GENOMIC AND TRANSCRIPTOMIC IMPACTS OF SMALL- AND LARGE-SCALE SPONTANEOUS MUTATIONS IN YEASTS

by

HOLLY CELINA MCQUEARY

(Under the Direction of David Hall)

ABSTRACT

Mutations are an essential evolutionary force for all living organisms on Earth. They occur at random and can be on the scale of nucleotides of chromosomes, or of genomes. Understanding the dynamics of mutation is an essential step in understanding biodiversity and evolution of life. Much is known about outside factors influencing mutation rates and spectra – i.e. UV radiation or mutagenic chemicals. However, there is little known about the effects of factors within genomes, such as copy number changes or presence of mobile genetic elements, on the genomic and transcriptomic scales. In this dissertation, I describe two studies on intragenomic factors: one on the transcriptomic consequences of aneuploidy and possibility of dosage compensation, and another on the effects of the presence of mobile genetic elements on mutation rate and spectra. In both of these studies, I use species from the genus Saccharomyces: S. cerevisiae and S. paradoxus, which are both incredibly useful model organisms for this sort of analysis. Each study involves a mutation accumulation experiment followed by whole-transcriptome or whole-genome sequencing and analysis. The studies presented here provide evidence that characteristics of genomes themselves have an effect on the transcription of genes and the accumulation of mutations over evolutionary time.

INDEX WORDS: mutation accumulation, aneuploidy, yeasts, mobile genetic elements, transposable elements, whole-genome sequencing, whole-transcriptome sequencing

GENOMIC AND TRANSCRIPTOMIC IMPACTS OF SMALL- AND LARGE-SCALE SPONTANEOUS MUTATIONS IN YEASTS

by

HOLLY CELINA MCQUEARY

BS, University of South Florida, 2015

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2020

GENOMIC AND TRANSCRIPTOMIC IMPACTS OF SMALL- AND LARGE-SCALE SPONTANEOUS MUTATIONS IN YEASTS

by

HOLLY CELINA MCQUEARY

Major Professor:	David Hall
Committee:	Kelly Dyer
	Douda Bensasson
	David Garfinkel
	Michael McEachern

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
August 2020

DEDICATION

I dedicate this dissertation to my adorable dog, Penelope, who has been my companion for nearly this entire process and has put up with trips to the lab, long hours at home alone, and my overall stress levels. I also would like to dedicate this to Lazarus, my box turtle who I adopted from the lab when we had to get rid of the lab pets. He has been my little companion the entire time I've been in the Hall lab and brings me joy daily. He now lives in the Dyer lab, where I know he will have a lovely lab life and be taken care of by many undergraduates and graduate students to come.

ACKNOWLEDGEMENTS

There are countless people who have helped me along the way on this endeavor. My mom has constantly been there for me to listen, provide encouragement, and overall support me in my academic pursuits. My dad and stepmom have also supported me along this journey, even if our views do not always align. My best friend Michael Nguyen has been my rock through this, even though we live far apart, we remain in contact almost daily and he has been an amazing friend. My best friend Karen Bobier has been my rock here in Athens, and without her and her husband Jeff I'm not sure I would be writing this today. Karen has not only been a fantastic life friend, but also a great work friend with whom I can bounce ideas off of and debate various topics in science. Sam Arsenault has been a fantastic friend throughout my grad school years and has been with me through the final processes of writing this dissertation; his love and support has made this work much easier and fun. My undergraduates from the Hall lab have been essential to my PhD work, as I have been the only grad student the majority of the time here. Shout-out to Sam Demario, Alexander Jamarillo, Brooke Hull, Alexandria Mulliken, Brittania Johnson, Ariella Tsfoni, Emma Fullet, and Anastacia Bankey. Our collaboration with Casey Bergman and his lab has been crucial to many of the analyses here, and Jingxuan Chen has been an immense help with this work. I would also like to acknowledge Audrey Ward and Nan Yao, who joined the lab right before I left. I wish we had spent more time together as labmates, but I look forward to remaining colleagues in the future!

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Mutations are an essential force in the evolution and adaptation of all living organisms on Earth. They occur at random and without regard to an organism's environment or biological needs. Mutations can be on the scale of nucleotides (nucleotide substitutions, small insertions or deletions, horizontal gene transfer events, transposable elements, and sequence duplications), of chromosomes (aneuploidies, inversions, translocations), or of genomes (whole-genome duplications). Depending on the scale of the mutation, it can have wildly different effects on the organism's genotype and phenotype (and thus their adaptation and evolution). To fully understand the evolutionary history and trajectory of organisms, understanding the mutational process and factors affecting it is essential. In particular, understanding the mutational dynamics of a "null" environment (i.e. without any selective bias) is essential. In order for selection to overcome genetic drift, the strength of selection must be greater than $1/nN_e$ where n is the ploidy of the organism (1 for haploids, 2 for diploids) and N_e is the effective population size. Therefore, in populations with small effective population size, genetic drift has a greater impact than natural selection (Charlesworth 2009). Mutation accumulation experiments aim to eliminate selection or to at least decrease it dramatically so that it has less of an impact on mutations fixing than genetic drift. This is accomplished by decreasing the effective population size (number of individuals giving rise to the next

generation). In this dissertation, I present two studies that utilize mutation accumulation experiments to assess the parameters and influences of spontaneous mutations in yeast.

Some factors affecting mutations are known – such as mutagens and other outside forces. However, the genomic and transcriptomic effects of intragenomic phenomena such as aneuploidy or mobile genetic elements are unclear. Aneuploidy, in which a cell contains a number of chromosomes that is not a multiple of the haploid state, can be considered an intragenomic factor capable of influencing gene expression. Another is the presence of mobile genetic elements such as transposable elements (TEs, or transposons). The effects of such aneuploidy events and mobile genetic elements on the transcription of genes or mutations in other parts of the genome not directly impacted are unclear. In this work, I will be focusing on the transcriptional impacts of spontaneous whole-chromosomal aneuploidies and the impact of mobile genetic element presence on small and large spontaneous genomic mutations in yeast.

Aneuploidy is the duplication or deletion of one or more chromosomes in an organism – a deviation from the normal number of chromosomes. This often has phenotypic consequences, including inviability and developmental issues as is seen in human aneuploidies such as trisomy X, 18, and 21 (Hassold and Hunt 2001). The exact reasoning for aneuploidy to cause defects is unclear, but it is related to the relative dosage of transcripts produced from the affected chromosomes (Torres et al. 2008). A way to circumvent this issue is to up- or down-regulate the expression of genes on the affected chromosomes, a phenomenon known as dosage compensation (DC). Dosage compensation is evident in species having heteromorphic sex chromosomes (Lyon 1962). In XY and ZW systems, one sex is hemizygous for a sex chromosome, resulting in

differential gene dose for sex-linked genes in the two sexes. Some species have evolved mechanisms to balance gene dose between the sexes. Three mechanisms are well-understood: C. *elegans* hermaphrodites (XX), genes on both of the X chromosomes are downregulated (Itoh et al. 2007); mammalian females (XX) condense and inactivate one X chromosome early in development (these structures are known as Barr bodies) (Lyon 1962; Ignacio Marin 2000); and Drosophila males (XY), which double the transcription rate of genes on the X chromosome (Ignacio Marin 2000; Stenberg and Larsson 2011). After gene duplication (through whole-chromosomal aneuploidies or other routes), there are three possible outcomes: nonfunctionalization, in which one copy is silenced by mutations that cause it to degrade; subfunctionalization, where both copies acquire mutations to the point that they both need to be expressed to obtain normal protein function; or neofunctionalization, where one copy gains a new function that is beneficial, while the other retains its original function (Lynch and Conery 2000). It is not clear how the cell tolerates the extra gene copies long enough for them to evolve one of the above possibilities, but a mechanism of dosage compensation employed during this time would allow the duplicate to persist long enough to accumulate mutations that would allow for new gene function.

Interestingly, not all taxa seem to have the same requirement of a dosage compensation mechanism for sex chromosomes. Birds, in which males are ZZ and females are ZW (female-heterogametic), have been shown to be lacking an effective dosage compensation mechanism (Itoh et al. 2007), and studies have shown that different bird species have varying degrees of dosage compensation, implying that avian taxa evolved different mechanisms to cope with dosage imbalance (Yuichiro Itoh 2010). No

dosage compensation has been seen in several other species including *Schistosoma mansoni* (Vicoso and Bachtrog 2011) and silkworms (Zha et al. 2009). A study has shown that dosage compensation in plants acts on a gene-by-gene basis (Alexander S. T. Papadopulos 2015).

Dosage compensation is not limited to sex chromosomes, however. Autosomal dosage compensation has been found to occur in fish and fruit flies. The triploid hybrid fish, *Squalius alburnoides*, employs a dosage compensation mechanism at the individual gene level to repair the imbalance caused by having three copies of each chromosome (Pala et al. 2008). *Drosophila* possess a dosage compensation mechanism that acts at the gene level to upregulate transcripts to repair the imbalance of gene dose caused by hemizygosity (Hangnoh Lee 2016).

Several studies have examined the effects of aneuploidy in *Saccharomyces cerevisiae,* and have found that aneuploids display several distinct phenotypes characterized by cell cycle defects, increased glucose uptake, and increased sensitivity to environments that interfere with the synthesis and folding of proteins (Eduardo M Torres 2016). These negative phenotypic effects suggest little to no intrinsic dosage compensation at the transcript level in this species. Laboratory strains seem to be less tolerant of aneuploidy than wild strains (James Hose 2015). Wild strains appear to use some type of compensation at the transcript level to buffer the deleterious effects of gene amplification (James Hose 2015), though this has been highly debated (Audrey P Gasch 2016; Eduardo M Torres 2016).

As for nucleotide-level mutations, transposable elements (transposons or TEs) are a type of mobile genetic element found in many taxa (Bourque et al. 2018). There are two

types of transposons: those that replicate through a DNA intermediate (DNA transposons) and those that replicate through an RNA intermediate (retrotransposons). It is known that transposons cause mutations when they move – either when they are excised from the genome, leaving mutational scars, or when they insert into the genome, causing homology-directed double strand break repair and mutations (Wicker et al. 2016). However, it is not known if the presence of transposable elements in the genome has an effect on mutations that occur independent of transposition events. One reason for this being possible is that a large proportion of the RNA in species is composed of transposon transcripts – in yeast, for example, Ty1 transcripts make up ~5-10% of the cell's poly-A RNA, despite TE DNA sequence being at most 3% of the genome (Elder et al. 1981; Kim et al. 1998; Liti et al. 2009; Carr et al. 2012). An excess of transcripts that are not necessary to a cell's own processes could increase stress to the cell and therefore could incur more mutations. Next-generation sequencing allows us to investigate large quantities of DNA and RNA from many samples, giving us the opportunity to discover the scope of mutations accumulated through genetic drift as well as investigate the expression of genes on duplicated or deleted chromosomes to find any anomalous effects.

In this dissertation I present two studies investigating what impacts aneuploidy and transposable elements have on the transcriptome and genome of two yeast species, *Saccharomyces cerevisiae* and *Saccharomyces paradoxus*, respectively. In chapter 2, we present an analysis of gene expression levels in *S. cerevisiae* lines that have undergone spontaneous aneuploidy events during a mutation accumulation experiment and investigate the impacts of heterozygosity on aneuploidy rate. In chapter 3, we present the effects of transposable element presence in a genome with a mutation accumulation study

of two strains of *S. paradoxus*: a haploid with 0 Ty1 elements, and a haploid with 1 Ty1 element. Chapter 4 contains a conclusion and where this research leads to in the future.

References

Alexander S. T. Papadopulos, M. C., Kate Ridout, Dmitry A. Filatov, 2015 Rapid Y degeneration and dosage compensation in plant sex chromosomes. PNAS 112**:** 13021-13026.

Audrey P Gasch, J. H., Michael A Newton, Maria Sardi, Mun Yong, Zhishi Wang, 2016 Further support for aneuploidy tolerance in wild yeast and effects of dosage compensation on gene copy-number evolution. eLIFE 5**:** 1-12.

Bourque, G., K. H. Burns, M. Gehring, V. Gorbunova, A. Seluanov *et al.*, 2018 Ten things you should know about transposable elements. Genome biology 19**:** 199.

Carr, M., D. Bensasson and C. M. Bergman, 2012 Evolutionary genomics of transposable elements in Saccharomyces cerevisiae. PloS one 7**:** e50978.

Charlesworth, B., 2009 Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics 10**:** 195-205.

Eduardo M Torres, M. S., Angelika Amon, 2016 No current evidence for widespread dosage compensation in S. cerevisiae. eLIFE 5**:** 1-19.

Elder, R., T. S. John, D. Stinchcomb and R. Davis, 1981 Studies on the transposable element Tyl of yeast I. RNA homologous to Tyl, pp. 581-591 in *Cold Spring Harbor symposia on quantitative biology*. Cold Spring Harbor Laboratory Press.

Hangnoh Lee, D.-Y. C., Cale Whitworth, Robert Eisman, Melissa Phelps, John Roote, Thomas Kaufman, Kevin Cook, Steven Russell, Teresa Przytycka, Brian Oliver, 2016 Effects of Gene Dose, Chromatin, and Network Topology on Expression in Drosophila melanogaster. PLoS Genetics 12.

Hassold, T., and P. Hunt, 2001 To err (meiotically) is human: the genesis of human aneuploidy. Nature Reviews Genetics 2**:** 280.

Ignacio Marin, M. L. S., Bruce S. Baker, 2000 The evolution of dosage-compensation mechanisms. BioEssays 22**:** 1106-1114.

Itoh, Y., E. Melamed, X. Yang, K. Kampf, S. Wang *et al.*, 2007 Dosage compensation is less effective in birds than in mammals. J Biol 6**:** 2.

James Hose, C. M. Y., Maria Sardi, Zhishi Wang, Michael A Newton, Audrey P Gasch, 2015 Dosage compensation can buffer copy-number variation in yeast. eLIFE 4**:** 1-27.

Kim, J. M., S. Vanguri, J. D. Boeke, A. Gabriel and D. F. Voytas, 1998 Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. Genome research 8**:** 464-478.

Liti, G., D. M. Carter, A. M. Moses, J. Warringer, L. Parts *et al.*, 2009 Population genomics of domestic and wild yeasts. Nature 458**:** 337-341.

Lynch, M., and J. S. Conery, 2000 The evolutionary fate and consequences of duplicate genes. Science 290**:** 1151-1155.

Lyon, M. F., 1962 Sex chromatin and gene action in the mammalian X-chromosome. American journal of human genetics 14**:** 135.

Pala, I., M. M. Coelho and M. Schartl, 2008 Dosage compensation by gene-copy silencing in a triploid hybrid fish. Current Biology 18**:** 1344-1348.

Stenberg, P., and J. Larsson, 2011 Buffering and the evolution of chromosome-wide gene regulation. Chromosoma 120**:** 213-225.

Torres, E. M., B. R. Williams and A. Amon, 2008 Aneuploidy: cells losing their balance. Genetics 179**:** 737-746.

Vicoso, B., and D. Bachtrog, 2011 Lack of global dosage compensation in Schistosoma mansoni, a female-heterogametic parasite. Genome Biology and Evolution 3**:** 230-235.

Wicker, T., Y. Yu, G. Haberer, K. F. Mayer, P. R. Marri *et al.*, 2016 DNA transposon activity is associated with increased mutation rates in genes of rice and other grasses. Nature communications 7**:** 1-9.

Yuichiro Itoh, K. R., Yong-Hwan Kim, Juli Wade, David F. Clayton, Arthur P. Arnold, 2010 Sex bias and dosage compensation in the zebra finch versus chicken genomes: General and specialized patterns among birds. Genome Research 20**:** 512-518.

Zha, X., Q. Xia, J. Duan, C. Wang, N. He *et al.*, 2009 Dosage analysis of Z chromosome genes using microarray in silkworm, Bombyx mori. Insect biochemistry and molecular biology 39**:** 315-321.

CHAPTER 2

ANEUPLOIDY CAUSES WIDESPREAD GENE EXPRESSION CHANGES AND IS NOT MEDIATED BY WHOLE-CHROMOSOME DOSAGE COMPENSATION IN YEAST

Introduction

Aneuploidy occurs when an organism contains an abnormal number of one or a few chromosomes. Familiar examples are those causing human disorders, such as Down's syndrome (trisomy 21) or Turner's syndrome (monosomy X) (HASSOLD AND HUNT 2001). While autosomal aneuploidies are generally deleterious in most organisms, presumably because of dosage problems (CHUNDURI AND STORCHOVA 2019), in some species, aneuploidies are surprisingly common, such as in some wild yeast (*Saccharomyces cerevisiae*) isolates (STROPE *et al.* 2015). There is debate as to why aneuploidy is tolerated, or even favored, in such populations. Some hypothesize there is an intrinsic mechanism of dosage compensation to buffer the deleterious effects of imbalanced gene dosage (HOSE *et al.* 2015; GASCH *et al.* 2016), similar to the mechanism of dosage compensation observed in sex chromosomes (MARIN *et al.* 2000). Such autosomal compensation has been observed in *Drosophila* and other species (BIRCHLER *et al.* 1990; MATOS *et al.* 2015). In yeast, the presence of such a mechanism has been contested, with some studies concluding that there is no evidence for dosage compensation at the whole-chromosome level (TORRES *et al.* 2010). Others suggest that aneuploid wild yeast attenuate protein levels by increasing protease activity or upregulating genes that are part of multiprotein complexes so that the relative dosages are more even (CHEN *et al.* 2003; VEITIA *et al.* 2008). It

has also been shown experimentally that the accumulation or loss of chromosomes can be adaptive in certain environments (SELMECKI *et al.* 2006; PAVELKA *et al.* 2010; CHEN *et al.* 2012; YONA *et al.* 2012; SELMECKI *et al.* 2015; DE VRIES *et al.* 2018). For example, yeast grown in an oxide-rich media accumulate an extra copy of chromosome XI as a response to oxygen stress (KAYA *et al.* 2015), and resistance to fluconazole in *Candida albicans* involves acquisition of aneuploidy (WAKABAYASHI *et al.* 2017; KOO *et al.* 2018).

Understanding dosage compensation is important for several reasons. Aneuploidy cannot be avoided because segregation machinery is not perfect. As such, determining whether there are intrinsic mechanisms of dosage compensation gives insight into the likely consequences of such aneuploidy. Dosage compensation is also critically important during the evolution of sex chromosomes from homomorphic autosomes (CHARLESWORTH 1991). Dosage compensation is thought to play a critical role in the evolution of sex chromosomes because of their different copy numbers in males versus females, a common example being the X-chromosome in XY systems. There are a variety of ways in which dosage compensation occurs in sex chromosomes to make up for differences in gene dosage between the sexes (CHANDLER 2017). However, the degree to which compensation evolves prior to, during, or after the evolution of dimorphism remains an open question (GU AND WALTERS 2017).

While dosage compensation has been observed for autosomes in *Drosophila* (DEVLIN *et al.* 1982; BIRCHLER *et al.* 1990; MCANALLY AND YAMPOLSKY 2009; CHEN AND OLIVER 2015; HANGNOH LEE 2016; LEE *et al.* 2016), it is unknown whether such an intrinsic mechanism exists in yeast. The fact that yeast are often found to be aneuploid in natural isolates (STROPE *et al.* 2015) could suggest that aneuploidy causes changes in gene expression that are adaptive, and no DC exists (KAYA *et al.* 2015; LINDER *et al.* 2017). Alternatively, yeast may be naturally robust to aneuploidy, so that aneuploid strains do not differ in fitness and thus occur in nature as neutral variants. The second hypothesis, coupled with the occurrence of aneuploid strains at reasonably

high frequencies, suggests that yeast may contain an innate mechanism for attenuating or compensating for differences in gene dose and that mutation to aneuploidy is relatively frequent.

To fully understand the effects of aneuploidy on yeast populations, we seek estimates of the rate of aneuploidy and the effects of aneuploidy on gene expression. Previous studies have observed the effects of aneuploidy in wild yeast populations, where selection is acting, and in chemically- or mitotically-induced aneuploids, where the rate of production of aneuploids is being manipulated (LINDER *et al.* 2017); (CAMPBELL *et al.* 1981; ANDERS *et al.* 2009; MULLA *et al.* 2014). In this study, we sought to determine the spontaneous rate of aneuploid formation for each chromosome and the effects of aneuploidy on gene expression in two strains of diploid yeast in the absence of selection. In each strain, aneuploid events were captured during a 2000-generation mutation accumulation (MA) experiment (Figure 1) with a single-cell bottleneck every 20 generations (JOSEPH AND HALL 2004a; ZHU *et al.* 2014). By passaging through a single-cell bottleneck the effective population size is kept small ($N_e \approx 11$), which minimizes the effects of selection; only mutations with heterozygous fitness effects (s) of approximately 5% or greater (i.e. $2s \geq 1/11$) will be efficiently acted on by selection. Using RNA sequencing, we analyzed the gene expression of 20 aneuploid and 18 euploid lines across both strains to find differentially expressed genes and to determine if there was evidence for dosage compensation at the whole-chromosome and individual-gene levels in yeast.

Methods

Estimating the spontaneous rate of aneuploid mutation

To determine the rate at which spontaneous aneuploidy occurs in yeast, we analyzed data from two previous mutation accumulation (MA) experiments (Figure 2.1) (JOSEPH AND HALL 2004b). In both, an ancestral strain was copied into multiple MA lines, which were then maintained separately for ~2000 cell generations (G) (2063 generations in the homozygous ancestor lines and 2108 in the heterozygous ancestor lines) via single-cell transfer every 48 hours

(± 1 hour) for 100 transfers. The actual number of generations that passed was estimated by measuring colony size after 48 hours of growth in a representative sample and then determining cell number by counting using a hemocytometer. To confirm that the vast majority of cells present at 48 hours were viable, we also estimated cell number by serial dilution and plating (data not shown).

The two diploid ancestral strains differed in their origin and degree of heterozygosity. One strain was obtained from a mating between NCYC 3631, which is a Matα derivative of YPS 606 (an oak strain from Pennsylvania, USA), and NCYC 3596, a Mat*a* derivative of DBPVG1106 (a wine strain isolated from a lici fruit in Indonesia). This highly heterozygous strain had a heterozygous site every ~250 bp and was homozygous for *ho* and *ura3* mutations (MA experiment and strain production performed by a previous graduate student, Megan Behringer).

The other strain was derived from a standard lab strain (S228C) and carried the following mutations: *ho ade2, lys2-801, his3-ΔD200, leu2-3.112,* and *ura 3-52* (JOSEPH AND HALL 2004a). The strain was obtained by transforming a Mat*a* haploid version of the strain with an *HO URA3* plasmid to generate a diploid version of the strain, followed by counterselection of the plasmid on 5FOA (JOSEPH AND HALL 2004a). This strain was thus homozygous at all loci except the mating type locus.

We used the number of aneuploid chromosomes in the MA lines at the end of the experiment to calculate the rate at which aneuploidy occurs in each of these strains. In brief, if the rate of aneuploidy for chromosome *c* is μ_c, then the probability that a line is not aneuploid for this chromosome is $(1-\mu_c)^G$, where G is the number of generations of MA. Thus if n_c MA lines show aneuploidy for this chromosome, implying that $(n - n_c)$ do not, where n is the total number of MA lines, then we can estimate the rate of aneuploidy per chromosome by solving $(1-\mu_c)^G = (1 - n_c/n)$ for μ_c. Similarly, we can estimate the overall aneuploidy rate, μ, which is the probability that a

cell will become aneuploid for any chromosome in a single cell division, by solving $(1-\mu)^G = (1 - n_a/(16\ n))$ for μ, where n_a is the number of aneuploid chromosomes across all MA lines.

Estimating the effects of aneuploidy on gene expression

To determine the effects of aneuploidy on gene expression, we collected and analyzed RNA sequencing data from a selection of euploid and aneuploid lines from each experiment. For aneuploid samples, we chose all the MA lines that were monosomic for a chromosome (3 lines), those that shared common aneuploidies (21 lines), and those that had more than one aneuploidy (4 lines). From the homozygous ancestor experiment, we selected 10 aneuploid and 12 euploid MA lines. From the heterozygous ancestor experiment, we selected 10 aneuploid and 6 euploid MA lines. Additionally, we collected RNA sequencing data for both ancestral lines, which had been stored at -80ºC since the beginning of the experiment. The homozygous strains were run in two separate RNA sequencing runs, separated by 2 years. In both sequencing runs, we included three replicates of each ancestor. For analysis, we kept these two datasets separate because we found that the ancestor was significantly different across the two sequencing runs. Across both strains and sequencing runs, we obtained RNA sequencing data for 38 strains representing two ancestor strains, 20 euploid lines and 16 aneuploid lines.

For obtaining RNA, for each MA line and ancestor we plated cells that had been stored at -80°C, and allowed growth for two days at 30ºC on YPD plates (1% yeast extract, 2% peptone, 2% glucose, 2% agarose), and then initiated 3ml liquid YPD cultures (no agarose) from three separate colonies (biological replicates) of each line. Liquid cultures were incubated on a rotator at 30°C for 24 hours, before being diluted into 50ml YPD and allowed to grow on a shaker at 30°C for 6 hours. Optical density (OD) measurements were taken to ensure all cultures were in the same log growth phase. Cells were then pelleted, and RNA was extracted from each replicate using the MasterPure Yeast RNA Purification Kit (Epicentre). Integrity, concentration, and

quality of RNA samples were assessed using a Qubit (Thermo Fisher Scientific). Libraries were prepared using the Illumina Stranded RNAseq Kit and were sequenced at the Georgia Genomics Facility on the Illumina NextSeq (75 cycles) single-end 75bp reads High Output flow cell. Samples were multiplexed and split across two sequencing lanes.

Raw reads were processed by the Georgia Genomics Facility to remove sequencing adapters and demultiplex samples. Quality control was performed using FastQC version 1.8.0_20 with default parameters (available at www.bioinformatics.babraham.ac.uk/projects/fastqc/). Low-quality bases were trimmed using Trimgalore version 0.4.4 using -phred 33, -q 20 (available at www.bioinformatics.babraham.ac.uk/projects/trim_galore/). RNA samples were aligned to the *Saccharomyces cerevisiae* reference genome (UCSC version sacCer3, available at support.illumina.com/sequencing/sequencing_software/igenome.html) and transcripts were annotated using Tophat v. 2.1.1 with -i 10 -I 10000 (TRAPNELL *et al.* 2012). Cufflinks v. 2.2.1 was used to assemble sample transcriptomes using default parameters (TRAPNELL *et al.* 2012) and Cuffnorm v. 2.2.1 was used with default parameters to normalize reads. Differential expression was determined using Cuffdiff v. 2.2.1 with default parameters, and read counts were found using HTseq v. 0.6.1pl (Python v. 2.7.8) (ANDERS *et al.* 2015). Finally, we used Samtools v. 1.3.1 to convert *.sam* files into *.bam* files and sort the resulting *.bam* files (LI *et al.* 2009). Scripts can be found at https://github.com/hollygene/Dosage_Compensation/tree/master/bash_scripts.

To compare chromosome-level changes in gene expression across strains Cuffnorm v. 2.2.1 (TRAPNELL *et al.* 2012) was used to calculate FPKM (fragments per kilobase per million reads) for each RNAseq data set. A custom bash script was then generated to join the FPKM values for each strain with the gene annotations file, convert the resulting file into a *.csv* formatted file, remove mitochondrial sequences (as we were not interested in gene expression changes in mitochondrial DNA), and change the chromosome names from Roman numerals to numbers (script can be found at

https://github.com/hollygene/Dosage_Compensation/blob/master/bash_scripts/DC_workflow_April2017.sh). For each gene, the average FPKM across the three replicates for each strain was calculated, followed by the average FPKM ratio (average FPKM in an MA line divided by the average FPKM in the ancestor). We noticed that the FPKM ratio was highly variable across MA line replicates for genes with an average FPKM < 5 across all euploid strains (ancestor + euploid MA lines), so we removed such genes, leaving a total of 6181 genes (Supplemental Figure 2.1, Appendix I). We also removed rRNA genes and tRNA genes, as these are challenging to map accurately and, because of their propensity to show extreme variation in copy number, can cause issues with data normalization.

To determine whether there was evidence for dosage compensation at the whole-chromosome level, we compared the average FPKM ratio for genes on an aneuploid chromosome to the expectation from gene dose. Thus, a trisomic chromosome would be expected to show a 1.5-fold increase in gene expression and an average FPKM ratio = 1.5 (log2ratio = 0.585). Similarly, monosomic and tetrasomic chromosomes should show average FPKM ratios of 0.5 (log2ratio = -1) and 2 (log2ratio = 1), respectively. We asked whether the observed distribution was consistent with the expected FPKM ratio by calculating the mean and confidence interval of the average FPKM ratio (a one-sample t-test). All analyses were done in RStudio (TEAM 2013). R scripts are available at https://github.com/hollygene/Dosage_Compensation/tree/master/R/scripts/R_scripts.

As we did for chromosome-level gene expression analysis, previous studies have almost exclusively used FPKM to measure gene expression to compare across strains or treatments. However, the use of FPKMs has been criticized because of loss of power due to relatively few replicates (three in our experiments). A possible loss of statistical power for individual genes is not an issue when comparing tens or hundreds of genes for each chromosome as we did in the chromosome-wide gene expression analysis above, and so there we used FPKMs to make our data more comparable to previous data. However, when examining individual genes, power

becomes a more serious concern. As such, we used a method, *DESeq2* (LOVE *et al.* 2014), that models the expression level for a gene in a particular replicate and treatment (in our case ancestor versus MA line). Importantly, the method also models the dispersion of the read depth, assuming the distribution of the read depth can be accurately represented by a negative binomial with genes of similar expression having similar dispersion. This method is thus expected to more accurately predict the actual read depth by explicitly considering the variance in the read depth across replicates. Thus, as a result, an unusually high or low depth for one replicate will not have equal weight compared to the depths for the other replicates. The method should thus be able to better detect genes that are differentially expressed (DE) in an MA line versus its ancestor.

Raw read counts obtained from htseq-count were used as input for DESeq2 (LOVE *et al.* 2014). Individual *DESeqDataSets* were produced for each strain, due to the high variation found across strains, as determined by principal component analysis (PCA) (Supplemental Figure 2.2, Appendix I). Genes expressed at low levels tend to have high variance and there is thus low power to detect changes in expression. Reads with counts less than 10 in every replicate were removed from further analysis. We used a more stringent cutoff in this analysis to focus on genes for which we have the most power for detecting a change in expression, since we are analyzing individual genes. Removing such genes from the data set resulted in 5532 genes being analyzed (Supplemental Figure 2.1, Appendix I).

The *DESeq()* function was implemented on all datasets with default parameters. Annotations were added using the *S. cerevisiae* database from Bioconductor (CARLSON M 2015). The *results()* function in *DESeq2* was implemented with default parameters, using a False Discovery Rate (FDR) of 0.1. Analyses were performed with one MA line and the ancestor at a time, since running all strains together would lead to an overestimate of dispersion because of the numerous aneuploid chromosomes in MA lines (see above). For one of the two batches of the homozygous strain, one of the ancestor replicates was substantially different based on a PCA, and so only 2 of the 3 ancestor replicates were used (Supplemental Figure 2.2, Appendix I). Similar to

the whole chromosome analysis, ratio distributions equal to the sample mean divided by the ancestral mean for the normalized counts were obtained from *DESeq2* estimated read counts. To visualize the data, histograms for both cis (present on aneuploid chromosome) and trans (present on remainder of chromosomes) genes were generated using ggplot2 in R (WICKHAM 2016).

In addition to looking at all genes in the genome to identify those that were differentially expressed, we also specifically concentrated on a few classes of genes that have been identified in previous work as either being dosage sensitive (DS) (115 genes, MAKANAE *et al.* 2013), or particularly likely to alter expression in response to stress. These latter genes include those in the environmental stress response (ESR) pathway (139 genes, GASCH *et al.* 2000) and those thought to play a role in aneuploidy stress response (ASR) (201 genes, TORRES *et al.* 2007). ASR genes were previously shown to be significantly differentially expressed in aneuploid but not euploid strains. To identify significant DE for genes from these categories, we tested each gene's expression against the expected expression for a disomic gene, and determined which genes were significantly different, then parsed those genes into what matches the ESR/DS/ASR genes. We then counted how many times each gene appeared as a measure of its degree of consistent DE across aneuploid MA lines.

Results

The rate of spontaneous aneuploidy is nearly twice as high in the heterozygous strain as the homozygous strain

The number of aneuploidy events by chromosome is shown in Table 2.1. We assume that aneuploidy is caused by mitotic nondisjunction since cells are kept asexual. A single non-disjunction event can produce both a monosomic and a trisomic chromosome in a diploid strain. Thus, two events would be required to obtain the one tetrasomic MA line. The total number of events in the homozygous ancestor strain varied between 0 and 5 per chromosome, which implies a maximal observed rate of nondisjunction for a single chromosome of 1.70×10^{-5} events/division

(obtained by solving $(1-\mu_c)^{2063} = 140/145$ for μ_c), and a minimum of zero. The observed rate of an event for any chromosome (i.e. the genome-wide rate) is 6.73×10^{-6} events/division (obtained by solving $(1-\mu)^{2063} = 1 - 32/(16*145)$ for μ_c). The total number of events in the heterozygous ancestor varied between 0 and 7 per chromosome, which implies a maximal observed rate of nondisjunction for a single chromosome of 1.56×10^{-4} events/division (obtained by solving $(1-\mu_c)^{2108} = 69/76$ for μ_c), and a minimum of zero. The observed rate of an event for any chromosome is 1.51×10^{-5} events/division (obtained by solving $(1-\mu)^{2108} = 1 - 38/(16*76)$ for μ_c), which is over two-fold higher than the homozygous strain. Examination of the number of euploid versus aneuploid lines indicates that this is a highly significant difference (Table 2.2, Fisher's Exact test, $p < 0.0001$).

We note that there were two monosomies and 30 trisomies, a 15-fold difference, in the homozygous experiment and one monosomy and 35 trisomy events in the heterozygous experiment. Since a single nondisjunction event creates both types of aneuploids in the daughter cells, this imbalance implies that monosomies are under-represented in the MA experiments. This finding suggests that monosomies have effects on fitness that are large enough to be seen by selection, even in the low-selection MA framework. Thus, the actual rate of aneuploidy might perhaps be better estimated as twice the trisomy event rate, giving 1.23×10^{-5} and 2.77×10^{-5} events per cell division for the homozygous and heterozygous ancestor strains respectively.

In addition, two chromosomes, 6 and 13, comprise 0 out of 70 observed aneuploidy events across the two experiments. If events occurred at random, each chromosome should have 1/16 of the observed events, or 4.4 each. Under a Poisson distribution, the probability of having a chromosome with no events when the expected number is 4.4 equals $e^{-4.4} = 0.013$. It thus seems clear that aneuploidy of chromosomes 6 and 13 either cause strongly deleterious fitness effects or are not tolerated (i.e. are lethal). These aneuploidies have been seen in aneuploid clinical yeast samples (ZHU *et al.* 2016), suggesting that differences in genetic background may alter the degree to which aneuploidy is deleterious.

To address whether one aneuploidy event increases the probability of another, we asked whether there was an excess of strains carrying two or more aneuploidies. For the homozygous strain, 28 of the 145 MA lines were found to be aneuploid. Of these, four lines contained two aneuploidies (i.e. two separate chromosomes had become aneuploid), which is not significantly different from the Poisson expectation of 3 (Fisher's Exact Test, $p > 0.99$). For the heterozygous strain, 29 out of 76 sequenced MA lines were found to be aneuploid. Of these, seven lines contained two aneuploidies, which is the same as the Poisson expectation.

To determine whether chromosome size effects the number of nondisjunction events we captured in our MA experiments, we plotted size versus number of nondisjunction events (Figure 2.2). While there is clearly variation in which chromosomes become aneuploid, there was no significant relationship with size. However, only two chromosomes, 1 and 9, were found to be monosomic and both of these are relatively small chromosomes (1 is the smallest at 230,218 bp, and 9 is the 4th smallest at 439,888 bp). This suggests that while there is no noticeable effect of chromosome length on aneuploidy occurrence, monosomy may be tolerated in smaller chromosomes better than larger chromosomes.

Little evidence for whole-chromosome dosage compensation in either strain

We performed RNAseq on 10 euploid and 12 aneuploid homozygous ancestor strain MA lines, and on 6 euploid and 10 aneuploid heterozygous ancestor strain MA lines. Whole-chromosome gene expression was analyzed by calculating the average and 95% confidence intervals of gene expression for each chromosome (Figure 2.3). ANOVAs were also run on each aneuploid sample, comparing the average gene expression from each chromosome to that of the other samples (lm(y~Line), where y is FPKM ratio and Line is the line number). If there were complete dosage compensation occurring on the whole-chromosome level, we would expect no difference between aneuploid and euploid chromosomes, such that ANOVAs would show no effect of chromosome number on gene expression. However, in the absence of dosage

compensation, chromosome number would have an effect, with aneuploid chromosomes underlying the significant difference among chromosomes. Further, in the absence of dosage compensation, we would expect gene expression to mirror gene dose such that aneuploid chromosomes would show 0.5 or 1.5-fold increases in expression for monosomic and trisomic chromosomes respectively.

ANOVAs indicated that the effect of chromosome was significant ($p < 0.01$) in every aneuploid MA line, as expected with no dosage compensation. For chromosomes that did not have any aneuploid lines represented in the dataset, we still found some differential expression in a few aneuploid lines. Specifically, for chromosome III, Line 76, 61, 59, 49, 18, 11 (heterozygous ancestor) are significantly different ($p<0.01$), suggesting that aneuploidy causes changes in gene expression of genes on chromosome III across aneuploid strains. However, ANOVAs on some euploid lines, also displayed significant p values for certain chromosomes, indicating that some chromosomes show changes in expression even in the absence of changes in dose. This could suggest an impact of the MA framework on gene expression in yeast.

If there is no chromosome-level dosage compensation, then the level of gene expression is expected to be proportional to chromosome copy number. For most aneuploid chromosomes in MA lines this prediction held: expression levels did not differ significantly from the expectation. However, in 4 MA lines (line numbers 18, 49, 59 and 61) from the heterozygous ancestor, the expected level of gene expression was less extreme than expected based on chromosome copy number (Figure 2.3). Chromosome 1 of line 18 had average expression change equal to 1.3-fold, chromosome 5 of line 49 had average expression change equal to 1.35-fold, chromosome 7 of line 59 had average expression change equal to 1.25-fold, and chromosome 7 of line 61 had average expression change equal to 1.39-fold. All these values were significantly different from the expected expression level of 1.5-fold ($p < 0.05$). The vast majority of gene expression changes, 65 of 69, are consistent with a lack of whole-chromosome dosage compensation

occurring in either strain, and together these findings support previous work showing no dosage compensation in aneuploid yeast (TORRES *et al.* 2010).

Distribution of gene expression from euploid versus aneuploid chromosomes

The previous analysis indicates that mean gene expression of aneuploid chromosomes seems to be predicted by gene dose. We next examined whether the mean expression for genes on the non-aneuploid (disomic) chromosomes is altered by aneuploidy. In addition, we examined whether the variance in gene expression for aneuploid chromosomes is the same as for euploid chromosomes in the 20 aneuploid MA lines, and whether the variance in gene expression differs between euploid MA lines and their euploid ancestor. The distribution of FPKM ratios (MA line FPKM / ancestor FPKM) for all genes in euploid samples (Figure 2.4), for genes on the aneuploid chromosome(s) in aneuploid samples (cis genes), and for genes not located on the aneuploid chromosome(s) in aneuploid samples (trans genes) were analyzed (Figures 2.5 and 2.6; Supplemental Figure 2.3, Appendix I).

The expected mean expression ratio in euploid lines is 1. In every euploid line analyzed, the expected distribution had a mean that was indistinguishable from 1 (pval > 0.1, Supplemental Figure 2.3, Appendix I). For aneuploid lines, the expected mean expression for trans genes (those not located on the aneuploid chromosome) is not equal to 1. This is because the aneuploid chromosome will have more (for trisomy) or fewer (for monosomy) reads mapping to it because the chromosome represents a different percentage of the genome in an aneuploid line. In Table 2.3, we indicate the expected mean expression level for trans genes in MA lines carrying a single trisomy. Similarly, for lines with monosomies, the expected mean expression of trans genes is higher. We tested the mean expression of trans genes against the expectation based on the chromosomes for which they were aneuploid and found that in no case were they significantly different (Supplemental Figure 2.3, Appendix I).

To examine whether the variance in gene expression is greater in aneuploid lines, we compared the variance in gene expression of both cis and trans genes to the variance of those same genes in a euploid line using a Levene's test since the distributions were heavily skewed. For comparisons, we randomly matched a euploid line with each aneuploid line. We determined whether the means and variances of these distributions differed from the expectation (the expectation being that both the means and the variances are equal between aneuploid and euploid lines). The variances of gene expression from cis genes were significantly different from the expectation in every case except for two – the comparison of homozygous line 15 (trisomic for chromosome 9) to homozygous line 5 (euploid) and the comparison of homozygous line 152 (trisomic for chromosomes 1 and 7) to homozygous line 1 (euploid) (Table 2.4). There is nothing immediately notable with these samples, though the ANOVA for chromosome 7 line 1 was significant ($p<0.05$), however there was no similar connection in line 5 for chromosome 9 (Supplemental Data; Appendix I). Further, the variances of trans genes in the aneuploid and euploid lines were significantly different from each other, which could be due to more reads going towards the cis (for trisomies) or trans (for monosomies) genes in the aneuploid lines as compared to the euploid lines, as discussed above.

Individual Dosage-Compensated Genes

Our analyses indicated that at the whole-chromosome level aneuploidy leads to changes in gene expression predicted by gene dose, such that there was no evidence for dosage compensation, and minor (or no) effects on expression of the rest of genome. Next, we investigated individual genes. We sought to group genes present on aneuploid chromosomes into five categories based on their gene expression, using similar metrics as a previous study (MALONE *et al.* 2012): 1. Not dosage compensated: these genes have expression levels not significantly different as those predicted by their gene dose. 2. Partially dosage compensated: these genes show less extreme gene expression changes than predicted by their dose. 3. Fully

dosage compensated: these genes show no change in expression in response to changes in gene dose. 4. Over-dosage compensated: these genes show changes in expression that are in the opposite direction of the change in gene dose. 5. Anti-dosage compensated genes show more extreme changes in expression than predicted by the change in gene dose (in the direction of the aneuploidy – i.e. monosomic genes would have lower gene expression than predicted by monosomy) (Table 2.5). Any gene that had expression levels different from the ancestor and different from the expectation based on gene dose was assigned to one of the categories depending on their level of expression. For the aneuploid strains we analyzed, we found several genes in each of these categories (Table 2.6). Since we are testing many genes (5587), power becomes limited due to the need to correct for multiple testing. For this reason, it is important to test for expression that is consistent both with respect to the ancestor and to the expectation based on gene dose. Many genes do not differ from either, in which case we cannot conclude the degree to which they are compensated – these genes were assigned as category 0 genes, or "unknown" compensation. Our analyses revealed that the power to distinguish whether a gene exhibits dosage compensation or not is low; the vast majority of genes are in category 0 (Table 2.6). For those genes in category 4 and 5, we find that there is little agreement between different strains in terms of the percentage of genes in these categories (Figure 2.6).

We compared the trans genes of aneuploid samples with those of samples with a different aneuploid chromosome(s) to determine if there was a common response to aneuploidy, as has been shown in previous studies (GASCH *et al.* 2000; ZILLIKENS *et al.* 2017a). We found that in lines from the heterozygous ancestor, at most, 8/10 aneuploid samples shared 15 DE trans genes (genes that were not located on an aneuploid chromosome) (Figure 2.9). In lines from the homozygous ancestor, at most 6 aneuploid lines shared 8 DE trans genes (Figure 2.14).

We then examined if euploid lines shared a common gene expression response and found that in lines from the homozygous ancestor, at most 5 euploid samples shared 8 common differentially expressed genes (Figure 2.15). In the heterozygous ancestor, at most 5 lines shared

54 DE genes (Figure 2.16). This result suggests a shared effect of the mutation accumulation experimental design on gene expression, particularly in the heterozygous ancestor lines.

Histone Genes

Histone genes H2A and H2B are known to possess a mechanism of dosage compensation in *S. cerevisiae* (OSLEY AND HEREFORD 1981; MEDICI *et al.* 2014). Our analyses did not include samples with aneuploidies on these chromosomes (II and IV), but we do have aneuploid samples for chromosomes containing other histone genes: XIV, XV, and XVI (containing histones 3,4, and linker, respectively). Six lines across both experiments are trisomic for chromosome XIV, 1 line is trisomic for chromosome XV, 13 lines are trisomic for XVI, and 1 line is tetrasomic for XVI. Previous studies have found that these genes do not display dosage compensation and we did also not find evidence for compensation (PETER R. ERIKSSON 2012) (Supplemental Table 2.1, Appendix I).

Stress Response Genes

Yeast are known to undergo what is known as the environmental stress response (GASCH *et al.* 2000; ZILLIKENS *et al.* 2017a), when conditions are unfavorable due to various factors, including temperature stress, oxidative stress, and nutrient limitation. We analyzed genes previously found to relate to the environmental stress response and found that our aneuploid samples did differentially express most of these genes (Figures 2.16-2.19), though there was no significant trend of a shared differential expression response of ESR genes between samples.

It has been found that similarly, aneuploid yeast undergo what is referred to as the "aneuploid stress response (ASR)," in which certain trans genes are differentially expressed (TORRES *et al.* 2010). A majority of these genes are also differentially expressed during the environmental stress response. To determine if we found the same pattern of differential expression in our spontaneously aneuploid samples, we investigated these ASR genes (201 genes

total) and found that in samples from the heterozygous ancestor, at most 7 lines shared 3 DE ASR genes. In the homozygous ancestor aneuploid lines, at most only 4 lines shared just 1 DE ASR gene (Figures 2.20 & 2.21). As expected, the euploid lines in both datasets did not show significant signatures of differential expression on ASR genes and as expected, did not share many DE ASR genes (Figures 2.22 & 2.23).

Dosage-Sensitive Genes

Previous studies have found that certain genes are more sensitive to changes in gene dose than others. Using the "genetic tug-of-war" method, Makanae et al 2013 found the copy-number limits of overexpression in all 5806 protein-coding genes in *S. cerevisiae*, and found 115 genes whose copy number limits were 10 or less (more than this amount caused cell death) (MAKANAE *et al.* 2013). Curious as to whether our samples exhibited a compensatory response for these dosage sensitive genes, we looked at the same set of genes and parsed out those that were significantly differentially expressed in our aneuploid samples. Most aneuploid samples had few differentially expressed dosage sensitive genes Figure 2.24 – 2.25). The euploid lines in both experiments had very few DE dosage-sensitive genes, consistent with the expectation that there would be zero (Figures 2.26 – 2.27).

Of particular interest were the genes on the aneuploid chromosomes, as they differ in copy number compared to the rest of the genes in the genome. Most samples showed a high level of compensation of dosage-sensitive genes on the aneuploid chromosome and elsewhere in the genome. However, samples with a trisomy for chromosome 9 appeared to be more tolerant of the duplication (likely due to individual gene compensation) than other chromosomes – samples ranged from 0 to 33% compensation (Table 2.6).

Discussion

Rate of aneuploidy

We calculated the rate of aneuploidy based on data from two previous yeast mutation accumulation experiments: one with a heterozygous strain and one with a homozygous strain (Figure 2.1). We found that the rate of aneuploidy is higher in the heterozygous strain than the homozygous strain ($p<0.0001$, Fisher's exact test, Table 2.2). The heterozygous ancestor strain MA lines had a total of 29 aneuploids and 47 euploids, whereas the homozygous ancestor MA lines had a total of 28 aneuploid and 117 euploid lines. Previous studies have found that hybrids of two yeast species have been shown to systematically lose all or part of one parent's genome (MARINONI *et al.* 1999). It is possible that the mating of distantly related *S. cerevisiae* strains to produce the heterozygous strain showed a milder version of genome incompatibility as exemplified by the higher rate of aneuploidy compared to the homozygous lab strain. However, the heterozygous strain did not show any growth defects (which could have indicated a phenotypic effect of genome incompatibility) compared to the homozygous strain. Further, since our homozygous strain was not isogenic with either of the heterozygous strain parents, it is possible that our findings are instead the result of differences in the genetic background of the strains. To examine the effect of heterozygosity per se in future experiments, diploids could be generated from each of the parent strains used to make the heterozygous strain and then used in mutation accumulation experiments to determine the rate at which aneuploidies arise.

In our experiment, we found 3 and 6 events for the homozygous and heterozygous strains involving chromosome V nondisjunction, implying a rate of 9.67×10^{-6} and 3.90×10^{-5} events per cell division, respectively. Previous studies have found that chromosome V is lost spontaneously by nondisjunction in *S. cerevisiae* at a rate of $2\text{-}8 \times 10^{-6}$ cell generations (MULLA *et al.* 2014). This previous study used a laboratory strain (A364A), which was highly homozygous – this likely explains the discrepancy in rates between the previous study and our heterozygous ancestor strain

rate and is not significantly different than the rate of aneuploidy of chromosome V found in both strains in our study.

We found a difference in aneuploidy rates at the individual chromosome level as well as overall. In the heterozygous ancestor strain MA lines, we found 10 (out of 29 total aneuploidies) trisomies of chromosome XVI, compared with 3 (out of 29 total aneuploidies) in the homozygous ancestor MA lines (Table 2.1). Previous studies have found a similar discrepancy between diploid and diploid-hybrid strains of yeast, with the hybrid strains showing a higher rate of aneuploidy at chromosome XVI (KUMARAN *et al.* 2013). These results suggest that heterozygosity influences either nondisjunction rate or tolerance of certain aneuploidies and that certain chromosomes are either more likely to become aneuploid or are better tolerated after becoming aneuploid, or both.

Due to the diploid nature of our initial MA ancestors, we were able to analyze trisomics, monosomics, and a tetrasomic to study the rate and effects of whole-chromosome aneuploidy. Contrary to most previous studies, we were able to observe the spontaneous rate and effects of monosomy, which is substantially less common than trisomy in our samples (Table 2.1). Considering nondisjunction events result in the production of both a trisomy and a monosomy, we would expect to see an equal number of each in our data. The lack of monosomies implies that there must be strong selection against them, implying that fewer copies of a chromosome is substantially more deleterious than additional copies. On explanation could be that a monosomy has a larger effect on gene expression, a two-fold difference, compared to trisomy, which results in a 1.5-fold difference.

No evidence for whole-chromosome dosage compensation at the transcript level

Our results suggest that there is no general mechanism for dosage-compensation in aneuploid yeast, either at the whole-chromosome or individual gene level (Figure 2.4, Table 2.6). Previous studies have found that the increase in a partner gene can rescue the sensitivity of a strain to another increased dosage. This may be occurring in the samples that had little to no

compensation of the dosage sensitive genes on the aneuploid chromosome. This mirrors previous findings that RNA level scales with DNA copy number (TORRES *et al.* 2010). It has also been found that aneuploid yeast samples utilized posttranscriptional methods of lowering protein levels and that no RNA-level compensation was occurring (TORRES *et al.* 2010). One explanation for previous studies that have found whole-chromosome dosage compensation effects is that they were using heterogeneous samples of yeast that were both aneuploid and euploid (JAMES HOSE 2015; AUDREY P GASCH 2016), causing gene expression ratios to be intermediate between what is expected for aneuploid and euploid DNA copy levels. Similarly, the apparent partial compensation we observed for some MA strains in our study may also be caused by heterogenous samples. To avoid this potential problem, future studies could employ the use of fluorescent activated cell sorting (FACS) to separate the aneuploid cells from the euploid cells and use only the aneuploid culture for RNA extraction. However, our evidence suggests that certain genes on the aneuploid chromosomes are partially compensated. It is possible that genes that are more deleterious in high numbers but are on chromosomes that also contain genes that are beneficial in high copy number are up- or down-regulated on a gene-by-gene basis in order to deal with the extra or missing chromosome, implying a robust stress response in relatively short evolutionary time.

Aneuploidy effects on trans genes

Previous studies have proposed that there is an effect of aneuploidy on the remainder of the genome, by looking at the peaks of the distributions and claiming that the apparent skew to the left of 1.00 indicated that the aneuploid chromosome was causing other expression effects in the genome (HOU *et al.* 2018). We investigated trans genes in our data and found that they showed the expected level of gene expression (Figure 2.7, Table 2.3), implying that aneuploidy does not cause a global change in gene expression. We also determined whether aneuploid lines shared any differentially expressed genes not located on aneuploid chromosomes. We compared

gene expression data between aneuploid samples and found in our heterozygous ancestor 15 commonly differentially expressed genes among 8 of our aneuploid lines. Similarly, in the homozygous ancestor lines, we found 8 commonly differentially expressed trans genes among 6 of the aneuploid lines.

Previous studies in yeast have found evidence of a transcriptional response to environmental stress as well as a transcriptional response to aneuploidy involving the "environmental stress response" genes and the "aneuploidy stress response" genes respectively (GASCH *et al.* 2000; TORRES *et al.* 2007; ZILLIKENS *et al.* 2017b). We investigated the environmental stress response (ESR) genes and found that most ESR genes were differentially expressed in our aneuploid samples, but not in euploid samples, suggesting that the state of aneuploidy has similar effects on the transcriptome to various environmental stresses including high salinity, high temperatures, and highly oxidative-species rich environments. It would be interesting to know if the yeast samples exposed to these environmental stresses had any copy number changes in their genomes – this would further add evidence to the hypothesis that aneuploidy is an adaptive state to changes in the environment and/or a consequence of stress. However, our aneuploid strains do not have many differentially expressed aneuploidy stress response genes, suggesting that each aneuploidy confers a different stress and therefore a different transcriptional stress response (Figures 2.20 and 2.21).

Conclusions and future directions

This study demonstrated that heterozygosity is correlated with a higher aneuploidy rate, that there is no evidence for whole-chromosome dosage compensation in aneuploid yeast, and that aneuploid chromosomes do not significantly influence the gene expression patterns among the rest of the genome. We did find evidence for compensation at the individual gene level for genes that are particularly toxic in high copy numbers, suggesting that cells are able to employ transcriptional compensatory mechanisms to tolerate aneuploidy at least at the individual gene

level. Further, our analyses demonstrated evidence for individual aneuploid lines to differentially express environmental and aneuploidy stress response genes. There were not many shared differentially expressed ESR/ASR genes among aneuploid lines, however, implying that each aneuploid line deals with its aneuploidy in a unique manner.

Our finding of no global effects of aneuploidy on gene expression is in direct opposition to a recent paper claiming this – however, we showed that the apparent skew of trans genes is actually due to sequencing bias from reads mapping to more (or less) copies of the aneuploid chromosome(s). Our analyses and evidence bring insights into the effects of aneuploidy on gene expression in budding yeast and can be applied to other species as well. Further, our findings provide insight into the evolution of sex chromosomes and dosage compensation – it would appear that evolution of a dosage compensation mechanism takes a longer evolutionary time than we allowed during our experiment, and stronger selection may also be required.

More insights into how wild yeast tolerate aneuploidy are required. A recent study found that the SSD1 gene in yeast is linked with aneuploidy tolerance in wild strains versus lab strains (HOSE *et al.* 2020). This gene is a translational repressor and is functional in wild yeast isolates but not in laboratory strains. This implies that wild aneuploid yeast strains can tolerate aneuploidy by attenuating translation of the duplicated genes. This reflected previous work in aneuploid yeast that showed compensation at the protein, but not RNA, level (NOAH DEPHOURE 2014). Our analyses provide further evidence for the lack of transcript-level dosage compensation, and future studies could use SSD1 knockout strains of yeast for mutation accumulation studies and determine rates and tolerance of aneuploidy in a similar manner as this study.

Tables

Table 2.1. The number of monosomies, disomies and tetrasomies seen for each MA experiment. The homozygous MA experiment had 32 events in 145 MA lines maintained for 2063 generations. The heterozygous MA experiment had 38 events in 76 MA lines maintained for 2103 generations.

	# monosomic		**# trisomic**		**# tetrasomic**		**total events**	
Chrom. #	homo	hetero	homo	hetero	homo	hetero	homo	hetero
1	0	1	1	3	0	0	1	4
2	0	0	3	0	0	0	3	0
3	0	0	2	0	0	0	2	0
4	0	0	3	0	0	0	3	0
5	0	0	3	4	0	0	3	4
6	0	0	0	0	0	0	0	0
7	0	0	1	5	0	0	1	5
8	0	0	4	0	0	0	4	0
9	2	0	3	2	0	0	5	2
10	0	0	1	1	0	0	1	1
11	0	0	1	0	0	0	1	0
12	0	0	1	7	0	0	1	7
13	0	0	0	0	0	0	0	0
14	0	0	4	2	0	0	4	2
15	0	0	0	1	0	0	0	1
16	0	0	3	10	0	1	3	11

Table 2.2. **Number of aneuploid versus euploid MA lines in each experiment.** The heterozygous ancestor strain gave rise to a substantially higher proportion of aneuploid MA lines, indicating a significantly higher mutation rate.

	Heterozygous ancestor	Homozygous ancestor
Aneuploid MA lines	29	28
Euploid MA lines	47	117

*p<0.0001 Fisher's exact test

Table 2.3. **Expected mean of gene expression in MA lines trisomic for single chromosome shown**. Because more reads map to trisomic chromosome, number mapping to rest of genome will decline slightly.

Trisomic Chromosome	Expected expression for trans genes
1	0.98
2	0.94
3	0.97
4	0.89
5	0.95
6	0.98
7	0.92
8	0.96
9	0.96
10	0.94
11	0.95
12	0.92
13	0.93
14	0.94
15	0.92
16	0.93

Table 2.4: Variance comparison between aneuploid and euploid chromosomes in aneuploid versus euploid MA lines. Note: MA line Het-76 carries a partial duplication of chromosome 14. Since it is a partial duplication, we did not examine the variance in expression for that chromosome.

MA line	Aneuploidy present*	Euploid line comparison	Aneuploid chromosomes Levene's Statistic p-value	Euploid chromosomes Levene's Statistic p-value
Hom-152	1, 7	Hom-1	Chr 1: 0.21 Chr 7: 0.58	0.49
Hom-117	5	Hom-2	3.9E-11	1.2E-12
Hom-123	5	Hom-3	3.5E-15	6.5E-10
Hom-108	7, 9^{m}	Hom-4	Chr 7: 6.0E-14 Chr 9: 4.2E-3	1.1E-55
Hom-15	9	Hom-5	0.15	2.5E-4
Hom-29	9^{m}	Hom-6	0.82	2.1E-9
Hom-88	9	Hom-7	4.4E-4	4.6E-10
Hom-119	9	Hom-8	2.1E-1	0.19
Hom-9	14	Hom-11	0.0041	7.8E-5
Hom-112	16	Hom-28	1.6E-5	0.98
Het-7	1	Het-1	0.16	6.8E-17
Het-11	1^{m}, 15	Het-2	Chr 1: 0.16 Chr 15: 0.11	2.2E-8
Het-18	1	Het-3	1.7E-4	0.02
Het-4	5	Het-5	2.3E-4	1.5E-9
Het-49	5	Het-9	1.4E-6	0.078
Het-59	7	Het-69	3.1E-30	3.8E-9
Het-61	7	Het-1	0.023	0.0024
Het-76	9, 10^{p}, 14	Het-2	Chr 9: 0.899 Chr 14: 0.0088	4.55E-12
Het-77	12	Het-3	5.6E-13	0.0032
Het-8	16^{tet}	Het-5	0.12	1.1E-16

* ^{m}monosomic. ^{p}partial duplication, ^{tet}tetrasomic. All others trisomic.

Table 2.5: Expected gene expression categories. 1. Not dosage compensated: these genes have expression levels not significantly different as those predicted by their gene dose. 2. Partially dosage compensated: these genes show less extreme gene expression changes than predicted by their dose. 3. Fully dosage compensated: these genes show no change in expression in response to changes in gene dose. 4. Over-dosage compensated: these genes show changes in expression that are in the opposite direction of the change in gene dose. 5. Anti-dosage compensated genes show more extreme changes in expression than predicted by the change in gene dose (in the direction of the aneuploidy – i.e. monosomic genes would have lower gene expression than predicted by monosomy).

Table 2.5. The expected expression of a gene for each expression category following trisomy (2→3 copies) or monosomy (2→1copy).

	Expression Category				
Copy number	1	2	3	4	5
2 → 3	1.5x	1-1.5x	1x	<1x	> 1.5x
2 → 1	0.5x	0.5-1x	1x	>1x	<0.5x

Table 2.6. The number of genes in each expression change category across the aneuploid strains for which we have RNAseq data. 0 = unknown, 1 = no dosage compensation (DC) , 2 = partial DC, 3 = full DC , 4 = over-compensation, 5 = anti-compensation.

		Category					
MA line	Aneuploidy	0	1	2	3	4	5
Hom-152	1, 7						
	Chr 1:	52	22	6	0	0	1
	Chr 7:	482	9	12	0	0	0
Hom-117	5	147	107	5	0	0	0
Hom-123	5	209	50	5	0	0	1
Hom-108	7, 9^{m}						
	Chr 7:	212	76	129	0	39	42
	Chr 9:	41	8	64	0	84	2
Hom-15	9	62	82	52	0	0	3
Hom-29	9^{m}	90	1	107	0	1	0
Hom-88	9	58	78	50	0	0	13
Hom-119	9	70	92	37	0	0	1
Hom-9	14	92	140	128	0	0	13
Hom-112	16	244	20	0	0	0	0
Het-7	1	51	29	1	0	0	0
Het-11	1^{m}, 15						
	Chr 1:	67	0	10	0	4	0
	Chr 15:	413	57	5	0	0	0
Het-18	1	45	29	6	0	1	0
Het-4	5	165	71	5	0	3	21
Het-49	5	135	89	10	0	0	0
Het-59	7	288	167	23	0	2	13
Het-61	7	160	193	20	0	2	18
Het-76	9, 10, 14^{p}						
	Chr 9:	76	77	43	0	0	3
	Chr 10:	250	62	20	0	0	5
	Chr 14:	165	132	81	0	0	2
Het-77	12	145	184	125	0	0	11
Het-8	16	117	126	31	0	16	156

Table 2.7. Example ANOVA table for chromosome I for heterozygous strain samples. Line 7, 11, and 18 are aneuploid for chromosome I (trisomic, monosomic, and trisomic, respectively), the other lines are disomic.

lm(formula = y ~ Line, data = chr1DataGC)

Residuals: Min: -3.9123 1Q: -0.2504 Median: -0.0161 3Q0.1929 Max: 6.5904

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.048111	0.060519	0.795	0.427
Line2	0.104434	0.085587	1.220	0.223
Line3	-0.067564	0.085587	-0.789	0.430
Line4	0.023977	0.085587	0.280	0.779
Line5	0.062101	0.085587	0.726	0.468
Line7	0.498164	0.085587	5.821	7.05e-09 ***
Line8	0.088339	0.085587	1.032	0.302
Line9	0.055628	0.085587	0.650	0.516
Line11	-0.405439	0.085587	-4.737	2.36e-06 ***
Line18	0.384978	0.085587	4.498	7.34e-06 ***
Line49	-0.023784	0.085587	-0.278	0.781
Line59	-0.032118	0.085587	-0.375	0.708
Line61	-0.045047	0.085587	-0.526	0.599
Line69	-0.002338	0.085587	-0.027	0.978
Line76	-0.066835	0.085587	-0.781	0.435
Line77	0.043517	0.085587	0.508	0.611

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6142 on 1632 degrees of freedom

Multiple R-squared: 0.08813, Adjusted R-squared: 0.07975

F-statistic: 10.51 on 15 and 1632 DF, p-value: < 2.2e-16

Figures

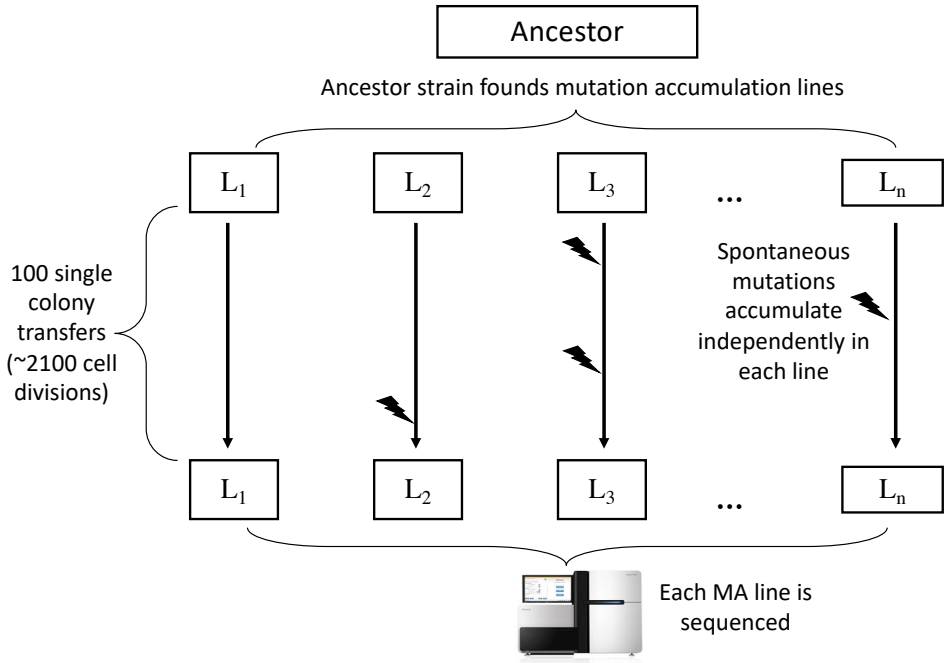


Figure 2.1. The mutation accumulation framework. An ancestor strain is used to found multiple MA lines that are then passaged by single cell transfer. Mutations with reasonably small fitness effects arise and fix at a rate that is independent of selection. The heterozygous strain MA experiment was performed by Megan Behringer, and the homozygous MA experiment was performed by Sarah Joseph and Dave Hall (JOSEPH AND HALL 2004a).

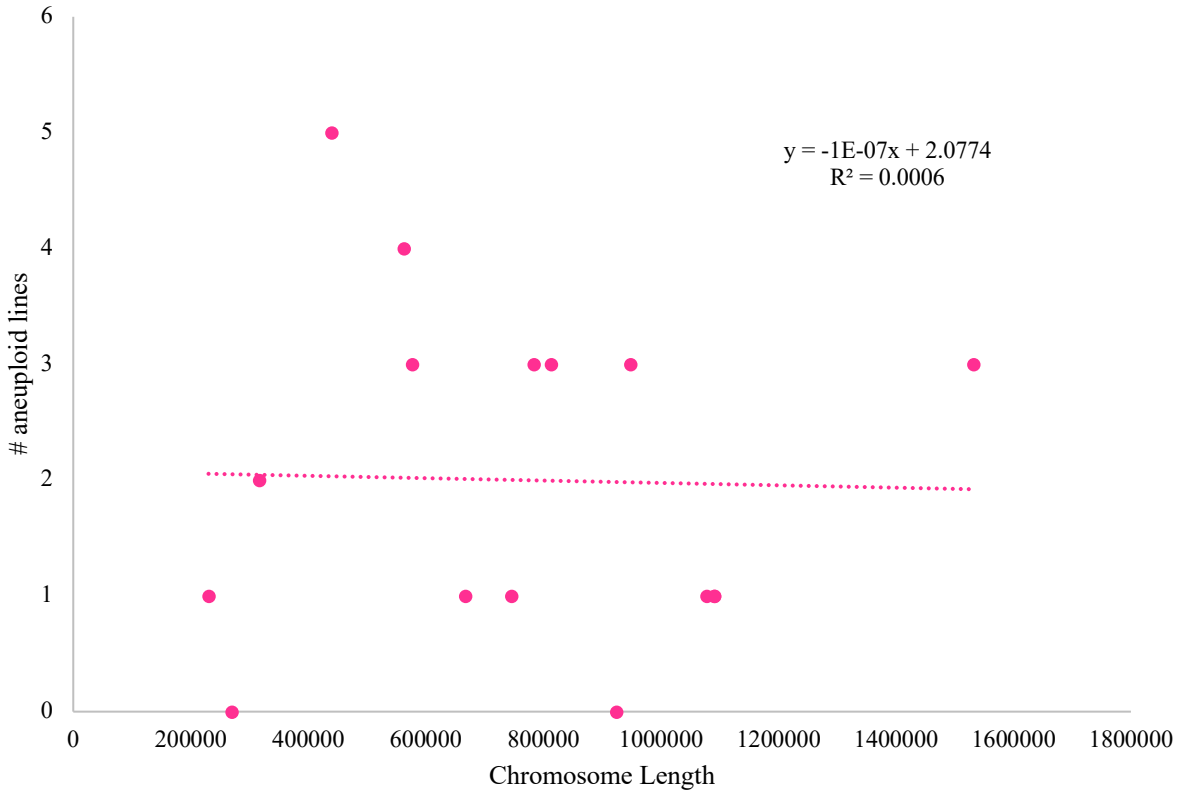


Figure 2.2: Relationship between chromosome size and aneuploidy events. There is no relationship between chromosome size and the number of aneuploidy (nondisjunction) events captured during mutation accumulation.

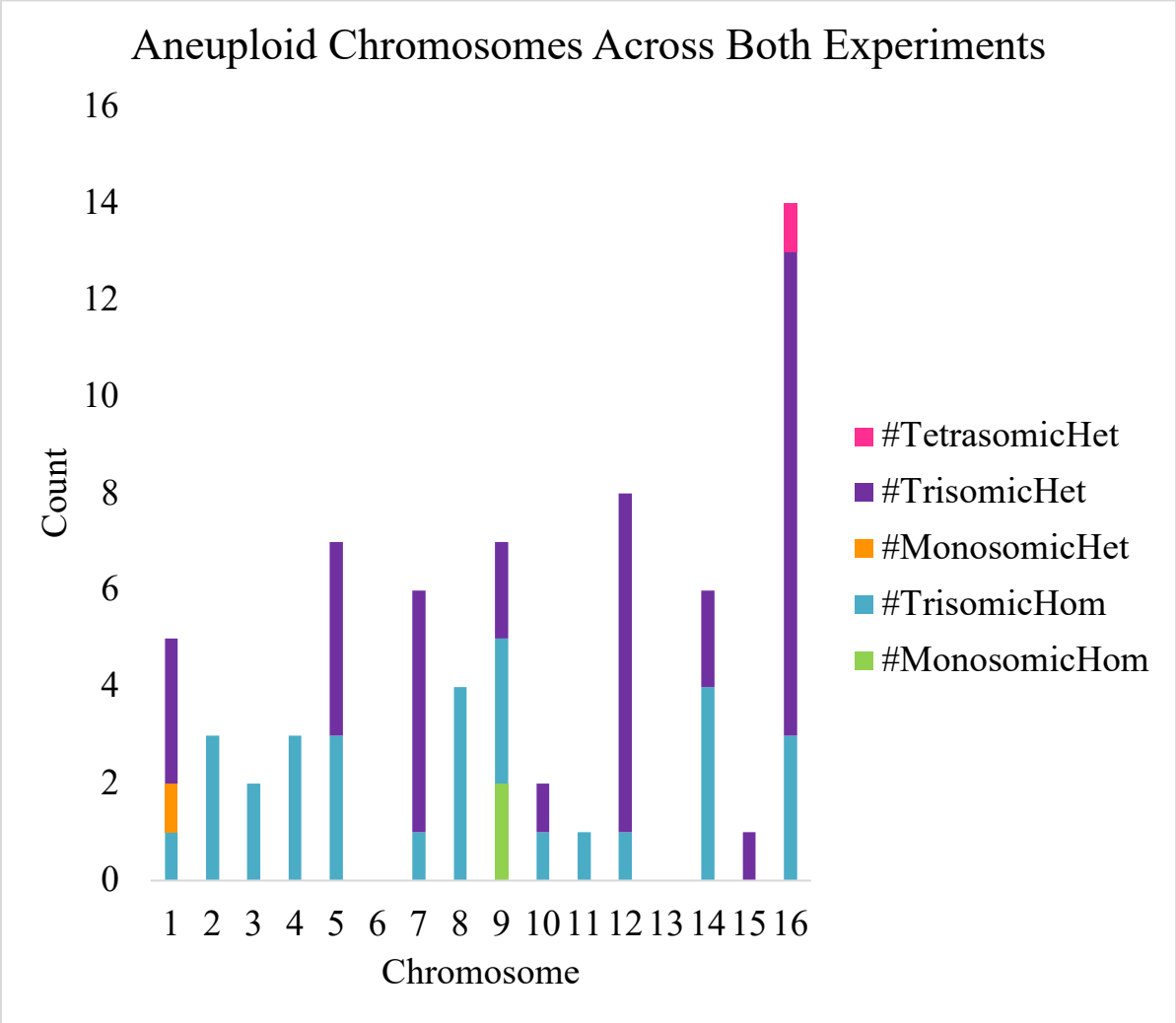
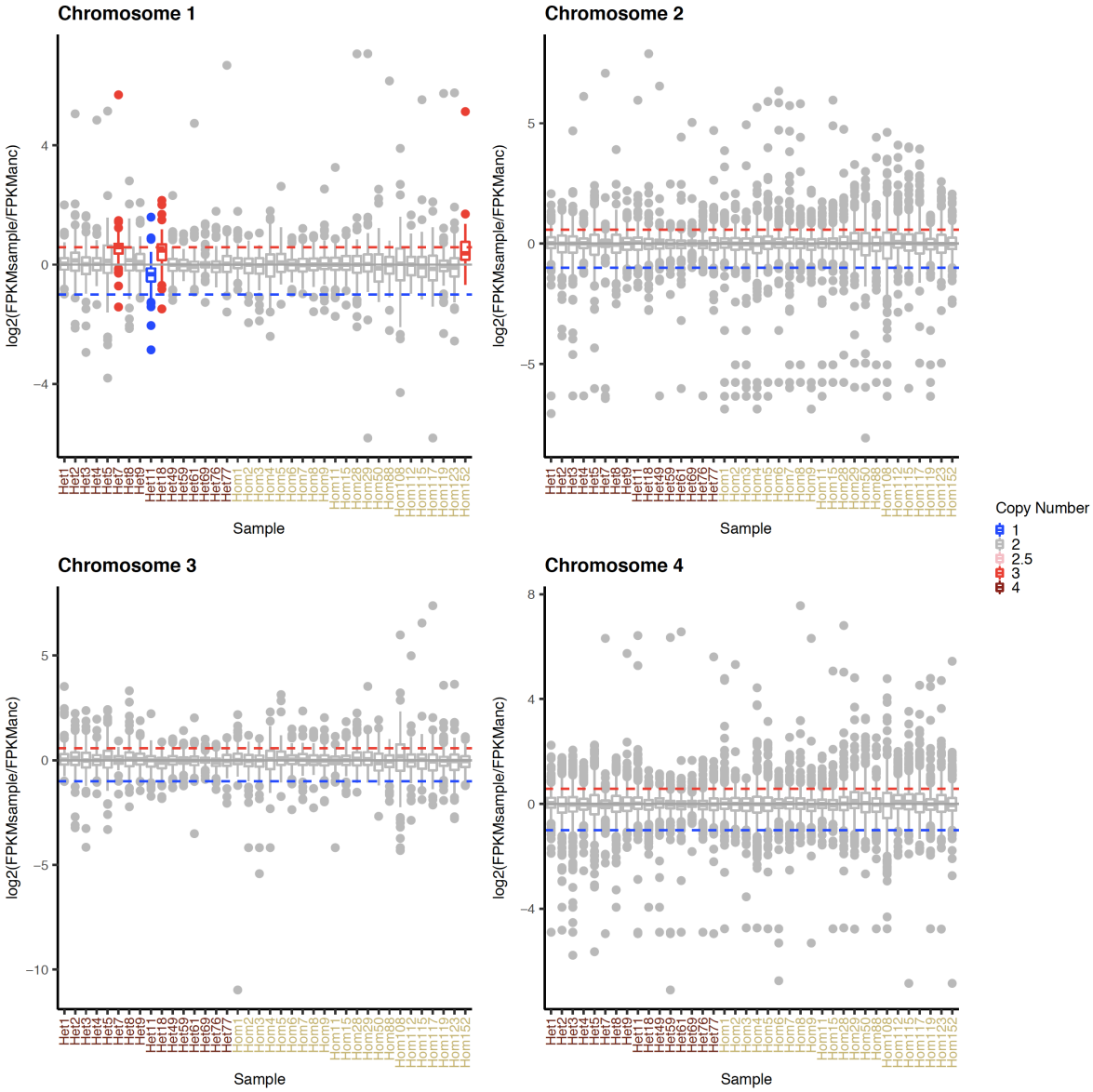
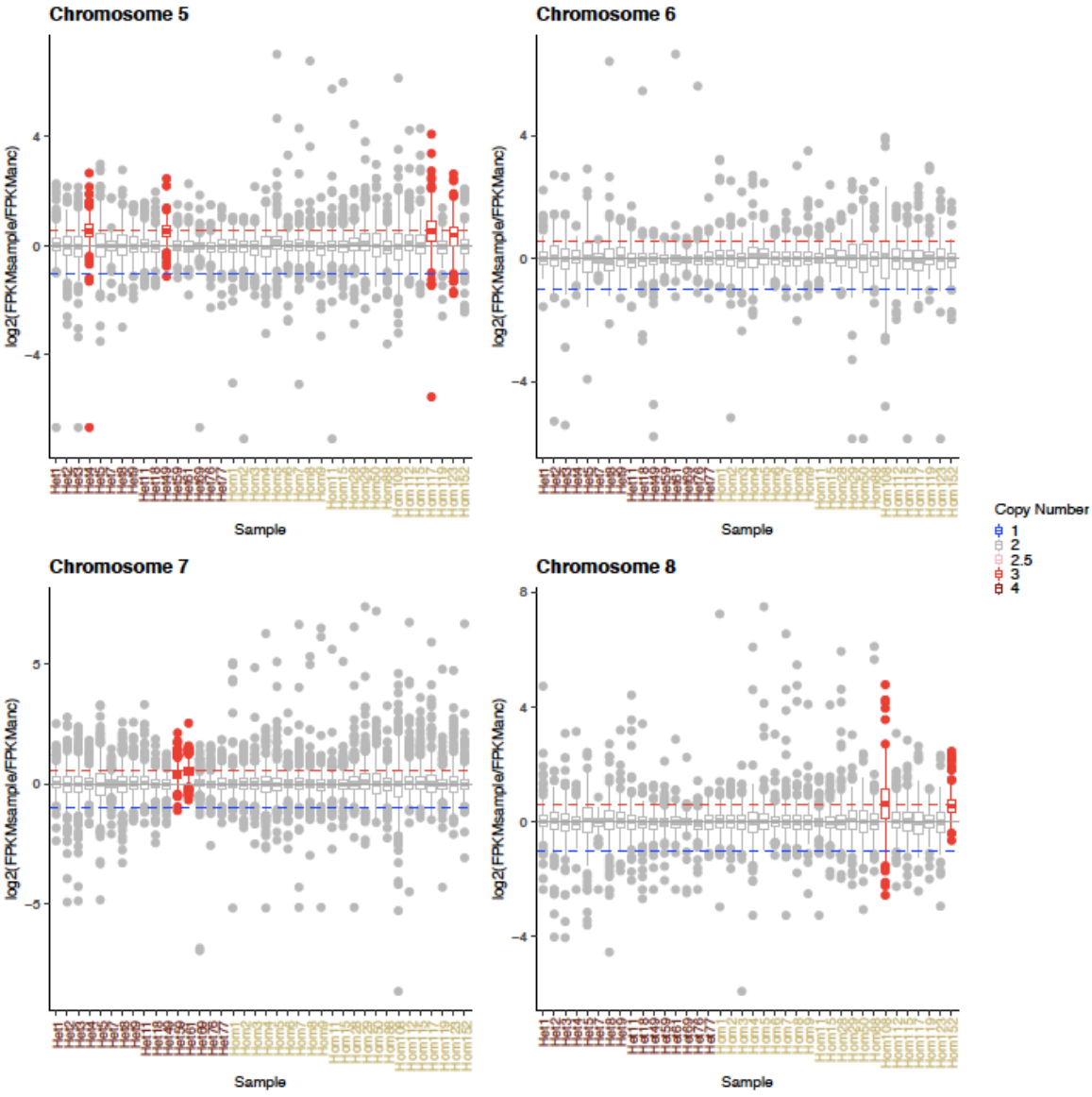
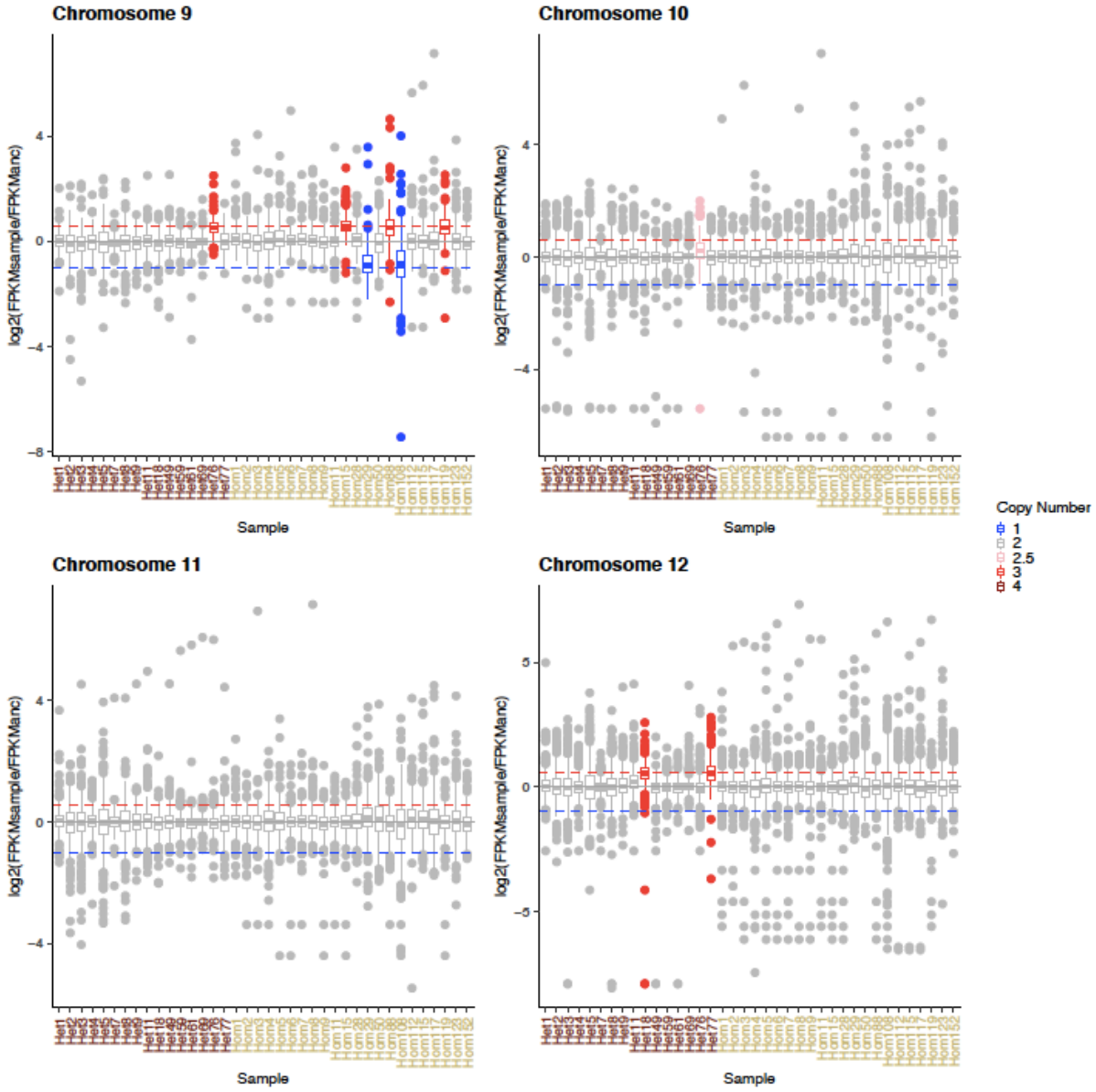


Figure 2.3: Distribution of aneuploidies in both heterozygous and homozygous ancestors across chromosomes. Chromosomes 6 and 13 never showed aneuploidy events in either strain.


Chromosome 1
Chromosome 2
Chromosome 3
Chromosome 4
log2(FPKMsample/FPKManc)
Sample
Copy Number
1
2
2.5
3
4


Chromosome 5
Chromosome 6
Chromosome 7
Chromosome 8
log2(FPKMsample/FPKMancr)
Sample
Copy Number
1
2
2.5
3
4


Chromosome 9
Chromosome 10
Chromosome 11
Chromosome 12
log2(FPKMsample/FPKManc)
Sample
Copy Number
1
2
2.5
3
4

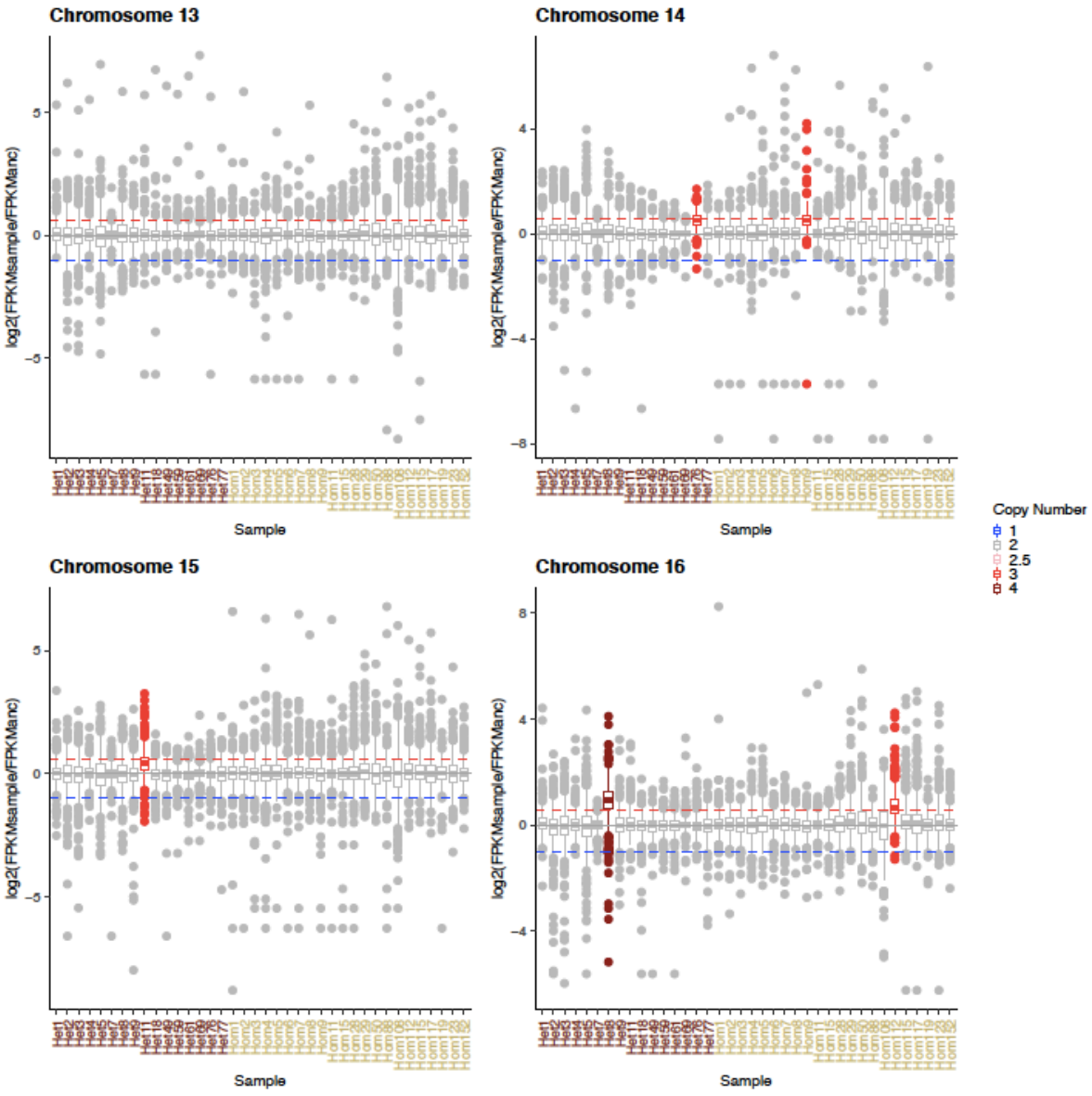


Figure 2.4: Boxplots showing gene expression levels measured by log2(FPKMratio) for 16 heterozygous ancestor (first 16 columns) and 22 homozygous ancestor derived MA lines. Horizontal gray line is expectation if there is no change in gene expression. Dashed red line is expectation for 1.5-fold increase and dashed blue line is for 2-fold decrease in gene expression. Boxes in blue indicates that MA line is monosomic for the chromosome, red indicates trisomy, dark red indicates tetrasomy, pink indicates a partially duplicated chromosome, and gray indicates disomy (the normal state).

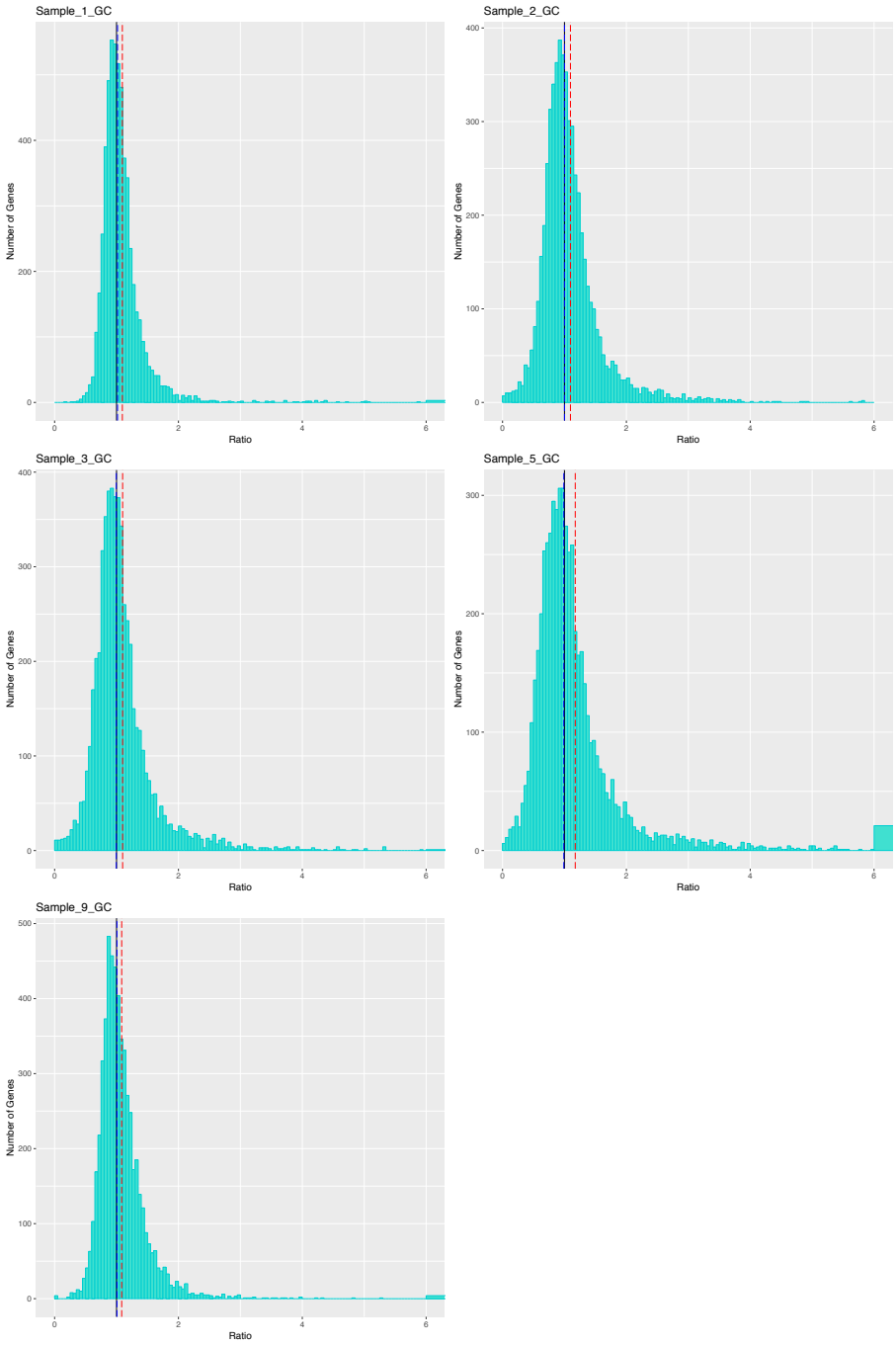


Figure 2.5. Example normalized count ratio distributions for genes in five euploid, heterozygous ancestor MA lines. Genes with normalized count ratios that were 6 or greater are binned together. Red line is mean of the MA line ratio to the ancestor, blue line is median of ratio to the ancestor, and the black line is intercept at 1 (ancestor = MA Line).

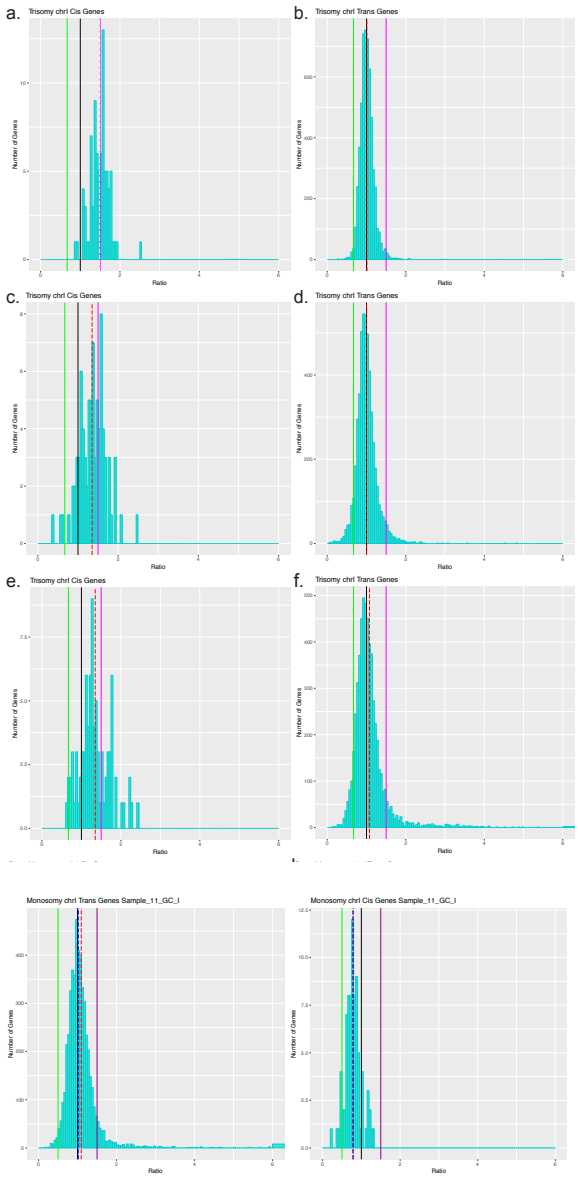


Figure 2.6. FPKM ratio distributions for cis and trans genes for four MA lines aneuploid (three trisomic and one monosomic) for chromosome 1. Heterozygous ancestor MA line 7 (trisomic) cis (panel a) and trans (panel b) genes. Heterozygous ancestor MA line 18 (trisomic) cis (panel c) and trans (panel d) genes. Homozygous ancestor MA line 152 (trisomic) cis (panel e) and trans (panel f) genes. (Note that line 152 is trisomic for chromosomes 1 and 7). Heterozygous ancestor MA line 11 (monosomic) cis (panel c) and trans (panel d) genes. Red dotted line: average ratio, black line: ratio of 1 (equal expression compared to ancestor), green line: ratio of 0.5 (expectation for monosomic genes), magenta line: ratio of 1.5 (expectation for trisomic genes).

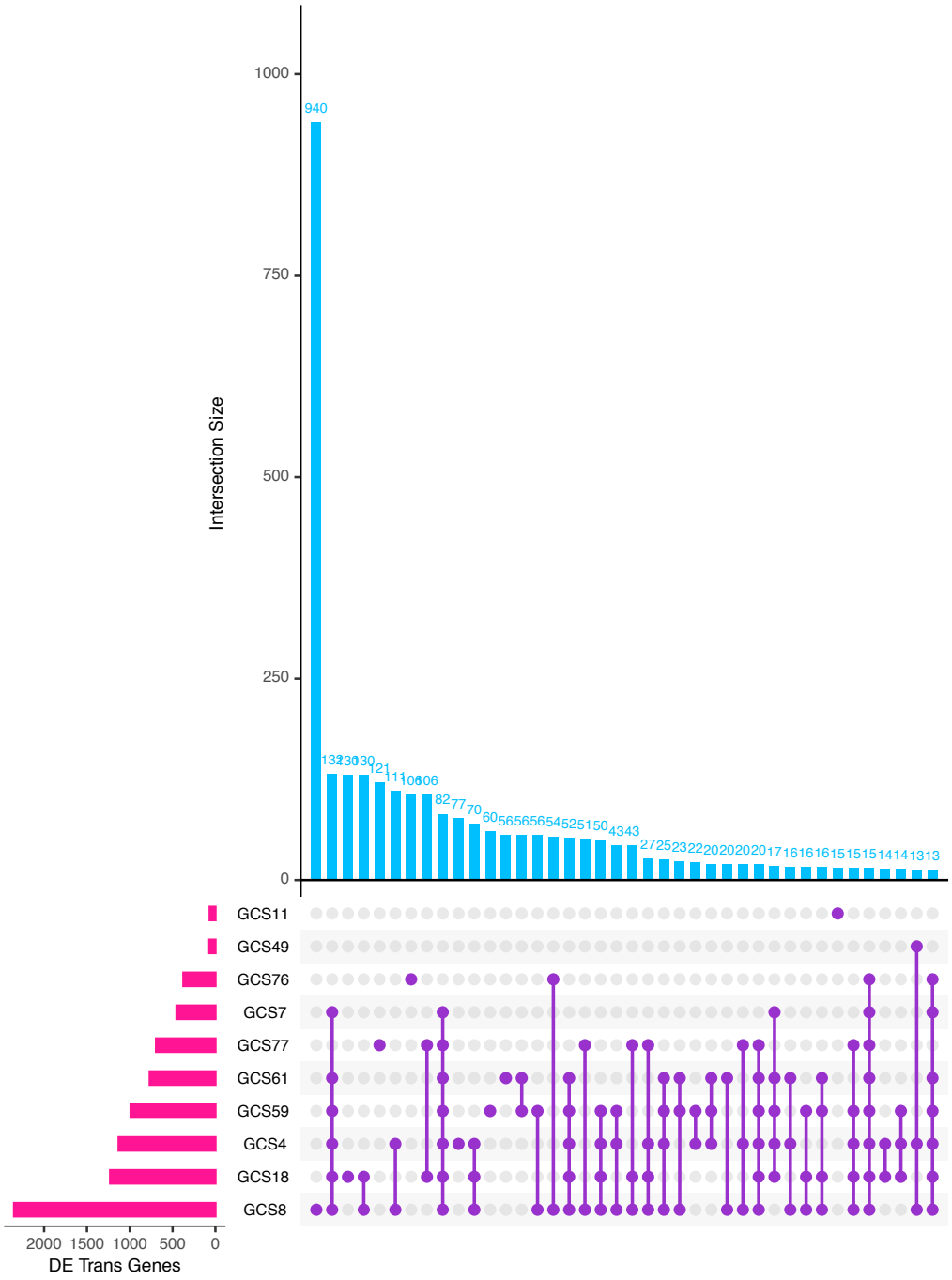


Figure 2.7: At most, 8/10 aneuploid samples from the heterozygous ancestor share 15 differentially expressed trans genes. Pink horizontal bars indicate the number of DE trans genes in a given line, while turquoise bars indicate the number of DE trans genes shared between the indicated lines (shown as purple circles).

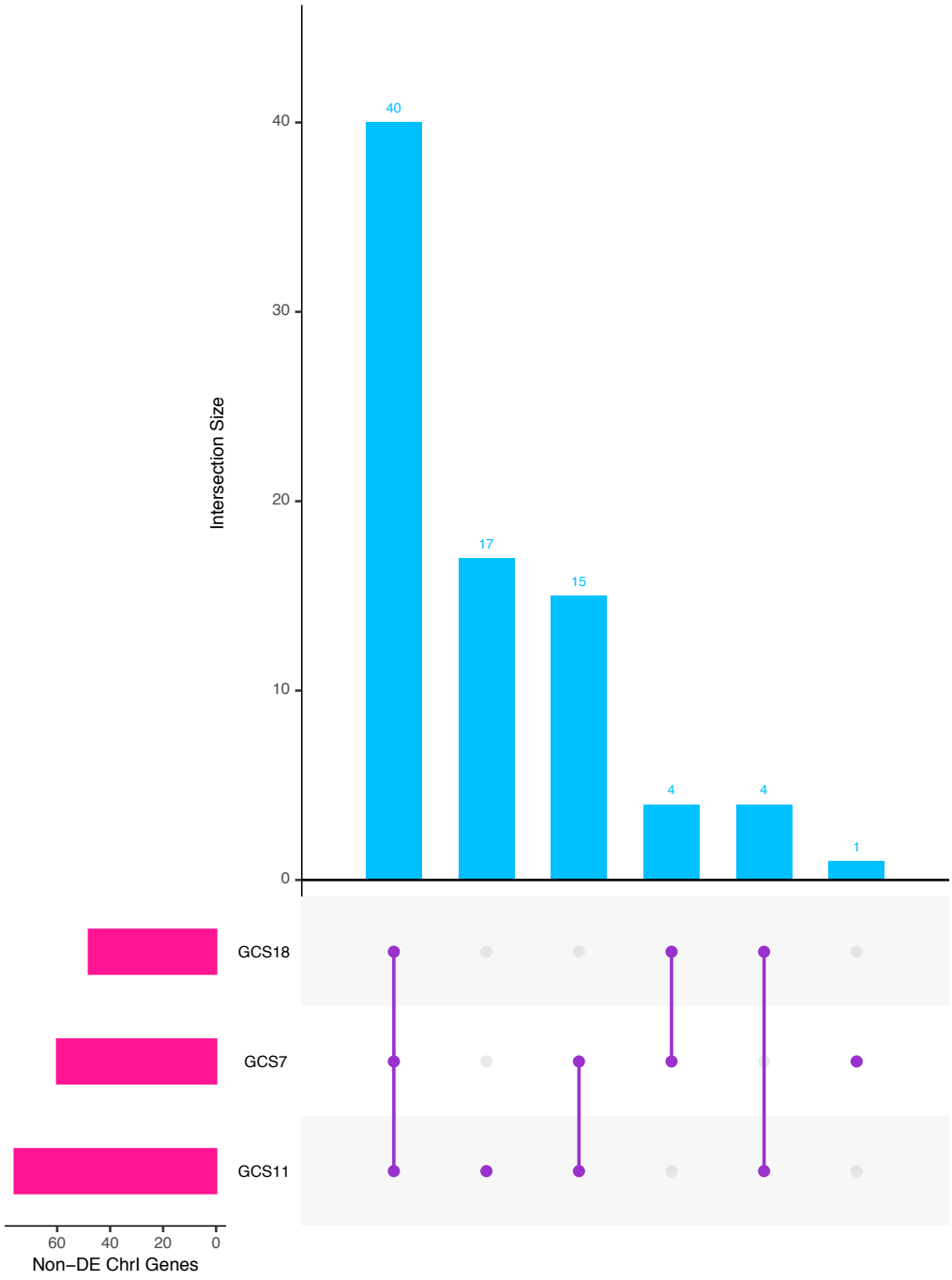


Figure 2.8: Non-differentially expressed genes on chromosome I that are shared between the heterozygous ancestor samples that are aneuploid for chromosome I (GCS11 is monosomic, the others are trisomic).

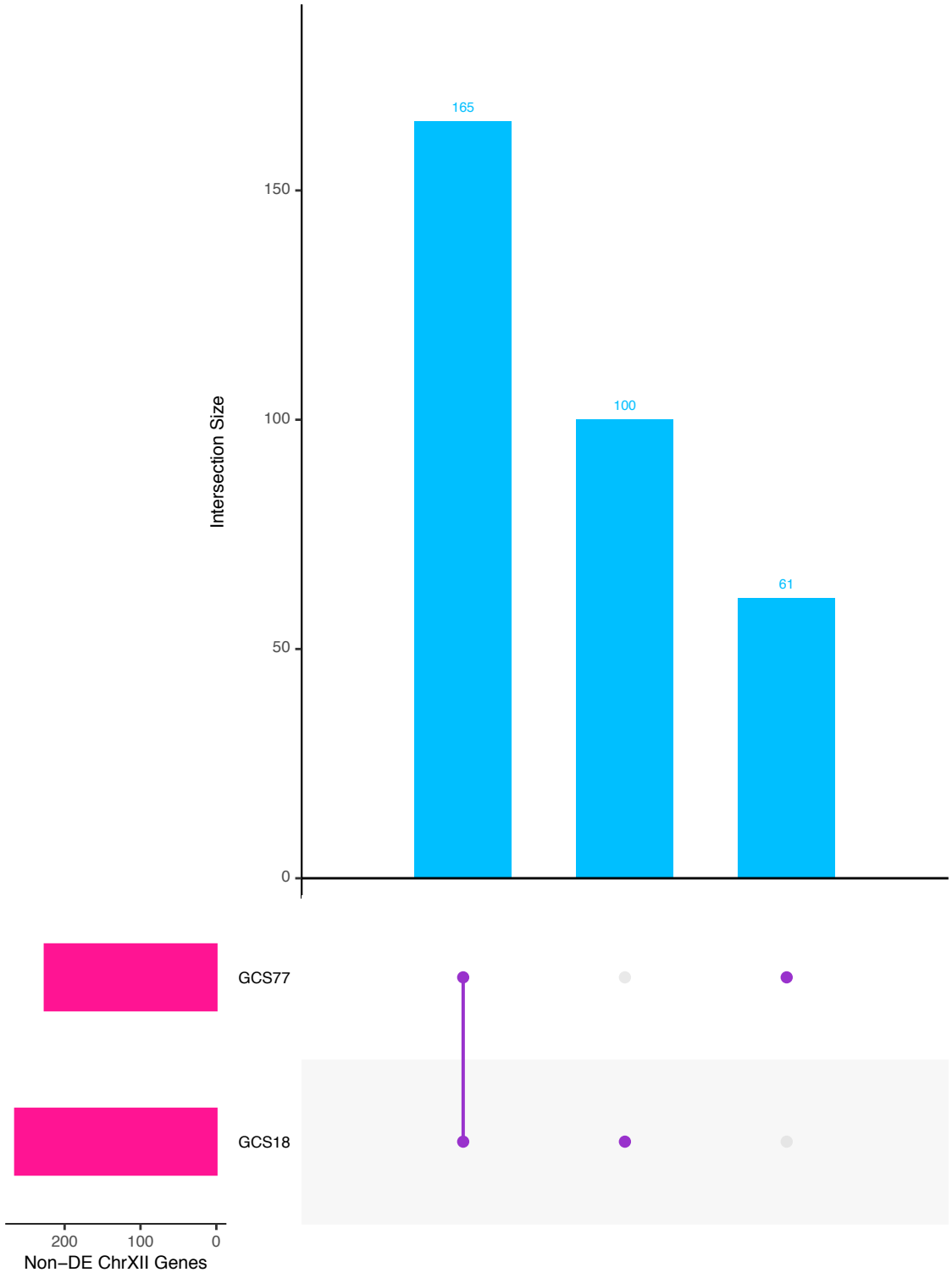


Figure 2.9: 165 non-DE genes shared between heterozygous ancestor samples trisomic for chromosome XII.

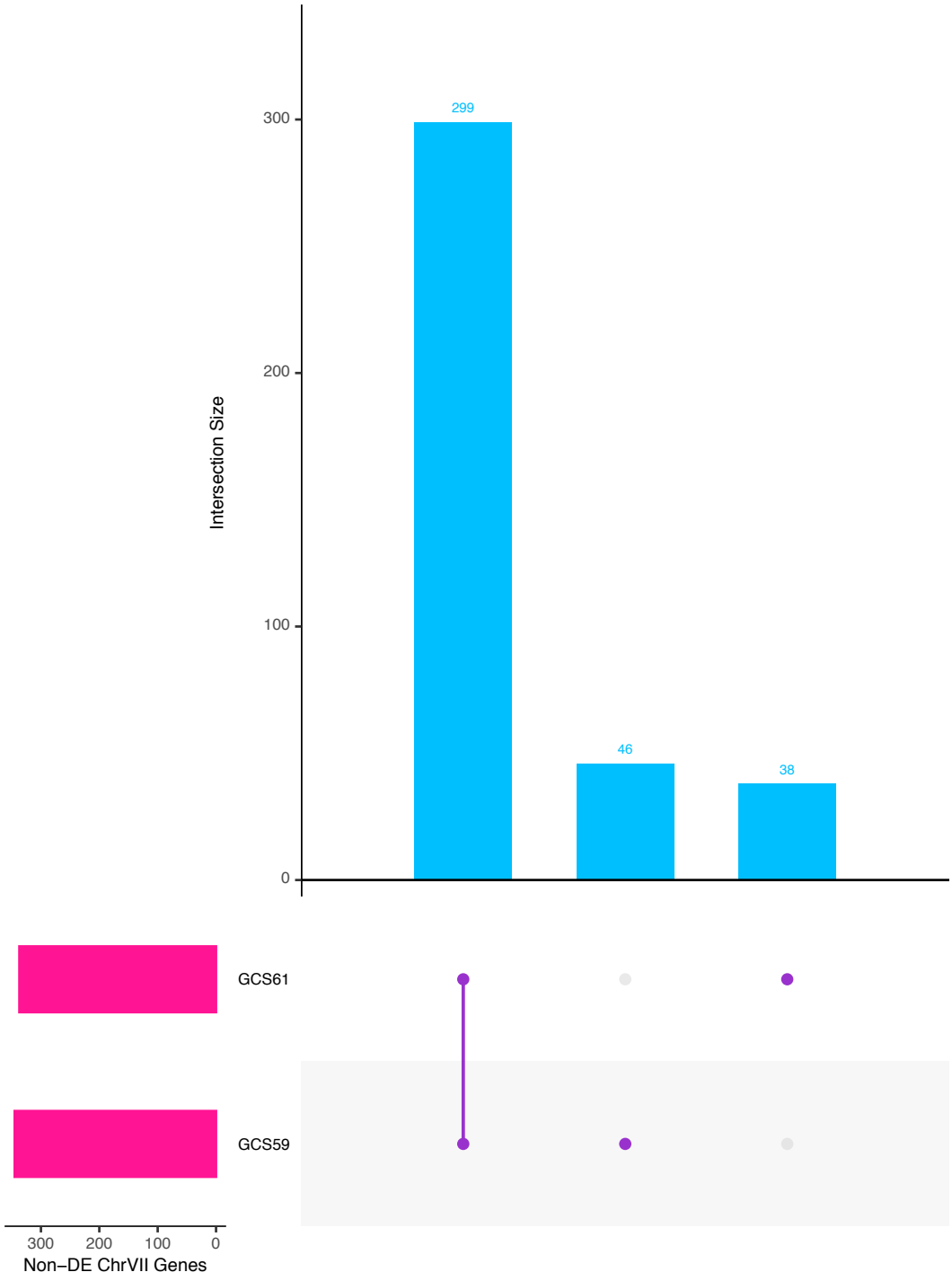


Figure 2.10: 299 non-DE genes shared between two heterozygous-ancestor samples trisomic for chromosome VII.

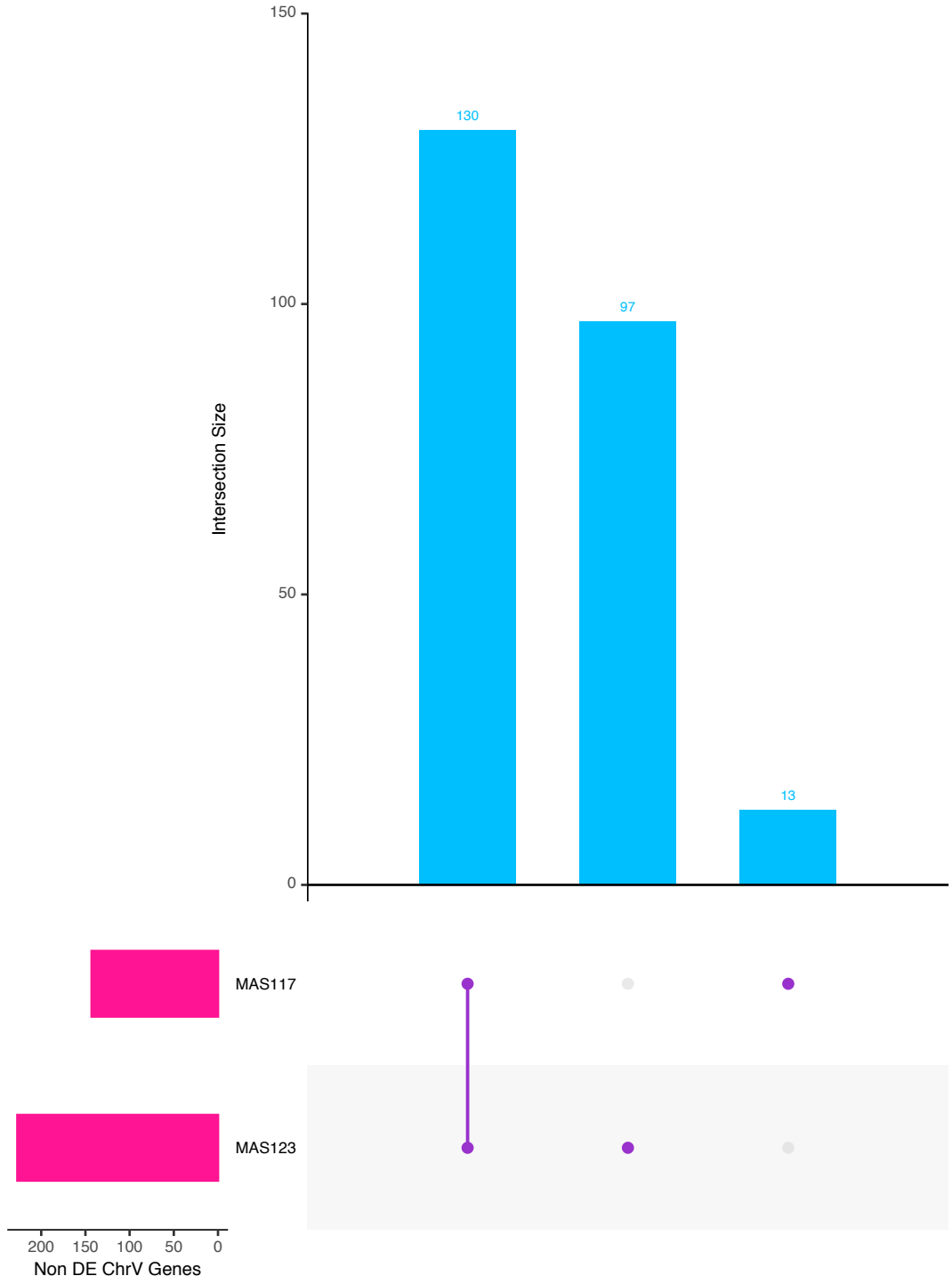


Figure 2.11: 130 shared non-DE genes on ChrV for two homozygous-ancestor derived samples trisomic for chr V.

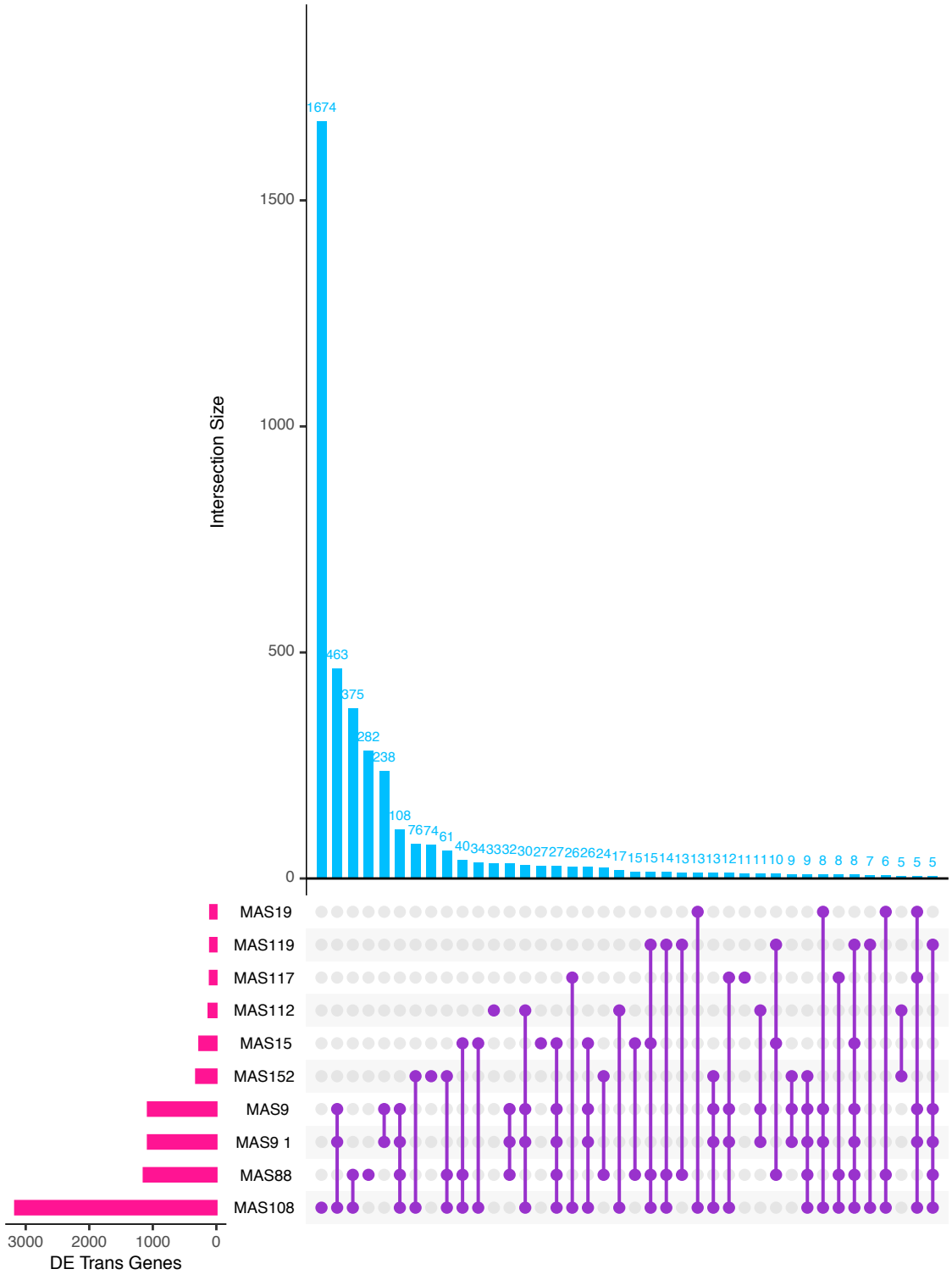


Figure 2.12: At most, 8 DE trans genes are shared between 6 aneuploid homozygous lines.

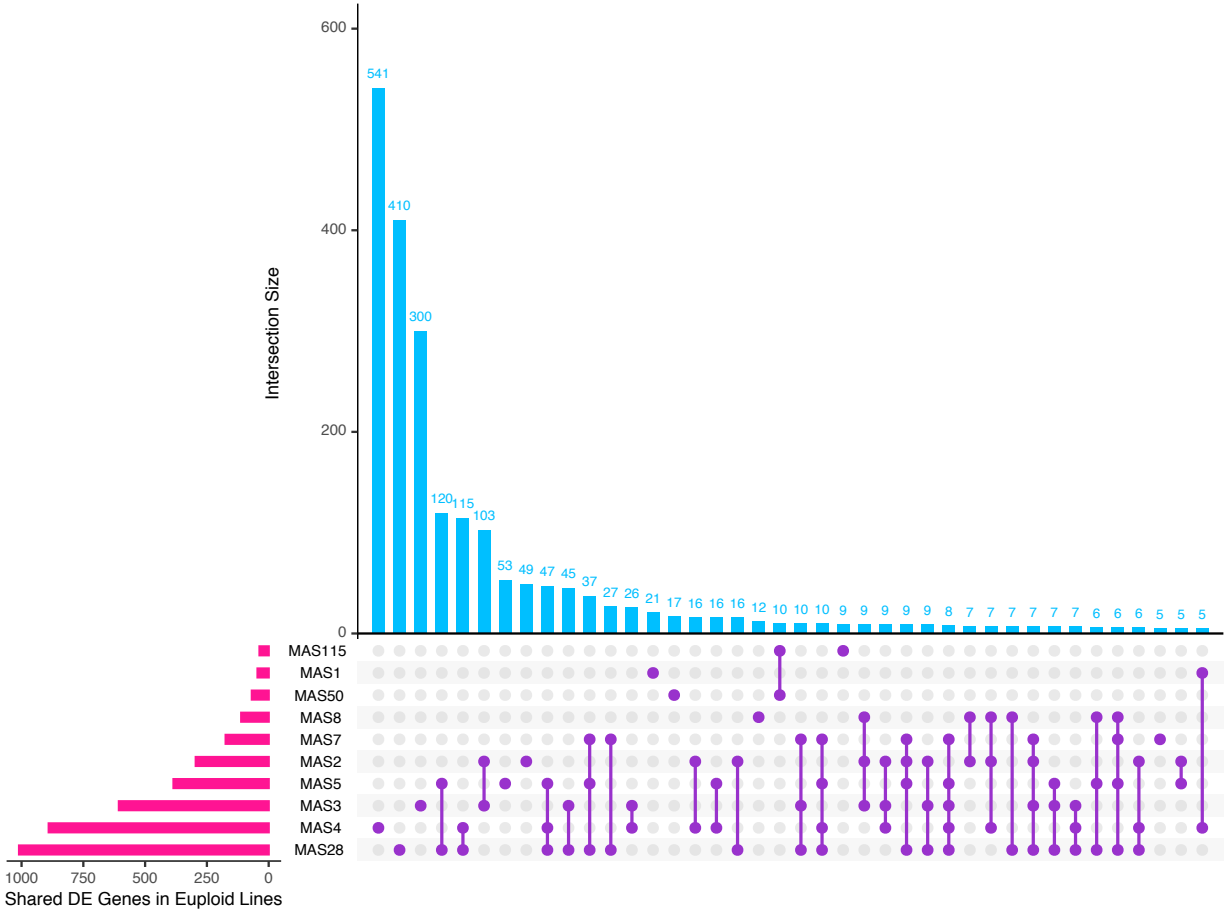


Figure 2.13: At most, 5 euploid homozygous ancestor lines share 8 DE genes.

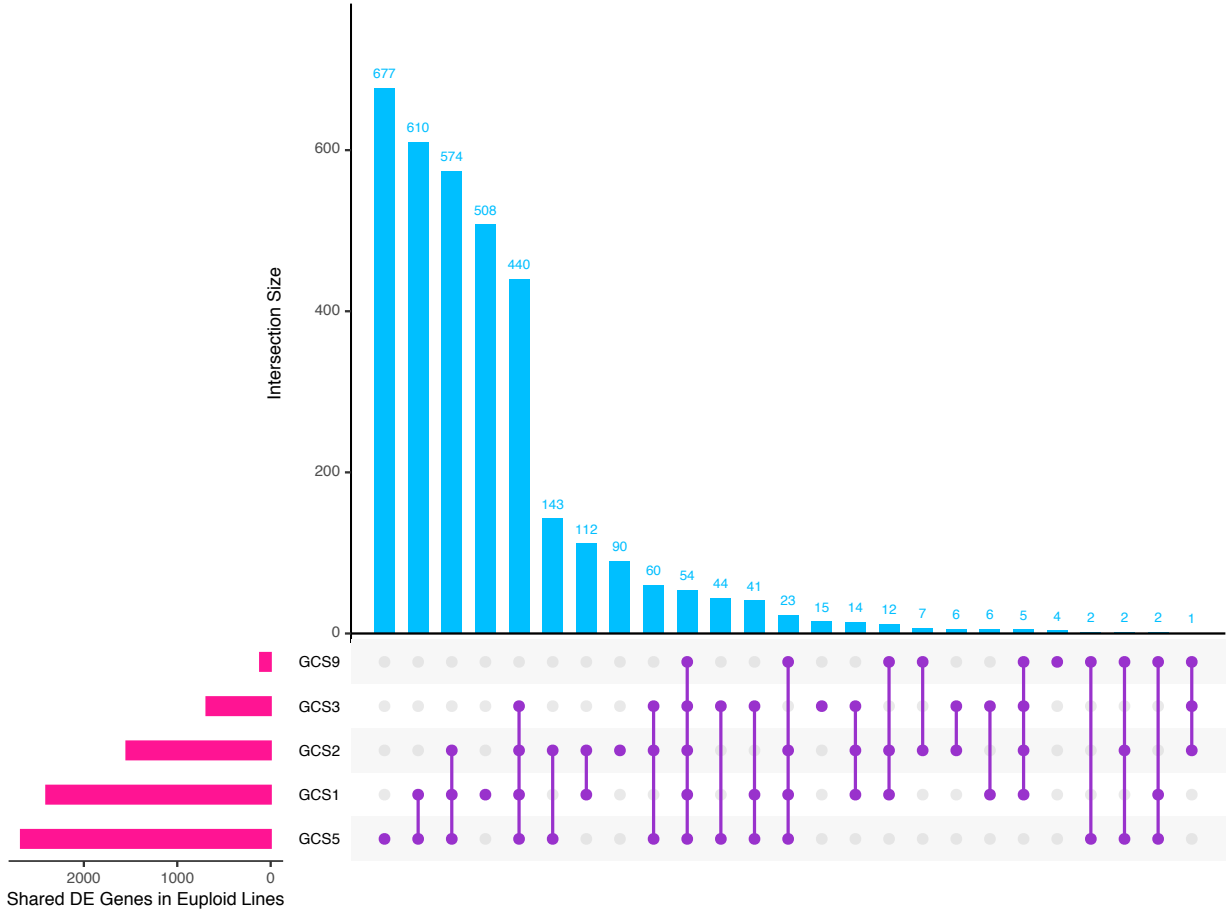


Figure 2.14: At most, 5 lines contain 54 shared DE genes in the heterozygous ancestor euploid lines.

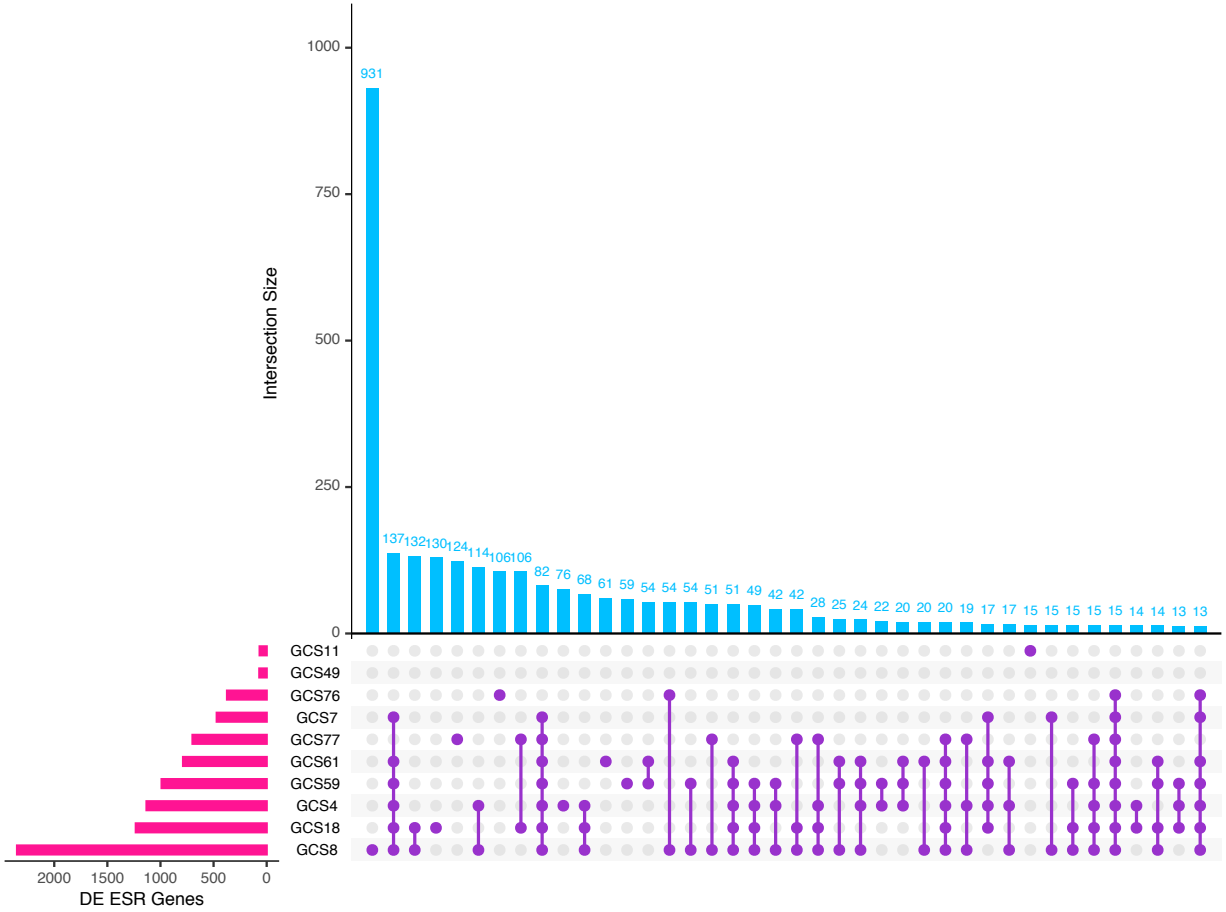


Figure 2.15: At most, 7 lines share 13 DE ESR genes heterozygous ancestor aneuploid lines.

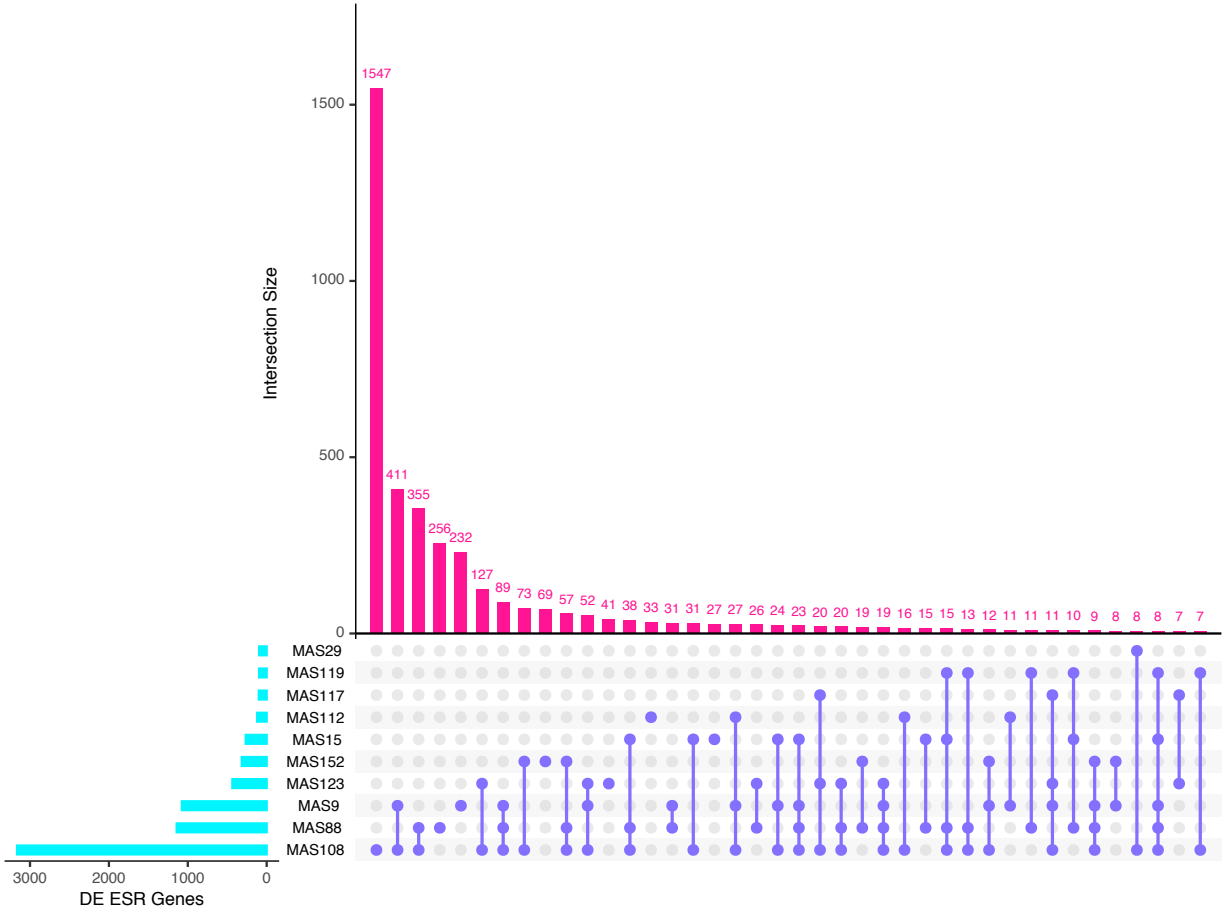


Figure 2.16: At most, 5 lines share 8 DE ESR genes in the homozygous ancestor aneuploid lines.

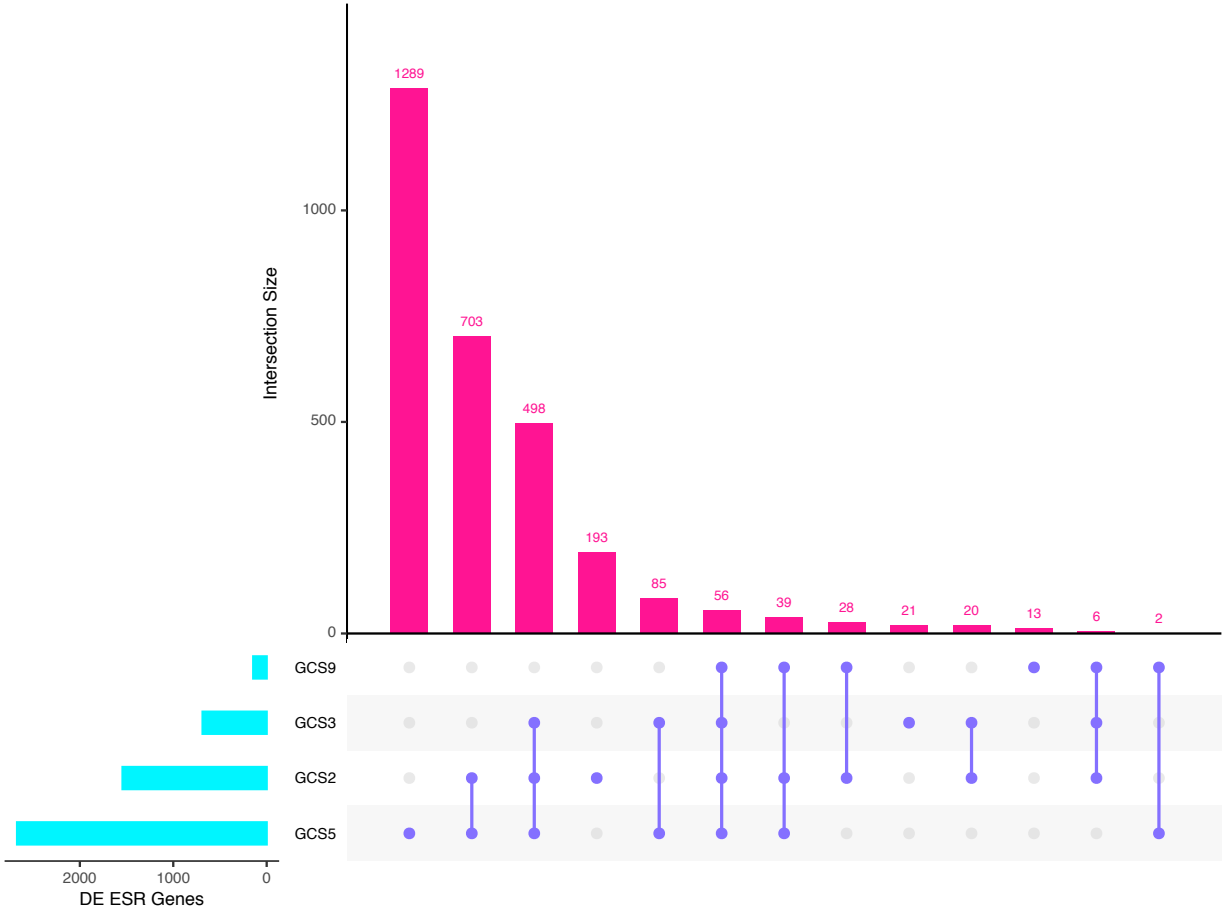


Figure 2.17: At most, 4 lines share 56 DE ESR genes in heterozygous ancestor euploid lines.

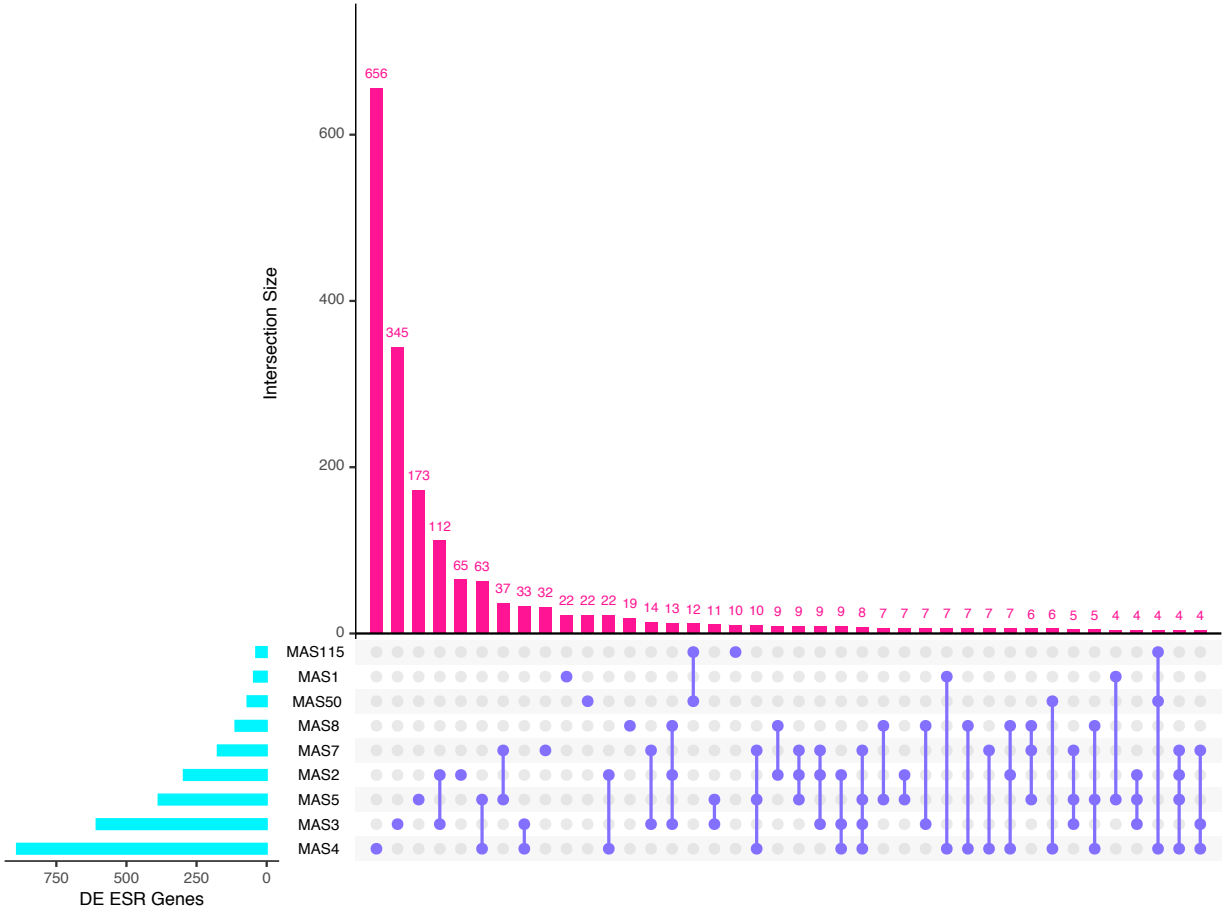


Figure 2.18: At most, 4 lines share 8 DE ESR genes in homozygous ancestor euploid lines.

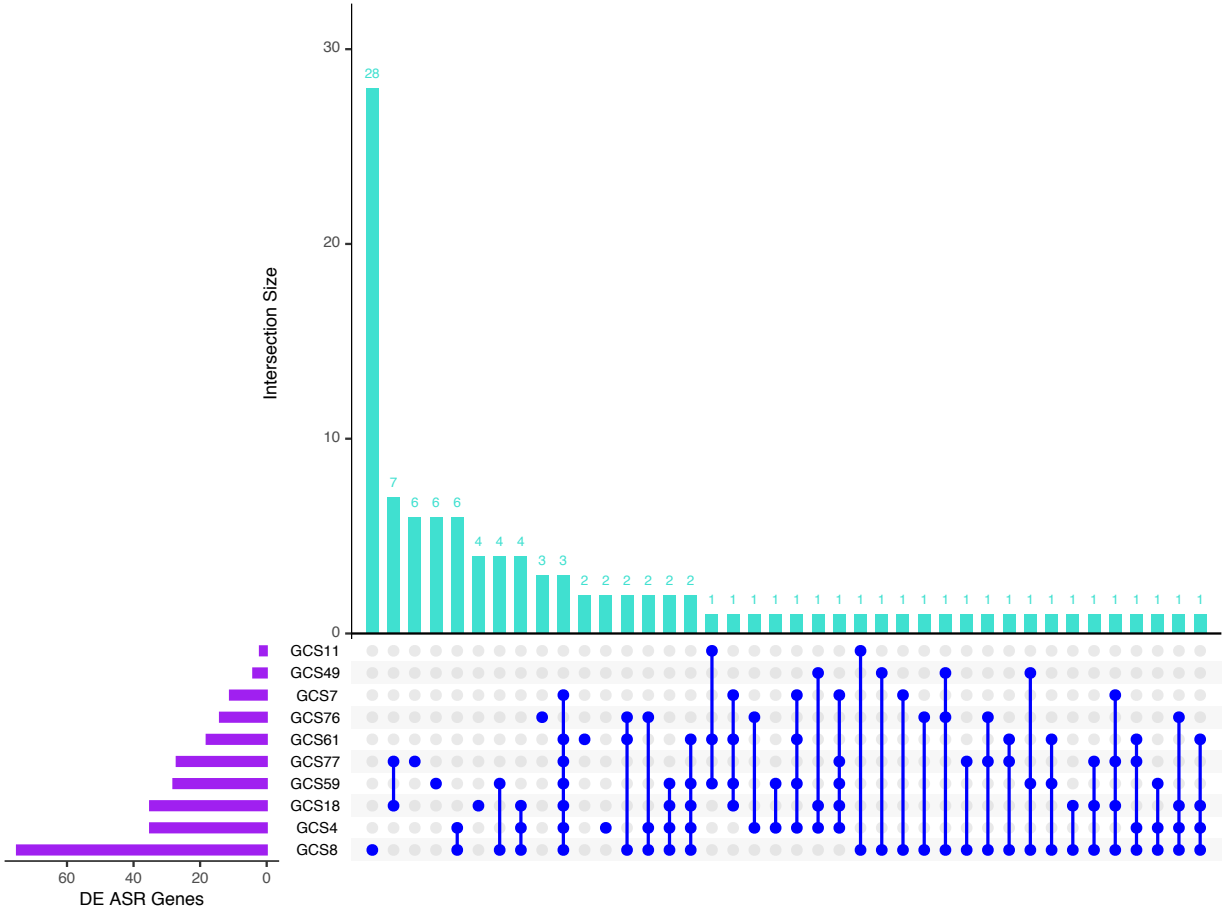


Figure 2.19: At most, 7 lines share 3 DE ASR genes in heterozygous ancestor aneuploid lines.

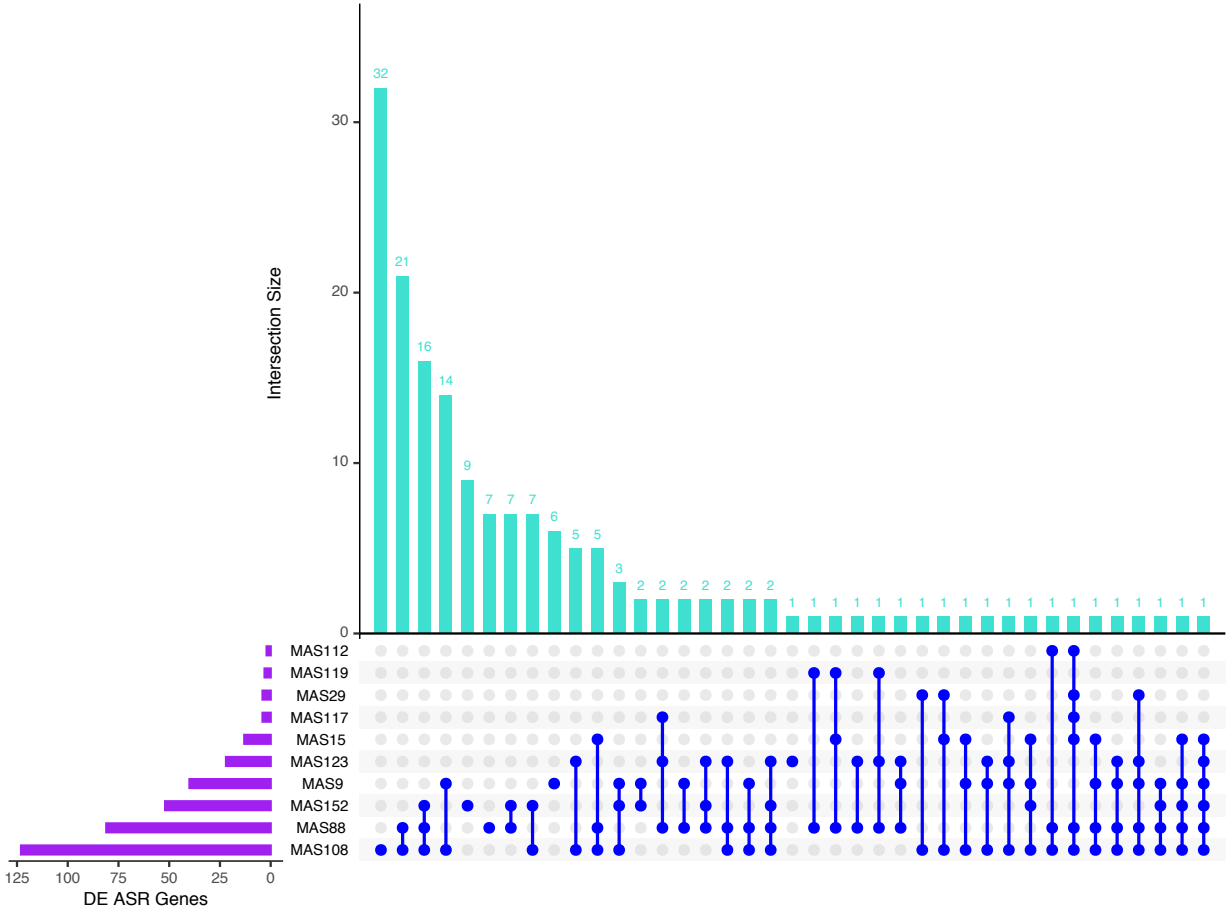


Figure 2.20: At most, 4 lines share 1 DE ASR gene in homozygous ancestor aneuploid lines.

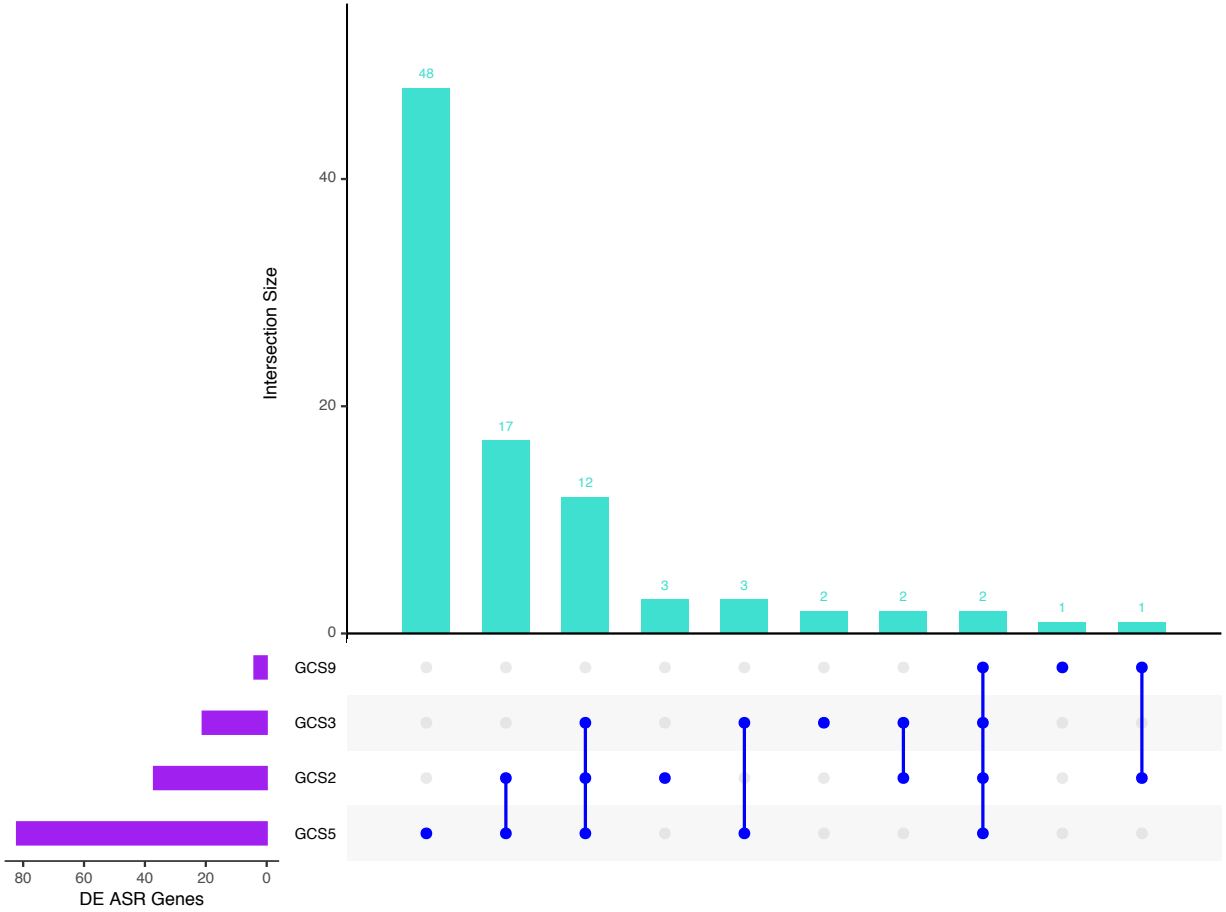


Figure 2.21: At most, 4 lines share 2 DE ASR genes in heterozygous ancestor euploid lines.

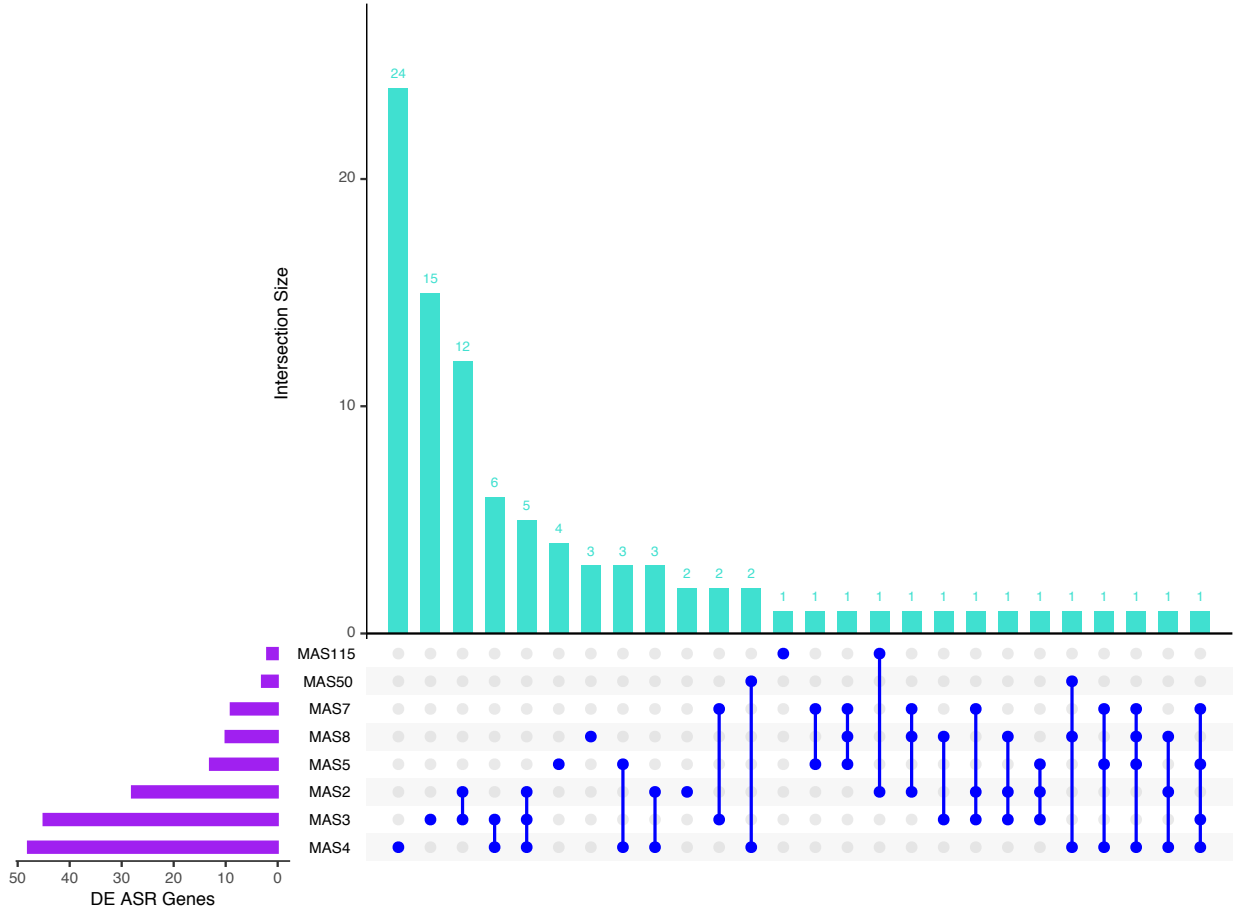


Figure 2.22: At most, 4 lines share 1 DE ASR gene in homozygous ancestor euploid lines.

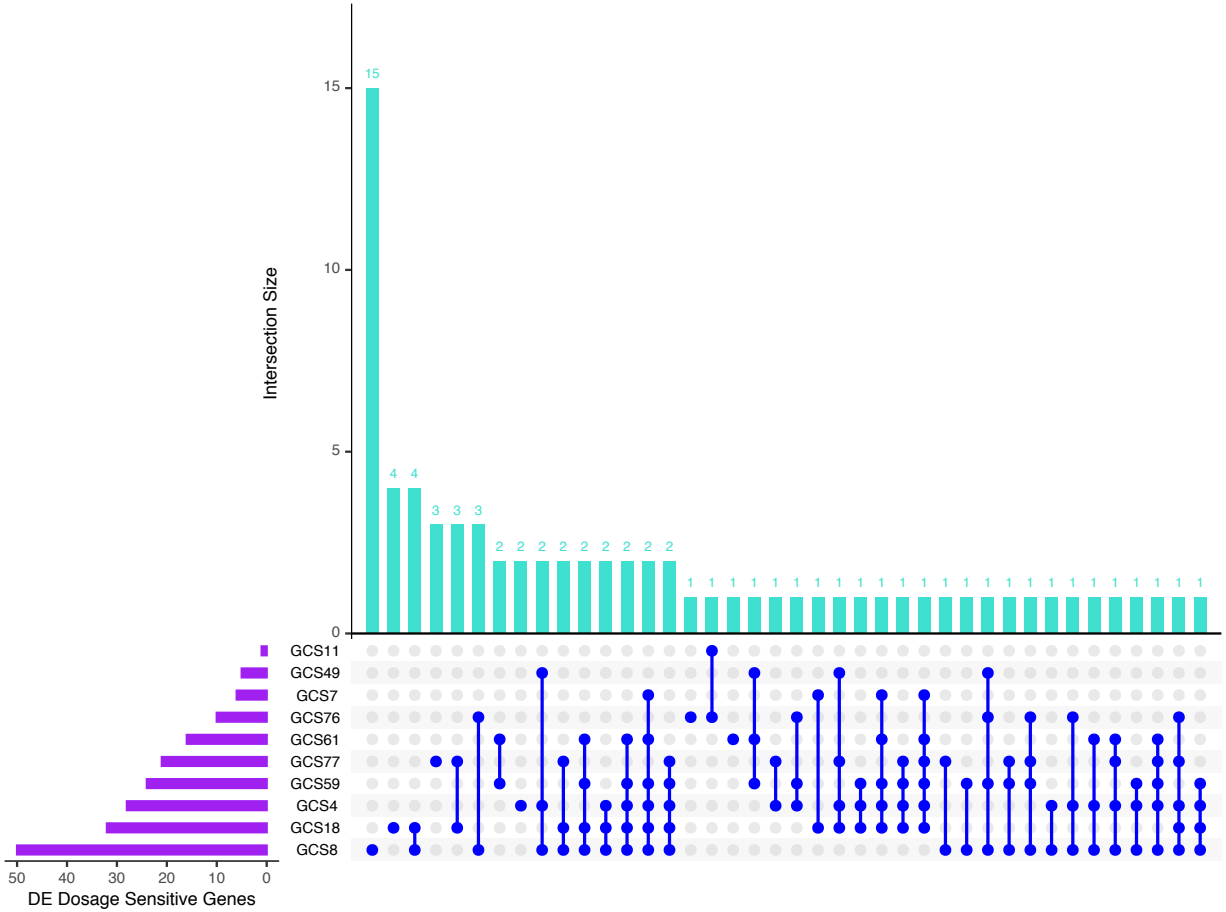


Figure 2.23: At most, 6 lines share 1 DE dosage sensitive gene in the heterozygous ancestor aneuploid lines.

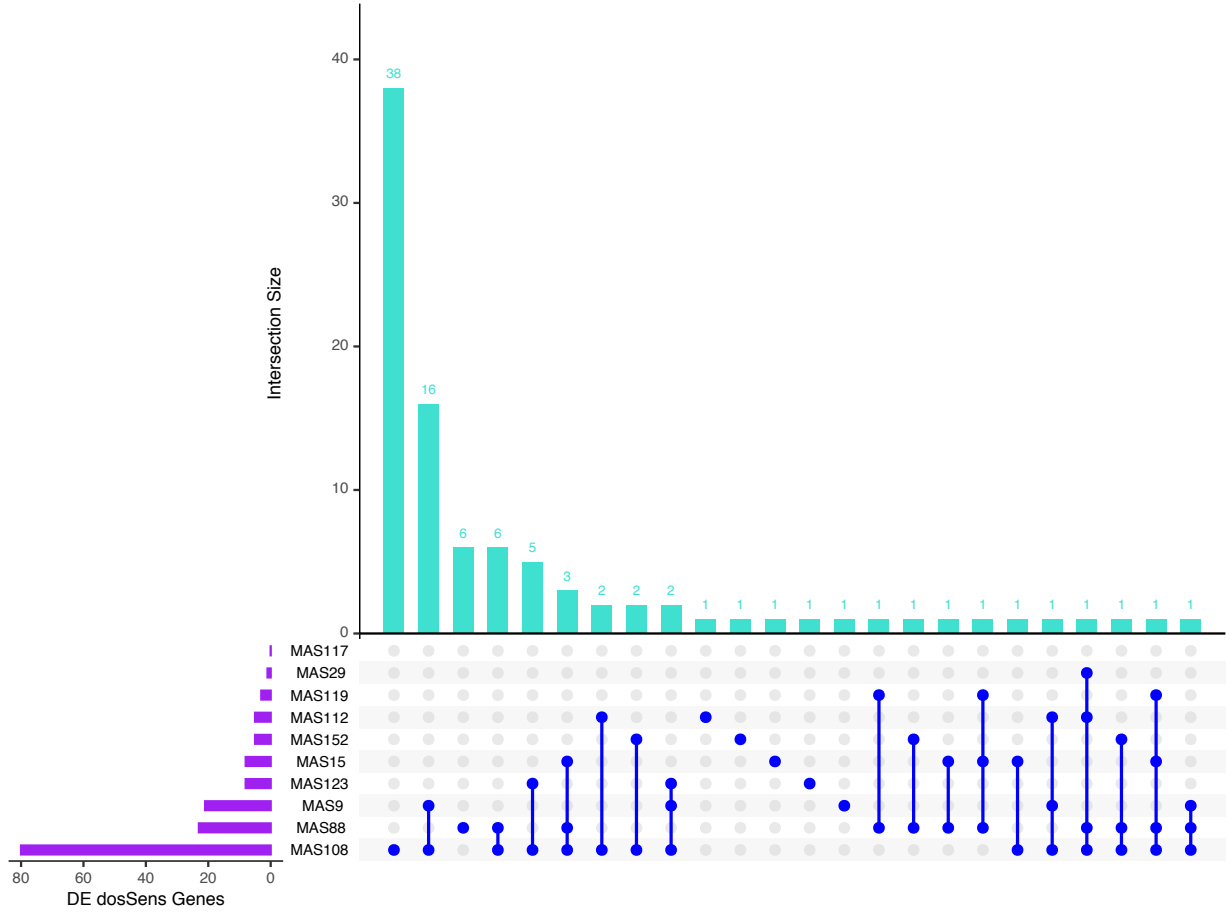


Figure 2.24: At most, 4 lines share 1 DE dosage sensitive gene in the homozygous ancestor aneuploid lines.

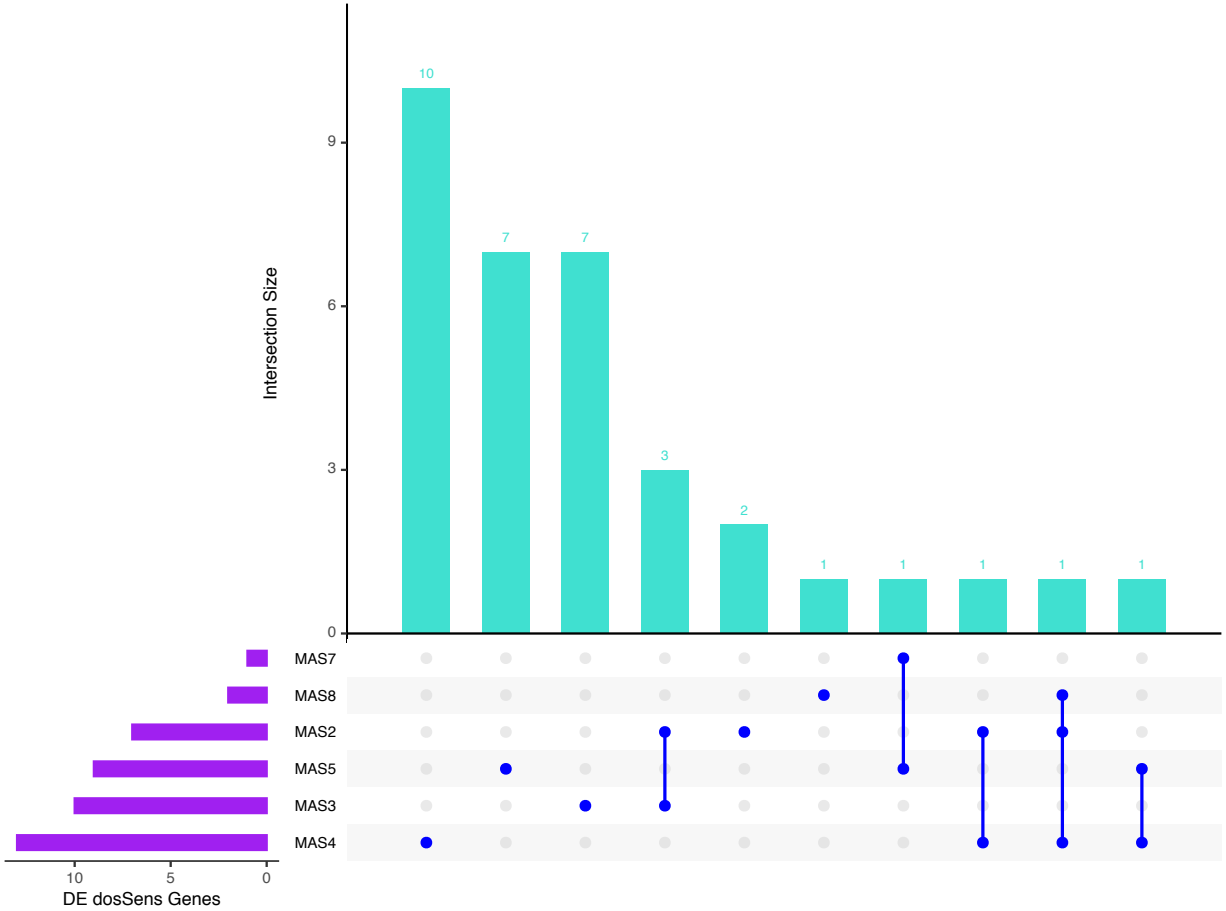


Figure 2.25: At most, 3 lines share 1 DE dosage sensitive gene in euploid lines from the homozygous ancestor.

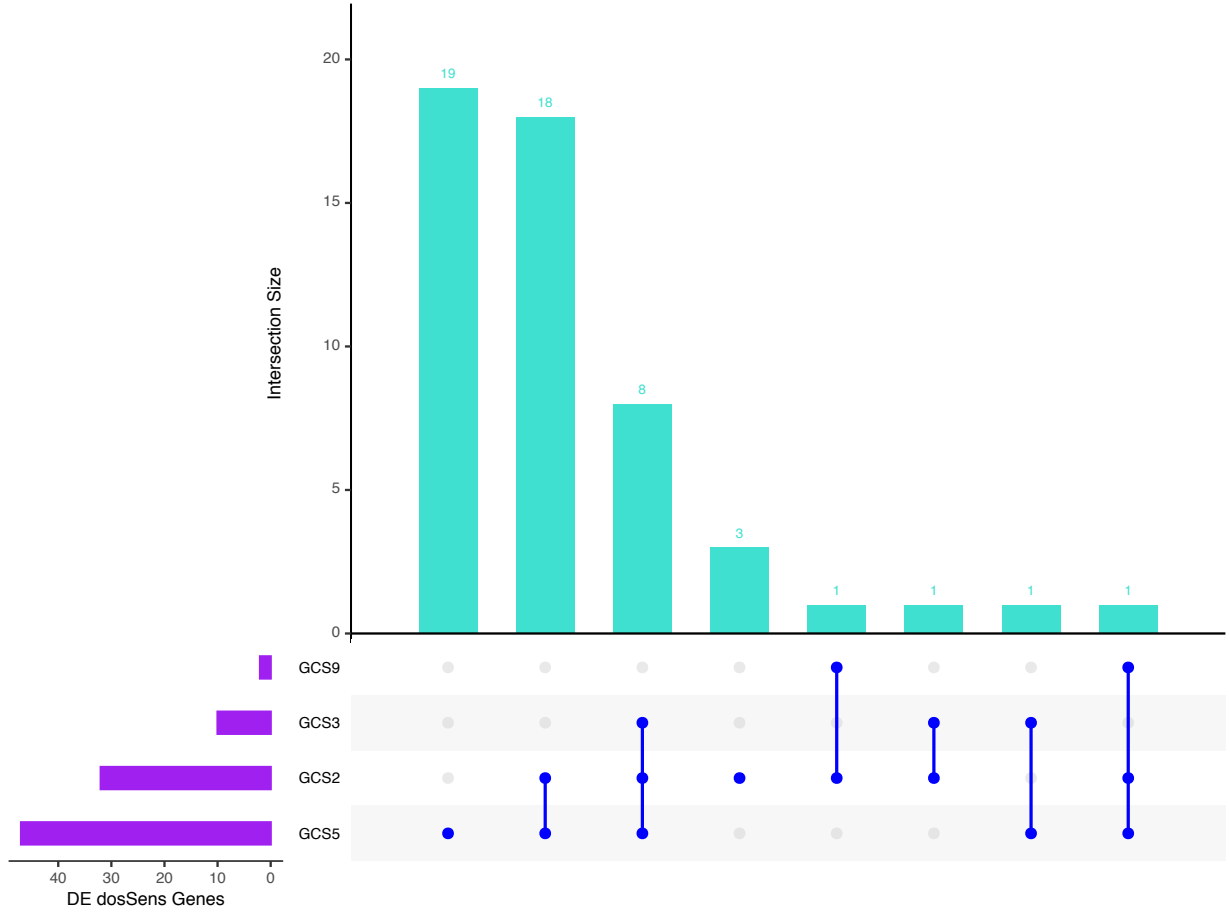


Figure 2.26: At most, 3 lines share 1 DE dosage-sensitive gene in euploid lines from the heterozygous ancestor.

References

Anders, K. R., J. R. Kudrna, K. E. Keller, B. Kinghorn, E. M. Miller *et al.*, 2009 A strategy for constructing aneuploid yeast strains by transient nondisjunction of a target chromosome. BMC Genet 10**:** 36.

Anders, S., P. T. Pyl and W. Huber, 2015 HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31**:** 166-169.

Audrey P Gasch, J. H., Michael A Newton, Maria Sardi, Mun Yong, Zhishi Wang, 2016 Further support for aneuploidy tolerance in wild yeast and effects of dosage compensation on gene copy-number evolution. eLIFE 5**:** 1-12.

Birchler, J. A., J. Hiebert and K. Paigen, 1990 Analysis of autosomal dosage compensation involving the alcohol dehydrogenase locus in Drosophila melanogaster. Genetics 124**:** 677-686.

Campbell, D., J. S. Doctor, J. H. Feuersanger and M. M. Doolittle, 1981 Differential mitotic stability of yeast disomes derived from triploid meiosis. Genetics 98**:** 239-255.

Carlson M, M. B., 2015 TxDb.Scerevisiae.UCSC.sacCer3.sgdGene: Annotation package for TxDb object(s). R package version 3.2.2.

Chandler, C. H., 2017 When and why does sex chromosome dosage compensation evolve? Ann N Y Acad Sci 1389**:** 37-51.

Charlesworth, B., 1991 The evolution of sex chromosomes. Science 251**:** 1030-1033.

Chen, D., W. M. Toone, J. Mata, R. Lyne, G. Burns *et al.*, 2003 Global transcriptional responses of fission yeast to environmental stress. Mol Biol Cell 14**:** 214-229.

Chen, G., W. D. Bradford, C. W. Seidel and R. Li, 2012 Hsp90 stress potentiates rapid cellular adaptation through induction of aneuploidy. Nature 482**:** 246-250.

Chen, Z. X., and B. Oliver, 2015 X Chromosome and Autosome Dosage Responses in Drosophila melanogaster Heads. G3 (Bethesda) 5**:** 1057-1063.

Chunduri, N. K., and Z. Storchova, 2019 The diverse consequences of aneuploidy. Nature Cell Biology 21**:** 54-62.

de Vries, A. R. G., M. A. Voskamp, A. C. van Aalst, L. H. Kristensen, L. Jansen *et al.*, 2018 Laboratory evolution of a Saccharomyces cerevisiae x S. eubayanus hybrid under simulated lager-brewing conditions: genetic diversity and phenotypic convergence. bioRxiv**:** 476929.

Devlin, R. H., D. G. Holm and T. A. Grigliatti, 1982 Autosomal dosage compensation Drosophila melanogaster strains trisomic for the left arm of chromosome 2. Proc Natl Acad Sci U S A 79**:** 1200-1204.

Gasch, A. P., J. Hose, M. A. Newton, M. Sardi, M. Yong *et al.*, 2016 Further support for aneuploidy tolerance in wild yeast and effects of dosage compensation on gene copy-number evolution. Elife 5**:** e14409.

Gasch, A. P., P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen *et al.*, 2000 Genomic expression programs in the response of yeast cells to environmental changes. Molecular biology of the cell 11**:** 4241-4257.

Gu, L., and J. R. Walters, 2017 Evolution of sex chromosome dosage compensation in animals: a beautiful theory, undermined by facts and bedeviled by details. Genome biology and evolution 9**:** 2461-2476.

Hangnoh Lee, D.-Y. C., Cale Whitworth, Robert Eisman, Melissa Phelps, John Roote, Thomas Kaufman, Kevin Cook, Steven Russell, Teresa Przytycka, Brian Oliver, 2016 Effects of Gene Dose, Chromatin, and Network Topology on Expression in Drosophila melanogaster. PLoS Genetics 12.

Hassold, T., and P. Hunt, 2001 To err (meiotically) is human: the genesis of human aneuploidy. Nature Reviews Genetics 2**:** 280.

Hose, J., L. E. Escalante, K. J. Clowers, H. A. Dutcher, D. Robinson *et al.*, 2020 The genetic basis of aneuploidy tolerance in wild yeast. Elife 9**:** e52063.

Hose, J., C. M. Yong, M. Sardi, Z. Wang, M. A. Newton *et al.*, 2015 Dosage compensation can buffer copy-number variation in wild yeast. Elife 4**:** e05462.

Hou, J., X. Shi, C. Chen, M. S. Islam, A. F. Johnson *et al.*, 2018 Global impacts of chromosomal imbalance on gene expression in Arabidopsis and other taxa. Proceedings of the National Academy of Sciences 115**:** E11321-E11330.

James Hose, C. M. Y., Maria Sardi, Zhishi Wang, Michael A Newton, Audrey P Gasch, 2015 Dosage compensation can buffer copy-number variation in yeast. eLIFE 4**:** 1-27.

Joseph, S. B., and D. W. Hall, 2004a Spontaneous Mutations in Diploid Saccharomyces cerevisiae. Genetics 168**:** 1817-1825.

Joseph, S. B., and D. W. Hall, 2004b Spontaneous mutations in diploid Saccharomyces cerevisiae: more beneficial than expected. Genetics 168**:** 1817-1825.

Kaya, A., M. V. Gerashchenko, I. Seim, J. Labarre, M. B. Toledano *et al.*, 2015 Adaptive aneuploidy protects against thiol peroxidase deficiency by increasing respiration via key mitochondrial proteins. Proc Natl Acad Sci U S A 112**:** 10685-10690.

Koo, D.-H., M. Jugulam, K. Putta, I. B. Cuvaca, D. E. Peterson *et al.*, 2018 Gene duplication and aneuploidy trigger rapid evolution of herbicide resistance in common waterhemp. Plant physiology 176**:** 1932-1938.

Kumaran, R., S.-Y. Yang and J.-Y. Leu, 2013 Characterization of chromosome stability in diploid, polyploid and hybrid yeast cells. PLoS One 8**:** e68094.

Lee, H., D.-Y. Cho, C. Whitworth, R. Eisman, M. Phelps *et al.*, 2016 Effects of gene dose, chromatin, and network topology on expression in Drosophila melanogaster. PLoS genetics 12**:** e1006295.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25**:** 2078-2079.

Linder, R. A., J. P. Greco, F. Seidl, T. Matsui and I. M. Ehrenreich, 2017 The Stress-Inducible Peroxidase TSA2 Underlies a Conditionally Beneficial Chromosomal Duplication in Saccharomyces cerevisiae. G3 (Bethesda) 7**:** 3177-3184.

Love, M. I., W. Huber and S. Anders, 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology 15**:** 550.

Makanae, K., R. Kintaka, T. Makino, H. Kitano and H. Moriya, 2013 Identification of dosage-sensitive genes in Saccharomyces cerevisiae using the genetic tug-of-war method. Genome research 23**:** 300-311.

Malone, J. H., D. Y. Cho, N. R. Mattiuzzo, C. G. Artieri, L. Jiang *et al.*, 2012 Mediation of Drosophila autosomal dosage effects and compensation by network interactions. Genome Biol 13**:** r28.

Marin, I., M. L. Siegal and B. S. Baker, 2000 The evolution of dosage-compensation mechanisms. system 18**:** 19.

Marinoni, G., M. Manuel, R. F. Petersen, J. Hvidtfeldt, P. Sulo *et al.*, 1999 Horizontal transfer of genetic material among Saccharomyces yeasts. Journal of Bacteriology 181**:** 6488-6496.

Matos, I., M. Machado, M. Schartl and M. Coelho, 2015 Gene expression dosage regulation in an allopolyploid fish. PloS one 10**:** e0116309.

McAnally, A. A., and L. Y. Yampolsky, 2009 Widespread transcriptional autosomal dosage compensation in Drosophila correlates with gene expression level. Genome Biol Evol 2**:** 44-52.

Medici, M., E. Porcu, G. Pistis, A. Teumer, S. J. Brown *et al.*, 2014 Identification of novel genetic Loci associated with thyroid peroxidase antibodies and clinical thyroid disease. PLoS Genet 10**:** e1004123.

Mulla, W., J. Zhu and R. Li, 2014 Yeast: a simple model system to study complex phenomena of aneuploidy. FEMS microbiology reviews 38**:** 201-212.

Noah Dephoure, S. H., Ciara O'Sullivan, Stacie E Dodgson, Steven P Gygi, Angelika Amon, Eduardo M Torres 2014 Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. eLIFE 3**:** 1-27.

Osley, M. A., and L. M. Hereford, 1981 Yeast histone genes show dosage compensation. Cell 24**:** 377-384.

Pavelka, N., G. Rancati, J. Zhu, W. D. Bradford, A. Saraf *et al.*, 2010 Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. Nature 468**:** 321-325.

Peter R. Eriksson, D. G., V. Nagarajavel, and David J. Clark, 2012 Regulation of Histone Gene Expression in Budding Yeast. Genetics 191**:** 7-20.

Selmecki, A., A. Forche and J. Berman, 2006 Aneuploidy and isochromosome formation in drug-resistant Candida albicans. Science 313**:** 367-370.

Selmecki, A. M., Y. E. Maruvka, P. A. Richmond, M. Guillet, N. Shoresh *et al.*, 2015 Polyploidy can drive rapid adaptation in yeast. Nature 519**:** 349-352.

Strope, P. K., D. A. Skelly, S. G. Kozmin, G. Mahadevan, E. A. Stone *et al.*, 2015 The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. Genome research 25**:** 762-774.

Team, R. C., 2013 R: A language and environment for statistical computing.

Thomas, P. D., M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak *et al.*, 2003 PANTHER: a library of protein families and subfamilies indexed by function. Genome research 13**:** 2129-2141.

Torres, E. M., N. Dephoure, A. Panneerselvam, C. M. Tucker, C. A. Whittaker *et al.*, 2010 Identification of aneuploidy-tolerating mutations. Cell 143**:** 71-83.

Torres, E. M., T. Sokolsky, C. M. Tucker, L. Y. Chan, M. Boselli *et al.*, 2007 Effects of aneuploidy on cellular physiology and cell division in haploid yeast. Science 317**:** 916-924.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim *et al.*, 2012 Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols 7**:** 562-578.

Veitia, R. A., S. Bottani and J. A. Birchler, 2008 Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. Trends in Genetics 24**:** 390-397.

Wakabayashi, H., C. Tucker, G. Bethlendy, A. Kravets, S. L. Welle *et al.*, 2017 NuA4 histone acetyltransferase activity is required for H4 acetylation on a dosage-compensated monosomic chromosome that confers resistance to fungal toxins. Epigenetics Chromatin 10**:** 49.

Wickham, H., 2016 *ggplot2: elegant graphics for data analysis*. Springer.

Yona, A. H., Y. S. Manor, R. H. Herbst, G. H. Romano, A. Mitchell *et al.*, 2012 Chromosomal duplication is a transient evolutionary solution to stress. Proceedings of the National Academy of Sciences 109**:** 21010-21015.

Zhu, Y. O., G. Sherlock and D. A. Petrov, 2016 Whole genome analysis of 132 clinical Saccharomyces cerevisiae strains reveals extensive ploidy variation. G3: Genes, Genomes, Genetics 6**:** 2421-2434.

Zhu, Y. O., M. L. Siegal, D. W. Hall and D. A. Petrov, 2014 Precise estimates of mutation rate and spectrum in yeast. Proceedings of the National Academy of Sciences 111**:** E2310-E2318.

Zillikens, M. C., S. Demissie, Y. H. Hsu, L. M. Yerges-Armstrong, W. C. Chou *et al.*, 2017a Erratum: Large meta-analysis of genome-wide association studies identifies five loci for lean body mass. Nat Commun 8**:** 1414.

Zillikens, M. C., S. Demissie, Y. H. Hsu, L. M. Yerges-Armstrong, W. C. Chou *et al.*, 2017b Large meta-analysis of genome-wide association studies identifies five loci for lean body mass. Nat Commun 8**:** 80.

CHAPTER 3

TRANSPOSON PRESENCE INCREASES RATE OF MULTINUCLEOTIDE MUTATIONS IN YEAST

Introduction

Mutation is the driver of genetic diversity and plays a vital role in the evolution and adaptation of species. Understanding how species evolve and adapt requires the characterization of the parameters of spontaneous mutations, including the rate, spectrum, and their potential phenotypic effects. There are many factors affecting mutation rate and spectra, including various environmental factors, such as radiation and chemicals. However, internal factors may also influence mutation rates. One such internal factor is mobile genetic elements, which include transposable elements (TEs or transposons), homing endonucleases, insertion sequences, endogenous retroviruses, and plasmids. Transposons are known to have expanded the genomes of numerous species, especially plants, in which they have been expanded dramatically – for example, the maize genome is comprised of ~80% transposable elements (BENNETZEN 2000). The mutational effects of transposition events themselves have been well characterized – they cause insertion mutations when they transpose to a new location in the genome and leave mutational scars when they are excised (WICKER *et al.* 2016).

Budding yeast (*Saccharomyces spp.*) contain 5 long-terminal repeat (LTR) retrotransposon families (Ty1 – Ty5), which are contained within the *copia* superfamily of retroelements (CURCIO *et al.* 2015). Of the 5 families, Ty1 has been studied the most,

with Ty1-H3 being the best-characterized Ty1. It is 5918 bp long, and each LTR is 334 bp in length. The retromobility of Ty1-H3 in the restrictive *Saccharomyces cerevisiae* strain S288C has been estimated to be approximately 1 x 10^{-6} events per generation, compared to a permissive strain (YPS606) at 1.5 x 10^{-4} transpositions per generation (CURCIO *et al.* 2015). *Saccharomyces* genomes are comprised of relatively little TE sequence, with most estimates around 3% (KIM *et al.* 1998; LITI *et al.* 2009; CARR *et al.* 2012).

Saccharomyces paradoxus and *Saccharomyces cerevisiae* diverged ~5-10 MYA, and they share ~90% sequence similarity (Naumov *et al.* 1992). *S. paradoxus* is often used in ecological genetics and transposon studies (Garfinkel *et al.* 2003; Moore *et al.* 2004; Garfinkel 2005; Matsuda and Garfinkel 2009). The two species share elements, though the Ty5 elements of *S. paradoxus* are transpositionally active, while those of *S. cerevisiae* S288C are defective (ZOU *et al.* 1995). In *Saccharomyces cerevisiae* S288C, there are 318 sites with Ty1 sequences – which include 279 solo LTRs, 2 truncated elements, and 32 full-length elements (CURCIO *et al.* 2015). *Saccharomyces paradoxus* (strain YPS138) has no full-length Ty1, 2, 3, 3_1p, 4, or 5 elements but does contain 22 full-length Tsu4 elements, 1 truncated Ty1, 232 Ty1/Ty2 solo LTRs, 10 Ty3 solo LTRs, 45 Ty3_1p solo LTRs, 49 Ty4 solo LTRs, 18 Tsu4 solo LTRs, and 34 Ty5 solo LTRs (BERGMAN 2018). Ty1 transposons replicate through an RNA intermediate via a copy-and-paste mechanism. The structure of a full-length Ty1 element includes 2 genes: GAG, which encodes the protein capsid, and POL, which encodes protease, integrase, and reverse transcriptase. Together, these transcripts compose ~5-10% of all poly-A RNA from expression of ~ 30 Ty1 elements in S288C (ELDER *et al.* 1981).

Many species suppress the activity and thus the number of copies of mobile genetic elements in the genome (HOLLISTER AND GAUT 2009). In yeast, *S. paradoxus* and *S. cerevisiae* utilize a mechanism known as copy number control (CNC) to limit transposition and thus the number of transposons in the genome. Previous studies have shown that the more transposons in the genome, the lower the rate of transposition (GARFINKEL *et al.* 2003). The evolution of such a mechanism suggests selection against excessive copy numbers of TEs - implying a deleterious impact of (many) TEs in the yeast genome. In addition to increasing the genome-wide insertion rate as copy number increases, another effect could be an increased mutation rate or change in spectrum not due to the transposition events themselves. In order to examine this possibility, a comparison of mutation rate and spectrum in a TE-free strain to that of a TE-containing strain is necessary.

There is little information on the effects of transposons on spontaneous mutations elsewhere in the genome that are unrelated to transposition events themselves. However, previous studies in bacteria have found evidence of an effect of transposon presence on mutation rate. Studies with insertion sequence (IS)-free bacteria showed reduced mutation rate using fluctuation and reporter gene assays (PÓSFAI *et al.* 2006; SUÁREZ *et al.* 2017). Suarez et al found reduced overall mutation rate in an engineered IS-free strain. The comparison included all mutations, where ~25% of them were due to IS movement in the control strain. The authors suggest that the differences in mutation rates are due to the lack of IS sequences. The location of transposon insertion had no effect on mutation rates and survivability, suggesting the presence of a transposon itself was the driving

factor and not potential sequences carried over from the donor strain or the effect of a particular insertion location.

In eukaryotes, a study in 2015 found that insertion of a Ty1 element into an *S. paradoxus* strain without endogenous Ty1s increased the time it takes for a cell to senesce in a nondividing state (its chronological lifespan) in oxidative-species rich media, but had no effect on mutation rate of the CAN1 reporter gene that confers resistance to canavanine (VANHOUTE AND MAXWELL 2014). This study suggests that Ty1 retrotransposons can have effects on yeast cells in stressful environments that are unrelated to transposition. One explanation is that Ty1s in the genome cause increased transcription of certain genes relevant to tolerance of reactive oxygen species by retrotransposing in proximity or by causing segmental duplications that increased the copy numbers of genes.

Determining the effects of mobile genetic elements on mutation rate and spectrum will enhance our overall understanding of mutation rates and the effects of these elements on evolution. Knowing how various factors alter mutation rates may be particularly useful in industrial settings where reduced rates are especially important. For example, in settings in which the homogeneity of the organism is essential for producing a certain compound, in vaccine development with attenuated viruses, when producing and utilizing genetically modified organisms, and to slow resistance to pesticides/antimicrobials/insecticides/etc.

In this study, we used 2 strains of *S. paradoxus*: 2 biological replicates of a haploid strain with 0 Ty1s (denoted TY-A and TY-B) and 1 haploid strain with 1 Ty1 element (TY+). We performed 200-day mutation accumulation experiments with ~35 lines of

each strain, for a total of ~1900 cell generations. We hypothesized that the large amount of transposon RNA (5-10% of all poly-A RNA) could stress cells, perhaps leading to increased mutation rate or a change in spectrum. We found no significant difference in single nucleotide mutation (SNM) or indel mutation rates between 0 and 1 Ty1-carrying strains but did find a significant impact on multinucleotide (several sites in close proximity) mutation (MNM) rates in these strains. In addition, we found a rate of retrotransposition in the 1-copy strain that is similar to previous estimates, with Ty1 elements preferentially inserting near or in genes that are transcribed by RNA polymerase III, as previously described (JI *et al.* 1993). Our results suggest an influence of TE presence on mutation rate and spectra in *S. paradoxus* and add to our understanding of the dynamics of transposable elements in eukaryotic genomes.

Methods

Mutation Accumulation

The initial ancestor strains, kindly gifted by Dr. David Garfinkel, were derivatives of *S. paradoxus* strain 337: DG1768 (referred to here as the 0-copy strain or TY-A/TY-B), a Ty1-less strain (*MATalpha his 3-*Δ*200hisG ura3 gal3* Spo-); and a strain containing one Ty1 element, DG4005 (referred to as 1-copy strain or TY+) (*MATalpha his3-*Δ*200hisG ura3 gal3 Ty1-4523* Spo-). The Ty1 in DG4005 (TY+) is on chromosome X between *RAD7* and *CDC8* and adjacent to a *Gly-tRNA* gene (GARFINKEL 2005). These original MA lines were the "ancestors" for all of the MA lines and were stored at -80°C. For each experiment, 48 initially-identical lines were produced from each of the ancestors and a mutation accumulation (MA) experiment was then carried out in the same manner

as previously described for *S. cerevisiae* (JOSEPH AND HALL 2004; ZHU *et al.* 2014) and *Sc. pombe* (BEHRINGER AND HALL 2016). Briefly, each ancestor was streaked onto solid YPD medium (1% yeast extract, 2% peptone, 2% dextrose, 2% agar) and incubated for 48 hours at 30°C. Subsequently, 48 randomly chosen colonies from each ancestor were transferred to new plates, for a total of 144 MA lines (96 for Ty1-less and 48 for Ty1-containing strain). Each of the three sets of 48 lines were each put through 200-days of passaging, streaking for single colonies, with a single-colony (assumed to have grown from a single cell) transferred every other day for a total of 100 transfers and approximately 1900 generations. MA lines were frozen in 15% glycerol at -80°C every 10 transfers. To screen for relatively common petite mutations (cells with defective mitochondria), which were not our interest, each line was streaked onto a YPG (1% yeast extract, 2% peptone, 3% glycerol, 2% agar) plate every 10 transfers. Petites do not grow on YPG so, if a line did not grow on this medium, a new colony from the previous transfer plate that was kept in the fridge was chosen at random to continue passaging.

Sequencing

After the MA experiment was completed, whole-genome sequencing was performed for all mutation accumulation lines plus the 3 original ancestors. One colony from each line was selected at random, inoculated into 2.5ml liquid YPD, and incubated on a rotator at 30°C for 24 hours. DNA was extracted using the Zymo YeaSTAR Genomic DNA kit protocol I with chloroform (Zymo Research). Whole-genome shotgun libraries were prepped using the protocol described previously, except without the bisulfite conversion step (URICH *et al.* 2015); in brief, gDNA was fragmented to 500bp,

bead-purified, adapters were ligated, and DNA was amplified using PCR primers that are specific to the adapters. MA lines were sequenced on a NovaSeq S4 flow cell PE 150 at Genewiz. Some MA lines were resequenced due to an error in the initial library preparation. The average coverage for each sample was ~300x, with the minimum being 90x.

Custom Reference Genome Production

To ensure we found only mutations that had arisen during mutation accumulation, we produced a custom reference genome and annotation. This was performed by Jingxuan Chen and Casey Bergman. In brief, HGAP assemblies from PacBio Sequel data of the TY-A ancestor (*S. paradoxus* strain 337) were Pilon polished with Illumina sequencing data generated from this study (WALKER *et al.* 2014). Four rounds of Pilon polishing were carried out, followed by scaffolding using RaGOO (ALONGE *et al.* 2019) and the YPS138 *S. paradoxus* reference genome. The resulting assembly was annotated using LRSDAY (YUE AND LITI 2018). Scripts for assembly, polishing, and annotation are located at https://github.com/bergmanlab/jingxuan/issues/12.

Quality control (QC), mapping, and identification of mutations

Quality control of reads was performed using fastQC (ANDREWS 2010) to determine quality of sequencing and identify any Illumina adapters that required removal. To identify variants, the Genome Analysis Toolkit (GATK) Best Practices were followed (DEPRISTO *et al.* 2011). In brief, an unmapped bam was produced from paired fastq files, this resulting bam file then had any remaining Illumina adapters removed, then this bam

was converted back to a fastq file, mapped to the reference using BWA, and merged with the unmapped bam to produce a mapped bam with all original information. This mapped bam file was then sorted, indexed, and duplicates were removed. After this, GATK's HaplotypeCaller was used in gVCF mode to produce VCF files for each of the MA lines. Eight gVCFs were chosen at random to recalibrate the bam files in order to remove any consensus variants (the ancestor was included in these MA lines in order to make sure ancestral variants were not present in the resulting VCFs). After recalibrated bams were created, HaplotypeCaller was run again, and the resulting gVCFs were combined together and jointly genotyped. A final VCF file with all MA lines and the ancestor was produced and subsequently filtered based on read depth, mappability, and the ancestor genotype. Depth cutoffs were determined by finding the average depth of the ancestor and using a Normal distribution. For a particular sequencing depth, if all parts of the genome were equally likely to be sequenced, we expect an approximate Normal distribution of depths for reasonably high sequencing depth. Assuming a Normal distribution we can determine the range of depths between which bases should fall. We calculated the depth range outside of which we expected not a single base to occur. Any base that had a depth that was outside of this range in each of the ancestor strains was discarded. The resulting VCFs were used in the final analysis.

During analysis, we noticed that the vcf genotype caller was somewhat unpredictable in how it called the genotype at particular bases. Genotype calls were thus redone using the frequency of each allele at each site. Because strains were haploid, any allele frequency less than 0.10 or greater than 0.90 was deemed to have the common allele. Sites with intermediate allele frequencies in the ancestors were discarded as such allele

frequencies are not possible in a haploid strain and thus might indicate the presence of a duplication. Sites in MA lines with allele frequencies that were intermediate for the frequency of a new mutation were examined one by one using the genome viewer.

New mutations were also used to identify any cross-contamination of MA lines during MA. If two MA lines shared one or more mutations that must be the result of contamination since the probability of an identical mutation occurring at the same site in two lines is very low (less than 10^{-16}). When contamination was identified, the higher numbered MA line was discarded from analysis as a likely cross contaminant. After contaminants were removed, the rate and spectrum of mutations was determined. All scripts can be found at https://github.com/hollygene/TE_MA.

For single nucleotide mutations (SNMs), the trinucleotide context was determined. To do so, the trinucleotide context for every possible site in the genome was found using the Biostrings R package function *trinucleotide frequency* (PAGÈS *et al.* 2020). A .bed file of the SNPs in each strain was produced using vcftoBed from the BEDOPS program (NEPH *et al.* 2012). The range of nucleotides was increased by adding or subtracting 1 on either side of the SNP position using an *awk* command in bash shell. To find trinucleotides of each 3bp position, getFASTA was used from the BEDTools package (QUINLAN AND HALL 2010). We then sorted the resulting trinucleotides into the 32 possible trinucleotides (32, not 64, because each trinucleotide also includes its complement on the other strand, i.e. AAA is also TTT) and then divided the number of mutations in a particular trinucleotide context by the frequency of that trinucleotide in the reference genome. Standard error was calculated by using the standard deviation of the number of mutations per line in each context divided by the square root of the number of MA lines,

divided by the number of generations, divided by the number of sites of that particular context. Trinucleotide frequency plus or minus 2 standard errors were then plotted in R using the ggplot2 package (WICKHAM 2016).

Ty1 Location and Movement

The McClintock package was used to identify and locate TE sequences within our sample genomes (NELSON *et al.* 2017) (analysis by Jingxuan Chen). We specifically used TELocate data from this analysis because we wanted to know where the Ty1 elements moved within the 1-copy strain. Briefly, reads were filtered for quality greater than 20 using Trim Galore (KRUEGER 2012), and McClintock was run with default parameters using the TE library produced from the 337-reference genome (scripts can be found at https://github.com/bergmanlab/jingxuan/blob/master/src/shell/holly_sep_rep.sh). Bar plots were produced for each strain to visualize the amount of Ty1 and LTR sequences in each sample using the R package ggplot2 (Supplemental Figures 3.3 and 3.4, Appendix II) (WICKHAM 2016). Locations of Ty1 elements in TY+ samples were identified and visualized using the Integrative Genomics Viewer (ROBINSON *et al.* 2011) (Supplemental Figure 3.5, Appendix II).

Results

Mutation Accumulation lines

We used the number of cells in an average colony (N_{cells}) at the time of transfer to calculate the number of cells generations (G) by solving $2^G = N_{cells}$. The number of cell generations was then used to calculate the harmonic mean (H) of the number of cells in a

colony (and thus the number in an MA line since every round of colony growth is assumed to be identical):

$$H = \frac{1}{\frac{1 + \sum_{i=1}^{G} \frac{1}{2^i}}{(1+G)}}$$

In TY-B and TY-A, 18.8 generations occurred between transfers, which represents an effective population size of 9.9 cells. In TY+, 19.2 generations occurred between transfers, which represents an effective population size of 10.1 cells.

Differences between 337 reference genome and each ancestor

We found few differences between the 337-reference genome and each ancestor (Supplemental Table 3.1). The TY-A ancestral genomic sequencing data generated from this study was used to construct the reference (see Methods), so we only looked for differences in the TY-B and TY+ ancestors. TY+ contained 4 ancestral variants, and TY-B contained 2. These were located within genes but did not cause any amino acid changes and thus we were not concerned of their effects on our data (Supplemental Table 3.1, Appendix II). We eliminated these from analysis before calling variants.

Intermediate mutation frequencies and contamination

Sites in MA lines with allele frequencies that were intermediate for the frequency of a new mutation were examined one by one using the genome viewer. In most cases, it was clear that such sites were the result of deletions in MA lines, which led to low depth (just a few reads incorrectly mapping to those regions). In some cases it could not be discerned why an intermediate allele frequency was obtained, but could perhaps be due to

mutation, though the depth was not particularly higher. In all cases an intermediate frequency site in an MA line was discarded. Shared newly-arising mutations were used to identify cross-contamination of MA lines. We found several contaminants, reducing the number of MA lines for analysis from 48 to 33 in TY+ (1-copy strain), to 39 in TY-B (0-copy strain), and to 36 in TY-A (0-copy strain).

Aneuploidy Events

We found one whole-chromosome disomy event of chromosome VII in line 43 of the TY-A strain, giving an estimated aneuploidy rate of 9.23×10^{-7} events/chromosome/cell generation (Figure 3.10). No aneuploidy events were detected in the other two strains. Given the rarity of aneuploidy events, it is not possible for us to determine whether the rates differ in a strain carrying a Ty1 transposon.

Small indels

Small indels were defined as mutations that were 1-50 bp in length, following previous definitions (LYNCH *et al.* 2008; ZHU *et al.* 2014; BEHRINGER AND HALL 2016). For TY-A (0-copy strain) MA lines, 12 indels across 36 lines were found, making the estimated indel mutation rate in this strain to be $1.48 \times 10^{-11} \pm 4.33 \times 10^{-12}$ indels/base/generation. The average insertion size was 5.375 and the average deletion size was 13.25. The combined insertions and deletions resulted in a net loss of 10 nucleotides across all TY-A lines during the experiment, which is not significantly

different from zero. For TY-B (0-copy strain) lines, 18 indels across 39 lines were found, for an estimated indel mutation rate of $2.05 \times 10^{-11} \pm 5.33 \times 10^{-12}$ indels/base/generation. The average insertion size was 6bp and the average size of deletions was 6.18bp. There was a net loss of 26 nucleotides across all lines during the experiment, which was not significantly different from zero. The TY+ (1-copy strain) lines had 15 indels across 33 lines, giving an estimated indel mutation rate of TY+ $1.98 \times 10^{-11} \pm 6.06 \times 10^{-12}$ indels/base/generation. The average insertion size was 16.4bp and the average deletion size was 6.75bp. There was a net gain of 9.7 nucleotides across all lines during the experiment for this strain, which was again not significantly different than zero. Given that the number of mutations per line follows a Poisson distribution (Supplemental Figure 3.2, Appendix II), we used the Rate Ratio Test R function to test for significant differences between the three experiments (FAY 2010). We found that the indel rates did not significantly differ from each other (pval > 0.1, Rate Ratio Test), and the indel rates of TY-A and TY-B datasets were nearly identical (pval > 0.8, Rate Ratio Test) (Figure 3.2).

Curious as to whether the indels in the 1-copy strain were caused by Ty1 endonuclease activity, we investigated where the indels mapped and if they were near tRNA genes or genes that were transcribed by RNA polymerase III, as Ty1 integrase has been shown to preferentially cut near these genes (EIGEL AND FELDMANN 1982; HANI AND FELDMANN 1998; KIM *et al.* 1998). We found only 5 out of the 15 indels in TY+ samples mapped in or near genes; the rest were intergenic. Of these 5, 3 were located in genes that were transcribed by RNA polymerase III, but none mapped in or near tRNA genes (Supplemental Table 3.4, Appendix II). If these were not facilitated by Ty1, we

would expect them to occur in RNA polymerase III-transcribed genes only 4% of the time (as 4% of the genome is transcribed by RNA polymerase III (ROBERTS *et al.* 2003)), giving us an expected number of indels in RNA polymerase III of less than 1 (0.6), which is not significantly different than what we see (p=0.604, Fisher's exact test). This suggests that the indels in our 1-copy strain are not likely facilitated by Ty1 transposition/integrase.

Single-nucleotide mutations

Single-nucleotide mutations (SNMs) are substitutions at a single site isolated in location from other such substitutions. In the TY-A (0-copy) MA lines, 152 SNMs were found across 36 lines, an average of 4.22 SNMs/line, for an estimated SNM mutation rate of $1.87 \times 10^{-10} \pm 1.80 \times 10^{-11}$ SNMs/base/generation. The TY-B (0-copy) MA lines had 153 mutations across 39 lines, an average of 3.44 mutations/line, for an estimated rate of $1.74 \times 10^{-10} \pm 1.86 \times 10^{-11}$ SNMs/base/generation. In the TY+ (1-copy) strain, there were 155 SNMs across 33 lines, an average of 4.69 mutations/line, giving an estimated mutation rate of $2.04 \times 10^{-10} \pm 1.74 \times 10^{-11}$ SNMs/base/generation. We found no significant difference in SNM rate among the 3 strains, though TY+ did display a slightly higher rate than TY-B or TY-A (pval > 0.1, Rate Ratio Test, Figure 3.1).

Double mutations, complex mutations, and structural variants

We also looked at multinucleotide mutations (MNMs) in our MA lines. Following previous work (SHARP *et al.* 2018), we defined these as any mutation (SNP or indel)

within 50 bp of another mutation. We found 8 MNMs in TY-A MA lines that involved a total of 94 single nucleotide changes, giving the MNM rate as $6.2 \pm 3.1 \times 10^{-12}$ MNMs/base/generation. The average length of an MNM in TY-A MA lines was 18.8 nucleotides. In TY-B MA lines, there were 9 MNMs across 39 lines that involved a total of 81 single nucleotide changes, for a rate of $1.0 \times 10^{-11} \pm 4.5 \times 10^{-12}$ MNMs/base/generation. Notably, 3 out of the 9 MNMs in TY-B were from one line (Ty-B 39). The average length of MNMs in TY-B MA lines was 9.67 bp.

In contrast, TY+ MA lines had 21 MNMs across 33 lines that involved a total of 413 single nucleotide changes, for a rate of $2.77 \pm 1.4 \times 10^{-11}$ MNMs/base/generation. Notably, one sample had 10 of the MNMs (Ty+ 20), 4 of which were on chromosome XII, and accounted for 220 of the single nucleotide changes. The MNM rate was statistically significantly different between the TY-A and TY+ ($p=0.008$, Rate Ratio Test) strain and the TY-B and TY+ strain ($p=0.027$, Rate Ratio Test) but not between the TY-A and TY-B strains ($p=0.867$, Rate Ratio Test). However, if the TY+ sample with 10 of the 21 MNMs is removed, the results are no longer statistically significant (TY-A v. TY+: $p=0.435$, TY-B v. TY+: $p=0.710$). This might suggest there is another factor influencing MNM rate besides TEs (any mutation that arises may have an impact on mutation rate in the cells). Further studies are needed to understand the basis of these multinucleotide mutations.

We wanted to know if there was any evidence that MNMs in TY+ lines were facilitated by Ty1 movement. Ty1 preferentially inserts upstream of tRNA genes or those that are transcribed by RNA pol III (Ji *et al.* 1993). We analyzed the regions where the MNMs were located and found that 3 of the 21 MNMs were located near genes that were

transcribed by RNA pol III (14% of MNMs) (Supplemental Table 3.3, Appendix II). In contrast, 4% of the yeast genome is transcribed by RNA polymerase III (ROBERTS *et al.* 2003). If the MNMs were not facilitated by Ty1, we would expect less than 1 MNM to be near RNA polymerase III. This suggests that a significantly larger portion of MNMs are likely facilitated by Ty1 transposition/integrase activity than would be expected by chance ($p=5.3 \times 10^{-8}$, Fisher's exact test).

Ty1 element movement in TY+ strain

Using the McClintock pipeline, we generated data for the locations of the Ty1 elements in our samples. We used the counts of Ty1 elements in each sample from TELocate data and our variant calling information to calculate an estimated rate of transposition of Ty1 elements in the TY+ genome. We found a total of 86 Ty1 insertions using TELocate data for a transposition rate of $1.3 \times 10^{-3} \pm 1.6 \times 10^{-4}$ transpositions/generation. This estimate does not control for increases in copy number of Ty1s over time, and thus is an overestimate of the rate per element. These insertions were not evenly distributed across lines (Supplementary Figure 3.7) or chromosomes (Supplementary Figure 3.6); one line did not have any Ty1 insertions, and 4 lines had only 1 Ty1 insertion.

It is notoriously difficult to call repetitive elements using short-read data (TREANGEN AND SALZBERG 2012). We were interested in knowing how accurate TELocate, the tool we used to find Ty1 insertions, was at finding Ty1 insertions, to achieve a baseline of sorts for our confidence in locating the Ty1 elements that transposed. The TY+ ancestor was engineered with 1 Ty1 element inserted on

chromosome X (see Methods for details). In order to double-check TELocate's accuracy, we looked at the location of this Ty1 in all TY+ samples using the Integrative Genomics Viewer (IGV) (ROBINSON *et al.* 2011). We found that most samples did have a Ty1 element in the region where it should have been, but some were up to 1 kb away from the insertion site or missing entirely (Supplementary Figure 3.5). Given these results, we are cautious about our Ty1 insertion results as the actual locations may be quite different.

We wanted to know where the Ty1 elements inserted into the genome and if they were located in regions that are expected (i.e. near tRNA genes or genes that are transcribed by RNA polymerase III (JI *et al.* 1993)). Using the TELocate data, we plotted the locations of Ty1 elements in TY+ samples on each chromosome (Figure 3.12). We also investigated what genes were in close proximity (within 1kb downstream) of these Ty1 insertions, and whether these were what we would expect. We found that out of 86 transposition events, 45 were intergenic, 23 were upstream of a gene (<1 kb away), and 18 were within genes (Supplemental Table 3.3, Appendix II). Given the recent pandemic, we were unable to run a PCR in the lab to confirm any of these sites. If the Ty1 locations are mapping correctly, of those that were upstream or lied within genes (41), 22 of those genes were transcribed by RNA polymerase III (29%). Only 4% of genes are transcribed by RNA polymerase III – we would expect <2 of the 41 insertions to be in/near genes transcribed by RNA polymerase III if the insertions were random, which is significantly different than what we see ($p=0.000256$, Fisher's exact test).

Using HaplotypeCaller data, we also found 20 indels of length 136bp in TY+ MA lines. Though they were all the same length, the sequences were not identical. However, they all mapped to Ty1 LTRs (Supplemental Table 3.3, Supplemental Figure 3.2,

Appendix II). We can therefore categorize these as "transposition attempts" rather than "transposition events." We compared the indels mapping to LTRs to TELocate data and found that 11 of the 20 we found through variant calling also matched full-length elements found using TELocate (NELSON *et al.* 2017) (Supplemental Figure 3.2, Appendix II). We are more confident about the locations of these elements since they have been confirmed with two different computational methods, though, again, we did not have the chance to obtain PCR data for these locations. We were curious if these LTRs had signatures of transposition by Ty1, so we investigated where they mapped and found that all 20 of them either were located upstream, downstream, or in a snoRNA or tRNA gene or a gene transcribed by RNA polymerase III (Supplementary Table 3.3, Appendix II). This suggests that these indels were indeed caused by Ty1 transposition attempts.

SNM mutation spectrum and biases

Mutation frequency per mutation type was normalized to the GC content of the genome (38% GC, 62% AT). Across all strains, C:G→T:A mutations had the highest frequency of SNMs. In TY-A MA lines, C:G→A:T was the next highest category. In both the TY-B and TY+ lines, the next highest category was A:T→G:C followed by C:G→A:T. If the mutation rate is the sole driving factor of G/C content, in TY-A we would expect the G/C content to be 37%, in TY-B we would expect 49%, and in TY+ we would expect 53%. The expectation for TY-A is almost exactly what we see; however, the expectations for TY-B and TY+ are substantially higher than the observed amount. We found that C:G sites mutated at significantly higher rates than A:T sites ($p=0.01155$,

two-sample t-test, Figure 3.4). We found no differences between strains in spectrum of mutations (p=0.976, ANOVA, Figure 3.4).

We found transition to transversion bias in each of our strains. If all mutations were equally likely, then transversions would outnumber transitions 2 to 1. TY-A, TY-B, and TY+ samples had a transition to transversion (Ts/Tv) ratio of 0.71, 1.15, and 1.01 respectively. This range of estimates (0.71 – 1.15) is inclusive of previous estimates in other yeast species, which have found Ts/Tv ratios of 0.95 (*S. cerevisiae*, Zhu 2014), 0.72 (*Sc. pombe*, Behringer 2016), and 0.82 (*S. cerevisiae*, Sharp 2018) (ZHU *et al.* 2014; BEHRINGER AND HALL 2016; SHARP *et al.* 2018).

We next examined the trinucleotide context of mutations. The trinucleotide context was found for every base in the genome and for each SNM (see Methods for details). We plotted the mutation rate at each trinucleotide, relative to its frequency in the genome, with the standard error of the number of mutations at that trinucleotide per line. We found that sites with A or T as the mutated base were significantly less likely to mutate than sites with a C or a G (p=0.000554, two-sample t-test). We found the same pattern in TY-B and TY+ strains (p=0.00612 and p=0.0018, respectively, two-sample t-test).

In TY-A MA lines, ACG (equivalent to CGT) and GCT (equivalent to AGC) trinucleotides were the most frequently mutated. The C in ACG is a CpG site. In previous studies CpG sites have been found to mutate at a higher frequency, even in species lacking DNA methylation such as *S. paradoxus* (BEHRINGER AND HALL 2016). We were curious as to whether there was a statistical difference in the mutation rates between the C in CpG sites and the C in GpC sites and found that there was no significant difference (TY-A: p=0.377, TY-B: p=0.307, TY+: p=0.253, two-sample t-test, Figure 3.5). In TY-

B, TCG was the most frequent to mutate, followed by GCG and ACG, all of which are CpG sites. TY+ MA lines showed the most dramatic example of this bias, with all 4 CpG sites showing higher mutation rates than others, but this was not significantly different than TY-A or TY-B strains (TY-A v. TY-B: p=0.363; TY-A v. TY+: p=0.151; TY-B v. TY+: p=0.436; two-sample t-test).

Comparison of mutation rates in this study to previous studies and other species

We compared our estimates to other MA studies of *S. cerevisiae* from a diploid MA experiment, *S. cerevisiae* from an experiment involving both haploid and diploid MA, and *S. pombe* MA (ZHU *et al.* 2014; BEHRINGER AND HALL 2016; SHARP *et al.* 2018). Compared to *S. cerevisiae* haploid MA data (SHARP *et al.* 2018), our mutation rates are relatively low, but were not significantly different (p>0.5, ANOVA, Figure 3.6). *S. cerevisiae* diploid MA data as well as *S. pombe* data were more similar to our rates (Figure 3.6) (ZHU *et al.* 2014; BEHRINGER AND HALL 2016; SHARP *et al.* 2018). The difference could be explained by different media conditions during MA: Sharp et al added adenine sulfate to the YPD media; this study and the other *S. cerevisiae* and *Sc. pombe* studies did not (ZHU *et al.* 2014; BEHRINGER AND HALL 2016; SHARP *et al.* 2018).

Discussion

We performed a 200-day mutation accumulation experiment with two different strains of *Saccharomyces paradoxus* differing in the presence of a Ty1 element to determine if transposable element presence in the genome affects mutation rate and/or

spectra. In our study, we characterized a total of 630 spontaneous mutations, including 460 SNMs, 38 MNMs, 45 indels, 1 whole-chromosome disomy, and 86 transposition events.

Rate of mutation at single nucleotides

We found no statistically significant difference among the 3 strains in rate of SNMs, though there was a trend toward a slightly elevated rate in TY+ MA lines (Figure 3.1). The presence of a transposable element thus seems to have no effect on the rate of SNMs. A previous study in *Actinetobacter baylyi* comparing an insertion-sequence-free strain and a wildtype strain in a reporter gene assay found similar results, with no significant effect on the point mutation rate (SUÁREZ *et al.* 2017). Based on our study and theirs, it appears that transposon presence does not have a significant effect on SNM mutation rate in haploid eukaryotic or prokaryotic genomes.

Spectrum of single nucleotide mutations

In TY-A MA lines, C:G → T:A mutations were the most frequent, followed by C:G →A:T mutations. This is similar to previous findings of C:G sites mutating more frequently than A:T sites (ZHU *et al.* 2014; BEHRINGER AND HALL 2016; SHARP *et al.* 2018). There was no significant difference between the 0-copy and 1-copy Ty1 strains on spectrum of SNMs as a whole ($p=0.976$, ANOVA; Figure 3.4). The expected equilibrium G/C frequency for TY-A is 37% G/C, almost exactly the same as the observed frequency for *S. paradoxus*. However, for the TY-B and TY+ strains, we calculated expected equilibrium G/C frequencies of 49% and 53%, respectively, which

are both substantially higher than the observed G/C content for the reference strain. These discrepancies suggest that there is another factor influencing G/C content in *S. paradoxus* aside from mutation pressure. However, given the differences between biological replicates here, there may have been experimental errors in the initial production of strains (i.e. non-homogeneous cell populations); from our data we cannot determine if 37% G/C or 49% G/C is the true expected equilibrium frequency. If the G/C equilibrium frequency in a 0-copy strain is 37% and not 49%, this would imply that the presence of transposable elements in the genome increases the G/C equilibrium frequency substantially; if the reverse is true, then transposable elements would seem to not have an impact on G/C equilibrium frequency. Future studies would benefit from performing a similar experiment to ours, ensuring that biological replicates are identical in genotype and possibly also using a diploid strain for comparison in addition to a strain with the wild-type load of transposable elements, to assess if transposition or copy number control is impacting G/C equilibrium frequency.

Context-specific SNM rates

Similar to previous studies, we found an effect of trinucleotide context on mutation rates (ZHU *et al.* 2014; BEHRINGER AND HALL 2016; SHARP *et al.* 2018) (Figure 3.5). CpG sites mutated at a higher frequency than non-CpG sites (Figure 3.5), though there was no statistically significant difference in CpG versus GpC mutation rates ($p>0.1$, see Results and Figure 3.5). It is well known that DNA methylation has an impact on mutation rate from C:G to T:A at CpG sites because deamination of a methylated C gives a T base (FREDERICO *et al.* 1993; XIA *et al.* 2012), but methylation seems unlikely to be a

driving factor in this species, as DNA methyltransferases not been found in *S. paradoxus* (CAPUANO *et al.* 2014). It is possible that the intrinsic mutability of CpG sites makes them perfect targets for the evolution of methylation, instead of DNA methylation causing the high mutability of CpG sites. Further studies in organisms without DNA methyltransferases, or in organisms engineered to have nonfunctional methyltransferases, would be useful to investigate this phenomenon further.

Rate and spectrum of indels

There was a significant difference between the TY-A strain and the TY+ strain in indel rate (Figure 3.2), with TY+ having a higher indel mutation rate. TY+ and TY-B indel rates were not statistically significant, though TY+ was higher. Both TY-A and TY-B strains experienced net nucleotide loss (10 bp and 26 bp, respectively), whereas the TY+ strain experienced a net nucleotide gain (9.7 bp). The increase in genome size combined with the higher indel rate of the Ty1-containing strain might suggest that the presence of TEs causes genomes to expand, even without including transposition events, but this study does not have enough data to show a significant difference.

We found that 5 out of the 15 indels in TY+ (1-copy) samples mapped in or near genes and of these, 3 indels mapped in genes that were transcribed by RNA polymerase III (Supplemental Table 3.4, Appendix II). We found that this was not significantly different than would be expected by chance ($p=0.604$, Fisher's exact test). This would suggest that these mutations are occurring independently of transposition events and/or integrase cutting. In similar previous studies, indel rates were increased in bacterial strains with more mobile genetic elements (SUÁREZ *et al.* 2017). Many species have

undergone genome expansion driven by mobile genetic elements (FESCHOTTE AND PRITHAM 2007), but our data suggest the transposition events themselves may not be the only factor in this expansion. A possible mechanism for this is if an endonuclease from a TE cuts the genome somewhere, but instead of a TE inserting, non-homologous end joining occurs, a repair mechanism known to cause indels (RODGERS AND MCVEY 2016).

Complex mutations, double mutations, and aneuploidy

We found that the TY+ strain had a statistically significant higher rate of multi-nucleotide mutations (MNMs) (Figure 3.3). One explanation for why MNMs are more frequent in the Ty1 containing TY+ strain is that they are due to gene conversion events (HICKS *et al.* 2010), which are known to produce runs of mutation events in close proximity to one another (HICKS *et al.* 2010). These gene conversion events may be facilitated by Ty1 elements (i.e. cutting by integrase), as previous studies have associated gene conversion with transposable elements (FAWCETT AND INNAN 2019). Future studies with heterozygous diploid strains would be useful to more accurately detect gene conversion events.

Transposition events in TY+ strain

We identified transposable elements and LTRs in our TY+ MA lines using McClintock and HaplotypeCaller (DEPRISTO *et al.* 2011; NELSON *et al.* 2017). With short-read data, identifying the locations of highly repetitive sequences is challenging (TREANGEN AND SALZBERG 2012). However, our estimate of transposition rate in this strain, $1.3 \times 10^{-3} \pm 1.6 \times 10^{-4}$ insertions/generation are not significantly different than

those measured using a plate-based assay from a single-copy strain closely related to our ancestral strain which was measured at 1.16 x 10^{-4} insertions/generation (MOORE *et al.* 2004). This suggests that the two methods for estimating rate of transposition (sequencing and plate-based assays) are comparable, even with the challenges of short-read sequencing, though the difficulty with short-read sequencing lies within finding locations of the elements. We found that, if our location data is accurate, a larger percentage of transposition events than would be expected by chance occurred near or in genes that are transcribed by RNA polymerase III, which supports previous data (EIGEL AND FELDMANN 1982; HANI AND FELDMANN 1998; KIM *et al.* 1998). This suggests that Ty1 elements in this strain are functioning normally, and the mutations they may have caused can be extrapolated to other scenarios. In addition, we found 20 indels of the same length (136bp) that all mapped to LTRs, and these were located in or near genes that were likely to be targeted by Ty1 (Supplementary Table 3.3), suggesting that Ty1 attempts to insert but is prevented from doing so, and instead either undergoes a recombination event or facilitates a double-strand break that is repaired by homology-directed repair.

Comparison to previous MA studies in yeast

We compared our results to two previous studies in *Saccharomyces cerevisiae* and one previous study in *Schizosaccharomyces pombe*. Our results are generally consistent with what has been observed in previous studies, though the mutation rates of the Sharp et al 2019 *Saccharomyces cerevisiae* study are significantly higher than those measured in this study and in the others – this may be due to differences in experimental

methods: Sharp et al used YPD media supplemented with adenine sulfate. Conversely, this study, the other *S. cerevisiae* studies, and the *Sch. pombe* study used YPD media with no added adenine sulfate.

Comparing MNM rates between our study and those of Sharp et al 2019 displayed noticeable differences – our 1-copy (TY+) strain had a significantly higher MNM rate than either of our 0-copy strains or the diploid or haploid strains in the Sharp et al study. This could be indicative of transposable element activity, as the strains used in Sharp et al likely contained the wild type numbers of Ty1 elements for *S. cerevisiae* and thus was under strong copy number control. This would decrease the amount of transposition and/or integrase activity in the genome, as copy number control limits the amount of Ty1 transcript produced and therefore the amount of integrase protein produced. This would suggest that multinucleotide mutations are facilitated by transposition activity and/or integrase activity in yeast genomes.

Conclusions and future directions

Our study supports the hypothesis that transposons play a role in mutations unrelated to transposition events as evidenced by the rate of multinucleotide mutations being significantly increased in the strain with one Ty1 element. However, these mutations may have been facilitated by the transposons – the transposition event may have failed, but the mutation was still possibly caused by Ty1 integrase. We found no significant difference between strains in SNP or indel rates, likely due to biological

differences, though our study lacked power to determine with confidence that indel rates are not different between strains. We found a similar rate of transposition to previous studies but acknowledge that our mapping and location data may not be accurate, as calling repeats using short-read sequencing is notoriously difficult. We found that Ty1s likely attempt to transpose but are sometimes thwarted, leaving mutational scars in their wake in the form of LTRs or MNMs.

Our results also recapitulate previous studies that CpG sites mutate at higher frequencies than non-CpG sites. However, we did not find a significant difference between rates of mutation at CpG versus GpC sites, suggesting that a C near a G is enough to cause high mutability, regardless of the order. Because our strains are haploid, we did not pick up any recessive lethal mutations that may have been caused by transposon presence. Future studies will need to utilize diploid strains to discern if there is any difference when recessive lethal mutations are masked by diploidy. A higher sample size in each strain would be useful to help mitigate the impacts of variance between samples and achieve more power for identifying differences between strains, and future studies would benefit from using a strain with a higher copy number of Ty1 elements to assess the impacts of copy number control on mutation rate and spectrum.

Figures

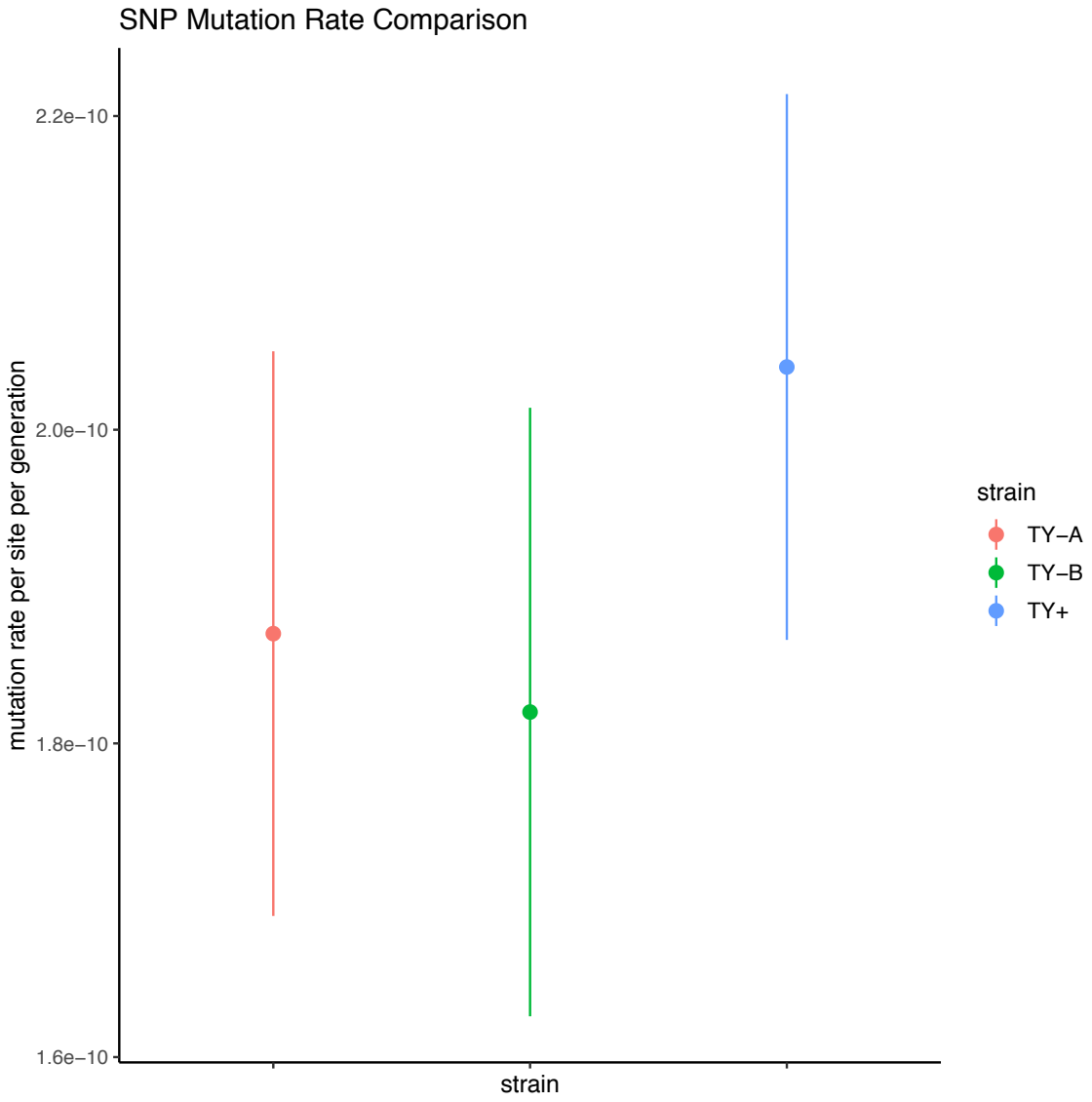


Figure 3.1: Single-nucleotide polymorphism mutation rates of the strains analyzed in this study. TY-A and TY-B are biological replicates of the 0-copy strain and TY+ is the 1-copy strain. D1 shows a slightly elevated SNP rate but is not statistically significant, due to the large standard error in this strain. Error bars are +/- 1 SE.

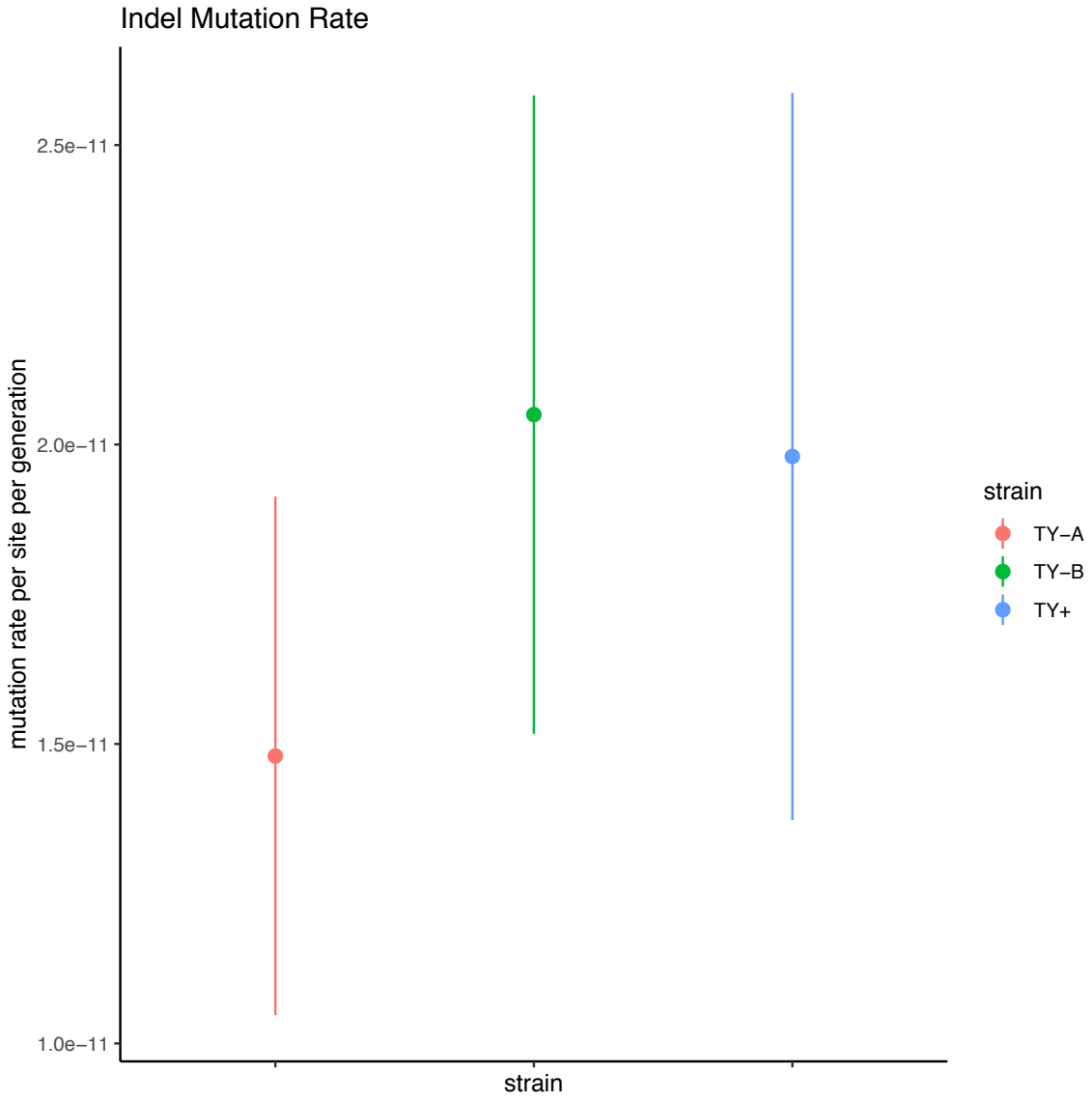


Figure 3.2: No significant difference of indel rates between strains. Error bars are +/- 1 SE. TY-A and TY-B are biological replicates of the 0-copy strain and TY+ is the 1-copy strain.

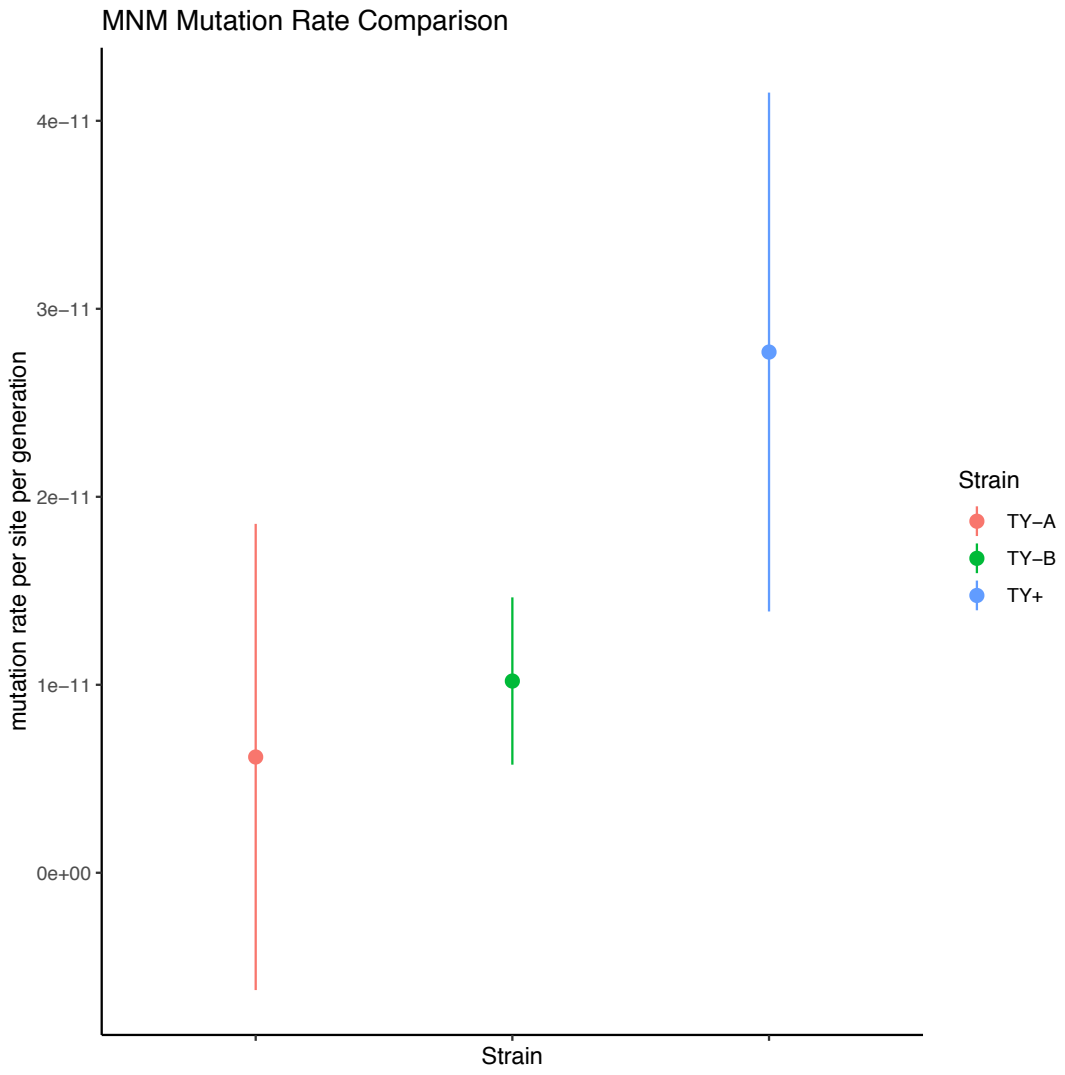


Figure 3.3: MNM rates are significantly higher in the TY+ strain compared to both the TY-B and TY-A strains. TY-A and TY-B are biological replicates of the 0-copy strain and TY+ is the 1-copy strain.

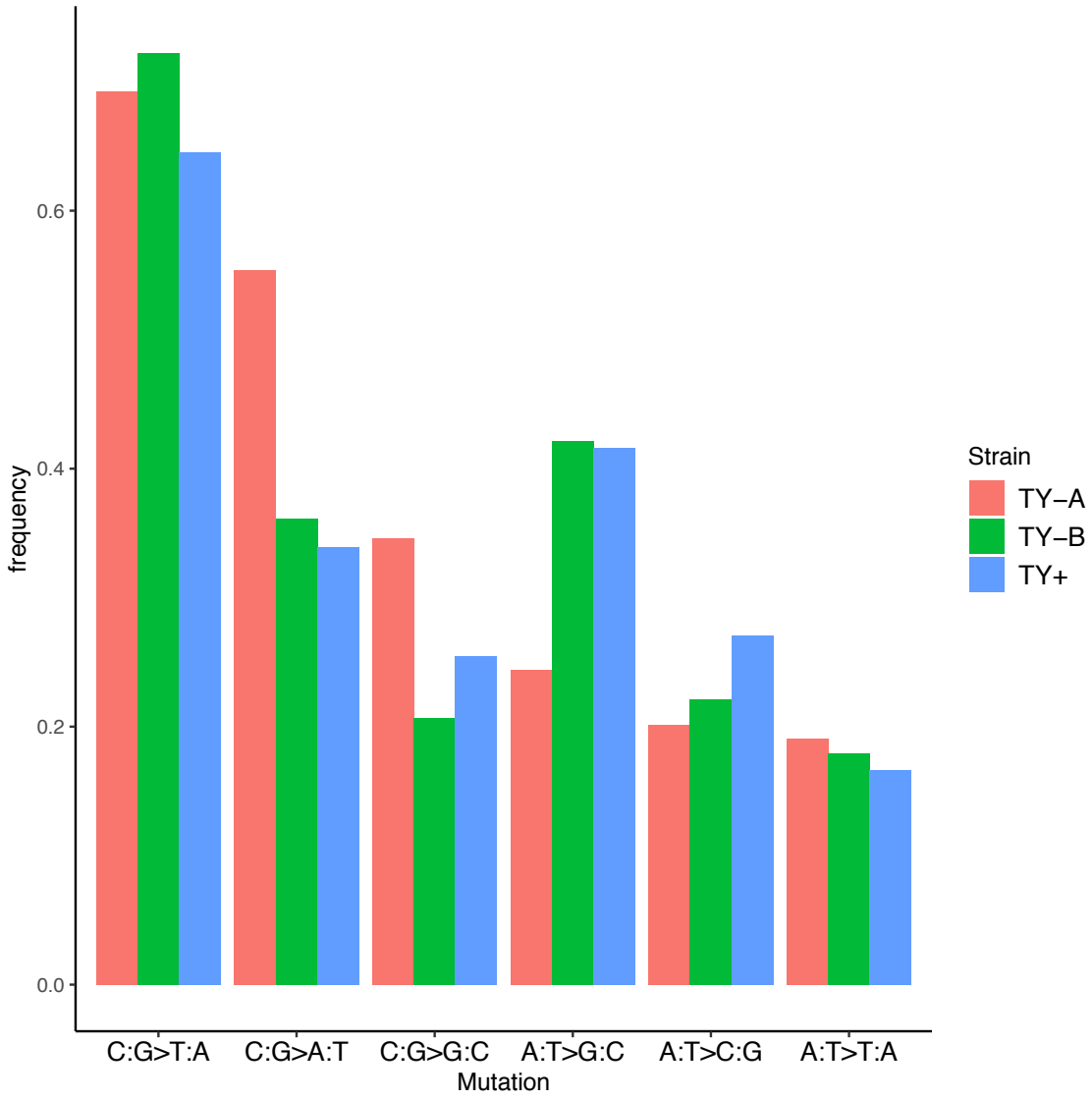
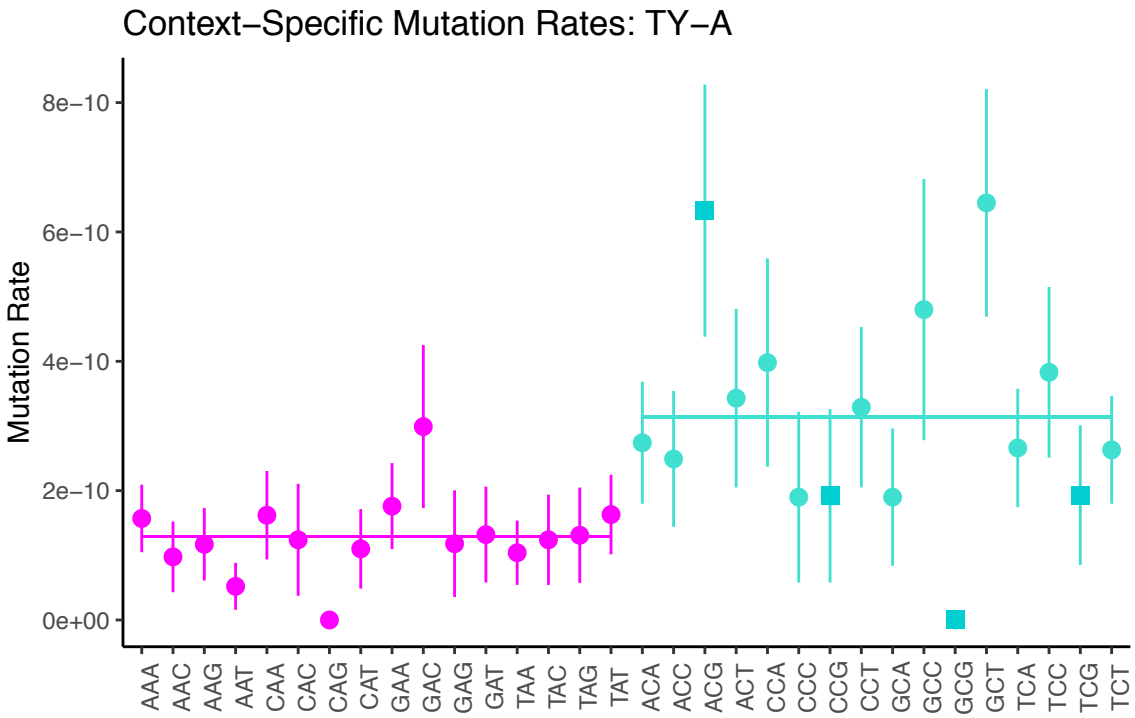
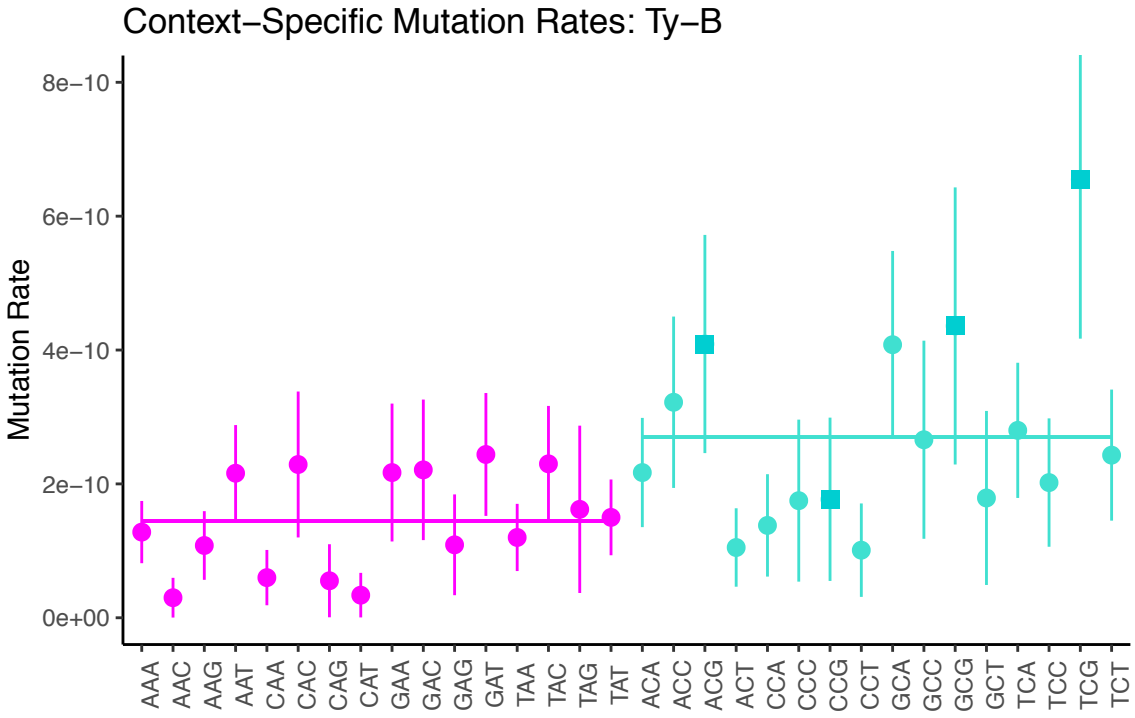


Figure 3.4: Spectrum of mutations in each strain, corrected for G/C content of the genome. C:G>A:T and C:G>T:A mutations were the two highest categories. TY-A and TY-B are biological replicates of the 0-copy strain and TY+ is the 1-copy strain. C:G sites mutate at significantly higher frequencies than A:T sites (p=0.01155, two-sample t-test). There was no significant difference between strains in spectrum of mutations (p=0.976, ANOVA).


Context−Specific Mutation Rates: TY−A
Mutation Rate
8e−10
6e−10
4e−10
2e−10
0e+00
AAA AAC AAG AAT CAA CAC CAG CAT GAA GAC GAG GAT TAA TAC TAG TAT ACA ACC ACG ACT CCA CCC CCG CCT GCA GCC GCG GCT TCA TCC TCG TCT


Context−Specific Mutation Rates: Ty−B
Mutation Rate
8e−10
6e−10
4e−10
2e−10
0e+00
AAA AAC AAG AAT CAA CAC CAG CAT GAA GAC GAG GAT TAA TAC TAG TAT ACA ACC ACG ACT CCA CCC CCG CCT GCA GCC GCG GCT TCA TCC TCG TCT

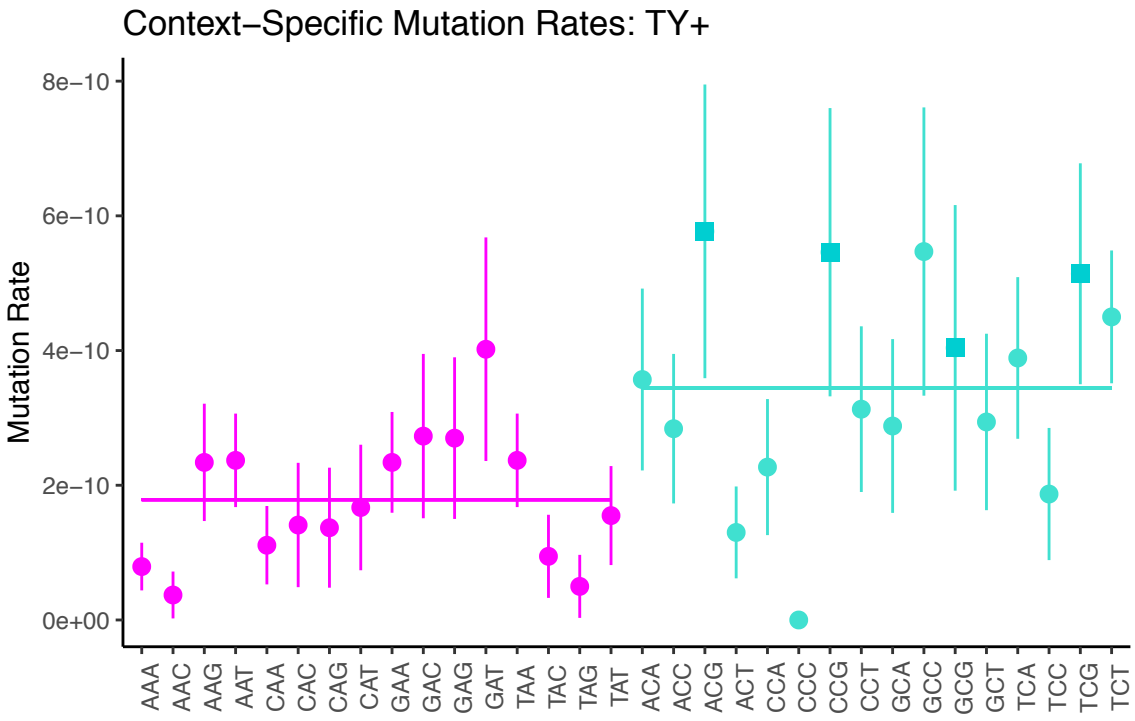


Figure 3.5: Context-specific mutation rates show higher mutation at CpG sites for TY-B and TY+ strains . Magenta colored points are those sites with a A/T as the middle position, and turquoise points are those sites with a C/G as the middle position. Squares indicate CpG sites. Horizontal lines in magenta and turquoise represent the averages for sites with an A/T and G/C, respectively. TY-A and TY-B are biological replicates of the 0-copy strain and TY+ is the 1-copy strain. A:T sites (magenta) are significantly less mutable than C:G sites (turquoise) (H0: p=0.00056, D0: p=0.0061, D1: p=0.0018; two-sample t-test).

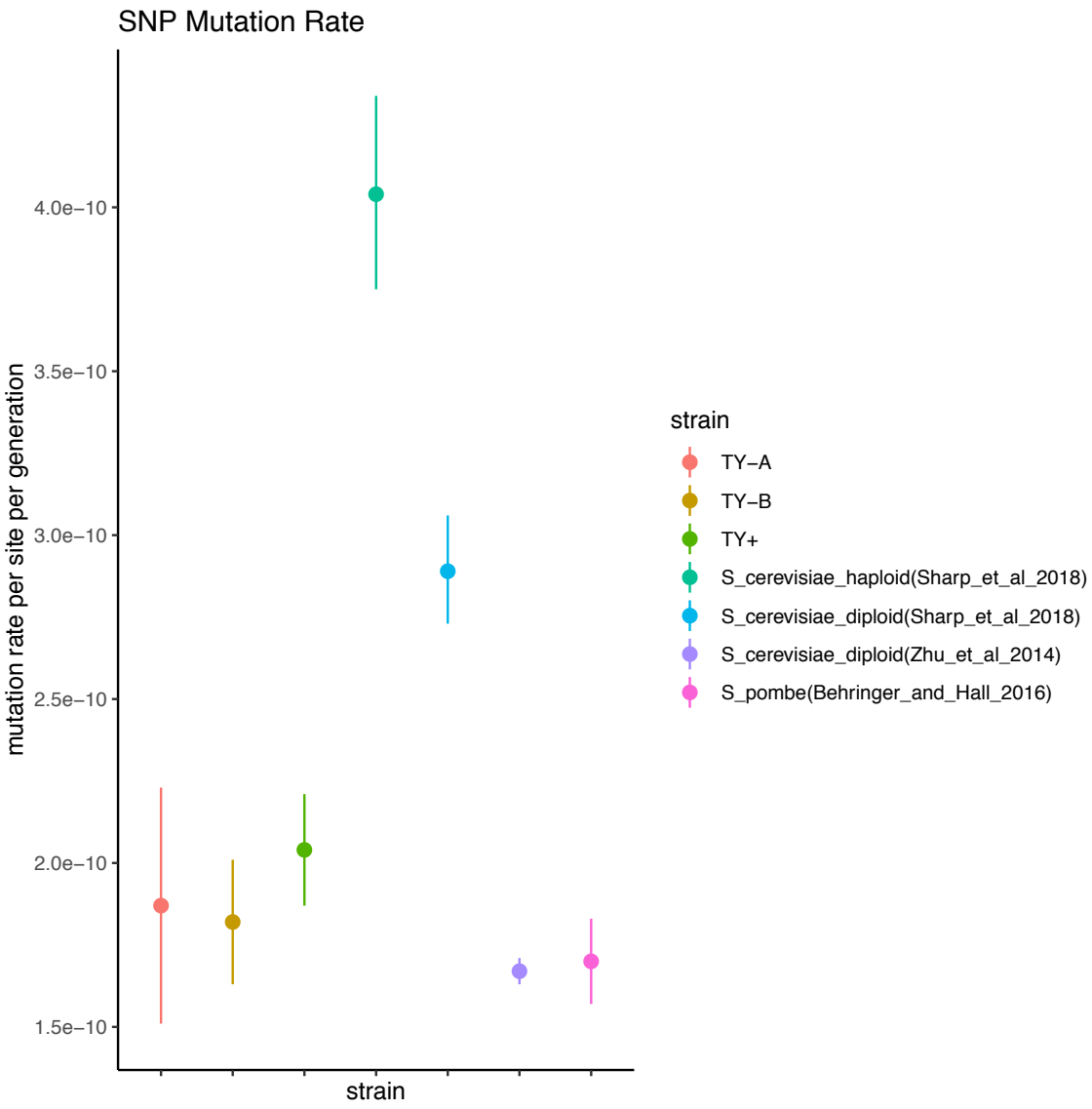


Figure 3.6: Comparison of SNP mutation rates across experiments. TY-A and TY-B are biological replicates of the 0-copy strain and TY+ is the 1-copy strain.

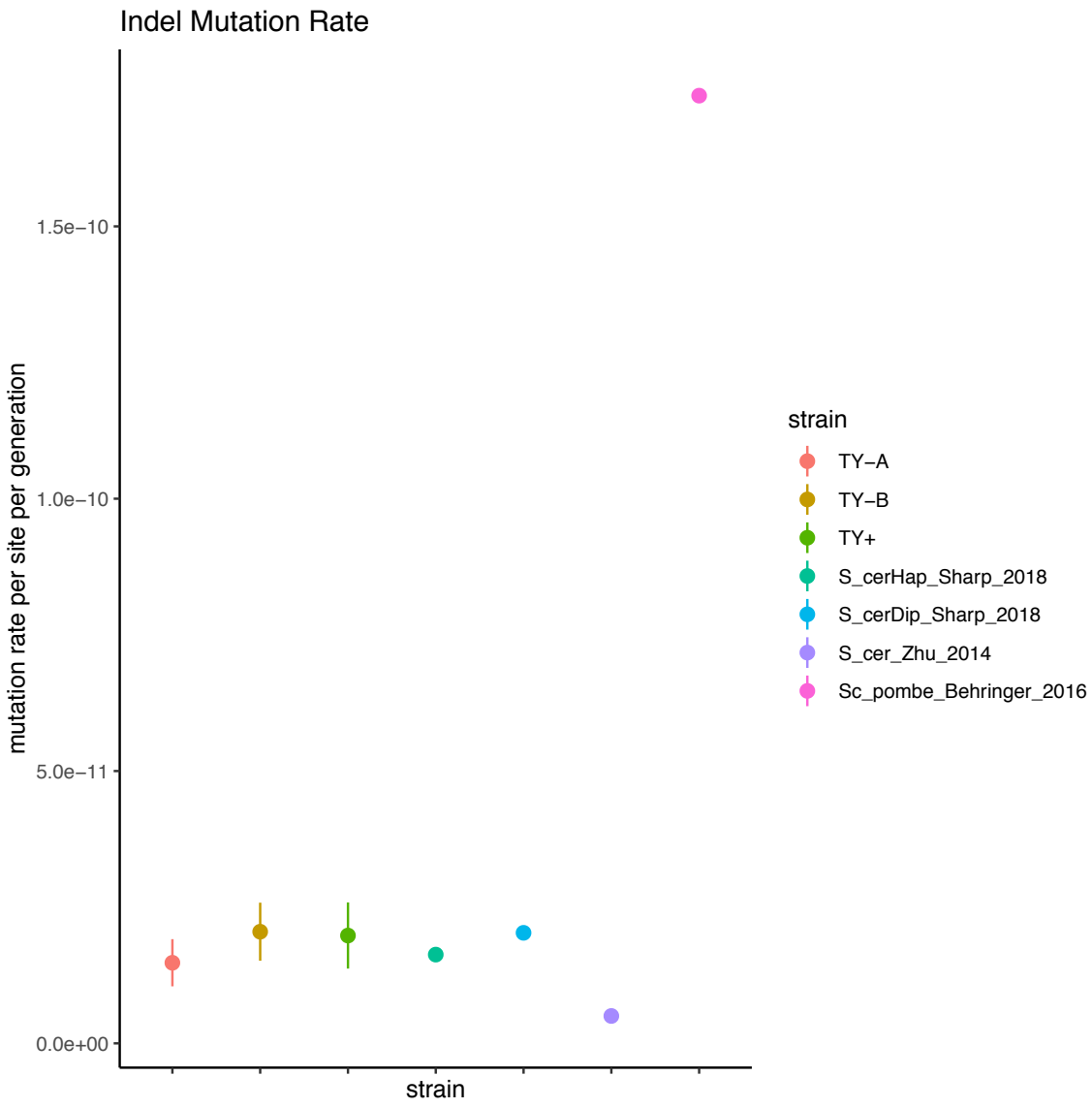


Figure 3.7: Comparison of indel mutation rates across experiments. TY-A and TY-B are biological replicates of the 0-copy strain and TY+ is the 1-copy strain.

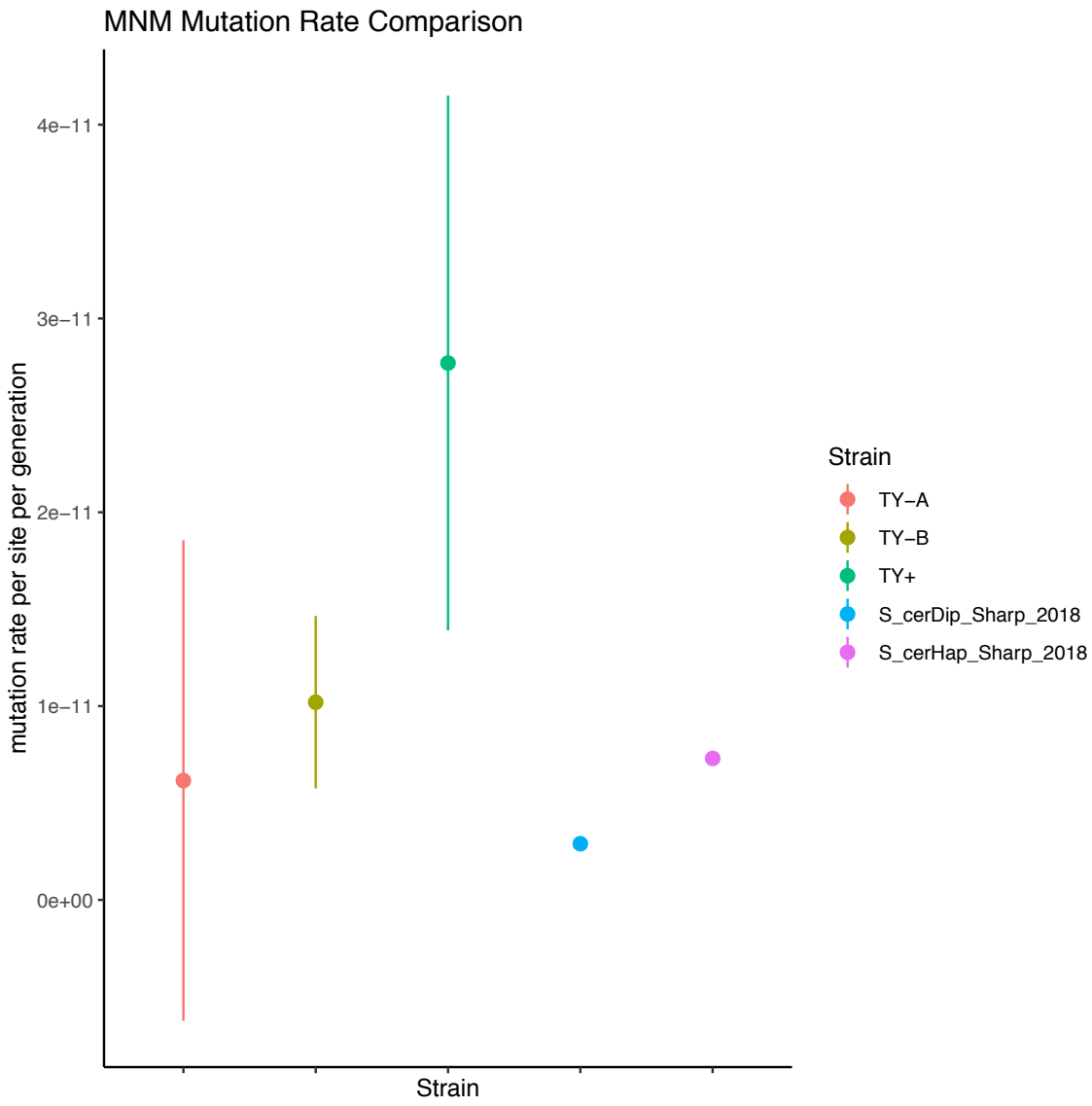


Figure 3.8: Comparison of MNM mutation rates across experiments. TY-A and TY-B are biological replicates of the 0-copy strain and TY+ is the 1-copy strain.

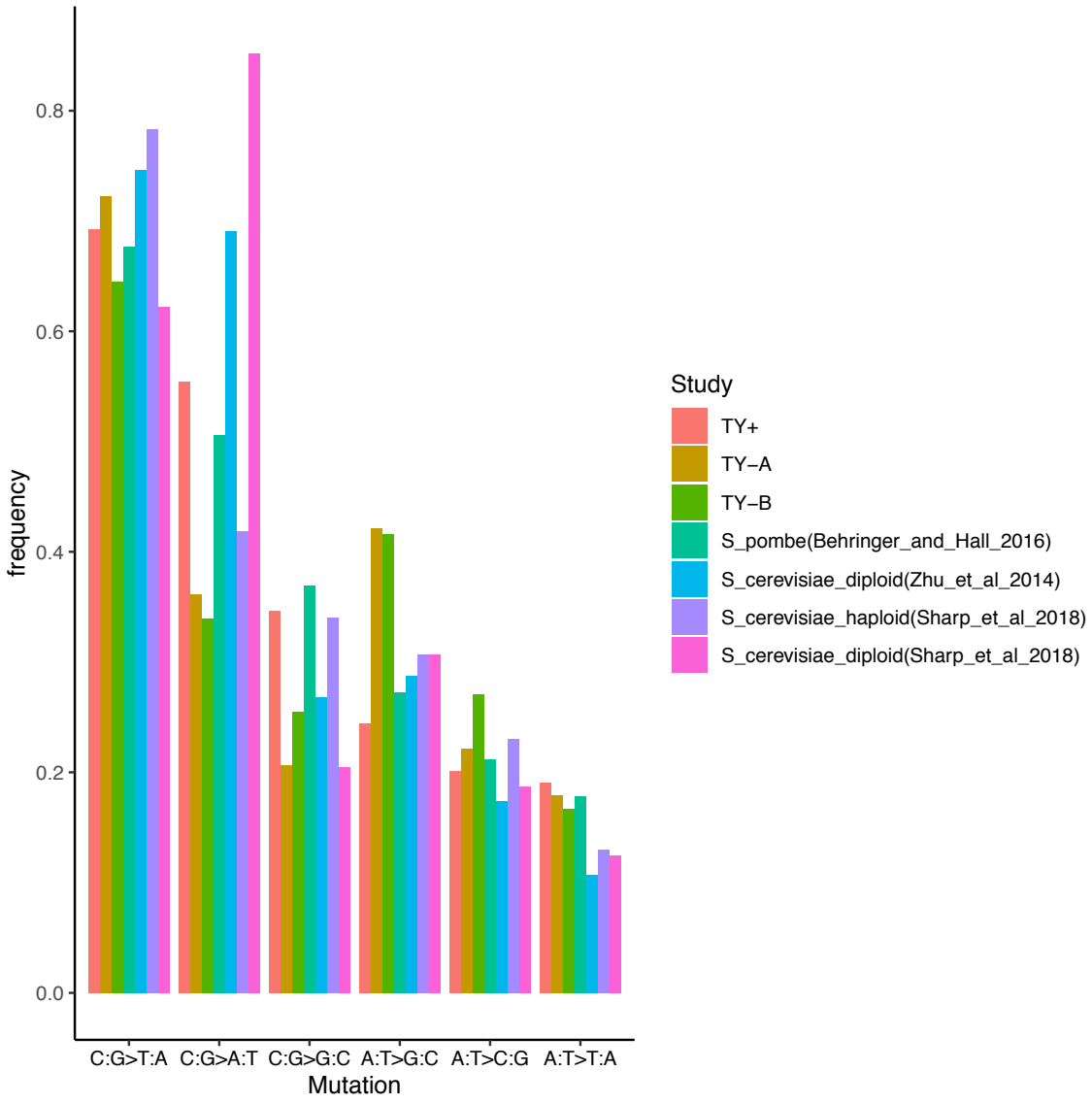


Figure 3.9: Comparison of spectrum of SNMs across experiments. TY-A and TY-B are biological replicates of the 0-copy strain and TY+ is the 1-copy strain. Y-axis is frequency of mutations relative to G/C content of the species' genome.

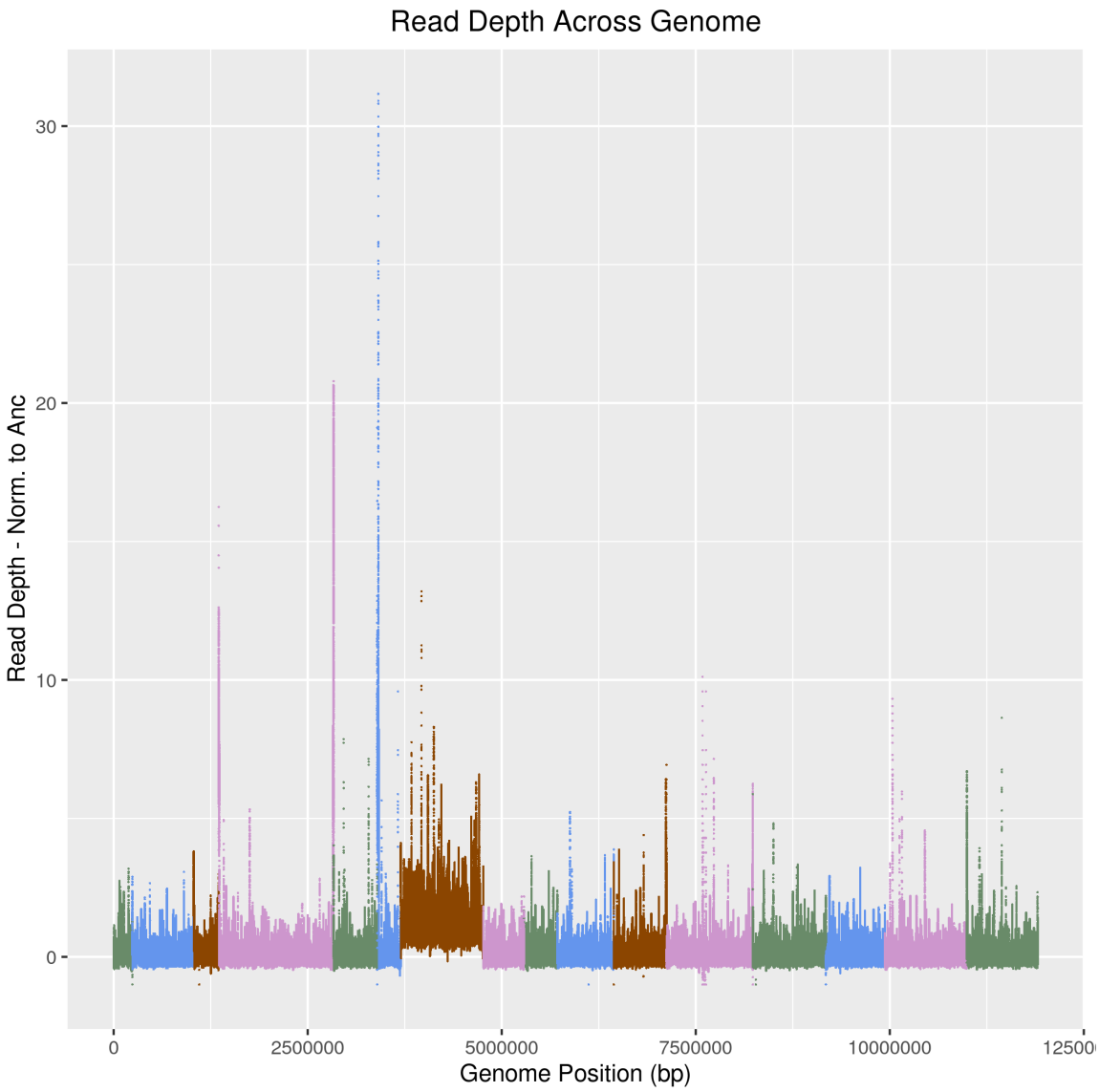


Figure 3.10: One whole-chromosome disomy event of chromosome VII was found in Line 43 of the TY-A strain.

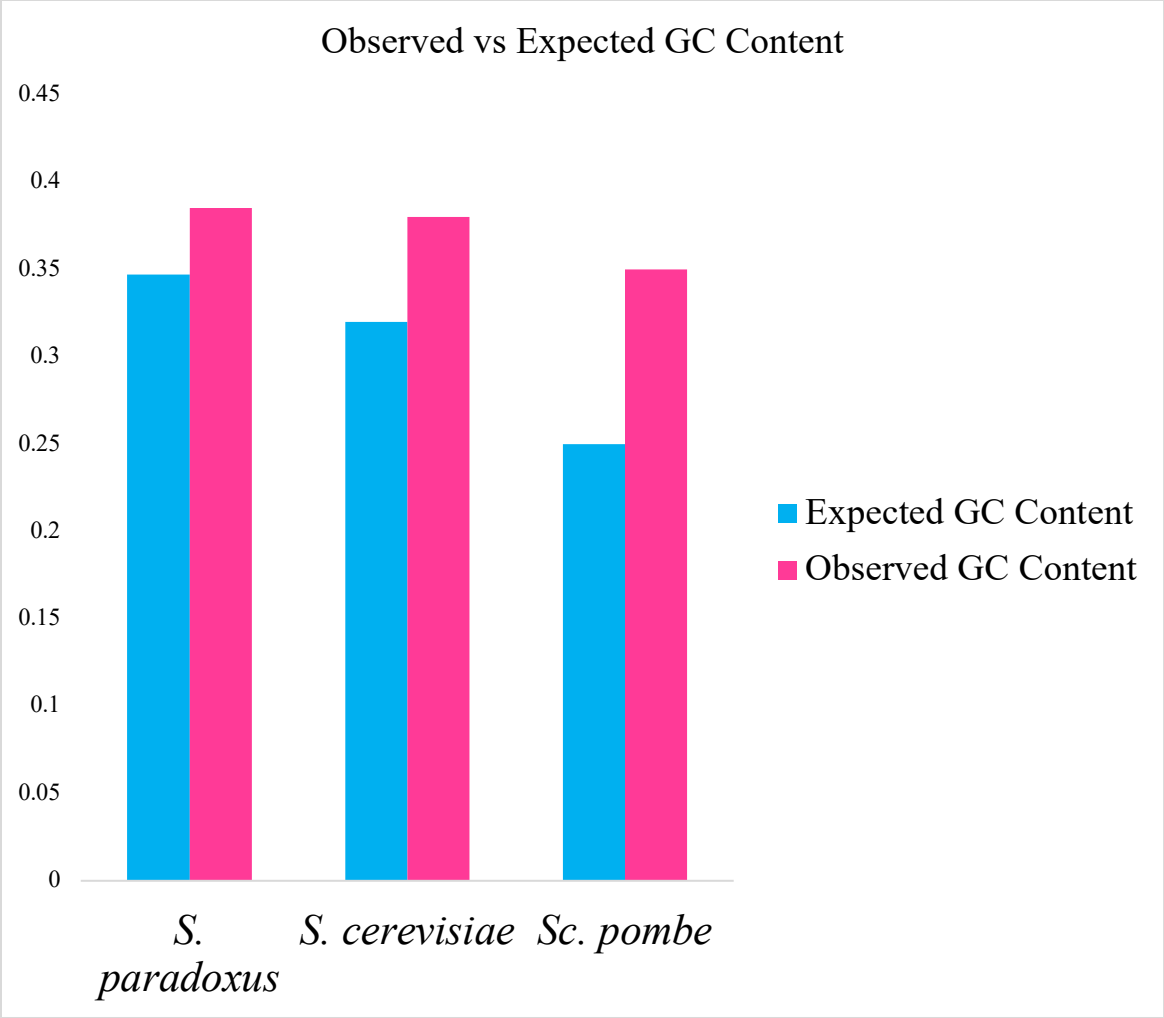
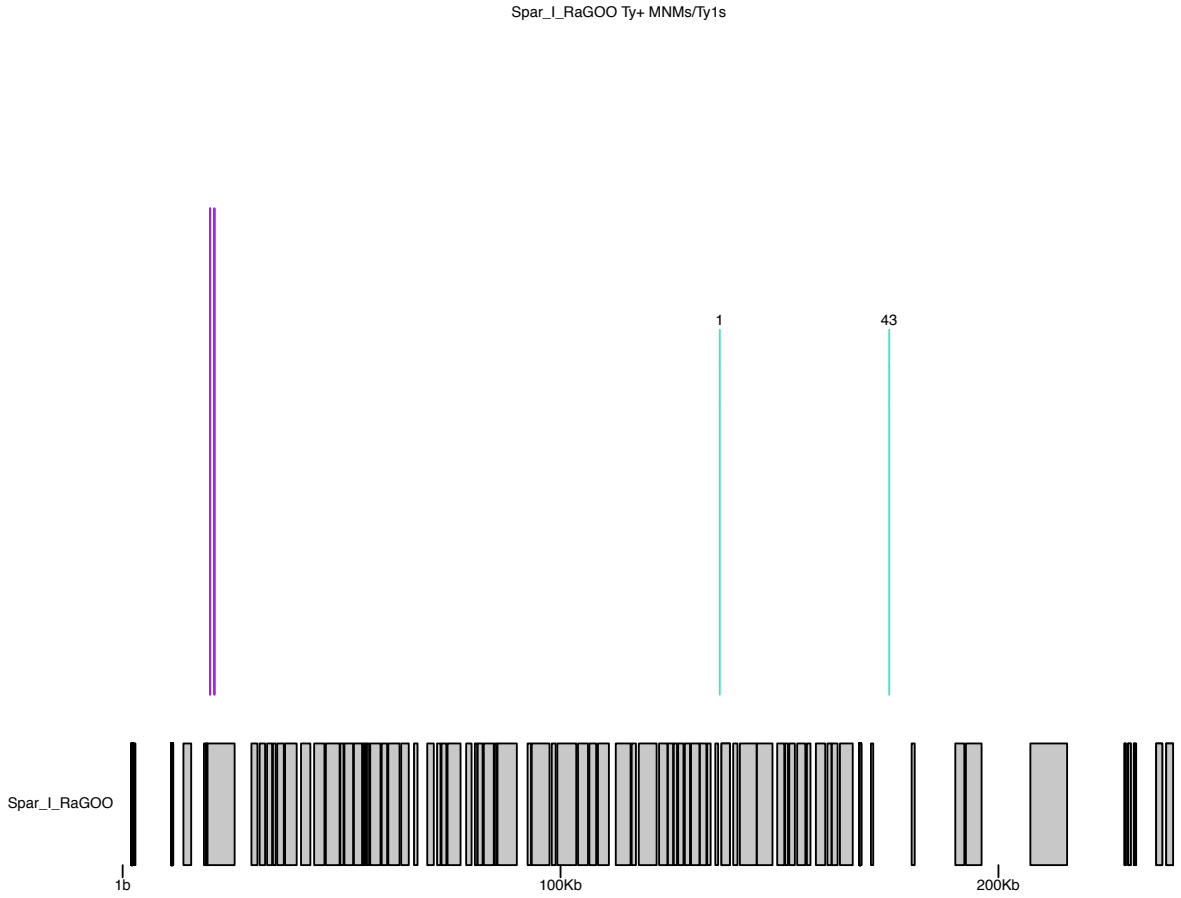
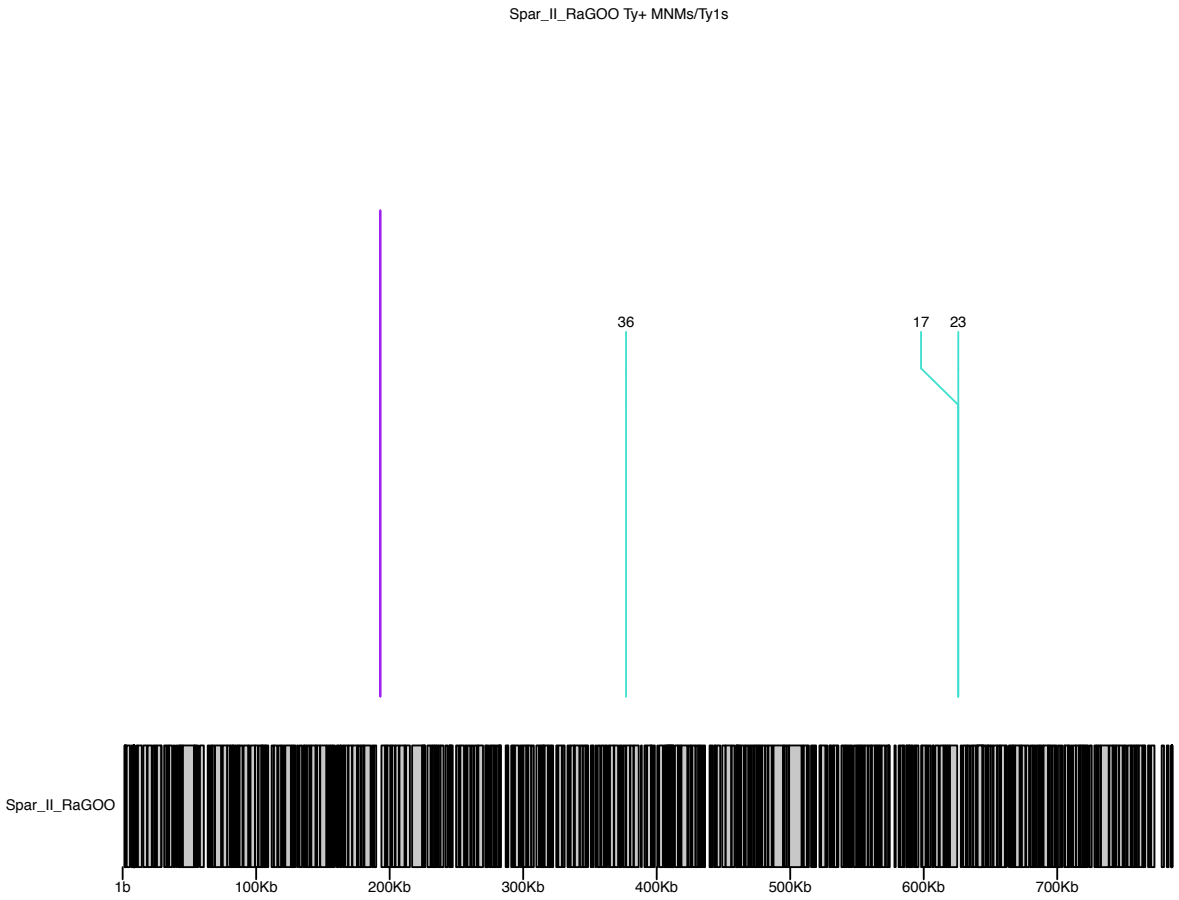
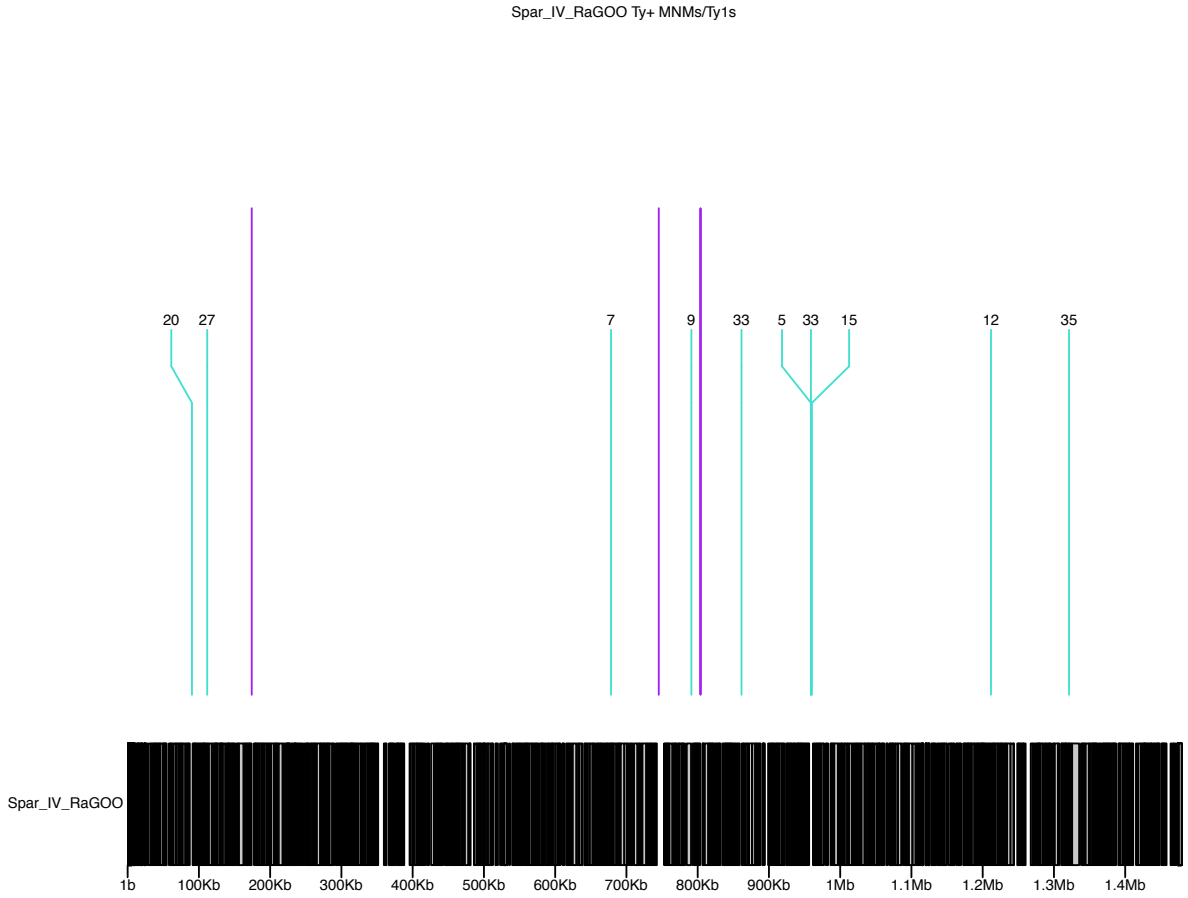


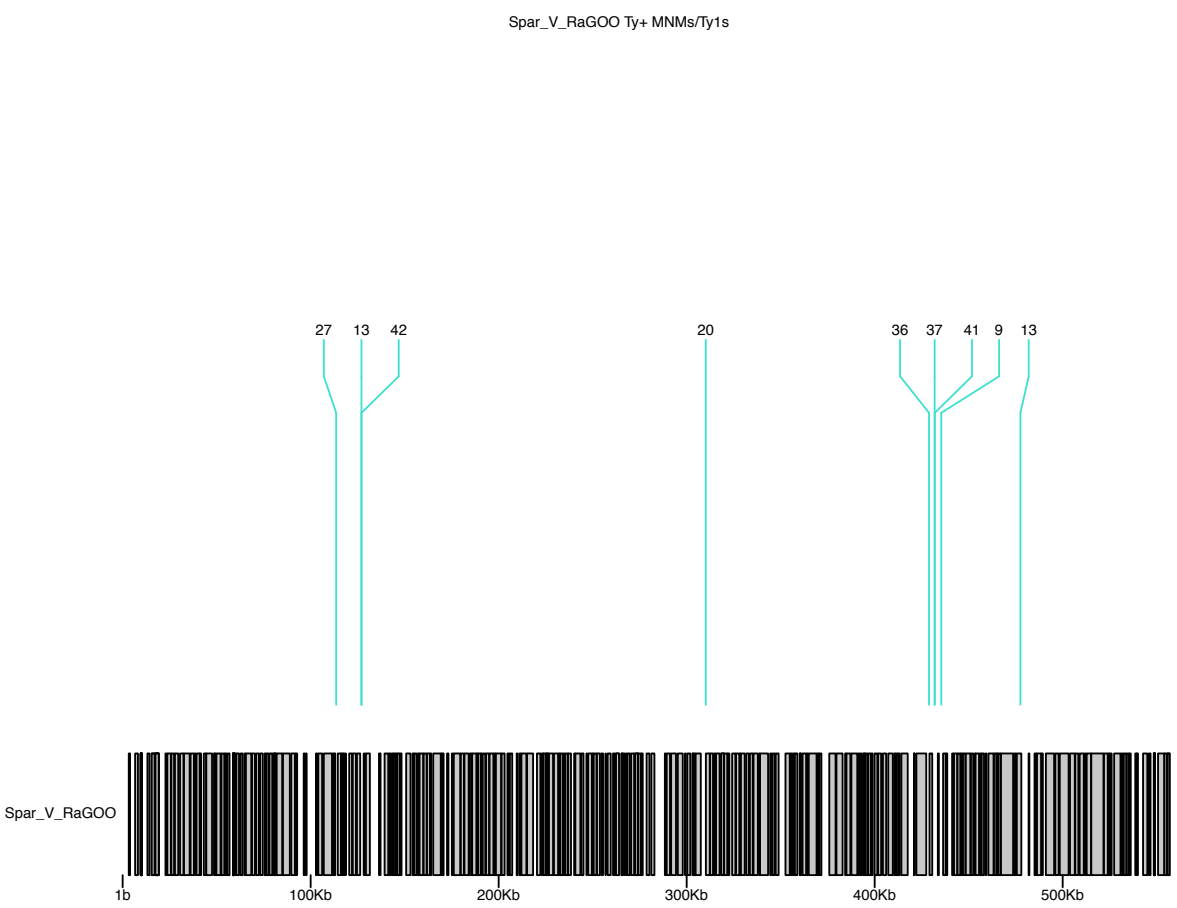
Figure 3.11: Observed versus expected GC content in 3 different yeast species.

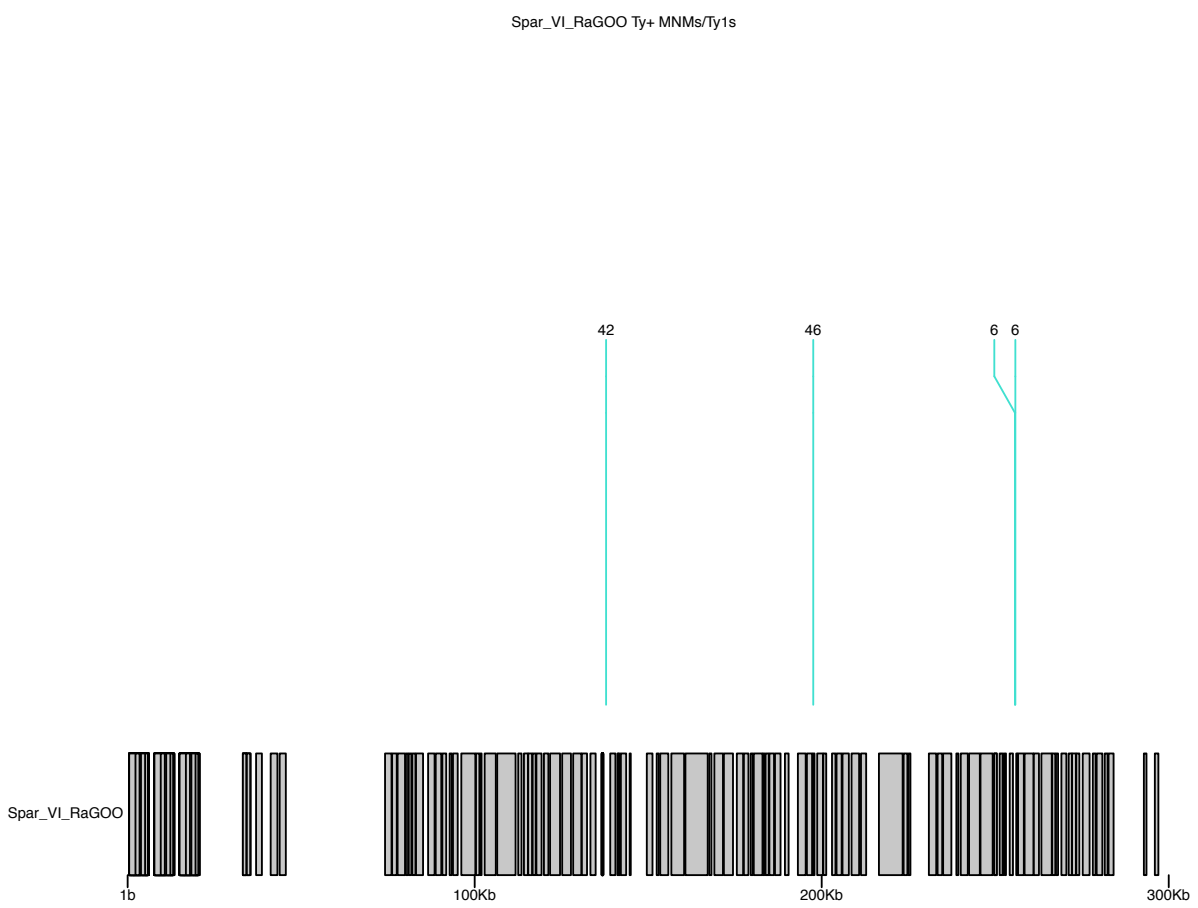

Spar_I_RaGOO Ty+ MNMs/Ty1s
1
43
Spar_I_RaGOO
1b
100Kb
200Kb

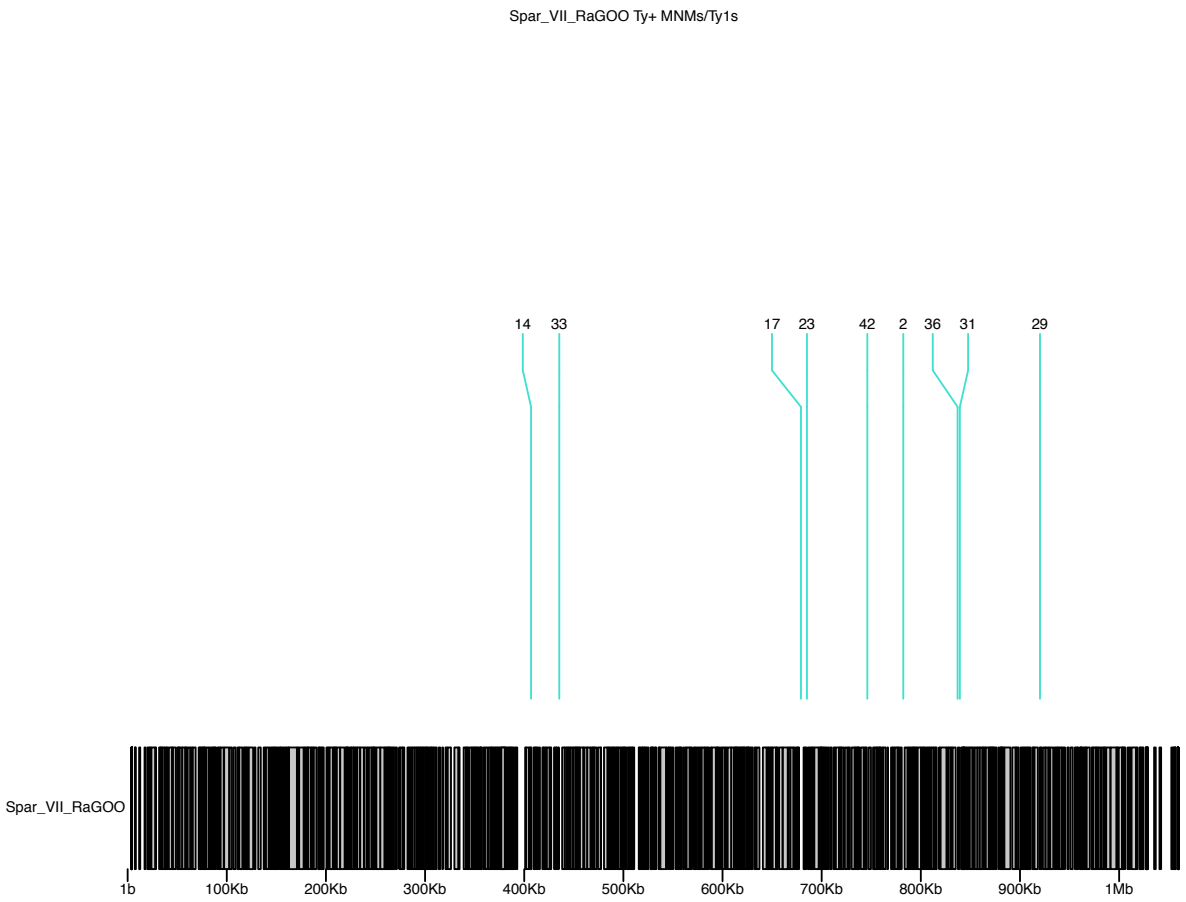

Spar_II_RaGOO Ty+ MNMs/Ty1s
36
17
23
Spar_II_RaGOO
1b
100Kb
200Kb
300Kb
400Kb
500Kb
600Kb
700Kb

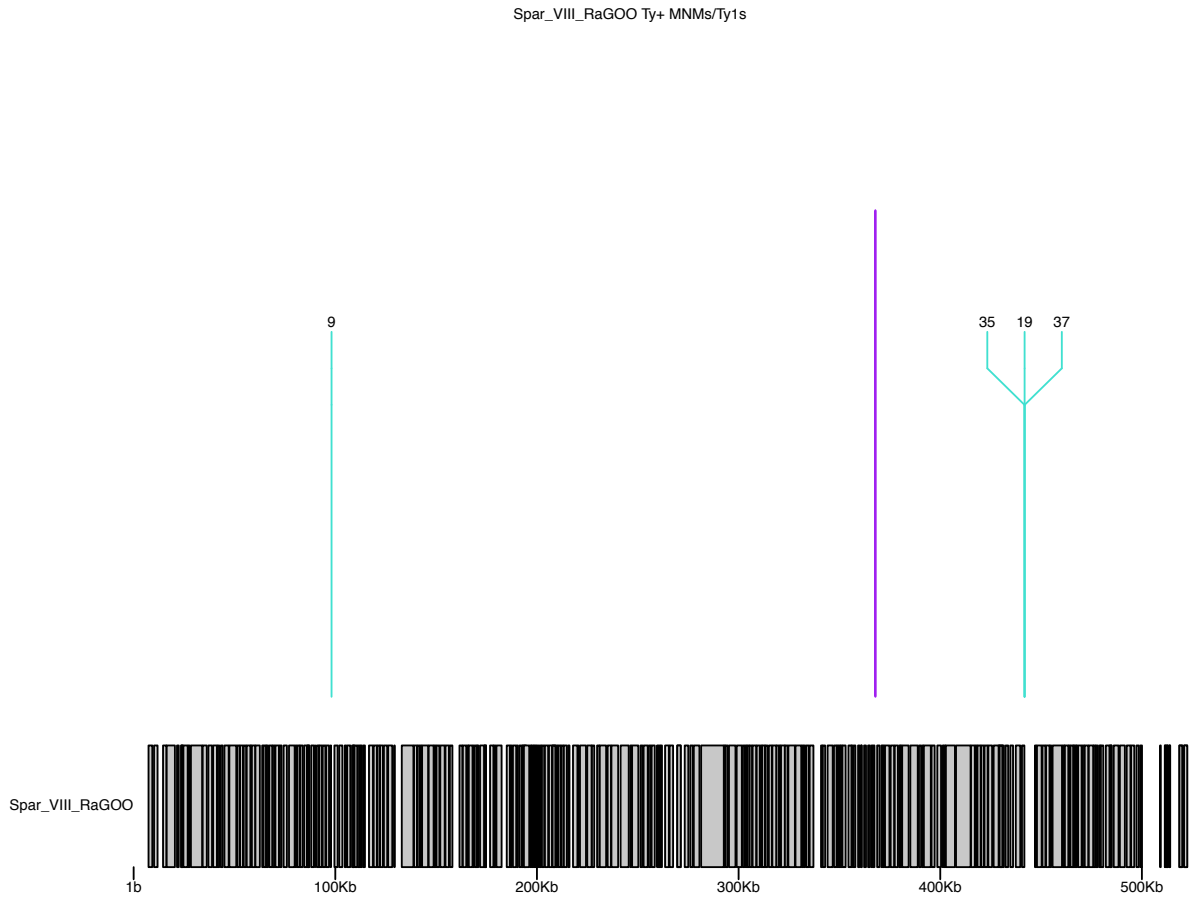

Spar_III_RaGOO Ty+ MNMs/Ty1s
20
46
Spar_III_RaGOO
1b
100Kb
200Kb
300Kb

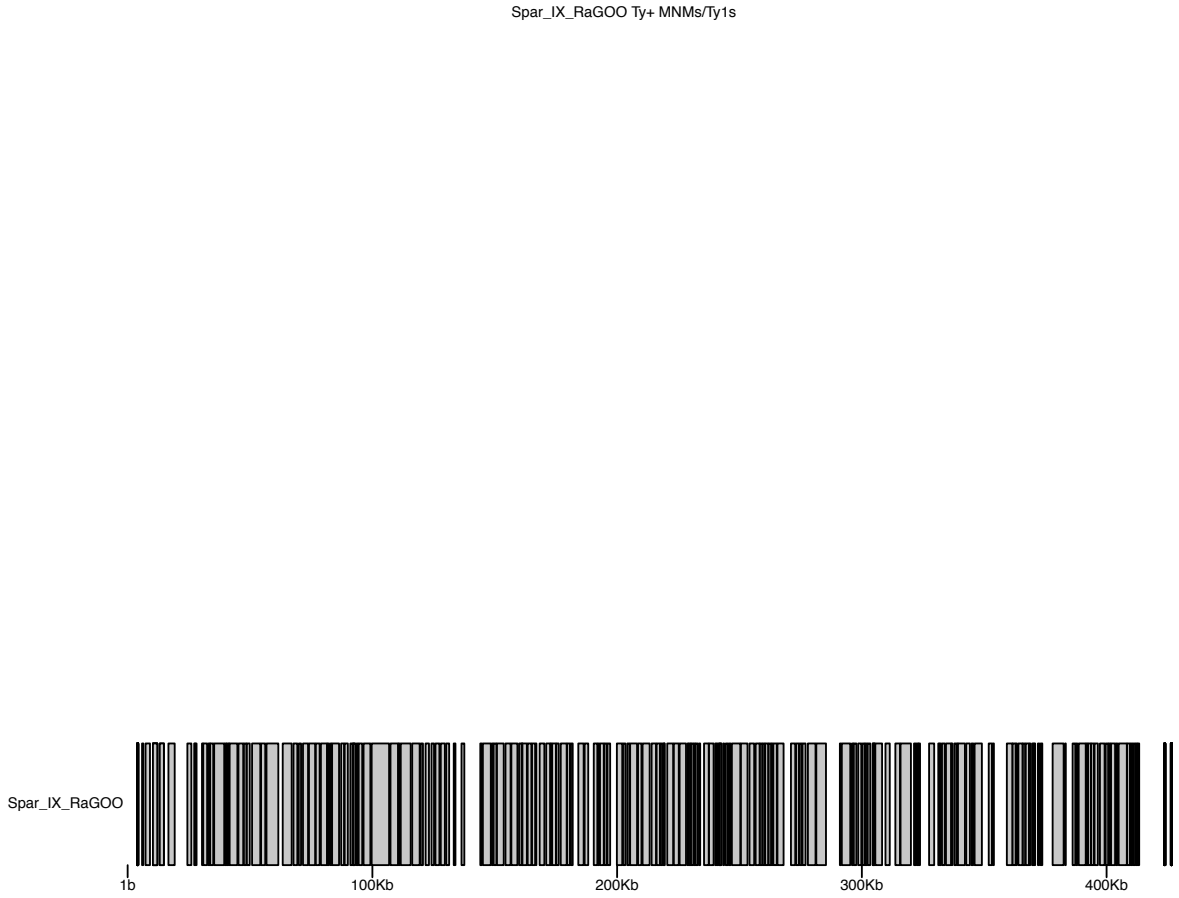

Spar_IV_RaGOO Ty+ MNMs/Ty1s
20
27
7
9
33
5
33
15
12
35
Spar_IV_RaGOO
1b
100Kb
200Kb
300Kb
400Kb
500Kb
600Kb
700Kb
800Kb
900Kb
1Mb
1.1Mb
1.2Mb
1.3Mb
1.4Mb

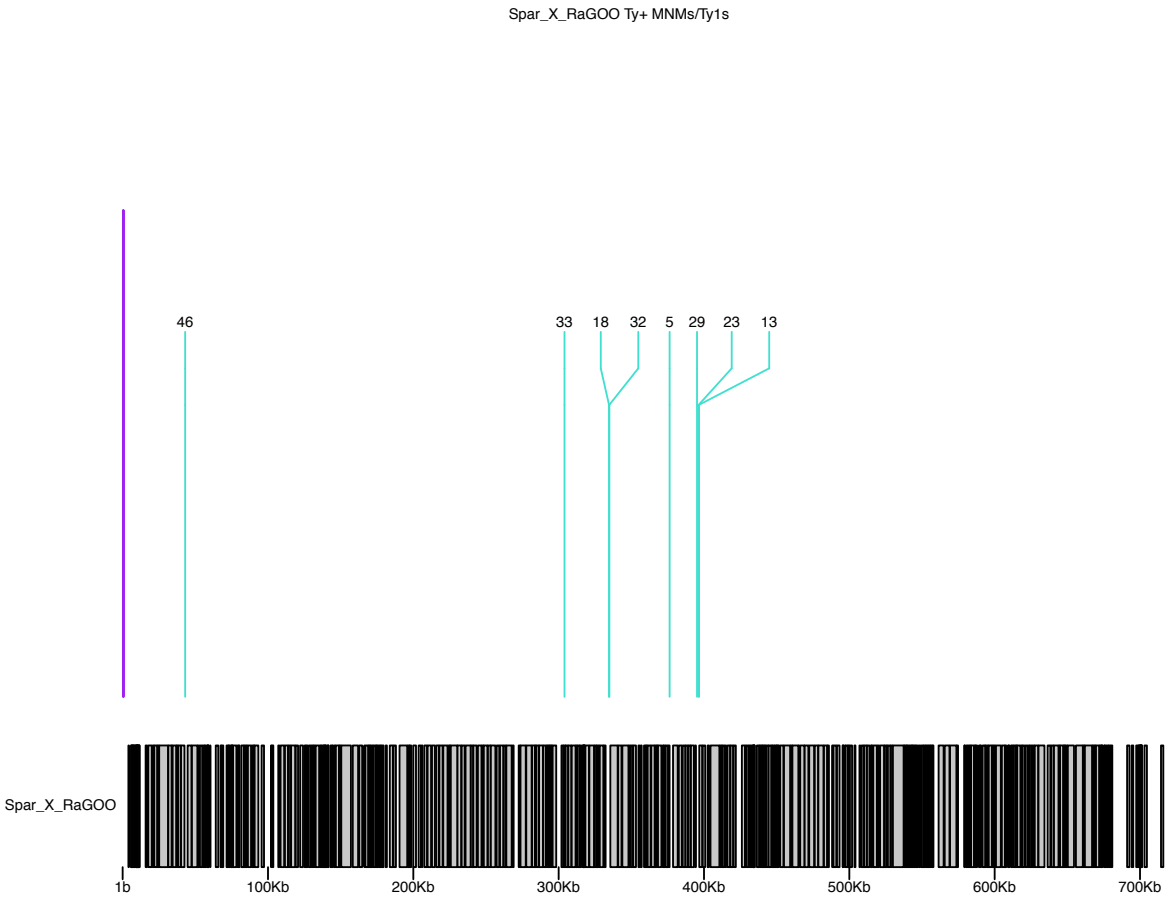

Spar_V_RaGOO Ty+ MNMs/Ty1s
27
13
42
20
36
37
41
9
13
Spar_V_RaGOO
1b
100Kb
200Kb
300Kb
400Kb
500Kb

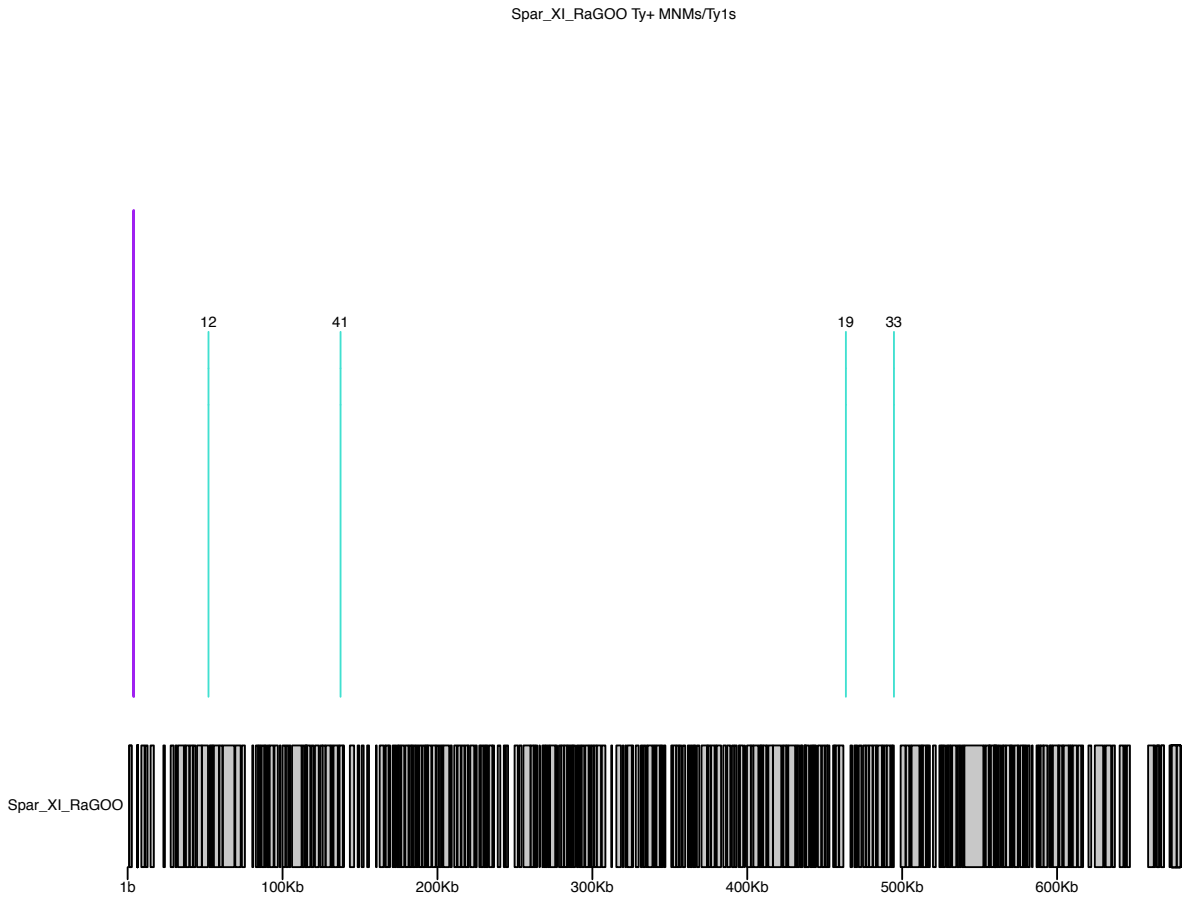

Spar_VI_RaGOO Ty+ MNMs/Ty1s
42
46
6
6
Spar_VI_RaGOO
1b
100Kb
200Kb
300Kb

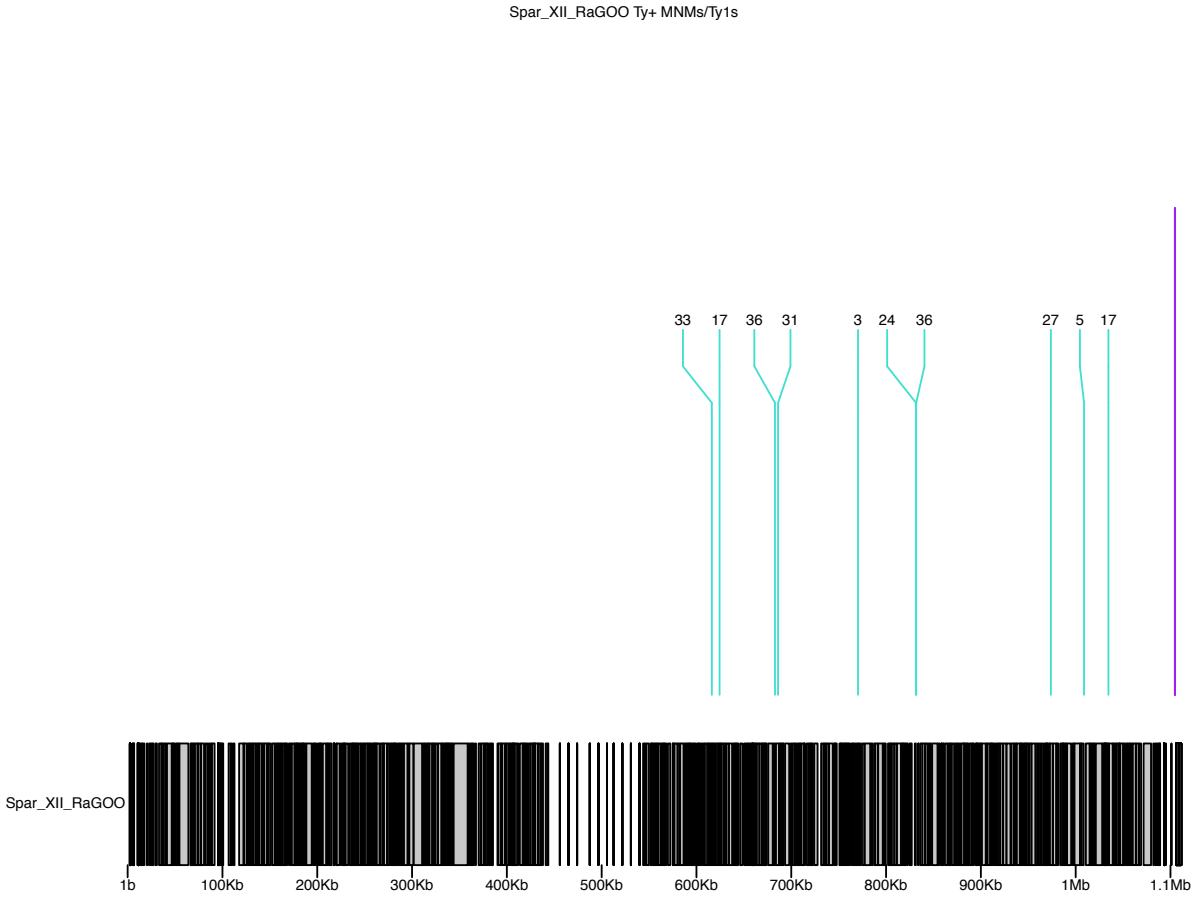

Spar_VII_RaGOO Ty+ MNMs/Ty1s
14
33
17
23
42
2
36
31
29
Spar_VII_RaGOO
1b
100Kb
200Kb
300Kb
400Kb
500Kb
600Kb
700Kb
800Kb
900Kb
1Mb

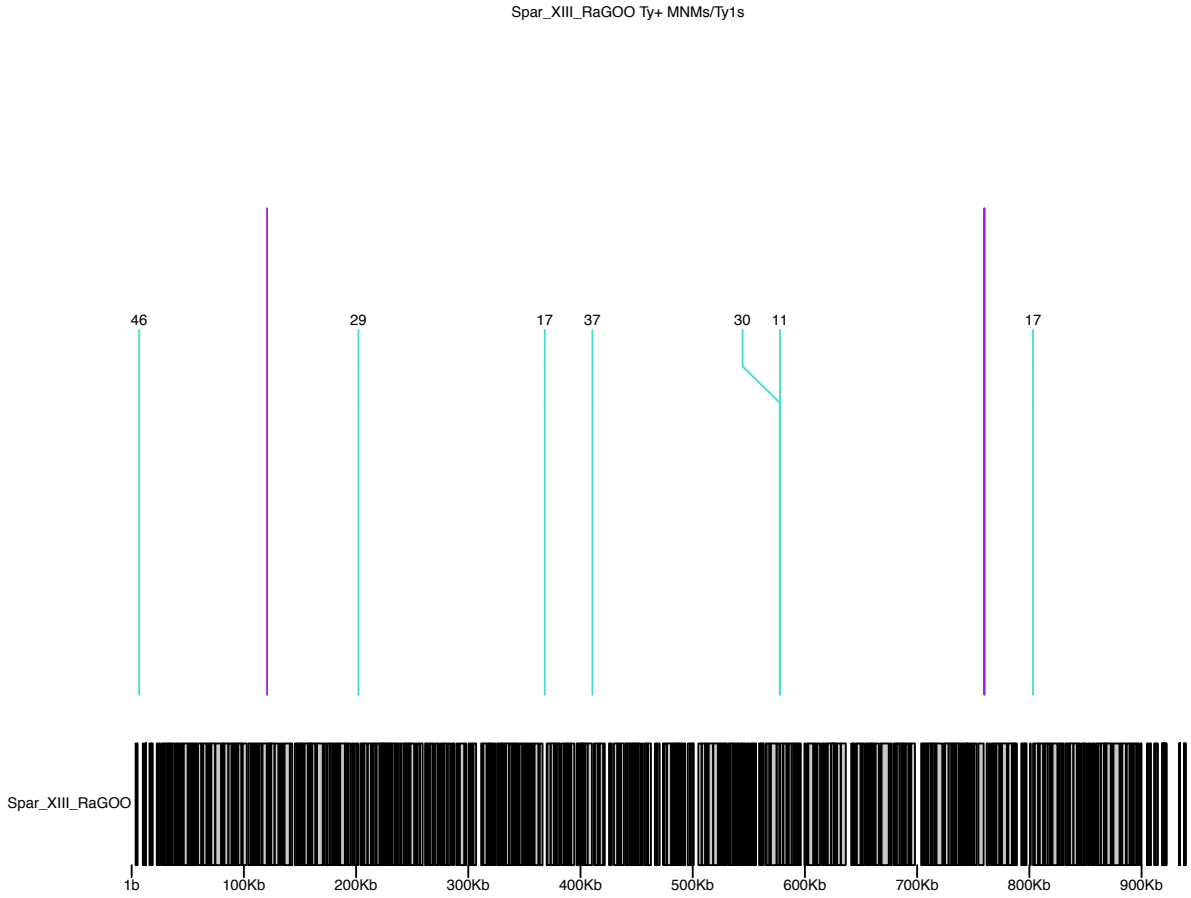

Spar_VIII_RaGOO Ty+ MNMs/Ty1s
9
35
19
37
Spar_VIII_RaGOO
1b
100Kb
200Kb
300Kb
400Kb
500Kb

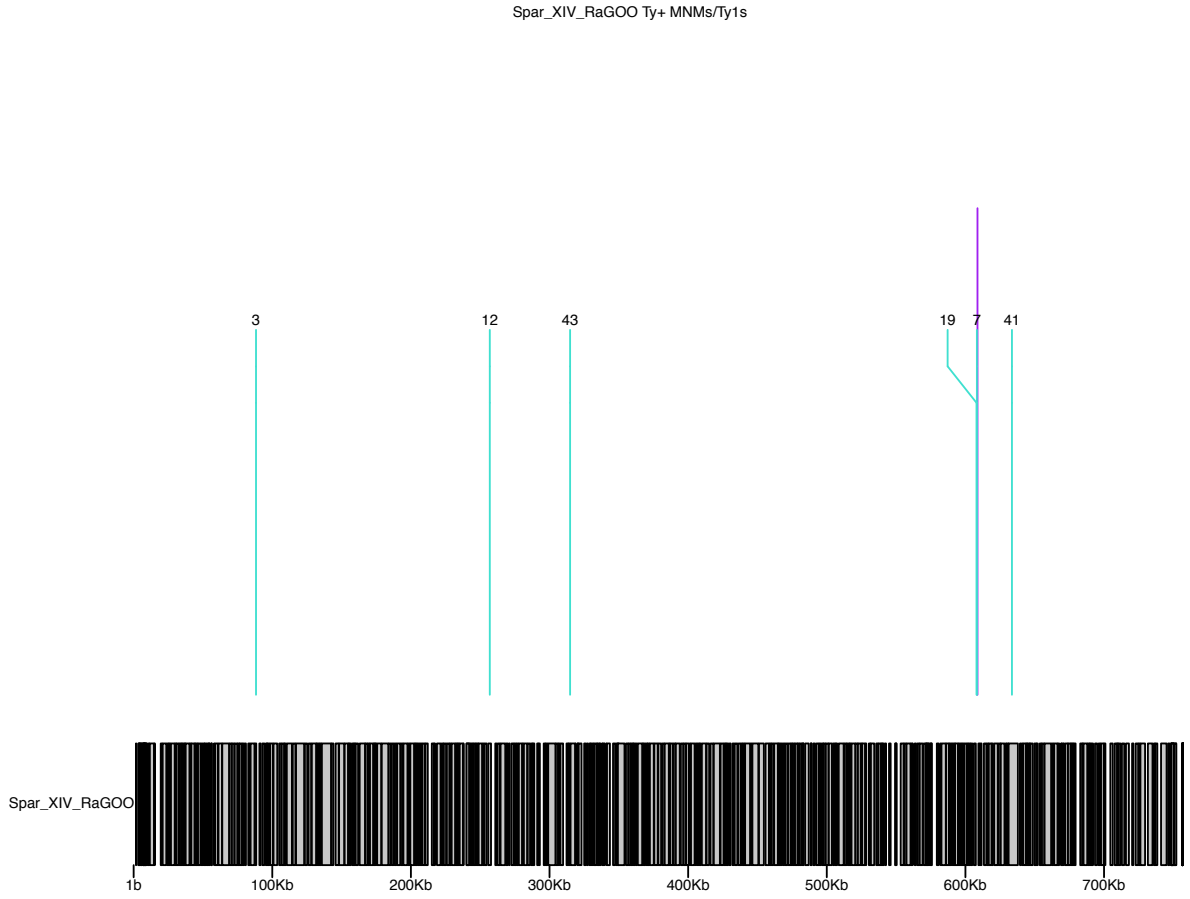

Spar_IX_RaGOO Ty+ MNMs/Ty1s
Spar_IX_RaGOO
1b
100Kb
200Kb
300Kb
400Kb

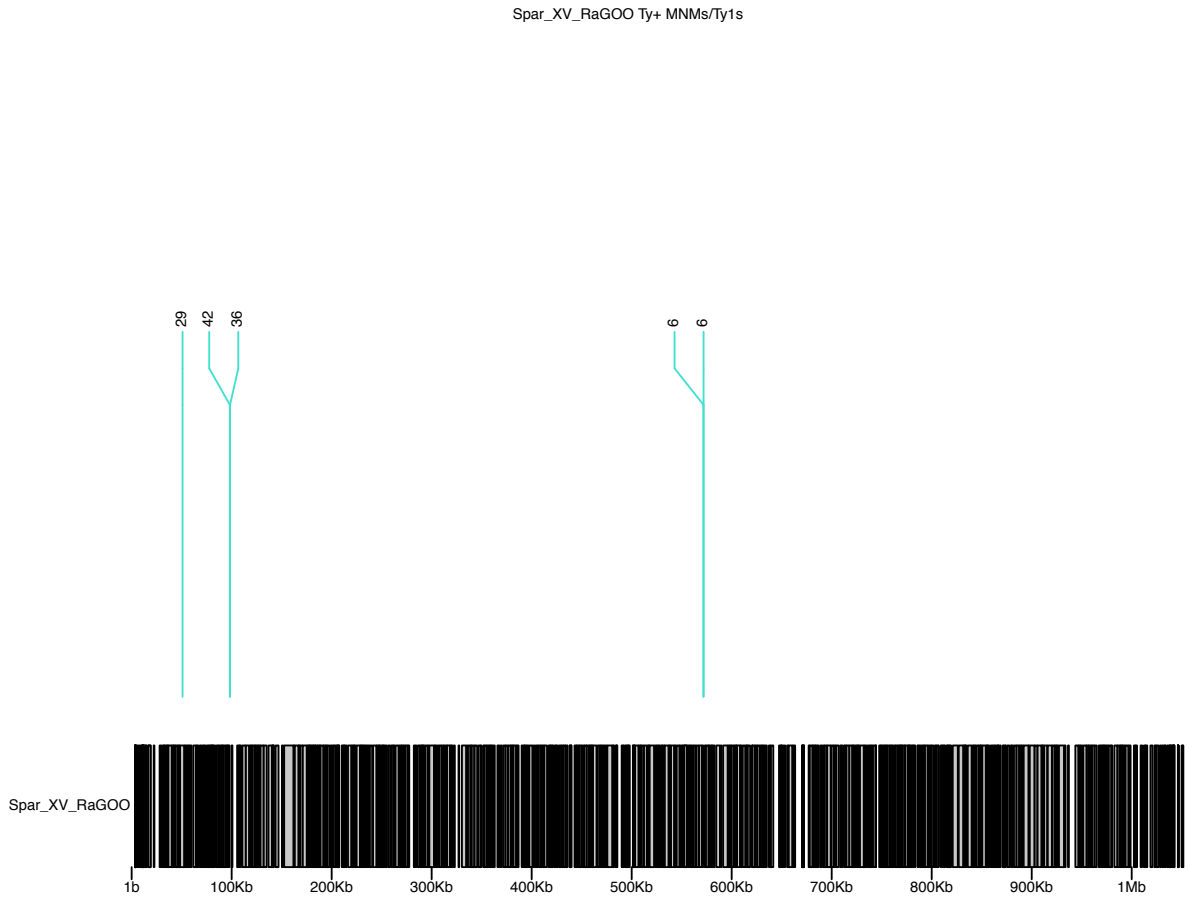

Spar_X_RaGOO Ty+ MNMs/Ty1s
46
33
18
32
5
29
23
13
Spar_X_RaGOO
1b
100Kb
200Kb
300Kb
400Kb
500Kb
600Kb
700Kb


Spar_XI_RaGOO Ty+ MNMs/Ty1s
12
41
19
33
Spar_XI_RaGOO
1b
100Kb
200Kb
300Kb
400Kb
500Kb
600Kb


Spar_XII_RaGOO Ty+ MNMs/Ty1s
33
17
36
31
3
24
36
27
5
17
Spar_XII_RaGOO
1b
100Kb
200Kb
300Kb
400Kb
500Kb
600Kb
700Kb
800Kb
900Kb
1Mb
1.1Mb


Spar_XIII_RaGOO Ty+ MNMs/Ty1s
46
29
17
37
30
11
17
Spar_XIII_RaGOO
1b
100Kb
200Kb
300Kb
400Kb
500Kb
600Kb
700Kb
800Kb
900Kb


Spar_XIV_RaGOO Ty+ MNMs/Ty1s
3
12
43
19
7
41
Spar_XIV_RaGOO
1b
100Kb
200Kb
300Kb
400Kb
500Kb
600Kb
700Kb


Spar_XV_RaGOO Ty+ MNMs/Ty1s
29
42
36
6
6
Spar_XV_RaGOO
1b
100Kb
200Kb
300Kb
400Kb
500Kb
600Kb
700Kb
800Kb
900Kb
1Mb

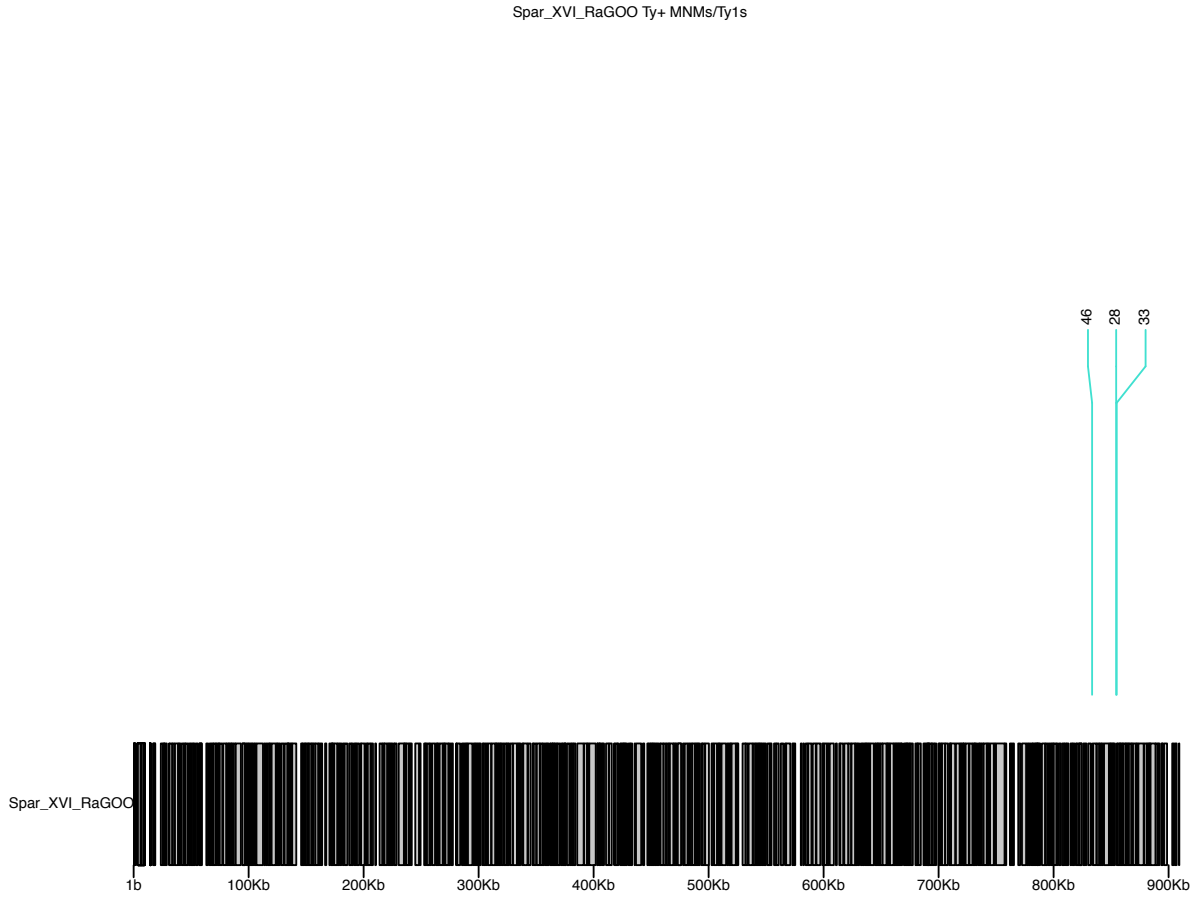

Spar_XVI_RaGOO Ty+ MNMs/Ty1s
46
28
33
Spar_XVI_RaGOO
1b
100Kb
200Kb
300Kb
400Kb
500Kb
600Kb
700Kb
800Kb
900Kb

Figure 3.12: Multinucleotide mutations and Ty1 locations across TY+ samples. Turquoise labels are Ty1 insertions from TELocate. Purple bars are MNM regions in TY+ samples. Plots are the chromosome represented in ideogram format: the gray boxes represent annotated genes along each chromosome.

References

Alonge, M., S. Soyk, S. Ramakrishnan, X. Wang, S. Goodwin *et al.*, 2019 RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome biology 20**:** 1-17.

Andrews, S., 2010 FastQC, pp.

Behringer, M. G., and D. W. Hall, 2016 Genome-wide estimates of mutation rates and spectrum in Schizosaccharomyces pombe indicate CpG sites are highly mutagenic despite the absence of DNA methylation. G3: Genes| Genomes| Genetics 6**:** 149-160.

Bennetzen, J. L., 2000 Transposable element contributions to plant gene and genome evolution. Plant molecular biology 42**:** 251-269.

Bergman, C. M., 2018 Horizontal transfer and proliferation of Tsu4 in Saccharomyces paradoxus. Mobile DNA 9**:** 1-8.

Capuano, F., M. Mülleder, R. Kok, H. J. Blom and M. Ralser, 2014 Cytosine DNA methylation is found in Drosophila melanogaster but absent in Saccharomyces cerevisiae, Schizosaccharomyces pombe, and other yeast species. Analytical chemistry 86**:** 3697-3702.

Carr, M., D. Bensasson and C. M. Bergman, 2012 Evolutionary genomics of transposable elements in Saccharomyces cerevisiae. PloS one 7**:** e50978.

Curcio, M. J., S. Lutz and P. Lesage, 2015 The Ty1 LTR-retrotransposon of budding yeast, Saccharomyces cerevisiae. Microbiology spectrum 3**:** 1.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics 43**:** 491.

Eigel, A., and H. Feldmann, 1982 Ty1 and delta elements occur adjacent to several tRNA genes in yeast. The EMBO journal 1**:** 1245-1250.

Elder, R., T. S. John, D. Stinchcomb and R. Davis, 1981 Studies on the transposable element Tyl of yeast I. RNA homologous to Tyl, pp. 581-591 in *Cold Spring Harbor symposia on quantitative biology*. Cold Spring Harbor Laboratory Press.

Fawcett, J. A., and H. Innan, 2019 The role of gene conversion between transposable elements in rewiring regulatory networks. Genome biology and evolution 11**:** 1723-1729.

Fay, M. P., 2010 Two-sided exact tests and matching confidence intervals for discrete data. R journal 2**:** 53-58.

Feschotte, C., and E. J. Pritham, 2007 DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. 41**:** 331-368.

Frederico, L. A., T. A. Kunkel and B. R. Shaw, 1993 Cytosine deamination in mismatched base pairs. Biochemistry 32**:** 6523-6530.

Garfinkel, D., 2005 Genome evolution mediated by Ty elements in Saccharomyces. Cytogenetic and genome research 110**:** 63-69.

Garfinkel, D. J., K. Nyswaner, J. Wang and J.-Y. Cho, 2003 Post-transcriptional cosuppression of Ty1 retrotransposition. Genetics 165**:** 83-99.

Hani, J., and H. Feldmann, 1998 tRNA genes and retroelements in the yeast genome. Nucleic acids research 26**:** 689-696.

Hicks, W. M., M. Kim and J. E. Haber, 2010 Increased mutagenesis and unique mutation signature associated with mitotic gene conversion. Science 329**:** 82-85.

Hollister, J. D., and B. S. Gaut, 2009 Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome research 19**:** 1419-1428.

Ji, H., D. Moore, M. Blomberg, L. Braiterman, D. Voytas *et al.*, 1993 Hotspots for unselected Ty1 transposition events on yeast chromosome III are near tRNA genes and LTR sequences. Cell 73**:** 1007-1018.

Joseph, S. B., and D. W. Hall, 2004 Spontaneous Mutations in Diploid Saccharomyces cerevisiae. Genetics 168**:** 1817-1825.

Kim, J. M., S. Vanguri, J. D. Boeke, A. Gabriel and D. F. Voytas, 1998 Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. Genome research 8**:** 464-478.

Krueger, F., 2012 Trim Galore!, pp.

Liti, G., D. M. Carter, A. M. Moses, J. Warringer, L. Parts *et al.*, 2009 Population genomics of domestic and wild yeasts. Nature 458**:** 337-341.

Lynch, M., W. Sung, K. Morris, N. Coffey, C. R. Landry *et al.*, 2008 A genome-wide view of the spectrum of spontaneous mutations in yeast. Proceedings of the National Academy of Sciences 105**:** 9272-9277.

Matsuda, E., and D. J. Garfinkel, 2009 Posttranslational interference of Ty1 retrotransposition by antisense RNAs. Proceedings of the National Academy of Sciences 106**:** 15657-15662.

Moore, S. P., G. Liti, K. M. Stefanisko, K. M. Nyswaner, C. Chang *et al.*, 2004 Analysis of a Ty1-less variant of Saccharomyces paradoxus: the gain and loss of Ty1 elements. Yeast 21**:** 649-660.

Naumov, G. I., E. S. Naumova, R. A. Lantto, E. J. Louis and M. Korhola, 1992 Genetic homology between Saccharomyces cerevisiae and its sibling species S. paradoxus and S. bayanus: electrophoretic karyotypes. Yeast 8**:** 599-612.

Nelson, M. G., R. S. Linheiro and C. M. Bergman, 2017 McClintock: an integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. G3: Genes, Genomes, Genetics 7**:** 2763-2778.

Neph, S., M. S. Kuehn, A. P. Reynolds, E. Haugen, R. E. Thurman *et al.*, 2012 BEDOPS: high-performance genomic feature operations. Bioinformatics 28**:** 1919-1920.

Pagès, H., P. Aboyoun, R. Gentleman and S. DebRoy, 2020 Biostrings: Efficient manipulation of biological strings. R package version 2.56.0.

Pósfai, G., G. Plunkett, T. Fehér, D. Frisch, G. M. Keil *et al.*, 2006 Emergent properties of reduced-genome Escherichia coli. science 312**:** 1044-1046.

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

Roberts, D. N., A. J. Stewart, J. T. Huff and B. R. Cairns, 2003 The RNA polymerase III transcriptome revealed by genome-wide localization and activity–occupancy relationships. Proceedings of the National Academy of Sciences 100**:** 14695-14700.

Robinson, J. T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander *et al.*, 2011 Integrative genomics viewer. Nature biotechnology 29**:** 24-26.

Rodgers, K., and M. McVey, 2016 Error-prone repair of DNA double-strand breaks. Journal of cellular physiology 231**:** 15-24.

Sharp, N. P., L. Sandell, C. G. James and S. P. Otto, 2018 The genome-wide rate and spectrum of spontaneous mutations differ between haploid and diploid yeast. Proceedings of the National Academy of Sciences 115**:** E5046-E5055.

Suárez, G. A., B. A. Renda, A. Dasgupta and J. E. Barrick, 2017 Reduced mutation rate and increased transformability of transposon-free Acinetobacter baylyi ADP1-ISx. Applied and environmental microbiology 83.

Treangen, T. J., and S. L. Salzberg, 2012 Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics 13**:** 36-46.

Urich, M. A., J. R. Nery, R. Lister, R. J. Schmitz and J. R. Ecker, 2015 MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nature protocols 10**:** 475.

VanHoute, D., and P. H. Maxwell, 2014 Extension of Saccharomyces paradoxus chronological lifespan by retrotransposons in certain media conditions is associated with changes in reactive oxygen species. Genetics 198**:** 531-545.

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel *et al.*, 2014 Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS one 9**:** e112963.

Wicker, T., Y. Yu, G. Haberer, K. F. Mayer, P. R. Marri *et al.*, 2016 DNA transposon activity is associated with increased mutation rates in genes of rice and other grasses. Nature communications 7**:** 1-9.

Wickham, H., 2016 *ggplot2: elegant graphics for data analysis*. Springer.

Xia, J., L. Han and Z. Zhao, 2012 Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome. BMC genomics 13**:** 1-9.

Yue, J.-X., and G. Liti, 2018 Long-read sequencing data analysis for yeasts. Nature protocols 13**:** 1213-1231.

Zhu, Y. O., M. L. Siegal, D. W. Hall and D. A. Petrov, 2014 Precise estimates of mutation rate and spectrum in yeast. Proceedings of the National Academy of Sciences 111**:** E2310-E2318.

Zou, S., D. A. Wright and D. F. Voytas, 1995 The Saccharomyces Ty5 retrotransposon family is associated with origins of DNA replication at the telomeres and the silent mating locus HMR. Proceedings of the National Academy of Sciences 92**:** 920-924.

CHAPTER 4

CONCLUSION

In this dissertation, I have presented two studies investigating the effects of two intragenomic factors: whole-chromosomal aneuploidy events and the presence of mobile genetic elements. We have seen that in both cases, there is widespread impacts on the transcriptome (for aneuploidy events) and other mutations (for mobile genetic elements). To assess effects on the transcriptome, we analyzed RNAseq data from 38 samples of both aneuploid and euploid yeast strains from two mutation accumulation experiments. We compared the gene expression of the derived lines to that of each ancestor to determine differential expression across the transcriptome. In light of recent debate on the presence of whole-chromosomal dosage compensation in yeast (JAMES HOSE 2015; AUDREY P GASCH 2016; EDUARDO M TORRES 2016), we sought to determine if our spontaneously-aneuploid yeast samples showed any evidence for whole-chromosomal dosage compensation, and found that they did not. We did find evidence for compensation on a gene-by-gene basis, however, especially for those genes that have been determined to be dosage-sensitive (MAKANAE *et al.* 2013). We also found evidence for differential expression of genes previously found to be associated with aneuploidy (TORRES *et al.* 2010) and the environmental stress response (CHEN *et al.* 2003).

In chapter 3, we asked whether the presence of transposable elements affects mutation rate and spectra in yeast. Using a mutation accumulation experiment, we

uncovered a total of 544 spontaneous mutations across 108 samples and determined the impacts of transposon presence. We determined that there is no statistically significant effect of TE presence on SNM or indel rate, but that TE presence does increase the MNM rate significantly. Potential mechanisms behind this are gene conversion events mediated by transposable elements, though much more work will be needed to assess this possibility.

Together, our results suggest that there is an impact of intragenomic factors on the genome and transcriptome in yeast. These insights provide new evidence for understanding the mutational dynamics in organisms and open up avenues for future studies. Future work would benefit from including an analysis of transposable elements on segmental aneuploidies, which we have observed both in our *S. cerevisiae* aneuploid samples and a diploid *S. paradoxus* (data not shown). In both datasets, the segmental breakpoints were in the same location as transposable elements.

In addition, it would be beneficial to compare diploid samples of S. paradoxus with differing numbers of TEs, as ours were haploid and as such we could not detect recessive lethal mutations. This analysis would also improve power to determine aneuploidy events as well. A comprehensive analysis of different mutational types, including SNMs, indels, MNMs, aneuploidies, and chromosomal rearrangements would be particularly impactful in determining the role of transposable elements on genome dynamics.

References

Audrey P Gasch, J. H., Michael A Newton, Maria Sardi, Mun Yong, Zhishi Wang, 2016 Further support for aneuploidy tolerance in wild yeast and effects of dosage compensation on gene copy-number evolution. eLIFE 5**:** 1-12.

Chen, D., W. M. Toone, J. Mata, R. Lyne, G. Burns *et al.*, 2003 Global transcriptional responses of fission yeast to environmental stress. Mol Biol Cell 14**:** 214-229.

Eduardo M Torres, M. S., Angelika Amon, 2016 No current evidence for widespread dosage compensation in S. cerevisiae. eLIFE 5**:** 1-19.

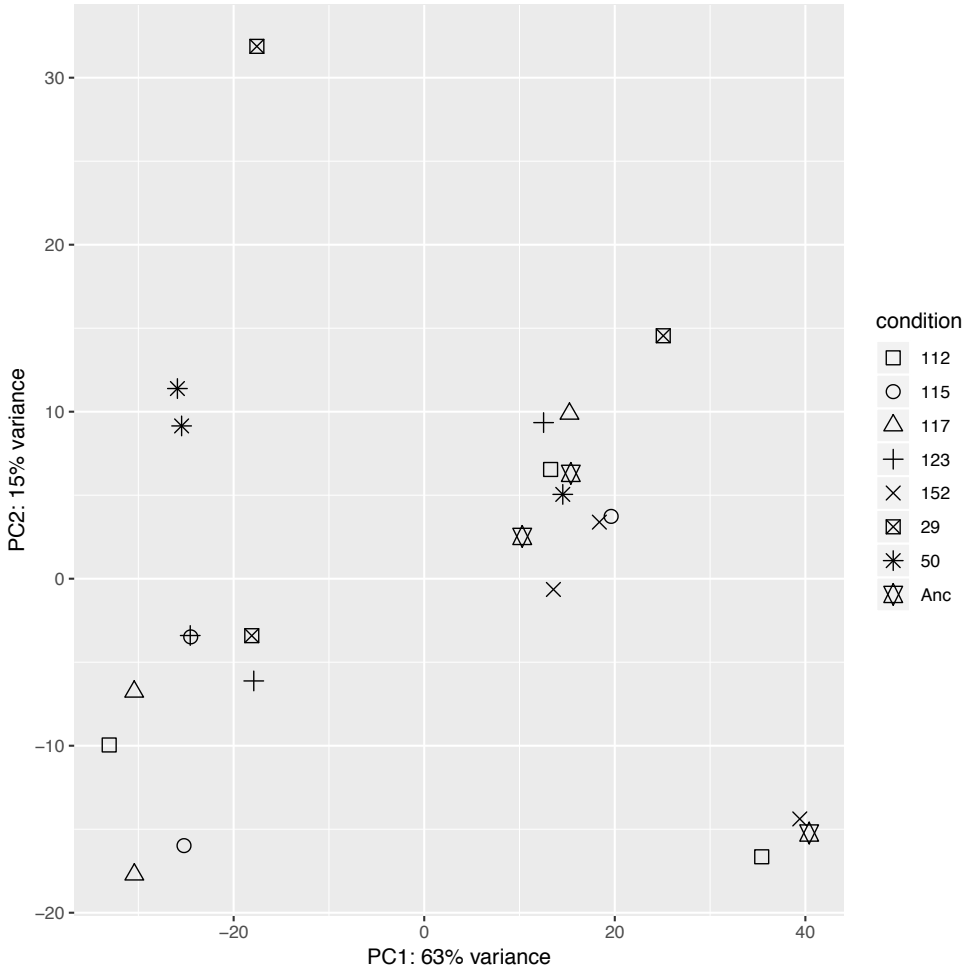
James Hose, C. M. Y., Maria Sardi, Zhishi Wang, Michael A Newton, Audrey P Gasch, 2015 Dosage compensation can buffer copy-number variation in yeast. eLIFE 4**:** 1-27.

Makanae, K., R. Kintaka, T. Makino, H. Kitano and H. Moriya, 2013 Identification of dosage-sensitive genes in Saccharomyces cerevisiae using the genetic tug-of-war method. Genome research 23**:** 300-311.

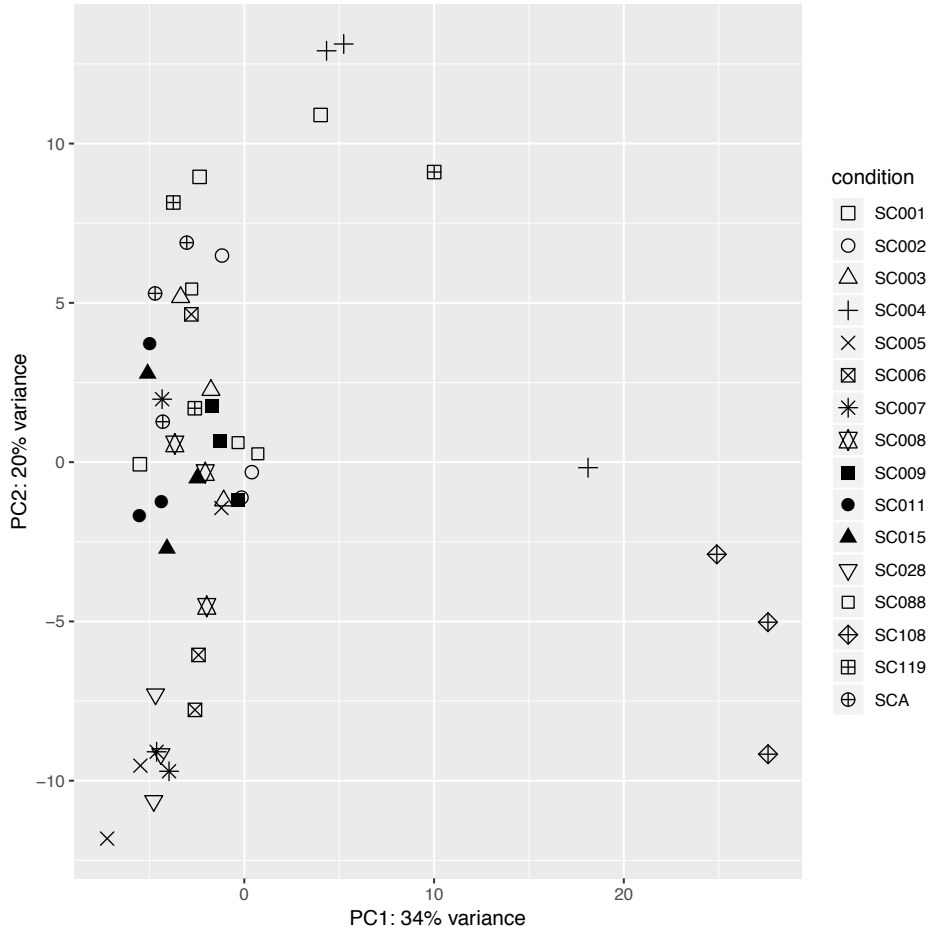
Torres, E. M., N. Dephoure, A. Panneerselvam, C. M. Tucker, C. A. Whittaker *et al.*, 2010 Identification of aneuploidy-tolerating mutations. Cell 143**:** 71-83.

Appendix A

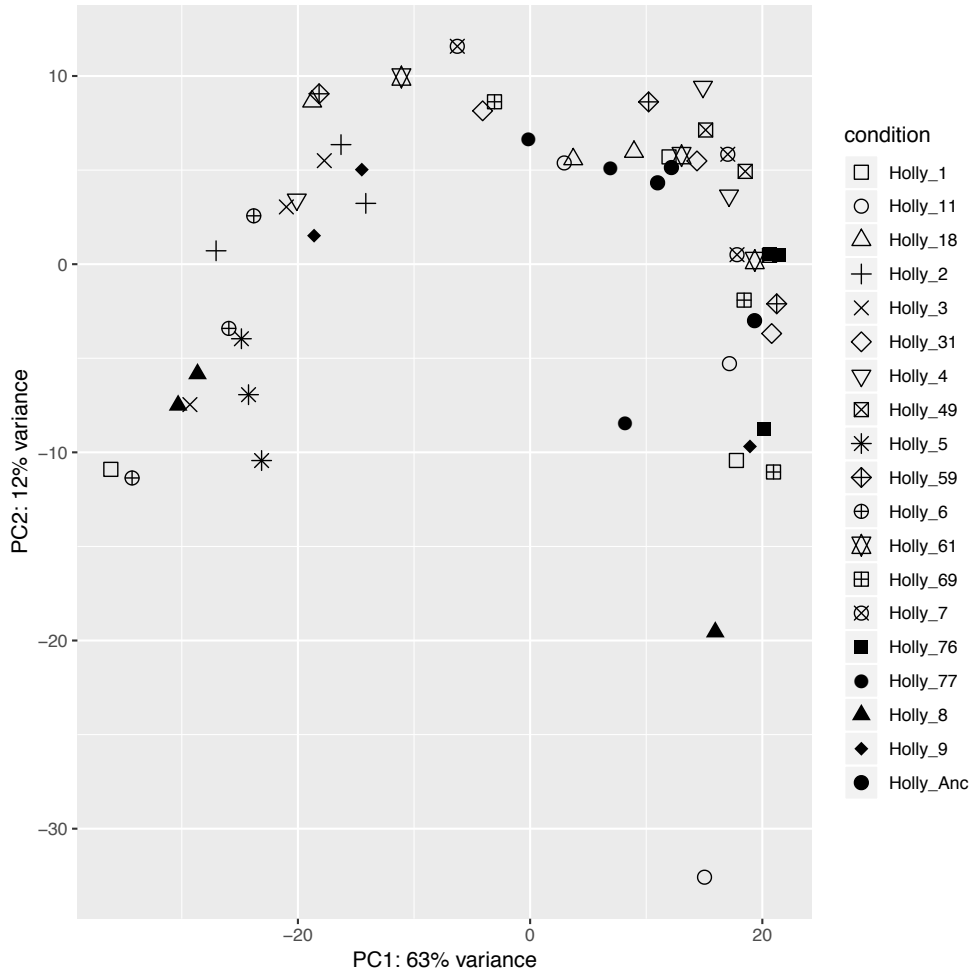
Supplemental Material for Chapter 2



PCA of second batch of homozygous ancestor samples. One ancestral replicate (near bottom right) was removed from analysis based on PC1.

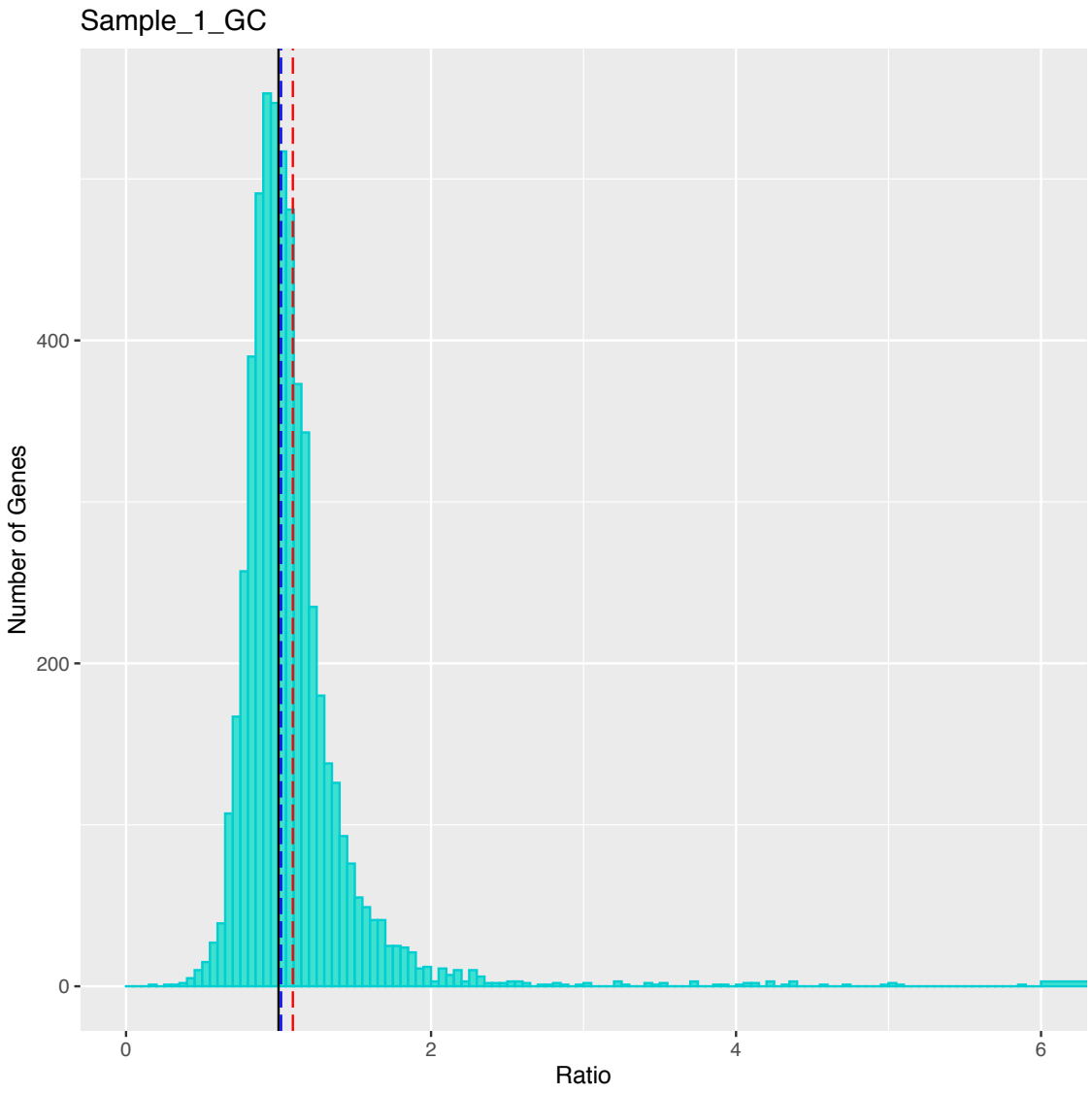


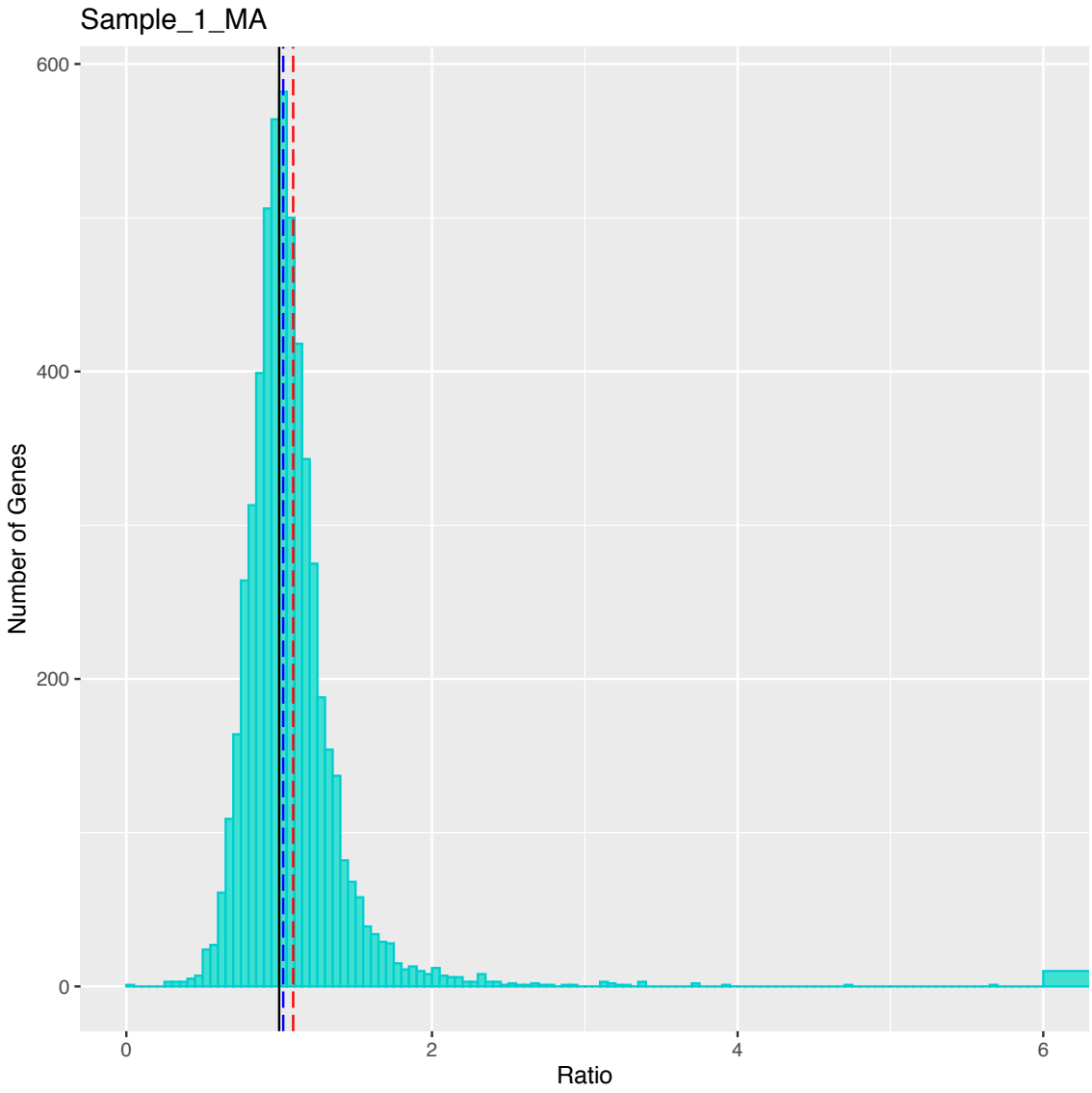
PCA of first homozygous ancestor sequencing run.

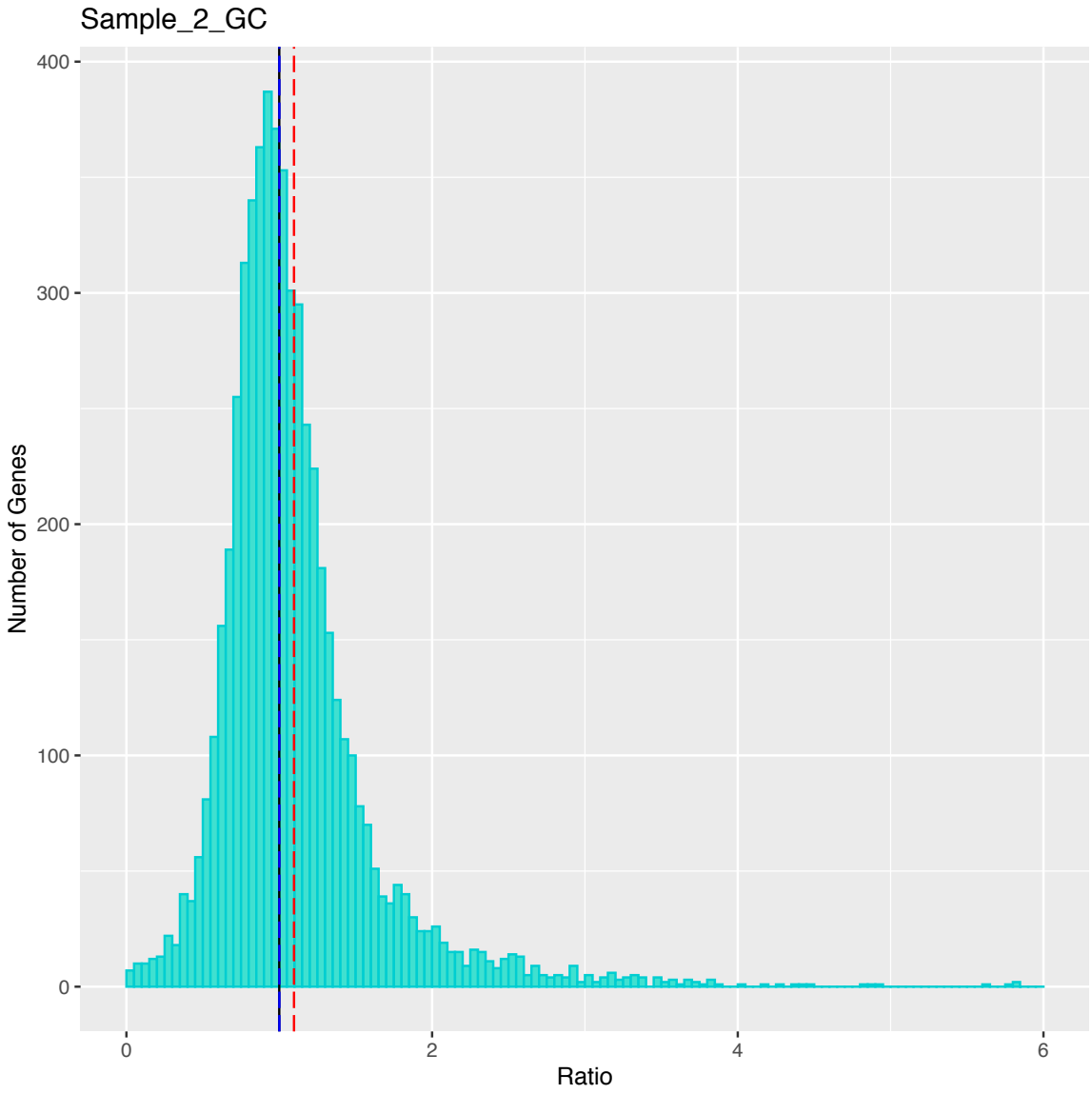


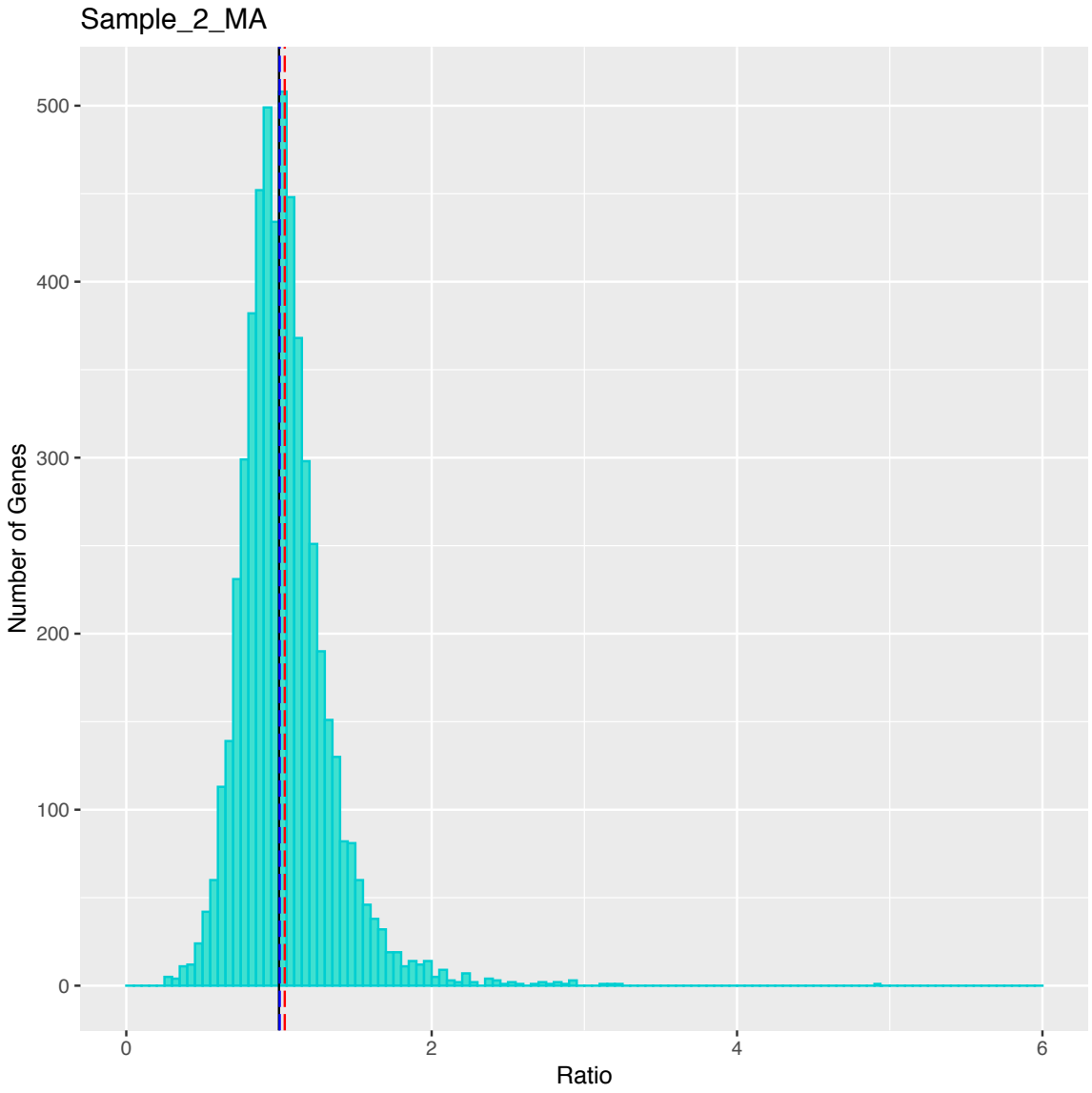
PCA of heterozygous ancestor sequencing run.

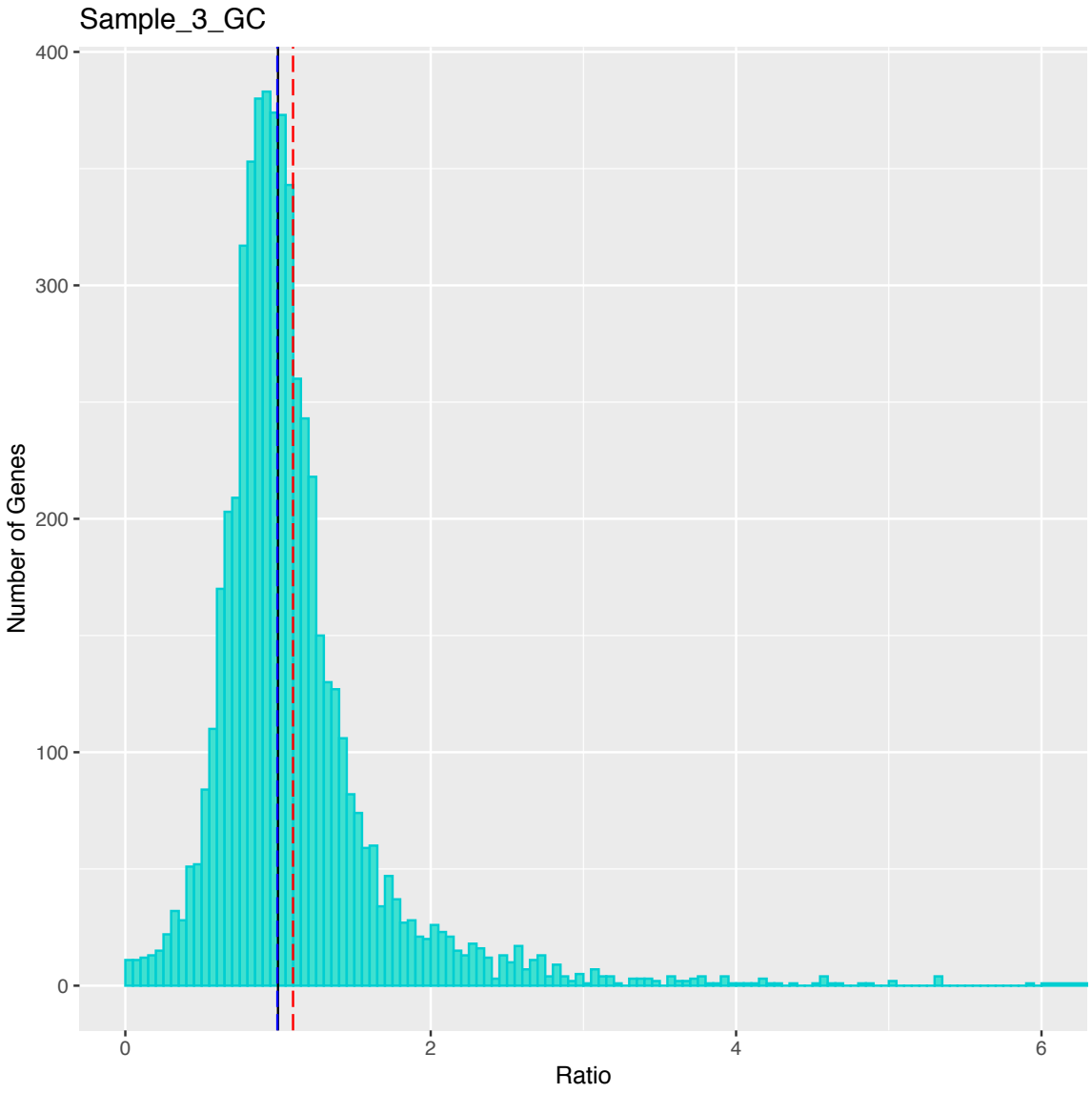
Supplemental Figure 2.1: Principal components analysis of all lines sequenced.

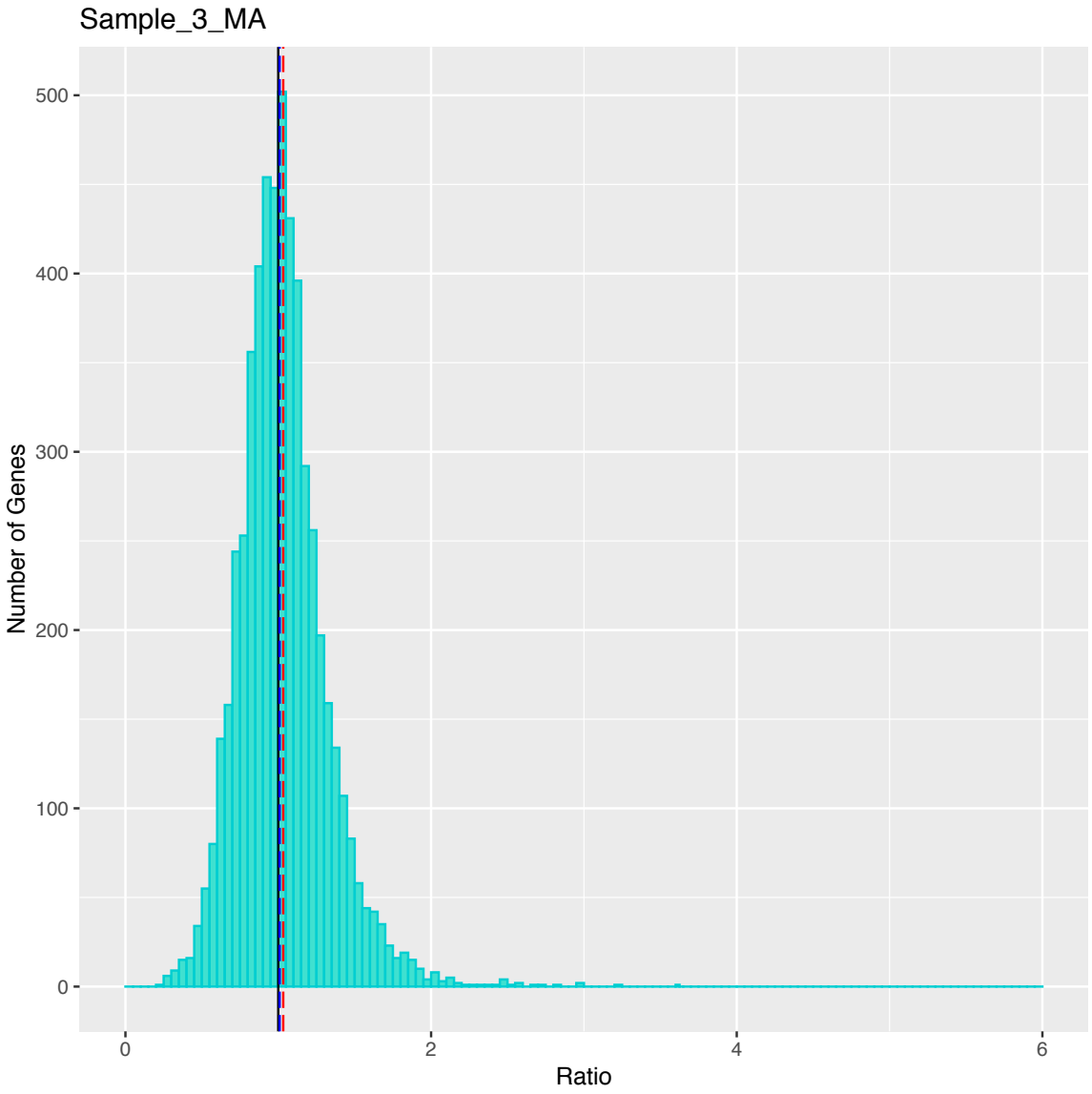

Sample_1_GC
Number of Genes
400
200
0
0
2
4
6
Ratio

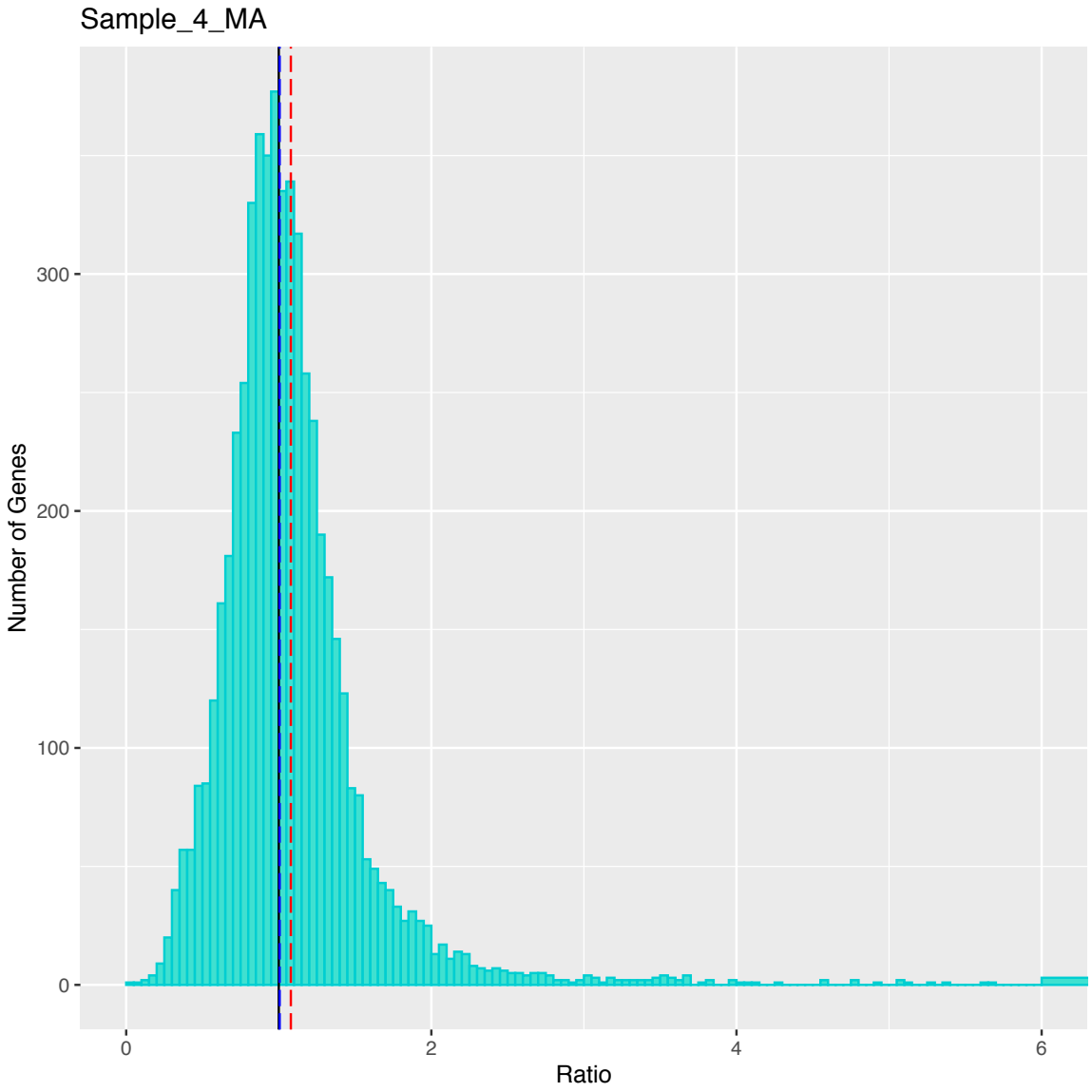

Sample_1_MA
600
400
200
0
Number of Genes
0
2
4
6
Ratio

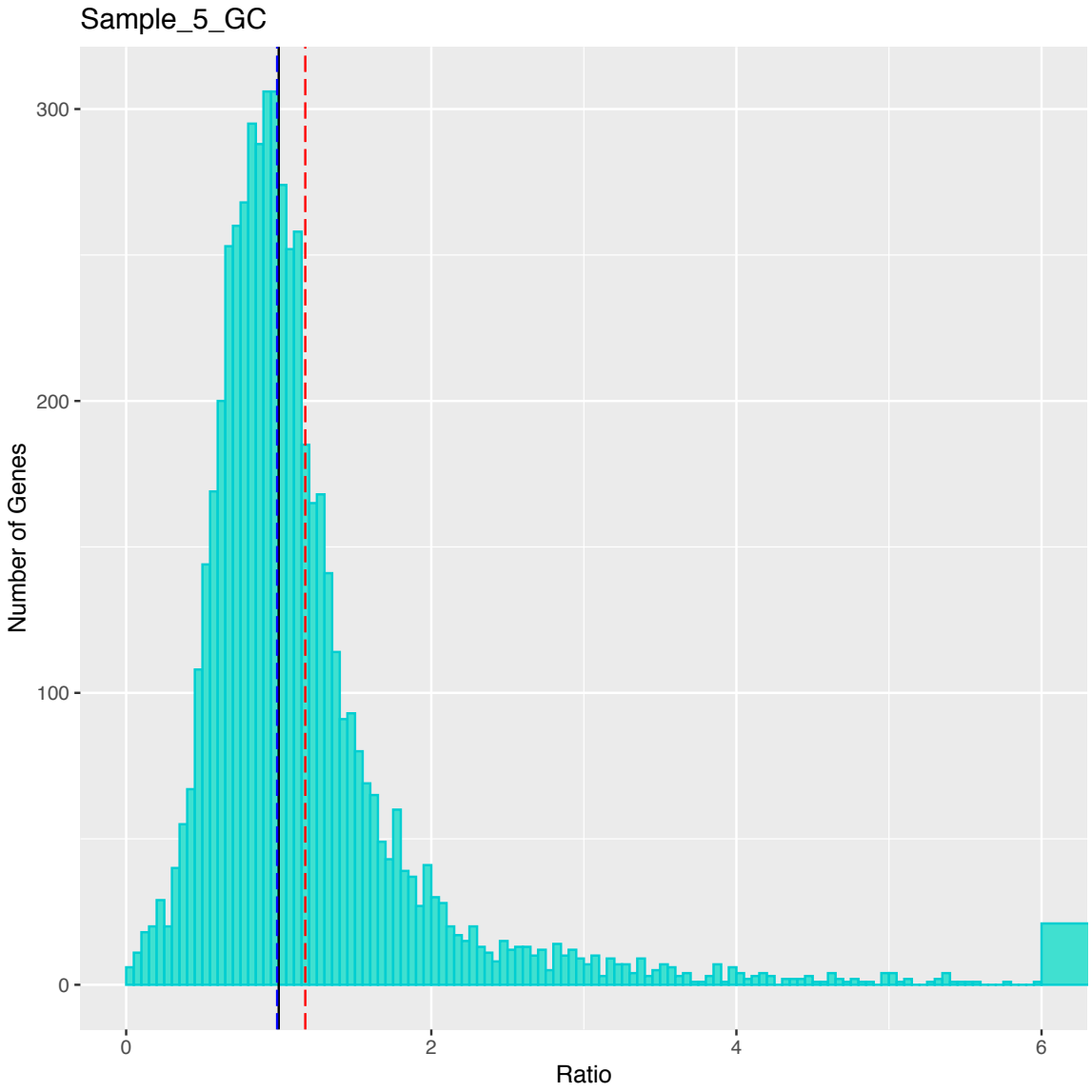

Sample_2_GC
Number of Genes
400
300
200
100
0
0
2
4
6
Ratio

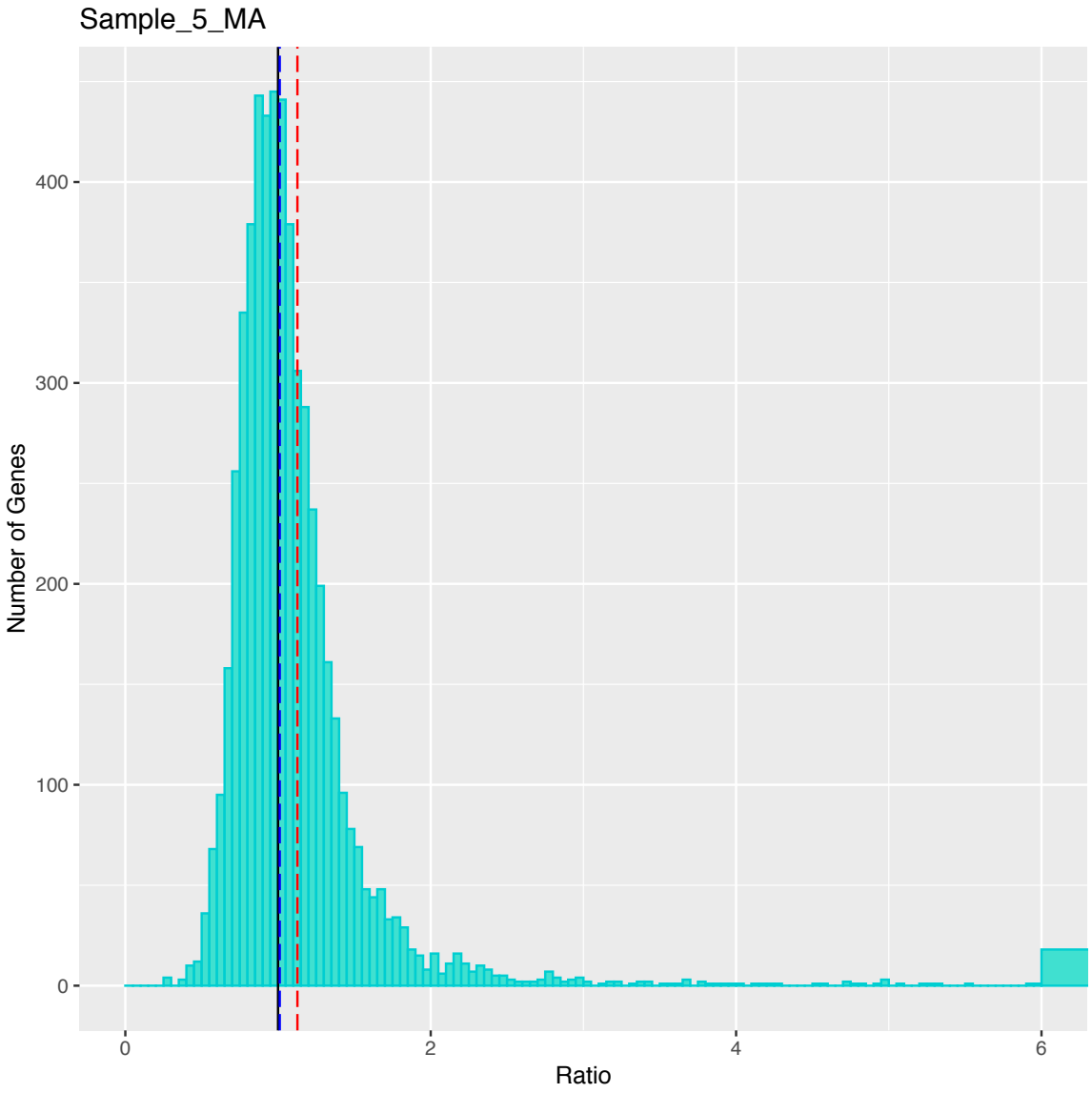

Sample_2_MA
Number of Genes
500
400
300
200
100
0
0
2
4
6
Ratio

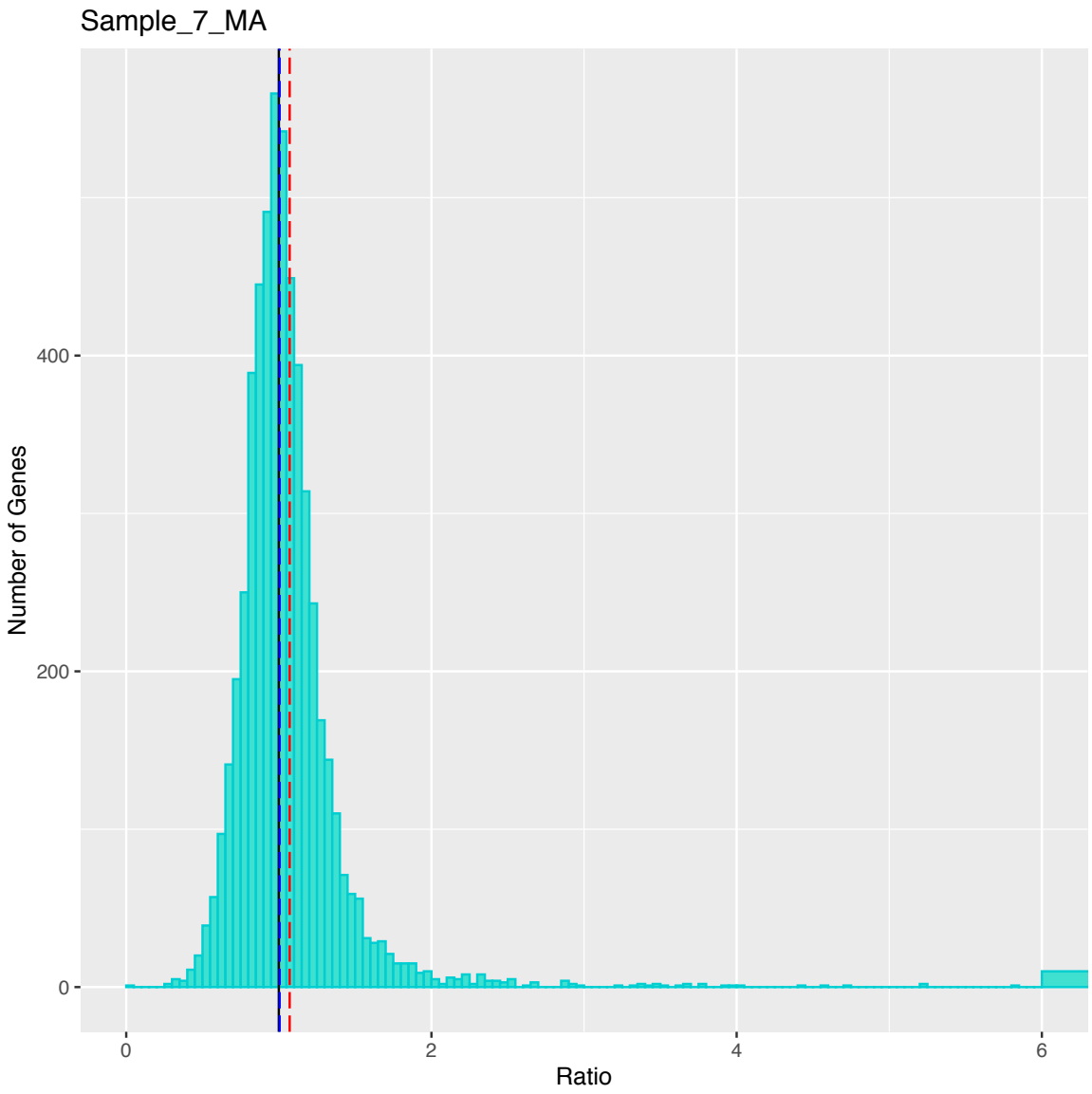

Sample_3_GC
Number of Genes
400
300
200
100
0
0
2
4
6
Ratio

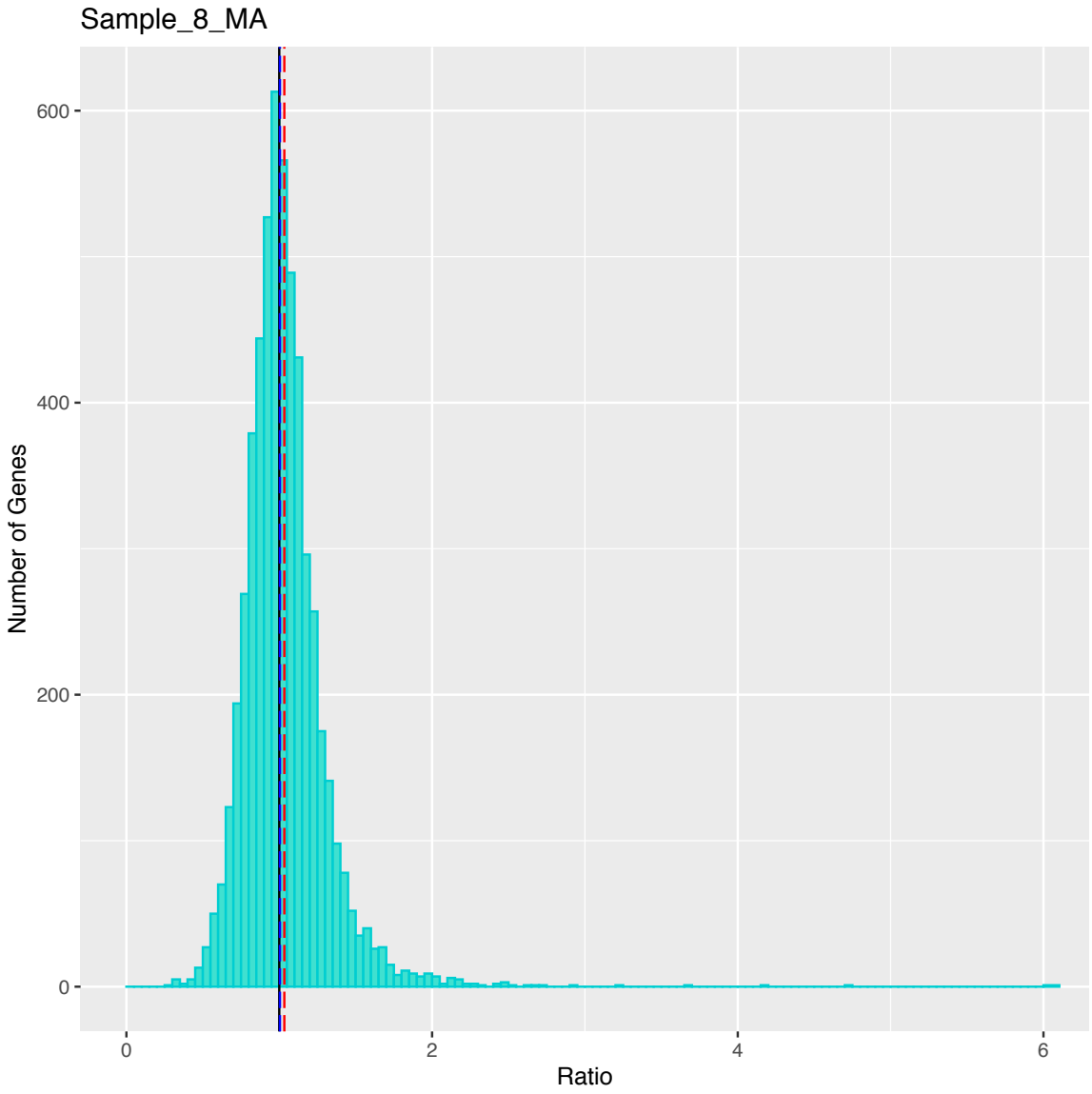

Sample_3_MA
Number of Genes
500
400
300
200
100
0
0
2
4
6
Ratio

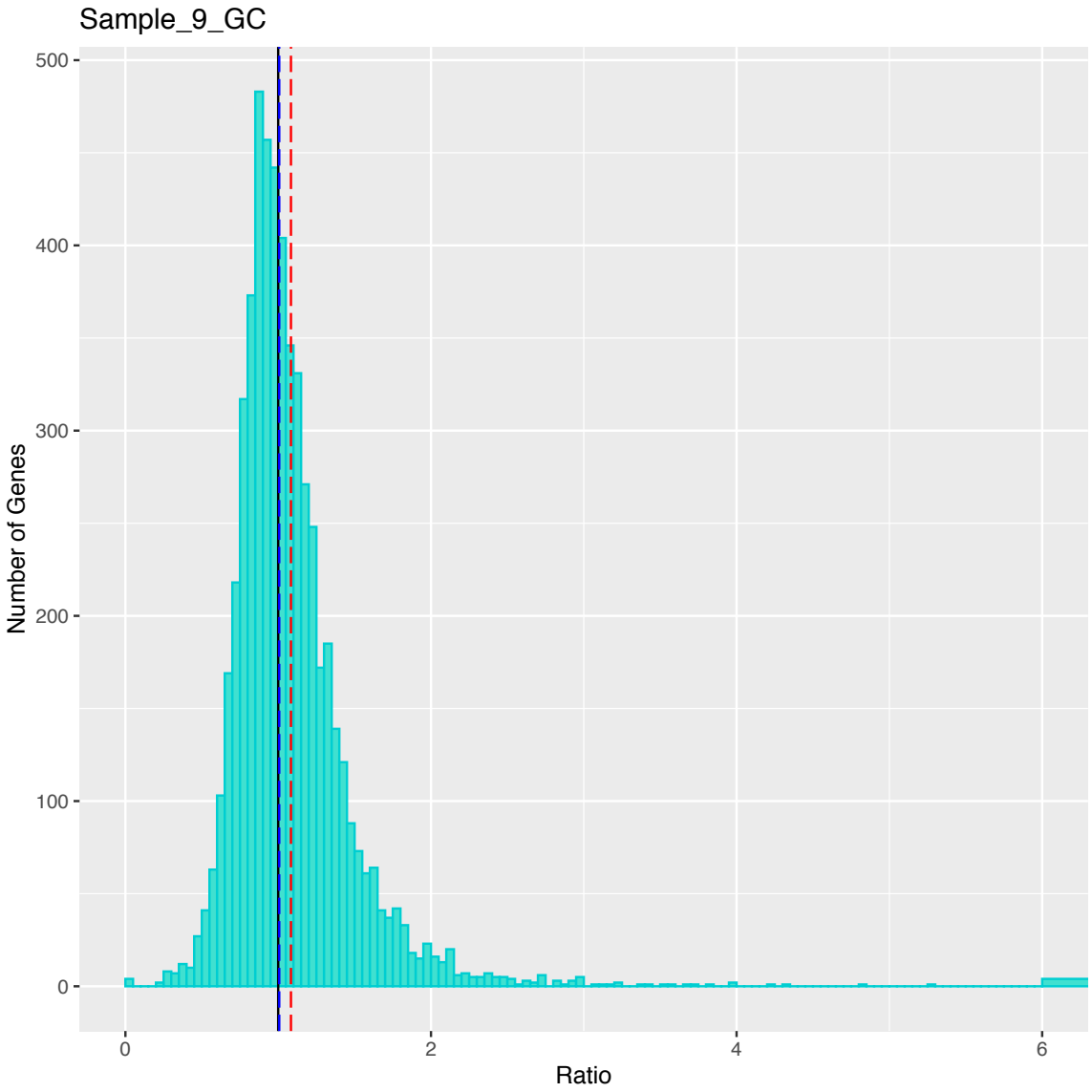

Sample_4_MA
Number of Genes
300
200
100
0
0
2
4
6
Ratio

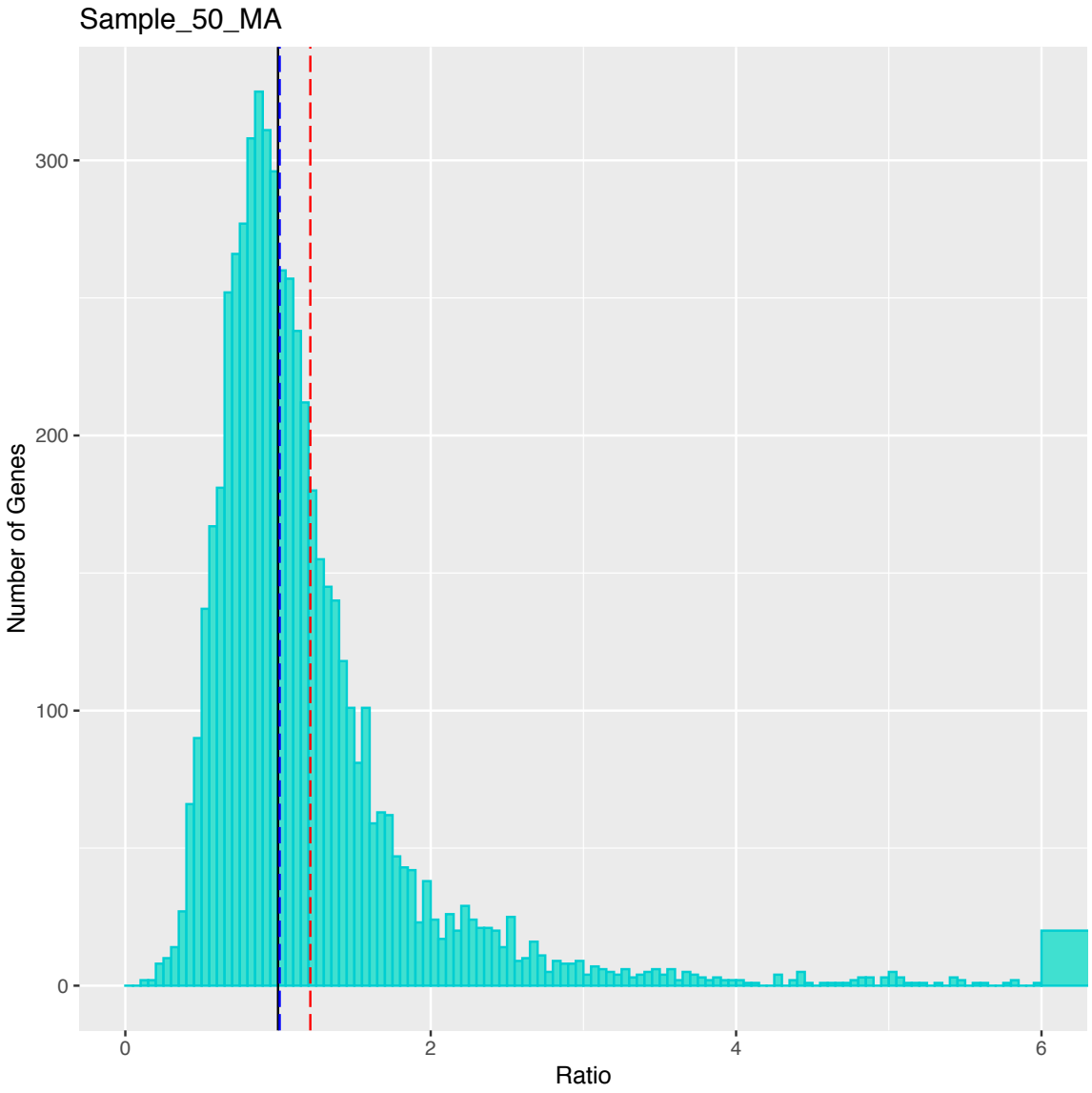

Sample_5_GC
Number of Genes
300
200
100
0
0
2
4
6
Ratio

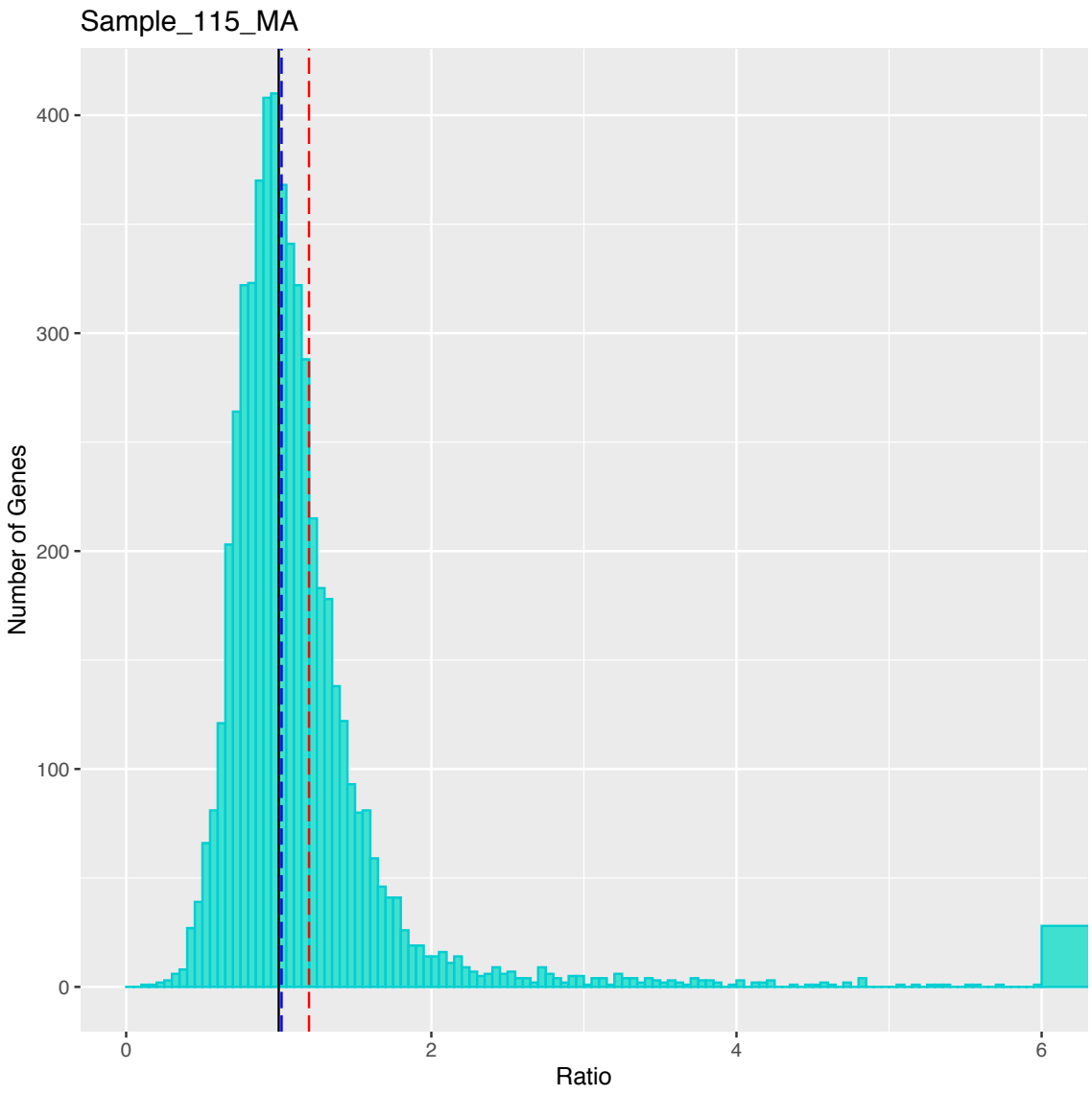

Sample_5_MA
Number of Genes
400
300
200
100
0
0
2
4
6
Ratio


Sample_7_MA
Number of Genes
400
200
0
0
2
4
6
Ratio

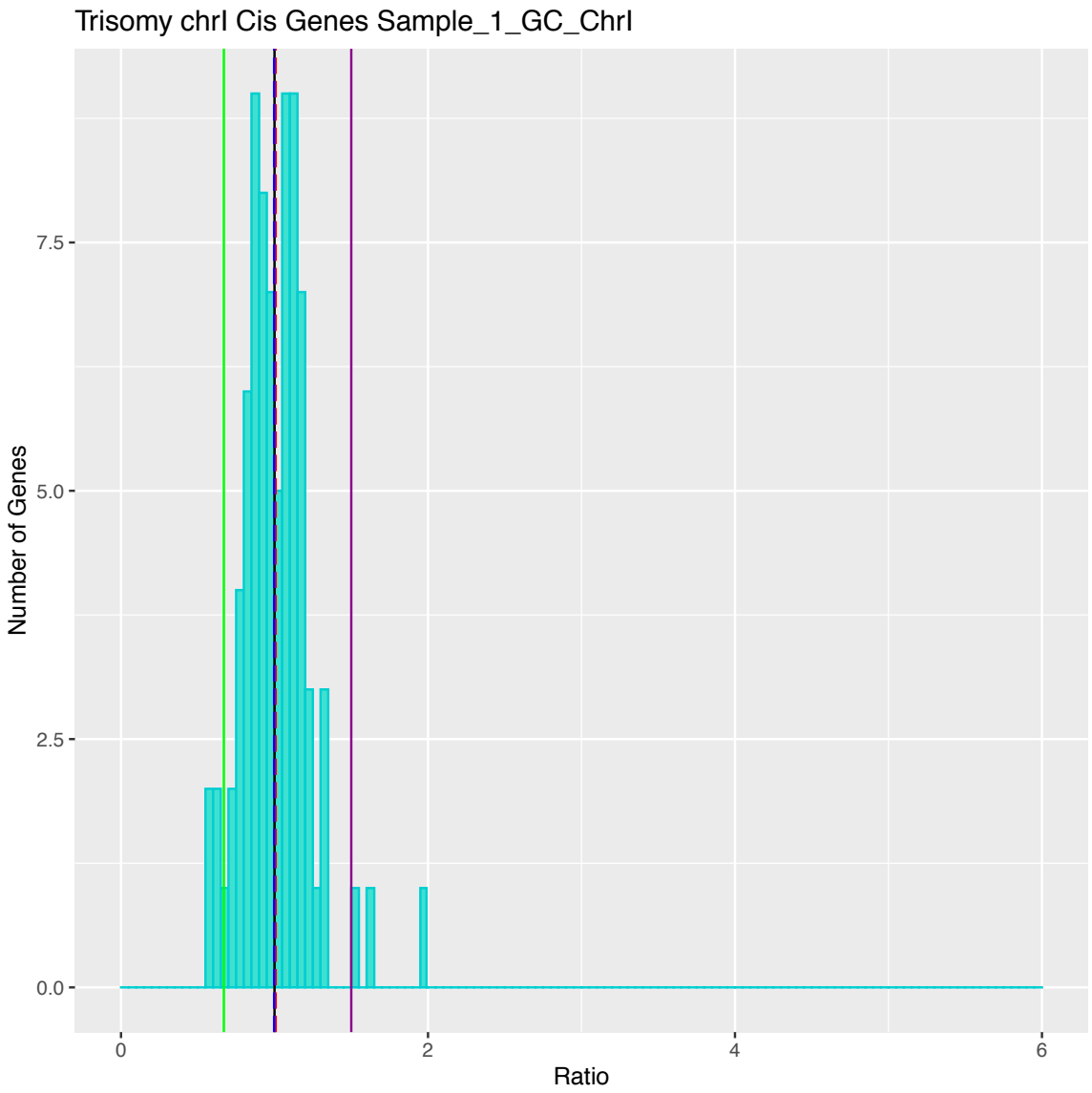

Sample_8_MA
600
400
200
0
Number of Genes
0
2
4
6
Ratio

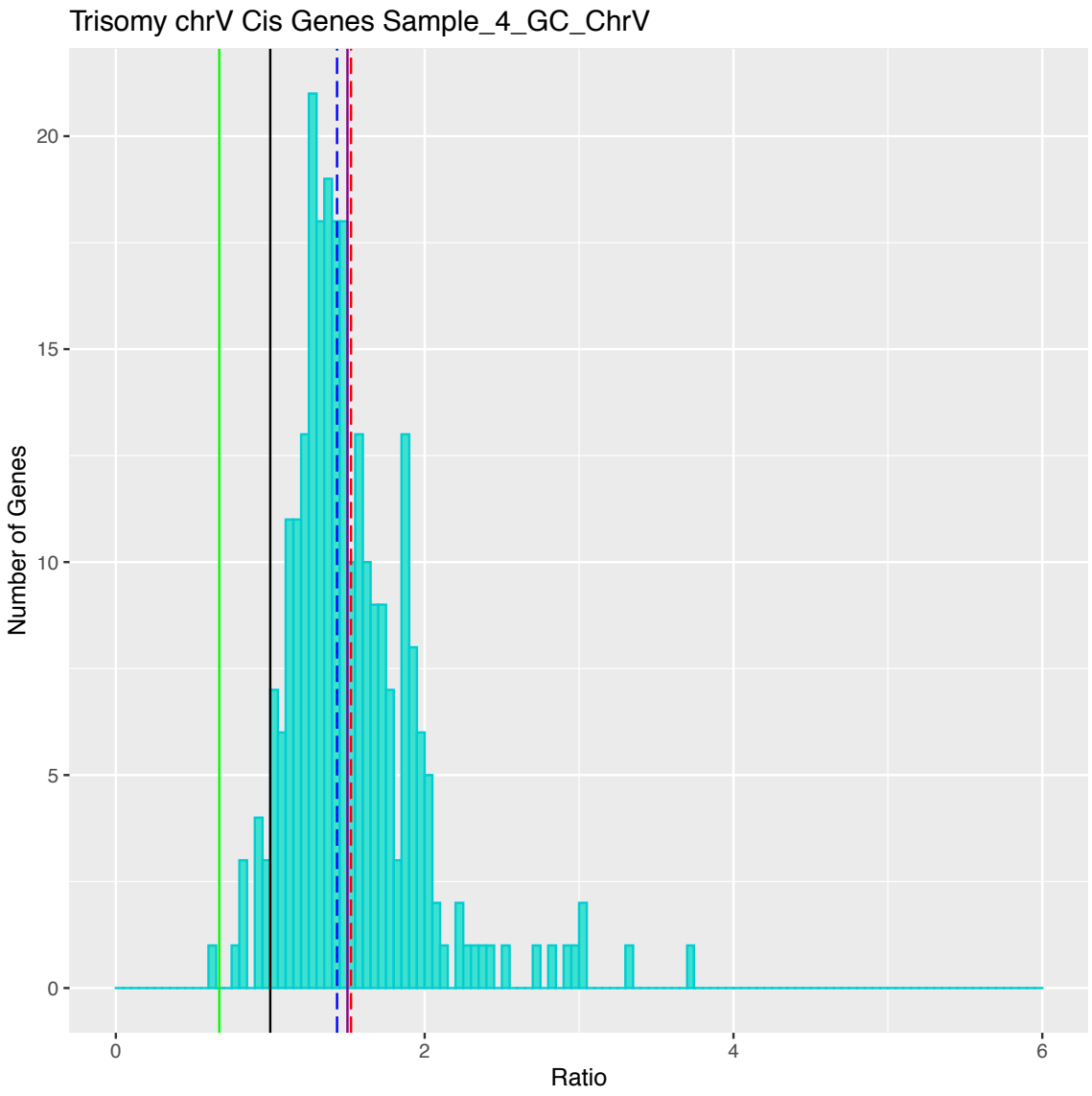

Sample_9_GC
Number of Genes
500
400
300
200
100
0
0
2
4
6
Ratio

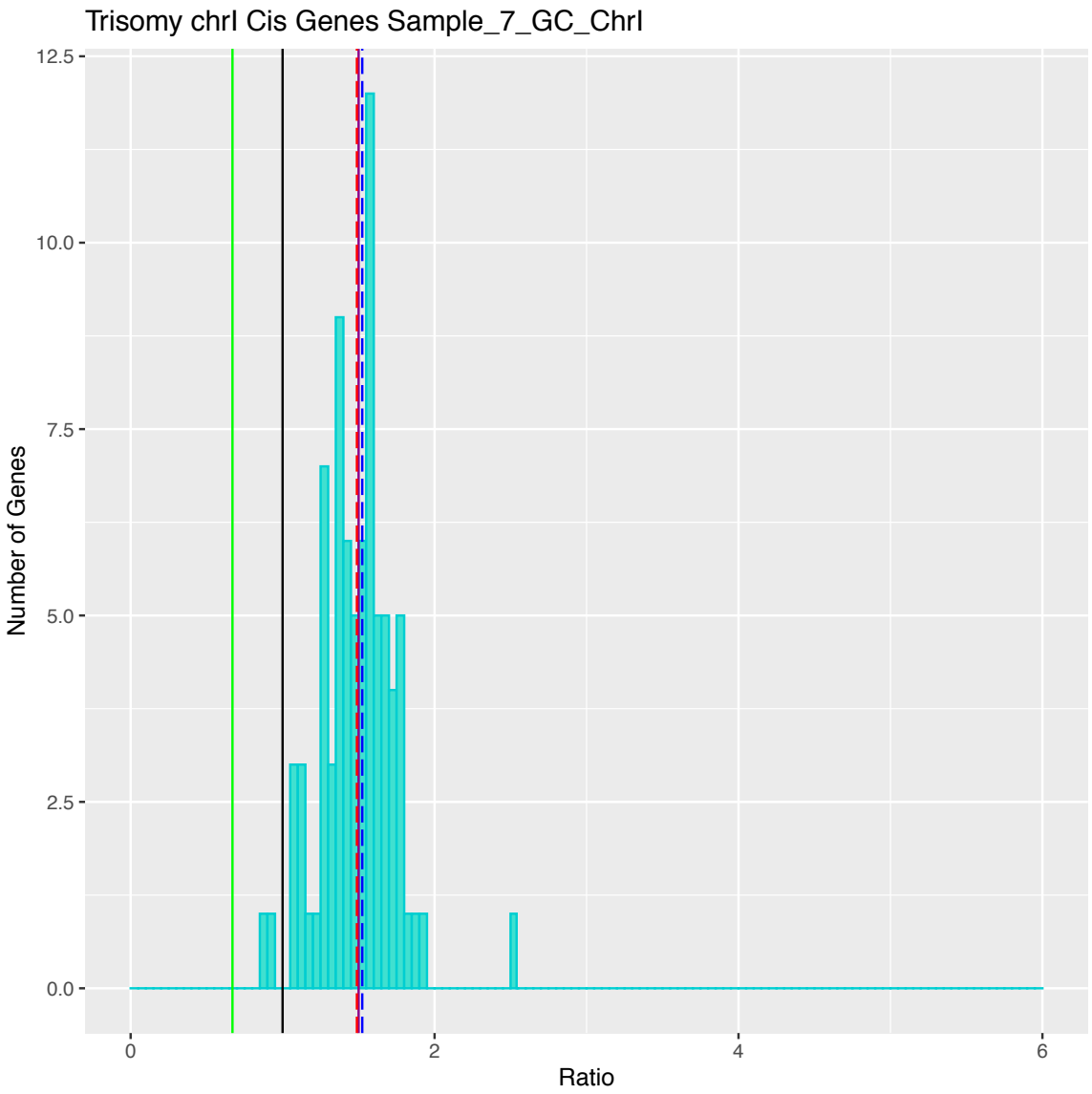

Sample_50_MA
Number of Genes
300
200
100
0
0
2
4
6
Ratio

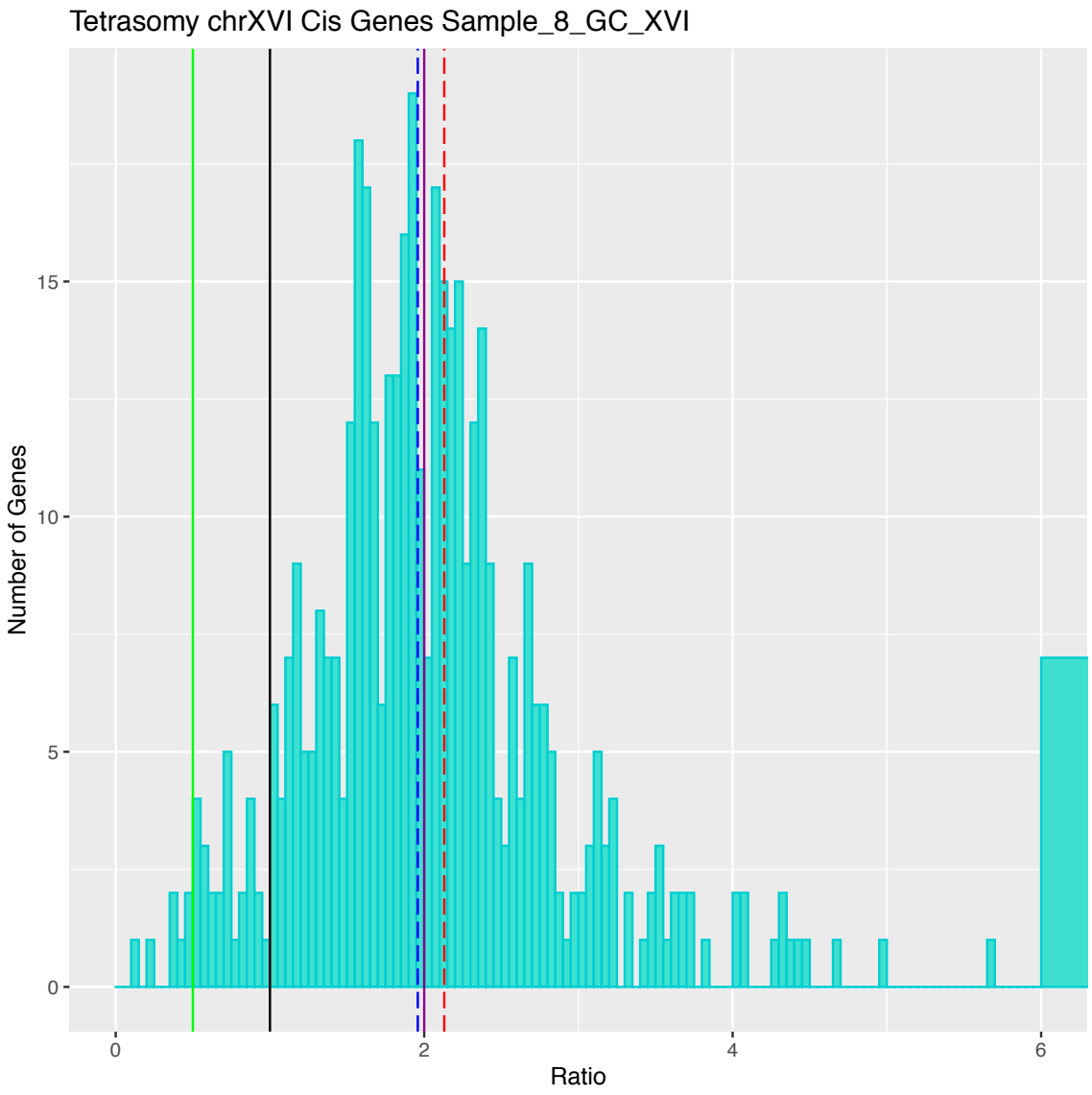


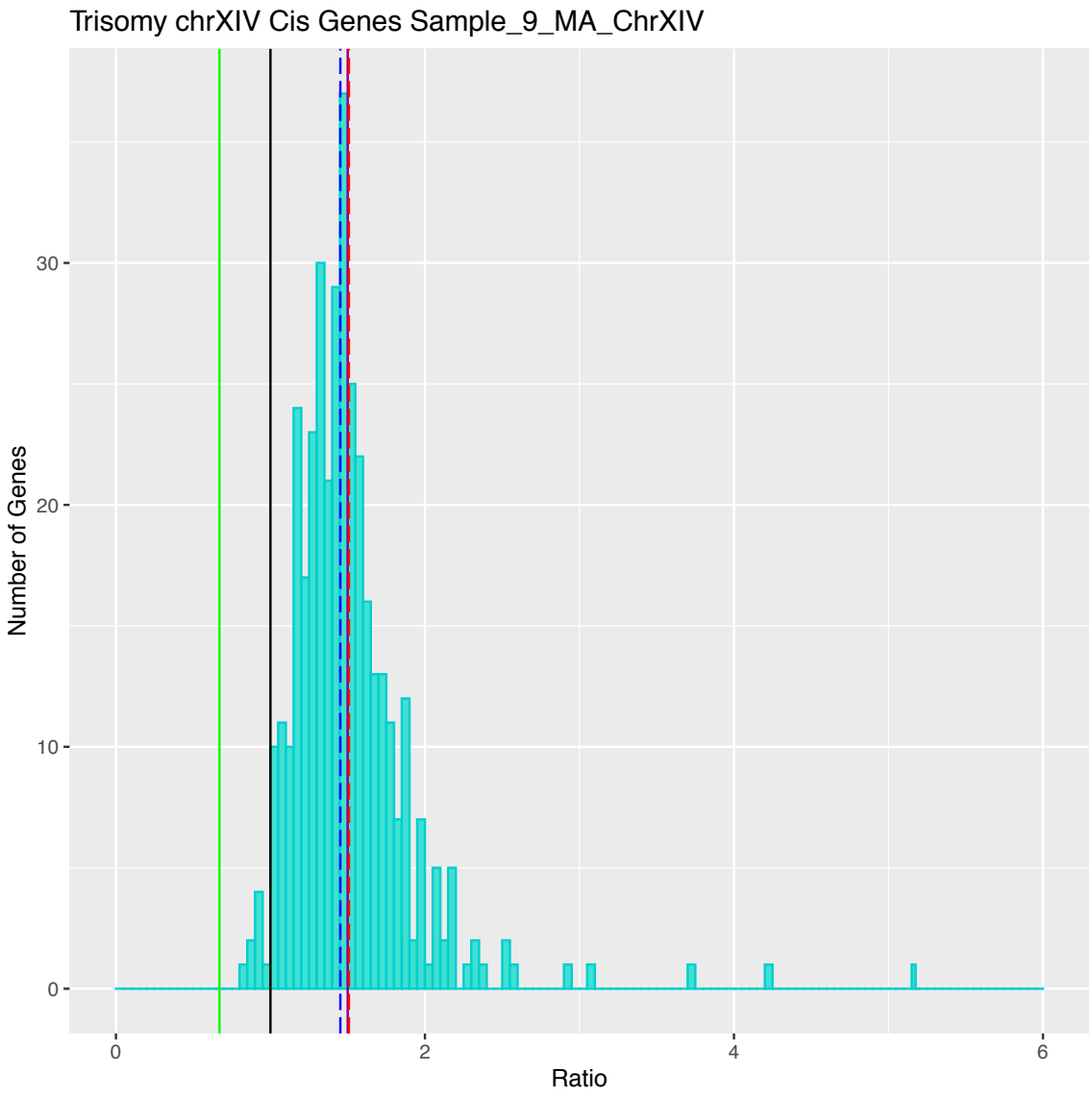
Supplemental Figure 2.2: Histograms of gene expression ratios for every euploid sample. Blue dashed line represents median gene expression ratio, dashed red line represents mean gene expression ratio, and black solid line represents the expected ratio if there is no difference between the ancestor and the given sample. Distributions produced after filtering.

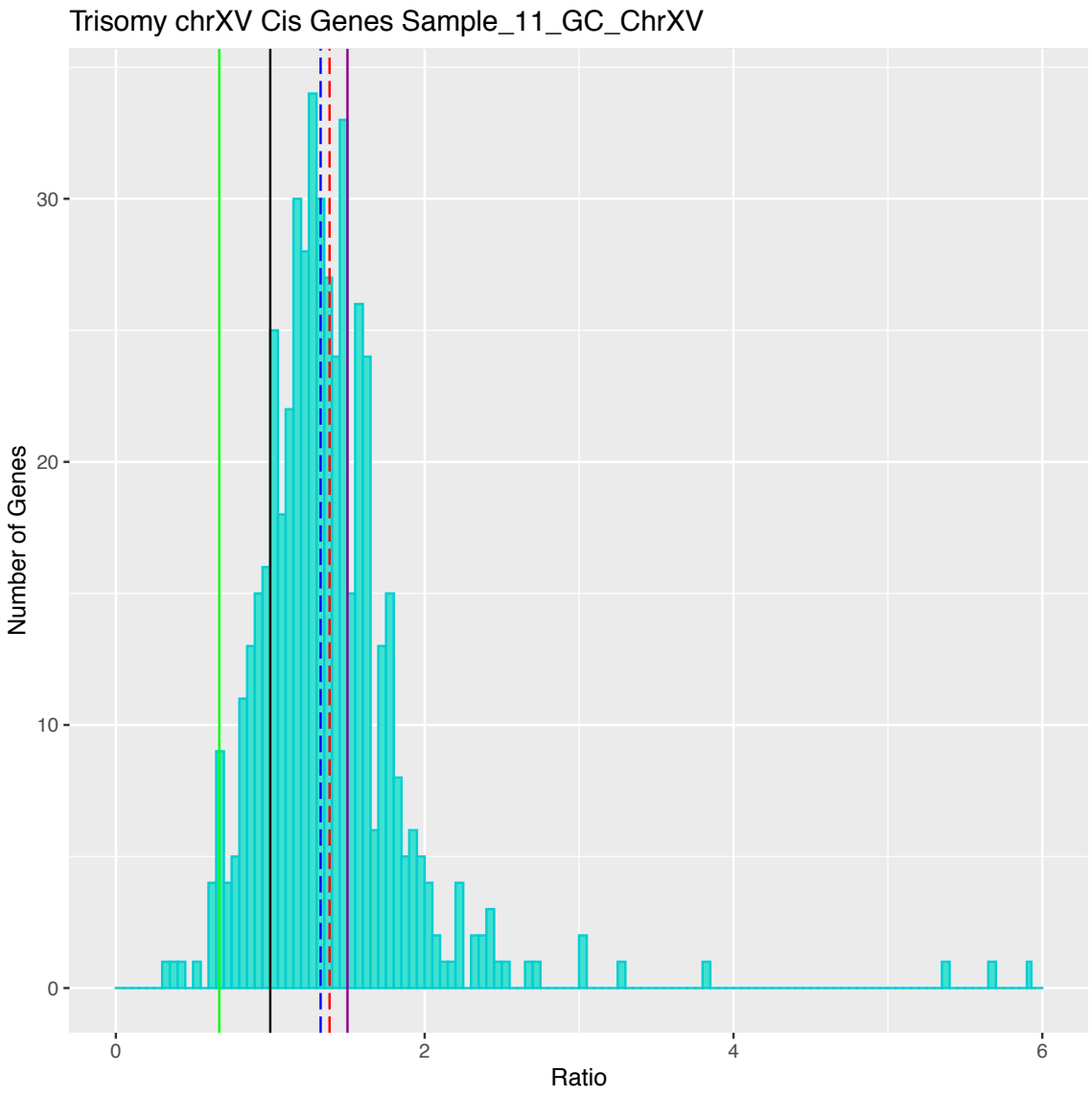

Trisomy chrI Cis Genes Sample_1_GC_ChrI
Number of Genes
7.5
5.0
2.5
0.0
0
2
4
6
Ratio

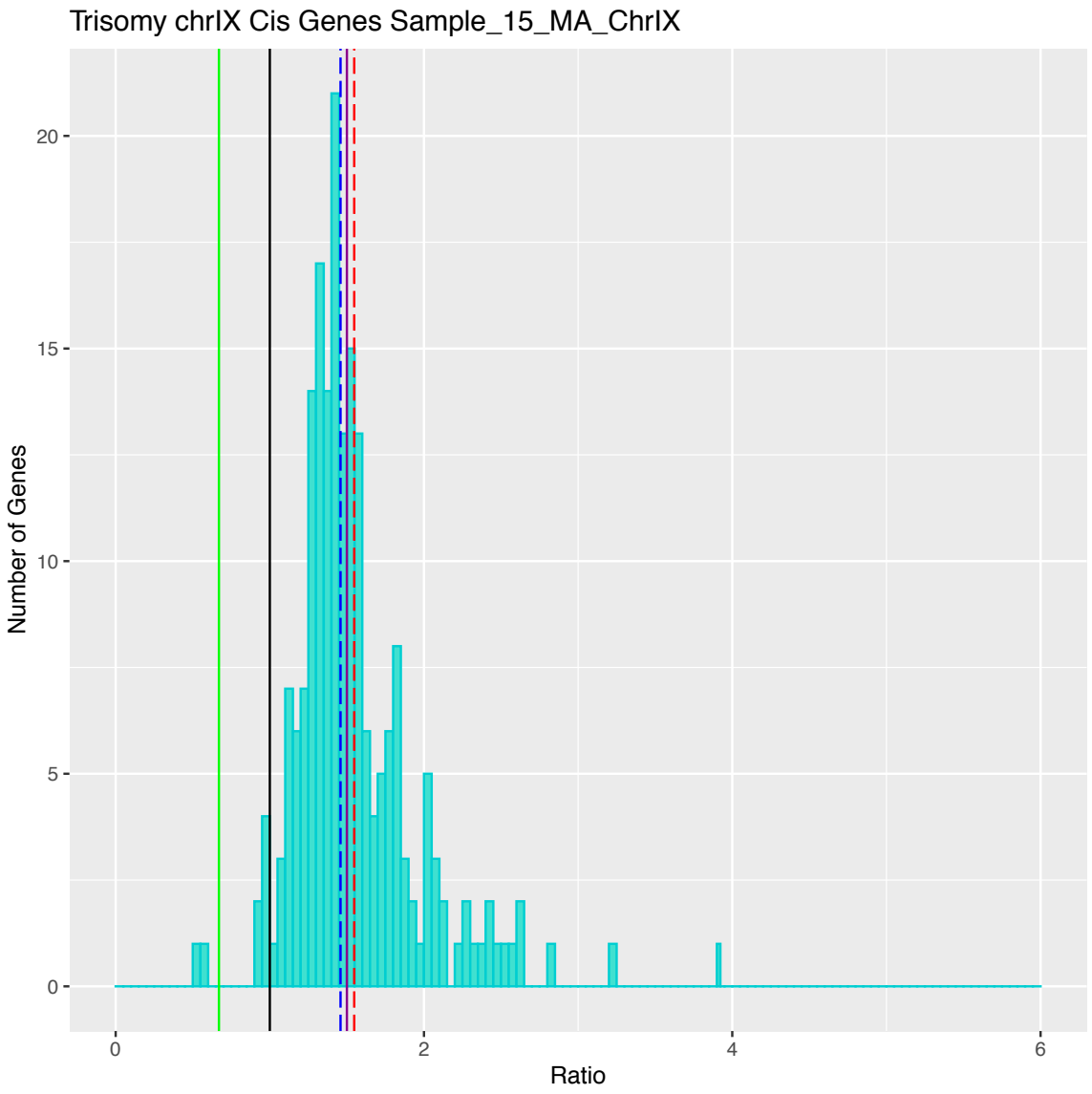

Trisomy chrV Cis Genes Sample_4_GC_ChrV
Number of Genes
0
5
10
15
20
0
2
4
6
Ratio

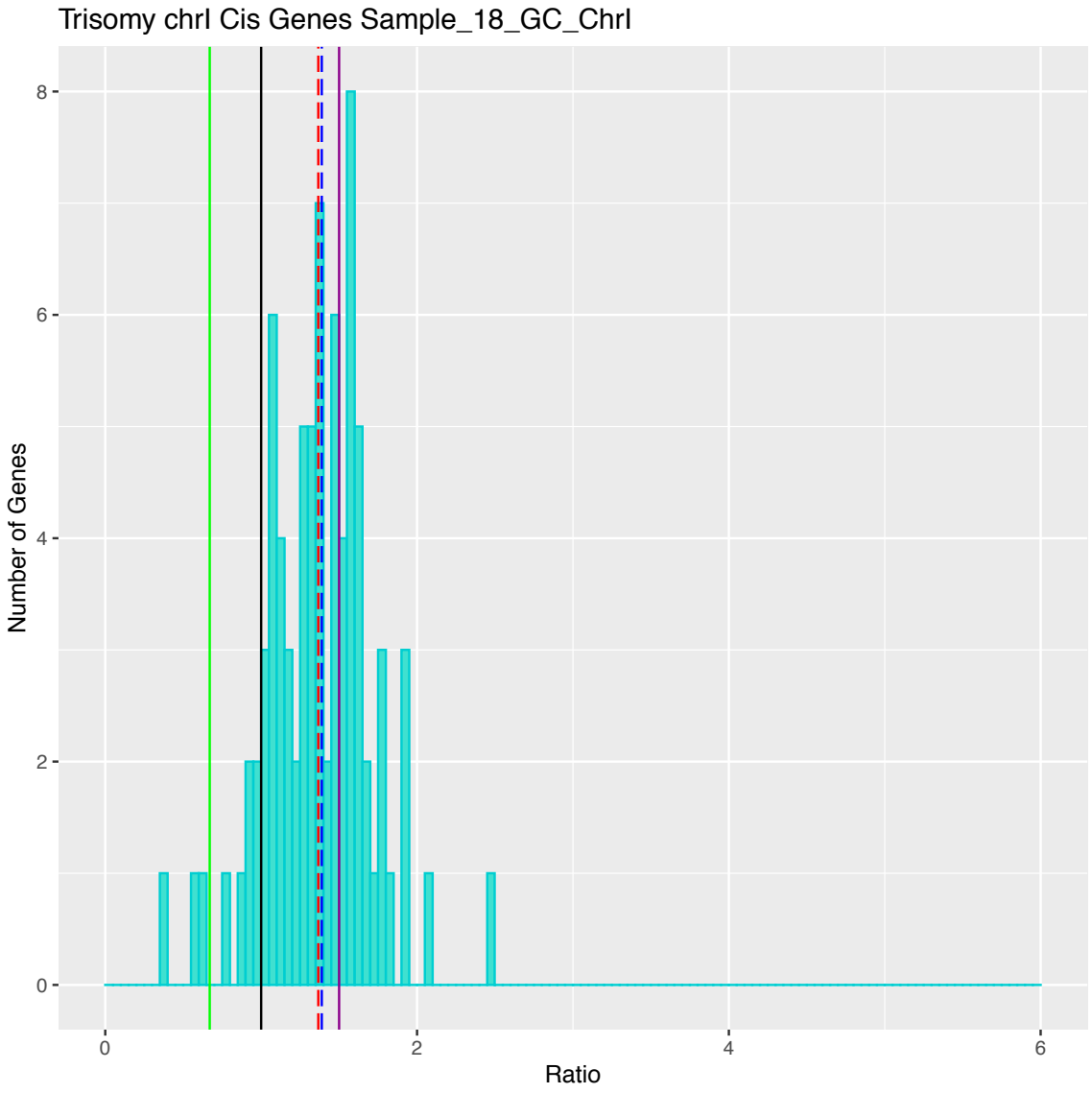

Trisomy chrI Cis Genes Sample_7_GC_ChrI
Number of Genes
12.5
10.0
7.5
5.0
2.5
0.0
0
2
4
6
Ratio

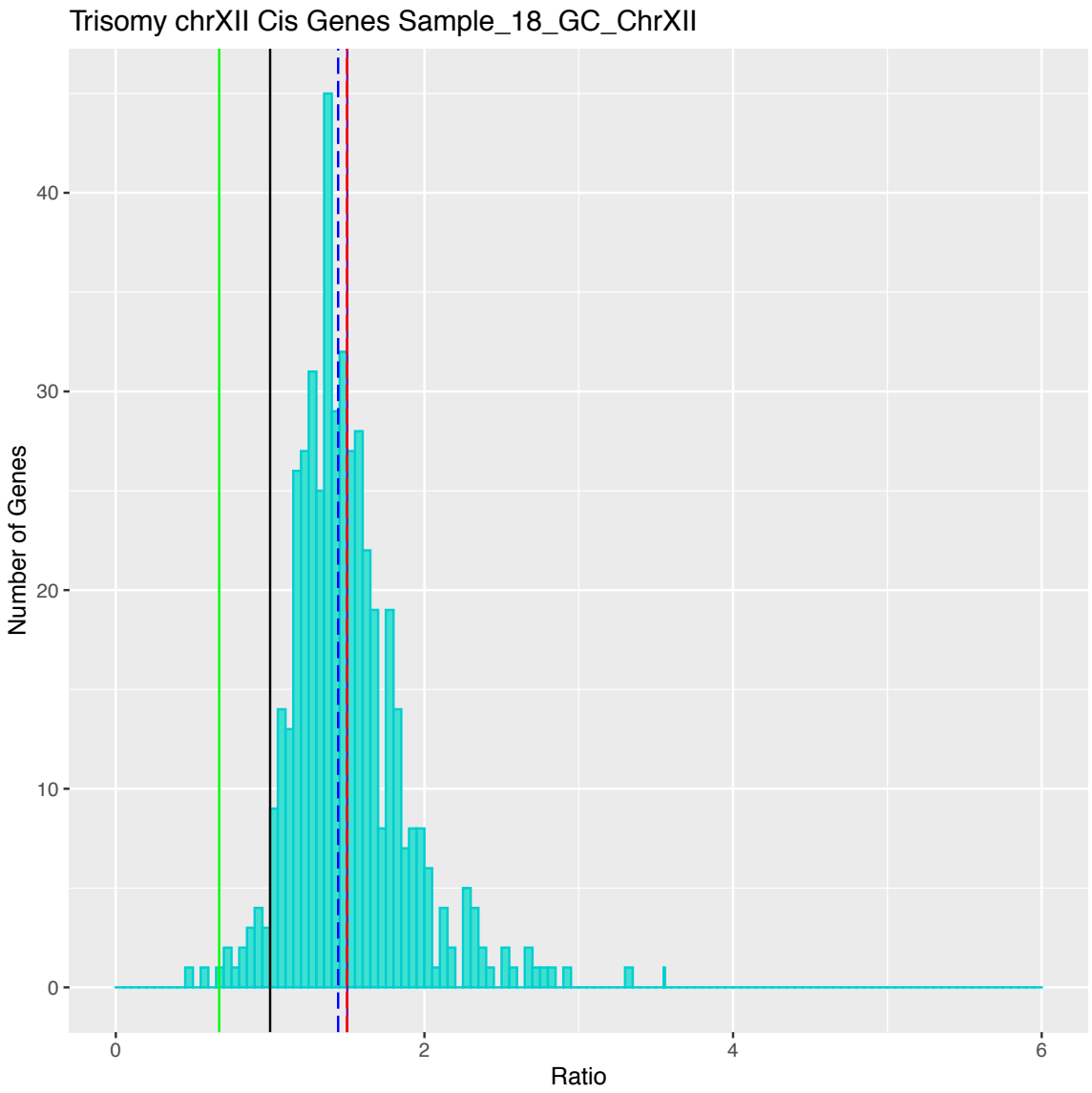

Tetrasomy chrXVI Cis Genes Sample_8_GC_XVI
Number of Genes
0
5
10
15
0
2
4
6
Ratio

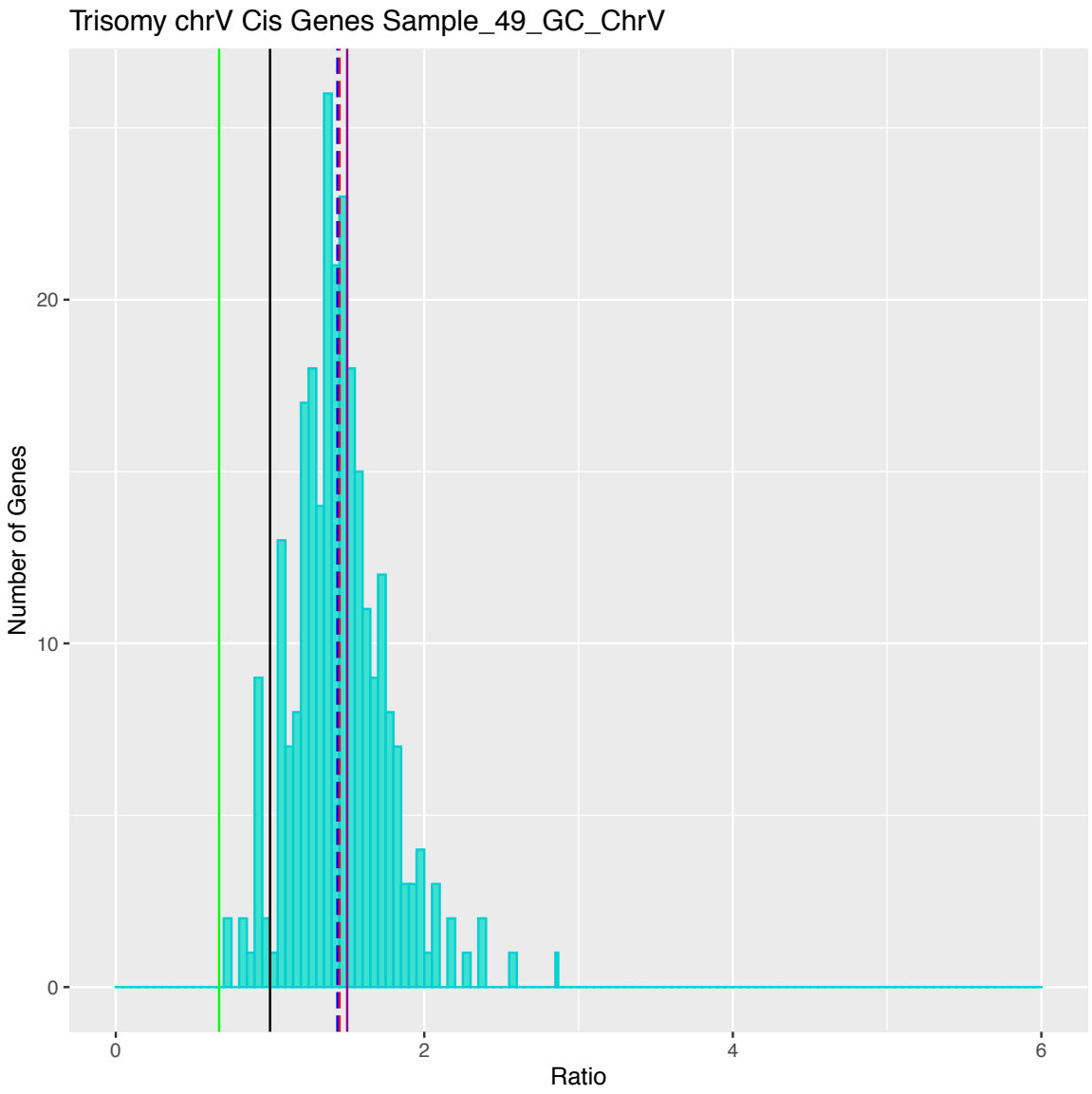

Trisomy chrXIV Cis Genes Sample_9_MA_ChrXIV
Number of Genes
30
20
10
0
0
2
4
6
Ratio

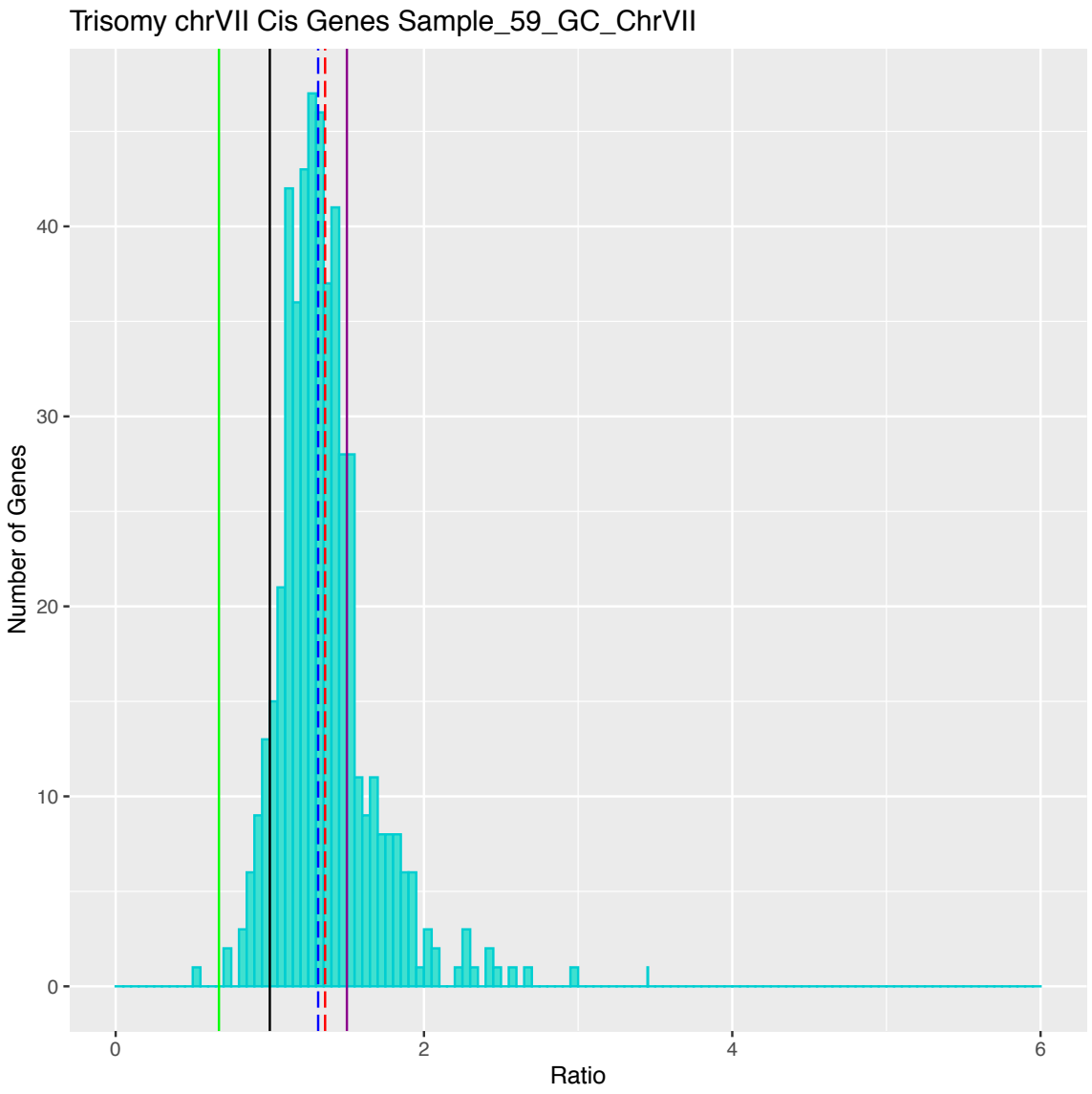

Trisomy chrXV Cis Genes Sample_11_GC_ChrXV
Number of Genes
30
20
10
0
0
2
4
6
Ratio

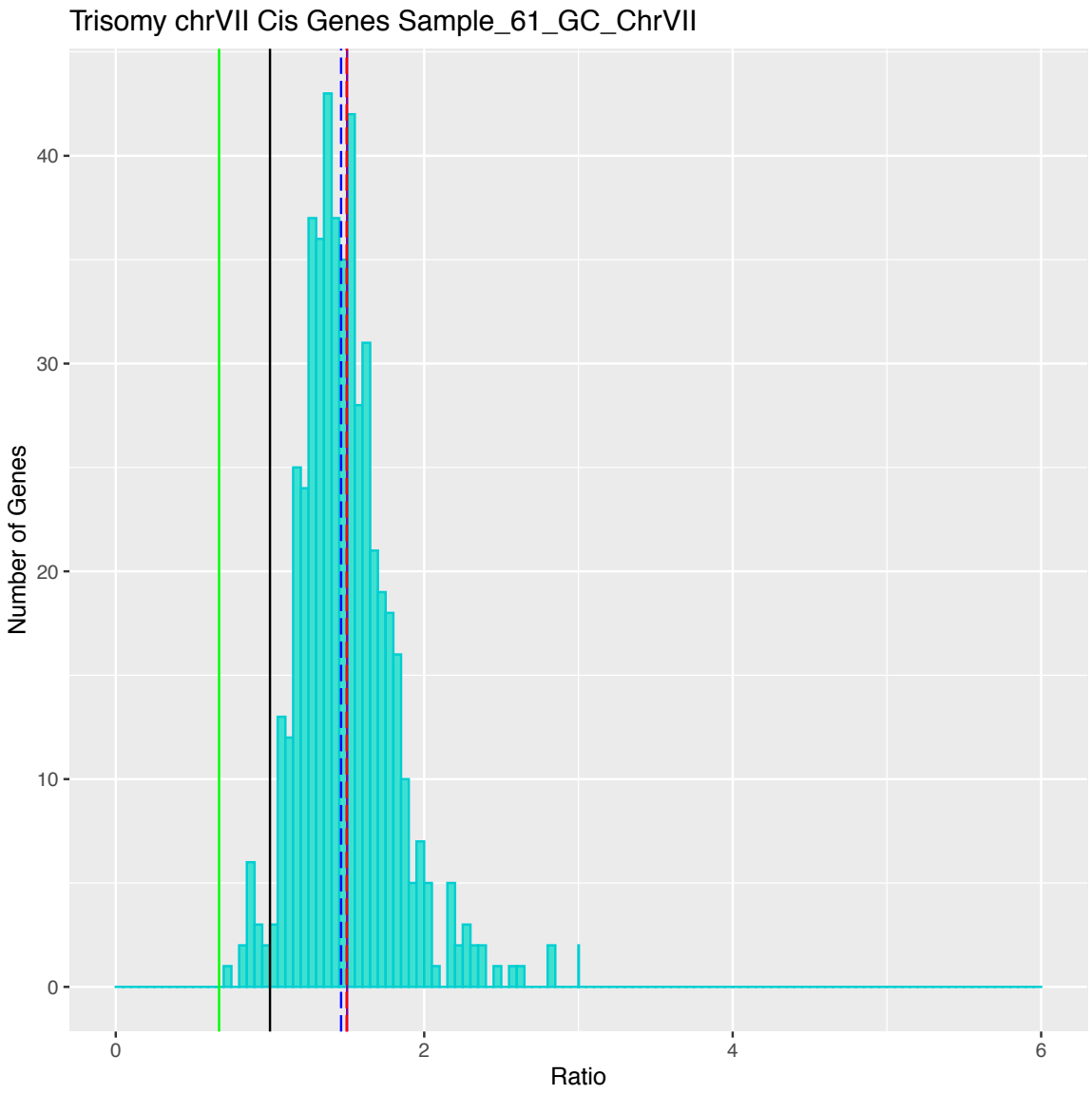

Trisomy chrIX Cis Genes Sample_15_MA_ChrIX
Number of Genes
0
5
10
15
20
0
2
4
6
Ratio

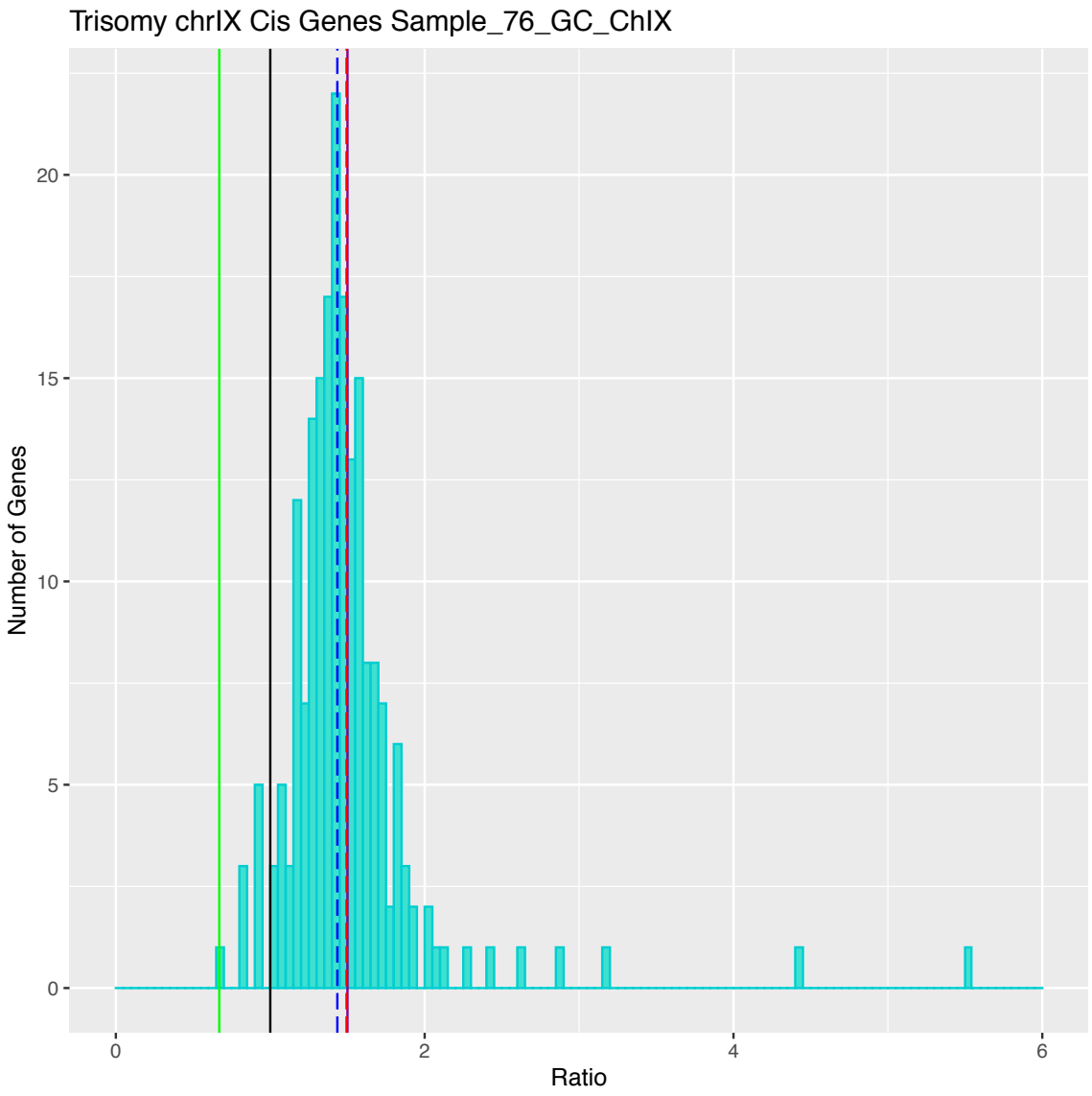

Trisomy chrI Cis Genes Sample_18_GC_ChrI
Number of Genes
0
2
4
6
8
0
2
4
6
Ratio

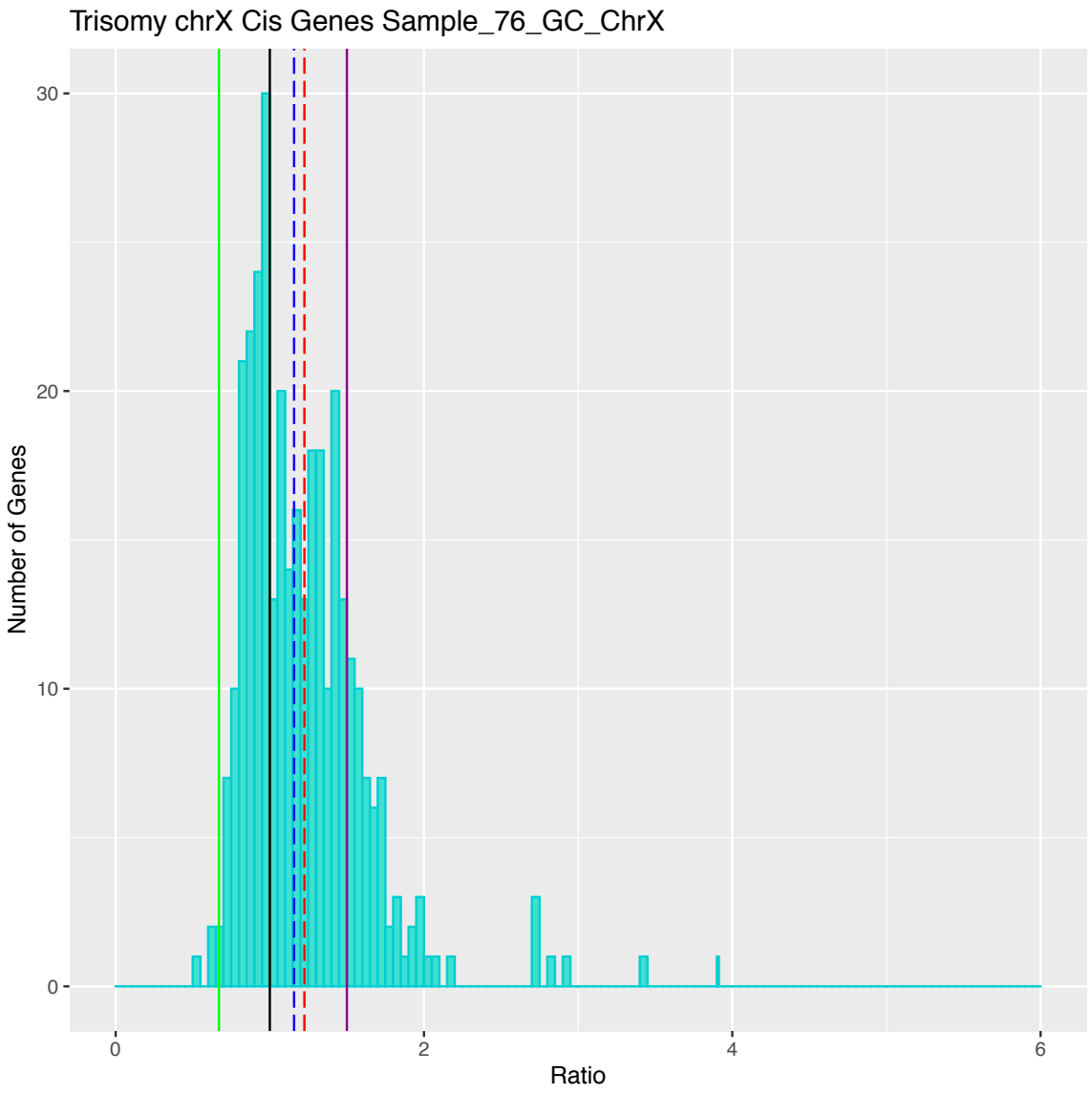

Trisomy chrXII Cis Genes Sample_18_GC_ChrXII
Number of Genes
0
10
20
30
40
0
2
4
6
Ratio

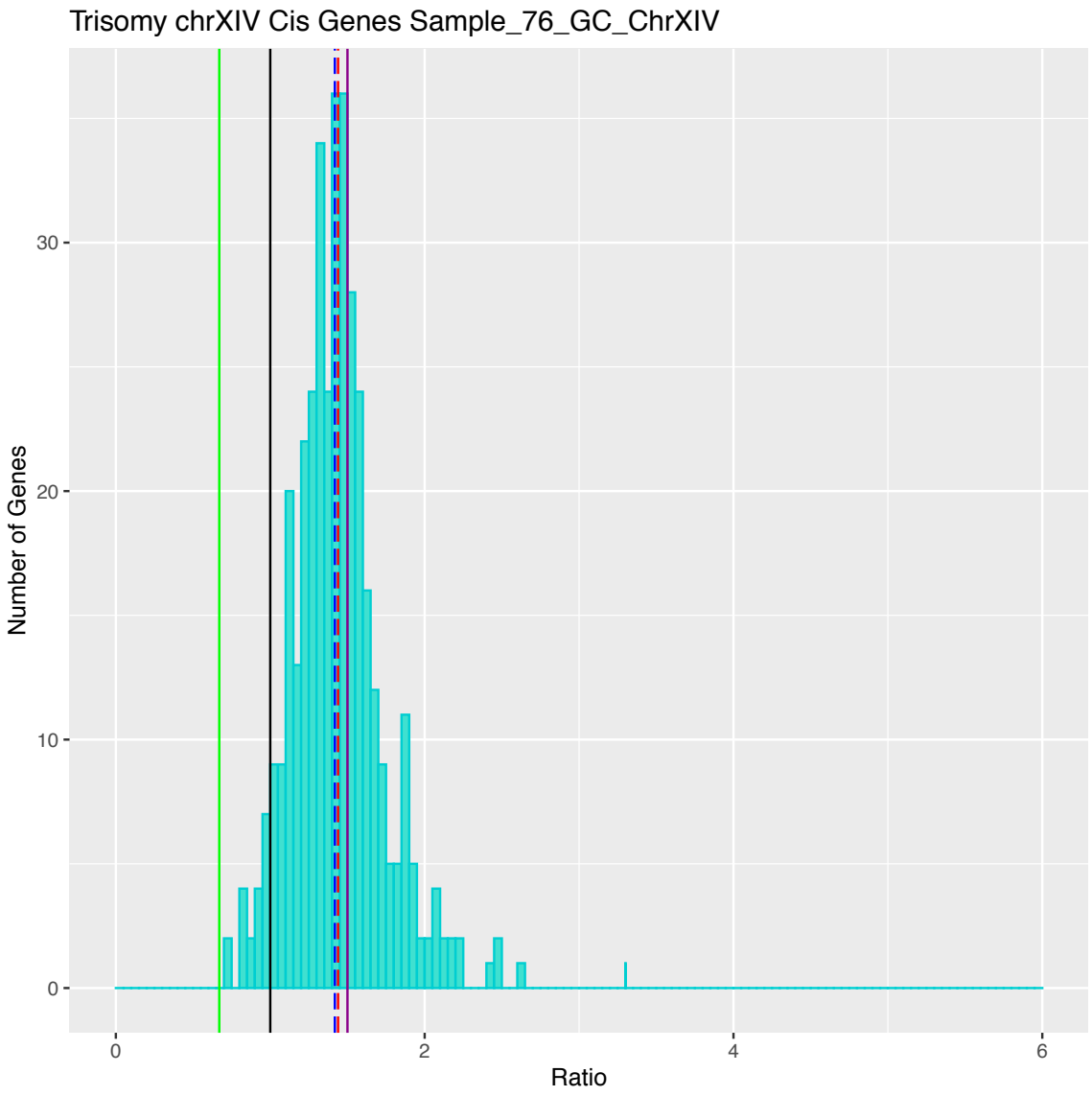

Trisomy chrV Cis Genes Sample_49_GC_ChrV
Number of Genes
20
10
0
0
2
4
6
Ratio

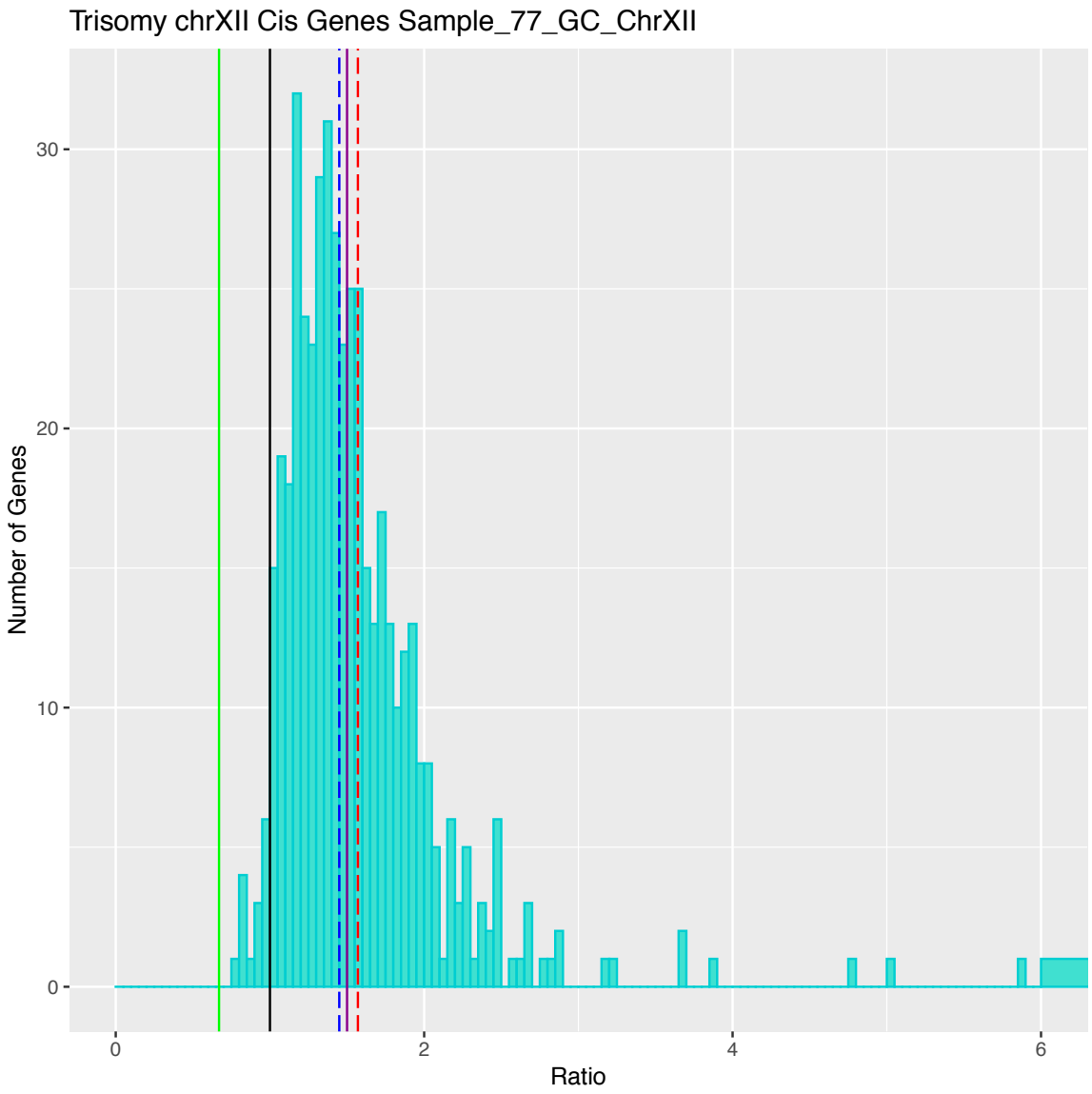

Trisomy chrVII Cis Genes Sample_59_GC_ChrVII
Number of Genes
0
10
20
30
40
0
2
4
6
Ratio

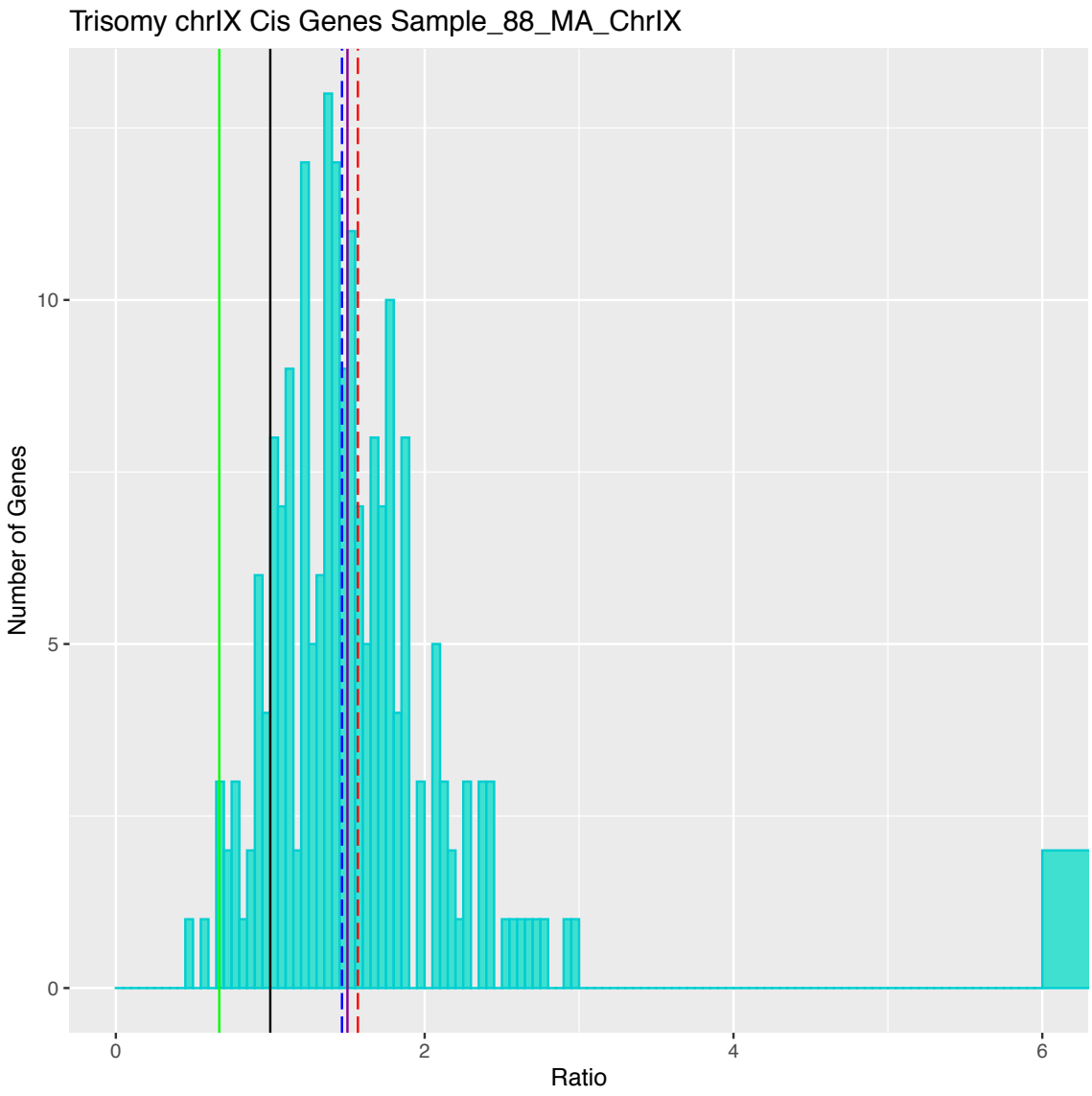

Trisomy chrVII Cis Genes Sample_61_GC_ChrVII
Number of Genes
0
10
20
30
40
0
2
4
6
Ratio

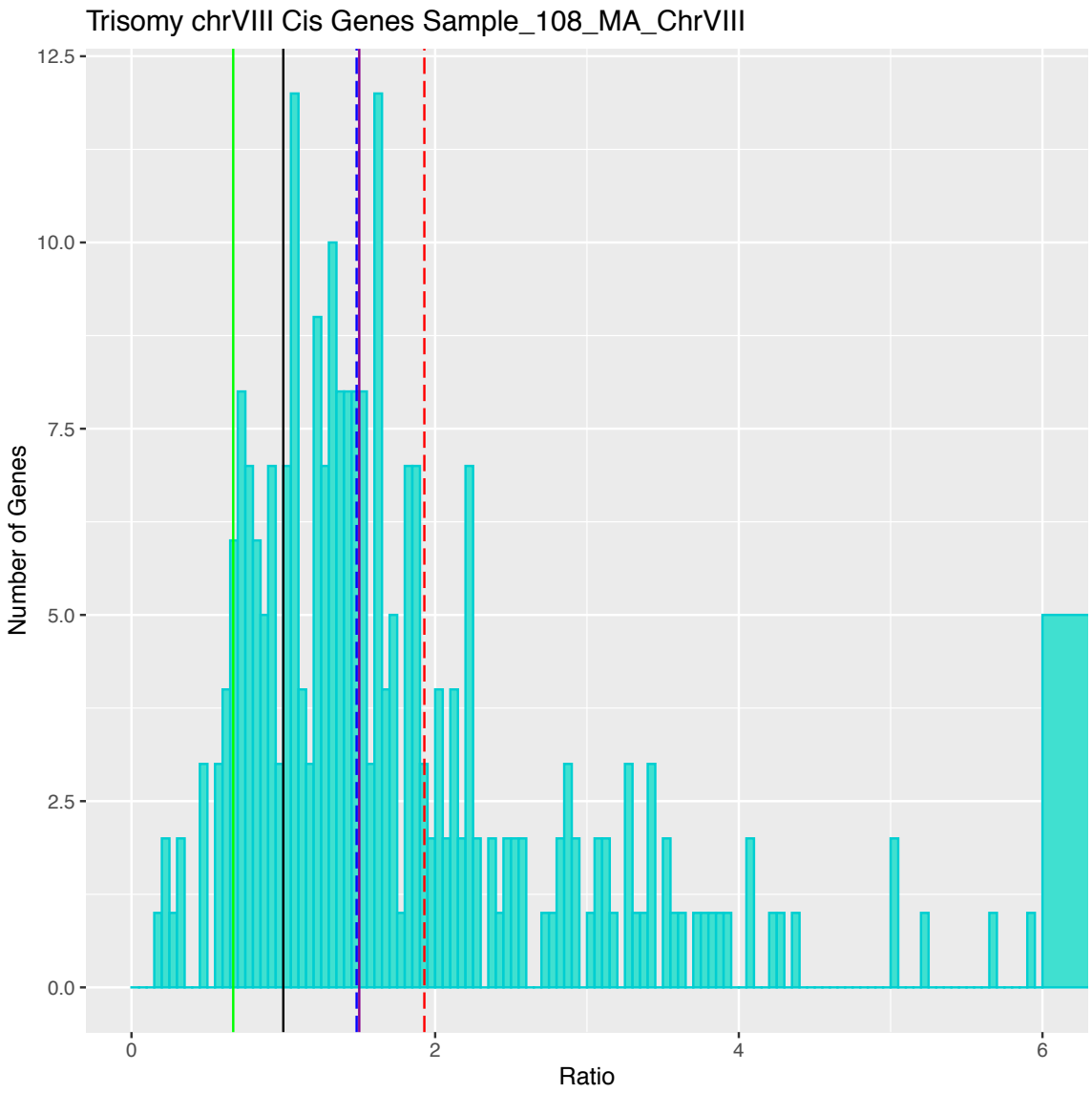

Trisomy chrIX Cis Genes Sample_76_GC_ChIX
Number of Genes
20
15
10
5
0
0
2
4
6
Ratio

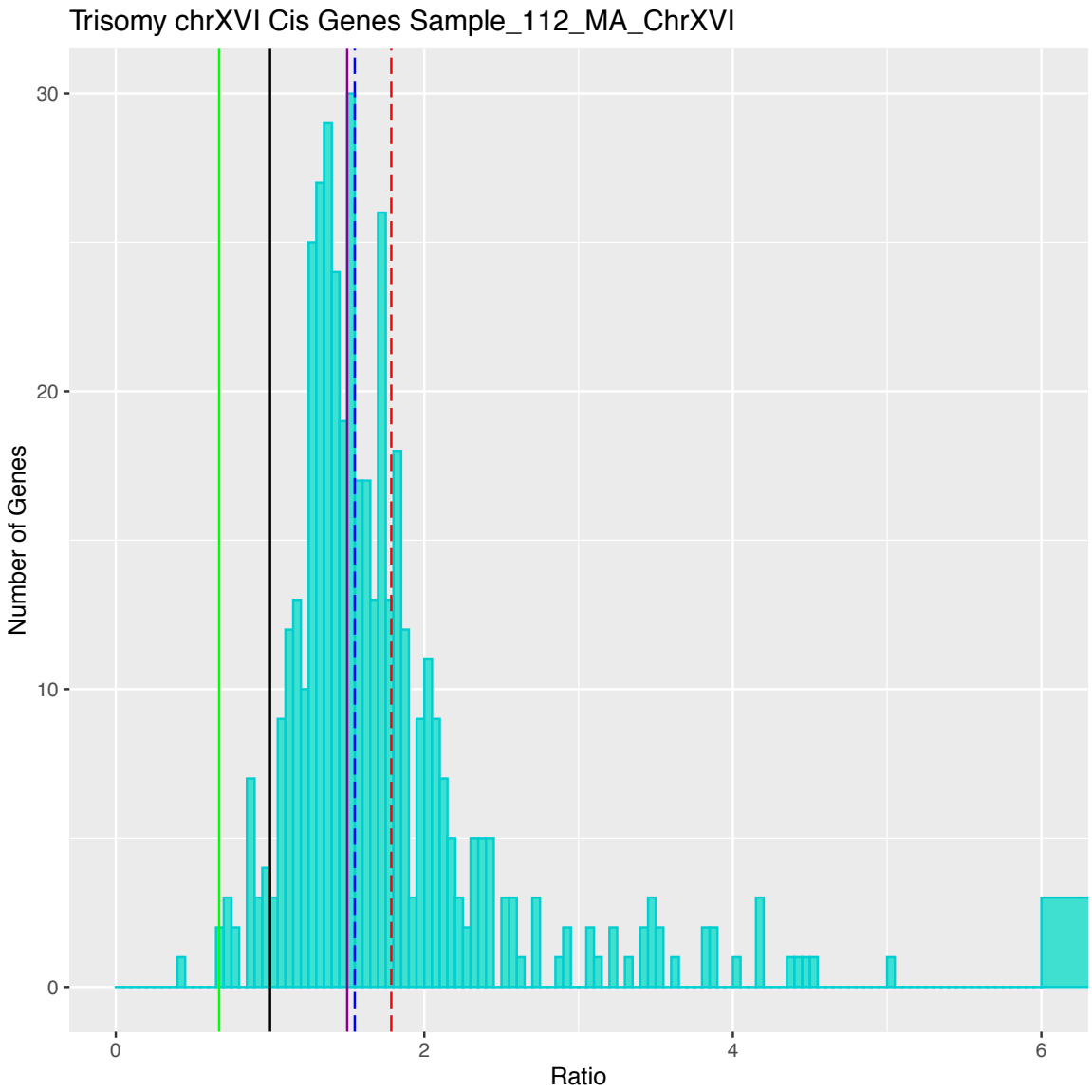

Trisomy chrX Cis Genes Sample_76_GC_ChrX
Number of Genes
30
20
10
0
0
2
4
6
Ratio

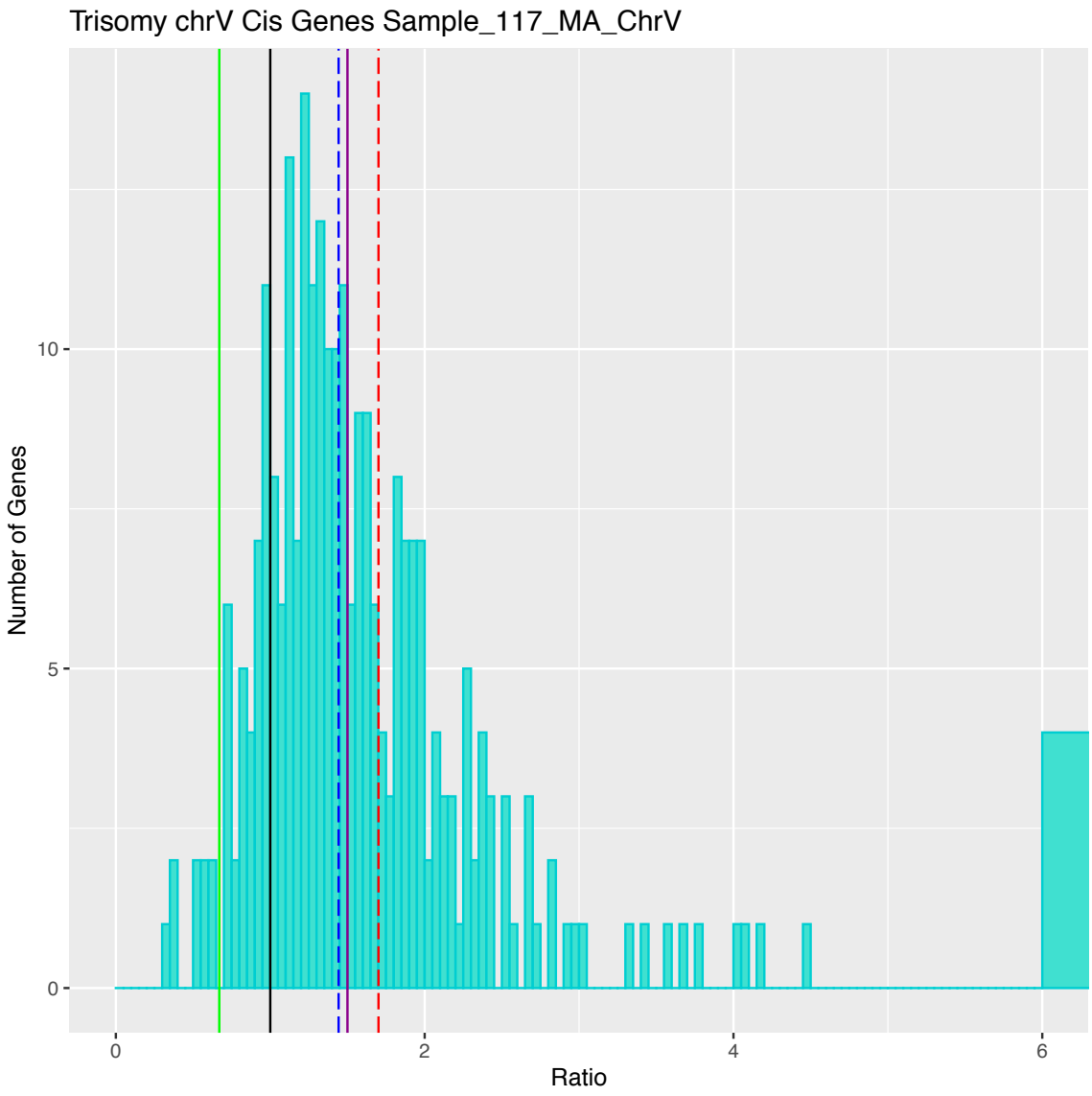

Trisomy chrXIV Cis Genes Sample_76_GC_ChrXIV
Number of Genes
30
20
10
0
0
2
4
6
Ratio

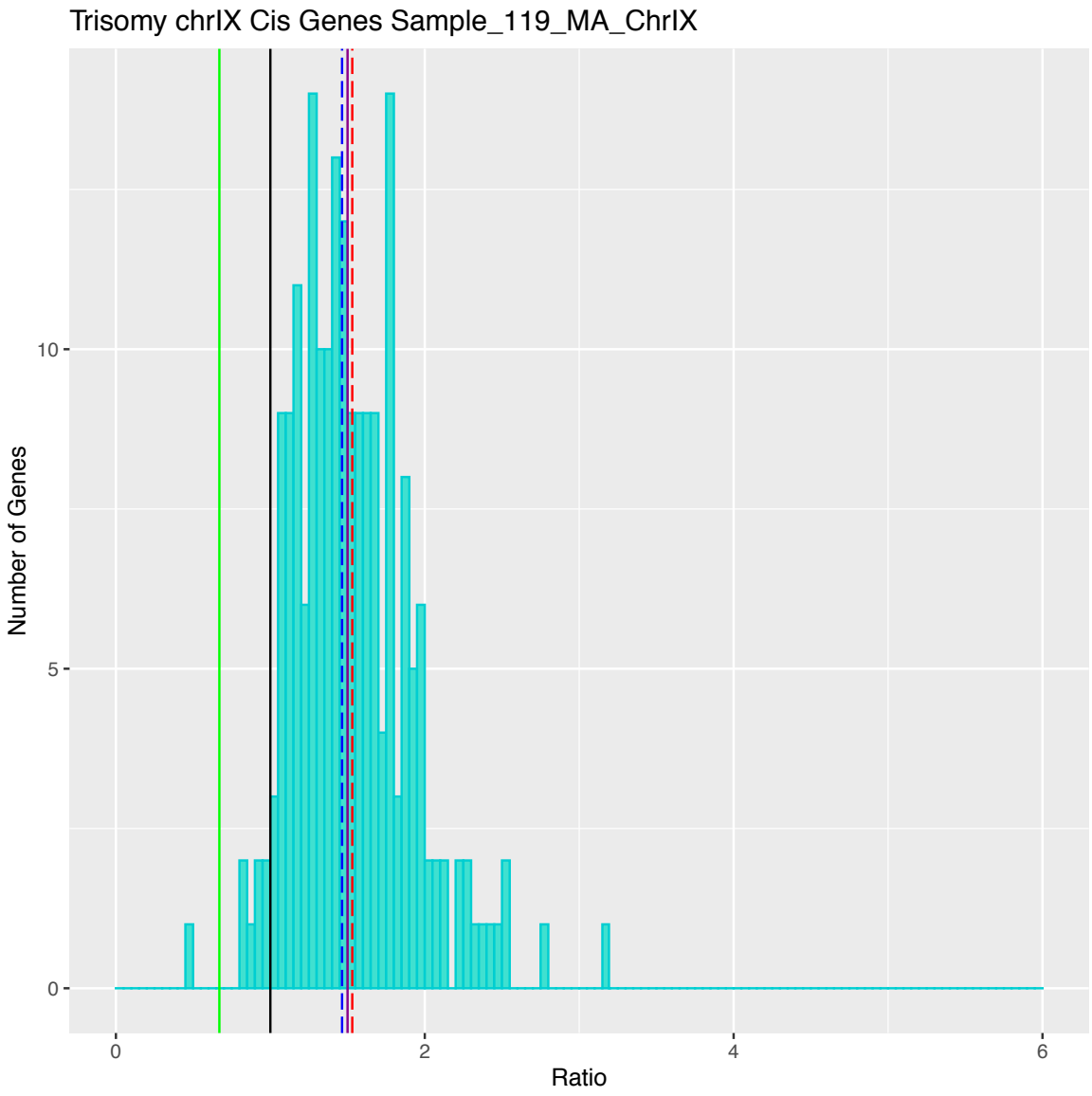

Trisomy chrXII Cis Genes Sample_77_GC_ChrXII
Number of Genes
30
20
10
0
0
2
4
6
Ratio

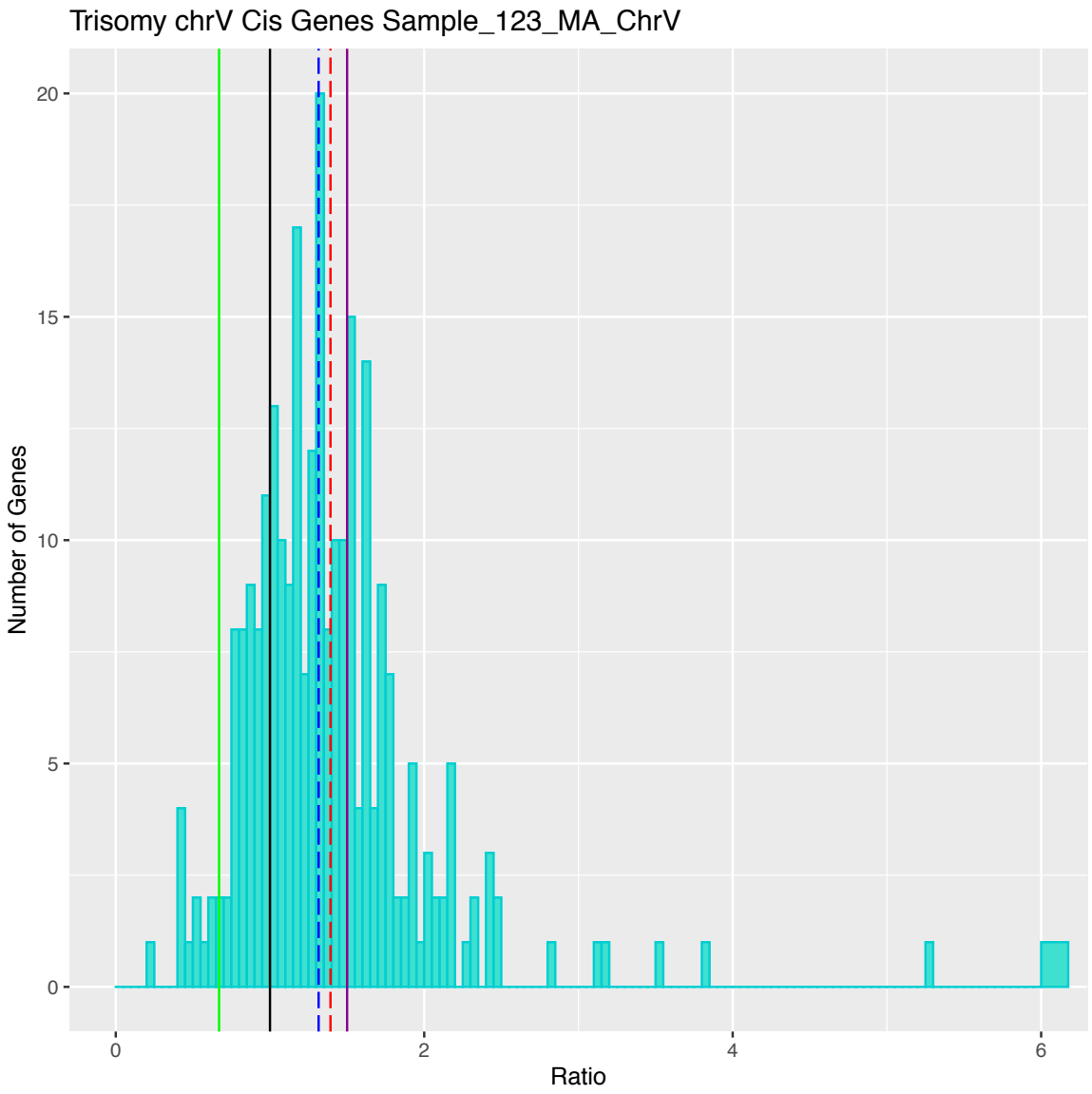

Trisomy chrIX Cis Genes Sample_88_MA_ChrIX
Number of Genes
10
5
0
0
2
4
6
Ratio

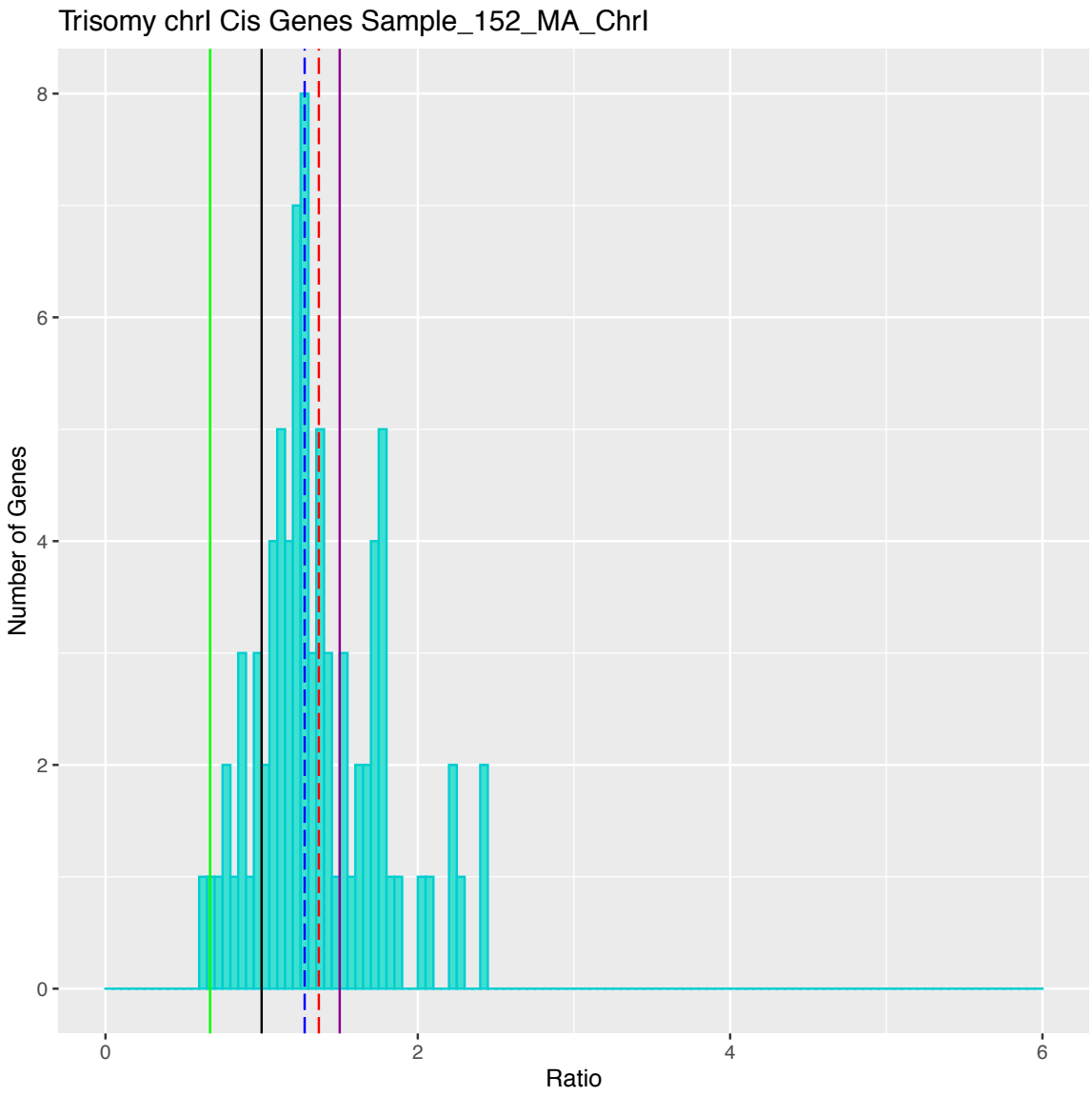

Trisomy chrVIII Cis Genes Sample_108_MA_ChrVIII
Number of Genes
12.5
10.0
7.5
5.0
2.5
0.0
0
2
4
6
Ratio

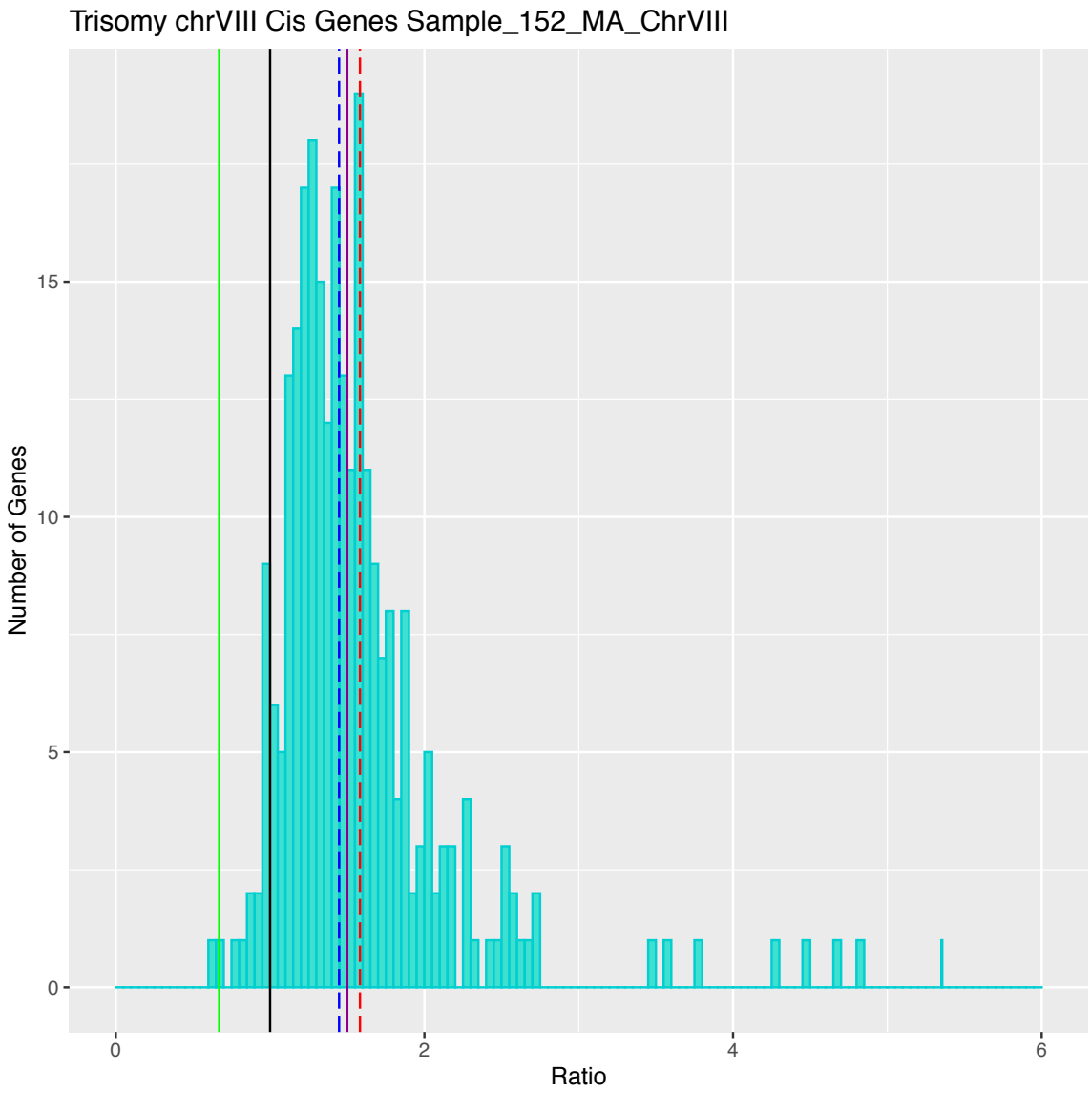

Trisomy chrXVI Cis Genes Sample_112_MA_ChrXVI
Number of Genes
30
20
10
0
0
2
4
6
Ratio


Trisomy chrV Cis Genes Sample_117_MA_ChrV
Number of Genes
10
5
0
0
2
4
6
Ratio

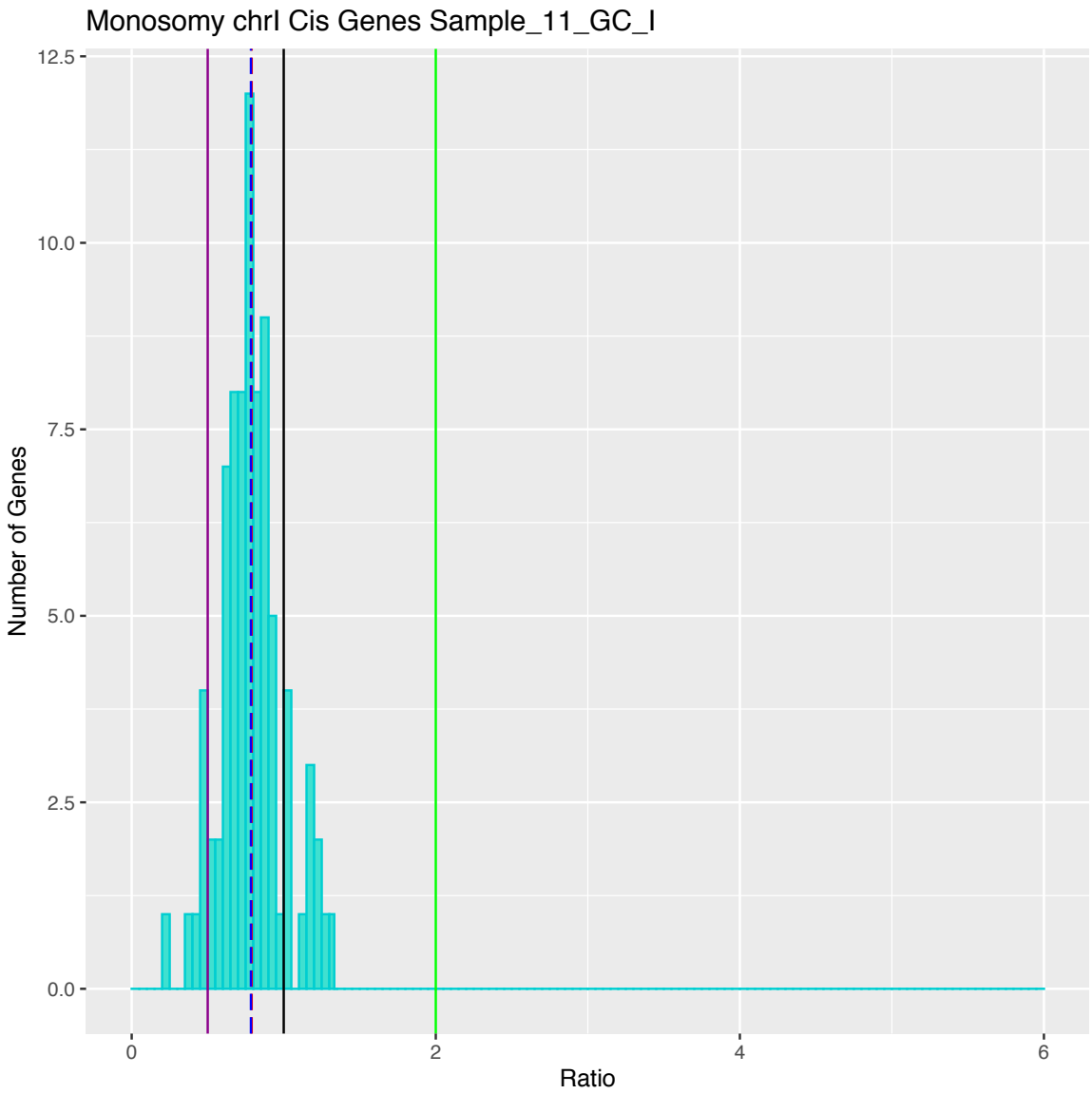

Trisomy chrIX Cis Genes Sample_119_MA_ChrIX
Number of Genes
10
5
0
0
2
4
6
Ratio

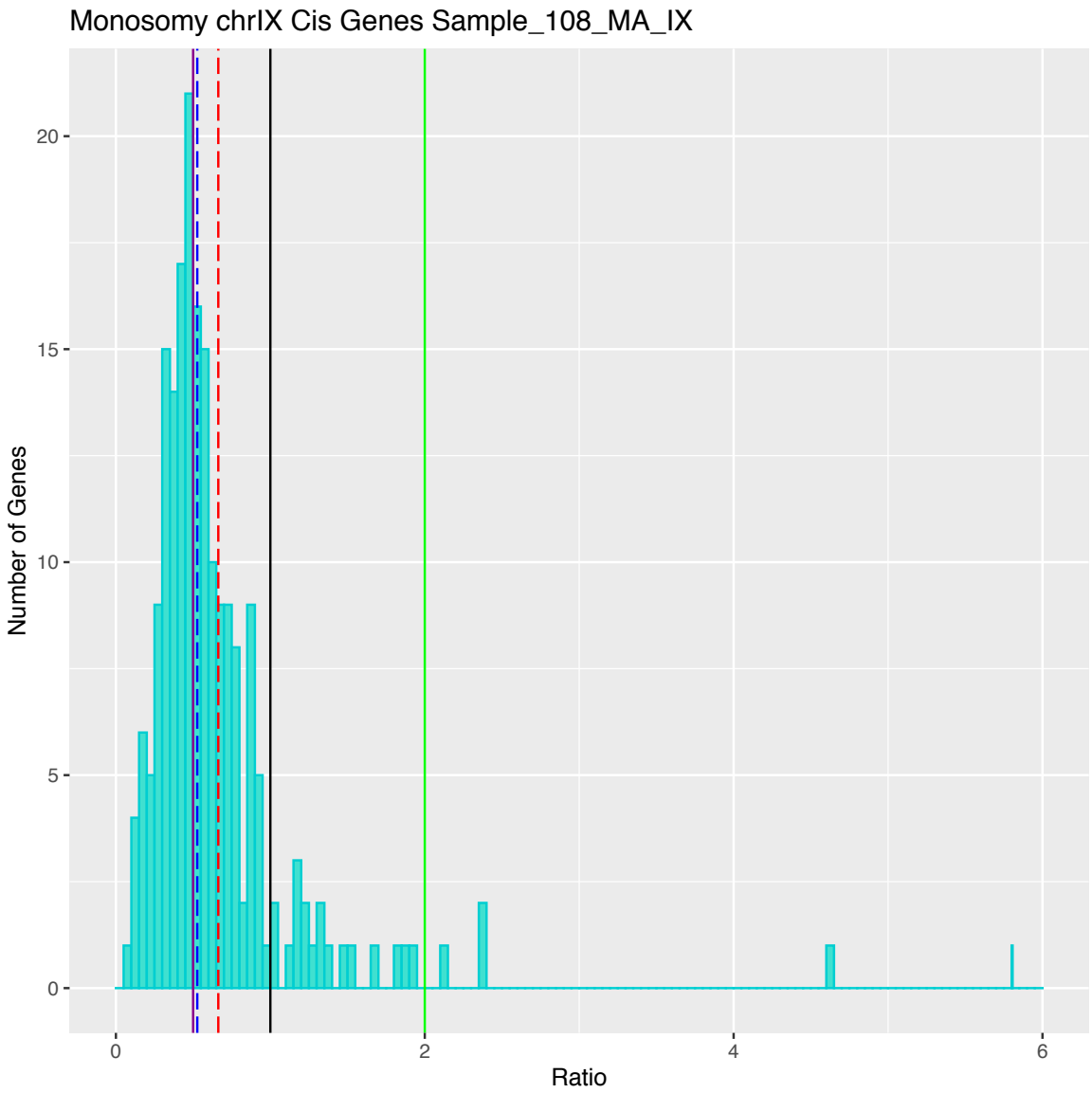

Trisomy chrV Cis Genes Sample_123_MA_ChrV
Number of Genes
20
15
10
5
0
0
2
4
6
Ratio

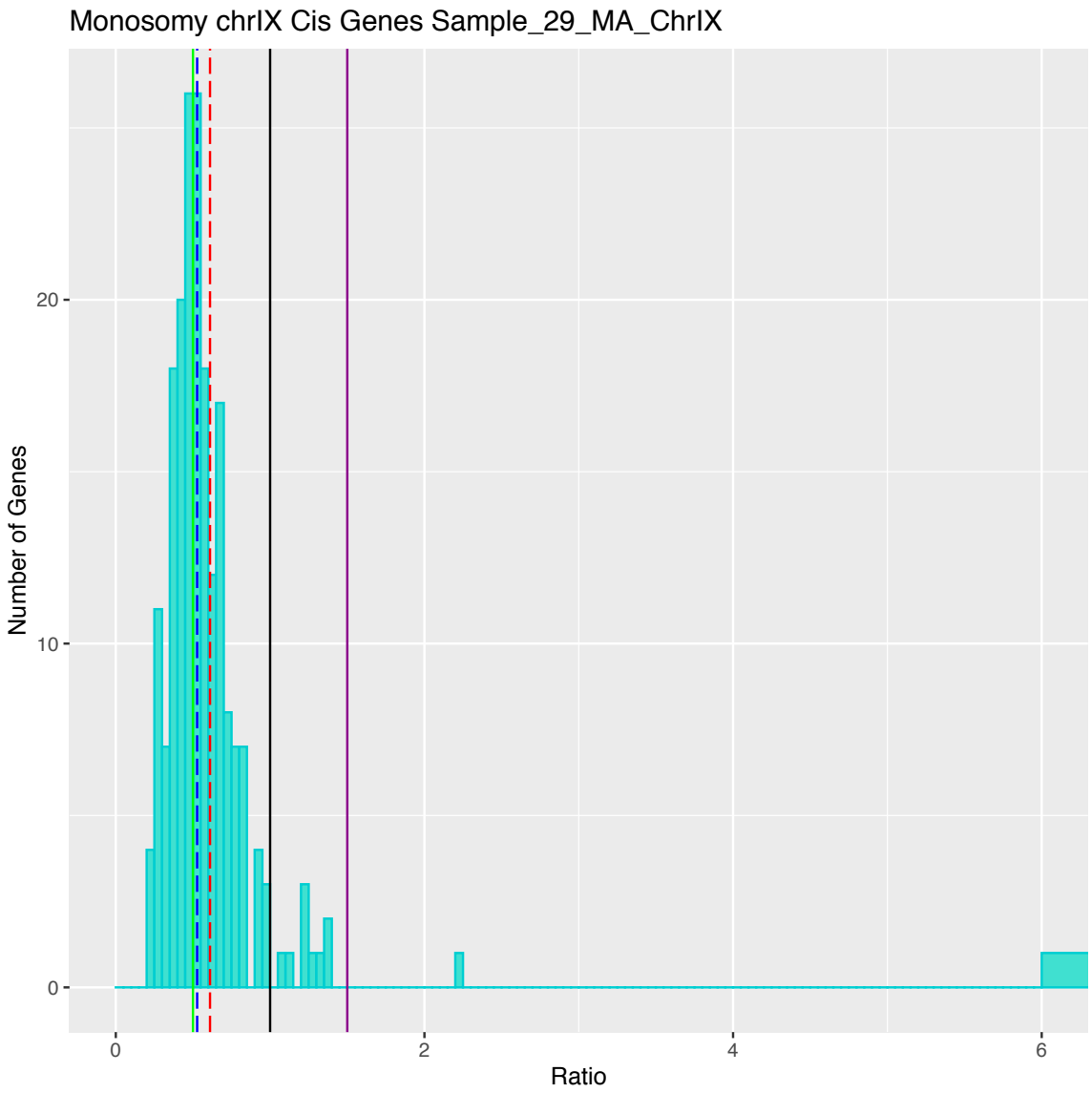

Trisomy chrI Cis Genes Sample_152_MA_ChrI
Number of Genes
0
2
4
6
8
0
2
4
6
Ratio



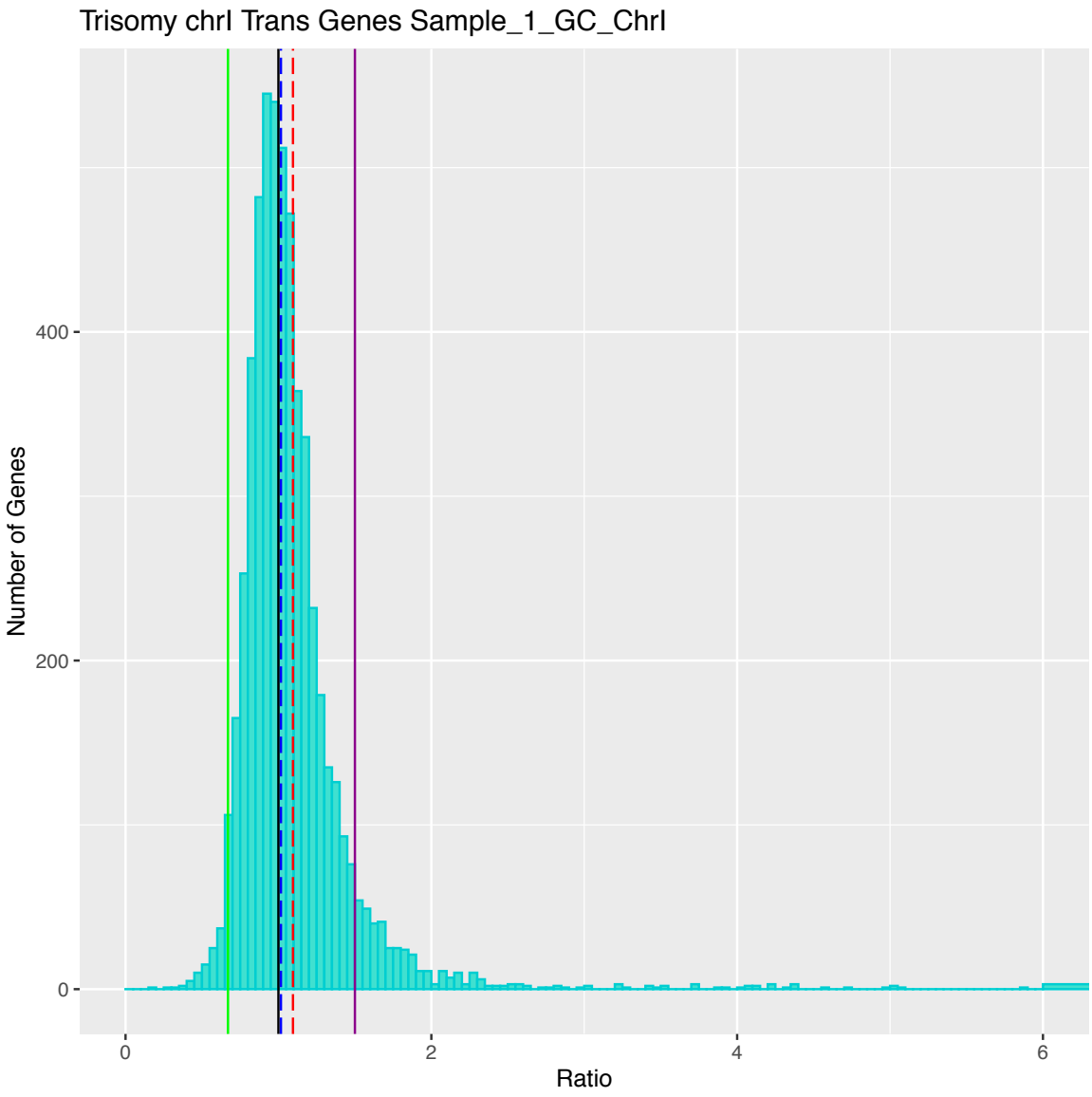
Supplemental Figure 2.3: Histograms of gene expression ratio for cis genes for aneuploid lines. Blue dashed line is median gene expression, red dashed line is mean gene expression; black solid line is expected if there is no difference between the sample and the ancestor, purple/magenta solid line is expected for trisomy, and green solid line is expected for monosomy.

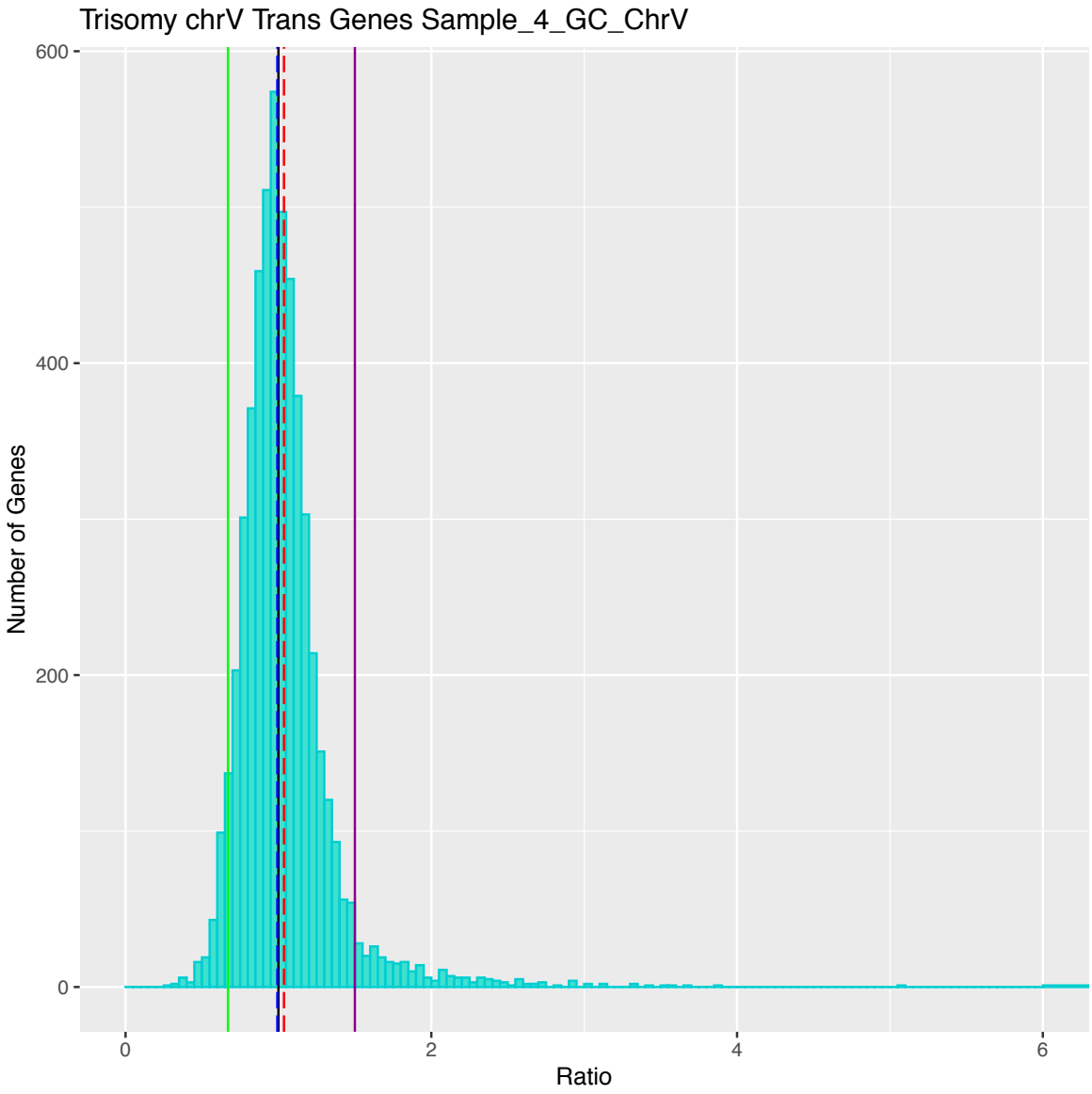

Monosomy chrI Cis Genes Sample_11_GC_I
Number of Genes
12.5
10.0
7.5
5.0
2.5
0.0
0
2
4
6
Ratio

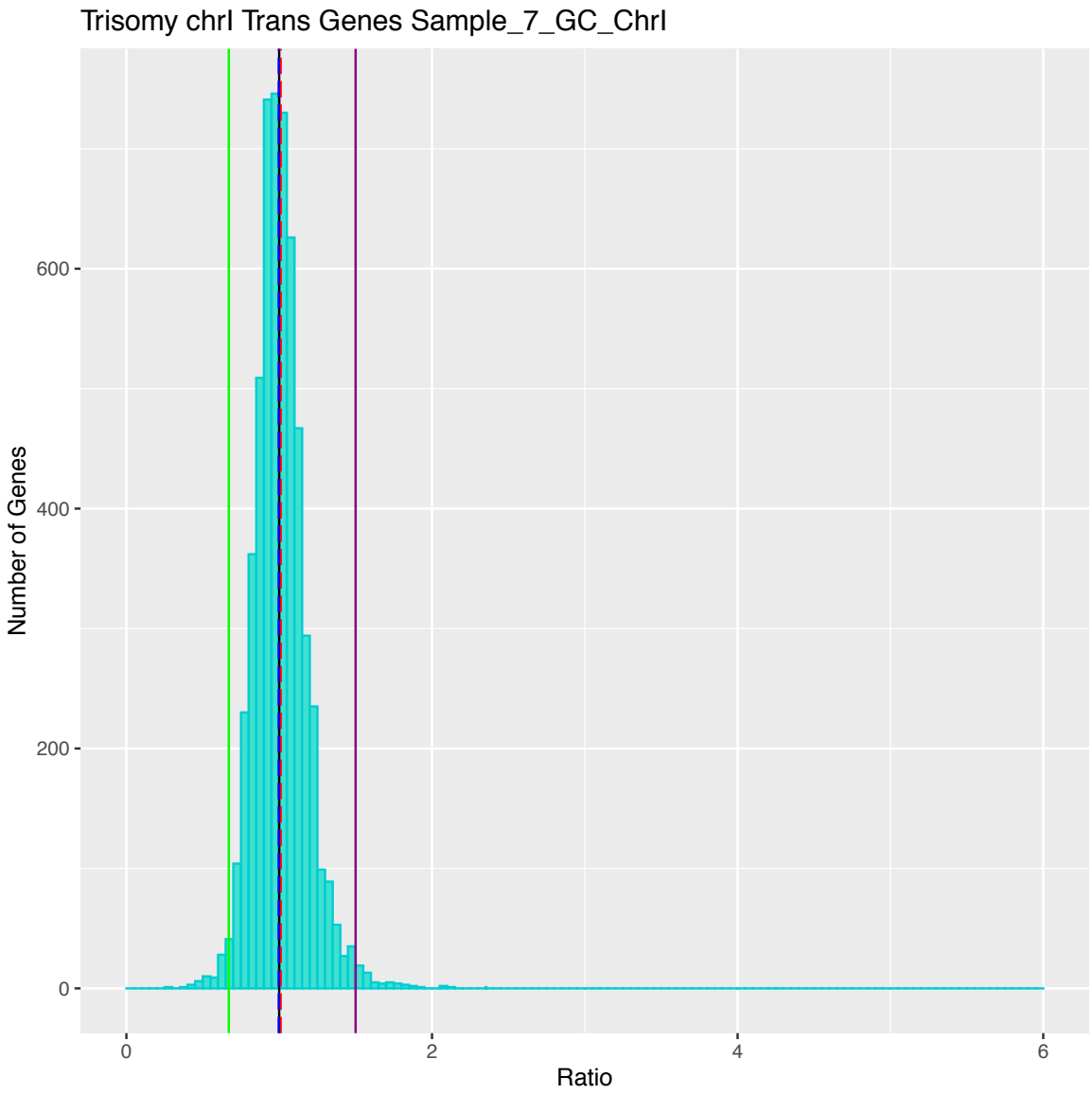

Monosomy chrIX Cis Genes Sample_108_MA_IX
Number of Genes
0
5
10
15
20
0
2
4
6
Ratio

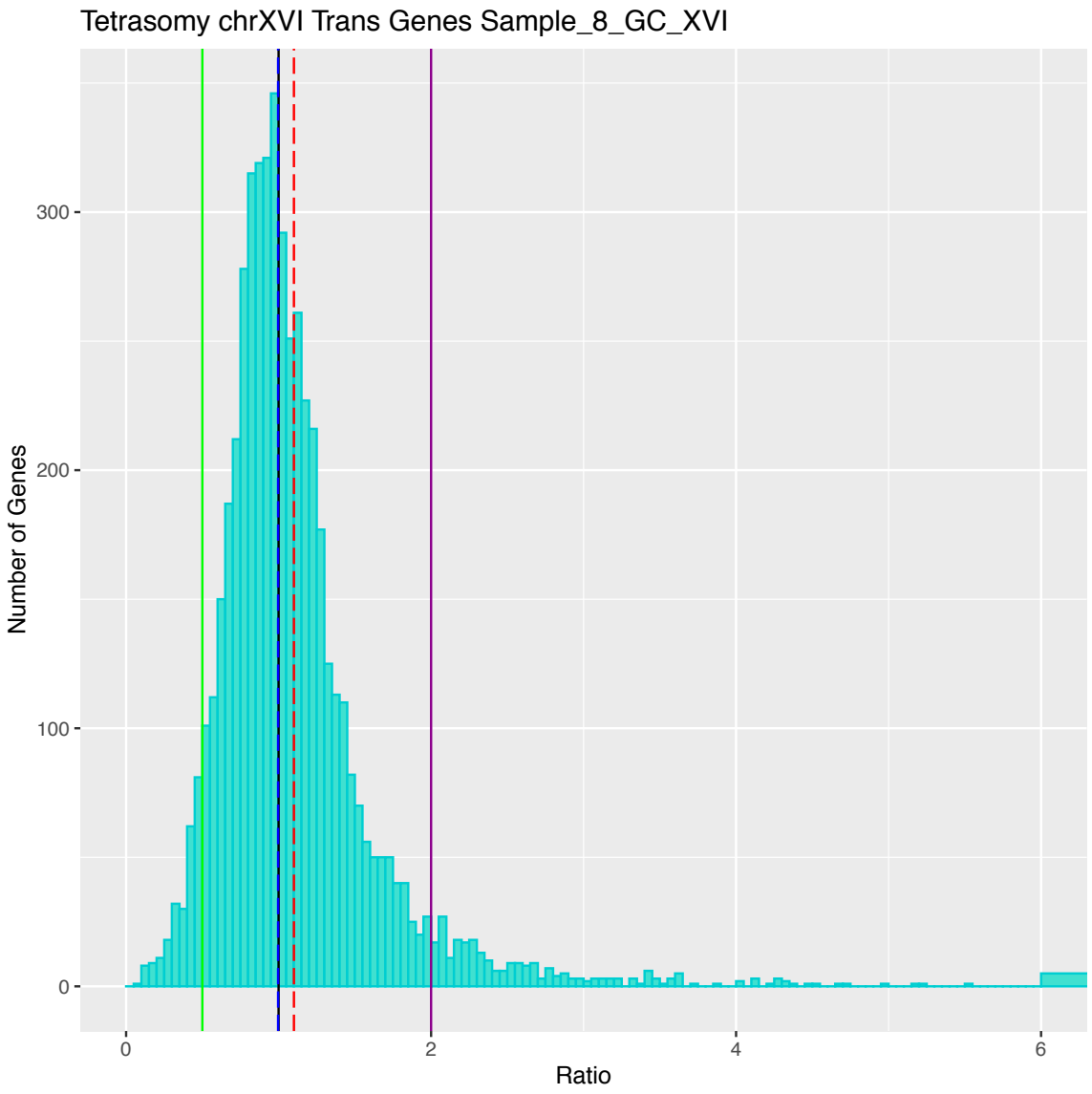


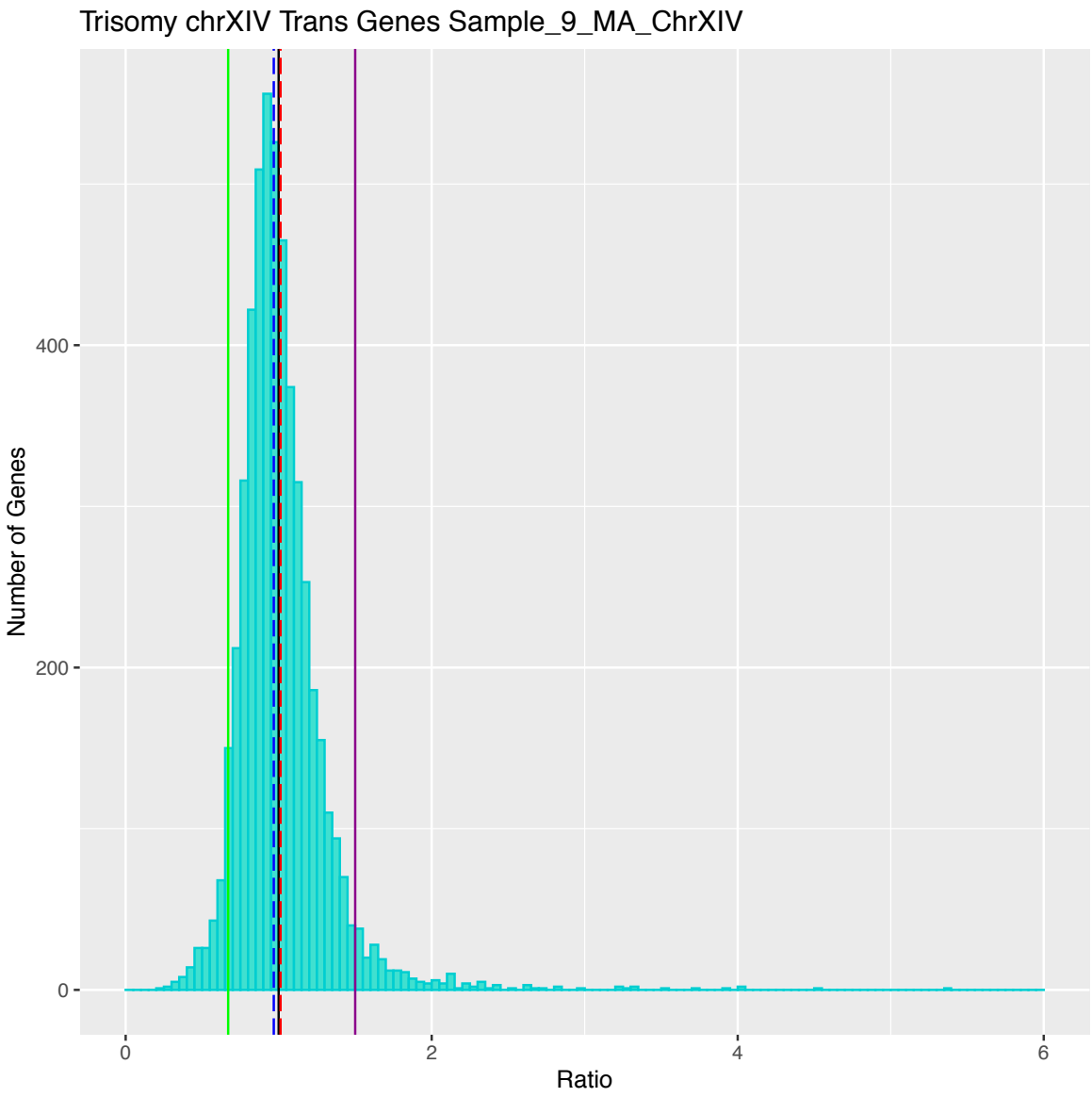
Supplemental Figure 2.3: Histograms of gene expression ratio for cis genes for monosomic aneuploid lines. Blue dashed line is median gene expression, red dashed line is mean gene expression; black solid line is expected if there is no difference between the sample and the ancestor, purple/magenta solid line is expected for monosomy, and green solid line is expected for trisomy.

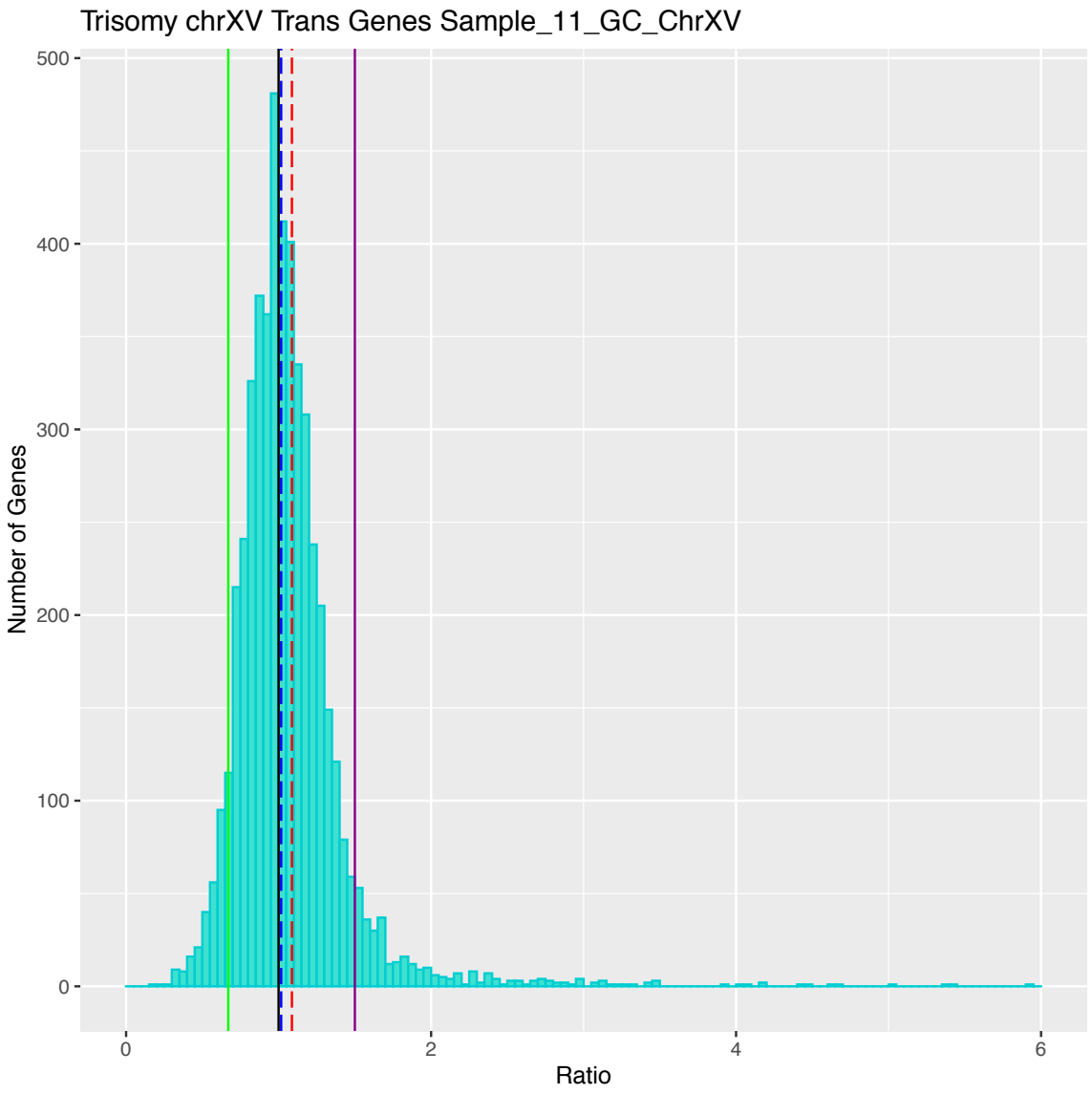

Trisomy chrI Trans Genes Sample_1_GC_ChrI
Number of Genes
400
200
0
0
2
4
6
Ratio

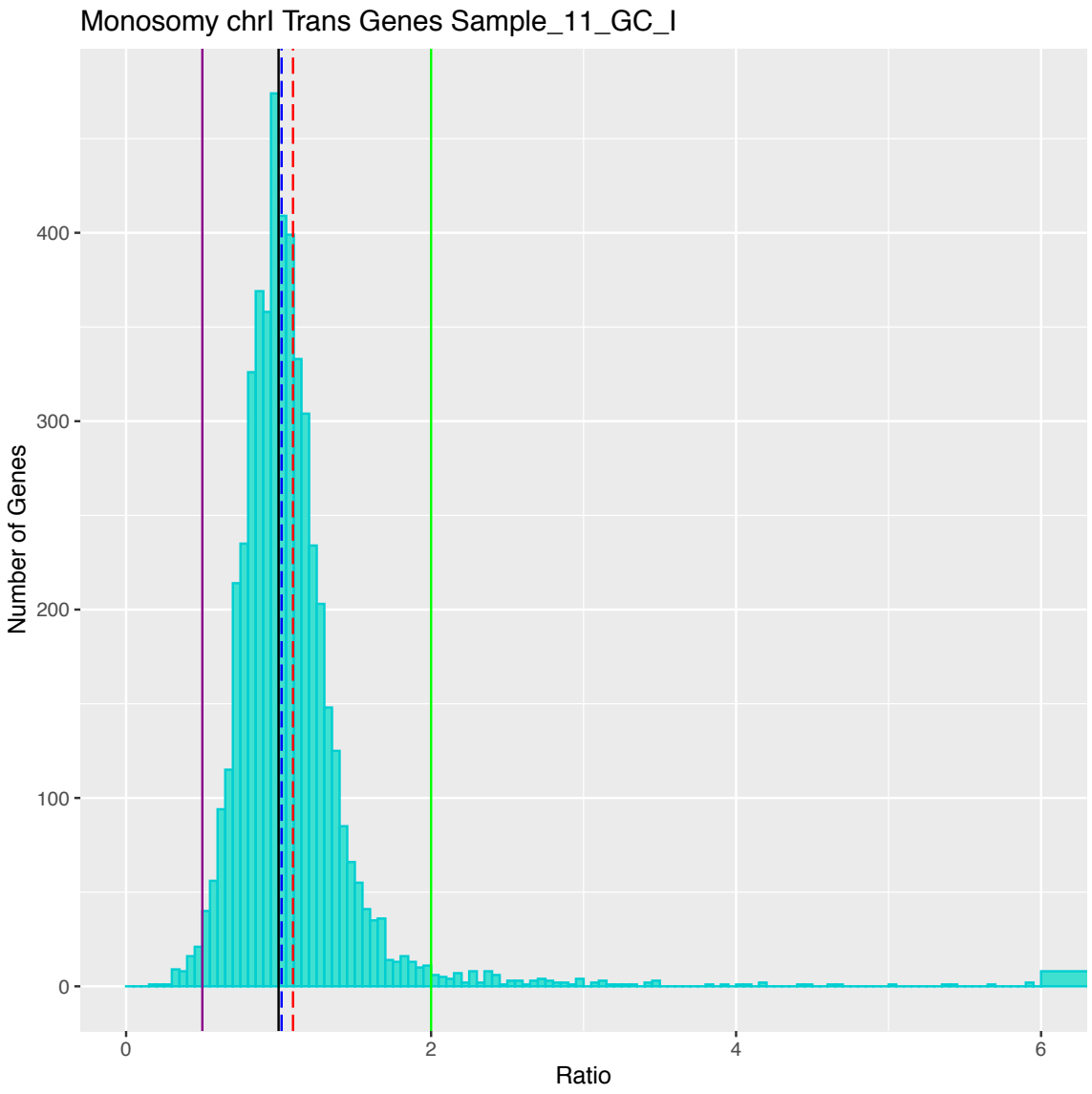

Trisomy chrV Trans Genes Sample_4_GC_ChrV
600
400
200
0
Number of Genes
0
2
4
6
Ratio

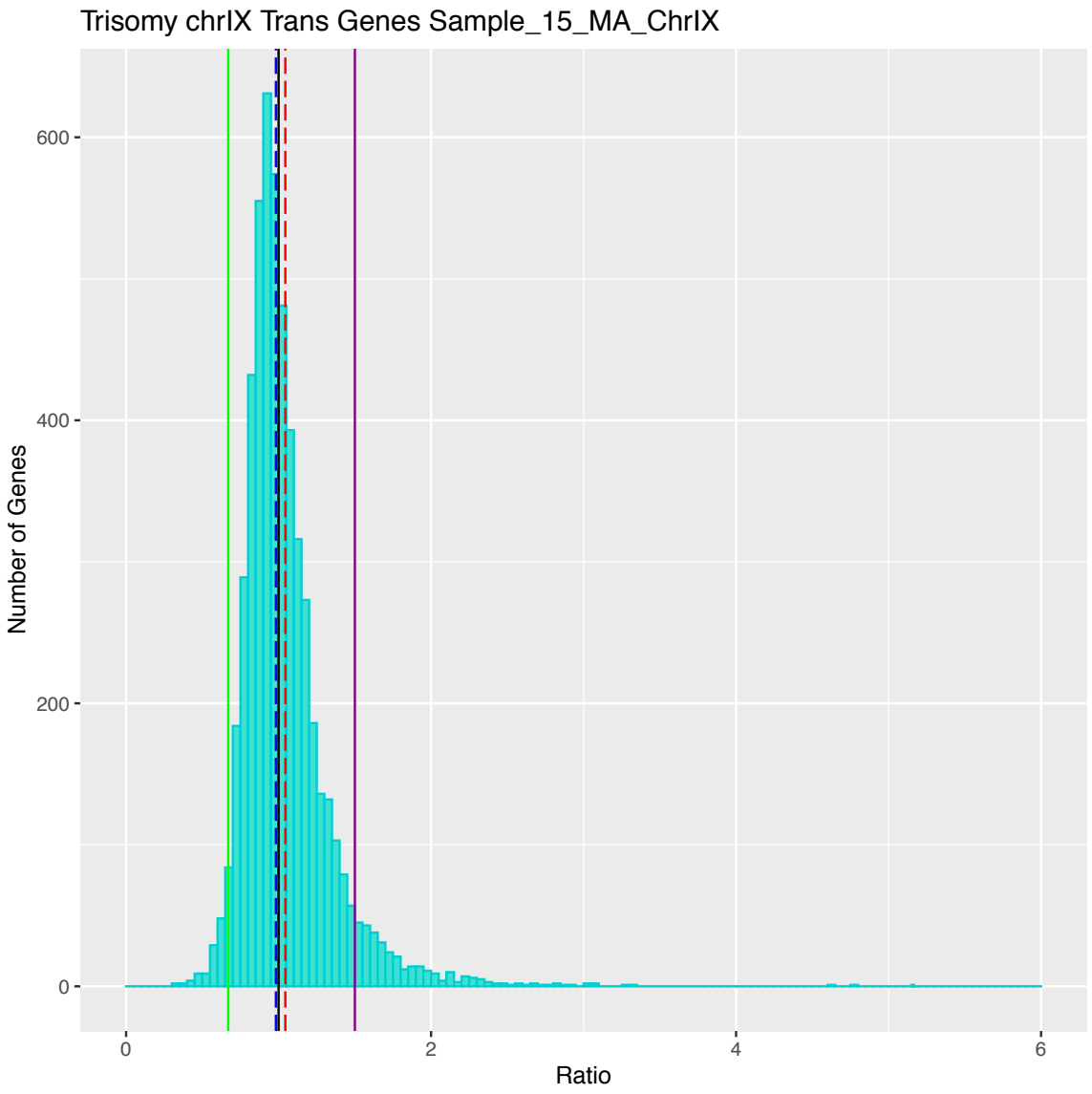

Trisomy chrI Trans Genes Sample_7_GC_ChrI
Number of Genes
600
400
200
0
0
2
4
6
Ratio

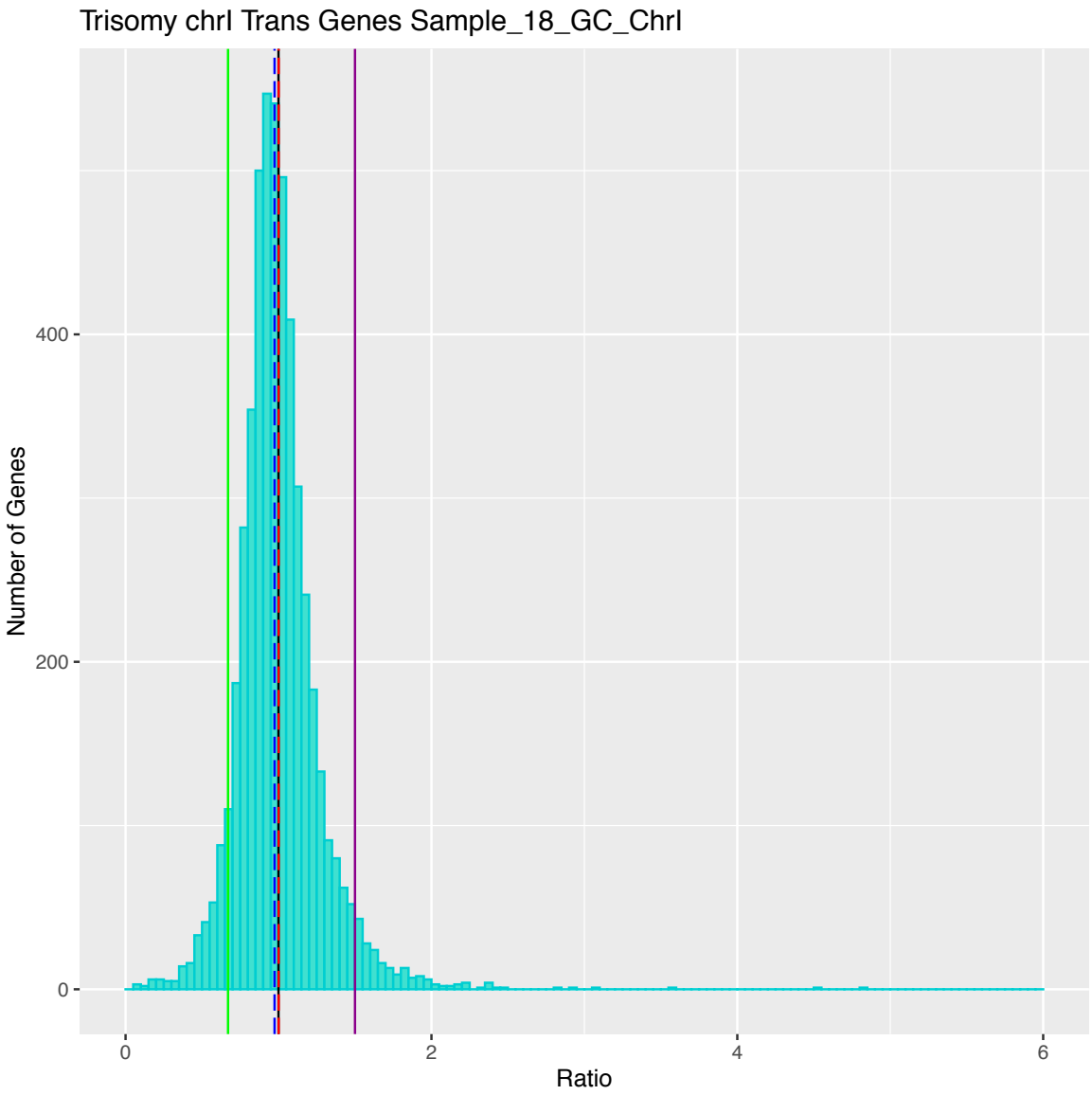

Tetrasomy chrXVI Trans Genes Sample_8_GC_XVI
Number of Genes
300
200
100
0
0
2
4
6
Ratio

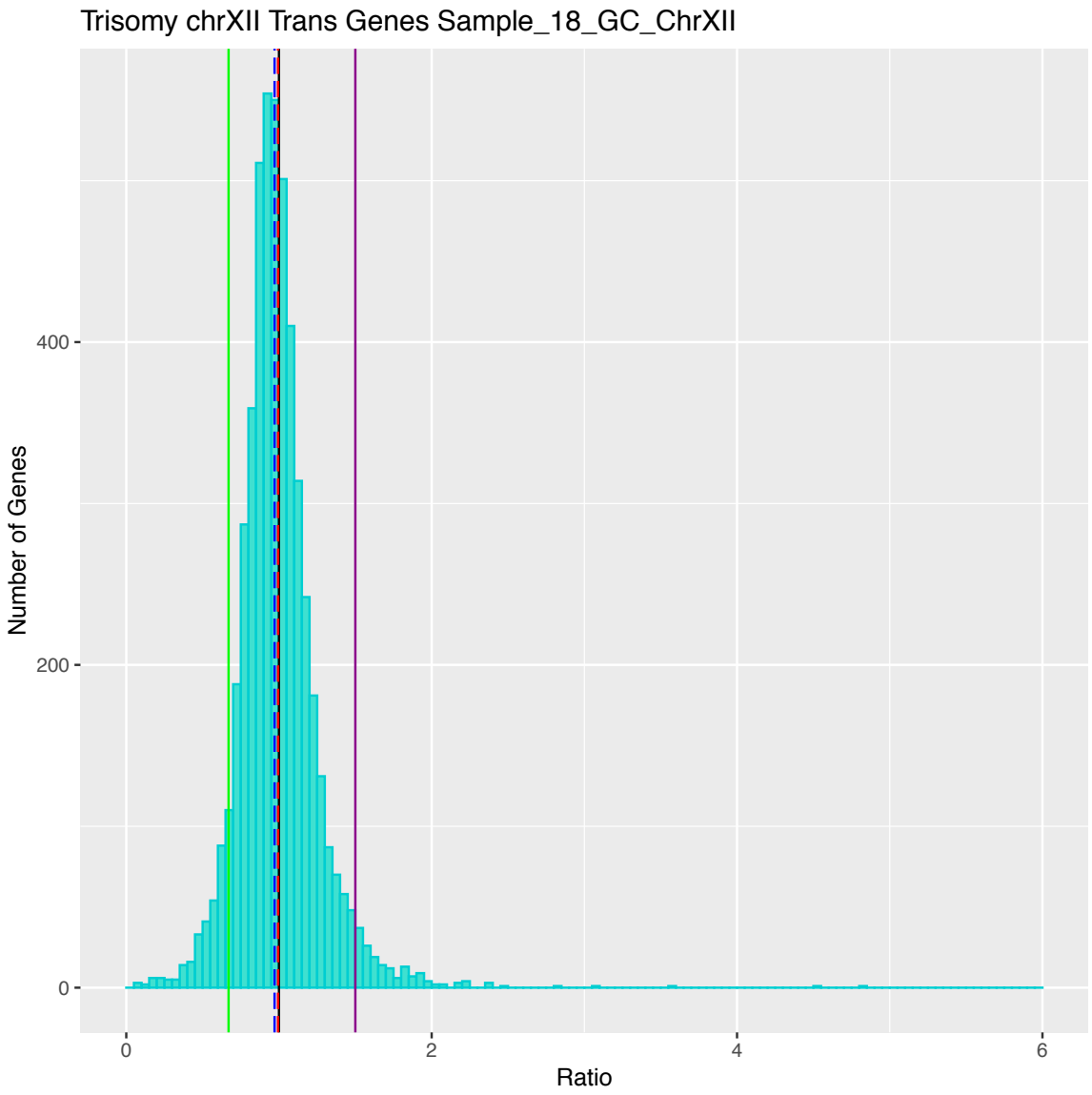

Trisomy chrXIV Trans Genes Sample_9_MA_ChrXIV
Number of Genes
400
200
0
0
2
4
6
Ratio

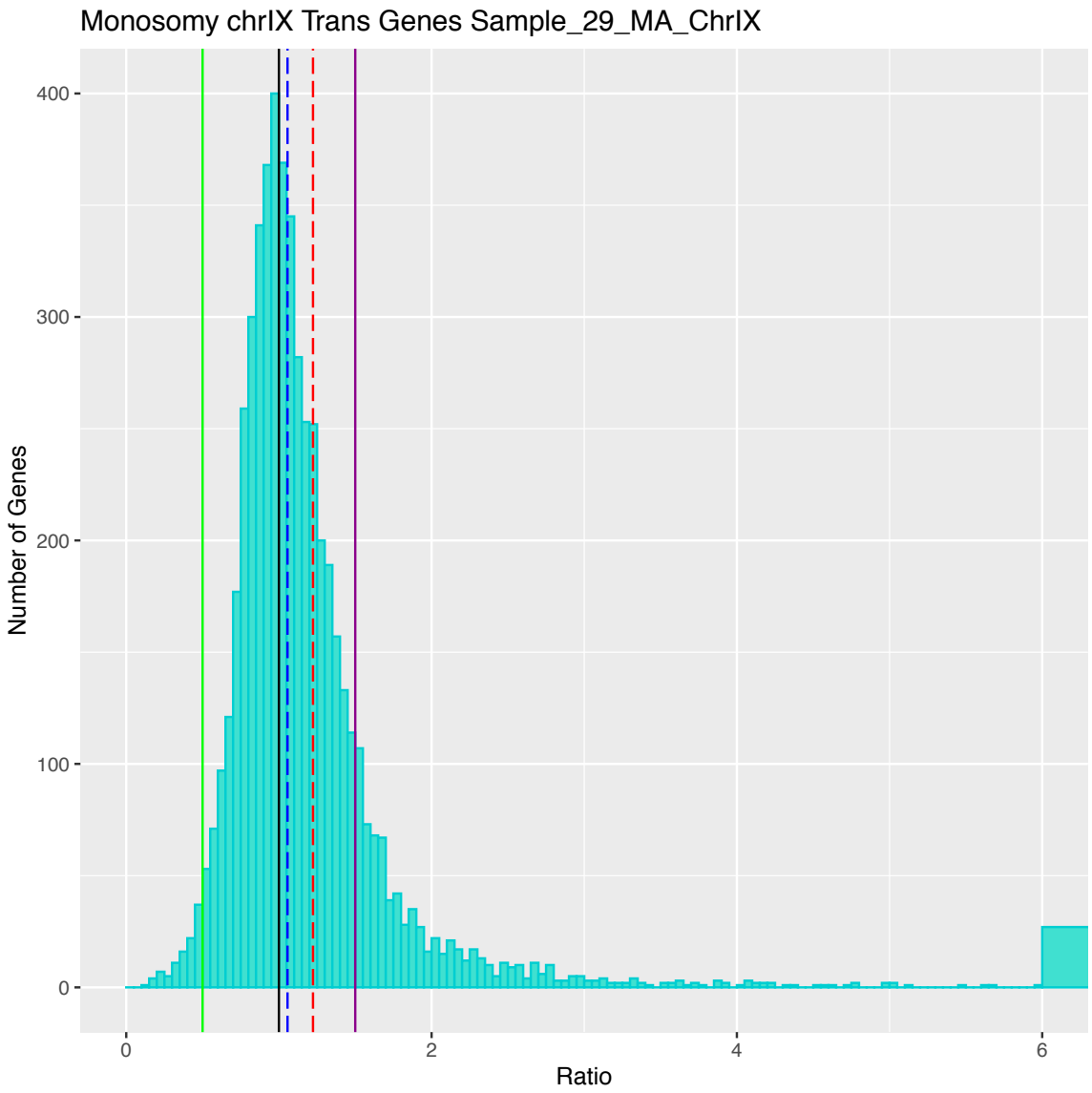

Trisomy chrXV Trans Genes Sample_11_GC_ChrXV
Number of Genes
500
400
300
200
100
0
0
2
4
6
Ratio

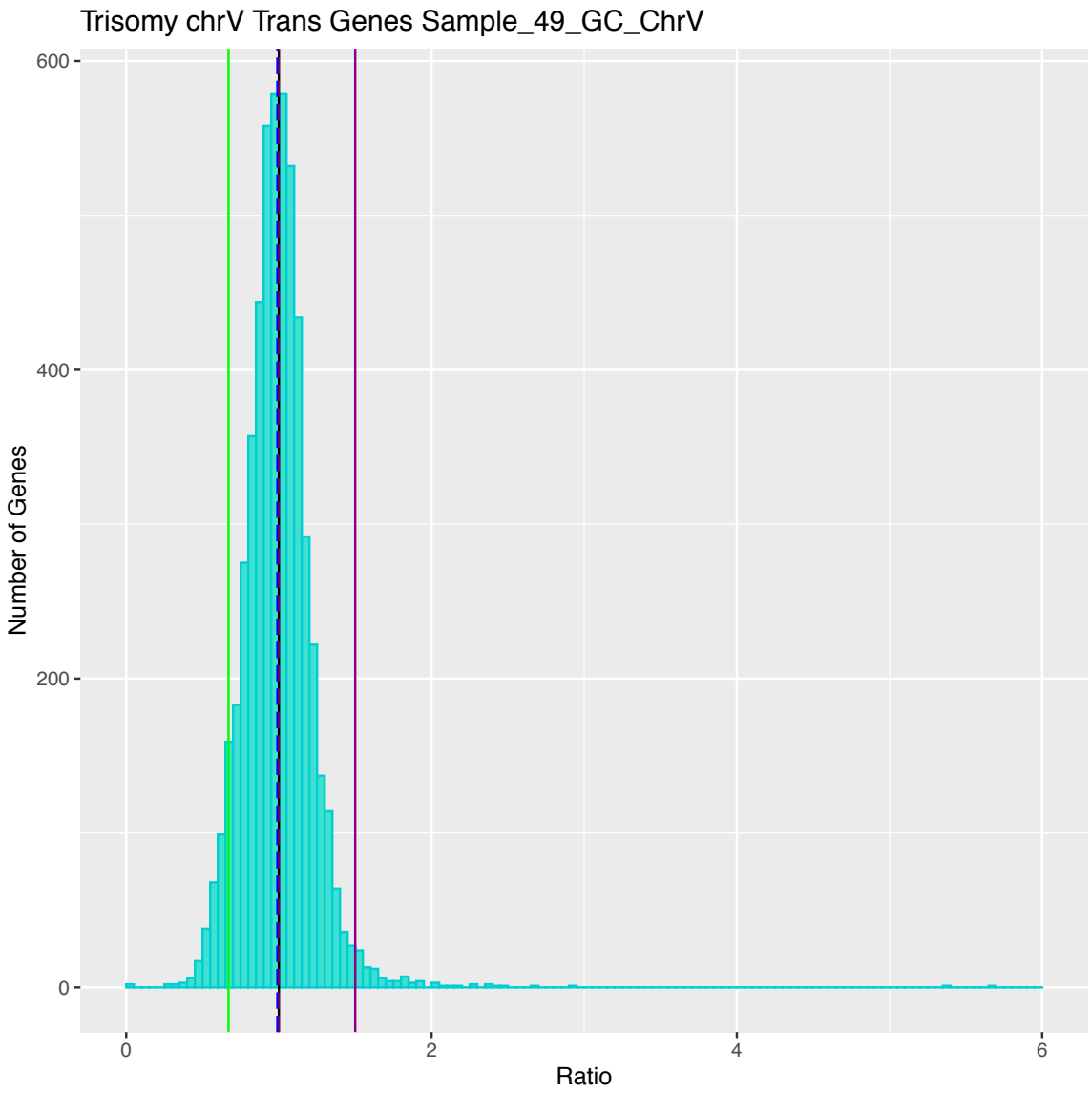

Monosomy chrI Trans Genes Sample_11_GC_I
Number of Genes
400
300
200
100
0
0
2
4
6
Ratio

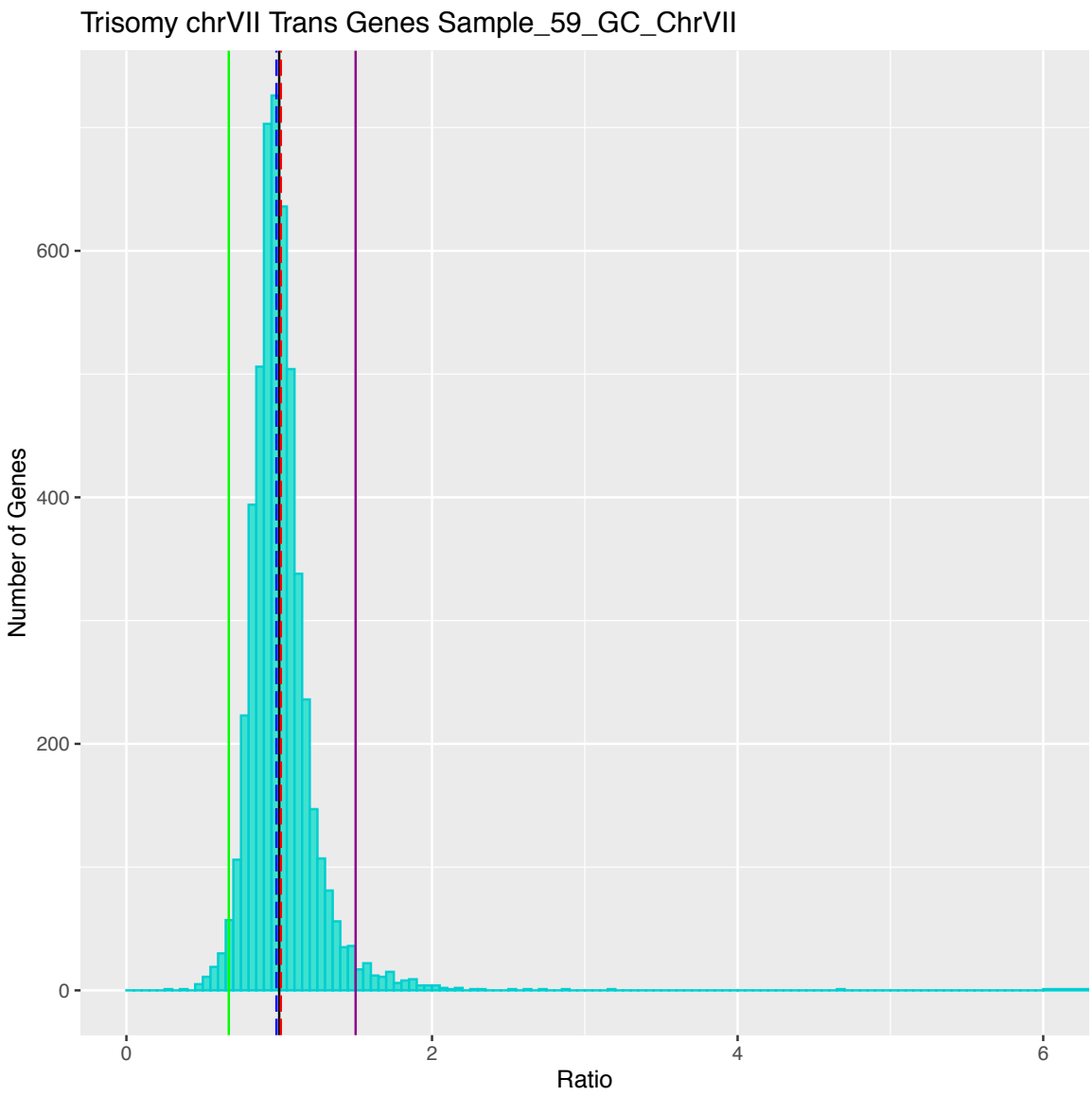

Trisomy chrIX Trans Genes Sample_15_MA_ChrIX
Number of Genes
600
400
200
0
0
2
4
6
Ratio

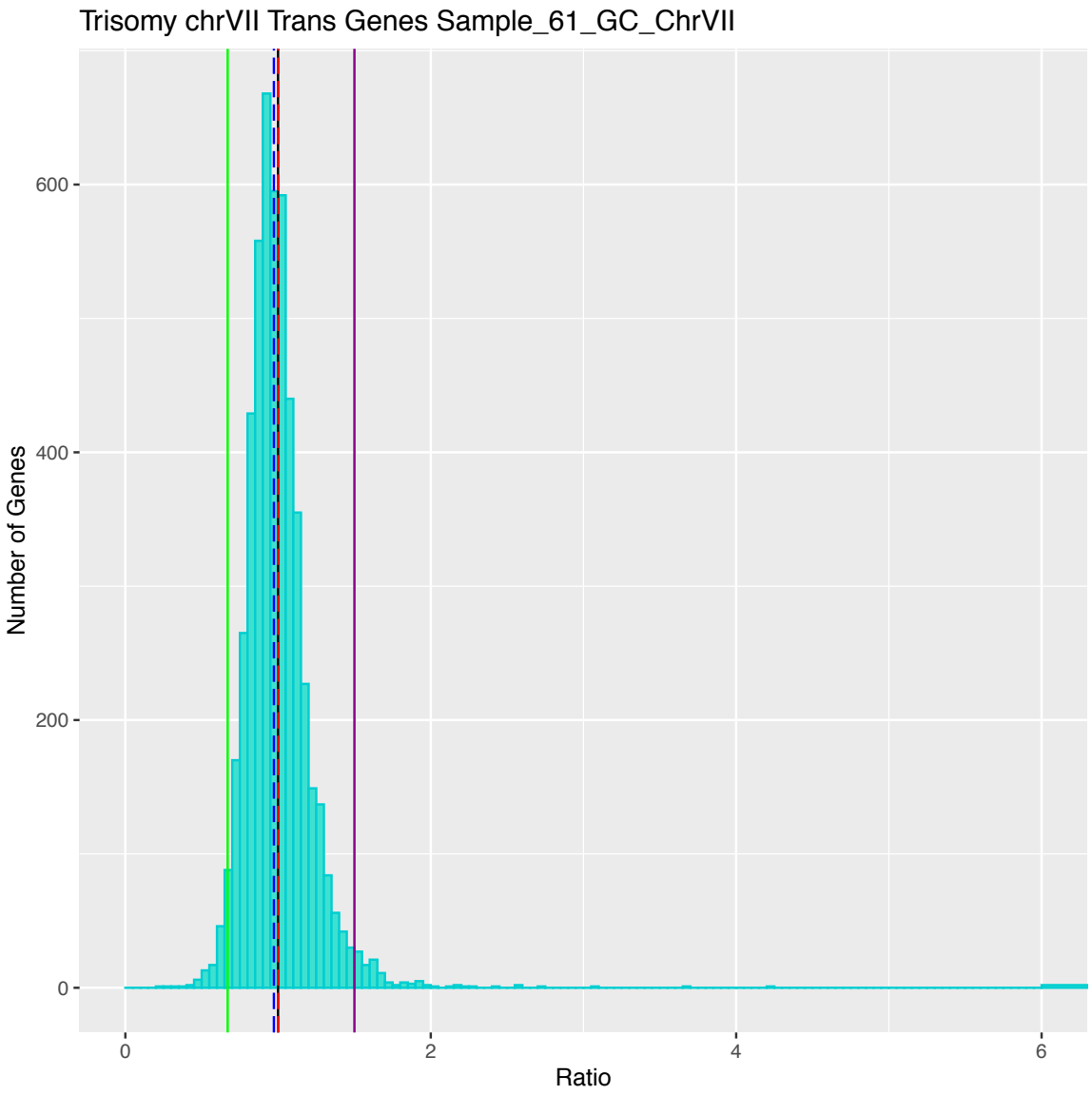

Trisomy chrI Trans Genes Sample_18_GC_ChrI
Number of Genes
400
200
0
0
2
4
6
Ratio

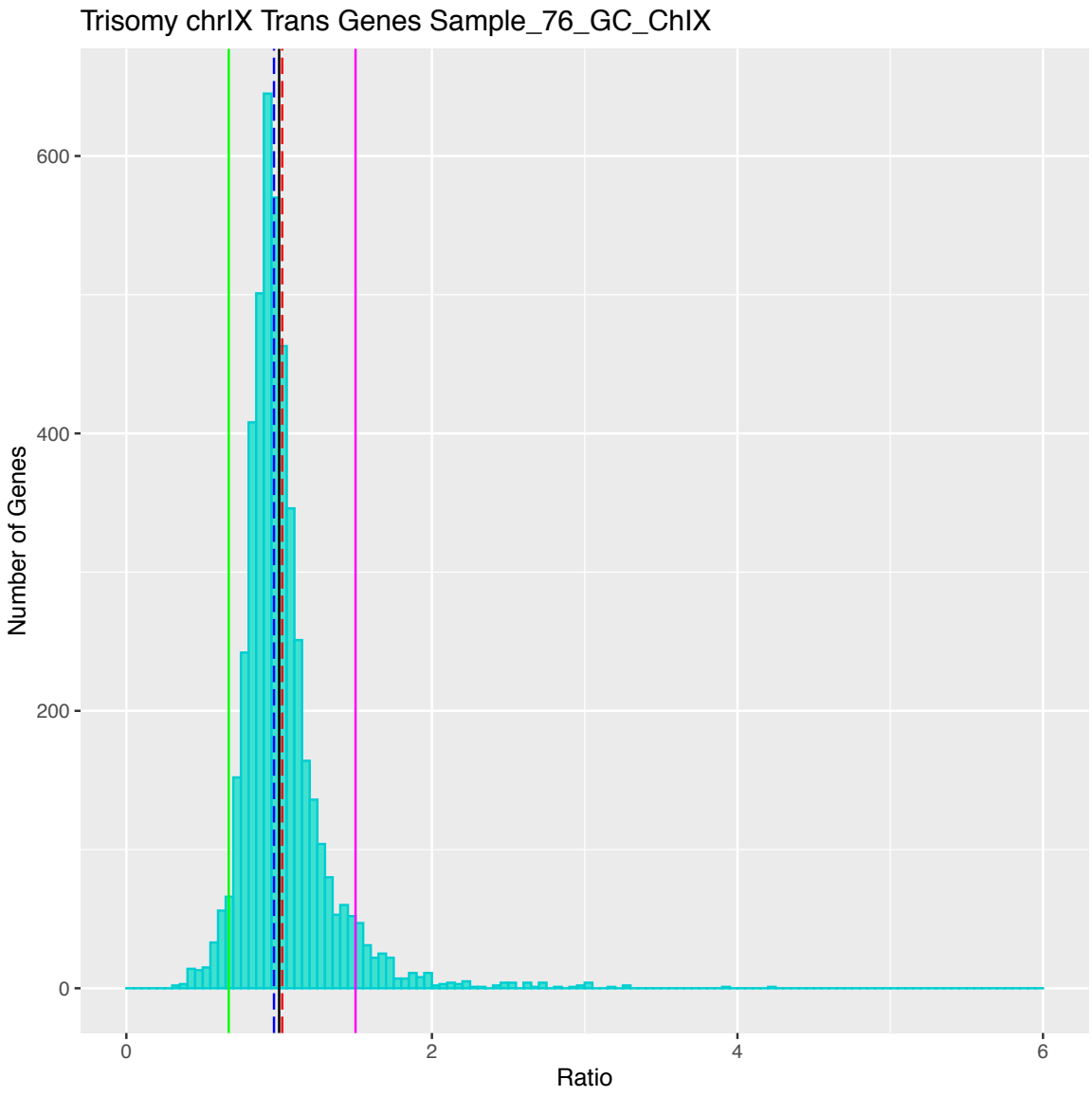

Trisomy chrXII Trans Genes Sample_18_GC_ChrXII
Number of Genes
400
200
0
0
2
4
6
Ratio

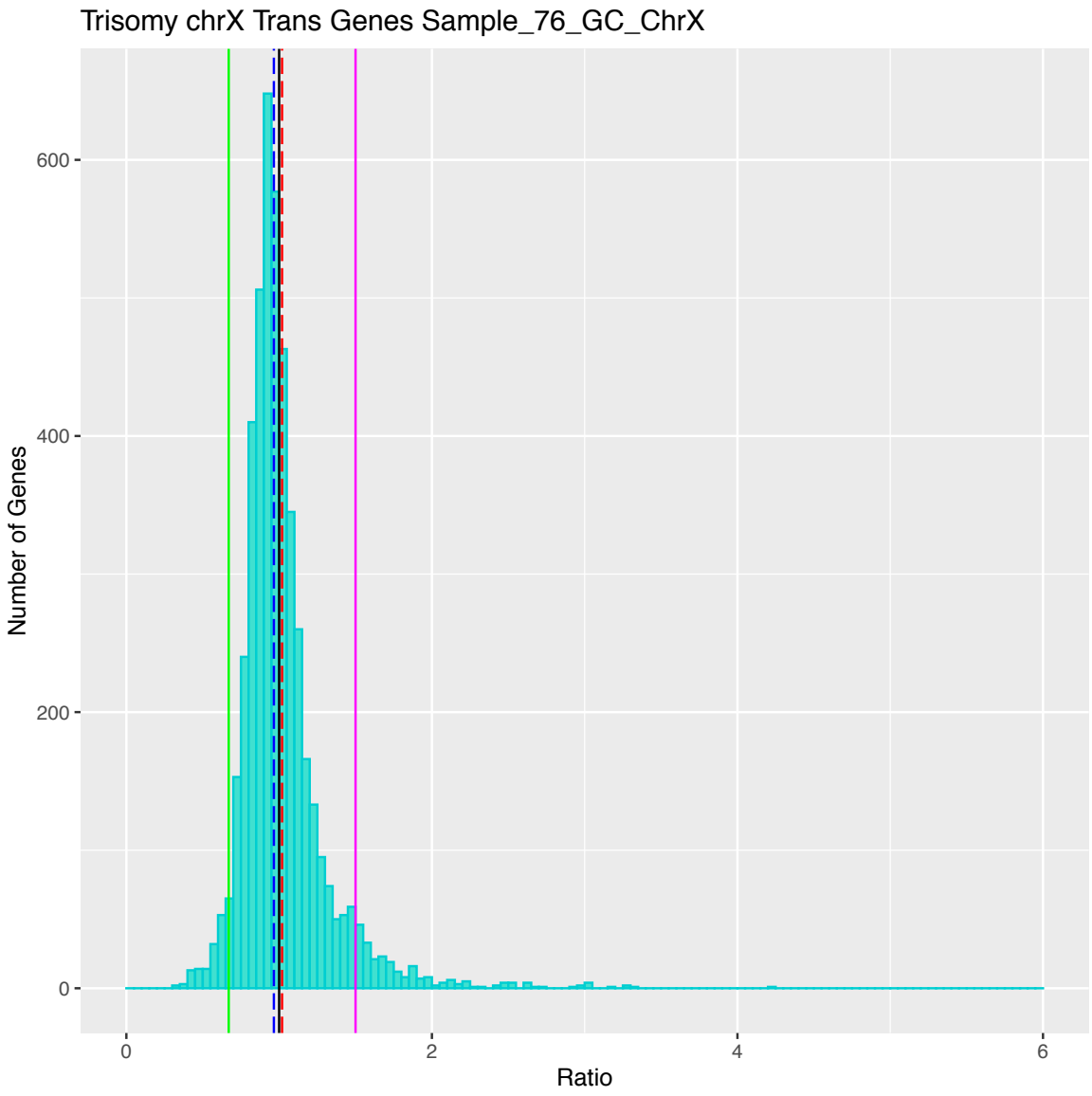

Monosomy chrIX Trans Genes Sample_29_MA_ChrIX
Number of Genes
400
300
200
100
0
0
2
4
6
Ratio

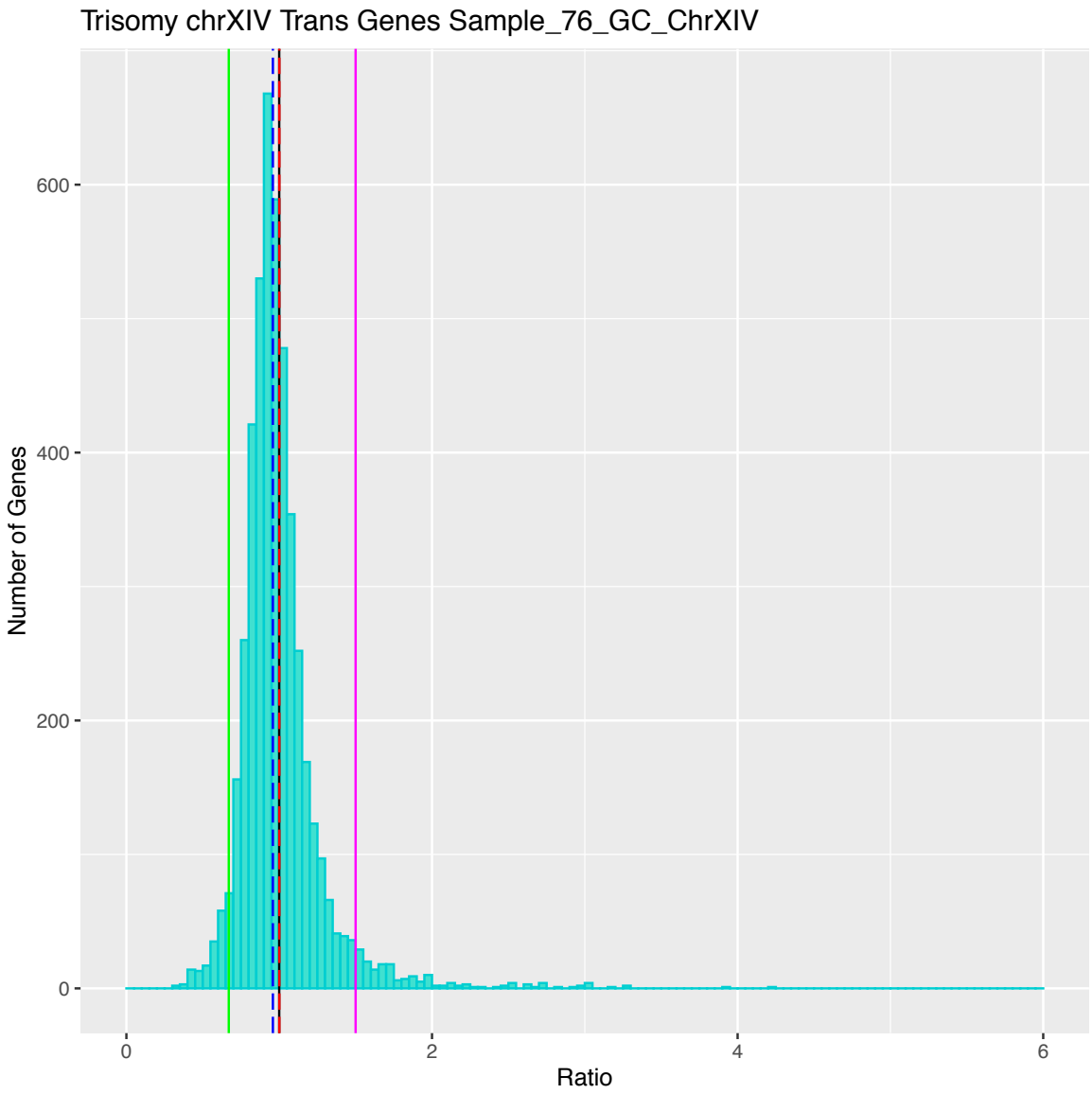

Trisomy chrV Trans Genes Sample_49_GC_ChrV
600
400
200
0
Number of Genes
0
2
4
6
Ratio

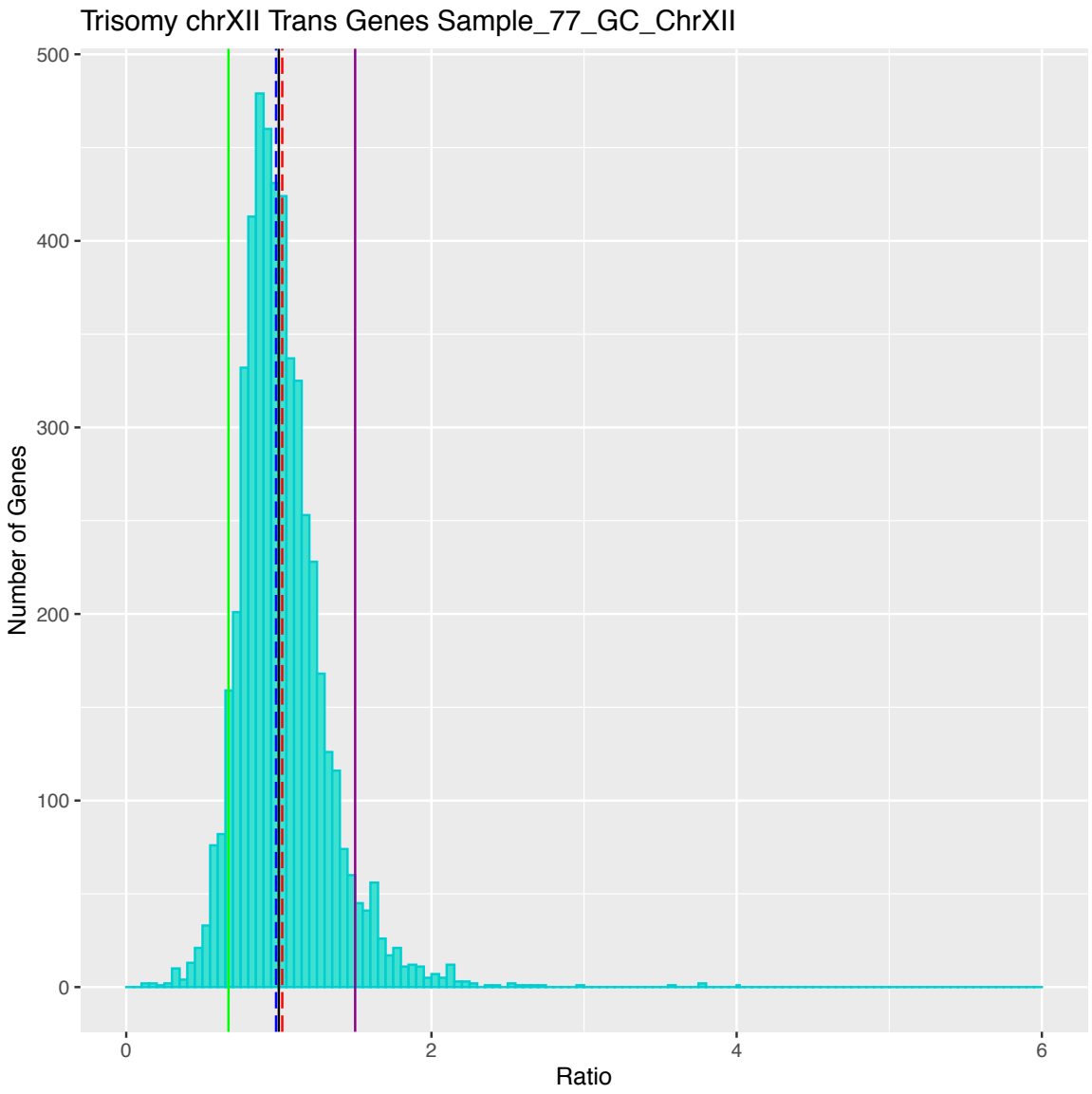

Trisomy chrVII Trans Genes Sample_59_GC_ChrVII
Number of Genes
600
400
200
0
0
2
4
6
Ratio

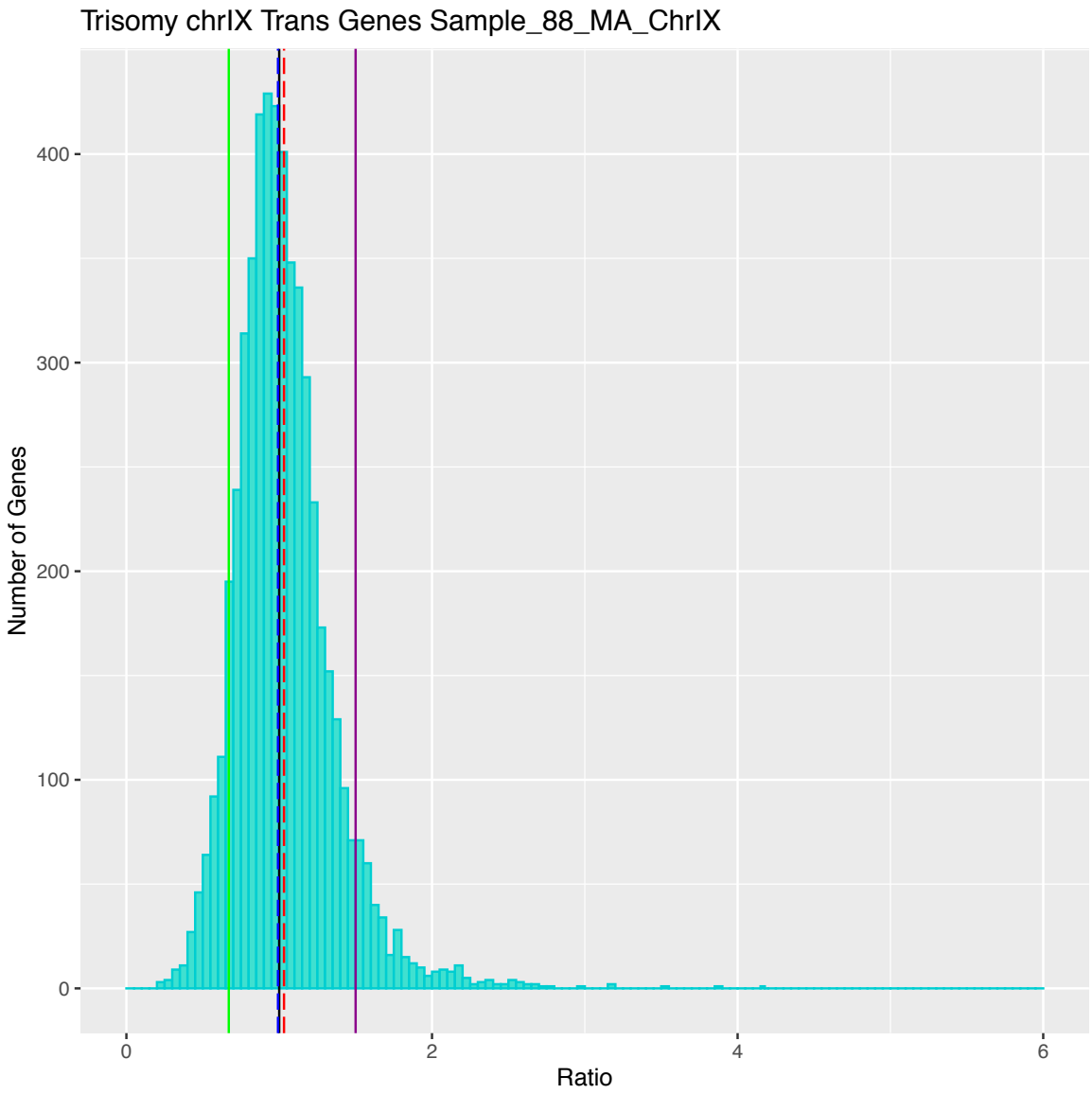

Trisomy chrVII Trans Genes Sample_61_GC_ChrVII
Number of Genes
600
400
200
0
0
2
4
6
Ratio

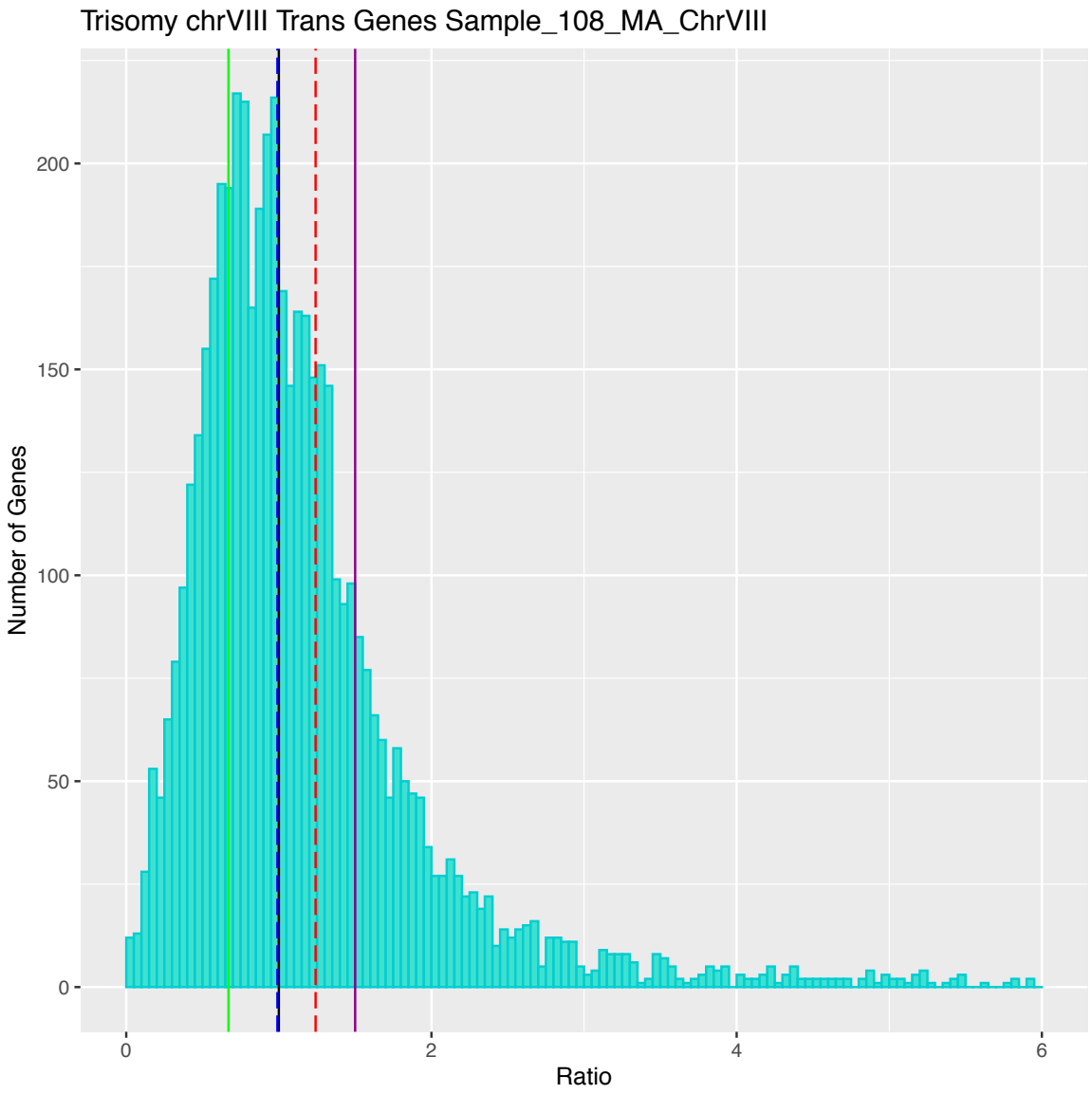

Trisomy chrIX Trans Genes Sample_76_GC_ChIX
Number of Genes
600
400
200
0
0
2
4
6
Ratio

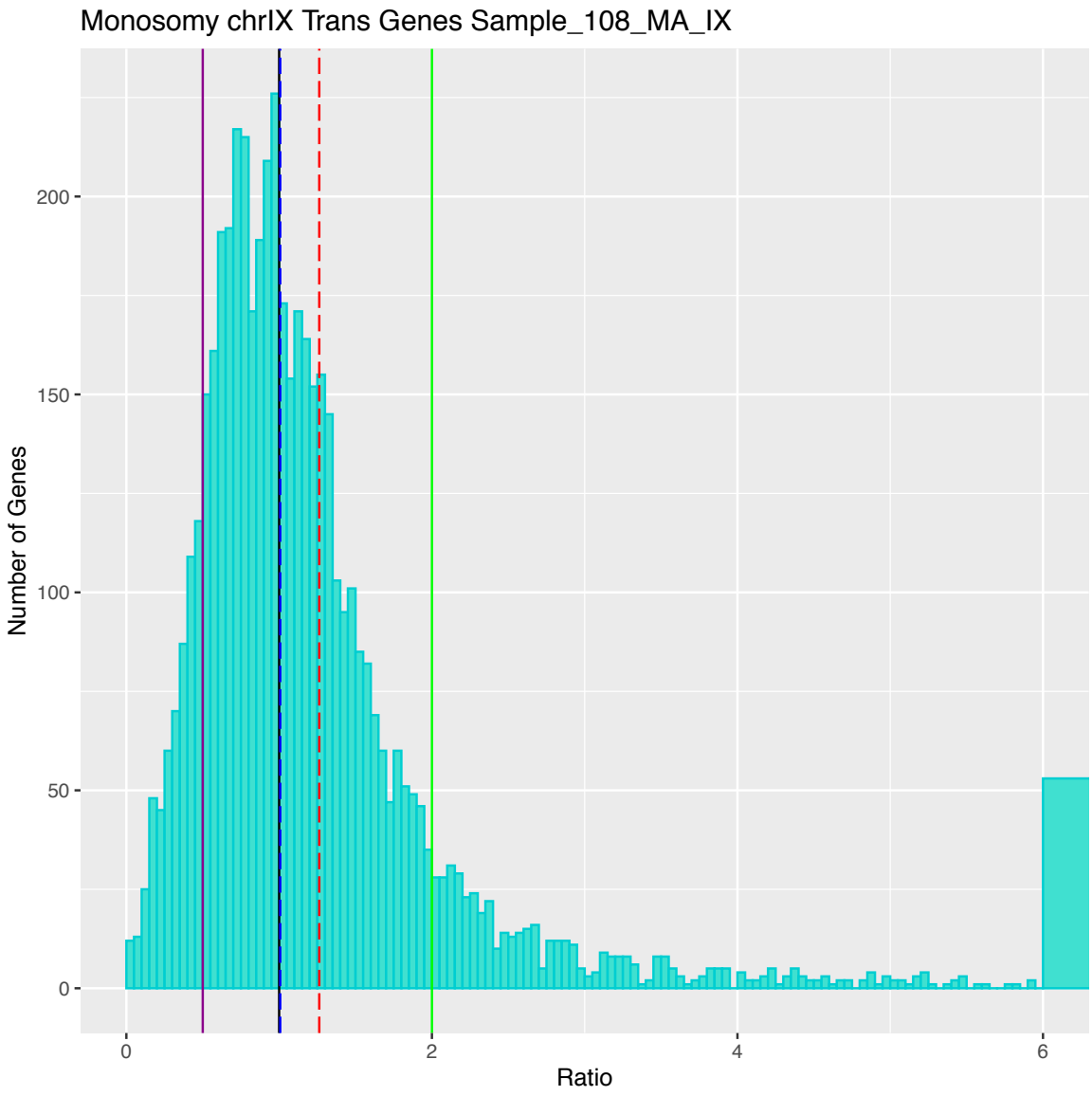

Trisomy chrX Trans Genes Sample_76_GC_ChrX
Number of Genes
600
400
200
0
0
2
4
6
Ratio

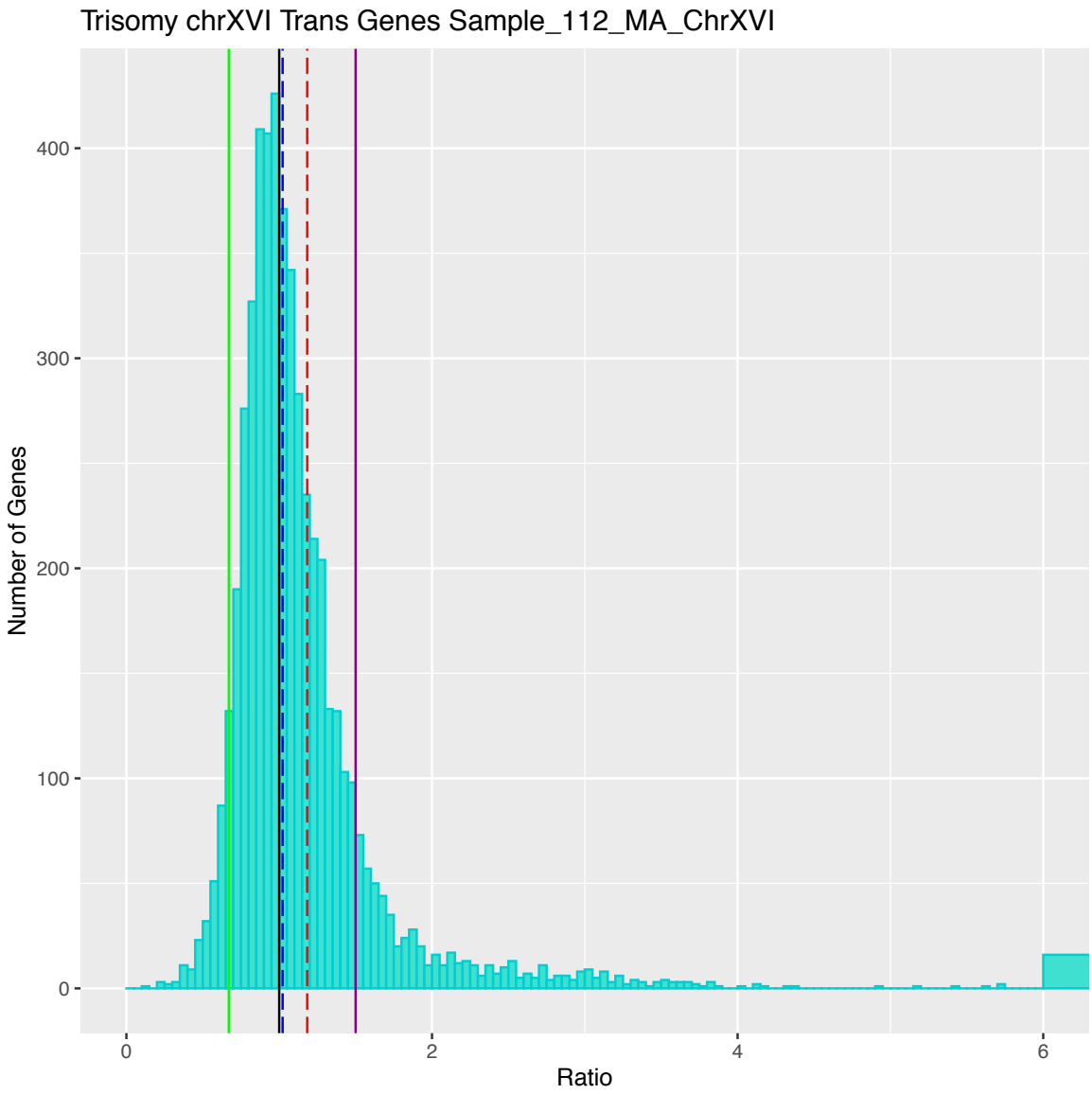

Trisomy chrXIV Trans Genes Sample_76_GC_ChrXIV
Number of Genes
600
400
200
0
0
2
4
6
Ratio

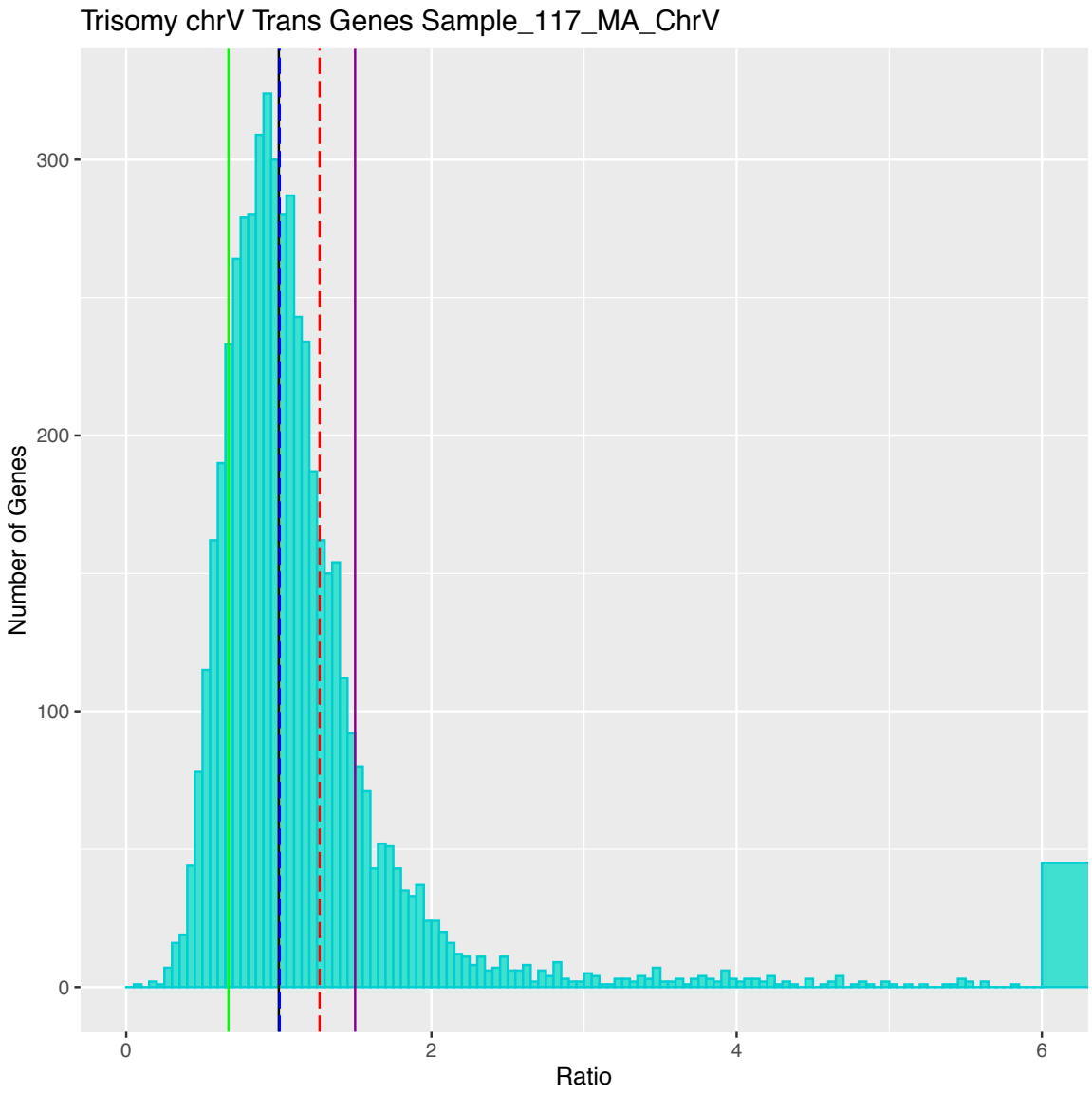

Trisomy chrXII Trans Genes Sample_77_GC_ChrXII
Number of Genes
500
400
300
200
100
0
0
2
4
6
Ratio

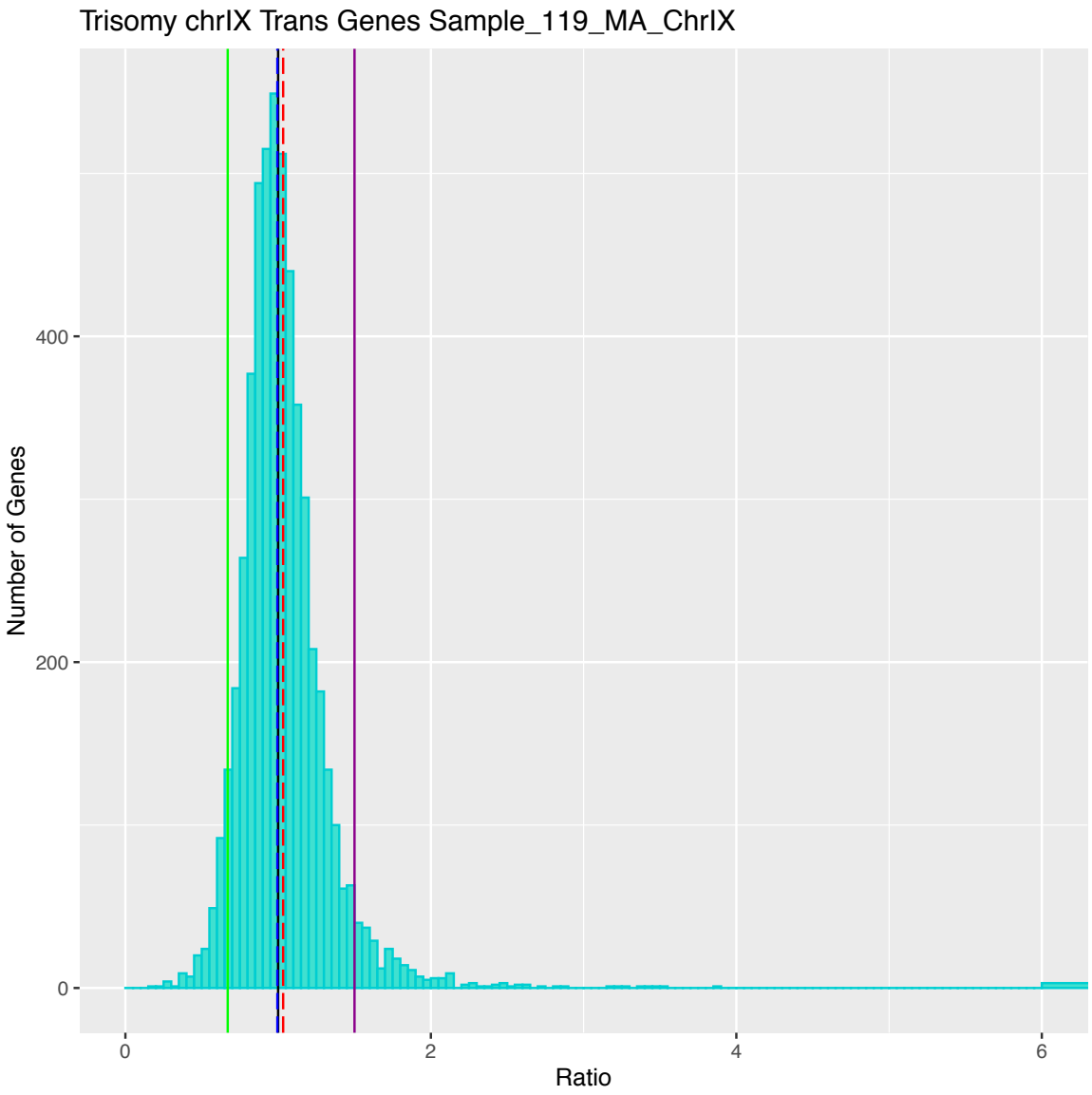

Trisomy chrIX Trans Genes Sample_88_MA_ChrIX
Number of Genes
400
300
200
100
0
0
2
4
6
Ratio

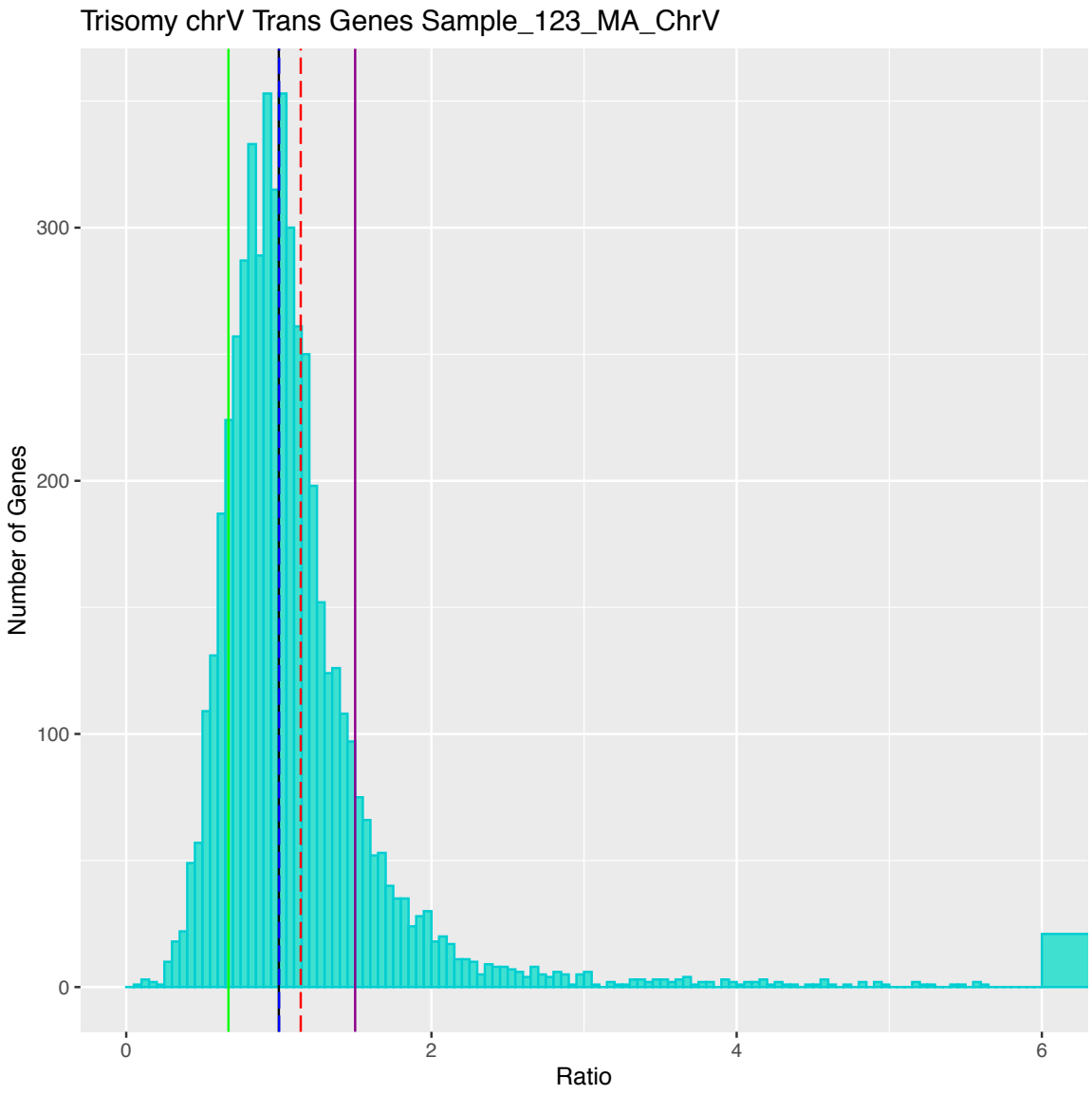

Trisomy chrVIII Trans Genes Sample_108_MA_ChrVIII
Number of Genes
200
150
100
50
0
0
2
4
6
Ratio

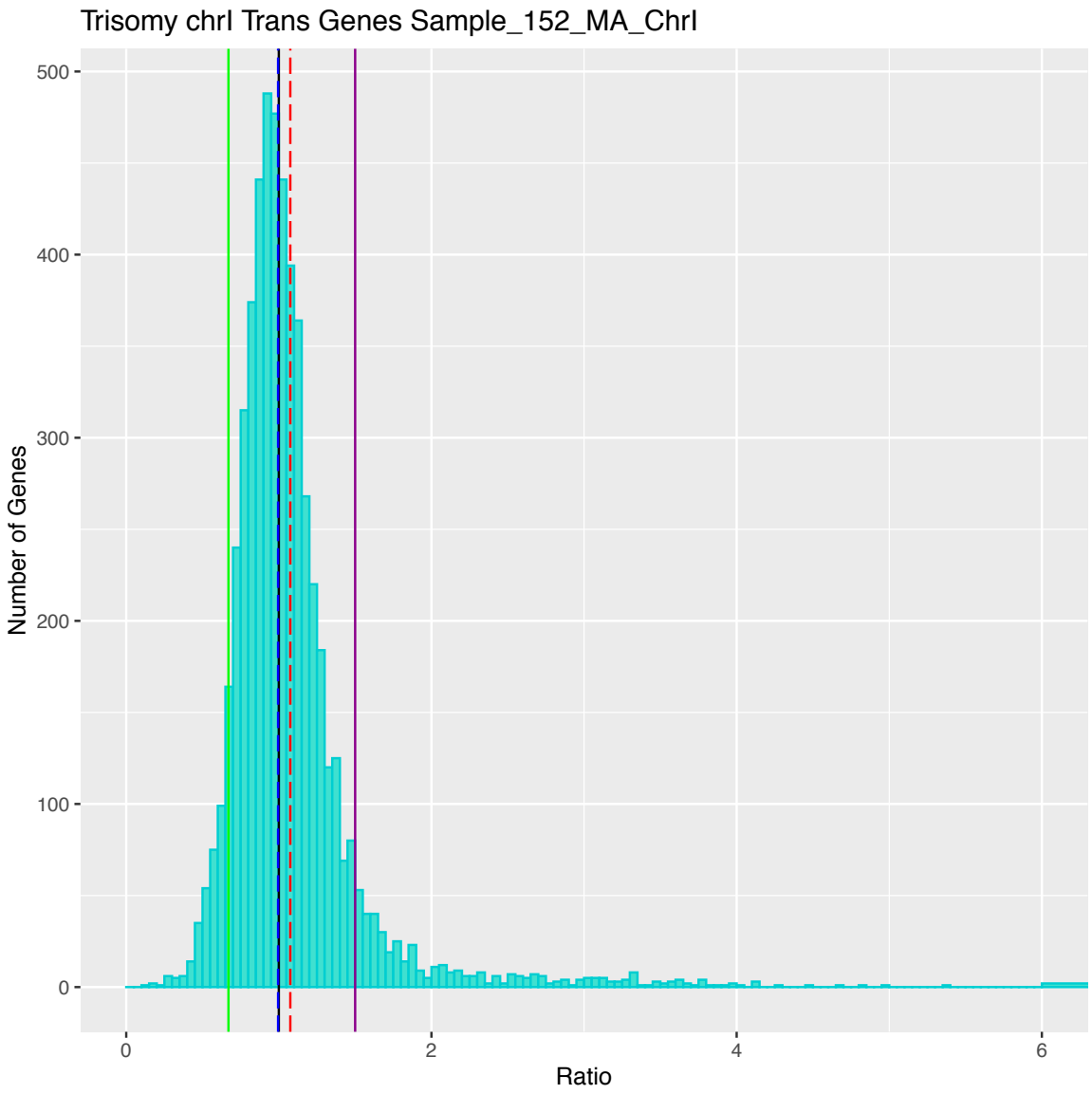

Monosomy chrIX Trans Genes Sample_108_MA_IX
Number of Genes
200
150
100
50
0
0
2
4
6
Ratio

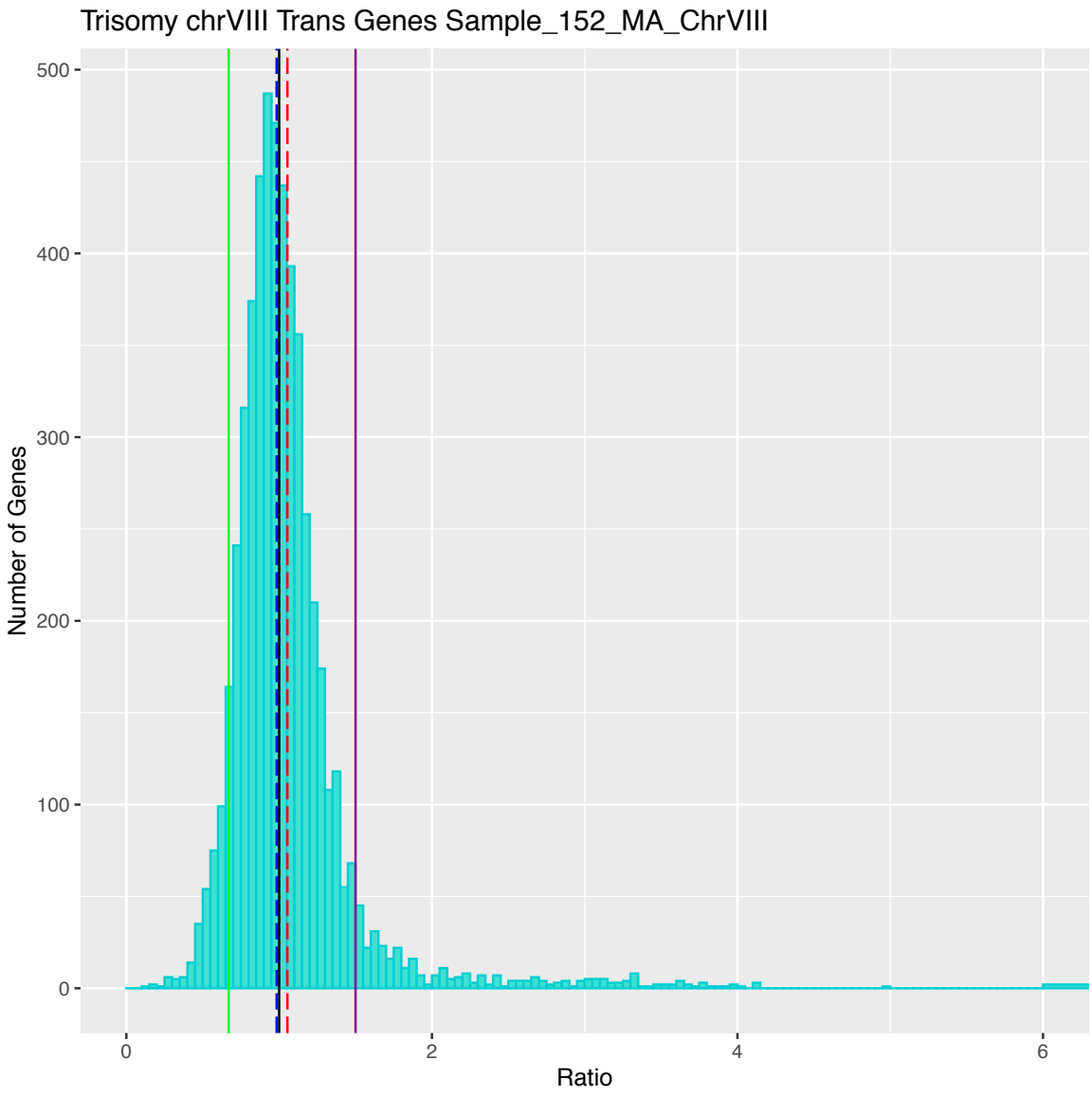

Trisomy chrXVI Trans Genes Sample_112_MA_ChrXVI
Number of Genes
400
300
200
100
0
0
2
4
6
Ratio


Trisomy chrV Trans Genes Sample_117_MA_ChrV
Number of Genes
300
200
100
0
0
2
4
6
Ratio

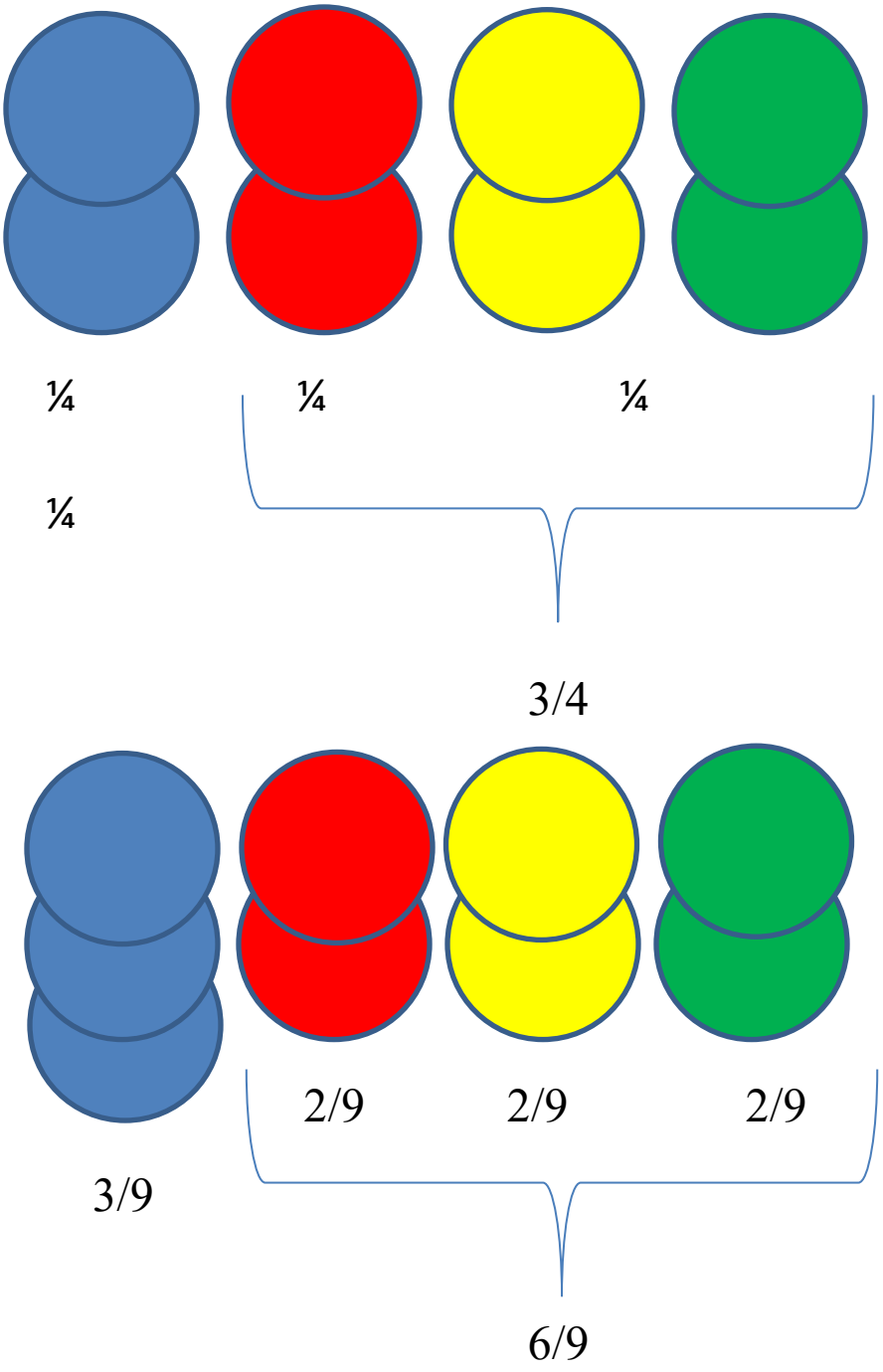

Trisomy chrIX Trans Genes Sample_119_MA_ChrIX
Number of Genes
400
200
0
0
2
4
6
Ratio


Trisomy chrV Trans Genes Sample_123_MA_ChrV
Number of Genes
300
200
100
0
0
2
4
6
Ratio


Trisomy chrI Trans Genes Sample_152_MA_ChrI
Number of Genes
500
400
300
200
100
0
0
2
4
6
Ratio



Supplemental Figure 2.4: Histograms of gene expression ratio for trans genes for aneuploid lines. Blue dashed line is median gene expression, red dashed line is mean gene expression; black solid line is expected if there is no difference between the sample and the ancestor, purple/magenta solid line is expected for monosomy, and green solid line is expected for trisomy.



Supplemental Figure 2.5: Graphic explaining why there are less reads to trans genes in aneuploid samples.

Supplemental Table 2.1: Histone genes do not show evidence for dosage compensation.

Sample 9 MA: 3n Chrom XIV	CHROM	p val
YNL030W	chrXIV	0.47345714
YNL031C	chrXIV	0.02953073
Sample 5 GC: Euploid		
YNL031C	chrXIV	-0.5735449
Sample 6 GC: Euploid		
YNL031C	chrXIV	-0.602327796
Sample 2 GC Euploid		
YNL031C	chrXIV	-0.783581586
Sample 4 MA Euploid		
YNL031C	chrXIV	-0.571639056
Sample 76 GC: 3n Chrom XIV		
YNL030W	chrXIV	1.30707478
Sample 5 GC Euploid		
YNL030W	chrXIV	-0.456516426
Sample 6 GC: Euploid		
YNL030W	chrXIV	-1.069145026
Sample 2 GC Euploid		
YNL030W	chrXIV	-1.163354563
Sample 5 MA Euploid		
YNL030W	chrXIV	1.260712151
Sample 8 MA Euploid		
YNL030W	chrXIV	1.015915942
Sample 76 GC: 3n Chrom XIV		
YNL031C	chrXIV	1.286739747
Sample 11 GC: 3n Chrom XV		
YOL012C	chrXV	0.02705394
Sample 5 GC Euploid		
YOL012C	chrXV	-0.746267705

Sample 6 GC: Euploid		
YOL012C	chrXV	-0.661943487
Sample 2 GC Euploid		
YOL012C	chrXV	-0.867495188
Sample 112 MA: 3n Chrom XVI		
YPL127C	chrXVI	0.60187799
Sample 5 GC Euploid		
YPL127C	chrXVI	-1.338035551
Sample 2 GC Euploid		
YPL127C	chrXVI	-1.031325188
Sample 7 MA Euploid		
YPL127C	chrXVI	-0.599943193
Sample 8 GC: 4n Chrom XVI		
YPL127C	chrXVI	-0.5278193

Supplemental Data: ANOVA results for heterozygous ancestor lines

```
Call:
lm(formula = y ~ Line, data = chr10DataGC)

Residuals:
   Min     1Q  Median     3Q    Max
-5.8764 -0.2104 -0.0039  0.2181  2.6883

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.043450   0.027753   1.566 0.117491
Line2       -0.071882   0.039248  -1.831 0.067080 .
Line3       -0.105918   0.039248  -2.699 0.006981 **
Line4       -0.043226   0.039248  -1.101 0.270788
Line5       -0.079235   0.039248  -2.019 0.043552 *
Line7       -0.055174   0.039248  -1.406 0.159845
Line8       -0.076579   0.039248  -1.951 0.051087 .
Line9       -0.013909   0.039248  -0.354 0.723060
Line11      -0.020249   0.039248  -0.516 0.605936
Line18      -0.130685   0.039248  -3.330 0.000875 ***
Line49      -0.092968   0.039248  -2.369 0.017882 *
Line59      -0.066252   0.039248  -1.688 0.091459 .
Line61      -0.087181   0.039248  -2.221 0.026370 *
Line69      -0.008323   0.039248  -0.212 0.832061
Line76       0.189169   0.039248   4.820 1.47e-06 ***
Line77      -0.051198   0.039248  -1.304 0.192124
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.536 on 5952 degrees of freedom
Multiple R-squared:  0.01681,       Adjusted R-squared:  0.01434
F-statistic: 6.786 on 15 and 5952 DF,  p-value: 8.272e-15


Call:
lm(formula = y ~ Line, data = chr11DataGC)

Residuals:
   Min     1Q  Median     3Q    Max
-4.0085 -0.2092 -0.0084  0.1985  6.0719

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.07613    0.02930   2.599  0.00938 **
Line2       -0.12361    0.04143  -2.984  0.00286 **
Line3       -0.10789    0.04143  -2.604  0.00924 **
Line4       -0.04171    0.04143  -1.007  0.31417
Line5       -0.07830    0.04143  -1.890  0.05883 .
Line7       -0.06313    0.04143  -1.524  0.12764
Line8       -0.08273    0.04143  -1.997  0.04591 *
Line9       -0.07338    0.04143  -1.771  0.07660 .
Line11      -0.03711    0.04143  -0.896  0.37043
```

```
Line18      -0.13973    0.04143  -3.373  0.00075 ***
Line49      -0.06644    0.04143  -1.604  0.10886
Line59      -0.07867    0.04143  -1.899  0.05766 .
Line61      -0.09706    0.04143  -2.343  0.01919 *
Line69      -0.05870    0.04143  -1.417  0.15662
Line76      -0.07057    0.04143  -1.703  0.08859 .
Line77      -0.09405    0.04143  -2.270  0.02325 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5378 on 5376 degrees of freedom
Multiple R-squared:  0.00372,       Adjusted R-squared:  0.0009402
F-statistic: 1.338 on 15 and 5376 DF,  p-value: 0.1696


Call:
lm(formula = y ~ Line, data = chr12DataGC)

Residuals:
   Min      1Q  Median      3Q     Max
-8.4229 -0.2265 -0.0140  0.2157  4.8866

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.10028    0.02566   3.908 9.36e-05 ***
Line2       -0.05953    0.03629  -1.640 0.100945
Line3       -0.12219    0.03629  -3.367 0.000762 ***
Line4       -0.10250    0.03629  -2.825 0.004742 **
Line5       -0.07449    0.03629  -2.053 0.040127 *
Line7       -0.09646    0.03629  -2.658 0.007865 **
Line8       -0.09899    0.03629  -2.728 0.006386 **
Line9       -0.01809    0.03629  -0.498 0.618163
Line11       0.11652    0.03629   3.211 0.001327 **
Line18       0.41606    0.03629  11.466  < 2e-16 ***
Line49      -0.20237    0.03629  -5.577 2.52e-08 ***
Line59      -0.11217    0.03629  -3.091 0.002000 **
Line61      -0.15248    0.03629  -4.202 2.67e-05 ***
Line69      -0.05896    0.03629  -1.625 0.104219
Line76      -0.12915    0.03629  -3.559 0.000374 ***
Line77       0.49135    0.03629  13.541  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5884 on 8400 degrees of freedom
Multiple R-squared:  0.0944,        Adjusted R-squared:  0.09278
F-statistic: 58.37 on 15 and 8400 DF,  p-value: < 2.2e-16


Call:
lm(formula = y ~ Line, data = chr13DataGC)

Residuals:
   Min      1Q  Median      3Q     Max
```

```
-5.7087 -0.2288 -0.0230  0.1967  7.2752

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.10826    0.02623   4.127 3.71e-05 ***
Line2       -0.13847    0.03709  -3.733 0.000191 ***
Line3       -0.14548    0.03709  -3.922 8.87e-05 ***
Line4       -0.07265    0.03709  -1.959 0.050206 .
Line5       -0.12587    0.03709  -3.393 0.000694 ***
Line7       -0.10293    0.03709  -2.775 0.005538 **
Line8       -0.07761    0.03709  -2.092 0.036446 *
Line9       -0.07164    0.03709  -1.931 0.053483 .
Line11      -0.07257    0.03709  -1.956 0.050444 .
Line18      -0.13354    0.03709  -3.600 0.000320 ***
Line49      -0.11282    0.03709  -3.041 0.002363 **
Line59      -0.09163    0.03709  -2.470 0.013526 *
Line61      -0.12742    0.03709  -3.435 0.000595 ***
Line69      -0.05003    0.03709  -1.349 0.177508
Line76      -0.11462    0.03709  -3.090 0.002008 **
Line77      -0.09452    0.03709  -2.548 0.010849 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5753 on 7680 degrees of freedom
Multiple R-squared:  0.00407,       Adjusted R-squared:  0.002125
F-statistic: 2.092 on 15 and 7680 DF,  p-value: 0.007888


Call:
lm(formula = y ~ Line, data = chr14DataGC)

Residuals:
   Min      1Q  Median      3Q     Max
-6.6706 -0.2183  0.0001  0.2052  3.8830

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.08537    0.02413   3.538 0.000406 ***
Line2       -0.01535    0.03412  -0.450 0.652930
Line3       -0.01504    0.03412  -0.441 0.659520
Line4       -0.07725    0.03412  -2.264 0.023622 *
Line5        0.01185    0.03412   0.347 0.728506
Line7       -0.08029    0.03412  -2.353 0.018662 *
Line8       -0.03715    0.03412  -1.089 0.276401
Line9       -0.02938    0.03412  -0.861 0.389281
Line11      -0.10141    0.03412  -2.972 0.002973 **
Line18      -0.12089    0.03412  -3.543 0.000399 ***
Line49      -0.14967    0.03412  -4.386 1.17e-05 ***
Line59      -0.08370    0.03412  -2.453 0.014204 *
Line61      -0.10711    0.03412  -3.139 0.001704 **
Line69      -0.06939    0.03412  -2.034 0.042036 *
Line76       0.42708    0.03412  12.515  < 2e-16 ***
Line77      -0.07181    0.03412  -2.104 0.035380 *
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4921 on 6640 degrees of freedom
Multiple R-squared:  0.06223,	Adjusted R-squared:  0.06011
F-statistic: 29.37 on 15 and 6640 DF,  p-value: < 2.2e-16


Call:
lm(formula = y ~ Line, data = chr15DataGC)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9311 -0.2113  0.0074  0.2269  3.3413

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03715    0.02200   1.689 0.091262 .
Line2       -0.09965    0.03111  -3.204 0.001362 **
Line3       -0.10116    0.03111  -3.252 0.001149 **
Line4       -0.03825    0.03111  -1.230 0.218831
Line5       -0.05921    0.03111  -1.903 0.057013 .
Line7       -0.07368    0.03111  -2.369 0.017873 *
Line8       -0.03093    0.03111  -0.994 0.320127
Line9       -0.07430    0.03111  -2.388 0.016939 *
Line11       0.38144    0.03111  12.263  < 2e-16 ***
Line18      -0.08076    0.03111  -2.596 0.009442 **
Line49      -0.10819    0.03111  -3.478 0.000507 ***
Line59      -0.08489    0.03111  -2.729 0.006363 **
Line61      -0.10178    0.03111  -3.272 0.001071 **
Line69      -0.03643    0.03111  -1.171 0.241557
Line76      -0.07287    0.03111  -2.343 0.019172 *
Line77      -0.05099    0.03111  -1.639 0.101213
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5242 on 9072 degrees of freedom
Multiple R-squared:  0.04413,	Adjusted R-squared:  0.04255
F-statistic: 27.92 on 15 and 9072 DF,  p-value: < 2.2e-16


Call:
lm(formula = y ~ Line, data = chr16DataGC)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0310 -0.2075  0.0006  0.2120  4.3841

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.07071    0.02519   2.807 0.005010 **
Line2       -0.14621    0.03562  -4.105 4.09e-05 ***
Line3       -0.12369    0.03562  -3.472 0.000519 ***
```

```
Line4       -0.02710    0.03562  -0.761 0.446880
Line5       -0.11288    0.03562  -3.169 0.001537 **
Line7       -0.06461    0.03562  -1.814 0.069741 .
Line8        0.80446    0.03562  22.584  < 2e-16 ***
Line9       -0.07048    0.03562  -1.979 0.047907 *
Line11      -0.03090    0.03562  -0.867 0.385771
Line18      -0.12181    0.03562  -3.419 0.000631 ***
Line49      -0.10457    0.03562  -2.936 0.003338 **
Line59      -0.07560    0.03562  -2.122 0.033851 *
Line61      -0.10737    0.03562  -3.014 0.002584 **
Line69      -0.04652    0.03562  -1.306 0.191580
Line76      -0.09823    0.03562  -2.758 0.005836 **
Line77      -0.09818    0.03562  -2.756 0.005861 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5518 on 7664 degrees of freedom
Multiple R-squared:  0.1353,        Adjusted R-squared:  0.1336
F-statistic: 79.93 on 15 and 7664 DF,  p-value: < 2.2e-16


Call:
lm(formula = y ~ Line, data = chr1DataGC)

Residuals:
   Min      1Q  Median      3Q     Max
-3.9123 -0.2504 -0.0161  0.1929  6.5904

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.048111   0.060519   0.795    0.427
Line2        0.104434   0.085587   1.220    0.223
Line3       -0.067564   0.085587  -0.789    0.430
Line4        0.023977   0.085587   0.280    0.779
Line5        0.062101   0.085587   0.726    0.468
Line7        0.498164   0.085587   5.821 7.05e-09 ***
Line8        0.088339   0.085587   1.032    0.302
Line9        0.055628   0.085587   0.650    0.516
Line11      -0.405439   0.085587  -4.737 2.36e-06 ***
Line18       0.384978   0.085587   4.498 7.34e-06 ***
Line49      -0.023784   0.085587  -0.278    0.781
Line59      -0.032118   0.085587  -0.375    0.708
Line61      -0.045047   0.085587  -0.526    0.599
Line69      -0.002338   0.085587  -0.027    0.978
Line76      -0.066835   0.085587  -0.781    0.435
Line77       0.043517   0.085587   0.508    0.611
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6142 on 1632 degrees of freedom
Multiple R-squared:  0.08813,       Adjusted R-squared:  0.07975
F-statistic: 10.51 on 15 and 1632 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = y ~ Line, data = chr2DataGC)

Residuals:
   Min     1Q  Median     3Q    Max
-7.0775 -0.2115  0.0029  0.2187  7.9057

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.017668   0.028117   0.628    0.530
Line2       -0.048387   0.039764  -1.217    0.224
Line3       -0.059357   0.039764  -1.493    0.136
Line4       -0.011688   0.039764  -0.294    0.769
Line5       -0.050971   0.039764  -1.282    0.200
Line7       -0.032734   0.039764  -0.823    0.410
Line8       -0.033988   0.039764  -0.855    0.393
Line9       -0.014371   0.039764  -0.361    0.718
Line11       0.015613   0.039764   0.393    0.695
Line18      -0.033646   0.039764  -0.846    0.398
Line49      -0.042802   0.039764  -1.076    0.282
Line59      -0.039665   0.039764  -0.998    0.319
Line61      -0.052958   0.039764  -1.332    0.183
Line69       0.018538   0.039764   0.466    0.641
Line76      -0.064105   0.039764  -1.612    0.107
Line77      -0.007319   0.039764  -0.184    0.854

Residual standard error: 0.5824 on 6848 degrees of freedom
Multiple R-squared:  0.001848,    Adjusted R-squared:  -0.0003383
F-statistic: 0.8453 on 15 and 6848 DF,  p-value: 0.6271


Call:
lm(formula = y ~ Line, data = chr3DataGC)

Residuals:
   Min     1Q  Median     3Q    Max
-4.1443 -0.2437 -0.0002  0.2351  3.4290

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.094173   0.039757   2.369 0.017919 *
Line2       -0.028822   0.056225  -0.513 0.608258
Line3       -0.106175   0.056225  -1.888 0.059080 .
Line4       -0.061657   0.056225  -1.097 0.272908
Line5       -0.007464   0.056225  -0.133 0.894392
Line7       -0.096421   0.056225  -1.715 0.086476 .
Line8        0.001606   0.056225   0.029 0.977209
Line9       -0.013958   0.056225  -0.248 0.803956
Line11      -0.156085   0.056225  -2.776 0.005540 **
Line18      -0.236045   0.056225  -4.198 2.78e-05 ***
Line49      -0.128924   0.056225  -2.293 0.021924 *
Line59      -0.111442   0.056225  -1.982 0.047571 *
```

```
Line61      -0.136831   0.056225  -2.434 0.015012 *
Line69      -0.082590   0.056225  -1.469 0.141968
Line76      -0.201466   0.056225  -3.583 0.000345 ***
Line77      -0.109070   0.056225  -1.940 0.052497 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5168 on 2688 degrees of freedom
Multiple R-squared:  0.0177,        Adjusted R-squared:  0.01222
F-statistic: 3.229 on 15 and 2688 DF,  p-value: 2.428e-05


Call:
lm(formula = y ~ Line, data = chr4DataGC)

Residuals:
   Min      1Q  Median      3Q     Max
-7.0639 -0.2002 -0.0024  0.1963  6.6083

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.05219    0.01735   3.008 0.002632 **
Line2       -0.09727    0.02453  -3.965 7.38e-05 ***
Line3       -0.08061    0.02453  -3.286 0.001020 **
Line4       -0.06206    0.02453  -2.530 0.011429 *
Line5       -0.06643    0.02453  -2.708 0.006780 **
Line7       -0.07875    0.02453  -3.210 0.001331 **
Line8       -0.06269    0.02453  -2.556 0.010615 *
Line9       -0.04707    0.02453  -1.919 0.055058 .
Line11      -0.01139    0.02453  -0.464 0.642351
Line18      -0.09985    0.02453  -4.070 4.73e-05 ***
Line49      -0.04351    0.02453  -1.774 0.076141 .
Line59      -0.08559    0.02453  -3.489 0.000487 ***
Line61      -0.09692    0.02453  -3.951 7.84e-05 ***
Line69      -0.04154    0.02453  -1.693 0.090469 .
Line76      -0.08729    0.02453  -3.558 0.000375 ***
Line77      -0.06242    0.02453  -2.544 0.010957 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4904 on 12768 degrees of freedom
Multiple R-squared:  0.003389,    Adjusted R-squared:  0.002218
F-statistic: 2.895 on 15 and 12768 DF,  p-value: 0.000138


Call:
lm(formula = y ~ Line, data = chr5DataGC)

Residuals:
   Min      1Q  Median      3Q     Max
-7.2076 -0.2305 -0.0114  0.2243  2.9353

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.07353    0.03048   2.412  0.01589 *
Line2       -0.05512    0.04311  -1.279  0.20111
Line3       -0.09706    0.04311  -2.251  0.02441 *
Line4        0.47164    0.04311  10.940  < 2e-16 ***
Line5       -0.01202    0.04311  -0.279  0.78042
Line7       -0.08066    0.04311  -1.871  0.06142 .
Line8       -0.04762    0.04311  -1.105  0.26935
Line9       -0.04835    0.04311  -1.122  0.26210
Line11      -0.06920    0.04311  -1.605  0.10851
Line18      -0.09618    0.04311  -2.231  0.02573 *
Line49       0.43073    0.04311   9.991  < 2e-16 ***
Line59      -0.09199    0.04311  -2.134  0.03290 *
Line61      -0.11620    0.04311  -2.695  0.00705 **
Line69      -0.08027    0.04311  -1.862  0.06267 .
Line76      -0.14149    0.04311  -3.282  0.00104 **
Line77      -0.08140    0.04311  -1.888  0.05908 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5306 on 4832 degrees of freedom
Multiple R-squared:  0.1002,        Adjusted R-squared:  0.09739
F-statistic: 35.87 on 15 and 4832 DF,  p-value: < 2.2e-16


Call:
lm(formula = y ~ Line, data = chr6DataGC)

Residuals:
   Min      1Q  Median      3Q     Max
-5.6564 -0.2300 -0.0128  0.2017  6.6094

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.060758   0.053282   1.140   0.2543
Line2        0.010294   0.075352   0.137   0.8913
Line3       -0.073950   0.075352  -0.981   0.3265
Line4       -0.047476   0.075352  -0.630   0.5287
Line5        0.006771   0.075352   0.090   0.9284
Line7       -0.066170   0.075352  -0.878   0.3800
Line8       -0.023501   0.075352  -0.312   0.7552
Line9        0.008857   0.075352   0.118   0.9064
Line11      -0.131298   0.075352  -1.742   0.0816 .
Line18      -0.079830   0.075352  -1.059   0.2895
Line49      -0.189594   0.075352  -2.516   0.0119 *
Line59      -0.084888   0.075352  -1.127   0.2601
Line61      -0.019293   0.075352  -0.256   0.7979
Line69      -0.064961   0.075352  -0.862   0.3887
Line76      -0.028100   0.075352  -0.373   0.7093
Line77      -0.048529   0.075352  -0.644   0.5196
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.5909 on 1952 degrees of freedom
Multiple R-squared:  0.007905,    Adjusted R-squared:  0.0002812
F-statistic: 1.037 on 15 and 1952 DF,  p-value: 0.413


Call:
lm(formula = y ~ Line, data = chr7DataGC)

Residuals:
   Min     1Q  Median     3Q    Max
-6.9356 -0.2159 -0.0103  0.2103  3.2824

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.08642    0.02124   4.069 4.76e-05 ***
Line2       -0.08930    0.03003  -2.974 0.002952 **
Line3       -0.09023    0.03003  -3.005 0.002667 **
Line4       -0.05858    0.03003  -1.951 0.051123 .
Line5       -0.07834    0.03003  -2.609 0.009109 **
Line7       -0.08903    0.03003  -2.964 0.003041 **
Line8       -0.07072    0.03003  -2.355 0.018559 *
Line9       -0.03257    0.03003  -1.084 0.278217
Line11      -0.01526    0.03003  -0.508 0.611400
Line18      -0.09476    0.03003  -3.155 0.001609 **
Line49      -0.11943    0.03003  -3.977 7.04e-05 ***
Line59       0.32743    0.03003  10.902  < 2e-16 ***
Line61       0.47031    0.03003  15.660  < 2e-16 ***
Line69      -0.07586    0.03003  -2.526 0.011558 *
Line76      -0.10798    0.03003  -3.595 0.000326 ***
Line77      -0.07257    0.03003  -2.416 0.015698 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5021 on 8928 degrees of freedom
Multiple R-squared:  0.09279,       Adjusted R-squared:  0.09126
F-statistic: 60.88 on 15 and 8928 DF,  p-value: < 2.2e-16


Call:
lm(formula = y ~ Line, data = chr8DataGC)

Residuals:
   Min     1Q  Median     3Q    Max
-4.5644 -0.2176  0.0048  0.2149  4.6639

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.05886    0.03105   1.896  0.05802 .
Line2       -0.06366    0.04391  -1.450  0.14713
Line3       -0.07981    0.04391  -1.818  0.06917 .
Line4       -0.07171    0.04391  -1.633  0.10250
Line5       -0.01711    0.04391  -0.390  0.69686
Line7       -0.07331    0.04391  -1.670  0.09505 .
```

```
Line8       -0.03550    0.04391  -0.809  0.41877
Line9        0.01237    0.04391   0.282  0.77810
Line11      -0.03954    0.04391  -0.901  0.36783
Line18      -0.11449    0.04391  -2.608  0.00914 **
Line49      -0.13559    0.04391  -3.088  0.00203 **
Line59      -0.06524    0.04391  -1.486  0.13739
Line61      -0.08516    0.04391  -1.940  0.05249 .
Line69      -0.05358    0.04391  -1.220  0.22237
Line76      -0.14273    0.04391  -3.251  0.00116 **
Line77      -0.04720    0.04391  -1.075  0.28238
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5241 on 4544 degrees of freedom
Multiple R-squared:  0.006497,    Adjusted R-squared:  0.003217
F-statistic: 1.981 on 15 and 4544 DF,  p-value: 0.01321


Call:
lm(formula = y ~ Line, data = chr9DataGC)

Residuals:
   Min      1Q  Median      3Q     Max
-5.1851 -0.2101  0.0108  0.2193  2.5443

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0270074  0.0304687   0.886 0.375461
Line2       -0.1472850  0.0430892  -3.418 0.000637 ***
Line3       -0.1627409  0.0430892  -3.777 0.000161 ***
Line4       -0.0589257  0.0430892  -1.368 0.171544
Line5       -0.0939892  0.0430892  -2.181 0.029227 *
Line7       -0.0942619  0.0430892  -2.188 0.028762 *
Line8       -0.0987465  0.0430892  -2.292 0.021981 *
Line9       -0.0661724  0.0430892  -1.536 0.124697
Line11      -0.0005015  0.0430892  -0.012 0.990715
Line18      -0.1986916  0.0430892  -4.611 4.14e-06 ***
Line49      -0.0353929  0.0430892  -0.821 0.411480
Line59      -0.0890176  0.0430892  -2.066 0.038909 *
Line61      -0.1136020  0.0430892  -2.636 0.008414 **
Line69      -0.0200296  0.0430892  -0.465 0.642074
Line76       0.5054586  0.0430892  11.731  < 2e-16 ***
Line77      -0.0612492  0.0430892  -1.421 0.155271
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4621 on 3664 degrees of freedom
Multiple R-squared:  0.09833,       Adjusted R-squared:  0.09464
F-statistic: 26.64 on 15 and 3664 DF,  p-value: < 2.2e-16
```

Supplemental Data: ANOVA results for homozygous ancestor lines.

```
Call:
lm(formula = y ~ Line, data = chr1DataMA)

Residuals:
   Min     1Q  Median     3Q    Max
-5.3734 -0.3819 -0.0722  0.2750  6.9357

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.13417    0.07986   1.680 0.093365 .
Line50      -0.13759    0.11294  -1.218 0.223524
Line112     -0.08817    0.11294  -0.781 0.435231
Line115     -0.10362    0.11294  -0.918 0.359163
Line117     -0.21546    0.11294  -1.908 0.056816 .
Line123     -0.12842    0.11294  -1.137 0.255866
Line152      0.40447    0.11294   3.581 0.000365 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8105 on 714 degrees of freedom
Multiple R-squared:  0.05276,       Adjusted R-squared:  0.0448
F-statistic: 6.629 on 6 and 714 DF,  p-value: 7.792e-07


Call:
lm(formula = y ~ Line, data = chr2DataMA)

Residuals:
   Min     1Q  Median     3Q    Max
-7.4584 -0.2692  0.0081  0.2810  5.9203

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0129161  0.0373682   0.346   0.7296
Line2       -0.0553960  0.0528466  -1.048   0.2946
Line3       -0.0841277  0.0528466  -1.592   0.1114
Line4       -0.1156188  0.0528466  -2.188   0.0287 *
Line5        0.0712320  0.0528466   1.348   0.1777
Line7        0.0197163  0.0528466   0.373   0.7091
Line8        0.0117606  0.0528466   0.223   0.8239
Line9       -0.0737775  0.0528466  -1.396   0.1627
Line11      -0.0238928  0.0528466  -0.452   0.6512
Line15       0.0278875  0.0528466   0.528   0.5977
Line28       0.0446541  0.0528466   0.845   0.3981
Line29       0.0384835  0.0528466   0.728   0.4665
Line50      -0.0470287  0.0528466  -0.890   0.3735
Line88      -0.0459765  0.0528466  -0.870   0.3843
Line108     -0.1213815  0.0528466  -2.297   0.0216 *
Line112      0.0936873  0.0528466   1.773   0.0763 .
Line115      0.0196062  0.0528466   0.371   0.7106
Line117     -0.0000458  0.0528466  -0.001   0.9993
```

```
Line119     -0.0743308  0.0528466  -1.407   0.1596
Line123     -0.0247336  0.0528466  -0.468   0.6398
Line152     -0.0512348  0.0528466  -0.970   0.3323
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.774 on 8988 degrees of freedom
Multiple R-squared:  0.005476,    Adjusted R-squared:  0.003263
F-statistic: 2.474 on 20 and 8988 DF,  p-value: 0.0002697


Call:
lm(formula = y ~ Line, data = chr3DataMA)

Residuals:
    Min      1Q  Median      3Q     Max
-10.9942  -0.2947   0.0014   0.3006   7.8459

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.019637   0.052724   0.372   0.7096
Line2       -0.079572   0.074563  -1.067   0.2860
Line3       -0.125993   0.074563  -1.690   0.0912 .
Line4       -0.013303   0.074563  -0.178   0.8584
Line5        0.096657   0.074563   1.296   0.1949
Line7       -0.007711   0.074563  -0.103   0.9176
Line8       -0.017923   0.074563  -0.240   0.8101
Line9       -0.043570   0.074563  -0.584   0.5590
Line11      -0.072688   0.074563  -0.975   0.3297
Line15      -0.003437   0.074563  -0.046   0.9632
Line28       0.083700   0.074563   1.123   0.2617
Line29       0.101905   0.074563   1.367   0.1718
Line50      -0.096868   0.074563  -1.299   0.1940
Line88      -0.095966   0.074563  -1.287   0.1982
Line108      0.065116   0.074563   0.873   0.3826
Line112     -0.042174   0.074563  -0.566   0.5717
Line115     -0.033274   0.074563  -0.446   0.6554
Line117     -0.089899   0.074563  -1.206   0.2280
Line119     -0.023888   0.074563  -0.320   0.7487
Line123     -0.117804   0.074563  -1.580   0.1142
Line152     -0.068622   0.074563  -0.920   0.3575
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6854 on 3528 degrees of freedom
Multiple R-squared:  0.009427,    Adjusted R-squared:  0.003812
F-statistic: 1.679 on 20 and 3528 DF,  p-value: 0.0297


Call:
lm(formula = y ~ Line, data = chr4DataMA)

Residuals:
```

```
   Min      1Q  Median      3Q     Max
-7.4796 -0.2538  0.0013  0.2542  7.5309

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0236639  0.0201918   1.172  0.24123
Line2       -0.0266121  0.0285555  -0.932  0.35138
Line3       -0.0656131  0.0285555  -2.298  0.02159 *
Line4       -0.0587243  0.0285555  -2.056  0.03975 *
Line5       -0.0202708  0.0285555  -0.710  0.47779
Line7       -0.0098373  0.0285555  -0.344  0.73048
Line8        0.0054828  0.0285555   0.192  0.84774
Line9       -0.0913638  0.0285555  -3.200  0.00138 **
Line11      -0.0153717  0.0285555  -0.538  0.59037
Line15      -0.0515555  0.0285555  -1.805  0.07102 .
Line28       0.0040727  0.0285555   0.143  0.88659
Line29       0.0726426  0.0285555   2.544  0.01097 *
Line50       0.0233862  0.0285555   0.819  0.41281
Line88      -0.0721350  0.0285555  -2.526  0.01154 *
Line108     -0.0856368  0.0285555  -2.999  0.00271 **
Line112      0.0555600  0.0285555   1.946  0.05171 .
Line115      0.0275490  0.0285555   0.965  0.33468
Line117     -0.0002823  0.0285555  -0.010  0.99211
Line119     -0.0396279  0.0285555  -1.388  0.16523
Line123     -0.0209933  0.0285555  -0.735  0.46224
Line152     -0.0705933  0.0285555  -2.472  0.01344 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5708 on 16758 degrees of freedom
Multiple R-squared:  0.005856,    Adjusted R-squared:  0.004669
F-statistic: 4.935 on 20 and 16758 DF,  p-value: 2.365e-12


Call:
lm(formula = y ~ Line, data = chr5DataMA)

Residuals:
   Min      1Q  Median      3Q     Max
-7.0796 -0.2753  0.0103  0.2698  6.8648

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.053867   0.037331  -1.443   0.1491
Line2       -0.008935   0.052794  -0.169   0.8656
Line3       -0.033448   0.052794  -0.634   0.5264
Line4       -0.032250   0.052794  -0.611   0.5413
Line5        0.213992   0.052794   4.053 5.11e-05 ***
Line7        0.057902   0.052794   1.097   0.2728
Line8        0.086235   0.052794   1.633   0.1024
Line9       -0.040397   0.052794  -0.765   0.4442
Line11       0.053846   0.052794   1.020   0.3078
Line15       0.078736   0.052794   1.491   0.1359
```

```
Line28       0.100687   0.052794   1.907   0.0565 .
Line29       0.128717   0.052794   2.438   0.0148 *
Line50       0.037280   0.052794   0.706   0.4801
Line88      -0.014415   0.052794  -0.273   0.7848
Line108      0.014469   0.052794   0.274   0.7840
Line112      0.147911   0.052794   2.802   0.0051 **
Line115      0.075746   0.052794   1.435   0.1514
Line117      0.535707   0.052794  10.147  < 2e-16 ***
Line119     -0.020049   0.052794  -0.380   0.7041
Line123      0.357888   0.052794   6.779 1.32e-11 ***
Line152     -0.020876   0.052794  -0.395   0.6925
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6498 on 6342 degrees of freedom
Multiple R-squared:  0.04339,       Adjusted R-squared:  0.04038
F-statistic: 14.38 on 20 and 6342 DF,  p-value: < 2.2e-16


Call:
lm(formula = y ~ Line, data = chr5DataMA)

Residuals:
   Min      1Q  Median      3Q     Max
-7.0796 -0.2753  0.0103  0.2698  6.8648

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.053867   0.037331  -1.443   0.1491
Line2       -0.008935   0.052794  -0.169   0.8656
Line3       -0.033448   0.052794  -0.634   0.5264
Line4       -0.032250   0.052794  -0.611   0.5413
Line5        0.213992   0.052794   4.053 5.11e-05 ***
Line7        0.057902   0.052794   1.097   0.2728
Line8        0.086235   0.052794   1.633   0.1024
Line9       -0.040397   0.052794  -0.765   0.4442
Line11       0.053846   0.052794   1.020   0.3078
Line15       0.078736   0.052794   1.491   0.1359
Line28       0.100687   0.052794   1.907   0.0565 .
Line29       0.128717   0.052794   2.438   0.0148 *
Line50       0.037280   0.052794   0.706   0.4801
Line88      -0.014415   0.052794  -0.273   0.7848
Line108      0.014469   0.052794   0.274   0.7840
Line112      0.147911   0.052794   2.802   0.0051 **
Line115      0.075746   0.052794   1.435   0.1514
Line117      0.535707   0.052794  10.147  < 2e-16 ***
Line119     -0.020049   0.052794  -0.380   0.7041
Line123      0.357888   0.052794   6.779 1.32e-11 ***
Line152     -0.020876   0.052794  -0.395   0.6925
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6498 on 6342 degrees of freedom
```

```
Multiple R-squared:  0.04339,       Adjusted R-squared:  0.04038
F-statistic: 14.38 on 20 and 6342 DF,  p-value: < 2.2e-16


Call:
lm(formula = y ~ Line, data = chr6DataMA)

Residuals:
   Min      1Q  Median      3Q     Max
-5.3621 -0.3027 -0.0049  0.2581  3.9222

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.108431   0.062052   1.747   0.0807 .
Line2       -0.129085   0.087754  -1.471   0.1414
Line3       -0.153480   0.087754  -1.749   0.0804 .
Line4       -0.058092   0.087754  -0.662   0.5080
Line5       -0.014594   0.087754  -0.166   0.8679
Line7       -0.103712   0.087754  -1.182   0.2374
Line8       -0.106896   0.087754  -1.218   0.2233
Line9       -0.032061   0.087754  -0.365   0.7149
Line11      -0.086260   0.087754  -0.983   0.3257
Line15       0.002664   0.087754   0.030   0.9758
Line28      -0.093758   0.087754  -1.068   0.2854
Line29      -0.046956   0.087754  -0.535   0.5926
Line50      -0.032637   0.087754  -0.372   0.7100
Line88      -0.090652   0.087754  -1.033   0.3017
Line108     -0.081861   0.087754  -0.933   0.3510
Line112     -0.089905   0.087754  -1.025   0.3057
Line115     -0.120181   0.087754  -1.370   0.1710
Line117     -0.103711   0.087754  -1.182   0.2374
Line119     -0.007226   0.087754  -0.082   0.9344
Line123     -0.069580   0.087754  -0.793   0.4279
Line152     -0.141146   0.087754  -1.608   0.1079
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6882 on 2562 degrees of freedom
Multiple R-squared:  0.004419,     Adjusted R-squared:  -0.003353
F-statistic: 0.5686 on 20 and 2562 DF,  p-value: 0.9356


Call:
lm(formula = y ~ Line, data = chr7DataMA)

Residuals:
   Min      1Q  Median      3Q     Max
-8.6286 -0.2708 -0.0183  0.2388  7.2742

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.06604    0.02742   2.408   0.0161 *
Line2       -0.02973    0.03878  -0.767   0.4433
```

```
Line3       -0.05234    0.03878  -1.350   0.1771
Line4       -0.07429    0.03878  -1.916   0.0554 .
Line5       -0.02810    0.03878  -0.725   0.4687
Line7       -0.02971    0.03878  -0.766   0.4436
Line8       -0.04427    0.03878  -1.141   0.2537
Line9       -0.07003    0.03878  -1.806   0.0710 .
Line11      -0.05168    0.03878  -1.333   0.1827
Line15      -0.04035    0.03878  -1.040   0.2982
Line28      -0.01507    0.03878  -0.389   0.6976
Line29       0.03339    0.03878   0.861   0.3893
Line50      -0.02525    0.03878  -0.651   0.5150
Line88      -0.07270    0.03878  -1.875   0.0609 .
Line108     -0.05322    0.03878  -1.372   0.1700
Line112      0.03127    0.03878   0.806   0.4200
Line115     -0.02382    0.03878  -0.614   0.5391
Line117     -0.04131    0.03878  -1.065   0.2868
Line119     -0.06196    0.03878  -1.598   0.1101
Line123     -0.06786    0.03878  -1.750   0.0802 .
Line152     -0.09111    0.03878  -2.349   0.0188 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6484 on 11718 degrees of freedom
Multiple R-squared:  0.002374,    Adjusted R-squared:  0.0006718
F-statistic: 1.395 on 20 and 11718 DF,  p-value: 0.1123


Call:
lm(formula = y ~ Line, data = chr8DataMA)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8911 -0.2930 -0.0259  0.2594  7.4064

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.042601   0.038278   1.113   0.2658
Line2       -0.027706   0.054134  -0.512   0.6088
Line3       -0.060945   0.054134  -1.126   0.2603
Line4       -0.063369   0.054134  -1.171   0.2418
Line5        0.044807   0.054134   0.828   0.4079
Line7       -0.010838   0.054134  -0.200   0.8413
Line8       -0.003586   0.054134  -0.066   0.9472
Line9       -0.051800   0.054134  -0.957   0.3387
Line11      -0.046988   0.054134  -0.868   0.3854
Line15      -0.025058   0.054134  -0.463   0.6435
Line28       0.008869   0.054134   0.164   0.8699
Line29      -0.007671   0.054134  -0.142   0.8873
Line50      -0.089729   0.054134  -1.658   0.0975 .
Line88      -0.026332   0.054134  -0.486   0.6267
Line108      0.584627   0.054134  10.800   <2e-16 ***
Line112      0.004753   0.054134   0.088   0.9300
Line115     -0.048766   0.054134  -0.901   0.3677
```

```
Line117     -0.128824   0.054134  -2.380   0.0174 *
Line119     -0.053209   0.054134  -0.983   0.3257
Line123     -0.120320   0.054134  -2.223   0.0263 *
Line152      0.500959   0.054134   9.254   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6462 on 5964 degrees of freedom
Multiple R-squared:  0.06896,       Adjusted R-squared:  0.06584
F-statistic: 22.09 on 20 and 5964 DF,  p-value: < 2.2e-16


Call:
lm(formula = y ~ Line, data = chr9DataMA)

Residuals:
   Min      1Q  Median      3Q     Max
-6.6107 -0.2948 -0.0375  0.2452  7.5298

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.14399    0.04189   3.438 0.000592 ***
Line2       -0.13341    0.05924  -2.252 0.024363 *
Line3       -0.17489    0.05924  -2.952 0.003169 **
Line4       -0.10577    0.05924  -1.786 0.074236 .
Line5        0.02021    0.05924   0.341 0.733042
Line7       -0.06519    0.05924  -1.100 0.271200
Line8       -0.07968    0.05924  -1.345 0.178635
Line9       -0.15104    0.05924  -2.550 0.010813 *
Line11      -0.12438    0.05924  -2.100 0.035807 *
Line15       0.45026    0.05924   7.601 3.52e-14 ***
Line28      -0.06830    0.05924  -1.153 0.248981
Line29      -1.01011    0.05924 -17.052  < 2e-16 ***
Line50      -0.17944    0.05924  -3.029 0.002466 **
Line88       0.40900    0.05924   6.904 5.70e-12 ***
Line108     -0.98399    0.05924 -16.611  < 2e-16 ***
Line112     -0.10504    0.05924  -1.773 0.076246 .
Line115     -0.13069    0.05924  -2.206 0.027415 *
Line117     -0.09210    0.05924  -1.555 0.120076
Line119      0.40545    0.05924   6.844 8.64e-12 ***
Line123     -0.13884    0.05924  -2.344 0.019128 *
Line152     -0.21804    0.05924  -3.681 0.000235 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6353 on 4809 degrees of freedom
Multiple R-squared:  0.2272,        Adjusted R-squared:  0.224
F-statistic: 70.68 on 20 and 4809 DF,  p-value: < 2.2e-16


Call:
lm(formula = y ~ Line, data = chr10DataMA)
```

```
Residuals:
   Min      1Q  Median      3Q     Max
-6.4665 -0.2639 -0.0052  0.2582  7.2217

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.029150   0.034072   0.856    0.392
Line2       -0.019959   0.048185  -0.414    0.679
Line3       -0.036097   0.048185  -0.749    0.454
Line4       -0.074748   0.048185  -1.551    0.121
Line5        0.001790   0.048185   0.037    0.970
Line7       -0.034818   0.048185  -0.723    0.470
Line8        0.005026   0.048185   0.104    0.917
Line9       -0.074034   0.048185  -1.536    0.124
Line11       0.014936   0.048185   0.310    0.757
Line15       0.017669   0.048185   0.367    0.714
Line28      -0.013893   0.048185  -0.288    0.773
Line29       0.054126   0.048185   1.123    0.261
Line50      -0.002511   0.048185  -0.052    0.958
Line88      -0.078711   0.048185  -1.634    0.102
Line108     -0.116090   0.048185  -2.409    0.016 *
Line112      0.061408   0.048185   1.274    0.203
Line115      0.048483   0.048185   1.006    0.314
Line117     -0.017116   0.048185  -0.355    0.722
Line119     -0.064576   0.048185  -1.340    0.180
Line123     -0.041089   0.048185  -0.853    0.394
Line152     -0.067294   0.048185  -1.397    0.163
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.658 on 7812 degrees of freedom
Multiple R-squared:  0.004941,    Adjusted R-squared:  0.002394
F-statistic:  1.94 on 20 and 7812 DF,  p-value: 0.00718


Call:
lm(formula = y ~ Line, data = chr11DataMA)

Residuals:
   Min      1Q  Median      3Q     Max
-5.3800 -0.2654 -0.0053  0.2548  7.1361

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.046100   0.034231   1.347  0.17811
Line2       -0.069521   0.048411  -1.436  0.15102
Line3       -0.079502   0.048411  -1.642  0.10058
Line4       -0.041052   0.048411  -0.848  0.39647
Line5       -0.002335   0.048411  -0.048  0.96154
Line7       -0.035455   0.048411  -0.732  0.46396
Line8       -0.011645   0.048411  -0.241  0.80992
Line9       -0.124375   0.048411  -2.569  0.01021 *
Line11      -0.045212   0.048411  -0.934  0.35038
```

```
Line15      -0.051030   0.048411  -1.054  0.29187
Line28       0.004407   0.048411   0.091  0.92747
Line29       0.081709   0.048411   1.688  0.09149 .
Line50       0.040056   0.048411   0.827  0.40803
Line88      -0.102530   0.048411  -2.118  0.03422 *
Line108     -0.137665   0.048411  -2.844  0.00447 **
Line112      0.047871   0.048411   0.989  0.32277
Line115      0.037727   0.048411   0.779  0.43583
Line117      0.018711   0.048411   0.387  0.69913
Line119     -0.071063   0.048411  -1.468  0.14217
Line123     -0.012930   0.048411  -0.267  0.78940
Line152     -0.105463   0.048411  -2.179  0.02940 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6284 on 7056 degrees of freedom
Multiple R-squared:  0.008612,    Adjusted R-squared:  0.005802
F-statistic: 3.065 on 20 and 7056 DF,  p-value: 4.792e-06


Call:
lm(formula = y ~ Line, data = chr12DataMA)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9937 -0.2588  0.0086  0.2757  7.2950

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.006544   0.032968   0.198   0.8427
Line2       -0.016368   0.046624  -0.351   0.7255
Line3       -0.055019   0.046624  -1.180   0.2380
Line4       -0.099940   0.046624  -2.144   0.0321 *
Line5        0.040056   0.046624   0.859   0.3903
Line7       -0.030373   0.046624  -0.651   0.5148
Line8        0.018685   0.046624   0.401   0.6886
Line9       -0.030881   0.046624  -0.662   0.5078
Line11      -0.004223   0.046624  -0.091   0.9278
Line15      -0.015575   0.046624  -0.334   0.7384
Line28      -0.015843   0.046624  -0.340   0.7340
Line29       0.061630   0.046624   1.322   0.1862
Line50      -0.063730   0.046624  -1.367   0.1717
Line88      -0.042324   0.046624  -0.908   0.3640
Line108     -0.005635   0.046624  -0.121   0.9038
Line112      0.025041   0.046624   0.537   0.5912
Line115     -0.025703   0.046624  -0.551   0.5814
Line117     -0.099652   0.046624  -2.137   0.0326 *
Line119     -0.034883   0.046624  -0.748   0.4544
Line123     -0.057875   0.046624  -1.241   0.2145
Line152     -0.030041   0.046624  -0.644   0.5194
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.7561 on 11025 degrees of freedom
Multiple R-squared:  0.002738,    Adjusted R-squared:  0.000929
F-statistic: 1.514 on 20 and 11025 DF,  p-value: 0.06584


Call:
lm(formula = y ~ Line, data = chr13DataMA)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2370 -0.2719 -0.0047  0.2510  6.4732

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0007587  0.0300654   0.025  0.97987
Line2        0.0096555  0.0425190   0.227  0.82036
Line3       -0.0566756  0.0425190  -1.333  0.18258
Line4       -0.0709358  0.0425190  -1.668  0.09528 .
Line5        0.0391742  0.0425190   0.921  0.35690
Line7        0.0011368  0.0425190   0.027  0.97867
Line8        0.0212880  0.0425190   0.501  0.61661
Line9       -0.0612518  0.0425190  -1.441  0.14974
Line11      -0.0078874  0.0425190  -0.186  0.85284
Line15       0.0088876  0.0425190   0.209  0.83443
Line28       0.0257850  0.0425190   0.606  0.54424
Line29       0.1008022  0.0425190   2.371  0.01777 *
Line50       0.0538299  0.0425190   1.266  0.20553
Line88      -0.0292602  0.0425190  -0.688  0.49136
Line108     -0.0778097  0.0425190  -1.830  0.06728 .
Line112      0.1257249  0.0425190   2.957  0.00311 **
Line115      0.0709096  0.0425190   1.668  0.09540 .
Line117      0.0520713  0.0425190   1.225  0.22073
Line119     -0.0103935  0.0425190  -0.244  0.80689
Line123      0.0261890  0.0425190   0.616  0.53795
Line152     -0.0404483  0.0425190  -0.951  0.34147
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6594 on 10080 degrees of freedom
Multiple R-squared:  0.006463,    Adjusted R-squared:  0.004492
F-statistic: 3.279 on 20 and 10080 DF,  p-value: 1.001e-06


Call:
lm(formula = y ~ Line, data = chr14DataMA)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8645 -0.2705 -0.0221  0.2515  6.3408

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.028990   0.032410   0.894    0.371
```

```
Line2        0.022777   0.045835   0.497    0.619
Line3       -0.011204   0.045835  -0.244    0.807
Line4       -0.006384   0.045835  -0.139    0.889
Line5        0.042464   0.045835   0.926    0.354
Line7        0.050682   0.045835   1.106    0.269
Line8        0.016167   0.045835   0.353    0.724
Line9        0.504157   0.045835  10.999   <2e-16 ***
Line11      -0.004188   0.045835  -0.091    0.927
Line15      -0.002585   0.045835  -0.056    0.955
Line28       0.036347   0.045835   0.793    0.428
Line29       0.057268   0.045835   1.249    0.212
Line50      -0.002264   0.045835  -0.049    0.961
Line88      -0.005671   0.045835  -0.124    0.902
Line108     -0.007048   0.045835  -0.154    0.878
Line112      0.012519   0.045835   0.273    0.785
Line115      0.025095   0.045835   0.548    0.584
Line117     -0.039718   0.045835  -0.867    0.386
Line119      0.012036   0.045835   0.263    0.793
Line123     -0.023506   0.045835  -0.513    0.608
Line152     -0.043892   0.045835  -0.958    0.338
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.661 on 8715 degrees of freedom
Multiple R-squared:  0.02659,       Adjusted R-squared:  0.02436
F-statistic: 11.91 on 20 and 8715 DF,  p-value: < 2.2e-16


Call:
lm(formula = y ~ Line, data = chr15DataMA)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8229 -0.2865 -0.0098  0.2702  6.7844

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.034629   0.029034   1.193   0.2330
Line2       -0.032954   0.041060  -0.803   0.4222
Line3       -0.058301   0.041060  -1.420   0.1557
Line4       -0.036715   0.041060  -0.894   0.3712
Line5        0.054686   0.041060   1.332   0.1829
Line7        0.018381   0.041060   0.448   0.6544
Line8        0.003219   0.041060   0.078   0.9375
Line9       -0.091813   0.041060  -2.236   0.0254 *
Line11      -0.017415   0.041060  -0.424   0.6715
Line15      -0.027566   0.041060  -0.671   0.5020
Line28       0.004058   0.041060   0.099   0.9213
Line29       0.075875   0.041060   1.848   0.0646 .
Line50      -0.033880   0.041060  -0.825   0.4093
Line88      -0.030914   0.041060  -0.753   0.4515
Line108     -0.004641   0.041060  -0.113   0.9100
Line112      0.095591   0.041060   2.328   0.0199 *
```

```
Line115      0.011767   0.041060   0.287   0.7744
Line117     -0.025635   0.041060  -0.624   0.5324
Line119     -0.029526   0.041060  -0.719   0.4721
Line123     -0.063314   0.041060  -1.542   0.1231
Line152     -0.062302   0.041060  -1.517   0.1292
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.692 on 11907 degrees of freedom
Multiple R-squared:  0.004192,     Adjusted R-squared:  0.002519
F-statistic: 2.506 on 20 and 11907 DF,  p-value: 0.0002177


Call:
lm(formula = y ~ Line, data = chr16DataMA)

Residuals:
   Min      1Q  Median      3Q     Max
-5.8128 -0.2733 -0.0122  0.2534  8.1818

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.059440   0.027832   2.136 0.032732 *
Line2       -0.044433   0.039361  -1.129 0.258985
Line3       -0.062596   0.039361  -1.590 0.111793
Line4       -0.088237   0.039361  -2.242 0.024999 *
Line5        0.018447   0.039361   0.469 0.639320
Line7       -0.025207   0.039361  -0.640 0.521926
Line8       -0.034723   0.039361  -0.882 0.377701
Line9       -0.123296   0.039361  -3.132 0.001739 **
Line11      -0.031360   0.039361  -0.797 0.425623
Line15      -0.042250   0.039361  -1.073 0.283112
Line28      -0.007897   0.039361  -0.201 0.840982
Line29       0.138458   0.039361   3.518 0.000437 ***
Line50       0.059020   0.039361   1.499 0.133782
Line88      -0.081872   0.039361  -2.080 0.037547 *
Line108     -0.116158   0.039361  -2.951 0.003174 **
Line112      0.661864   0.039361  16.815  < 2e-16 ***
Line115      0.103540   0.039361   2.631 0.008538 **
Line117      0.075854   0.039361   1.927 0.053989 .
Line119     -0.052582   0.039361  -1.336 0.181610
Line123      0.046593   0.039361   1.184 0.236541
Line152     -0.065598   0.039361  -1.667 0.095631 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6098 on 10059 degrees of freedom
Multiple R-squared:  0.06429,       Adjusted R-squared:  0.06243
F-statistic: 34.56 on 20 and 10059 DF,  p-value: < 2.2e-16
```

Appendix B

Supplemental Figures and Tables for Chapter 3

Supplemental Table 3.1: Variants between TY-B ancestor and *S. paradoxus* 337 reference genome (see Methods for details).

Chr	Pos	Ref	Alt	GT	Type	Gene
IV	641630	GCAACAA	G,GCAA	2	Indel	YDR099W
VII	61782	T	TATC	1	SNM	YGL229C

Supplemental Table 3.2: Variants between TY+ ancestor and *S. paradoxus* 337 reference genome (see Methods for details).

Chr	Pos	Ref	Alt	GT	Type	Gene
IV	641630	GCAACAA	G,GCAA	2	Indel	YDR099W
VII	61782	T	TATC	1	SNM	YGL229C
VII	986079	G	A	1	SNM	YGR268C
XV	299062	G	T	1	SNM	YOL044W
XVI	186933	TAA	T,TA	2	SNM	

Supplemental Table 3.3: Indels mapping to LTRs in TY+ lines. Almost all of them are 136bp in length, but differ in sequence.

#CHROM	POS	REF	ALT	% match	name	description	length	indel size	matches TE locate?	Location
Spar_I_RaGOO	172829	T	TGATAATGAGAATTGGGTGAATGT TGAGATAATTGTTGGGATTCCATTG TTGATAAAGGCTATAATATTAGGT ATACAGAATATACTAGTAATGTAA ATACTAGTTAGTAGATGATAGTTG ATTTCTATTCCAACA	98%	YPRCdelta23	Ty1 LTR	338 bp	136		<500 bp upstream of tRNA gene
Spar_III_RaGOO	133181	T	TAAATGTGAGAATTGGGTGAATGT TGAGATAATTGTTGGGATTCCATTG TTGATAAAGGCTATAATATTAGGT ATACAGAATATACTAGTAATGTAA ATACTAGTTAGTAGATGATAGTTG ATTTCTATTCCAACA	98%	YLRCdelta7	Ty1 LTR	335 bp	136	Y	<200 bp upstream of snoRNA gene
Spar_IX_RaGOO	312016	T	TCACTATATGAGAATTGGGTGAAT GTTGAGATAATTGTTGGGATTCCAT TGTTGATAAAGGCTATAATATTAG GTATACAGAATATACTAGTAATGT AAATACTAGTTAGTAGATGATAGT TGATTTCTATTCCAA	98%	YCLWdelta15	Ty1 LTR	337 bp	136		In snoRNA gene
Spar_V_RaGOO	127186	G	GCTGGCTGAGAATTGGGTGAATGT TGAGATAATTGTTGGGATTCCATTG TTGATAAAGGCTATAATATTAGGT ATACAGAATATACTAGTAATGTAA ATACTAGTTAGTAGATGATAGTTG ATTTCTATTCCAACA	98%	YCRCdelta6	Ty1 LTR	332 bp	136	Y	Upstream of snoRNA gene
Spar_V_RaGOO	431571	T	TGTTCATGAGAATTGGGTGAATGT TGAGATAATTGTTGGGATTCCATTG TTGATAAAGGCTATAATATTAGGT ATACAGAATATACTAGTAATGTAA ATACTAGTTAGTAGATGATAGTTG ATTTCTATTCCAACA	98%	YORCdelta25	Ty1 LTR	332 bp	136		<500 bp upstream of tRNA gene
Spar_V_RaGOO	438466	T	TTCTTTTGAGAATTGGGTGAATGTT GAGATAATTGTTGGGATTCCATTGT TGATAAAGGCTATAATATTAGGTA TACAGAATATACTAGTAATGTAAA TACTAGTTAGTAGATGATAGTTGA TTTCTATTCCAACA	98%	YNLWdelta4	Ty1 LTR	334 bp	136		In YER137C

Spar_VI_RaGOO	197951	T	TATTACTGAGAATTGGGTGAATGT TGAGATAATTGTTGGGATTCCATTG TTGATAAAGGCTATAATATTAGGT ATACAGAATATACTAGTAATGTAA ATACTAGTTAGTAGATGATAGTTG ATTTCTATTCCAACA	97%	YPLWdelta4	Ty1 LTR	337 bp	136	Y	In snoRNA gene
Spar_VII_RaGOO	407618	C	CCTTAGTGAGAATTGGGTGAATGT TGAGATAATTGTTGGGATTCCATTG TTGATAAAGGCTATAATATTAGGT ATACAGAATATACTAGTAATGTAA ATACTAGTTAGTAGATGATAGTTG ATTTCTATTCCAACA	97%	YLRCdelta7	Ty1 LTR	335 bp	136	Y	In snoRNA gene
Spar_VII_RaGOO	839634	T	TCTAATTGAGAATTGGGTGAATGT TGAGATAATTGTTGGGATTCCATTG TTGATAAAGGCTATAATATTAGGT ATACAGAATATACTAGTAATGTAA ATACTAGTTAGTAGATGATAGTTG ATTTCTATTCCAACA	98%	YKRCdelta8	Ty1 LTR	332 bp	136		In snoRNA gene
Spar_VIII_RaGOO	442242	T	TCGAGTTGAGAATTGGGTGAATGT TGAGATAATTGTTGGGATTCCATTG TTGATAAAGGCTATAATATTAGGT ATACAGAATATACTAGTAATGTAA ATACTAGTTAGTAGATGATAGTTG ATTTCTATTCCAACA	98%	YNLCdelta1	Ty1 LTR	332 bp	136	Y	Upstream of tRNA
Spar_X_RaGOO	43017	C	CGATAATGTTGGAATAGAAATCAA CTATCATCTACTAACTAGTATTTAC ATTACTAGTATATTCTGTATACCTA ATATTATAGCCTTTATCAACAATGG AATCCCAACAATTATCTCAACATTC ACCCAATTCTCA	98%	YERCdelta20	Ty1 LTR	337 bp	136	Y	Intergenic
Spar_X_RaGOO	183598	G	GGTATCTGAGAATTGGGTGAATGT TGAGATAATTGTTGGGATTCCATTG TTGATAAAGGCTATAATATTAGGT ATACAGAATATACTAGTAATGTAA ATACTAGTTAGTAGATGATAGTTG ATTTCTATTCCAACA	98%	YCRCdelta6	Ty1 LTR	332 bp	136		Upstream of YJL112W
Spar_X_RaGOO	376694	T	TGGCCATGTTGGAATAGAAATCAA CTATCATCTACTAACTAGTATTTAC ATTACTAGTATATTCTGTATACCTA ATATTATAGCCTTTATCAACAATGG	98%	YERCdelta20	Ty1 LTR	337 bp	136	Y	Upstream of snoRNA

			AATCCCAACAATTATCTCAACATTC ACCCAATTCTCA							
Spar_X_RaGOO	395751	G	GGTTAATGTTGGAATAGAAATCAA CTATCATCTACTAACTAGTATTTAC ATTACTAGTATATTCTGTATACCTA ATATTATAGCCTTTATCAACAATGG AATCCCAACAATTATCTCAACATTC ACCCAATTCTCA	98%	YERCdelta20	Ty1 LTR	338 bp	136	Y	<300 bp downstream tRNA
Spar_XII_RaGOO	773415	T	TGAACGTGTTGGAATAGAAATCAA CTATCATCTACTAACTAGTATTTAC ATTACTAGTATATTCTGTATACCTA ATATTATAGCCTTTATCAACAATGG AATCCCAACAATTATCTCAACATTC ACCCAATTCTCA	98%	YERCdelta20	Ty1 LTR	337 bp	136	Y	Downstream snoRNA; upstream of tRNA
Spar_XII_RaGOO	856583	T	TATTATTTGTTGGAATAGAAATCA ACTATCATCTACTAACTAGTATTTA CATTACTAGTATATTCTGTATACCT AATATTATAGCCTTTATCAACAATG GAATCCCAACAATTATCTCAACAT TCACCCAATTCTC	98%	YERCdelta20	Ty1 LTR	337 bp	136	Y	Upstream of snoRNA
Spar_XIII_RaGOO	802251	T	TCTCCTTGAGAATTGGGTGAATGTT GAGATAATTGTTGGGATTCCATTGT TGATAAAGGCTATAATATTAGGTA TACAGAATATACTAGTAATGTAAA TACTAGTTAGTAGATGATAGTTGA TTTCTATTCCAACA	98%	YCLWdelta15	Ty1 LTR	337 bp	136		Downstream of snoRNA
Spar_XIV_RaGOO	543184	G	GGATCTTGAGAATTGGGTGAATGT TGAGATAATTGTTGGGATTCCATTG TTGATAAAGGCTATAATATTAGGT ATACAGAATATACTAGTAATGTAA ATACTAGTTAGTAGATGATAGTTG ATTTCTATTCCAACA	98%	YORCdelta25	Ty1 LTR	332 bp	136		Downstream of tRNA
Spar_XV_RaGOO	98450	T	TCACATATTGAGAATTGGGTGAAT GTTGAGATAATTGTTGGGATTCCAT TGTTGATAAAGGCTATAATATTAG GTATACAGAATATACTAGTAATGT AAATACTAGTTAGTAGATGATAGT TGATTTCTATTCCAA	98%	YKRCdelta8	Ty1 LTR	332 bp	136	Y	Downstream of snoRNA, upstream of tRNA

Spar_XV_RaGOO	674966	T	TGATTCTGTTGGAATAGAAATCAA CTATCATCTACTAACTAGTATTTAC ATTACTAGTATATTCTGTATACCTA ATATTATAGCCTTTATCAACAATGG AATCCCAACAATTATCTCAACATTC ACCCAATTCTCA	98%	YERCdelta20	Ty1 LTR	337 bp	136		Downstream of tRNA, upstream of snoRNA
Spar_VII_RaGOO	679651	T	TGATAATGTTGGAATAGAAATCAA CTATCATCTACTAACTAGTATTTAC ATTACTAGTATATTATCATATACGG TGTTAGAAGATGACGCAAATGATG AGAAATAGTCATCTAAATTAGTGG AAGCTGAAACGCAAGGATTGATAA TGTAATAGGATCAATGAATATAAA CATATAAAATGATGATAATAATAT TTATAGAATTGTGTAGAATTGCAG ATTCCCTTTTATGGATTCCTAAATC CTTGAGGAGAACTTCTAGTATATTC TGTATACCTAATATTATAGCCTTTA TCAACAATGGAATCCCAACAATTA TCTCAACATTCACCCAATTCTCA	99%	YLRWdelta11	Ty1 LTR	334 bp	340	Y	Downstream of tRNA

Supplemental Table 3.4: Genes that multinucleotide mutations in 1-copy strain either lie within or are upstream of. Three of these are transcribed by RNA polymerase III (highlighted in yellow).

GENE ID	Common Name	Description	Additional Info
YAR050W	FLO1	FLO1 lectin-like protein involved in flocculation; cell wall protein	
YAL062W	GDH3	GDH3; NADP+ dependent glutamate hydrogenase; synthesizes glutamate from ammonia and alpha-ketoglutarate	
YBL014C	RRN6	RRN6; component of the core factor rDNA transcription factor complex; required for transcription of 35S rRNA genes by RNApol I	
YNL339C	YRF1-6	helicase encoded by the Y' element of subtelomeric regions	
YBL111C	YBL111C	helicase-like protein encoded within the telomeric Y' element	transcribed by RNA pol III
YHL049C	YHL049C	putative protein of unknown function	
YCR019W	MAK32	protein necessary for stability of L-A dsRNA-containing particles; maintenance of Killer	
YCR020W	HTL1	component of the RSC chromatin remodeling complex	transcribed by RNA pol III

YDL160C	DHH1	DHH1; cytoplasmic DEAD-box helicase, stimulates mRNA decapping	transcribed by RNA pol III
YDR151C	CTH1	CTH1; member of the CCCH zinc finger family; activates transcription and has a role in mRNA degradation	
YHR146W	CRP1	CRP1; binds to cruciform DNA structures	
YOL159C-A	YOL159C-A	protein of unknown function; overexpression affects endocytic protein trafficking	
YBR301W	PAU24	cell wall mannoprotein; member of the seripauperin multigene family encoded mainly in subtelomeric regions;completely repressed in aerobic conditions	
YLR466W	YFR1-4	helicase encoded by the Y' element of subtelomeric regions, highly expressed in mutants lacking telomerase component TLC1	
YPR204W	YPR204W	DNA helicases encoded within the telomeric Y' element	
YHR219W	YHR219W	putative protein of unknown function similarity to helicases, located in the telomere region on the right arm of chromosome VIII	

YFL068W	YFL068W	putative protein of unknown function
YML080W	DUS1	dihydrouridine synthase; member of a widespread family of conserved proteins; modifies pre-tRNA
YML079W	YML079W	non-essential protein of unknown function
YMR250W	GATY+	glutamate decarboxylase; converts glutamate into gamma-aminobutylic acid during glutamate catabolism; involved in response to oxidative stress
YNR002C	ATO2	putative transmembrane protein involved in export of ammonia;

Supplemental Table 3.5: Transposition events in TY+ samples. Data was generated using the McClintock pipeline and TELocate data (performed by Jingxuan Chen). Table includes locations and genes that are either upstream or downstream of the insertion or are the locations into which the Ty1 element inserted. There were 86 Ty1 insertions from TELocate total.

Chrom	Pos	Strain	Sample	Type	Gene	RNA pol III?
Spar_I_RaGOO	136357	TY+	1	< 2kb upstream of	YLL024C	transcribed by RNA pol III
Spar_I_RaGOO	175059	TY+	43	intergenic		
Spar_II_RaGOO	625824	TY+	17	intergenic		
Spar_II_RaGOO	625871	TY+	23	intergenic		
Spar_II_RaGOO	377039	TY+	36	< 2kb upstream of	YBR079C	transcribed by RNA pol III
Spar_III_RaGOO	99695	TY+	20	< 2kb upstream of	YCL018W	transcribed by RNA pol III
Spar_III_RaGOO	132856	TY+	46	intergenic		
Spar_IV_RaGOO	958941	TY+	5	intergenic		
Spar_IV_RaGOO	678366	TY+	7	< 2kb upstream of	YDR119W-A	
Spar_IV_RaGOO	791189	TY+	9	intergenic		
Spar_IV_RaGOO	1211595	TY+	12	< 2kb downstream of	YDR390C	
Spar_IV_RaGOO	960366	TY+	15	intergenic		
Spar_IV_RaGOO	90240	TY+	20	< 2kb upstream of	YDL210W	transcribed by RNA pol III
Spar_IV_RaGOO	111686	TY+	27	< 2kb upstream of	YDL197C	transcribed by RNA pol III

Spar_IV_RaGOO	861506	TY+	33	< 2kb upstream of	YDR211W	
Spar_IV_RaGOO	958946	TY+	33	intergenic		
Spar_IV_RaGOO	1321147	TY+	35	in	YDR453C	transcribed by RNA pol III
Spar_V_RaGOO	435434	TY+	9	intergenic		
Spar_V_RaGOO	127022	TY+	13	< 2kb downstream of	YEL013W	
Spar_V_RaGOO	477540	TY+	13	in	YER157W	transcribed by RNA pol III
Spar_V_RaGOO	310145	TY+	20	< 2kb upstream of	YER077C	
Spar_V_RaGOO	113586	TY+	27	< 2kb downstream of	YEL020W-A	
Spar_V_RaGOO	428907	TY+	36	< 2kb upstream of	YER133W	transcribed by RNA pol III
Spar_V_RaGOO	431913	TY+	37	intergenic		
Spar_V_RaGOO	431981	TY+	41	intergenic		
Spar_V_RaGOO	127054	TY+	42	< 2kb downstream of	YEL013W	
Spar_VI_RaGOO	255801	TY+	6	< 2kb upstream of	YFR036W	transcribed by RNA pol III
Spar_VI_RaGOO	255689	TY+	6	< 2kb upstream of	YFR036W	transcribed by RNA pol III

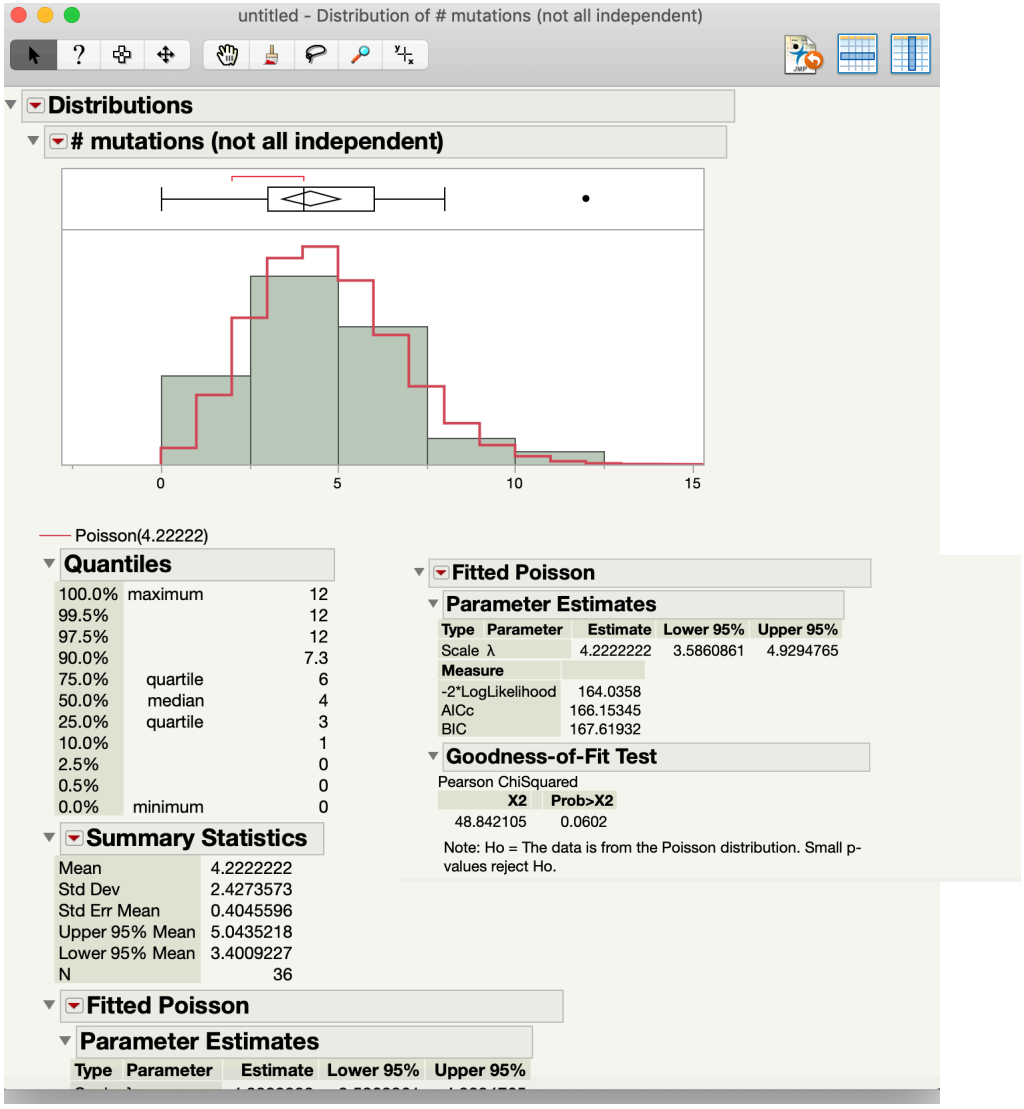
Spar_VI_RaGOO	137869	TY+	42	intergenic		
Spar_VI_RaGOO	197562	TY+	46	in	YFR011C	transcribed by RNA pol III
Spar_VII_RaGOO	782295	TY+	2	intergenic		
Spar_VII_RaGOO	406875	TY+	14	in	YGL047W	
Spar_VII_RaGOO	679055	TY+	17	intergenic		
Spar_VII_RaGOO	685143	TY+	23	< 2kb downstream of	YGR109C	
Spar_VII_RaGOO	920183	TY+	29	in	YGR234W	transcribed by RNA pol III
Spar_VII_RaGOO	839353	TY+	31	intergenic		
Spar_VII_RaGOO	435346	TY+	33	intergenic		
Spar_VII_RaGOO	837042	TY+	36	< 2kb upstream of	YGR189C	transcribed by RNA pol III
Spar_VII_RaGOO	746036	TY+	42	intergenic		
Spar_VIII_RaGOO	98130	TY+	9	intergenic		
Spar_VIII_RaGOO	441829	TY+	19	intergenic		
Spar_VIII_RaGOO	441767	TY+	35	intergenic		
Spar_VIII_RaGOO	441846	TY+	37	intergenic		
Spar_X_RaGOO	376343	TY+	5	intergenic		
Spar_X_RaGOO	396575	TY+	13	< 2kb upstream of	YJL010C	transcribed by RNA pol III
Spar_X_RaGOO	334656	TY+	18	< 2kb upstream of	YJL039C	transcribed by RNA pol III
Spar_X_RaGOO	396338	TY+	23	< 2kb	YJL010C	transcribed by RNA pol

				upstream of		III
Spar_X_RaGOO	395204	TY+	29	intergenic		
Spar_X_RaGOO	334893	TY+	32	< 2kb upstream of	YJL039C	transcribed by RNA pol III
Spar_X_RaGOO	304012	TY+	33	in	YJL059W	
Spar_X_RaGOO	43018	TY+	46	intergenic		
Spar_XI_RaGOO	52190	TY+	12	intergenic		
Spar_XI_RaGOO	137441	TY+	41	in	YKL168C	
Spar_XI_RaGOO	463667	TY+	19	intergenic		
Spar_XI_RaGOO	494690	TY+	33	intergenic		
Spar_XII_RaGOO	361180	TY+	33	upstream of	YLR108C	
Spar_XII_RaGOO	616233	TY+	33	in	YLR211C	
Spar_XII_RaGOO	624392	TY+	17	in	YLR216C	
Spar_XII_RaGOO	682846	TY+	36	in	YLR248W	
Spar_XII_RaGOO	686070	TY+	31	in	YLR249W	transcribed by RNA pol III
Spar_XII_RaGOO	770435	TY+	3	upstream of	YLR301W	
Spar_XII_RaGOO	831625	TY+	24	intergenic		
Spar_XII_RaGOO	831796	TY+	36	intergenic		
Spar_XII_RaGOO	973846	TY+	27	in	YLR409C	transcribed by RNA pol III
Spar_XII_RaGOO	1008835	TY+	5	intergenic		
Spar_XII_RaGOO	1034569	TY+	17	in	YLR435W	
Spar_XIII_RaGOO	6731	TY+	46	intergenic		
Spar_XIII_RaGOO	202127	TY+	29	intergenic		

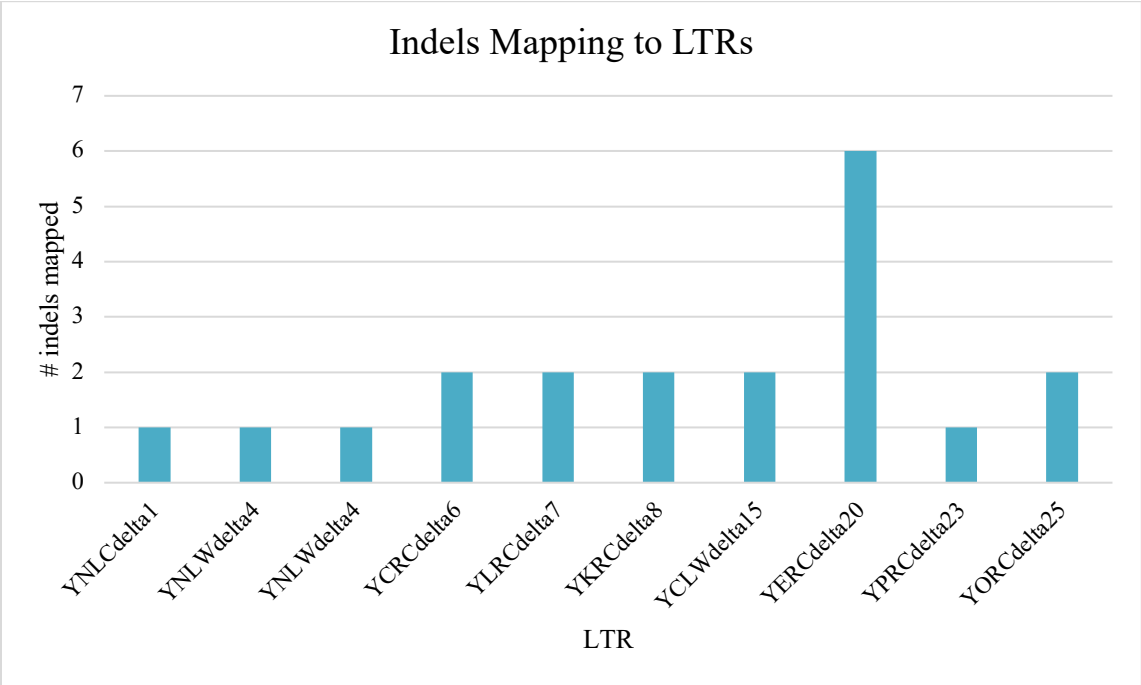
Spar_XIII_RaGOO	368105	TY+	17	intergenic		
Spar_XIII_RaGOO	410625	TY+	37	intergenic		
Spar_XIII_RaGOO	577574	TY+	30	upstream of	YMR164C	
Spar_XIII_RaGOO	577725	TY+	11	upstream of	YMR164C	
Spar_XIII_RaGOO	803267	TY+	17	upstream of	YMR272C	transcribed by RNA pol III
Spar_XIV_RaGOO	88271	TY+	3	in	YNL286W	
Spar_XIV_RaGOO	256913	TY+	12	in	YNL197C	
Spar_XIV_RaGOO	314827	TY+	43	upstream of	YNL163C	transcribed by RNA pol III
Spar_XIV_RaGOO	608022	TY+	19	intergenic		
Spar_XIV_RaGOO	608271	TY+	7	intergenic		
Spar_XIV_RaGOO	633563	TY+	41	in	YNR016C	transcribed by RNA pol III
Spar_XV_RaGOO	50906	TY+	29	in	YOL138C	
Spar_XV_RaGOO	98231	TY+	42	intergenic		
Spar_XV_RaGOO	98367	TY+	36	intergenic		
Spar_XV_RaGOO	571785	TY+	6	intergenic		
Spar_XV_RaGOO	571887	TY+	6	intergenic		
Spar_XVI_RaGOO	833497	TY+	46	in	YPR159C-A	
Spar_XVI_RaGOO	854414	TY+	28	upstream of	YPR170W-B	
Spar_XVI_RaGOO	854867	TY+	33	upstream of	YPR170W-B	

Supplemental Table 3.6: Locations of indels in 1-copy strain. 5 out of the 15 map in or near genes, and of these 5, 3 are transcribed by RNA pol III.

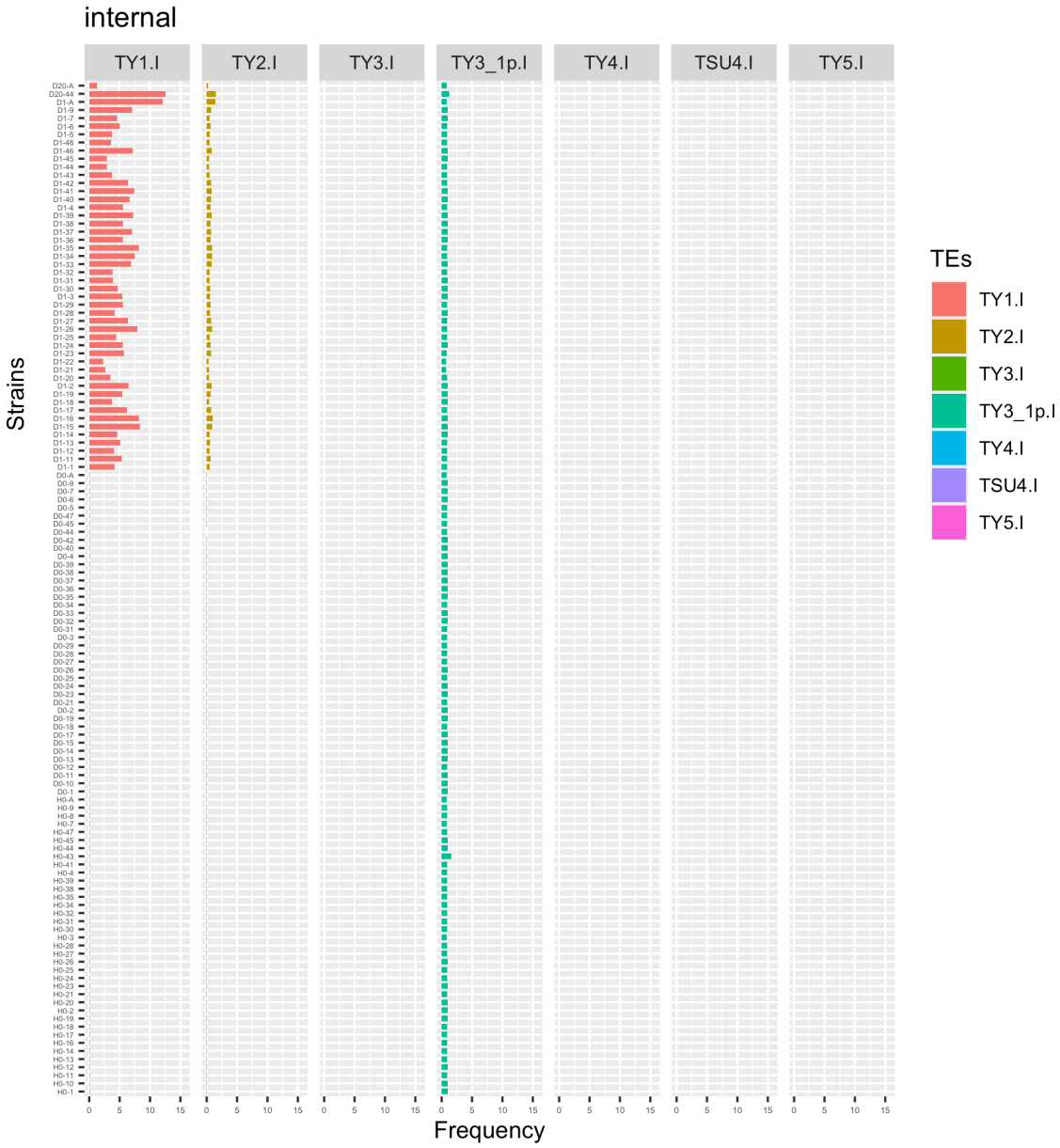
#CHROM	POS	REF	ALT	Near Genes?	
Spar_I_RaGOO	182697	GCATTTCATTTATATTG TCATTTCATTTATATTGT	G	intergenic	
Spar_II_RaGOO	146031	CGTT	C	in YBL035C	transcribed by RNA pol III
Spar_III_RaGOO	40092	AGAAG	A	upstream of YCL057W	
Spar_IV_RaGOO	746765	T	TG	intergenic	
Spar_IV_RaGOO	1048011	GA	G	in YDR310C	transcribed by RNA pol III
Spar_IX_RaGOO	430080	GA	G	intergenic	
Spar_VII_RaGOO	706669	CT	C	in YGR122W	
Spar_VII_RaGOO	1063702	G	GGTGTGTGGT	intergenic	
Spar_X_RaGOO	357921	C	CTGAGA	intergenic	
Spar_XIII_RaGOO	939823	G	GTTAGATTTGTTTACA	intergenic	
Spar_XIV_RaGOO	90041	C	CAATTATCTCAACATTCACC CAATTCTCA	intergenic	
Spar_XIV_RaGOO	415201	C	CT	intergenic	
Spar_XIV_RaGOO	608826	T	TTGAGAATTGGGTGAATGT TGAGATAATTGTTGGGATT CCATTGTTGATA	intergenic	
Spar_XV_RaGOO	74074	GA	G	intergenic	
Spar_XV_RaGOO	1013351	GC	G	in YOR378W	transcribed by RNA pol III



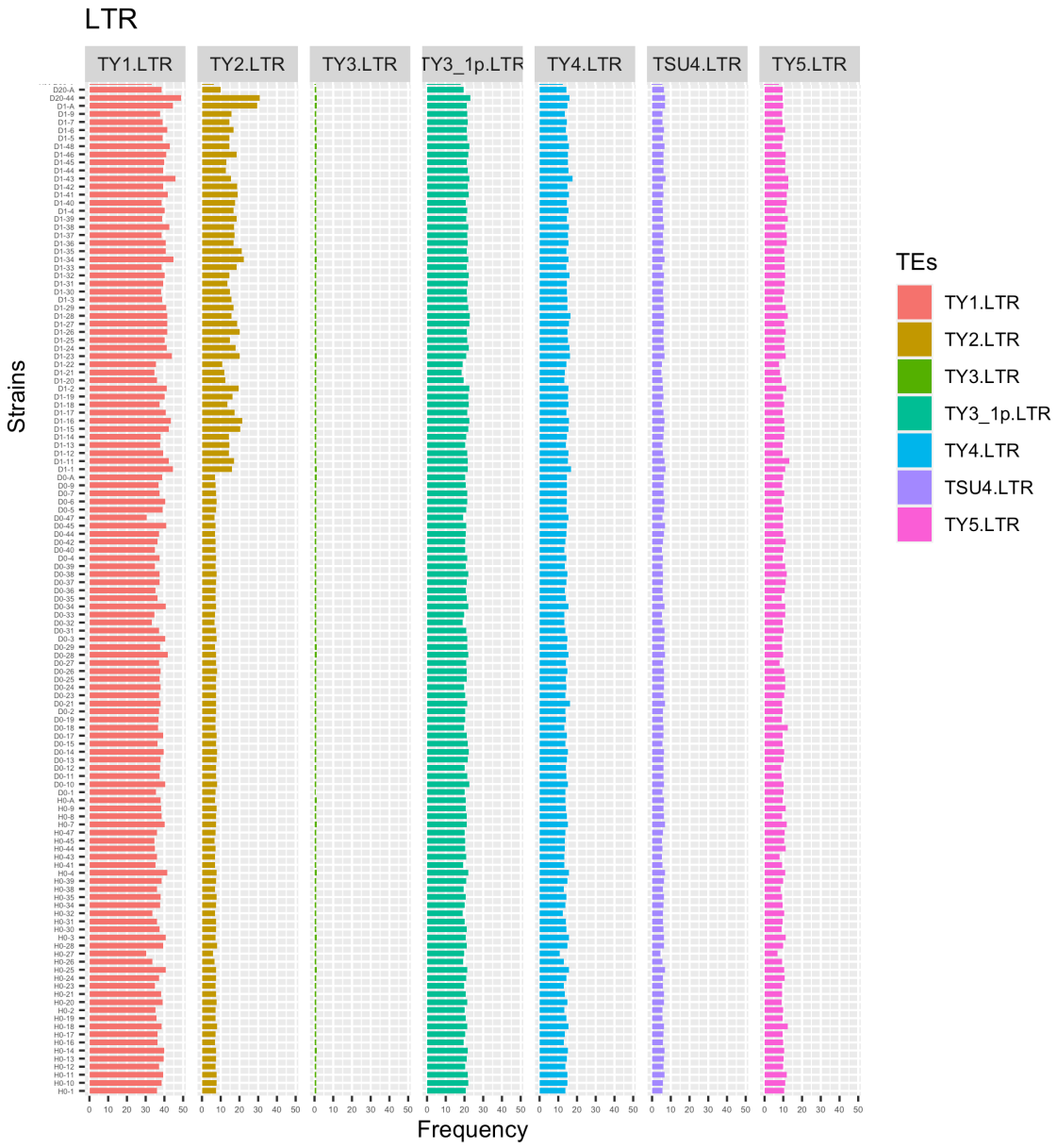
Supplemental Figure 3.1: Average number of mutations per line in TY-A samples. This distribution is not significantly different from a Poisson distribution ($p>0.05$).



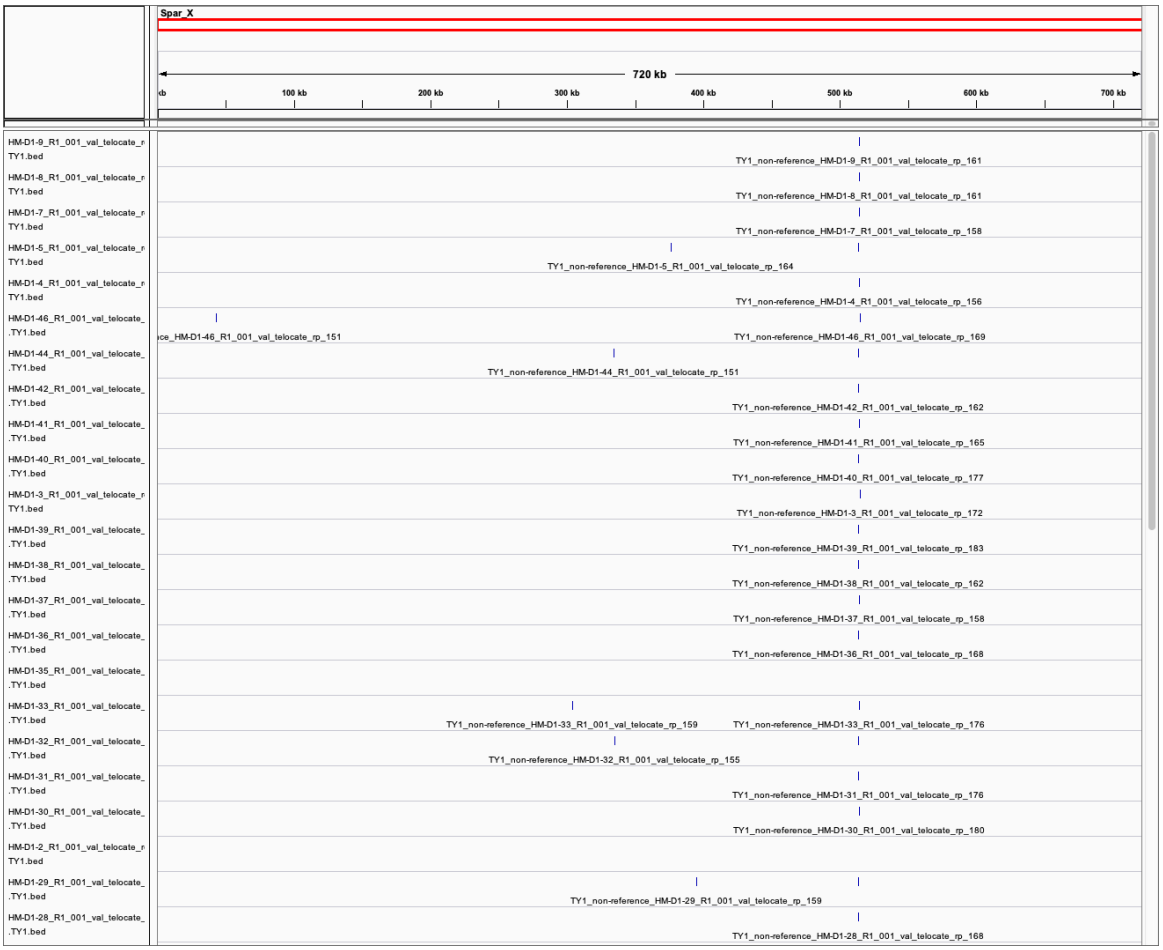
Supplemental Figure 3.2: Number of indels in TY+ lines that map to different LTRs.

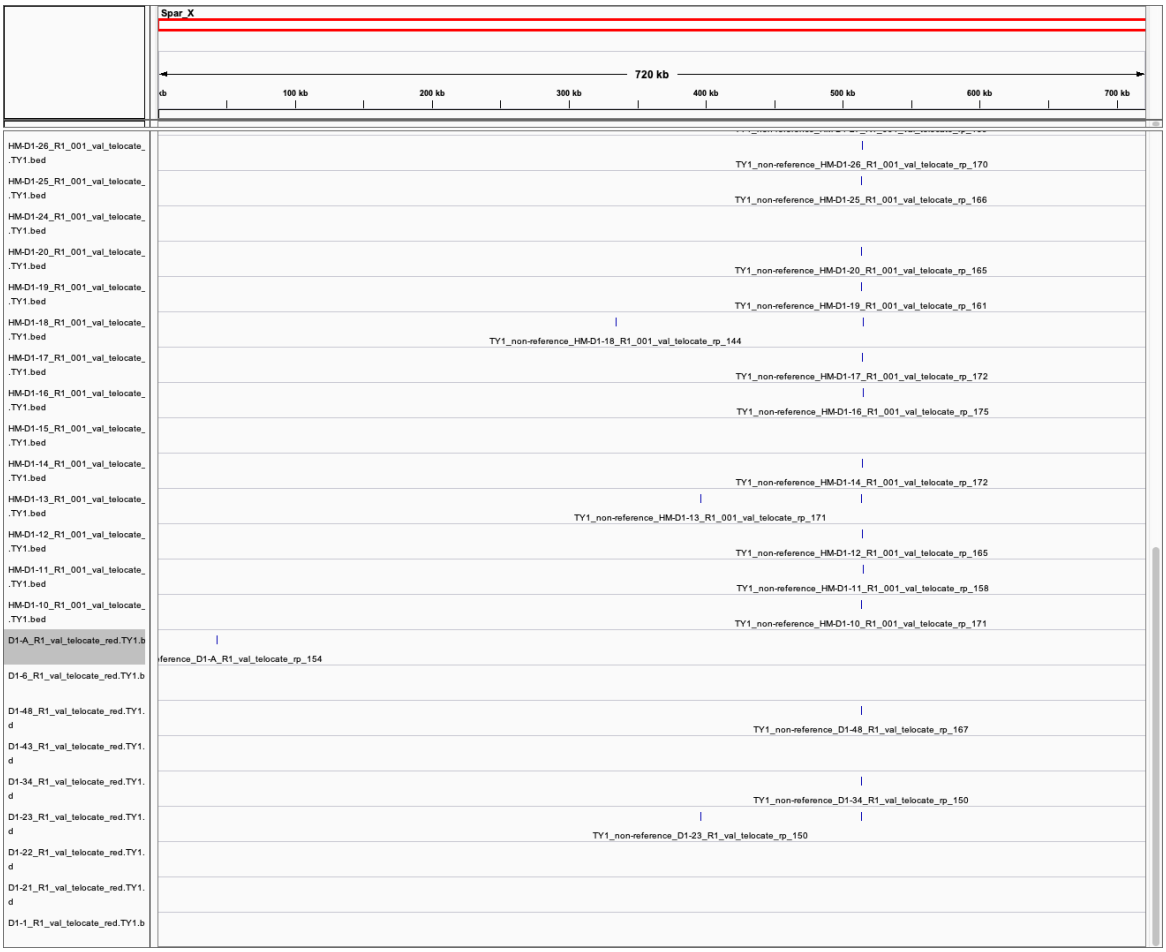


Supplemental Figure 3.3: Ty1 copy numbers in strains TY-A, TY-B, and TY+. Barplots produced using the R package ggplot2. See methods for details and locations of scripts. Line 33 from the TY-A ancestor (HM-H0-33), which should have zero transposable elements, was clearly mislabeled as it has two elements and it was thus removed from all analyses.

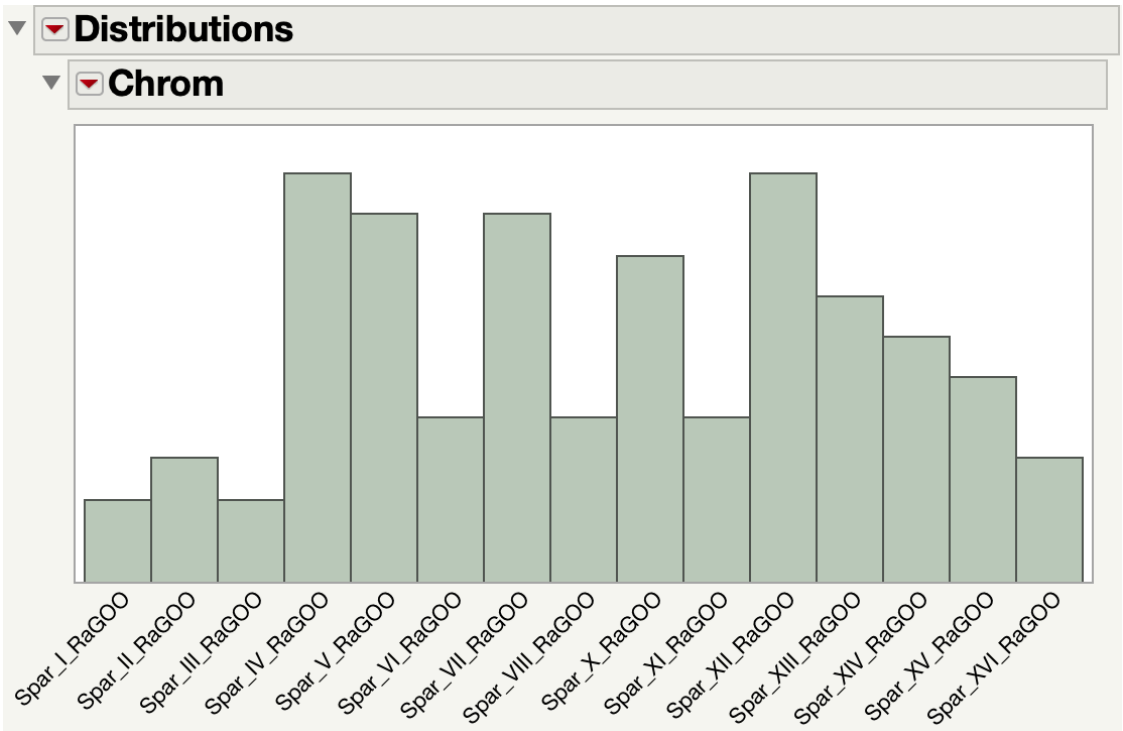


Supplemental Figure 3.4: Barplots for LTR copy numbers (not including LTRs found in full-length elements) in each strain.


Spar_X
720 kb
kb
100 kb
200 kb
300 kb
400 kb
500 kb
600 kb
700 kb
HM-D1-9_R1_001_val_telocate_r
TY1.bed
TY1_non-reference_HM-D1-9_R1_001_val_telocate_rp_161
HM-D1-8_R1_001_val_telocate_r
TY1.bed
TY1_non-reference_HM-D1-8_R1_001_val_telocate_rp_161
HM-D1-7_R1_001_val_telocate_r
TY1.bed
TY1_non-reference_HM-D1-7_R1_001_val_telocate_rp_158
HM-D1-5_R1_001_val_telocate_r
TY1.bed
TY1_non-reference_HM-D1-5_R1_001_val_telocate_rp_164
HM-D1-4_R1_001_val_telocate_r
TY1.bed
TY1_non-reference_HM-D1-4_R1_001_val_telocate_rp_156
HM-D1-46_R1_001_val_telocate_
.TY1.bed
ce_HM-D1-46_R1_001_val_telocate_rp_151
TY1_non-reference_HM-D1-46_R1_001_val_telocate_rp_169
HM-D1-44_R1_001_val_telocate_
.TY1.bed
TY1_non-reference_HM-D1-44_R1_001_val_telocate_rp_151
HM-D1-42_R1_001_val_telocate_
.TY1.bed
TY1_non-reference_HM-D1-42_R1_001_val_telocate_rp_162
HM-D1-41_R1_001_val_telocate_
.TY1.bed
TY1_non-reference_HM-D1-41_R1_001_val_telocate_rp_165
HM-D1-40_R1_001_val_telocate_
.TY1.bed
TY1_non-reference_HM-D1-40_R1_001_val_telocate_rp_177
HM-D1-3_R1_001_val_telocate_r
TY1.bed
TY1_non-reference_HM-D1-3_R1_001_val_telocate_rp_172
HM-D1-39_R1_001_val_telocate_
.TY1.bed
TY1_non-reference_HM-D1-39_R1_001_val_telocate_rp_183
HM-D1-38_R1_001_val_telocate_
.TY1.bed
TY1_non-reference_HM-D1-38_R1_001_val_telocate_rp_162
HM-D1-37_R1_001_val_telocate_
.TY1.bed
TY1_non-reference_HM-D1-37_R1_001_val_telocate_rp_158
HM-D1-36_R1_001_val_telocate_
.TY1.bed
TY1_non-reference_HM-D1-36_R1_001_val_telocate_rp_168
HM-D1-35_R1_001_val_telocate_
.TY1.bed
HM-D1-33_R1_001_val_telocate_
.TY1.bed
TY1_non-reference_HM-D1-33_R1_001_val_telocate_rp_159
TY1_non-reference_HM-D1-33_R1_001_val_telocate_rp_176
HM-D1-32_R1_001_val_telocate_
.TY1.bed
TY1_non-reference_HM-D1-32_R1_001_val_telocate_rp_155
HM-D1-31_R1_001_val_telocate_
.TY1.bed
TY1_non-reference_HM-D1-31_R1_001_val_telocate_rp_176
HM-D1-30_R1_001_val_telocate_
.TY1.bed
TY1_non-reference_HM-D1-30_R1_001_val_telocate_rp_180
HM-D1-2_R1_001_val_telocate_r
TY1.bed
HM-D1-29_R1_001_val_telocate_
.TY1.bed
TY1_non-reference_HM-D1-29_R1_001_val_telocate_rp_159
HM-D1-28_R1_001_val_telocate_
.TY1.bed
TY1_non-reference_HM-D1-28_R1_001_val_telocate_rp_168



Supplemental Figure 3.5: Visualization of ancestral Ty1 locations in TY+ samples using the Integrative Genomics Viewer (IGV). See methods for details.

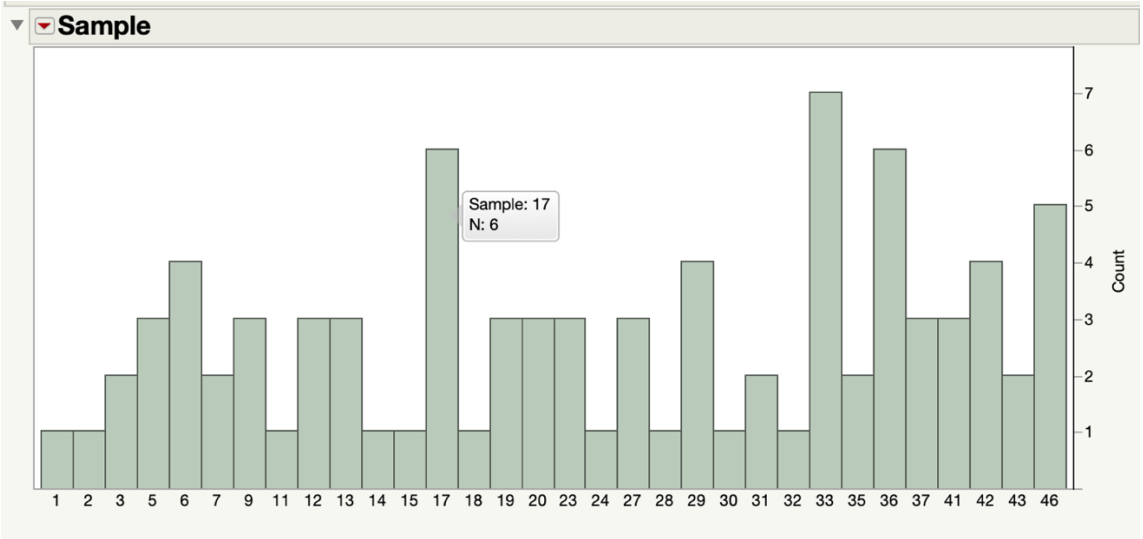


Frequencies

Level	Count	Prob
Spar_I_RaGOO	2	0.02326
Spar_II_RaGOO	3	0.03488
Spar_III_RaGOO	2	0.02326
Spar_IV_RaGOO	10	0.11628
Spar_V_RaGOO	9	0.10465
Spar_VI_RaGOO	4	0.04651
Spar_VII_RaGOO	9	0.10465
Spar_VIII_RaGOO	4	0.04651
Spar_X_RaGOO	8	0.09302
Spar_XI_RaGOO	4	0.04651
Spar_XII_RaGOO	10	0.11628
Spar_XIII_RaGOO	7	0.08140
Spar_XIV_RaGOO	6	0.06977
Spar_XV_RaGOO	5	0.05814
Spar_XVI_RaGOO	3	0.03488
Total	86	1.00000

N Missing 0
15 Levels

Supplemental Figure 3.6: Distribution of Ty1 insertions across chromosomes. There were no Ty1 insertions found on chromosome IX.



Supplemental Figure 3.7: Distribution of Ty1 insertions across TY+ lines. Line numbers are on the x-axis and counts of Ty1 insertions are on the y-axis. The average number of Ty1 insertions per line is 2.4 and the median number of insertions per line is 2, with the maximum number of insertions in one line being 7. Note: lines 8, 10, 16, 21, 22, 25, 26, 34, 38, 39, 44, 45, and 48 are missing from the analysis due to cross-line contamination.