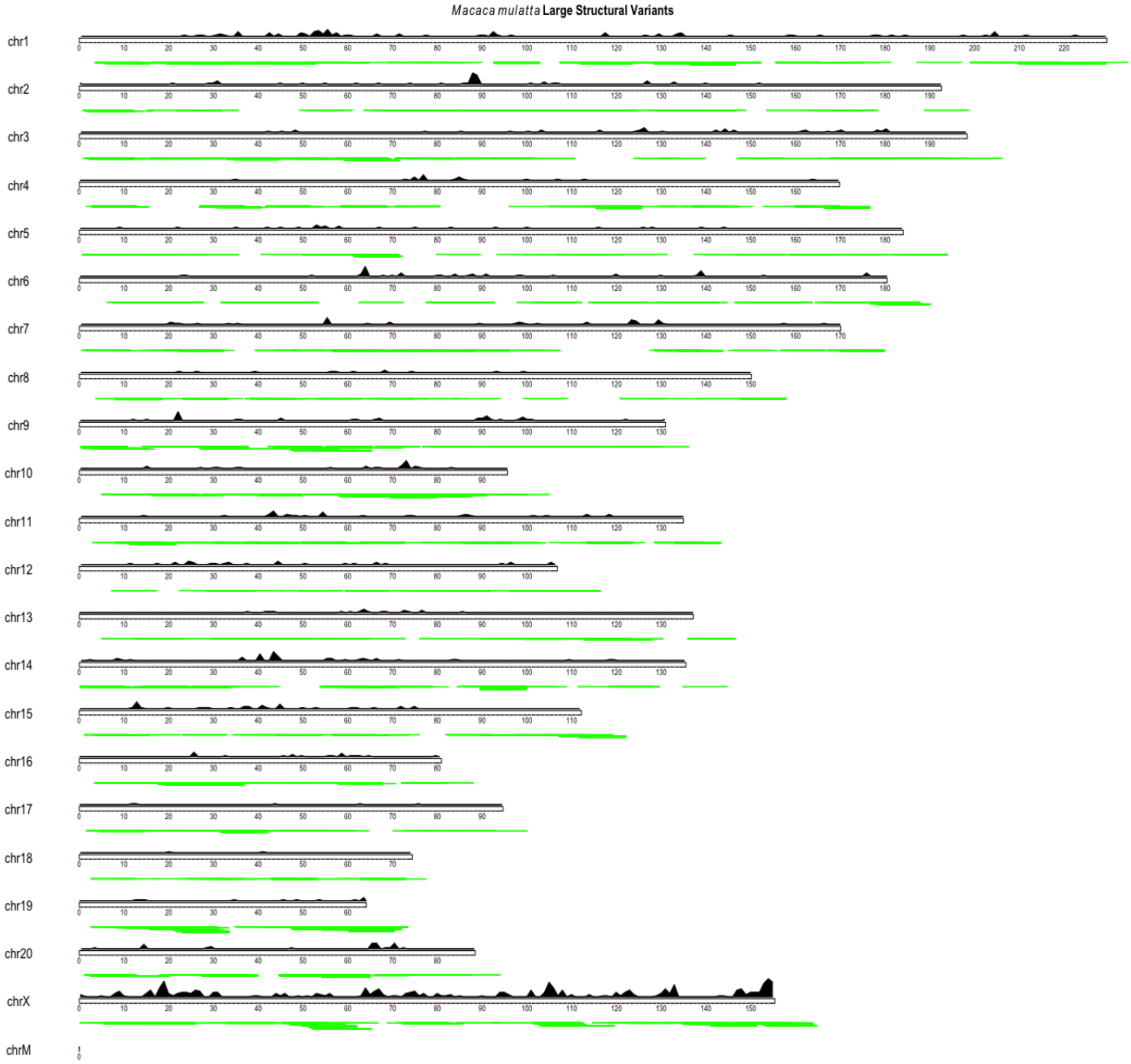


Figure 15: M. mulatta Large variants filtered for 3/5 match overlaid with areas of high FST

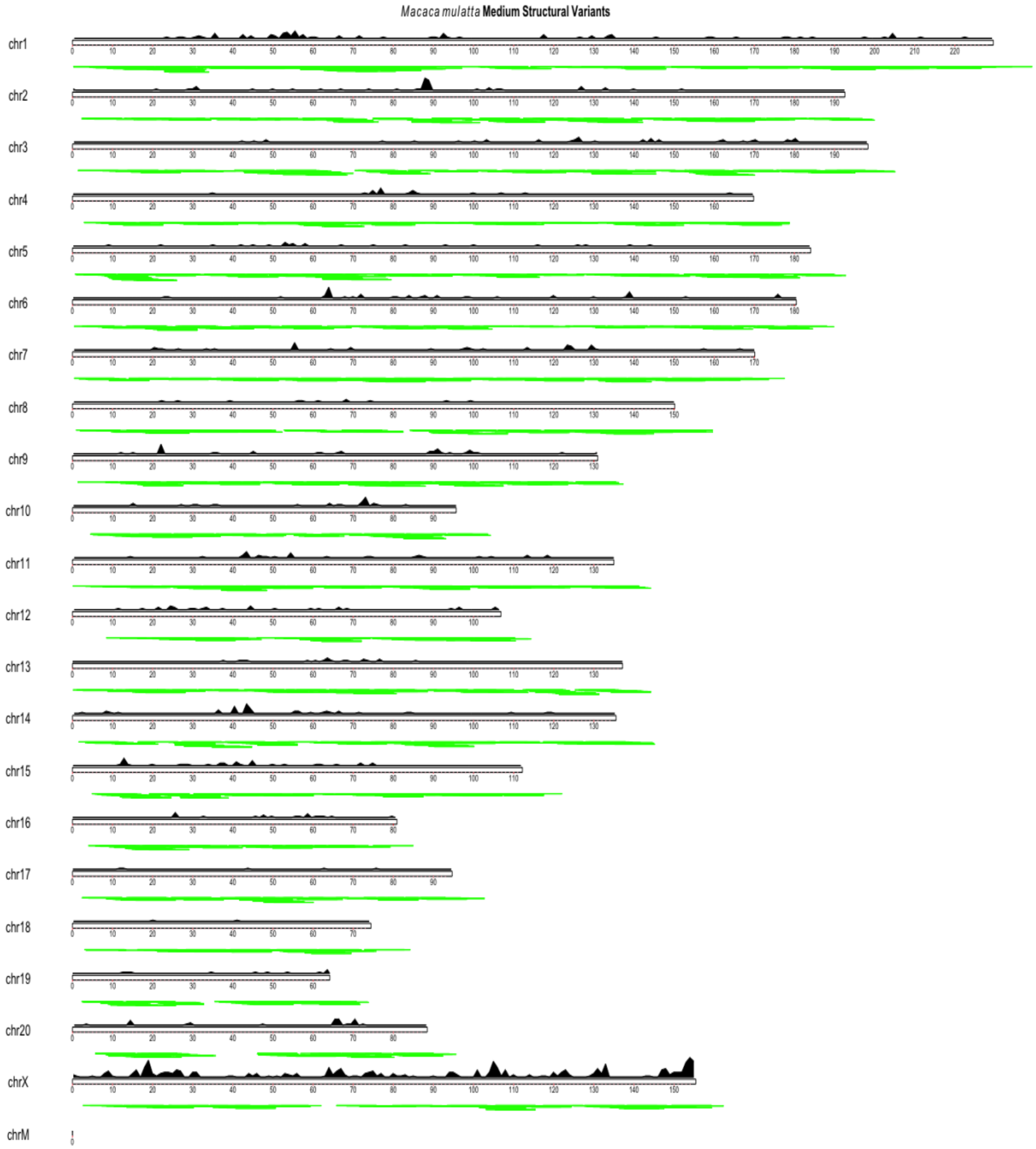


Figure 16: M. mulatta Medium variants filtered for 3/5 match overlaid with areas of high FST

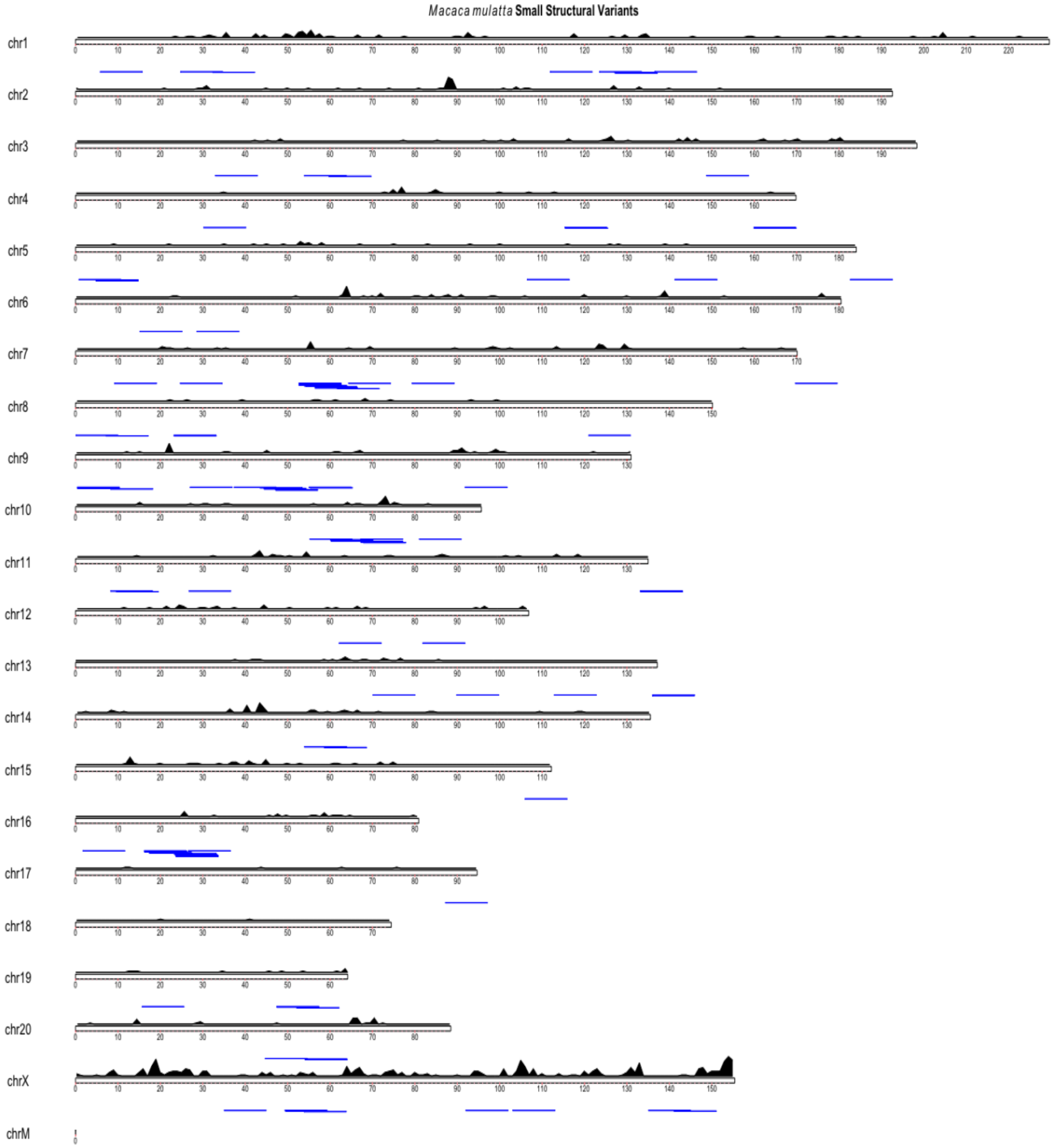


Figure 17:M. mulatta small variants filtered for 3/5 match overlaid with areas of high FST

References

Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, *21*(6), 974–984. https://doi.org/10.1101/gr.114876.110

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*. https://doi.org/10.1093/nar/25.17.3389

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Bisby, F., Bisby, F. A., & Coddington, J. (3AD). Characterization of Biodiversity, (22). Retrieved from https://repository.si.edu/bitstream/handle/10088/16769/ent_Bisby_Al_1995_CharacterizationOfBiodiversity.pdf

Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F., & Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biology*, *13*(12), R122. https://doi.org/10.1186/gb-2012-13-12-r122

Brodin, P., Jojic, V., Gao, T., Bhattacharya, S., Lopez Angel, C. J., Furman, D., … Davis, M. M. (2015). Variation in the Human Immune System Is Largely Driven by Non-Heritable Influences. https://doi.org/10.1016/j.cell.2014.12.020

Castresana, J., Lübben, M., Saraste, M., & Higgins, D. G. (1994). Evolution of cytochrome oxidase, an enzyme older than atmospheric oxygen. *The EMBO Journal*, *13*(11), 2516–2525. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8013452

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., … Mardis, E. R. (2009). BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, *6*(9), 677–681. https://doi.org/10.1038/nmeth.1363

Chong, Z., Ruan, J., Gao, M., Zhou, W., Chen, T., Fan, X., … Chen, K. (2017). novoBreak: local assembly for breakpoint detection in cancer genomes. *Nature Methods*, *14*(1), 65–67. https://doi.org/10.1038/nmeth.4084

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … 1000 Genomes Project Analysis Group, 1000 Genomes Project Analysis. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Elango, N., Lee, J., Peng, Z., Loh, Y.-H. E., & Yi, S. V. (n.d.). Evolutionary rate variation in Old World monkeys. *Biol. Lett.* https://doi.org/10.1098/rsbl.2008.0712

Escaramís, G., Docampo, E., & Rabionet, R. (2015). A decade of structural variants: description, history and methods to detect structural variation. *Briefings in Functional Genomics*, *14*(5), 305–314. https://doi.org/10.1093/bfgp/elv014

Feulner, P. G. D., & De-Kayne, R. (2017). Genome evolution, structural rearrangements and speciation. *Journal of Evolutionary Biology*, *30*(8), 1488–1490. https://doi.org/10.1111/jeb.13101

Fooden, J. (1983). *The Macaques: Studies in Ecology, Behavior, and Evolution*. *NAT. HIST. BULL. SIAM Soc* (Vol. 31). Retrieved from http://www.siamese-heritage.org/nhbsspdf/vol031-040/NHBSS_031_1k_Ali_TheMacaquesStudiesInE.pdf

Fooden, J., & Fooden, J. (2000). *Systematic review of the rhesus macaque, Macaca mulatta (Zimmermann, 1780) /*. Chicago, Ill. : Field Museum of Natural History,. Retrieved from https://www.biodiversitylibrary.org/item/30778

Frankham, R. (2015). Genetic rescue of small inbred populations: meta-analysis reveals large and consistent benefits of gene flow. *Molecular Ecology*, *24*(11), 2610–2618. https://doi.org/10.1111/mec.13139

Garrigan, D., Kingan, S. B., Geneva, A. J., Vedanayagam, J. P., & Presgraves, D. C. (2014). Genome Diversity and Divergence in Drosophila mauritiana : Multiple Signatures of Faster X Evolution. *Genome Biology and Evolution*, *6*(9), 2444–2458. https://doi.org/10.1093/gbe/evu198

Ge, X., Kwok, P.-Y., & Shieh, J. T. C. (2015). Prioritizing genes for X-linked diseases using population exome data. *Human Molecular Genetics*, *24*(3), 599–608. https://doi.org/10.1093/hmg/ddu473

Gel, B., & Serra, E. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics*, *33*(19), 3088–3090. https://doi.org/10.1093/bioinformatics/btx346

Harris, R. S. (2007). IMPROVED PAIRWISE ALIGNMENT OF GENOMIC DNA. Retrieved from http://www.bx.psu.edu/~rsharris/rsharris_phd_thesis_2007.pdf

Hatem, A., Bozdağ, D., Toland, A. E., & Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, *14*(1), 184. https://doi.org/10.1186/1471-2105-14-184

Higashino, A., Sakate, R., Kameoka, Y., Takahashi, I., Hirata, M., Tanuma, R., … Osada, N. (2012). Whole-genome sequencing and analysis of the Malaysian cynomolgus macaque (Macaca fascicularis) genome. *Genome Biology*, *13*(7), R58. https://doi.org/10.1186/gb-2012-13-7-r58

Holland, A. J., & Cleveland, D. W. (2012). Chromoanagenesis and cancer: Mechanisms and consequences of localized, complex chromosomal rearrangements. *Nature Medicine*. https://doi.org/10.1038/nm.2988

Hosack, D. A., Dennis, G., Sherman, B. T., Lane, H., & Lempicki, R. A. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biology*, *4*(10), R70. https://doi.org/10.1186/gb-2003-4-10-r70

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. https://doi.org/10.1038/nprot.2008.211

Kirkpatrick, M., & Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*, *173*(1), 419–434. https://doi.org/10.1534/genetics.105.047985

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, *33*(7), 1870–1874. https://doi.org/10.1093/molbev/msw054

Lam, H. Y. K., Mu, X. J., Stütz, A. M., Tanzer, A., Cayting, P. D., Snyder, M., … Gerstein, M. B. (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature Biotechnology*, *28*(1), 47–55. https://doi.org/10.1038/nbt.1600

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6). https://doi.org/10.1186/gb-2014-15-6-r84

Leinonen, R., Sugawara, H., Shumway, M., & International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(Database issue), D19-21. https://doi.org/10.1093/nar/gkq1019

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbq015

Li, W.-H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution*, *36*(1), 96–99. https://doi.org/10.1007/BF02407308

Li, W. H., Wu, C. I., & Luo, C. C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution*, *2*(2), 150–174. https://doi.org/10.1093/oxfordjournals.molbev.a040343

Lindberg, M. R., Hall, I. M., & Quinlan, A. R. (2015). Population-based structural variation discovery with Hydra-Multi. *Bioinformatics*, *31*(8), 1286–1289. https://doi.org/10.1093/bioinformatics/btu771

Liu, B., Conroy, J. M., Morrison, C. D., Odunsi, A. O., Qin, M., Wei, L., … Wang, J. (2015). Structural variation discovery in the cancer genome using next generation sequencing: Computational solutions and perspectives. *Oncotarget*, *6*(8), 5477–5489. https://doi.org/10.18632/oncotarget.3491

Mardis, E. R. (2008). Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, *9*(1), 387–402. https://doi.org/10.1146/annurev.genom.9.081307.164359

Mashl, R. J., Scott, A. D., Huang, K.-L., Wyczalkowski, M. A., Yoon, C. J., Niu, B., … Ding, L. (2017). GenomeVIP: a cloud platform for genomic variant discovery and interpretation. *Genome Research*, *27*(8), 1450–1459. https://doi.org/10.1101/gr.211656.116

Mayr, E. (1982). The Growth of Biological Thought: Diversity, Evolution, and Inheritance. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*, 616–617. https://doi.org/10.1016/0162-3095(84)90038-4

McGaugh, S. E., & Noor, M. A. F. (2012). Genomic impacts of chromosomal inversions in parapatric Drosophila species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1587), 422–429. https://doi.org/10.1098/rstb.2011.0250

Metzker, M. L. (2010). Sequencing technologies the next generation. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2626

Nei, M. (1973). Analysis of Gene Diversity in Subdivided Populations. *Proc. Nat. Acad. Sci. USA*, *70*(12), 3321–3323. https://doi.org/10.1073/pnas.70.12.3321

Nei, M., & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, *3*(5), 418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(10), 5269–5273. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/291943

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2015). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, *32*(2), btv566. https://doi.org/10.1093/bioinformatics/btv566

Pamilo, P., & Bianchi, N. O. (1993). Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Molecular Biology and Evolution*, *10*(2), 271–281. https://doi.org/10.1093/oxfordjournals.molbev.a040003

Prager, E. M., & Wilson, A. C. (1975). Slow Evolutionary Loss of the Potential for Interspecific Hybridization in Birds: A Manifestation of Slow Regulatory Evolution (anatomical evolution/protein evolution/chromosomal evolution/frogs/mammals/immunology), *72*(1), 200–204. Retrieved from http://europepmc.org/backend/ptpmcrender.fcgi?accid=PMC432270&blobtype=pdf

Presgraves, D. C. (2008). Sex chromosomes and speciation in Drosophila. *Trends in Genetics*, *24*(7), 336–343. https://doi.org/10.1016/j.tig.2008.04.007

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18). https://doi.org/10.1093/bioinformatics/bts378

Ruffalo, M., Laframboise, T., & Koyutürk, M. (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, *27*(20), 2790–2796. https://doi.org/10.1093/bioinformatics/btr477

Scally, A. (2016). Mutation rates and the evolution of germline structure. *BioRxiv*, 034298. https://doi.org/10.1101/034298

Ségurel, L., Wyman, M. J., & Przeworski, M. (2014). Determinants of Mutation Rate Variation in the Human Germline. *Annual Review of Genomics and Human Genetics*, *15*(1), 47–70. https://doi.org/10.1146/annurev-genom-031714-125740

Shang, J., Zhu, F., Vongsangnak, W., Tang, Y., Zhang, W., & Shen, B. (2014). Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. *BioMed Research International*, *2014*, 309650. https://doi.org/10.1155/2014/309650

Singh, M., & Sinha, A. (2004). *Macaque Societies: A Model for the Study of Social Organization*. Cambridge: Cambridge University Press. Retrieved from http://nias.res.in/publication/life-history-traits-ecological-adaptations-or-phylogenetic-relics

Smith, D. G., & Mcdonough, J. (2005). Mitochondrial DNA variation in Chinese and Indian Rhesus macaques (Macaca mulatta). *American Journal of Primatology*, *65*(1), 1–25. https://doi.org/10.1002/ajp.20094

Stanyon, R., Fantini, C., Camperio-Ciani, A., Chiarelli, B., & Ardito, G. (1988). Banded karyotypes of 20 Papionini species reveal no necessary correlation with speciation. *American Journal of Primatology*, *16*(1), 3–17. https://doi.org/10.1002/ajp.1350160103

Stevison, L. S., Hoehn, K. B., & Noor, M. A. F. (2011). Effects of inversions on within- and between-species recombination and divergence. *Genome Biology and Evolution*, *3*(1), 830–841. https://doi.org/10.1093/gbe/evr081

Stevison, L. S., & Kohn, M. H. (2008). Determining genetic background in captive stocks of cynomolgus macaques (*Macaca fascicularis*). *Journal of Medical Primatology*, *37*(6), 311–317. https://doi.org/10.1111/j.1600-0684.2008.00292.x

Stevison, L. S., & Kohn, M. H. (2009). Divergence population genetic analysis of hybridization between rhesus and cynomolgus macaques. *Molecular Ecology*, *18*(11), 2457–2475. https://doi.org/10.1111/j.1365-294X.2009.04212.x

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595. https://doi.org/PMC1203831

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen,

S. G. (2012). Primer3--new capabilities and interfaces. *Nucleic Acids Research*, *40*(15), e115. https://doi.org/10.1093/nar/gks596

Wang, J., Mullighan, C. G., Easton, J., Roberts, S., Heatley, S. L., Ma, J., … Zhang, J. (2011). CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nature Methods*, *8*(8), 652–654. https://doi.org/10.1038/nmeth.1628

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, *38*(6), 1358. https://doi.org/10.2307/2408641

Wickham, H. (2011). ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(2), 180–185. https://doi.org/10.1002/wics.147

Wijaya, E., Shimizu, K., Asai, K., & Hamada, M. (2014). Reference-free prediction of rearrangement breakpoint reads. *Bioinformatics*, *30*(18), 2559–2567. https://doi.org/10.1093/bioinformatics/btu360

Wilson, A. C., Maxson, L. R., & Sarich, V. M. (1974). Two types of molecular evolution. Evidence from studies of interspecific hybridization. *Proceedings of the National Academy of Sciences of the United States of America*, *71*(7), 2843–2847. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4212492

Wong, K., Keane, T. M., Stalker, J., & Adams, D. J. (2010). Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biology*, *11*(12), R128. https://doi.org/10.1186/gb-2010-11-12-r128

Yan, G., Zhang, G., Fang, X., Zhang, Y., Li, C., Ling, F., … Wang, J. (2011). Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nature Biotechnology*, *29*(11), 1019–1023. https://doi.org/10.1038/nbt.1992

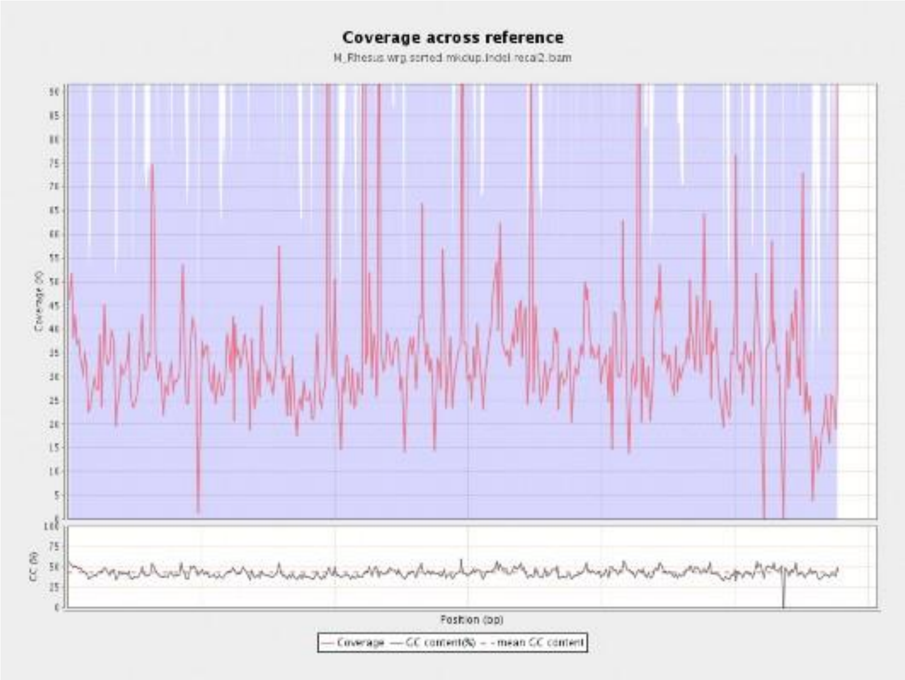
Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, *25*(21), 2865–2871. https://doi.org/10.1093/bioinformatics/btp394

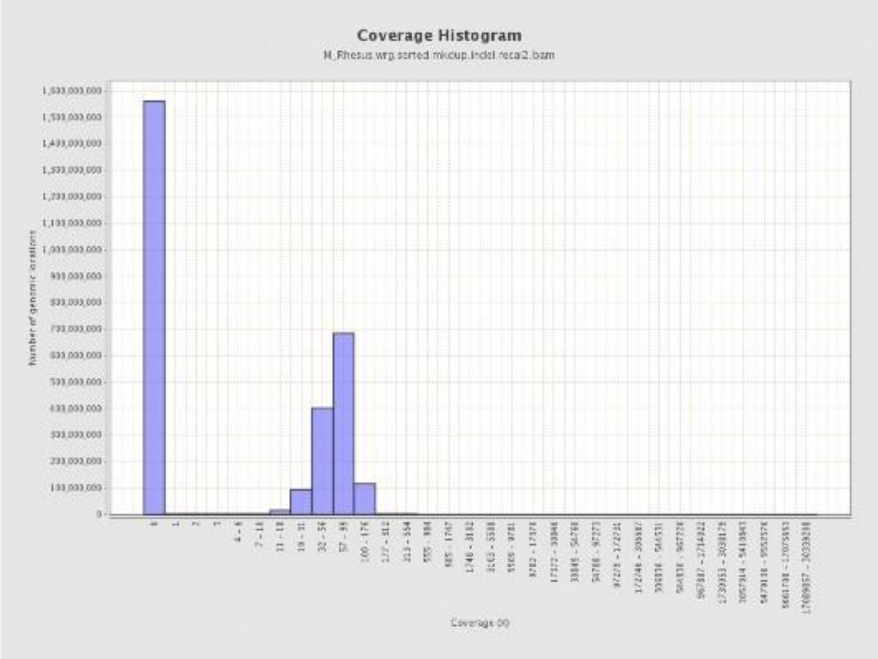
Yu, X., Guda, K., Willis, J., Veigl, M., Wang, Z., Markowitz, S., … Sun, S. (2012). How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? *BioData Mining*, *5*(1), 6. https://doi.org/10.1186/1756-0381-5-6

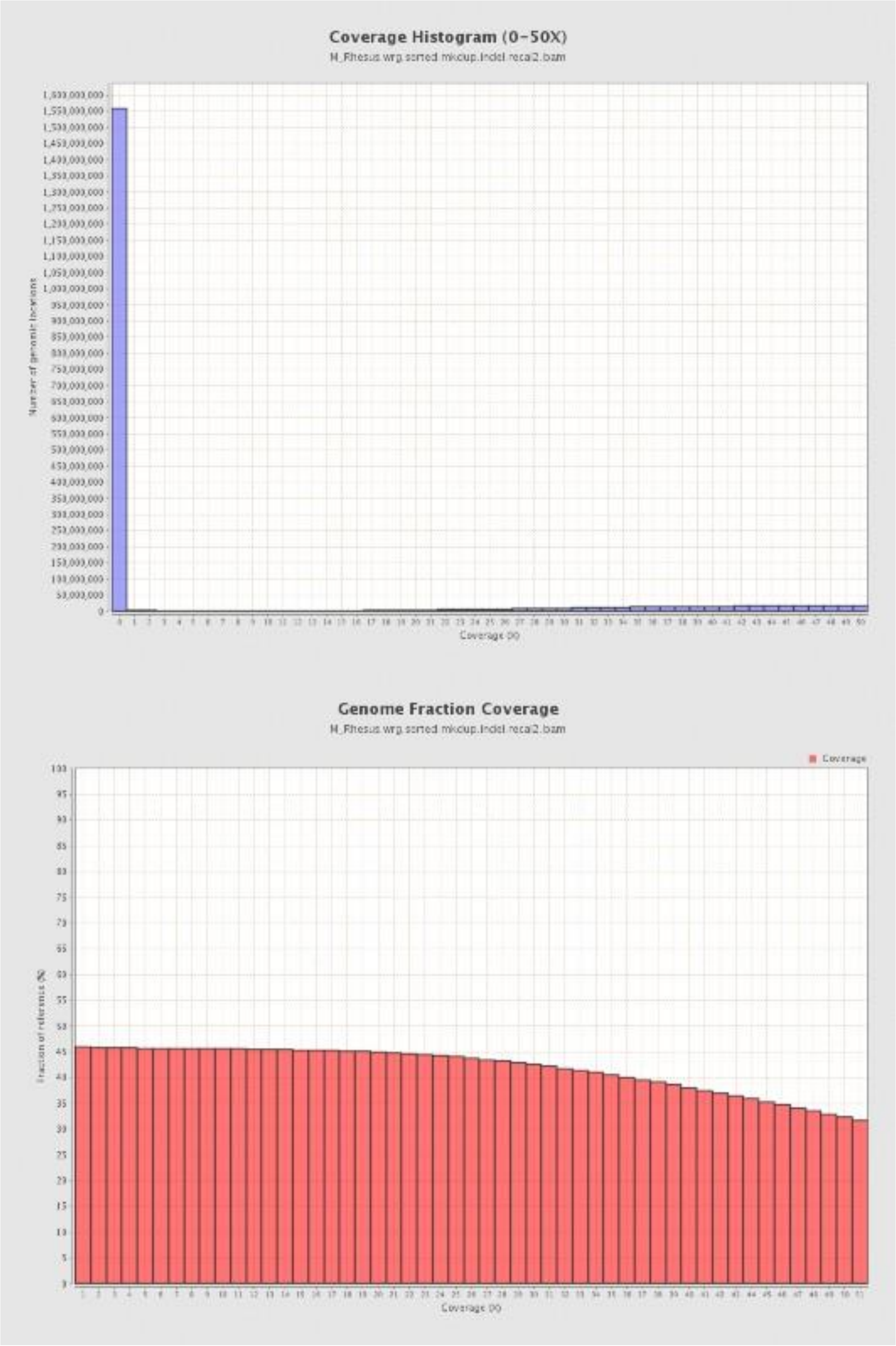
Ziemann, M. (2016). Accuracy, speed and error tolerance of short DNA sequence aligners. *BioRxiv*, 053686. https://doi.org/10.1101/053686

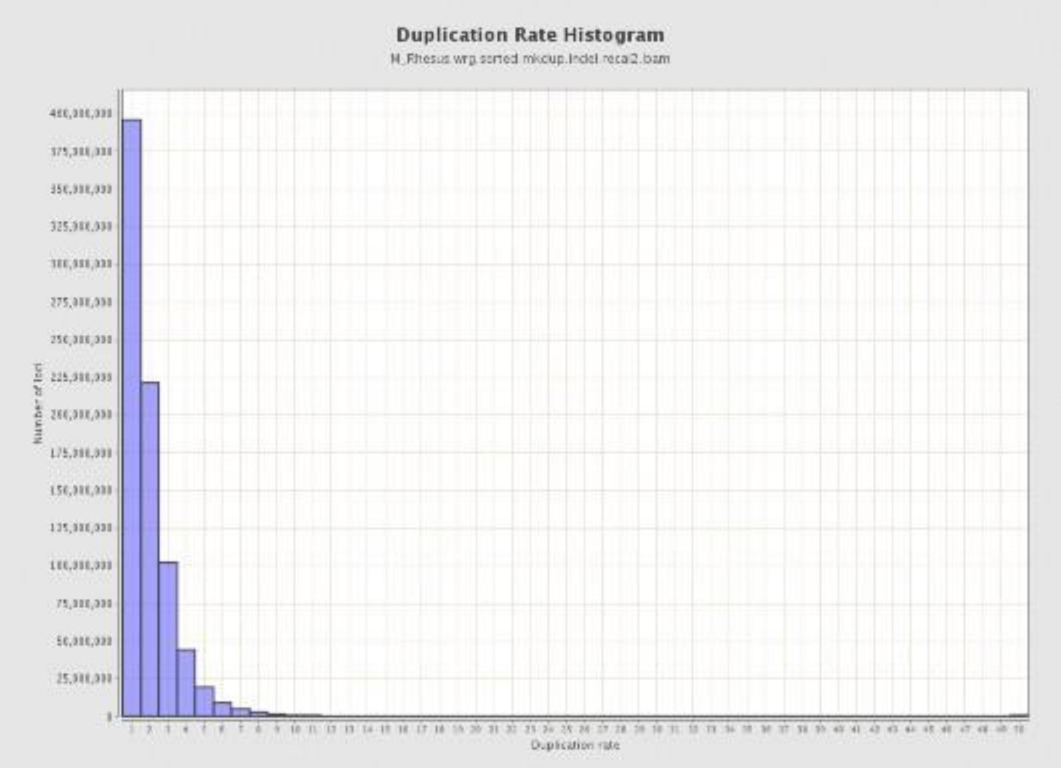
Appendix

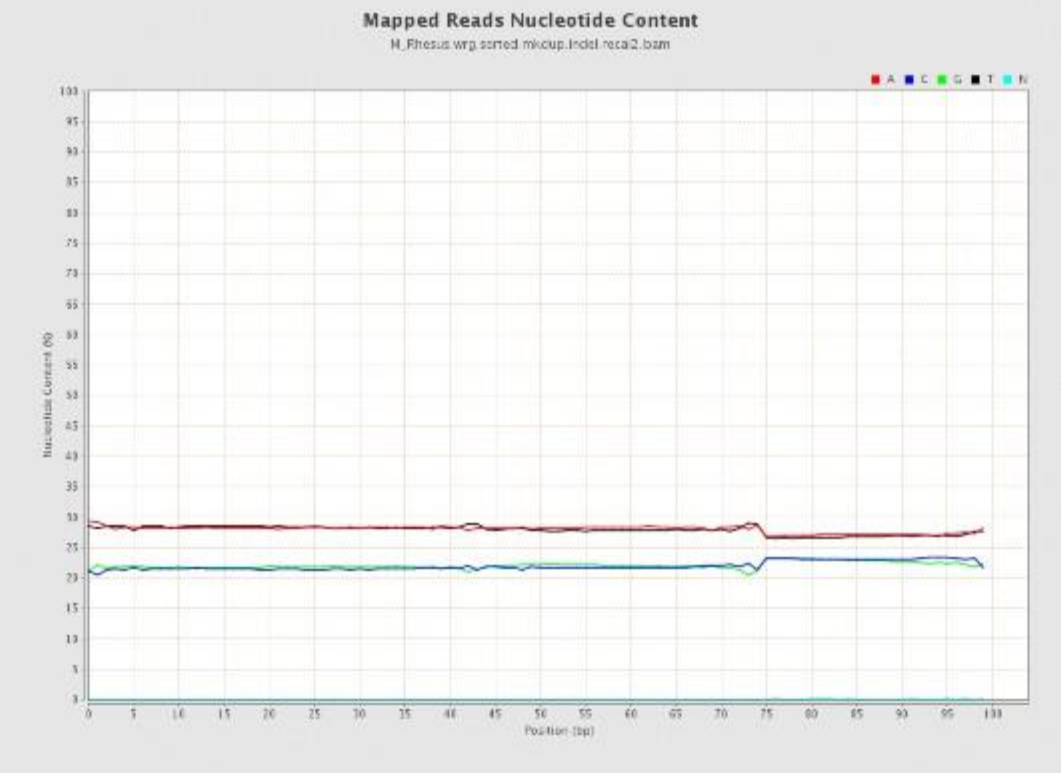
M. mulatta BAMQC

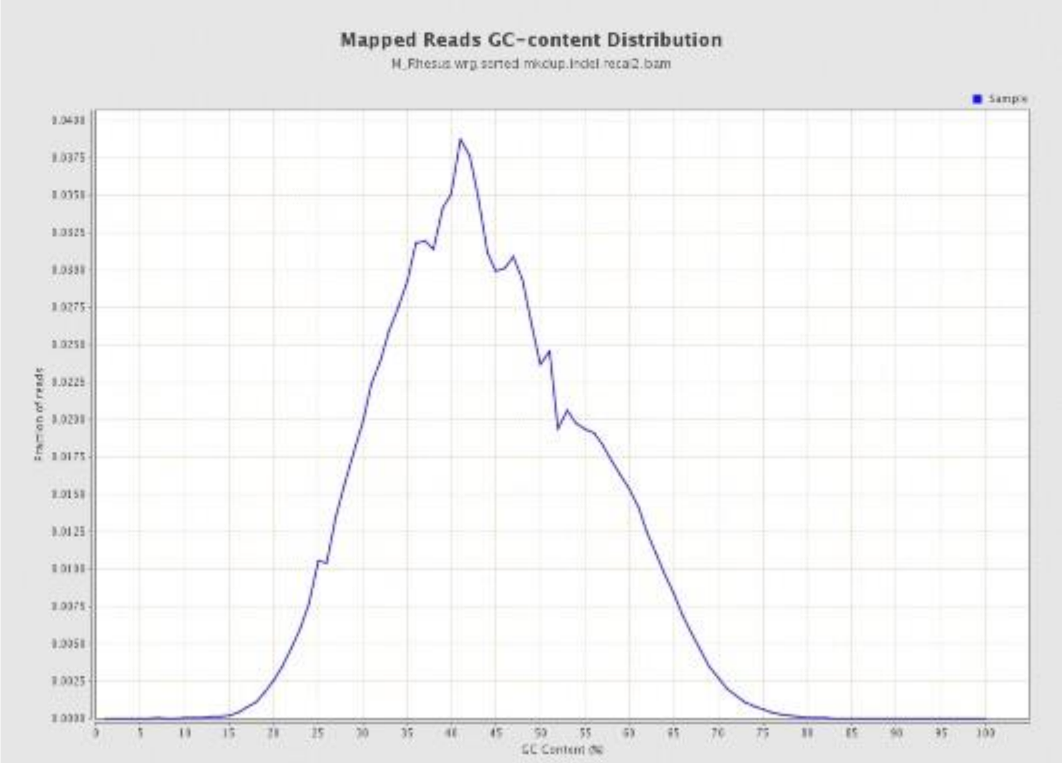


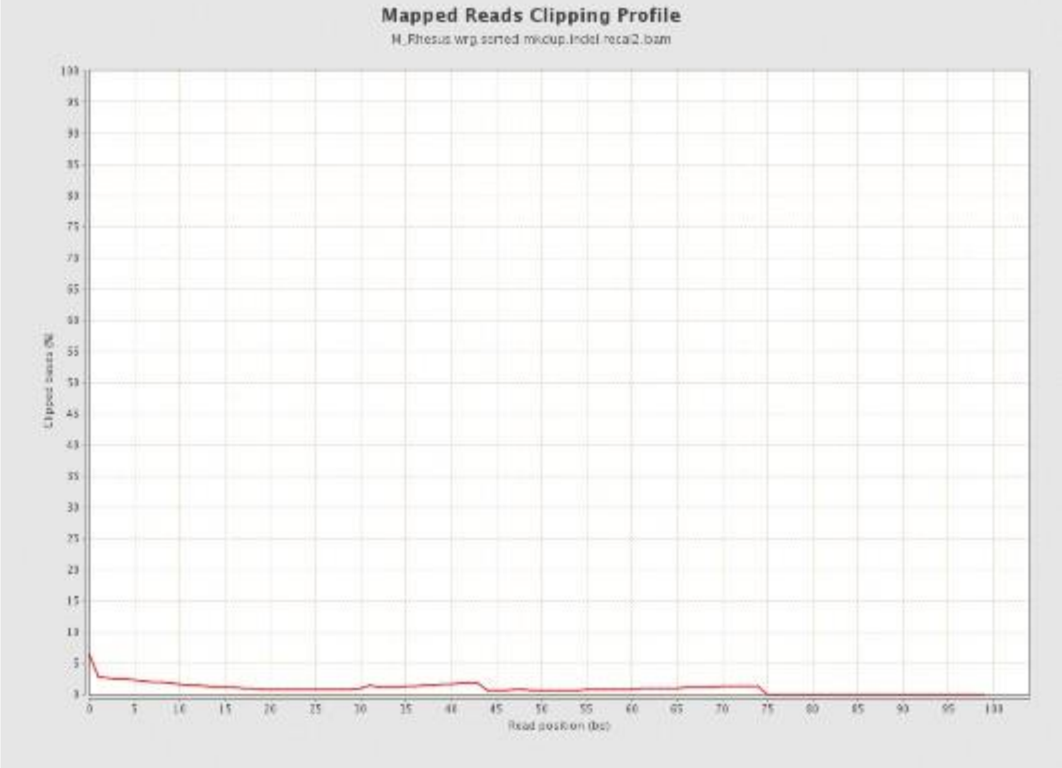


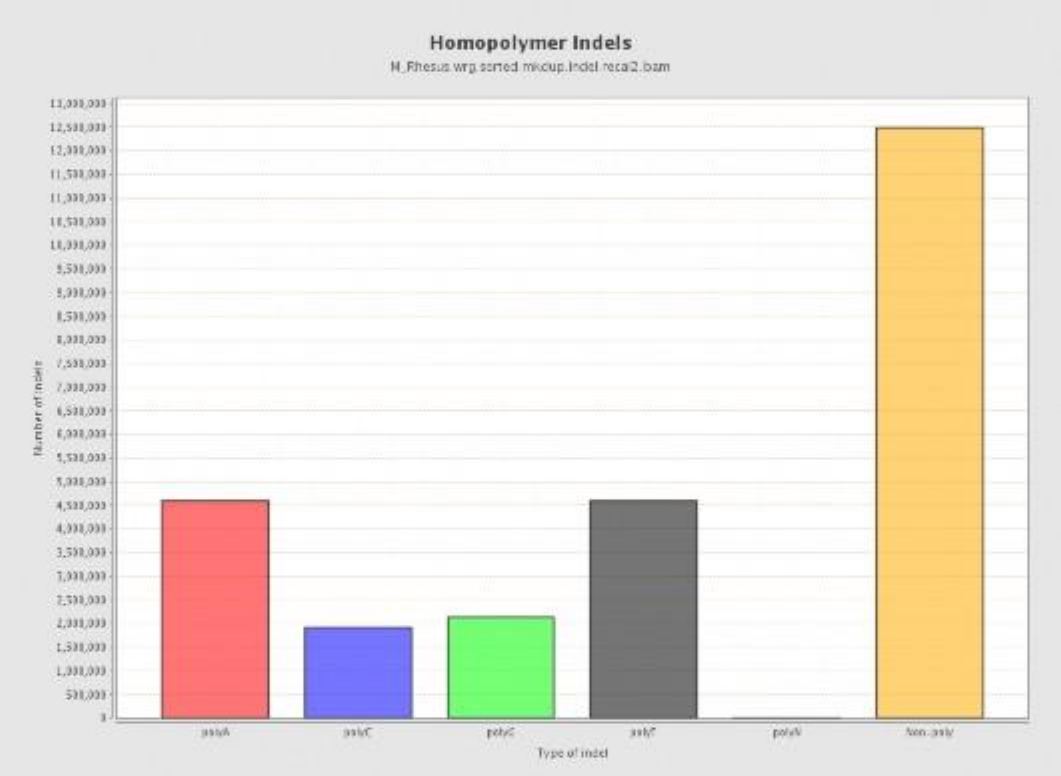

Coverage Histogram (0–50X)
M_Rhesus.wrg.sorted.mkdup.indel.recal2.bam
Number of genomic locations
1,600,000,000
1,550,000,000
1,500,000,000
1,450,000,000
1,400,000,000
1,350,000,000
1,300,000,000
1,250,000,000
1,200,000,000
1,150,000,000
1,100,000,000
1,050,000,000
1,000,000,000
950,000,000
900,000,000
850,000,000
800,000,000
750,000,000
700,000,000
650,000,000
600,000,000
550,000,000
500,000,000
450,000,000
400,000,000
350,000,000
300,000,000
250,000,000
200,000,000
150,000,000
100,000,000
50,000,000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Coverage (X)
Genome Fraction Coverage
M_Rhesus.wrg.sorted.mkdup.indel.recal2.bam
Coverage
Fraction of reference (%)
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
Coverage (X)

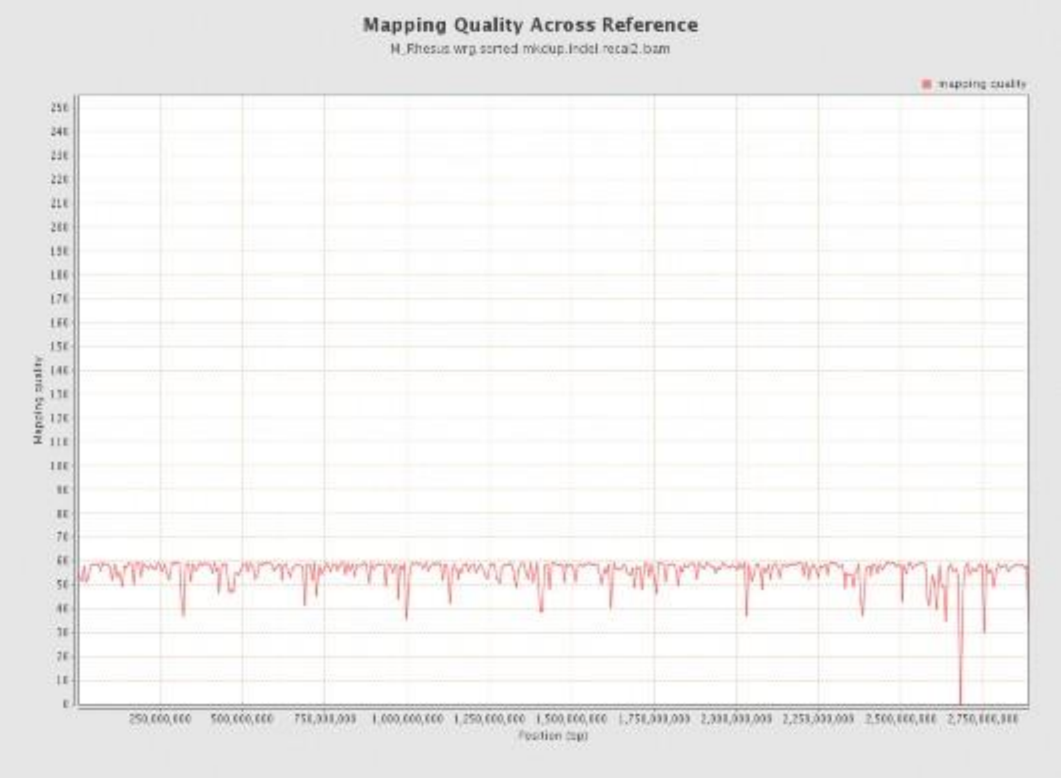

Duplication Rate Histogram
M_Rhesus.wrg.sorted.mkdup.indel.recal2.bam
Number of loci
400,000,000
375,000,000
350,000,000
325,000,000
300,000,000
275,000,000
250,000,000
225,000,000
200,000,000
175,000,000
150,000,000
125,000,000
100,000,000
75,000,000
50,000,000
25,000,000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Duplication rate

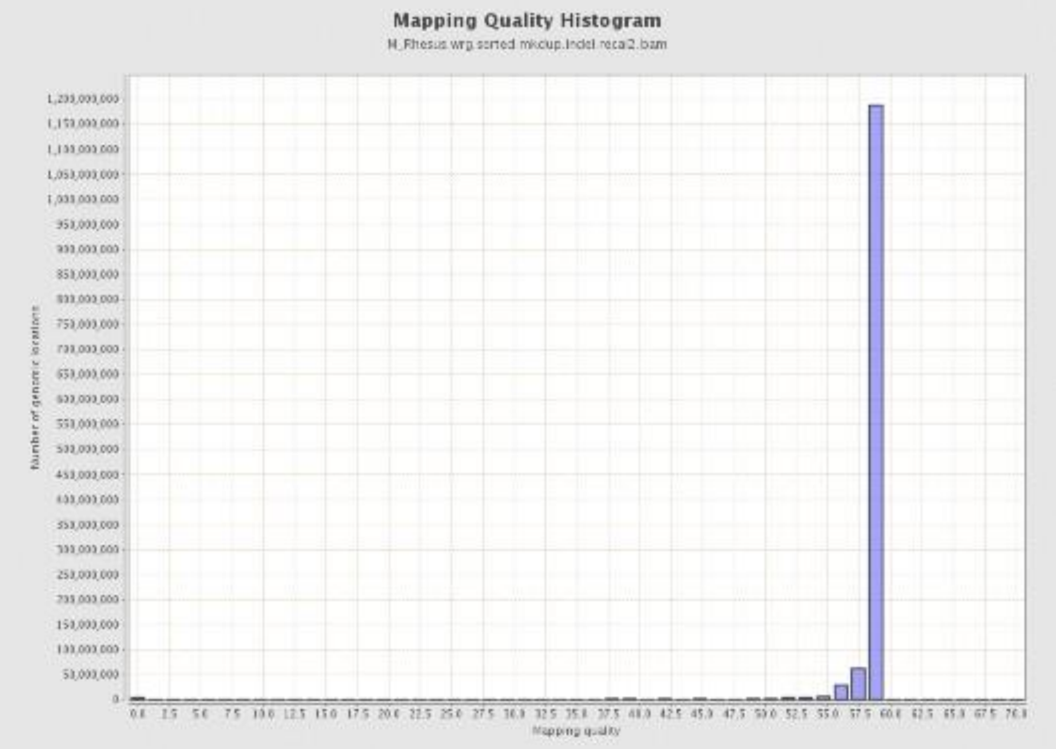

Mapped Reads Nucleotide Content
M_Rhesus.wrg.sorted.mkdup.indel.recal2.bam
A C G T N
Nucleotide Content (%)
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Position (bp)

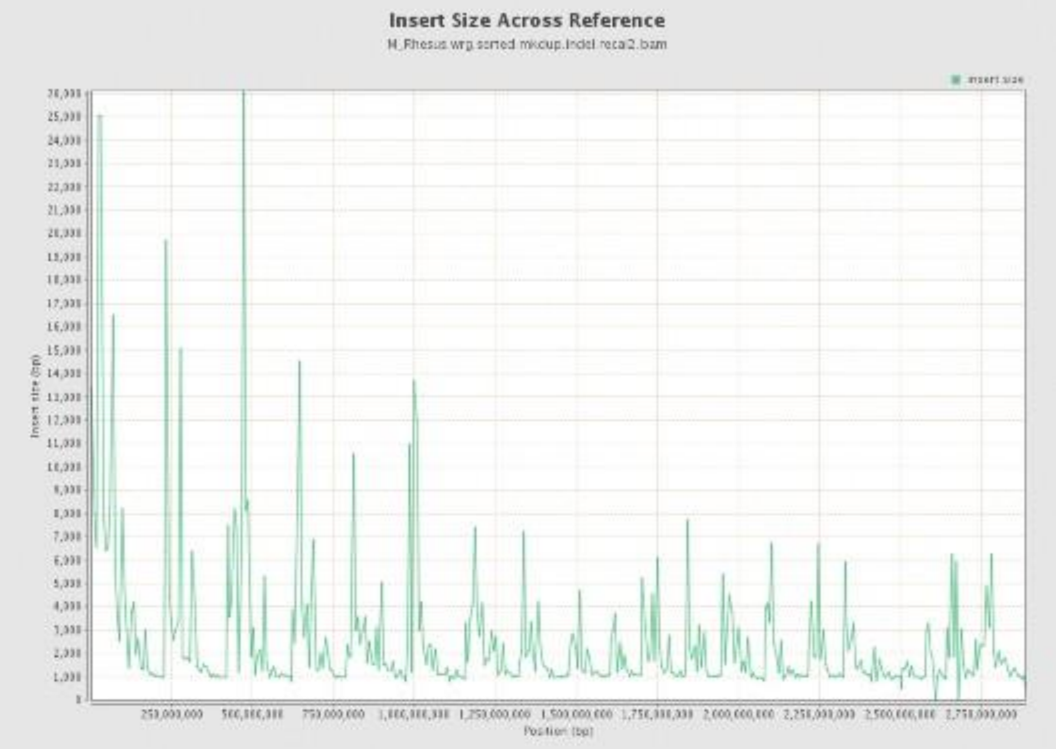

Mapped Reads GC-content Distribution
M_Rhesus.wrg.sorted.mkdup.indel.recal2.bam
Sample
Fraction of reads
0.0400
0.0375
0.0350
0.0325
0.0300
0.0275
0.0250
0.0225
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
GC Content (%)

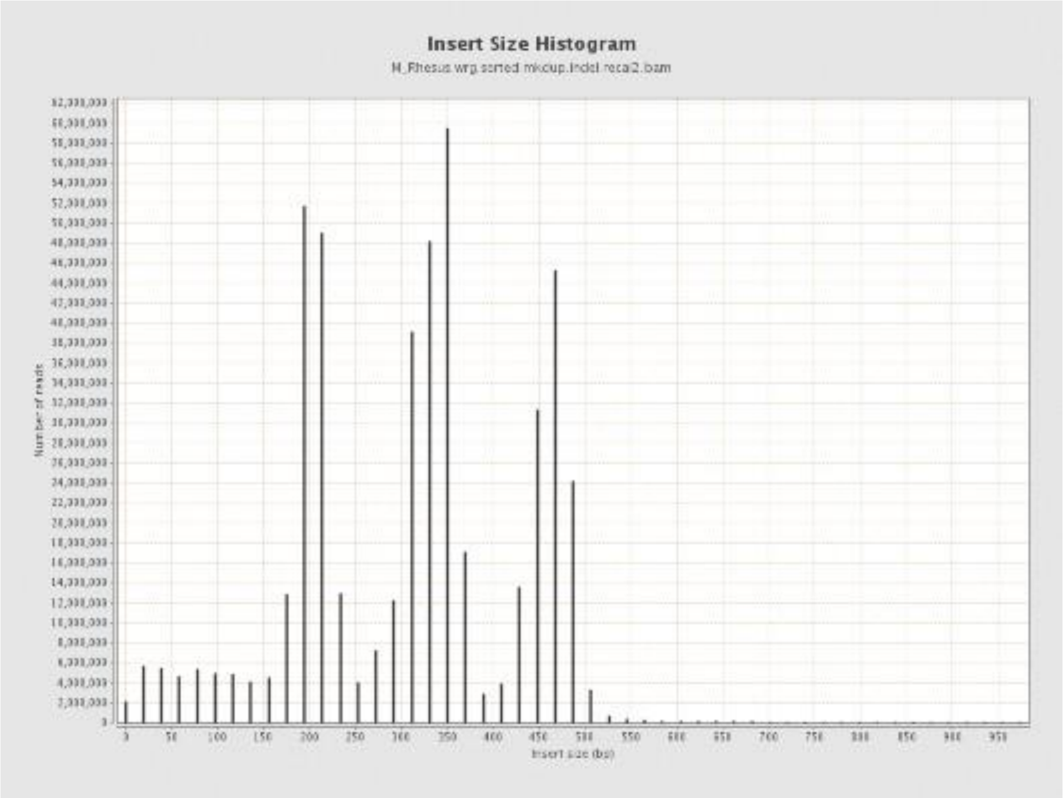

Mapped Reads Clipping Profile
M_Rhesus.wrg.sorted.mkdup.indel.recal2.bam
Clipped bases (%)
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
Read position (bp)


Homopolymer Indels
M_Rhesus.wrg.sorted.mkdup.indel.recal2.bam
Number of indels
13,000,000
12,500,000
12,000,000
11,500,000
11,000,000
10,500,000
10,000,000
9,500,000
9,000,000
8,500,000
8,000,000
7,500,000
7,000,000
6,500,000
6,000,000
5,500,000
5,000,000
4,500,000
4,000,000
3,500,000
3,000,000
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
polyA
polyC
polyG
polyT
polyN
Non-poly
Type of indel

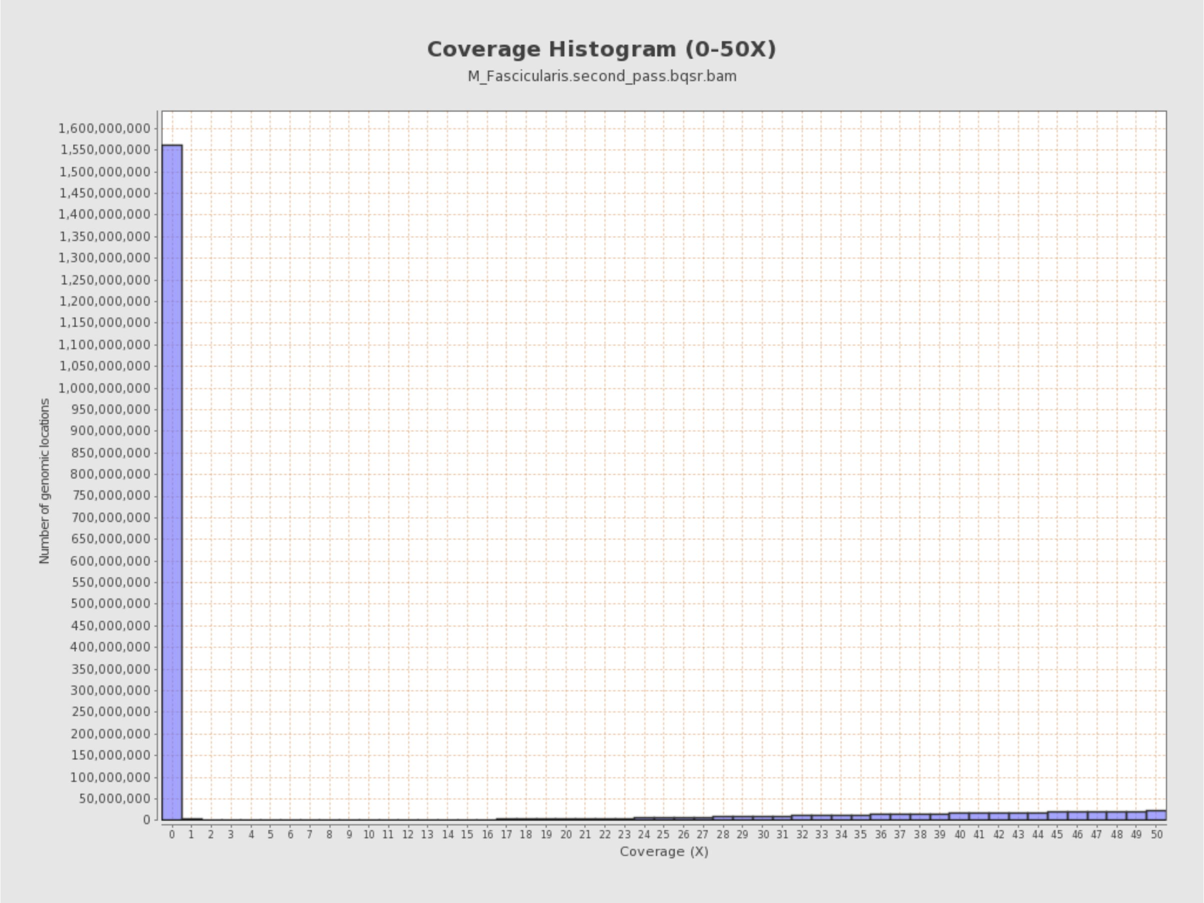

Mapping Quality Across Reference
M_Rhesus.wrg.sorted.mkdup.indel.recal2.bam
mapping quality
Mapping quality
250
240
230
220
210
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
250,000,000
500,000,000
750,000,000
1,000,000,000
1,250,000,000
1,500,000,000
1,750,000,000
2,000,000,000
2,250,000,000
2,500,000,000
2,750,000,000
Position (bp)

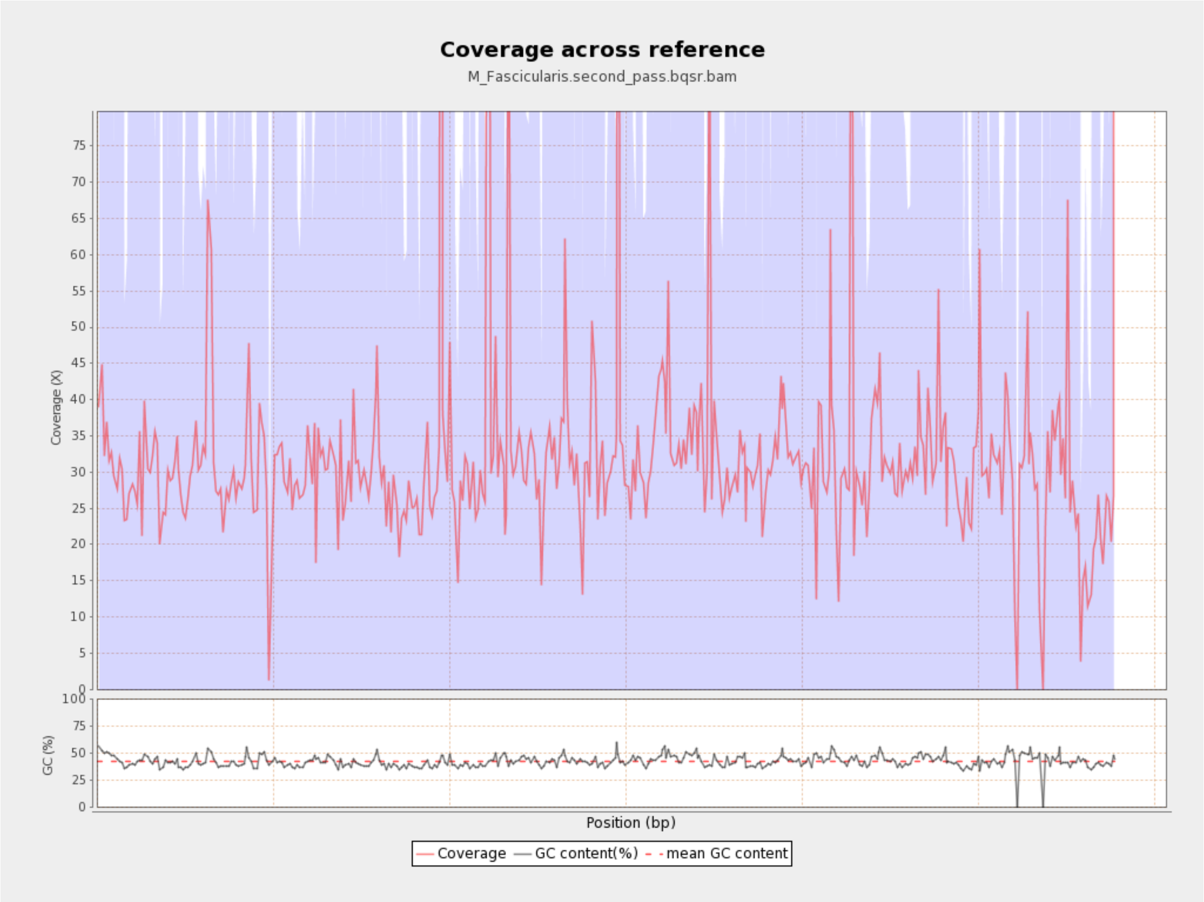

Mapping Quality Histogram
M_Rhesus.wrg.sorted.mkdup.Indel.recal2.bam
Number of genomic locations
1,200,000,000
1,150,000,000
1,100,000,000
1,050,000,000
1,000,000,000
950,000,000
900,000,000
850,000,000
800,000,000
750,000,000
700,000,000
650,000,000
600,000,000
550,000,000
500,000,000
450,000,000
400,000,000
350,000,000
300,000,000
250,000,000
200,000,000
150,000,000
100,000,000
50,000,000
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5
40.0
42.5
45.0
47.5
50.0
52.5
55.0
57.5
60.0
62.5
65.0
67.5
70.0
Mapping quality

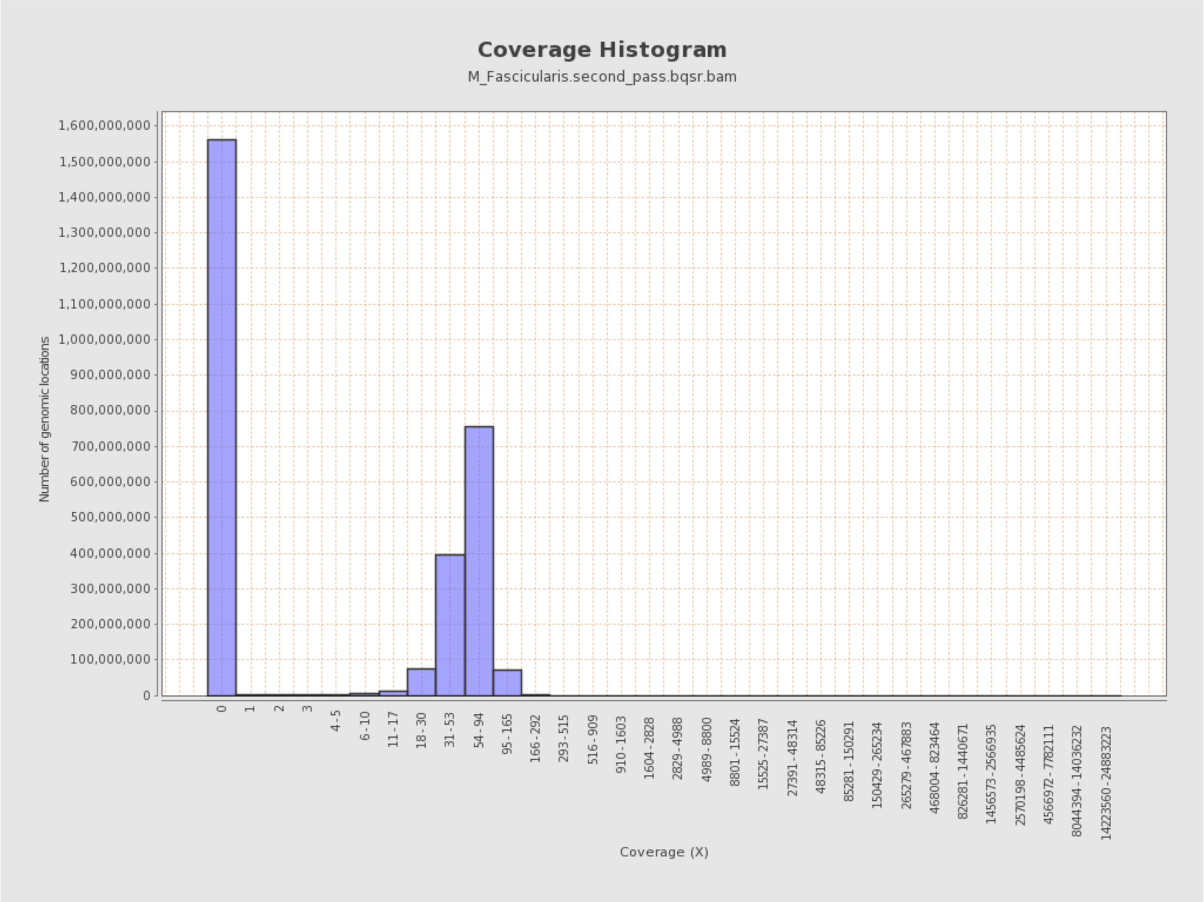

Insert Size Across Reference
M_Rhesus.wrg.sorted.mkdup.Indel.recal2.bam
insert size
Insert size (bp)
26,000
25,000
24,000
23,000
22,000
21,000
20,000
19,000
18,000
17,000
16,000
15,000
14,000
13,000
12,000
11,000
10,000
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
250,000,000
500,000,000
750,000,000
1,000,000,000
1,250,000,000
1,500,000,000
1,750,000,000
2,000,000,000
2,250,000,000
2,500,000,000
2,750,000,000
Position (bp)

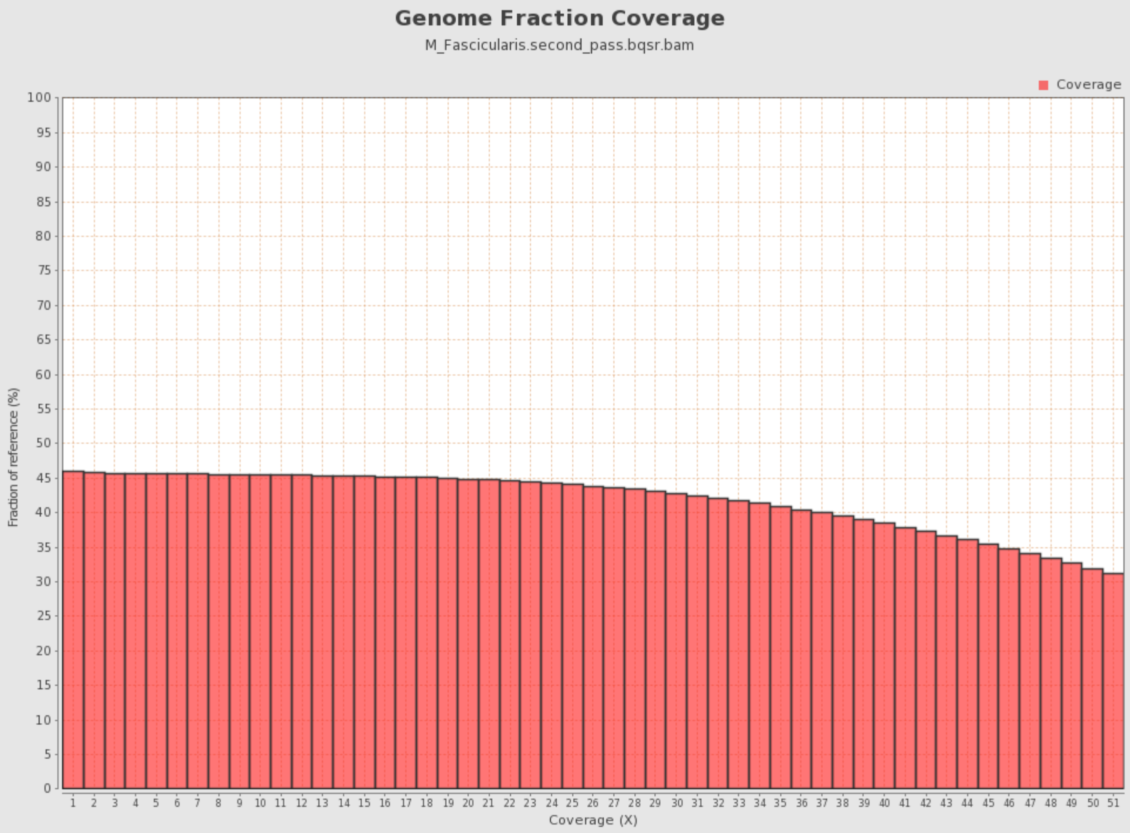


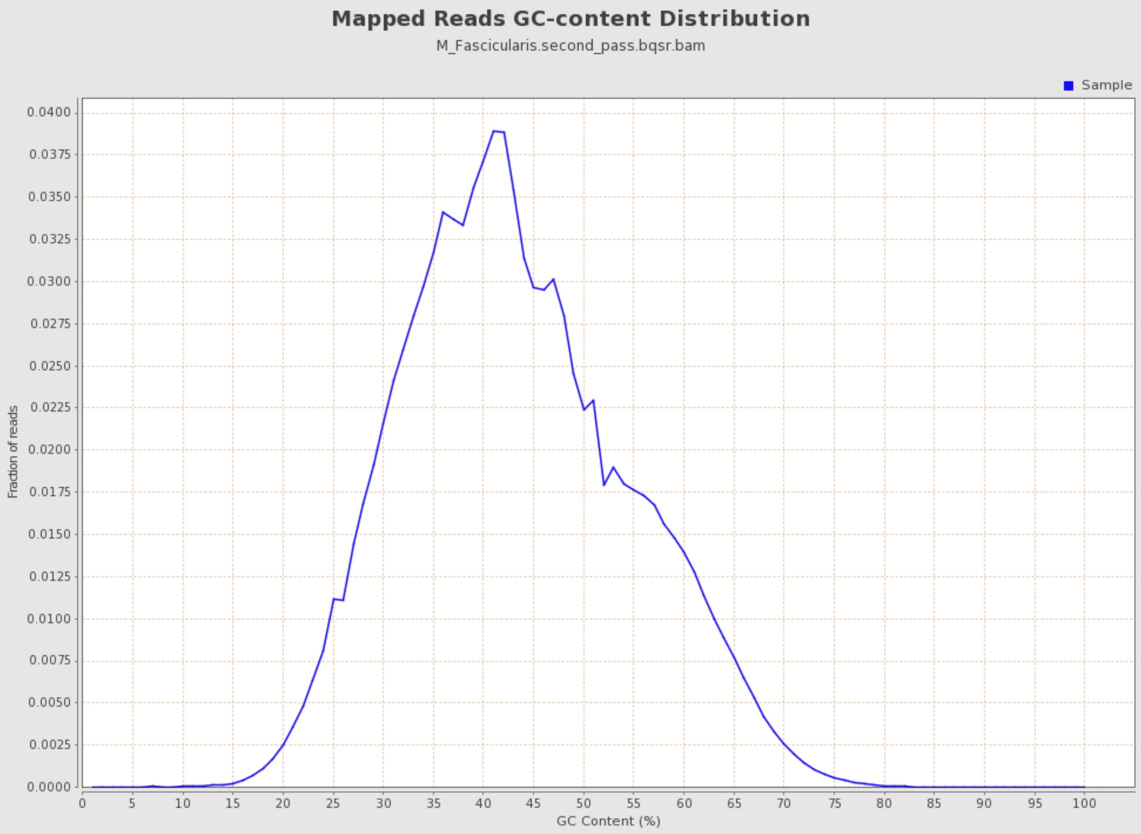
M.fascicularis BAMQC

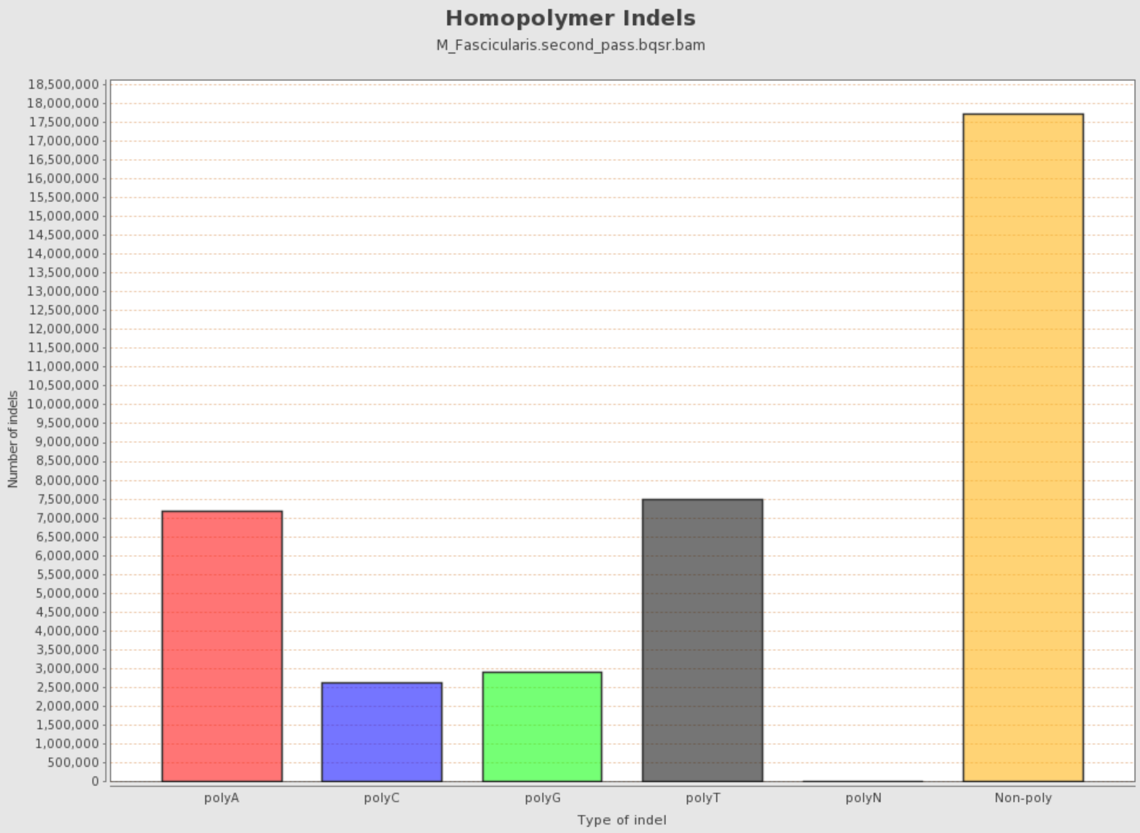


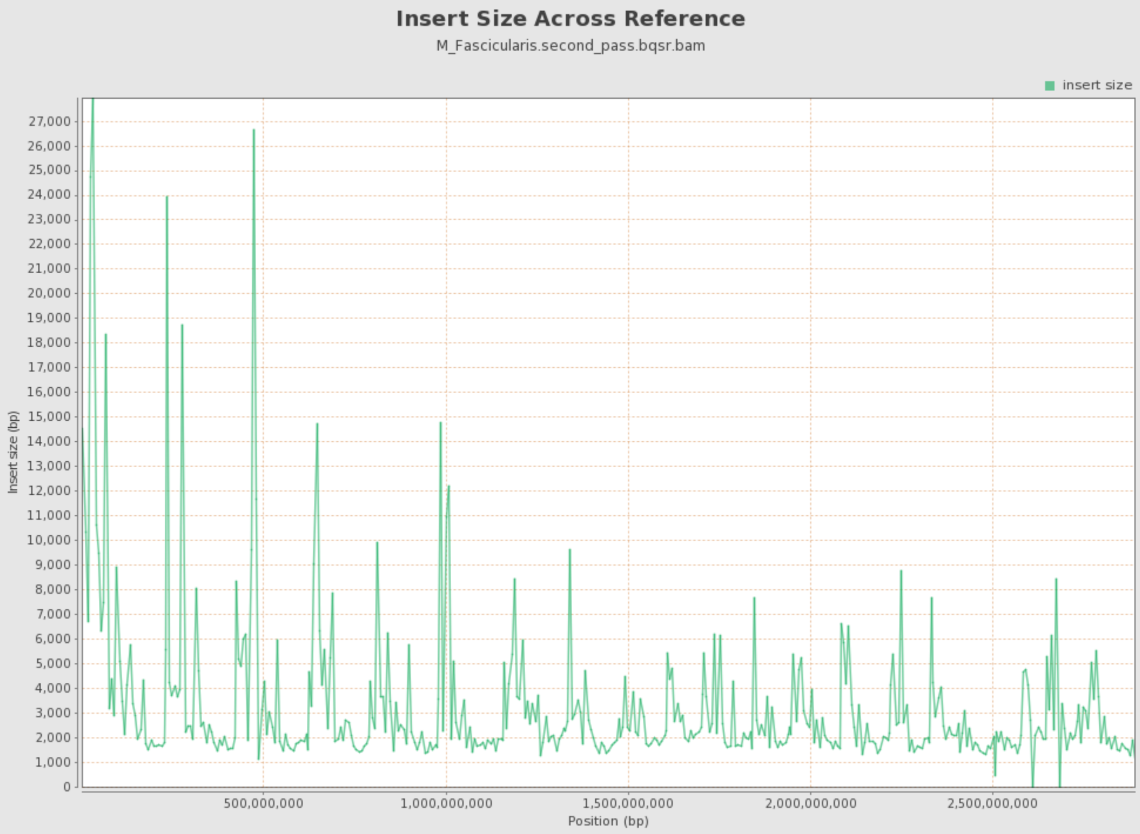

Coverage across reference
M_Fascicularis.second_pass.bqsr.bam
Coverage (X)
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
GC (%)
100
75
50
25
0
Position (bp)
Coverage
GC content(%)
mean GC content

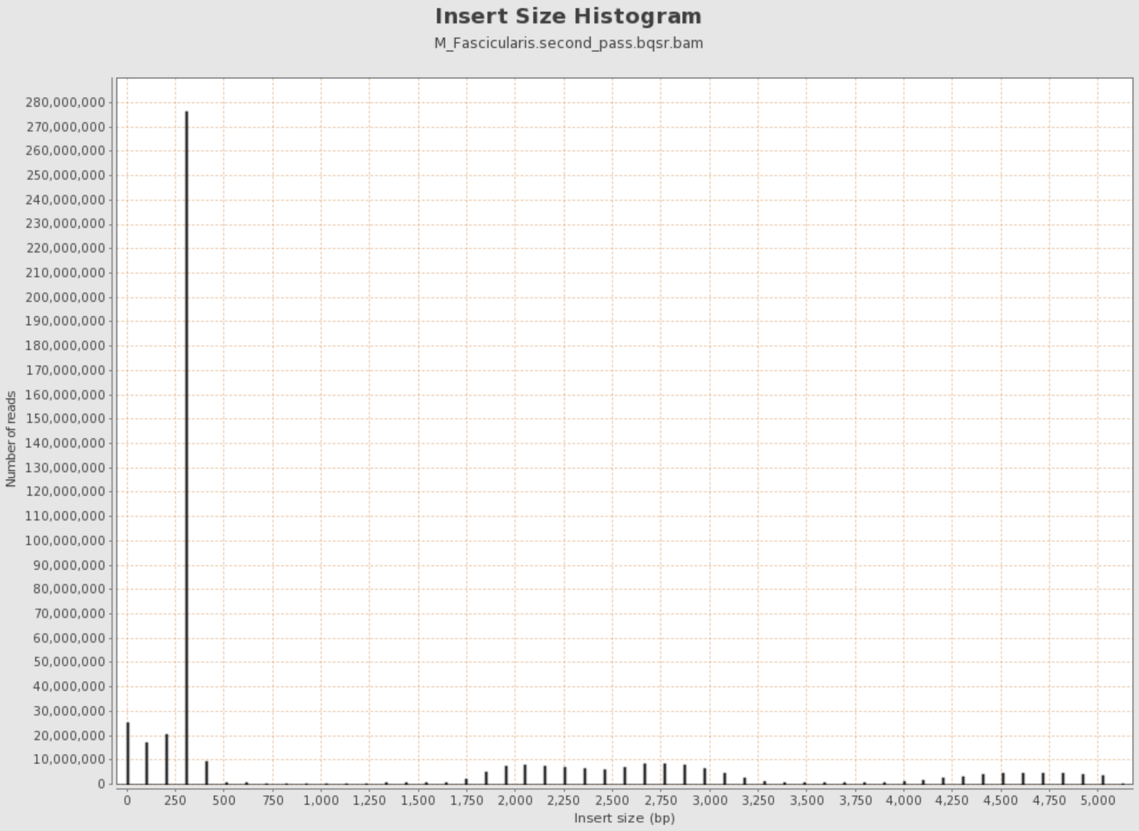

Coverage Histogram
M_Fascicularis.second_pass.bqsr.bam
Number of genomic locations
1,600,000,000
1,500,000,000
1,400,000,000
1,300,000,000
1,200,000,000
1,100,000,000
1,000,000,000
900,000,000
800,000,000
700,000,000
600,000,000
500,000,000
400,000,000
300,000,000
200,000,000
100,000,000
0
0
1
2
3
4 - 5
6 - 10
11 - 17
18 - 30
31 - 53
54 - 94
95 - 165
166 - 292
293 - 515
516 - 909
910 - 1603
1604 - 2828
2829 - 4988
4989 - 8800
8801 - 15524
15525 - 27387
27391 - 48314
48315 - 85226
85281 - 150291
150429 - 265234
265279 - 467883
468004 - 823464
826281 - 1440671
1456573 - 2566935
2570198 - 4485624
4566972 - 7782111
8044394 - 14036232
14223560 - 24883223
Coverage (X)

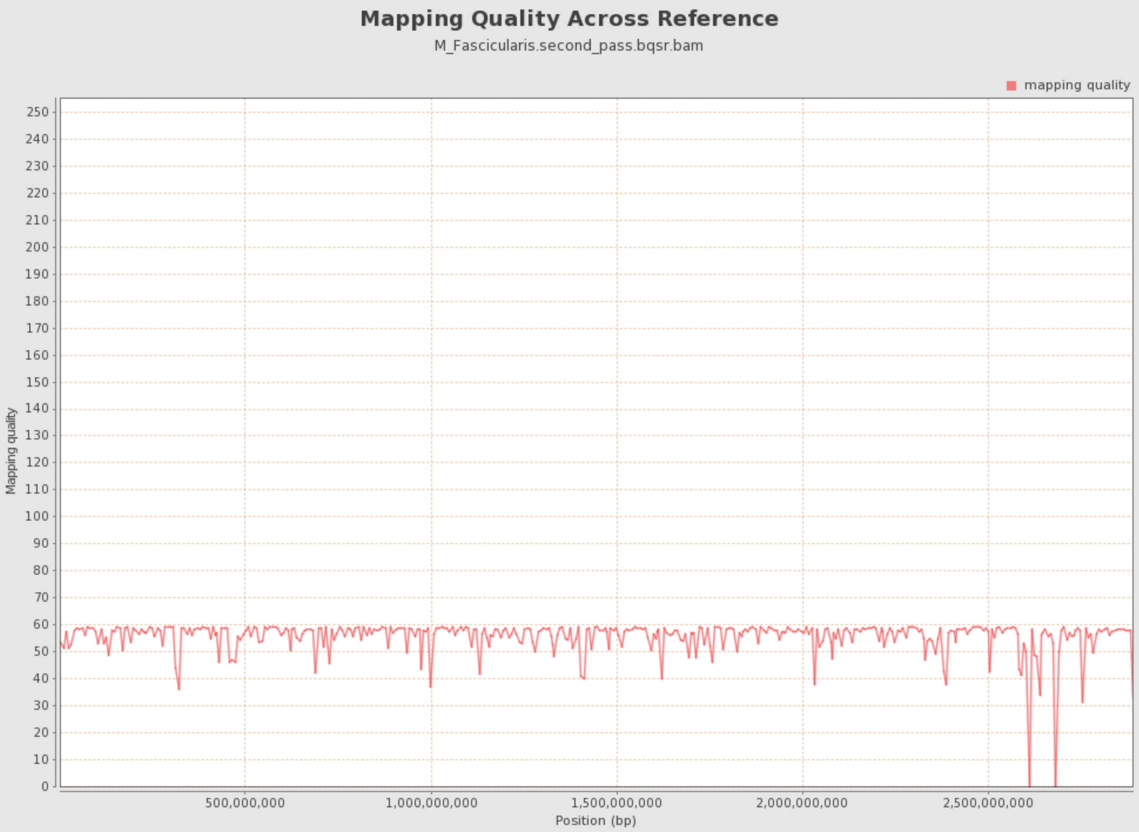

Genome Fraction Coverage
M_Fascicularis.second_pass.bqsr.bam
Coverage
Fraction of reference (%)
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
Coverage (X)

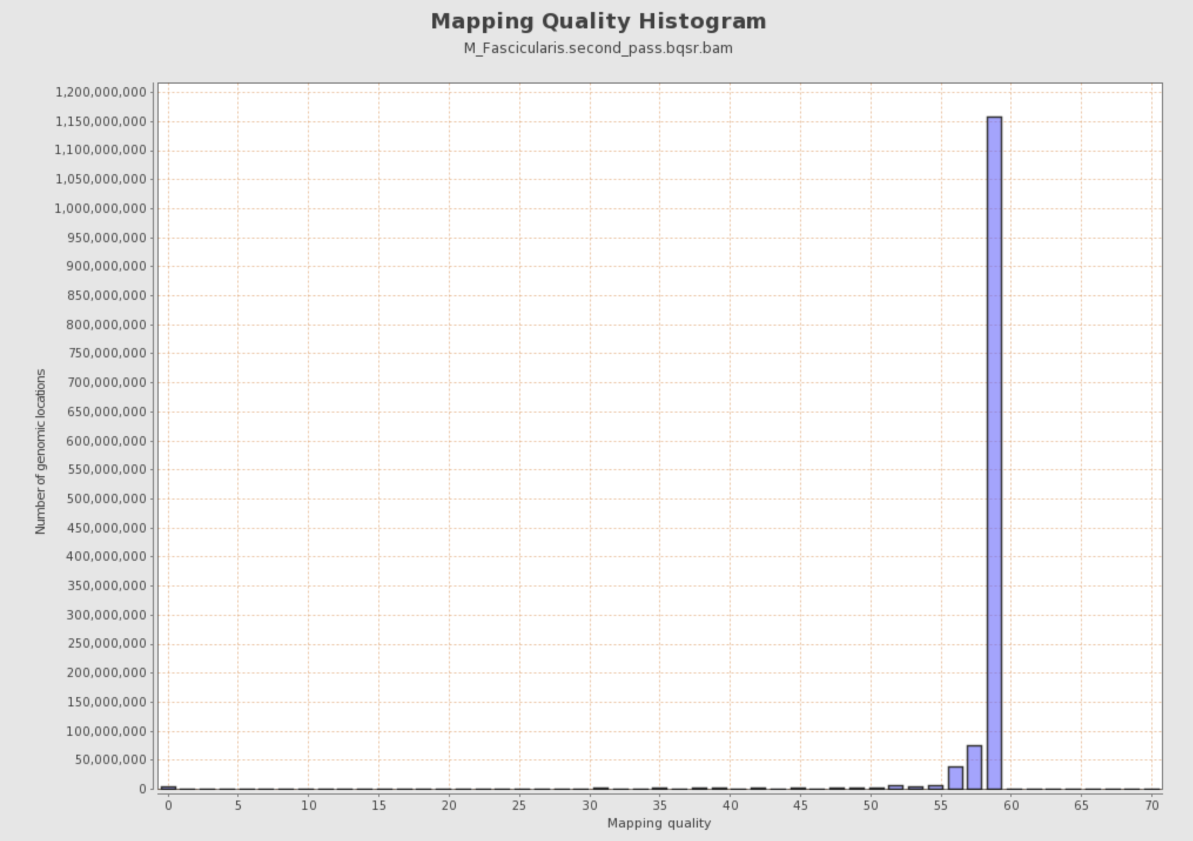

Mapped Reads GC-content Distribution
M_Fascicularis.second_pass.bqsr.bam
Sample
Fraction of reads
0.0400
0.0375
0.0350
0.0325
0.0300
0.0275
0.0250
0.0225
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
GC Content (%)

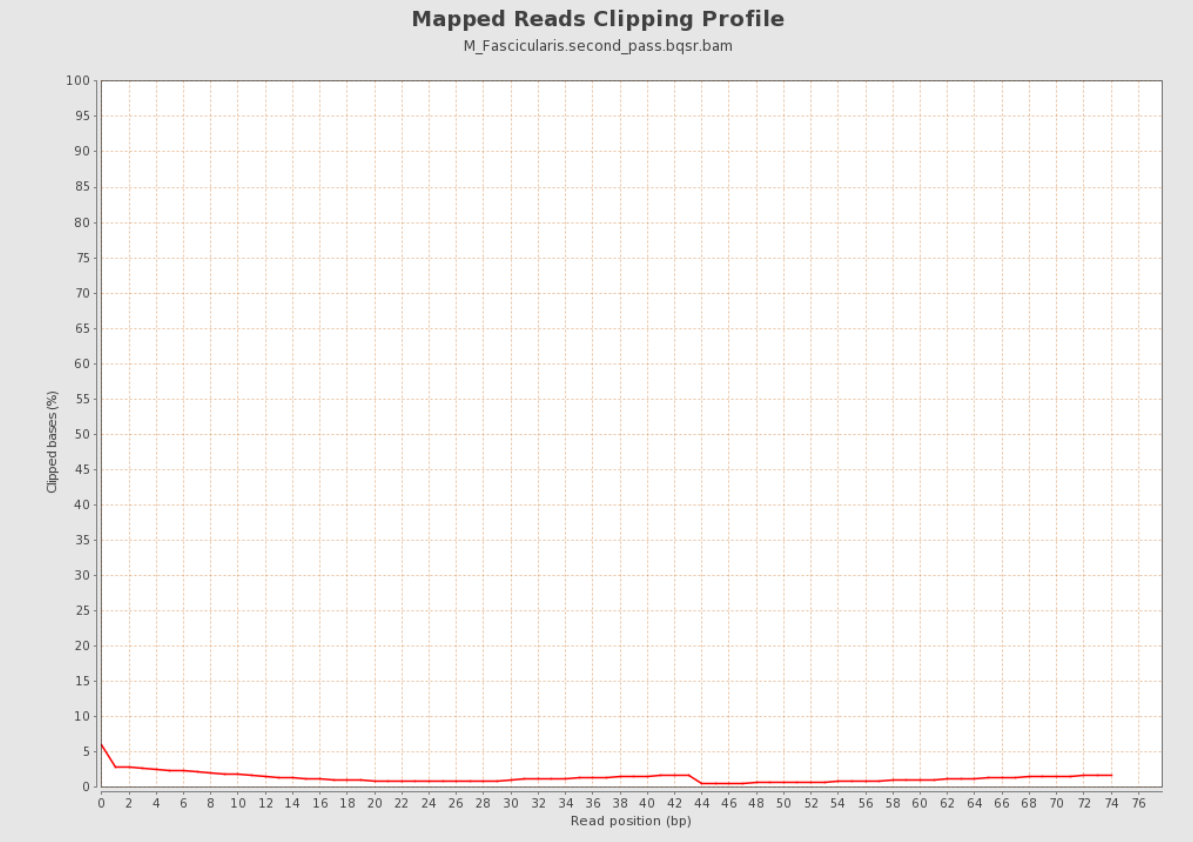

Homopolymer Indels
M_Fascicularis.second_pass.bqsr.bam
Number of indels
18,500,000
18,000,000
17,500,000
17,000,000
16,500,000
16,000,000
15,500,000
15,000,000
14,500,000
14,000,000
13,500,000
13,000,000
12,500,000
12,000,000
11,500,000
11,000,000
10,500,000
10,000,000
9,500,000
9,000,000
8,500,000
8,000,000
7,500,000
7,000,000
6,500,000
6,000,000
5,500,000
5,000,000
4,500,000
4,000,000
3,500,000
3,000,000
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
polyA
polyC
polyG
polyT
polyN
Non-poly
Type of indel

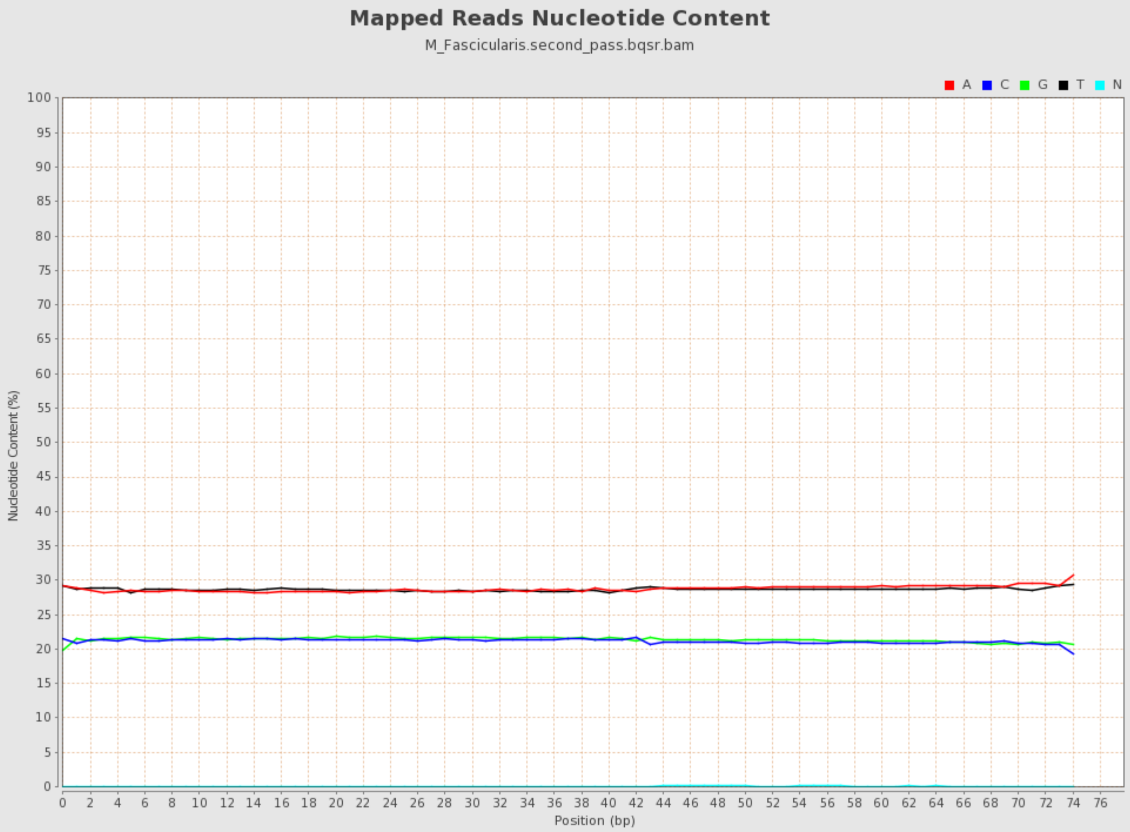

Insert Size Across Reference
M_Fascicularis.second_pass.bqsr.bam
insert size
Insert size (bp)
27,000
26,000
25,000
24,000
23,000
22,000
21,000
20,000
19,000
18,000
17,000
16,000
15,000
14,000
13,000
12,000
11,000
10,000
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
500,000,000
1,000,000,000
1,500,000,000
2,000,000,000
2,500,000,000
Position (bp)

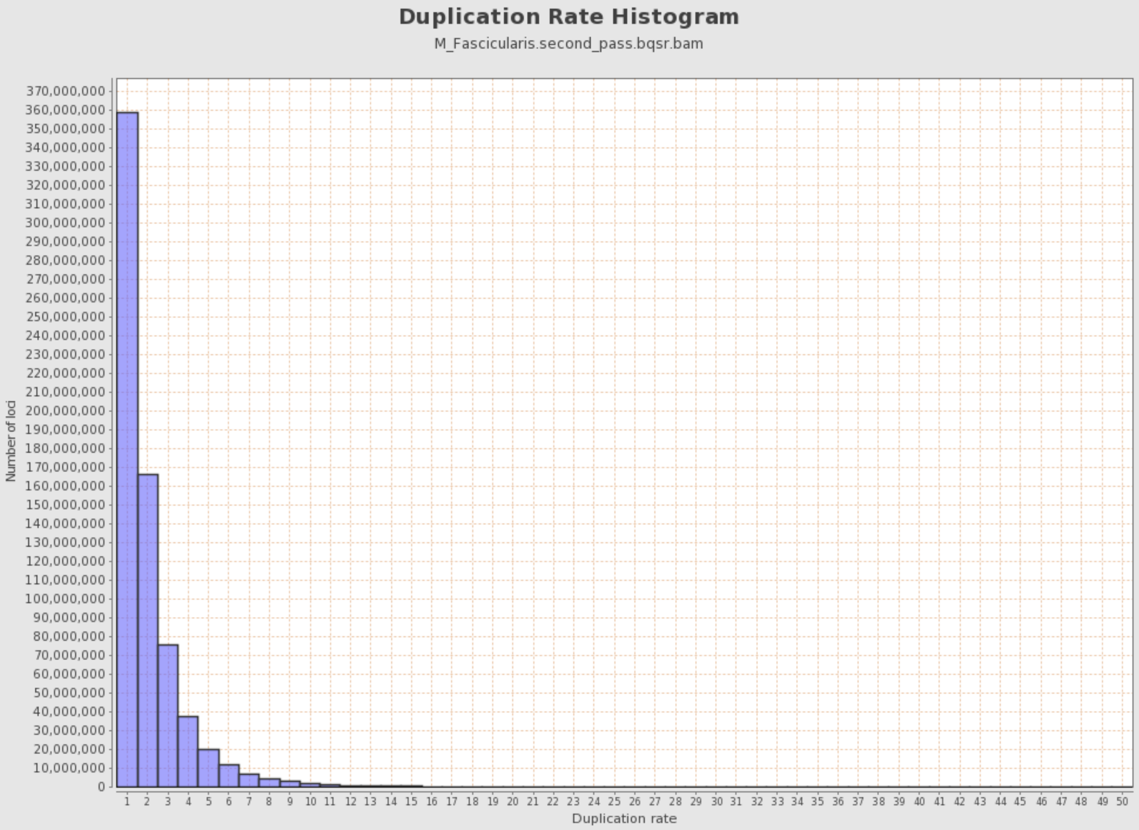

Insert Size Histogram
M_Fascicularis.second_pass.bqsr.bam
Number of reads
280,000,000
270,000,000
260,000,000
250,000,000
240,000,000
230,000,000
220,000,000
210,000,000
200,000,000
190,000,000
180,000,000
170,000,000
160,000,000
150,000,000
140,000,000
130,000,000
120,000,000
110,000,000
100,000,000
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
0
250
500
750
1,000
1,250
1,500
1,750
2,000
2,250
2,500
2,750
3,000
3,250
3,500
3,750
4,000
4,250
4,500
4,750
5,000
Insert size (bp)


Mapping Quality Across Reference
M_Fascicularis.second_pass.bqsr.bam
mapping quality
Mapping quality
250
240
230
220
210
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
500,000,000
1,000,000,000
1,500,000,000
2,000,000,000
2,500,000,000
Position (bp)


Mapping Quality Histogram
M_Fascicularis.second_pass.bqsr.bam
Number of genomic locations
1,200,000,000
1,150,000,000
1,100,000,000
1,050,000,000
1,000,000,000
950,000,000
900,000,000
850,000,000
800,000,000
750,000,000
700,000,000
650,000,000
600,000,000
550,000,000
500,000,000
450,000,000
400,000,000
350,000,000
300,000,000
250,000,000
200,000,000
150,000,000
100,000,000
50,000,000
0
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
Mapping quality


Mapped Reads Clipping Profile
M_Fascicularis.second_pass.bqsr.bam
Clipped bases (%)
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50
52
54
56
58
60
62
64
66
68
70
72
74
76
Read position (bp)


Mapped Reads Nucleotide Content
M_Fascicularis.second_pass.bqsr.bam
🟥 A 🟦 C 🟩 G ⬛ T 🟦 N
Nucleotide Content (%)
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76
Position (bp)


Duplication Rate Histogram
M_Fascicularis.second_pass.bqsr.bam
Number of loci
370,000,000
360,000,000
350,000,000
340,000,000
330,000,000
320,000,000
310,000,000
300,000,000
290,000,000
280,000,000
270,000,000
260,000,000
250,000,000
240,000,000
230,000,000
220,000,000
210,000,000
200,000,000
190,000,000
180,000,000
170,000,000
160,000,000
150,000,000
140,000,000
130,000,000
120,000,000
110,000,000
100,000,000
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
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Duplication rate