EVOLUTIONARY DYNAMICS OF GENOME ARCHITECTURE

by

MEGAN G. BEHRINGER

(Under the Direction of DAVID W. HALL)

ABSTRACT

With sequencing technology becoming more advanced and widely available, so has the ability to describe the forces and processes shaping genome architecture. Spontaneous mutations such as nucleotide substitutions, indels, duplications, large-scale insertions and deletions, inversions, and translocations provide the input that is then acted on by adaptive (natural selection) and non-adaptive (recombination, random genetic drift) processes to create the genetic variation within and among species. Through this dissertation, I use a combination of experimental evolution and comparative genomics to examine these four forces and how they act to create and maintain the genome. First, through two mutation accumulation experiments, I investigate the rates, biases and spectra of mutation and recombination in the absence of selection. Lastly, I discuss the effect of genome wide selection on premature termination codons in introns across seven model organisms.

INDEX WORDS: Mutation, Recombination, Gene Conversion, Nonsense Mediated Decay

EVOLUTIONARY DYNAMICS OF GENOME ARCHITECTURE

by

MEGAN G. BEHRINGER

BS, Auburn University, 2009

MS, Auburn University, 2011

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

2015

© 2015

Megan G. Behringer

All Rights Reserved

EVOLUTIONARY DYNAMICS OF GENOME ARCHITECTURE

by

MEGAN G BEHRINGER

Major Professor:	David Hall
Committee:	Jessica Kissinger
	Michael McEachern
	Jan Mrazek
	Paul Scheilkleman

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
May 2015

ACKNOWLEDGEMENTS

Throughout the process of writing this dissertation I have encountered so many amazing and supportive individuals that I would like to take this time to thank. First I'd like to thank my advisor, Dave Hall, for creating a lab atmosphere that I was excited about coming to work at everyday: without being able to daily talk out my ideas with him I would still be stuck on chapter 1. My beautiful and wonderful lab-mate Sarah Sander, for always being there for encouragement when I need it and for being so welcoming when I joined the lab. My awesome undergrads and technicians: Moon-tae Kim, Cameron Story, John Chamberlin, Justin Vinomon, Katherine Korones, and Ian Milton for helping me with all aspects of the mutation accumulation projects so that I could spend more time on the computer and analyzing data.

I would like to thank my committee: Jessie, Mike, Jan, and Paul, for taking the time to write qualifying exams for me, come to my committee meetings and even read this dissertation. I appreciate all the advice and conversation you all gave for my projects giving me an outside eye and helping me think more deeply. I'd also like to thank the evolution community at the University of Georgia, members of the EvolGen reading group and everyone who ever attended one of my EDGE talks, for all the support, ideas, discussion, making UGA such a great place to conduct my doctoral research.

Mostly, I'd like to thank my family: my parents, David and Kathy, and my brother Alex, for being so supportive through the years, and understanding when I explained to you I was going to attend college for 10 years. You guys made me who I am and I love you all dearly. Finally, my wonderful husband Blane for loving me and keeping me together while I attempted my Ph.D., and for spreading plates on Saturday when I know you'd much rather be eating lunch.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION TO THESIS

With sequencing technology becoming more advanced and widely available, so has the ability to describe the forces and processes shaping genome architecture. Spontaneous mutations such as nucleotide substitutions, indels, duplications, large-scale insertions and deletions, inversions, and translocations provide the input that is then acted on by adaptive (natural selection) and non-adaptive (recombination, random genetic drift) processes to create the genetic variation within and among species. Changes in genome architecture have direct implications for adaptation and speciation. While theoretical approaches have provided methods to estimate rates of mutation (1-5), high-throughput sequencing techniques coupled with experimental evolution can provide direct quantification and assessment of mutations at the genomic level. These same high-throughput sequencing techniques can be applied to investigate loss of heterozygosity as a result of recombination. Additionally, experiments examining the rates and biases of mutation and recombination can be performed in the absence of selection so that only random genetic drift is operating. Throughout this introduction, I will briefly discuss the four major evolutionary forces acting on genome architecture and how they apply to the presented research.

Mutation

Mutation is the ultimate source of genetic variation, specifically; mutation is the permanent change of a nucleotide sequence. It is a random process and occurs independent of an organism's adaptive needs. In reference to genome architecture, mutations come in many flavors. At the nucleotide sequence level, mutations include: nucleotide substitutions, small-scale insertions or deletions (indels), medium to large-scale insertions caused by horizontal gene transfers, transposable elements, duplications, and medium to large-scale deletions. Mutations on the chromosome level may include: inversions, translocations, chromosomal fusions/fissions, or whole chromosome deletions and duplications. Finally, at the genomic level, whole genome duplications may occur.

Each of these mutations occur at different rates and have different biases as to the location and direction of change (6-8). Examples include transition bias (9) (or as in the case of the grasshopper species *Podisma pedestris*, lack thereof (10)), indel bias (11), retrotransposon insertion bias (12), deletion bias (13), and inversion bias (14, 15). In addition to rate and directional bias, there may also be locational bias or "hotspots". Hotspots may be more susceptible to mutation due to nucleotide modifications (CpG dinucleotides in many eukaryotes), strand slippage (simple nucleotide repeat regions), or double helix instability (complex repeat regions; regions with high transcription rates). These mutation rates and biases directly contribute to the diverse genomic landscape that breeds evolutionary change.

Experiments estimating genome-wide mutation rates, spectrum and biases have only been performed in a handful of organisms: *Saccharomyces cerevisiae (16, 17)*, *Caenorhabditis elegans (18)*, *Arabidopsis thaliana (19)*, *Dictyostelium dicsoideum (20)*, *Escherichia coli (21)*,

Paramecium tetraurelia (22), *Chlamydomonas reinhardtii* (23), *Daphnia pulex (24)*, *Tetrahymena thermophila (25)*, and *Drosophila melanogaster* (26).

Recombination

The exchange of genetic information within the genome is referred to as recombination. Namely the result of double-stranded break repair, recombination is a common occurrence in genome dynamics: through the repair of double-stranded breaks either programmed in meiosis, due to DNA damage during mitotic replication, or stress occurring at any thing throughout the cell cycle.

Recombination has many roles in genome organization and architecture. It is able to both increase and decrease genetic diversity. Recombination can act to break up alleles that are in linkage disequilibrium, which increases genetic diversity and allows the alleles to respond independently to natural selection. This is incredibly important in the light of Muller's ratchet where deleterious alleles accumulate due to lack of recombination (27). However, recombination can also decrease genotypic diversity through gene conversion. Gene conversion is a process where one allele replaces another during molecular recombination. When it occurs between nonsister, homologous chromatids it results in a loss of heterozygosity.

These events are not always random and result in many direct effects on genome architecture including: G/C-biased gene conversion (28), hotspot drive (29), indel drive (30), and mutagenic recombination (31). Theoretical approaches to understanding the effects of biased gene conversion in both large populations and finite populations have been conducted (32). It has been shown that gene conversion not only drastically decreases genetic variation, but also it must occur frequently to contribute to divergence of isolated populations. In fungi, gene conversion is reported to occur frequently, with the estimated rate to be between 0 and 0.5 /locus /generation,

with the majority of the rates being between 0.002 and 0.1/locus /generation (33). Much like mutation, the ability to characterize recombination rates and the frequency of gene conversion bias allows us to better understand the origin and maintenance of diversity amongst genomes.

Natural Selection

The process by which specific variants become more or less common in the population as a result of their effect on fitness is natural selection. Natural selection is an adaptive force, acting on the genetic variation created by mutation and recombination. It occurs on the phenotypic level and is unable to distinguish between variants as long as their effect on fitness is the same. Within protein coding regions, natural selection may act upon nucleotide substitutions causing obvious phenotypic changes such as non-synonymous substitutions resulting in differences in protein activity. However, in the context of genome architecture, natural selection has been implicated in other specific phenomena such as: codon bias (34, 35), head to head gene arrangement (36), and other genome streamlining processes (37-39) However, the strength of selective force in the general shaping within and among genomes is largely considered weak consistently out competed by random genetic drift except in large populations (40, 41).

Random Genetic Drift

The final force responsible for the shaping of genome architecture is random genetic drift. Genetic drift is the stochastic fluctuation in frequency of variants due to sampling. Acting in conflict with natural selection, the effects of drift are greater at smaller effective population sizes (42). In order for natural selection to overcome genetic drift in a diploid organism, the strength of selection must be greater than $1/(nN_e)$, where n is the ploidy of the genome (n = 1 in haploids and 2 in diploids) and N_e is the effective population size. As a result, genomes with smaller effective population sizes are expected to harbor more slightly deleterious mutations

leading to effects such as "genome bloat", manifesting in the form of expanded UTRs, intragenic regions (43-45), and additional introns (46). The accumulations of these mutations are not always terminally deleterious; they can provide the variation from which *de novo* genes and alternative splicing might evolve. Additionally, since many genomic features such as genome size, codon bias, gene number, intron size, transposable element content are under weak selection (47), we can observe a gradient across the size and/or degree of these features where random genetic drift outcompetes natural selection as effective population size decreases.

Through this dissertation I will present three studies that examine the role of the four major evolutionary forces affecting genome architecture. The first two studies investigate mutation and recombination's role in maintaining the genome, in the absence of selection, directed only by random genetic drift. In chapter 2, I present the first study of genome-wide parameters of mutation in *S. pombe*, a species that is about 600 million years diverged from *S. cerevisiae*, approximately the same amount of divergence as *Drosophila* and humans. In chapter 3, I present the first unbiased, genome-wide study of mitotic gene conversion in *S. cerevisiae*, which elucidates the rates and biases of gene conversion by examining loss of heterozygosity due to mitotic recombination. Lastly, to describe a specific example of genome-wide selection, in chapter 4 I present a comparative study of genome-wide selection on in-frame premature termination codons within introns across seven major model organisms.

Literature Cited

1. Bateman A (1959) The viability of near-normal irradiated chromosomes. *International Journal of Radiation Biology* 1(2):170-180.
2. Mukai T (1964) The genetic structure of natural populations of Drosophila melanogaster. I. Spontaneous mutation rate of polygenes controlling viability. *Genetics* 50(1):1.
3. Keightley PD (1994) The distribution of mutation effects on viability in Drosophila melanogaster. *Genetics* 138(4):1315-1322.
4. García-Dorado A (1997) The rate and effects distribution of viability mutation in Drosophila: minimum distance estimation. *Evolution*:1130-1139.
5. Shaw FH, Geyer CJ, & Shaw RG (2002) A comprehensive model of mutations affecting fitness and inferences for Arabidopsis thaliana. *Evolution* 56(3):453-463.
6. Baer CF, Miyamoto MM, & Denver DR (2007) Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nature Reviews Genetics* 8(8):619-631.
7. Lynch M (2010) Evolution of the mutation rate. *Trends in Genetics* 26(8):345-352.
8. Hodgkinson A & Eyre-Walker A (2011) Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics* 12(11):756-766.
9. Wakeley J (1996) The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends in ecology & evolution* 11(4):158-162.
10. Keller I, Bensasson D, & Nichols RA (2007) Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. *PLoS genetics* 3(2):e22.
11. Gregory TR (2003) Is small indel bias a determinant of genome size? *TRENDS in Genetics* 19(9):485-488.
12. Pereira V (2004) Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome. *Genome biology* 5(10):R79.
13. Mira A, Ochman H, & Moran NA (2001) Deletional bias and the evolution of bacterial genomes. *Trends in Genetics* 17(10):589-596.
14. Stenger JE*, et al.* (2001) Biased distribution of inverted and direct Alus in the human genome: implications for insertion, exclusion, and genome stability. *Genome research* 11(1):12-27.
15. Achaz G, Coissac E, Netter P, & Rocha EP (2003) Associations between inverted repeats and the structural evolution of bacterial genomes. *Genetics* 164(4):1279-1289.
16. Lynch M*, et al.* (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences* 105(27):9272-9277.
17. Zhu YO, Siegal ML, Hall DW, & Petrov DA (2014) Precise estimates of mutation rate and spectrum in yeast. *Proceedings of the National Academy of Sciences* 111(22):E2310-E2318.
18. Denver DR*, et al.* (2012) Variation in base-substitution mutation in experimental and natural lineages of Caenorhabditis nematodes. *Genome biology and evolution* 4(4):513-522.
19. Ossowski S*, et al.* (2010) The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science* 327(5961):92-94.

20. Saxer G*, et al.* (2012) Whole Genome Sequencing of Mutation Accumulation Lines Reveals a Low Mutation Rate in the Social Amoeba Dictyostelium discoideum. *PloS one* 7(10):e46759.
21. Lee H, Popodi E, Tang H, & Foster PL (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proceedings of the National Academy of Sciences* 109(41):E2774-E2783.
22. Sung W*, et al.* (2012) Extraordinary genome stability in the ciliate Paramecium tetraurelia. *Proceedings of the National Academy of Sciences* 109(47):19339-19344.
23. Ness RW, Morgan AD, Colegrave N, & Keightley PD (2012) Estimate of the spontaneous mutation rate in Chlamydomonas reinhardtii. *Genetics* 192(4):1447-1454.
24. Tucker AE, Ackerman MS, Eads BD, Xu S, & Lynch M (2013) Population-genomic insights into the evolutionary origin and fate of obligately asexual Daphnia pulex. *Proceedings of the National Academy of Sciences* 110(39):15740-15745.
25. Long H-A, Paixão T, Azevedo RB, & Zufall RA (2013) Accumulation of Spontaneous Mutations in the Ciliate Tetrahymena thermophila. *Genetics* 195(2):527-540.
26. Schrider DR, Houle D, Lynch M, & Hahn MW (2013) Rates and Genomic Consequences of Spontaneous Mutational Events in Drosophila melanogaster. *Genetics*.
27. Muller HJ (1964) The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 1(1):2-9.
28. Duret L & Galtier N (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual review of genomics and human genetics* 10:285-311.
29. Jeffreys AJ & Neumann R (2002) Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nature genetics* 31(3):267-271.
30. Leushkin EV & Bazykin GA (2013) SHORT INDELS ARE SUBJECT TO INSERTION−BIASED GENE CONVERSION. *Evolution*.
31. Hicks WM, Kim M, & Haber JE (2010) Increased mutagenesis and unique mutation signature associated with mitotic gene conversion. *Science* 329(5987):82-85.
32. Nagylaki T (1983) Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences* 80(20):6278-6281.
33. Lamb B & Helmi S (1982) The extent to which gene conversion can change allele frequencies in populations. *Genetical Research* 39(02):199-217.
34. Hershberg R & Petrov DA (2008) Selection on codon bias. *Annual review of genetics* 42:287-299.
35. Kliman RM & Hey J (1994) The effects of mutation and natural selection on codon bias in the genes of Drosophila. *Genetics* 137(4):1049-1056.
36. Li Y-Y*, et al.* (2006) Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. *PLoS computational biology* 2(7):e74.
37. Singer GA, Lloyd AT, Huminiecki LB, & Wolfe KH (2005) Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Molecular Biology and Evolution* 22(3):767-775.
38. Bradeen JM, Timmermans M, & Messing J (1997) Dynamic genome organization and gene evolution by positive selection in geminivirus (Geminiviridae). *Molecular Biology and Evolution* 14(11):1114-1124.
39. Al-Shahrour F*, et al.* (2010) Selection upon genome architecture: conservation of functional neighborhoods with changing genes. *PLoS computational biology* 6(10):e1000953.

40. Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences* 104(Suppl 1):8597-8604.
41. Koonin EV (2009) Darwinian evolution in the light of genomics. *Nucleic acids research* 37(4):1011-1034.
42. Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* 10(3):195-205.
43. Hahn MW, Stajich JE, & Wray GA (2003) The effects of selection against spurious transcription factor binding sites. *Molecular biology and evolution* 20(6):901-906.
44. Lynch M, Scofield DG, & Hong X (2005) The evolution of transcription-initiation sites. *Molecular biology and evolution* 22(4):1137-1146.
45. Froula JL & Francino MP (2007) Selection against spurious promoter motifs correlates with translational efficiency across bacteria. *PLoS One* 2(8):e745.
46. Lynch M (2002) Intron evolution as a population-genetic process. *Proceedings of the National Academy of Sciences* 99(9):6118-6123.
47. Lynch M & Conery JS (2003) The origins of genome complexity. *Science* 302(5649):1401-1404.

CHAPTER 2

MUTATION RATE AND SPECTRUM IN THE FISSION YEAST SCHIZOSACCHAROMYCES POMBE[1]

[1] Behringer, M.G. and D. W. Hall. To be submitted to *Genetics*.

Abstract

Mutation rates have been characterized across all domains of life. However, to gain a complete understanding of mutation rate a comparative process must be employed. Here we compared the mutation rate and spectrum in the fission yeast *Schizosaccharomyces pombe* to the budding yeast *Sacchaomyces cerevisiae*. While the two organisms are approximately 600 million years diverged from each other, they share similar life histories, genome size and genomic G/C content. We found that *S. pombe* and *S. cerevisiae* have similar mutation rates, contrary to what would be expected due to *S. pombe*'s smaller published effective population size. Additionally, in contrast to *S. cerevisiae*, *S. pombe*'s genome, under relaxed selection, possesses an insertion bias and an increased mutation rate at cytosine nucleotides in the absence of methylation.

Introduction

Spontaneous mutation, the source of all genetic variation, is the fuel for evolution and the foundation of all genetic differences within and between species. A complete understanding of the mutational process, both in terms of the rate at which different mutations arise, including kind and location, and the determination of their fitness effects, is thus critical to understanding genetic variation at all levels. However, elucidating the important parameters of mutation is difficult for two reasons. The main difficulty is that spontaneous mutations occur infrequently, which means it can be difficult to obtain large enough samples of mutations to robustly detect patterns. One workaround is to artificially increase the mutation rate using chemical or other methods such as mutagens or X-rays (1-3), or genetic methods such as repair pathway knock outs (4, 5). However, it is clear that such methods bias the mutational spectrum in various ways (2, 6). Another way to

deal with the rarity of spontaneous mutations is to examine genetic differences between individuals, populations or species that have arisen via mutation. However, because many mutations are acted upon by natural selection, this raises the second main difficulty with studying spontaneous mutations. The observed genetic variation has been acted upon by selection and is thus a biased sample of spontaneous mutations (7, 8). Exacerbating this problem is the finding that sites in the genome previously thought to be essentially free of selection, such as those in intronic regions, intergenic regions and four-fold redundant codon positions are in fact often constrained by selection, making the study of mutation at these sites biased by selection (9, 10).

One method that has been employed to overcome the problems of rarity and selection in studying spontaneous mutation is the mutation accumulation (MA) experiment. MA experiments maintain multiple, initially identical lines at very low effective population sizes for many generations (11). Lines accumulate spontaneous mutations at rates proportional to their occurrence since selection is ineffective at enriching for beneficial or decreasing deleterious mutations, unless their fitness effects are large (12). Each line accumulates only a few mutations but, by having numerous lines, several hundred mutations can be captured across the lines. As whole genome sequencing becomes increasingly inexpensive, a large number of MA lines can be sequenced and mutations within the lines can be identified and used to more accurately estimate the frequency and spectrum of spontaneous mutation. This approach has been used to examine spontaneous mutations in MA lines in a number of eukaryotic species including *Arabidopsis thaliana, Caenorhabditis elegans, Chlamydomonas reinhardtii, Dictyostelium discoideum, Drosophila melanogaster, Paramecium tetraurelia, and*

Saccharomyces cerevisiae (13-19). The largest study of this kind to date, in *S. cerevisiae,* yielded 924 spontaneous mutations (14).

The species that have been utilized in MA experiments are a somewhat random group. While we have learned a great deal by comparing mutation rates and spectrum estimates across them (20), it is difficult to test hypotheses for differences in the observed spectra using such a disparate group of species. To begin to remedy this problem, we chose to perform a MA experiment in the haploid fission yeast *Schizosaccharomyces pombe* in order to compare results to *S. cerevisiae*. Originally isolated from millet beer, *Sc. pombe* contains elongated centromeres, like humans, and highly conserved pathways for DNA checkpoint activation, making it an important model organism for studies of the cell cycle and genomic stability (21, 22). While *Sc. pombe* and *S. cerevisiae* are distantly related evolutionarily, with a divergence time of 600-1200 million years (23, 24), they both exhibit similar life histories as single-celled yeasts and are cultured in the lab using essentially identical methods. In addition, they have similar G/C content (36.06% in *Sc. pombe* versus 38.29% in *S. cerevisiae*) and sequenced genome size (12.57 Mb in *Sc. pombe* versus 12.1 Mb in *S. cerevisiae*). However, they do differ in ways that allow tests of hypotheses concerning differences in mutation rates and spectra.

We expected the mutation rate in *Sc. pombe* to be larger than in *S. cerevisiae* because its effective population size is 2.6 times smaller (25, 26). Since most mutations are deleterious, natural selection favors reduced mutation rate. The observed mutation rate is expected to be limited by the drift barrier (12); the point at which the fitness advantage of an additional reduction in the mutation rate is the same magnitude as the strength of genetic drift, on the order of the reciprocal of the effective population size.

Using the relationship reported in Sung *et. al.* (2012), we predicted that *Sc. pombe* will have a mutation rate that is approximately 2.4 times larger than *S. cerevisiae*. Estimates of mutation rate using reporter genes (summarized in the Appendix of Lynch, 2010) indicate that the base substitution rate in *Sc. pombe* is ~0.82 x 10^{-9} per base per cell division, which is about four fold higher than the prediction given the mutation rate estimate of 0.167 x 10^{-9} per base per cell division in *S. cerevisiae* (14). The deviation from the expectation may be due to the previous use of reporter genes. We expected to estimate a mutation rate in *Sc. pombe* that was approximately 0.4 x 10^{-9} per base per cell division.

Since the genomic G/C content of *Sc. pombe* is essentially identical to *S. cerevisiae*, we conjectured that the SNM bias, which includes the relative rates of mutations among the different base pairs, would be substantially similar to *S. cerevisiae*. One important finding in *S. cerevisiae* was an elevated mutation rate of G:C base pairs (14). In *S. cerevisiae*, G:C base pairs mutate at a very high rate compared to A:T base pairs. Surprisingly, the mutation rate bias is actually too large to explain the actual G/C content of *S. cerevisiae*, suggesting at least one other force affects G/C content. Further, in *S. cerevisiae*, G:C base pairs are especially mutagenic when they occur in certain trinucleotide contexts, specifically when they are the middle base pair in CCG and TCG trinucleotides. This was interpreted as indicating a very low occurrence of DNA methylation in *S. cerevisiae*, in agreement with previous work (27). *Sc. pombe* is believed to lack DNA methylation so we did not expect to see an elevation of the mutation rate at G:C base pairs that was context dependent.

In *S. cerevisiae*, analysis of small insertions and deletions of less than 50bp indicated a bias towards loss of genetic material across the diploid MA lines (14). In contrast, a MA experiment using haploid MA lines of *S. cerevisiae* observed a bias towards gain of genetic material (28). Since *Sc. pombe* is a haploid and was passaged as such in our MA experiment, we hypothesized that we would see a bias towards a gain of genetic material across small indels in this haploid species.

We predicted that we would find evidence for a higher rate of mutations caused by non-recombinational repair of double strand breaks in *Sc. pombe* because it is haploid compared to diploid *S. cerevisiae*. While double strand breaks cannot be directly observed or always inferred with certainty in the MA framework, their occurrence can be inferred by the presence of multiple mutations in close proximity (29, 30). In *S. cerevisiae*, three double mutations, which consist of two single-nucleotide mutations (SNMs) adjacent to one another, were observed across MA lines that had accumulated 800 single nucleotide mutations, giving a relative occurrence of double SNMs of about 0.25% compared to the single nucleotide substitution rate. In addition, there were five "complex mutations" in the *S. cerevisiae* data, in which multiple SNMs, and often small indels, were in close proximity, giving a relative occurrence of complex mutations of about 0.25% compared to the single nucleotide substitution rate. We predicted that we would see a substantial increase in the rate of double mutations and complex mutations in *Sc. pombe* compared to *S. cerevisiae* if these types of events are indeed caused by repair of double stranded breaks.

We also predicted that the observed rate of aneuploidy would be very low in *Sc. pombe*. Since *Sc. pombe* is haploid, loss of a single chromosome results in loss of all

copies of the genes on that chromosome, which would likely be lethal and thus unobservable. In addition, gain of a single chromosome would result in a doubling of gene dose for all genes on the chromosome, which is a greater increase than the 1.5 fold increase when trisomy occurs in diploid *S. cerevisiae*. If effects due to gene dosage are increased with larger differences in dosage across genes, this effect would further reduce the likelihood of observing aneuploidy. In addition, *Sc. pombe* has only three chromosomes, implying that many more genes would be affected by an aneuploidy event than in *S. cerevisiae*, which has a similar genome size but 16 chromosomes. The only instance of aneuploidy reported in previous work in *Sc. pombe* is disomy of chromosome III (31). For these reasons, we predicted we would see no chromosome loss and probably no chromosome gain across the MA experiment.

This study presents the first genome wide estimates of mutational parameters for *Sc. pombe*. We examined 79 MA lines, cultivated as haploids for an average of 1952 generations. We were able to identify a total of 697 mutations. These mutations allow us to calculate precise estimates of the mutation rate and spectrum for *Sc. pombe.*

Results

A derivative of the haploid fission yeast *Sc. pombe* strain 972h-, which is the sequenced *Sc. pombe* reference strain was acquired from ATCC and used as the ancestor of the MA lines. From this ancestor, a set of 96 clonal MA lines was created and passaged independently for approximately 1952 generations (Materials and Methods). Lines were bottlenecked by transferring a single random colony every 48 ±1 hours (~20 generations). Each MA line was frozen in 15% glycerol and stored at -80ºC every 10 transfers. Photographs of colonies from all MA lines taken every seven transfers were

used to estimate the average number of cell generations per transfer. Over the course of the MA experiment, average growth rate declined approximately 5%, so that colonies near the end of the MA experiment contained approximately half the number of cells as at the beginning. The average number of cells in a colony was used to estimate the effective population size in an average MA line as 10.26 cells. This small effective population size implies that deleterious (beneficial) mutations with a fitness effect greater than ~0.097 would be underrepresented (overrepresented) due to the action of selection (28, 32). Mutations with smaller effects are expected to accumulate approximately at random in the MA lines. 100-bp, paired-end Illumina, whole genome shotgun DNA libraries were constructed for 96 MA lines and two biological replicates of the ancestral strain and run on two lanes of Illumina HiSeq2000 with an aim of ~ 60x coverage per line. Except for telomeric regions in lines with sequencing depth below 43x read depth coverage was uniform across all three chromosomes in all lines (Appendix A).

Sequencing reads were mapped to the *Sc. pombe* reference genome using the Burrows-Wheeler aligner (BWA) software (33) and nucleotide variants were identified using the Genome Analysis Toolkit (GATK) (34). Assuming *a priori* that the per base mutation rate would resemble that of *S. cerevisiae* (~ 1.67×10^{-10} per base per generation), the probability of acquiring the same mutation in two lines over 1952 generations was ~1.06×10^{-13} (14). If we observed the same mutation across lines, it thus implied across-line contamination. Nineteen of the 96 MA lines were lost due either to contamination by a different microbial species (6 lines) or to across MA line contamination (11 lines), which left 79 lines for further analysis.

Approximately 2%, equal to 248kbp, of the genome was excluded from the analysis to ensure precise identification of new mutations. This included centromeres, telomeres, mating type loci and the representative rDNA repeat belonging to the two tandem rDNA repeat arrays on chromosome III, which were already excluded from the 12.47 Mb reference genome (35).

Differences between the MA ancestor and *Sc. pombe* reference genome

A total of 272 fixed mutational differences were identified between our MA ancestor line and the reference genome for *Sc. pombe* strain 972h-. These fixed differences included 85 single nucleotide mutations (SNMs), 129 small insertions (<50 bp), 47 small deletions (< 50 bp), 6 double SNMs and 5 complex mutations (Appendix A). Double SNMs are SNMs that occur adjacent to one another. Complex mutations are 3 or more SNMs and/or indels occurring within 50 nucleotides of one another. Both complex mutations and double SNMs were deemed to be non-independent events and were thus analyzed separately. Of the 85 SNMs, 41 were transitions and 44 were transversions, giving a transition to transversion ratio of 0.93. Within transitions, after correcting for genomic G/C content and assuming that all changes were mutations from the reference strain to the ancestral strain, G:C → A:T mutations were 1.38 times more frequent than A:T → G:C mutations (Figure 2.1).

The distribution of SNMs across chromosomes was not significantly different from the expectation based on chromosome length ($X^2 = 5.45$, $p = 0.06$). This result held whether we tested all three chromosomes, or just chromosomes I and II ($X^2 = 3.05$, $p = 0.06$), which together represent 81% of the genome. Chromosome I (5.58 Mb) contained 75 small insertions, which is greater than the 57 expected based on chromosome length

(X^2 = 12.21, p = 0.002). All other mutations, including small deletions, complex mutations, and double SNMs, did not show bias across chromosomes, after accounting for length (Appendix A).

Differences between the MA lines and the MA ancestor

Single Nucleotide Mutations

Across the MA lines, 327 SNMs arose during MA, which gives the genome-wide mutation rate for *Sc. pombe* as 1.70 x 10^{-10} ± 0.32 per base per generation (Figure 2.2). This estimate does not include those single nucleotide changes that were present in double or complex mutations, which were analyzed separately. Including SNMs in double and complex mutations increases the genome-wide mutation rate estimate to 3.02 x 10^{-10} per base per generation.

The number of SNMs varied across the 79 MA lines from zero (eight lines) to 13 (one line) with an average of 4.14 SNMs per line (ignoring doubles and complex mutations). Surprisingly, the distribution of mutations across MA lines was not consistent with a Poisson distribution ($p < 0.001$). Instead, the distribution was consistent with a negative binomial distribution (gamma-Poisson) ($\lambda = 4.16$, $\gamma = 2.10$; $p = 0.797$) (Figure 2.3). SNMs appeared to occur at random with respect to protein encoding genes: 52.4% of SNMs occur in the 57% of the genome that is protein coding (Fisher's exact, p = 0.27), and 3.6% of SNMs occur in the 3% of the genome that is intronic sequence (Fisher's exact, p = 0.83), suggesting that selection was inefficient during MA.

The distribution of SNMs across the three chromosomes was not significantly different from the expectation based on chromosome length (X^2 = 5.58, p= 0.06). However, when we tested just chromosomes I and II we found that there were more

mutations than expected on chromosome II (X^2 = 5.06, p = 0.02) (Appendix A). The distributions of all other mutation classes (small insertions, deletions, complex, double SNMs) were not significantly different from expected based on chromosome length.

Single Nucleotide Mutation Biases

We investigated possible biases within the mutational process. We observed a transition to transversion (Ts/Tv) bias of 0.72 (Figure 2.1). Within transitions, G:C → A:T mutations were 2.02 times more frequent than A:T → G:C. If we assume that G/C content in *Sc. pombe* is determined solely by the mutational process, the mutation biases we observed indicate that the equilibrium genome G/C content is expected to be 25.14%, far less than the 36.06% in the reference genome.

G/C content did not appear to have an effect on local mutation rate. For this analysis we determined the G/C content across the genome in 10kb windows. We then ranked windows based on GC content, and then split the ranked list into groups so that each group had approximately the same numbers of SNMs. We then calculated the number of bases in each group and determined the mutation rate for that group. Regardless of how many groups we used, we found no significant difference in mutation rate across groups, suggesting that G/C has no effect on local mutation rate. Data for the case where we split into two groups is shown in Appendix A (Pearson's r; all base pairs: r =0.384, p = 0.175; A:T base pairs: r = -0.135, p = 0.645; G:C base pairs: r =0.523, p = 0.054) (Appendix A).

Replication time had a complex effect on mutation rate. SNMs were assigned a replication time based on the average time at which the closest ori to the SNM fires during mitosis (36). There is not very much variation in the firing time of different oris

in *Sc. pombe*, but we again made a ranked list of firing times. We split this list into three groups, with each containing approximately the same number of SNMs, counted the total number of bases in each groups and calculated the mutation rate. The three groups were early, with replication times before 75 minutes, mid, with replication between 75 and 76 minutes, and late, with replication times later than 76 minutes. Across all base pairs, mid replication times were associated with smaller mutation rates however this difference was not statistically significant. (one sample t-test; all base pairs: $t = 0.89$, $p = 0.469$; A:T base pairs: $t = 1.57$, $p = 0.256$; G:C base pairs: $t = 2.11$, $p = 0.170$) (Appendix A).

Using RNA-seq data collected from our ancestor line, we investigated if gene expression affected local mutation rate. SNMs were separated into three equal groups: low expression (logFPKM < 1.25), medium expression (logFPKM: 1.25 - 1.75), and high expression (logFPKM > 1.75). Since our protocol for creating RNA libraries employs poly-A capture, tRNA and rRNA which normally have very high expression levels are not represented in our transcriptome. When SNMs that occur in tRNA and rRNA are excluded there is no difference in relative mutation rate due to gene expression (Pearson's $r = 0.069$, $p = 0.860$). However, adding these SNMs back, and assuming that they have high levels of expression (35), highly expressed regions have a borderline significantly higher relative mutation rate (Pearson's $r = 0.639$, $p = 0.064$) (Appendix A).

Lastly we examined the effect of the trinucleotide context, which is the identity of the nucleotides immediately before and after the mutated nucleotide, on mutation rate (14). Each nucleotide position within the genome, along with its neighboring bases, was assigned to one of 32 trinucleotide possibilities. Strand orientation was ignored and trinucleotide groups were defined such that complementary trinucleotides belong to the

same group. Additionally, within G:C base pairs, mutation rates at CpG sites were increased specifically at the CCG (CGG) and GCG (CGC) trinucleotide classes, which had the highest mutation rates among all classes (Figure 2.4). Mutations that occured at CpG sites were not biased in the direction of C → T which would be expected if the increase mutation rate was due to methylation. Of the 44 mutations occurring at CpG sites, 23 were C → A mutations (52.3%), 18 were C → T (40.1%) and 3 were C → G (6.6%).

Small Insertion and Deletion Mutations

Across the MA lines, we identified 335 small indels of less than 50 bp, including 288 insertions and 47 deletions (list of indels are in Appendix). There was a mutational bias for increased genome size: the observed indels resulted in a net gain across all lines of 340 bp. Small indels occured as frequently as SNMs across the MA lines and the resulting spontaneous indel rate, 0.174×10^{-9} /base/generation, is essentially identical to the mutation rate calculated for SNMs, ignoring double SNMs and complex mutations. Indels however are not randomly distributed throughout the genome. They were substantially under represented in coding regions (observed: 33, expected: 191; Fisher's Exact Test: $p < 0.001$), occurred as frequently as expected based on target size in introns (observed 21, expected 10; $p = 0.064$), and were over represented in non-coding regions (observed 314, expected 134; $p < 0.001$).

Effects of SNMs and Indels

We annotated the expected effects of our SNM and small indel mutations using Ensembl's variant effect predictor (VEP) software (37). For the 183 SNMs that occurred within coding regions, 53 were synonymous, 113 were missense, 12 were within introns,

1 was within a splice donor site, 1 was within a splice acceptor site, 2 were nonsense and 1 changed a termination codon (UAA) into another termination codon (UAG) (Figure 2.5). The ratio of synonymous to missense changes (0.398) was not statistically different from the expected (0.323), computed using randomly generated protein coding sequences (Z-test for proportions: 1.104, $p = 0.271$) (Appendix) (38). This suggests selection did not reduce the number of missense mutations captured during MA.

Among the 54 small indels that occurred within coding regions 7 were deletions, of which 5 were frameshift deletions and 2 were within introns. For the remaining 47 insertions, 23 were frameshift insertions, 5 were in-frame insertions ranging from 6-12 bp in length, and 19 were within introns (Figure 2.6, Appendix A). The proportion of indels in coding regions that were in-frame was not significantly different from 1/3 (Fisher's Exact Test $p > 0.2$).

Double Mutations, Complex Mutations and Structural Variants

As expected, no aneuploids or copy number variants were detected among the haploid MA lines. We specifically searched for other structural variants including medium-sized deletions (50-1000 bp), inversions, translocations and duplications using the DELLY software that predicts structural variants from short insert paired-end sequences (39). Of these structural variants, we detected 3 medium-sized deletions, two of which were flanked by a number of SNMs (Appendix). In addition to runs of SNMs associated with the three medium deletions, 16 other complex mutations and 17 separate, double SNMs were also detected. Complex mutations thus occurred at a rate of 8.32×10^{-12} per base per generation and double SNMs occurred at a rate of 8.84×10^{-12} per base per generation.

Error in Identification of Mutations

Our *Sc. pombe* lines were haploids and our sequencing coverage exceeded 25x in every line (average = 44x). Given previous work in diploid *S. cerevisiae*, we predicted we would have a very low likelihood of incorrectly calling a SNM or indel. To confirm our prediction, we randomly chose five lines and checked all called mutations using Sanger sequencing (Appendix). Together, these lines contained 15 SNMs, 20 small insertions, 3 small deletions, 1 double SNM and 1 complex mutation. Sanger sequencing confirmed all of the mutations identified from next-generation sequencing. Assuming that the probability that we obtained no false positives was not a rare event, i.e. the probability was greater than 5%, these confirmations indicate that our false positive error rate is no larger than 0.076 for SNMs and small indels combined, and could be close to zero. We did find that in the region of the complex mutation (MA Line 58, Chromosome II, positions 4273519 – 4274319) additional mutations not identified in next-genetation sequencing data were present (i.e. false negatives). These mutations were probably not detected due to the difficulty in reference mapping sequencing reads that are too divergent from the reference genome. This is supported by the extreme loss of coverage that was observed in this 800 bp region. Just 80 bp on either side of the region coverage was 48x, while across the complex mutation coverage averaged 9x, with 3 nucleotides in the center only being represented by a single read. Our results for this complex mutation suggest that the numbers of detected changes present in the 5 complex mutations we identified between the ancestor and reference strain and the 18 in the MA lines, are underestimates of the number actually present.

Discussion

Rate of mutation at single nucleotides

We expected to observe a higher mutation rate in *Sc. pombe* than has been reported in *S. cerevisiae*, based on estimates of effective population size (26) and on data using reporter genes (40). Our 79 *Sc. pombe* MA lines, cultured for approximately 1952 generations, accumulated 697 mutations, including 327 SNMs, 288 insertions of less than 50 bp, 47 small deletions of less than 50 bp, 18 complex mutations, 17 double SNMs, and 3 medium deletions (190 bp, 205 bp, 628 bp). The estimate of the base pair mutation rate in *Sc. pombe* based on the 327 accumulated SNMs is $0.170 \times 10^{-9} \pm 0.32$ per base per generation, which is almost identical to the mutation rate of 0.167×10^{-9} per base per generation based on 867 accumulated SNM mutations in *S. cerevisiae* (14). The effective population size in *Sc. pombe* was estimated as 1.0×10^{7}, however this estimate uses the mutation rate estimate from haploid *S. cerevisiae* of 0.33×10^{-9} per base per generation (28) and the standing genetic variation among *Sc. pombe* strains (26). With our estimate of the mutation rate, which is half as large, the resulting estimated effective population size should be twice as large or 2×10^{7}, much closer to the *S. cerevisiae* estimate of effective population size.

Spectrum of single nucleotide mutations

If differences between the ancestor and the reference represent both neutral and selected mutations, then the mutation spectrum might differ from that observed among the MA lines. In the MA lines SNMs represent 46.9% of the observed mutations, while in the MA ancestor they are significantly less common. The lower proportion of SNMs among differences between the ancestor and reference suggest that ~1/3 of the SNMs that

arose in the lineages linking the two strains are deleterious and thus prevented from accumulation. Additionally, the location of SNMs within the genome with respect to exon versus intron, and non-coding regions in both the ancestor and MA lines are similar suggesting that many of SNMs retained in the ancestor are neutral (Figure 2.2, Appendix A).

Bias in single nucleotide mutations

Within SNMs there isn't a significant difference between the rate of transitions and transversions in the ancestor compared to the reference or in the MA lines compared to the ancestor. However, there is a difference in mutational bias. Within MA lines, G:C → A:T transitions occurred at a rate that was 2.02 greater than A:T → G:C transitions. This transition bias was greater than the bias of 1.39 observed between the ancestor and the reference genome. If the variation within the ancestor had been exposed to selection, the lower transition bias may be evidence of a selective force maintaining G/C content within the genome. When comparing *Sc. pombe* to *S. cerevisiae*, even though the G/C content of both genomes are similar (35% in *Sc. pombe*; 38% in *S. cerevisiae*) the equilibrium G/C content of both genomes, calculated from the mutation spectra, are very different (25% in *Sc. pombe*; 32% in *S. cerevisiae*). In both species, the observed G/C content is higher than predicted from mutation biases suggesting that there is either selection for lower G/C content or some other mechanism, perhaps biased gene conversion, that is causing the increase in equilibrium G/C content. G/C content in *Sc. pombe* is further from its expected equilibrium, perhaps suggesting stronger selection or more biased gene conversion in this species. The lower equilibrium G/C content of *Sc. pombe* is due to different mutational biases in the two species. In particular, G:C → A:T

or T:A is 2.97 times more frequent than A:T → G:C or C:G in *Sc. pombe* but only 2.25 times more frequent in *S. cerevisiae*.

Rate and Spectrum of Indels

Small insertions or deletions may have been favored by selection since 17.3% of the differences between the ancestor and reference are indels compared to 6.7% in the MA lines, which is a significant difference. Small deletions particularly, may be favored in the ancestor since excess DNA is a mutational hazard, making the genome more vulnerable to mutations that may cause spurious transcription (41). This is in contrast to our MA lines where we observed an insertion bias with a net loss of 340 bp. A similar bias towards insertions has also been observed in *C. elegans* (3.75 insertions to deletions) and implies that the insertion/deletion balance alone does not account for *Sc. pombe*'s genome size (42).

In addition to nucleotide substitution biases, *Sc. pombe* experiences small indels at a rate than is almost 35 fold higher than *S. cerevisiae* (1.74×10^{-10} vs. 5.0×10^{-12} per base pair per generation). One possibility may be due to the differences in genome complexity between the two species. Most indels identified in *Sc. pombe* were within low complexity, intergenic regions such as microsatellites and mononucleotide runs. *Sc. pombe*'s genome is 60.2% protein encoding (57% excluding introns) and *S. cerevisiae*'s genome is 71% protein encoding (70.5% excluding introns) indicating more noncoding DNA. However, in an analysis of short simple repetitive sequences, there is little observed difference in the amount of repetitive sequence in *S. cerevisiae* and *Sc. pombe* (43). A second possibility is that there may be differences in the mismatch repair pathway in between these two species resulting in an increase of insertions in *Sc. pombe*.

Complex mutations, double mutations and aneuploidy

We also predicted an increase of mutations associated with double stranded breaks; particularly double SNMs and complex mutations. Double stranded breaks are lethal to a cell unless repaired. Repair can involve homologous recombination, which tends to be the preferred mechanism (44), but can also utilize nonhomologous end-joining (NHEJ). Recombinational repair requires homologous copies of DNA, which in a haploid organism are present during the S and G2 phases of the cell cyle, when sister chromatids are present. In the absence of homologous DNA, double stranded breaks are repaired through NHEJ. Homologous recombination is substantially more likely to be an error-free method of repair when sister chromatids are involved, while non-homologous end joining is considered to be error-prone introducing small insertions or deletions when unclean ends exist inhibiting precise repair (45, 46). Throughout our experiment we observed 18 instances of complex mutations (combination of a SNM and Indel or 3+ SNMs within a 50bp region). Six of the 18 observed complex mutations occurred within flocculin genes, specifically within the characteristic tandem repeats contained in their gene sequence (47). These tandem repeats make flocculation genes highly prone to recombination. It is not unsurprising then that should double stranded breaks commonly occur in regions containing flocculation genes that we should repeatedly observe this mutation signature potentially associated with repair events in absence of a homolog in such a non-random fashion.

As we expected given the haploid state, only three chromosomes, and previous work (31), no instances of aneuploidy were detected during MA. Aneuploidy has only been observed as a disomic haploid of chromosome III in *Sc. pombe*. Even if a disomic

haploid had occurred and been fixed in a line during MA, it would likely have been unstable (31, 48) and thus lost, and would thus have been difficult to observe in the experiment. Interestingly, when *S. cerevisiae* is passaged as a mitotic haploid it tends to be very unstable and at large population size, where selection is effective, it reverts to a diploid state (49).This instability is in contrast to the relative stability of diploid strains (14, 50). This perhaps implies that while aneuploidy is more difficult in haploids than in diploids if all else is equal, when diploidy is the natural state, disomies, or other steps towards diploidy, may be tolerated and perhaps even favored in haploids. It would be interesting to see if the instability of haploidy in *S. cerevisiae*, which is a natural diploid, would also be present in a diploid *Sc. pombe*, which is a natural haploid.

Cytosine mutation in absence of methylation

One of the major surprises of the *S. cerevisiae* mutation spectrum is the high mutation rate observed at some C:G base pairs, especially when the C is adjacent to a G in some CpG dinucleotides, especially CCG and TCG. We observe a similar unexpectedly high mutation rate at some C:G base pairs, especially CCG and GCG. In many eukayotes a major cause of mutation at cytosine nucleotides is spontaneous de-amination of methylated cytosines (m^5C) to thymine, which results in a T-G mismatch that can then be repaired, or replicated, to give a CG → TA substitution. Methylated cytosines usually reside within CpG sites as a form of epigenetic gene-regulation. *Sc. pombe* is not thought to contain DNA methylation and thus cytosines are expected to be unmethylated within genomic DNA (51). Mutation rates within *Sc. pombe* at A:T base pairs are relatively uniform with respect to adjacent base pairs. However, at G:C bases, the local mutation rates at CCG and GCG are markedly increased. In *S. cerevisiae*, the

increased mutation rates at CCG and TCG were attributed to very low levels of methylation of cytosine at those CpG sites. Such sites would then be susceptible to spontaneous de-amination and high C:G → T:A substitution rates. If the existence of previously undetected, very low levels of methylation is also a viable hypothesis in *Sc. pombe*, then there should be an excess number of C:G → T:A mutations at the CCG and GCG sites. However, there are significantly more C:G → A:T mutations at CCG and GCG sites in *Sc. pombe*. Additionally, when the genomic mutation biases for *Sc. pombe* are compared to the mutations observed at CpG sites, the cytosines mutate as expected (number of mutations observed at CpG sites: C → T 18, C → A 23, C → G 3; number of mutations expected based on genomic mutation biases: C → T 16, C → A 22, C → G 6; $X^2 = 1.76$, $p = 0.41$, suggesting a different mechanism than deamination of methylated cytosine.

Materials and Methods

Mutation Accumulation Lines

Sc.pombe MA lines were passaged in the same manner as described for *S. cerevisiae* (Joseph and Hall 2004). Briefly, the haploid ancestral line, 972 h- (ATCC 26189), was streaked onto rich, solid YPD medium (1% yeast extract, 2% peptone, 2% dextrose, 2% agar) and incubated at 30ºC. From the streaked ancestor, 96 random isolated colonies were selected after 48 hours and used to found 96 MA lines. Lines were cultured six to a YPD plate and bottlenecked by randomly selecting one isolated colony per line every 48 hours and transferring to a new plate. Lines were passaged for a total of 100 transfers (200 days). Every ten transfers, a random colony from each line was frozen and stored in 15% glycerol at -80ºC.

Every 10 transfers, photographs were taken of all 96 MA lines. From these photographs, colony size was measured for five random colonies per line using ImageJ (52). In addition, the number of cells per colony for colonies of various sizes were recorded at transfer 10 (T10) and 100 (T100) by suspending a single colony in 1ml of water and counting individual cells using a haemocytometer. These measurements were used to determine a standard curve of colony size versus cell number. Measurements of average colony sizes per transfer were then used to calculate the average number of cells at each transfer during the experiment. From the average number of cells present in a colony at transfer, the number of cell generations per transfer, and across the entire experiment, could be estimated, and the effective population size across the experiment could be determined.

Sequencing

MA lines were cultured from frozen T100 stock on solid YPD medium at 30ºC for 48h. A single colony from each line was selected, inoculated into 3mL liquid YPD, and incubated on a rotator at 30ºC for 48h. Cells were then pelleted and DNA was extracted using the YeaSTAR kit (Zymo Research) protocol with chloroform and an extended digestion time with zymolase of 2.5h at 37ºC. Whole genome shotgun libraries were prepared by the Georgia Genomics Facility using the Kapa Library Low-Throughput Library Preparation Kit with Standard PCR Amp Module KK8232 with dual SPRI size selection clean-up to generate 100bp paired end fragments with ~300bp inserts. After 7 cycles of PCR the libraries (96 haploid MA lines and 2 haploid ancestor samples) were pooled across two lanes of Illumina HiSeq 2000 machines.

QC, Mapping, and Identification of Mutations

Sequence reads from each library were quality controlled with the ea-utils and fastx toolkit software in order to remove low quality reads and residual adaptor sequence (53, 54)(Commands in Appendix). Following the workflow outlined in Zhu et al, 2014 (14) adjusted for a haploid dataset, QCed reads were then mapped to the *Sc. pombe* reference genome ASM294v2.24 with BWA v1.1.2, sorted and indexed with SAMtools v1.0, and assigned line identification numbers with Picard Tools v1.87 (21, 33). Duplicated reads were marked with Picard Tools and removed, and then the remaining sequence reads were locally realigned with GATK v3.2.2 (34). SNM and indel variants for each line and the ancestor were identified simultaneously using GATK's Unified Genotyper tool with parameter settings for haploid organisms. The resulting VCF files were converted to tab delimited text using VCFtools v0.1.12a vcf-to-tab function (55). Fixed differences between the MA ancestor sequences and the reference sequence were noted and removed. In order to call a variant, four reads with 75% of the reads favoring the alternate allele was needed. Regions of the genome that corresponded to centromeres, telomeres, and mating type loci (Approximately 248kbp) were excluded from the analysis to avoid inaccurate mapping. This is in addition to the two rDNA repeat arrays on chromosome III which are already excluded from the reference genome. Identified SNMs and small indels were annotated using Ensembl's variant effect predictor (VEP) while flanking regions were determined using the fill-fs program from the VCFtools package (33, 37).

Presence of medium and large structural variants (SV) were investigated using the Delly software package (39). Sorted and indexed bam files for each MA line as well as

the ancestor were used as input and coordinates of possible deletions, inversions, translocations, and duplications were output as VCF files. SVs that passed Delly's QC were investigated further using the integrated genome viewer IGV v2.1.23 (56). If IGV supported the SV call, the region was confirmed with PCR (Primers in Appendix).

Sequencing also allowed the detection of across-line and other microbial contamination. Across line contamination was deemed to have occurred if any two lines shared an identical new mutation. When this happened, one of the lines (chosen by coin flip) was discarded from the remainder of the analysis.

Gene Expression

To obtain accurate estimates of gene expression levels for our ancestor strain, we sequenced mRNA from 10 biological replicates. We selected 10 colonies, inoculated into 3mL liquid YPD medium, and incubated on a rotator at 30ºC for 48h. After 48h, mRNA was extracted using the MasterPure Yeast RNA Purification kit (Epicentere). mRNA libraries were constructed using the Illumina Truseq mRNA Stranded Kit, amplified using 13 cycles of PCR and sequenced on an Illumina HiSeq 2500. Libraries were sequenced as 100bp single-end reads. Sequenced reads were QCed in the same manner as genomic sequencing reads , reference mapped with TopHat v.2.0.13(57) , and assembled with Cufflinks (58). The median FPKM for each site across ancestor replicates was chosen to represent transcription at that site.

Verification of Identified Mutations

From the 79 MA lines analyzed in this study, five lines were randomly selected to verify the mutations that were identified bioinformatically with Sanger sequencing (Primers in Appendix). Additionally, all potential structural variants were compared to

the ancestor with PCR and gel electrophoresis. Primers were designed using Primer3 and PCR products destined for sequencing were cleaned using standard Exo-SAP protocol, and sequenced with ABI BigDye Terminator Cycle Sequencing Kit. Completed sequencing recations were submitted to the Georgia Genomics Facility and analyzed using an Applied Biosystems 3730xl 96-capillary DNA Analyzer.

References

1. Muller HJ (1930) Types of visible variations induced by X-rays inDrosophila. *Journal of Genetics* 22(3):299-334.
2. Greene EA*, et al.* (2003) Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in Arabidopsis. *Genetics* 164(2):731-740.
3. Blumenstiel JP*, et al.* (2009) Identification of EMS-induced mutations in Drosophila melanogaster by whole-genome sequencing. *Genetics* 182(1):25-32.
4. Hoffman PD, Leonard JM, Lindberg GE, Bollmann SR, & Hays JB (2004) Rapid accumulation of mutations during seed-to-seed propagation of mismatch-repair-defective Arabidopsis. *Genes & development* 18(21):2676-2685.
5. Denver DR, Feinberg S, Steding C, Durbin M, & Lynch M (2006) The relative roles of three DNA repair pathways in preventing Caenorhabditis elegans mutation accumulation. *Genetics* 174(1):57-65.
6. Koornneeff M, Dellaert L, & Van der Veen J (1982) EMS-and relation-induced mutation frequencies at individual loci in Arabidopsis thaliana (L.) Heynh. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 93(1):109-123.
7. Nachman MW & Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics* 156(1):297-304.
8. Ho SY, Phillips MJ, Cooper A, & Drummond AJ (2005) Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Molecular biology and evolution* 22(7):1561-1568.
9. Andolfatto P (2005) Adaptive evolution of non-coding DNA in Drosophila. *Nature* 437(7062):1149-1152.
10. Hershberg R & Petrov DA (2008) Selection on codon bias. *Annual review of genetics* 42:287-299.
11. Halligan DL & Keightley PD (2009) Spontaneous mutation accumulation studies in evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics* 40:151-172.
12. Sung W, Ackerman MS, Miller SF, Doak TG, & Lynch M (2012) Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences* 109(45):18488-18492.
13. Denver DR*, et al.* (2012) Variation in base-substitution mutation in experimental and natural lineages of Caenorhabditis nematodes. *Genome biology and evolution*:evs028.
14. Zhu YO, Siegal ML, Hall DW, & Petrov DA (2014) Precise estimates of mutation rate and spectrum in yeast. *Proceedings of the National Academy of Sciences* 111(22):E2310-E2318.
15. Sung W*, et al.* (2012) Extraordinary genome stability in the ciliate Paramecium tetraurelia. *Proceedings of the National Academy of Sciences* 109(47):19339-19344.
16. Keightley PD*, et al.* (2009) Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome research*:gr. 091231.091109.

17. Ness RW, Morgan AD, Colegrave N, & Keightley PD (2012) Estimate of the spontaneous mutation rate in Chlamydomonas reinhardtii. *Genetics* 192(4):1447-1454.
18. Rutter MT*, et al.* (2012) Fitness of Arabidopsis thaliana mutation accumulation lines whose spontaneous mutations are known. *Evolution* 66(7):2335-2339.
19. Saxer G*, et al.* (2012) Whole genome sequencing of mutation accumulation lines reveals a low mutation rate in the social amoeba Dictyostelium discoideum. *PLoS One* 7(10):e46759.
20. Lynch M & Conery JS (2003) The origins of genome complexity. *Science* 302(5649):1401-1404.
21. Wood V*, et al.* (2002) The genome sequence of Schizosaccharomyces pombe. *Nature* 415(6874):871-880.
22. Humphrey T (2000) DNA damage and cell cycle control in Schizosaccharomyces pombe. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 451(1):211-226.
23. Douzery EJ, Snell EA, Bapteste E, Delsuc F, & Philippe H (2004) The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? *Proceedings of the National Academy of Sciences of the United States of America* 101(43):15386-15391.
24. Heckman DS*, et al.* (2001) Molecular evidence for the early colonization of land by fungi and plants. *Science* 293(5532):1129-1133.
25. Brown WR*, et al.* (2011) A Geographically Diverse Collection of *Schizosaccharomyces pombe* Isolates Shows Limited Phenotypic Variation but Extensive Karyotypic Diversity. *G3: Genes, Genomes, Genetics* 1(7):615-626.
26. Skelly DA, Ronald J, Connelly CF, & Akey JM (2009) Population genomics of intron splicing in 38 Saccharomyces cerevisiae genome sequences. *Genome biology and evolution* 1:466-478.
27. Tang Y, Gao X-D, Wang Y, Yuan B-F, & Feng Y-Q (2012) Widespread existence of cytosine methylation in yeast DNA measured by gas chromatography/mass spectrometry. *Analytical chemistry* 84(16):7249-7255.
28. Lynch M*, et al.* (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences* 105(27):9272-9277.
29. Strathern JN, Shafer BK, & McGill CB (1995) DNA synthesis errors associated with double-strand-break repair. *Genetics* 140(3):965-972.
30. Holbeck SL & Strathern JN (1997) A role for REV3 in mutagenesis during double-strand break repair in Saccharomyces cerevisiae. *Genetics* 147(3):1017-1024.
31. Niwa O, Tange Y, & Kurabayashi A (2006) Growth arrest and chromosome instability in aneuploid yeast. *Yeast* 23(13):937-950.
32. Hall DW, Mahmoudizad R, Hurd AW, & Joseph SB (2008) Spontaneous mutations in diploid Saccharomyces cerevisiae: another thousand cell generations. *Genetics research* 90(03):229-241.
33. Li H*, et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078-2079.

34. McKenna A, *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20(9):1297-1303.
35. Wood V, *et al.* (2011) PomBase: a comprehensive online resource for fission yeast. *Nucleic acids research*:gkr853.
36. Eshaghi M, *et al.* (2007) Global profiling of DNA replication timing and efficiency reveals that efficient replication/firing occurs late during S-phase in S. pombe. *PLoS One* 2(8):e722.
37. McLaren W, *et al.* (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26(16):2069-2070.
38. Graur D (2003) Single-base Mutation. *Nature encyclopedia of the human genome*:287.
39. Rausch T, *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28(18):i333-i339.
40. Lynch M (2010) Evolution of the mutation rate. *Trends in Genetics* 26(8):345-352.
41. Hahn MW, Stajich JE, & Wray GA (2003) The effects of selection against spurious transcription factor binding sites. *Molecular biology and evolution* 20(6):901-906.
42. Denver DR, Morris K, Lynch M, & Thomas WK (2004) High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome. *Nature* 430(7000):679-682.
43. Karaoglu H, Lee CMY, & Meyer W (2005) Survey of simple sequence repeats in completed fungal genomes. *Molecular biology and evolution* 22(3):639-649.
44. Raji H & Hartsuiker E (2006) Double-strand break repair and homologous recombination in Schizosaccharomyces pombe. *Yeast* 23(13):963-976.
45. Cavero S, Chahwan C, & Russell P (2007) Xlf1 is required for DNA repair by nonhomologous end joining in Schizosaccharomyces pombe. *Genetics* 175(2):963-967.
46. Shrivastav M, De Haro LP, & Nickoloff JA (2008) Regulation of DNA double-strand break repair pathway choice. *Cell research* 18(1):134-147.
47. Verstrepen KJ, Jansen A, Lewitter F, & Fink GR (2005) Intragenic tandem repeats generate functional variability. *Nature genetics* 37(9):986-990.
48. Niwa O & Yanagida M (1985) Triploid meiosis and aneuploidy in Schizosaccharomyces pombe: an unstable aneuploid disomic for chromosome III. *Current genetics* 9(6):463-470.
49. Glazunov A, Boreĭko A, & Esser A (1988) [Relative competitiveness of haploid and diploid yeast cells growing in a mixed population]. *Mikrobiologiia* 58(5):769-777.
50. Nishant K, *et al.* (2010) The baker's yeast diploid genome is remarkably stable in vegetative growth and meiosis. *PLoS genetics* 6(9):e1001109.
51. Antequera F, Tamame M, Villanueva J, & Santos T (1984) DNA methylation in the fungi. *Journal of Biological Chemistry* 259(13):8033-8036.
52. Schneider CA, Rasband WS, & Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. *Nature methods* 9(7):671-675.

53. Aronesty E (2011) ea-utils: Command-line tools for processing biological sequencing data.
54. Gordon A & Hannon G (2010) Fastx-toolkit. *FASTQ/A short-reads preprocessing tools (unpublished)* http://hannonlab. *cshl. edu/fastx_toolkit*.
55. Danecek P*, et al.* (2011) The variant call format and VCFtools. *Bioinformatics* 27(15):2156-2158.
56. Robinson JT*, et al.* (2011) Integrative genomics viewer. *Nature biotechnology* 29(1):24-26.
57. Trapnell C, Pachter L, & Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105-1111.
58. Trapnell C*, et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* 28(5):511-515.

Figures

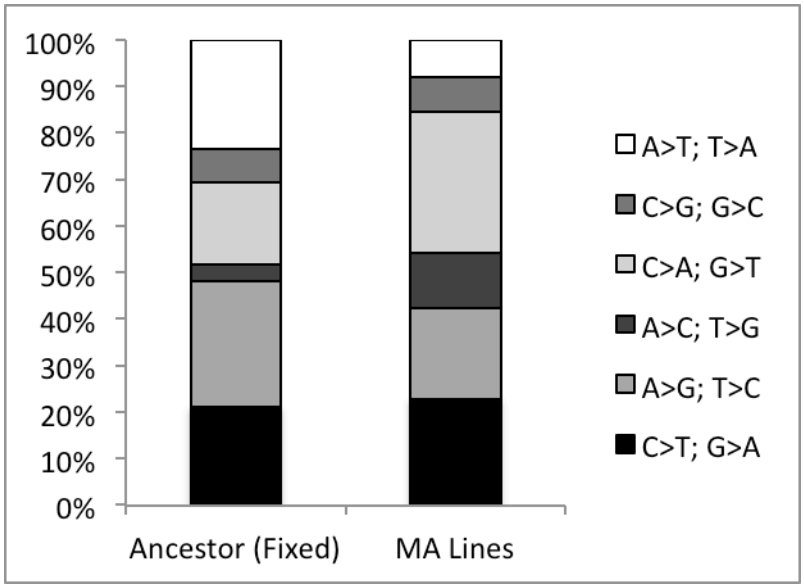

Figure 2.1. **Summary of mutations for each of six possible nucleotide changes for both the MA Ancestor and the MA lines compared to the reference genome.**

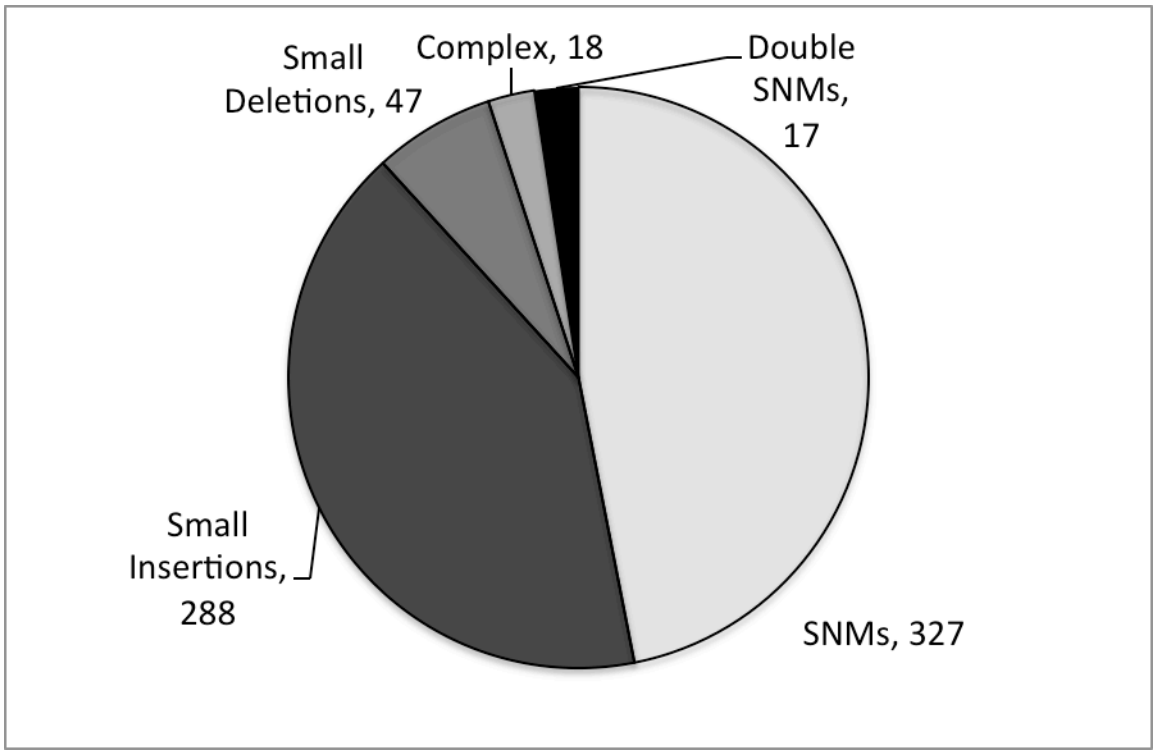

Figure 2.2. **Summary of all mutations identified across 79 MA lines.**

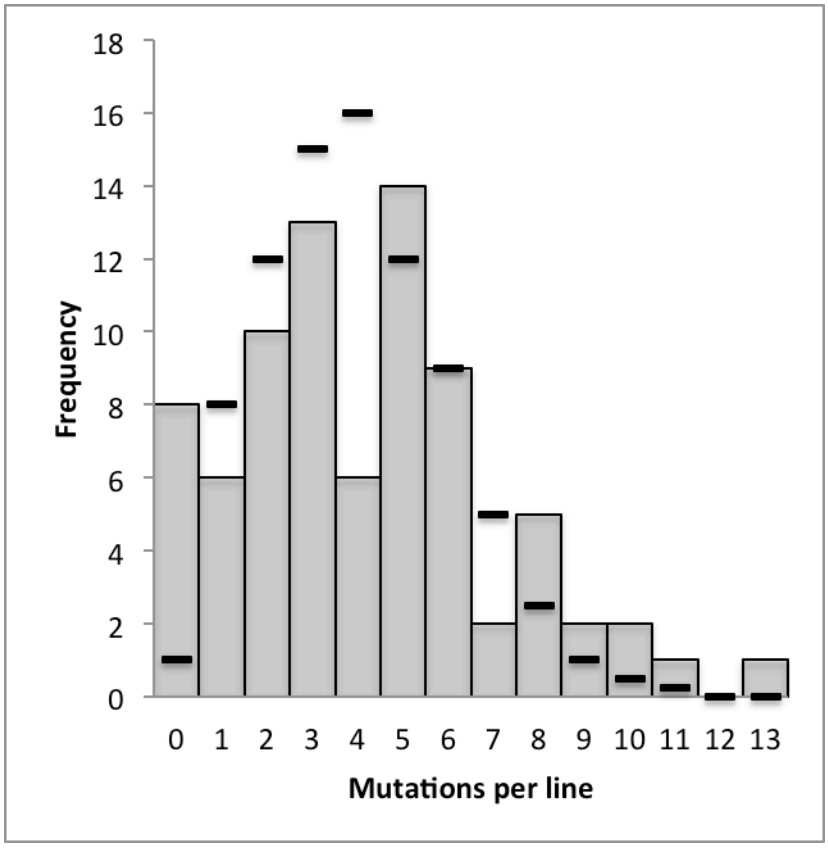

Figure 2.3. **Histogram of SNM counts per line in 79 MA lines.** Distribution is consistent with a negative binomial (P = 0.79, X^2 test). Bars represent the per line count for the 327 SNMs called across the 79 lines, Dashes represent the negative binomial distribution with $\lambda = 4.16$, $\sigma = 2.10$.

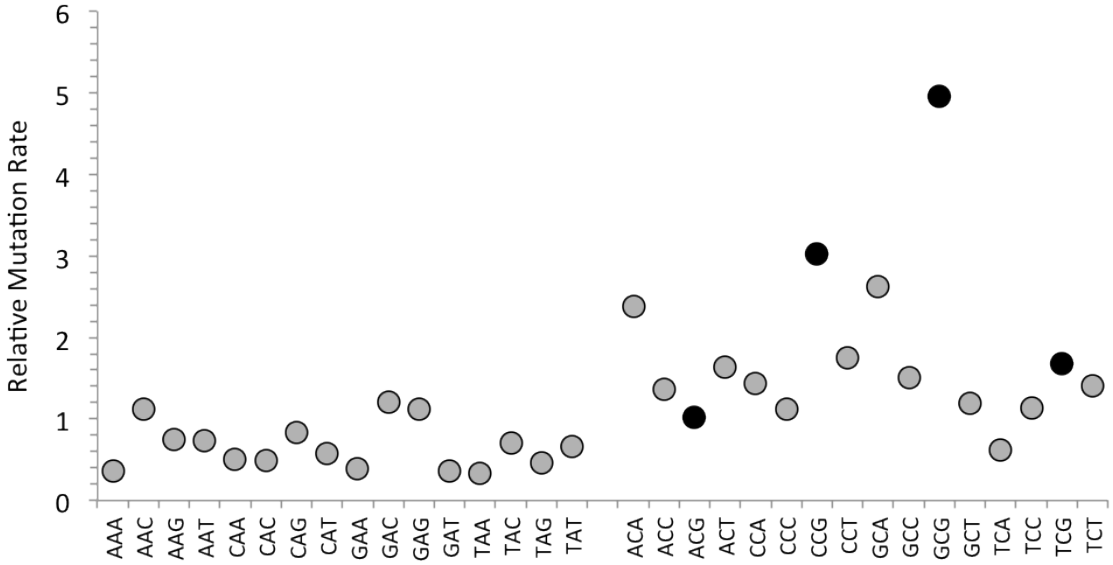

Figure 2.4. **Neighbor-dependent mutation rate defined as the effects of immediate flanking nucleotides on mutation rate at a central base of interest** (ex. C in an aCa trinucleotide). Trinucleotide classes represent mutation rates of both strand orientations (aCa trinucleotide class includes overall mutation rate at aCa and tGt sites). Mutation rate is shown relative to the average single nucleotide mutation rate across all sites (=1.7 x 10^{-10} per base, per generation). The average, relative mutation rate of 1.8 at G:C bases shows clear overall elevation over the corresponding rate of 0.66 at A:T bases. Black-filled points are those trinucleotiodes with a C in the central position with a G as the 3' neighbor.

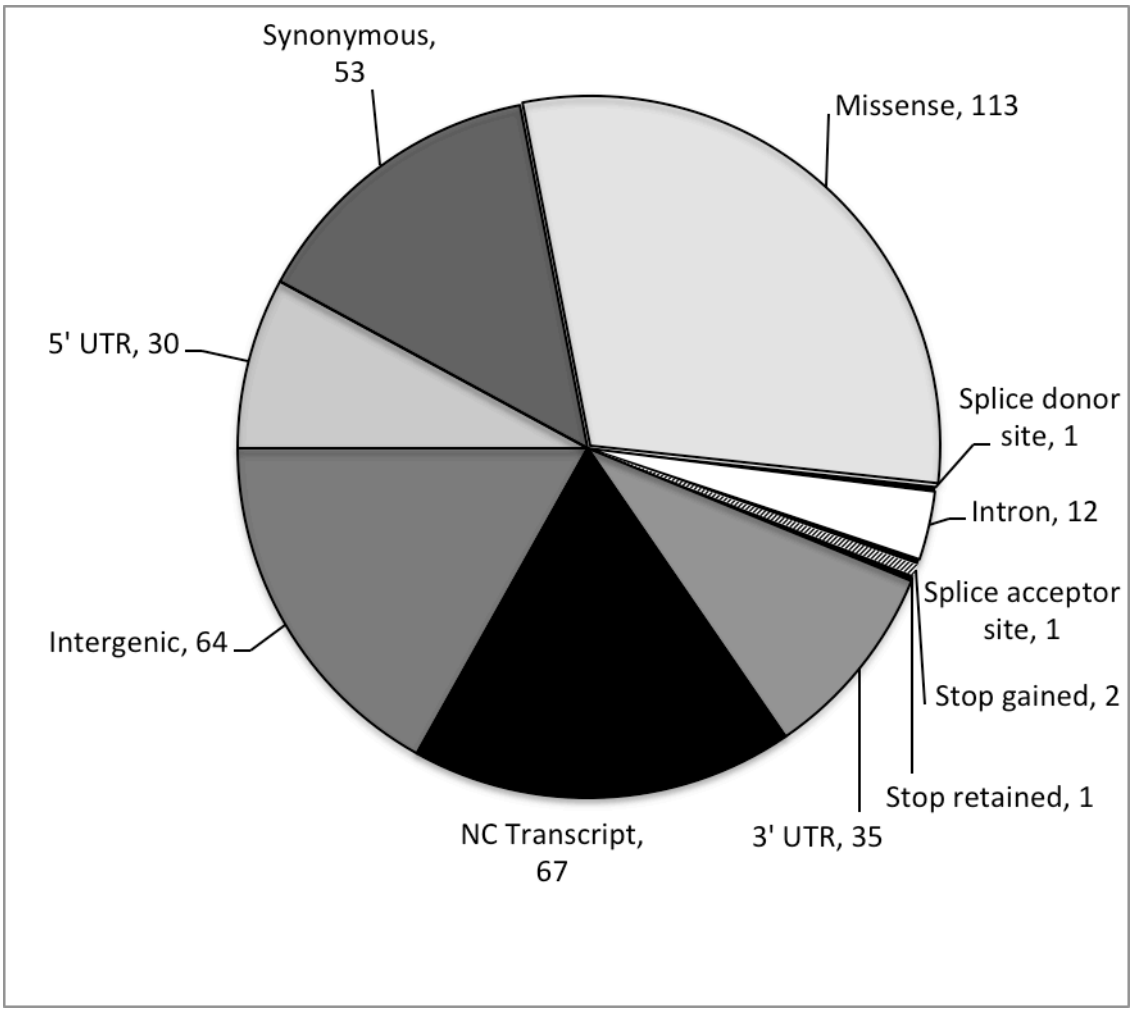

Figure 2.5. **Summary of the locations including predicted functional consequences of all SNMs**. Numbers of SNMs observed across all 79 MA lines are shown for each category. Sum of numbers is greater than the observed number of mutations across all MA lines because some mutations had multiple predicted functional effects.

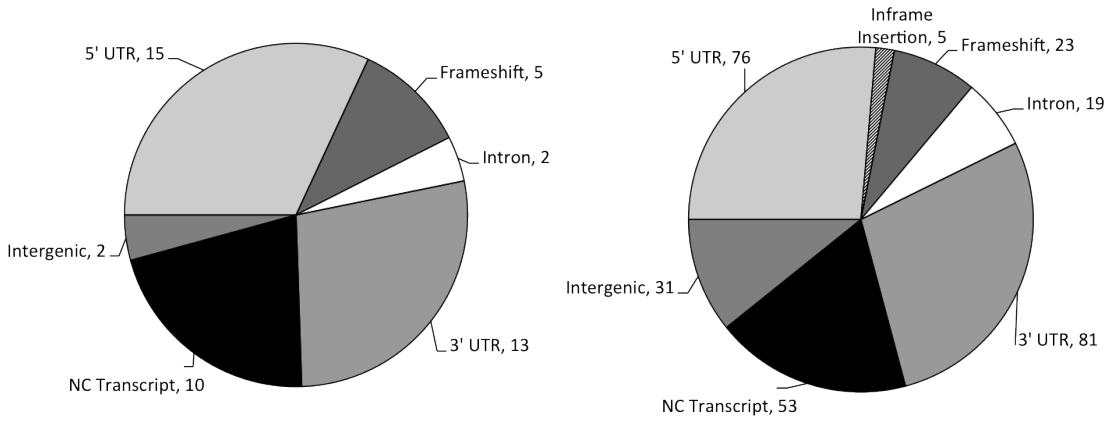

Figure 2.6. **Summary of the locations including predicted functional consequences of all small indels** (less than 50bp in size). Numbers of indels identified across all 79 lines are shown for each category. Sum of numbers is greater than the observed number of mutations across all MA lines because some indels had multiple predicted functional effects

CHAPTER 3

RATE AND BIASES OF MITOTIC GENE CONVERSION IN *SACCHAROMYCES CEREVISIAE*[1]

[1] Behringer, M.G. and D.W. Hall. To be submitted to *Genetics*.

Abstract

Recombination in concert with mutation, natural selection and random genetic drift is one of the major forces shaping genome architecture. Gene conversion, a common side effect of recombination, is responsible for the loss of heterozygosity associated with heteroduplex repair. To investigate the long-term genome wide effects of gene conversion we preformed a mutation accumulation experiment with a highly-heterozygous strain of *Saccharomyces cerevisiae*. After 2108 mitotic cell divisions, across 25 lines, 274 gene conversion events are observed; while across 85 lines, 39 chromosome duplication events were observed. Median gene conversion tracks length is observed to be 6.2 kb for crossovers and non crossovers with ~13% of events observed extending longer than 50,000 bp. While no G/C bias is observed for gene conversion tracts, deletion bias is observed for all tract lengths.

Introduction

Recombination, one of the four major forces affecting genetic variation, is responsible for assorting variation within and among chromosomes. During meiosis, recombination is a programmed event resulting in independent assortment of alleles (1). While in mitotically dividing cells, recombination mainly refers to homologous recombination; a method of repairing double stranded breaks (2). When repairing double stranded breaks through homologous recombination, the broken chromosome will use either the sister chromatid or a homologous chromosome as a template for repair (3). This often results in a region of non-reciprocally transferred DNA in addition to the heteroduplex created from repair. If the transferred DNA and/or resolved heteroduplex alter the original DNA sequence of the repaired chromosome, a gene conversion event

has occurred (4). Essentially, gene conversion refers to the stretch of loss of heterozygosity (LOH) following the repair of double stranded breaks.

While generally regarded as a non-adaptive force, recombination and as an end result gene conversion, can have a large impact on adaptation (5, 6). This is especially important in asexually reproducing organisms, where the potential to make recessive beneficial mutations homozygous, would be particularly advantageous (7, 8).This mitotic LOH has been observed in a number of organisms resulting in rapid adaptation (9-11).

Three main pathways are responsible for homologous recombination: synthesis-dependent strand annealing, double Holliday junction, and break induced replication (3). All double stranded breaks repaired by synthesis-dependent strand annealing resolve as non-crossovers in contrast to breaks repaired by double Holliday junction preferably resolve as crossovers but can resolve as non-crossovers as well (12). Whether double stranded breaks resolve as crossover or non-crossover has a direct effect on length of the gene conversion event. One genome wide study reported a median length of 7.6kb for gene conversions associated with chromosomal crossovers (13-15). While another reported equal lengths for both crossovers and non-crossovers (15). Further, due to the nature of break induced replication where the chromosome end is lost and replaced entirely by template DNA, associated gene conversion events may be especially long (16).

In addition to length of gene conversion events, there are also many reported biases in the location and the direction of gene conversion: G/C biased gene conversion (17-19), hotspot drive (20, 21), and indel drive. These biases, specifically G/C biased gene conversion and indel drive, may act in concert with spontaneous mutation shaping

the genome in terms of size and G/C content (18). However, these biases have mainly been discussed in the context of meiotic recombination and the effects of mitotic gene conversion have not been explicitly considered.

In this study, we present a whole-genome investigation of mitotic gene conversion by examining LOH in a highly-heterozygous strain of *S. cerevisiae*. We employed a mutation accumulation (MA) framework (Chapter 2) to examine gene conversion in an unbiased manner in the presence of reduced selection. The final data set includes whole-genome sequence data from 93 *S. cerevisiae* MA lines, which have been allowed to divide mitotically for 2108 generations. Here, we present the analysis of the first 25 MA lines. Using 51,895 heterozygous markers, we have detected a total of 274 LOH events across these 25 lines, ranging in length from 27 to 569,061 nucleotides. Overall a total of 13,422,425 bases experienced gene conversion allowing us to accurately investigate the presence/absence of G/C bias, insertion/deletion bias, and to map gene conversion hotspots.

Results

The diploid ancestor was generated by crossing *S. cerevisiae* haploid strains NCYC 3631 (designated strain PA (Y)) and NCYC 3596 (designated strain IN (D)), which are derivatives of strains YPS606 (oak tree, Pennsylvania, US) and DBVPG1106 (lici fruit, Indonesia) (22). Strains were chosen because they are genetically highly divergent generating 51,895 heterozygous sites when mated. From this ancestor, 96 initially identical MA lines were created and then passaged through single-cell transfer for 2108 generations following established protocols (Joseph and Hall 2004). Single-cell bottlenecks were created every 48h (~21 generations). Three lines were abandoned

during the course of the mutation accumulation experiment due to petite mutation. From the remaining 93 lines, as well as the two haploid parental and single diploid ancestor strains, 75bp paired-end whole genome shotgun libraries were constructed and sequenced in 3 runs of an Illumina NextSeq500 for an average of 77x coverage per line. Except for centromeres, telomeres, repetitive elements, and lines which experienced trisomy events (Table 1) (Appendix B), coverage was uniform across all chromosomes in all lines.

Quality controlled sequencing reads were reference assembled using BWA (23) and gene conversion events were detected using GATK (24). We define mitotic gene conversion as loss of heterozygosity that occurs in cells that are mitotically dividing regardless of whether the event occurred during the mitosis phase of the cell cycle. Gene conversion was detected by the conversion of sites that were originally identified as heterozygous in the ancestor, to homozygous in the direction of either haploid parent. In order for a heterozygous marker to be identified the site must be represented by at least 10 sequencing reads.

Number and length of conversion events

Throughout the mutation accumulation experiment, a total of 274 gene conversion events were detected for an estimated gene conversion rate of 5.06×10^{-10} per base per generation. This rate is the probability of a gene conversion event commencing at any particular base in the genome. Since we do not know the sequence direction in which gene conversion events were initiated, we arbitrarily call the left breakpoint (closest to position 0) the beginning of the event. Thus, each gene conversion event has exactly one beginning, which is to the left to one ending, or right-most downstream marker. Gene conversion events ranged in minimum length from 27 to 569,061 bp with a median of

7949 bp for an estimated loss of heterozygosity rate of 2.11 x 10^{-5} per base per generation. Gene conversion tracts were identified as uninterrupted stretches of heterozygosity loss in favor of a single parental haplotype. Tracts ended when there was either a change in the direction of loss of heterozygosity or a heterozygous marker with a quality score greater than 200. Since the power to detect gene conversion events relied on the spacing of naturally occurring markers, the resolution varied throughout the genome (Figure 3.1). The median distance between markers was 69 bp allowing very fine mapping of some gene conversion events. However, 382 markers representing 5.8% of the genome were located more than 1000 bp apart, therefore small gene conversion events within these regions could have gone undetected. In addition, once a gene conversion event occurred in a region, additional events could no longer be scored since heterozygosity would be gone. For these reasons, the estimated rate of gene conversion across our MA lines is likely to underestimate the actual gene conversion rate by a few percent. There was no detected bias in direction in terms of the preferentially converting to the parental strain sequences (IN or PA).

Types of conversion events

Gene conversion events can be categorized in terms of the type of recombination or repair events that may be responsible for their creation. A stretch of LOH that appears to continue through the end of the chromosome may be a result of crossover conversion or break induced repair (class A, Figure 3.2A), while a contained region of LOH surrounded by heterozygous markers are consistent with patterns expected from synthesis-dependent strand annealing or double Holiday junction (class B, Figure 3.2B). Additionally, adjacent stretches of LOH that seemingly originate from different parents

may be the result of a reciprocal crossover that preceded synthesis-dependent strand annealing or double Holiday junction (class C, Figure 3.2C), while adjacent stretches of LOH from seemingly different parents which extend to the end of the chromosome on one side and are flanked by heterozygous markers on the other may be the result of synthesis-dependent strand annealing or double Holiday junction that preceded a crossover recombination or break induced repair event (class D, Figure 3.2D). Across event classes gene conversions differed in both number and length (Table 3.2). Class B contains the majority of gene conversions accounting for 219 out of 274 events, with a median tract length of 6,189 bp, while class D contains the longest gene conversions with a median tract length of 211,673 bp.

Location of conversion events

The number of gene conversion events correlated strongly with chromosome length, with longer chromosomes experiencing more gene conversions than shorter chromosomes (Figure 3.3). To ensure this correlation was not skewed by longer gene conversions events, we repeated this analysis with gene conversion tract lengths shorter than 7,900 bp, approximately the median tract length, and the correlation remained unchanged (Appendix B).

Gene conversion hotspots were identified by gene conversion events having the same starting position (Figure 3.4)(Appendix B). In our observed 274 Gene conversion events, 68 events shared either a beginning or an end with another event. Five gene conversion events occurred at position 490579 on chromosome XII, four of which shared breakpoints at both the beginning and end of the tract, however, 3 of these events converted in the direction of strain IN, and the other in the direction of strain PA.

G/C biases in conversion events

No evidence of biased gene conversion towards G/C bases was detected. For all events, a total of 31066 A/T bases were converted to G/C and 31098 G/C bases were converted to A/T for a G/C bias ratio of 0.999. When track lengths are arranged into three groups each containing approximately equal numbers of events based on log scaled tract length (small tracts: 2.5 -3.5; medium tracts: 3.5 - 4.5; and large tracts: >4.5) small tracts seemed to favor conversion toward A/T bases with a G/C bias of 0.85 ± 0.03, while the other tract lengths were not significantly different from 1 (Figure 3.5). The G/C bias within small tracts appears to be driven by conversion tracts that are less than 317 bp long (log scaled value 2.5), where a G/C bias of 0.68 exists.

Indel biases in conversion events

A total of 534 nucleotides were lost across the 25 MA lines through gene conversion. To determine if there is any bias amongst gene conversion events in gain or loss of nucleotides due to tract length, we compared net gain/loss of nucleotides for each subset of tract lengths to the possible number of nucleotides that could be gained or lost within each gene conversion event of those tract lengths. Nucleotides were lost across all tract lengths ((bases lost/possible bases lost) small tracts: 9/113; medium tracts: 69/379), however substantial gain was observed in long tracts where a net 456 nucleotides were lost out of a possible 914. When looking at number of gene conversion events responsible for gain and loss of nucleotides, 50 long tract conversions resulted in net bases lost while only 21 resulted in bases gained creating a deletion bias of 2.38 for long tract conversions. If you combine these results with losses and gains across all tract sizes there exists a deletion bias of 1.39 for gene conversion events resulting in net nucleotide loss.

Discussion

Throughout this study, 93 highly-heterozygous MA lines were cultured for 2108 generations. We scored gene conversion events for 25 MA lines and in an additional 84 we have also scored aneuploidy events. In total, across the 18 MA lines, 220 gene conversion events were observed with a total of 11,563,514 nucleotides converted. Additionally, we have detected 24 trisomy events in 56 lines. We observed a probability of 1.03 x 10-5/base/generation for any base to be involved in a gene conversion event caused by heteroduplex repair (class B). This is a 10-fold increase from the gene conversion rate of 1.3 x 10^{-6} observed at the URA3 locus on chromosome IV by Yim, *et. al*.(15). This discrepancy is not just due to a reduced rate of gene conversion on chromosome IV where our observed rate of class B gene conversions is 1.93 x 10^{-5}, almost double our genome wide observation. Further, when only considering class B events that include the URA3 locus, we observe a gene conversion rate of 7.91 x 10^{-5}.

Trisomy in gene conversion lines

Mitotic trisomy can be caused by nondisjunction during anaphase or anaphase lag (25, 26). Through the course of this experiment, 39 trisomy events were observed across 85 lines. This is a large, significant, increase (2.4x) compared to the number of trisomy events observed in isogenic *S. cerevisiae* mutation accumulation lines (27), where 28 trisomy events were observed across 145 lines. One possibility for the difference in trisomy rate may be attributed to the genetic distance between the two haploids used to create our lines. There was no significant correlation between the % identity of the centromeres in the two haploid ancestor parental lines with the number of trisomy events across chromosomes, suggesting that nondisjunction is not being driven by centromere

divergence. Another possibility for the increased rate of trisomy may be related to the wild nature of the parent strains. In both oak and wine strains aneuploidy has been identified and explained as an advantage in stressful conditions (28-30). It is unknown however if wild strains experience spontaneous aneuploidy at higher rates than laboratory strains. Additionally, the probability that a chromosome becomes trisomic appears to be non random, with the distribution of trisomic events across chromosomes poorly fitting a Poisson distribution (Kolmogorov's $D = 0.344$, $p = 0.009$). Although 39 trisomy events occurred, seven chromosomes did not experience trisomy in any line, while chromosome V and VII had six, chromosome XII had seven, and chromosome XVI had ten trisomy events.

It has been suggested that mitotic recombination and trisomy may be linked (26). In a previous study, 5% of cells with a recombination event on chromosome V were also trisomic for chromosome 5. However when examining chromosome IV, the chromosome with the largest number of gene conversion events relative to its length, not a single trisomic event was observed. Additionally, no correlation was found when investigating all chromosomes for an association between trisomy and gene conversion ($r = 0.098$, $p = 0.72$).

Double Holliday junctions and synthesis-dependent strand annealing are responsible for long gene conversions

While it is not surprising that class A and D gene conversions accounted for the longest stretches of LOH, as break induced replication will replace entire ends of chromosomes. It is surprising however, that 13% of class B gene conversions extended over 50 kb. While we can rule out break-induced replication, since these gene

conversions do not extend to the chromosome end. It has been suggested that these extremely long gene conversions are due to the creation of a double stranded gap (13, 15, 31). The decision to repair through homologous recombination is regulated by the cyclin-dependent kinase CDK1 (32). This CDK, responsible for activating DNA damage check points and promoting repair by homologous recombination, is at lower concentrations during G1 and higher concentrations during G2/M (33). Because DNA damage checkpoints are not efficiently activated during G1, double stranded breaks can exist in S phase and encountered by the replication machinery. When this encounter occurs a stronger checkpoint response is induced resulting in extreme resectioning of both the 5´and 3´ strands leaving a double stranded gap (31). These gaps are shown to be repairable by homologous recombination(34) . It is also possible that a third unknown mechanism for mitotic recombination is responsible for long gene conversions.

Fragile sites leading to increased Gene Conversion

A phenomenon that may lead to recombination "hotspots" is fragile sites in the DNA (35). In the context of mitotic cell division, fragile sites have been associated with genome instability and DNA damage due to stalled replication machinery. Two potential recombination hotspots were identified: one on chromosome V and one on chromosome XII. The hotspot on XII at position 490579 appeared five times and was often associated with a more downstream hotspot at position 507626. This particular region has been described recently as a fragile site, with no report of gene conversion (36). However when this region is converted in the direction of strain IN, we observe an average tract length of 344133 converted nucleotides. The one event where this region converted in the direction of strain PA, the gene conversion tract was only 16634 nucleotides long. This

region is particularly interesting due to its proximity to the rRNA array. It has been previously determined that in the vicinity of these rRNA repeats collisions between replication and transcription machinery is often responsible for recombination (37). The other potential recombination hotspot on chromosome V appeared three times, each time converting in the direction of strain IN with tracts ranging from 942 to 22234 nucleotides. This hotspot spans the gene TFP1, which is also highly-transcribed (38).

Impact of Gene Conversion on Genome Maintenance

Although observed in meiotic gene conversion, G/C bias does not seem to appear exist in mitotic gene conversion. In fact, short gene conversion events (< 316 nt) appear to be biased in the A/T direction. This is unlikely to have a major effect on genome wide G/C content however, since gene conversions < 316 bp only account for 7% of gene conversion events and 0.025% of converted bases. More striking is the apparent bias towards deletions, which was observed across all ranges of gene conversion lengths. Specifically, in long gene conversions (> 31600 bp) where a net 456 bases were deleted out of a possible 914. Given that long gene conversions account for 25% of all events and 88% of bases converted, gene conversion may be an important factor in maintaining genome size.

Materials and Methods

Gene Conversion Lines

A highly-heterozygous *S. cerevisiae* ancestor line was created by mating two diverged *S. cerevisiae* haploids from the *Saccharomyces* Genome Resequencing project: NCYC 3631, a Mat alpha derivative of YPS606, and NCYC 3596, a Mat a derivative of DBVPG1106 (22). When mated, the lines contain 51,895 heterozygous sites or

approximately one site every 233 bases. The starting diploid, a highly-heterozygous ancestor was used to found 96, initially identical mutation accumulation (MA) lines, which were passaged by single-cell transfer every two days (~21 generations), as described in detail in Joseph and Hall (39). Briefly, the diploid ancestral line was streaked onto rich, solid YPD medium (1% yeast extract, 2% peptone, 2% dextrose, 2% agar) and incubated at 30ºC. From the streaked ancestor, 96 random isolated colonies were selected after 48 hours and used to found the 96 gene conversion lines. Lines were cultured six to a single YPD plate and bottlenecked by randomly selecting one isolated colony per line every 48 hours and transferring to a new plate. Lines were passaged for a total of 100 transfers (200 days). Every ten transfers, a copy of each mutation accumulation line was plated on YPG to screen for the petite mutation while a random colony from each line was frozen and stored in 15% glycerol at -80ºC. Pictures of each line were taken every ten generations and used in conjunction with standard curves to estimate colony size, in terms of numbers of cells, at the time of transfer. Colony number estimates were then used to estimate the effective population size of the average MA line.

Library Preparation and Whole Genome Sequencing

Mutation accumulation lines were cultured from frozen T100 stock on solid YPD medium at 30ºC for 48h. A single colony from each line was selected, inoculated into 3mL of liquid YPD medium (1% yeast extract, 2% peptone, 2% dextrose), and incubated on a rotator at 30ºC for 48h. Cells were then pelleted and DNA was extracted using the YeaSTAR kit (Zymo Research) protocol with chloroform and an extended digestion time with zymolase overnight at 37C.

Genomic libraries were constructed using a standard protocol. Purification

between steps was performed using magnetic purification beads (MPB) coupled with 80% ethanol washes. To purify, MPB were mixed with samples at a volume ratio of 1.4:1 prior to adapter ligation and 1.0:1.0 post adapter ligation. To wash, MPB solution was mixed with samples by pipetting and left at room temperature for 10 minutes. Afterwards, a magnet was used to isolate magnetic beads from the solution. Supernatant was removed and beads were washed with 2 cycles of 80% ethanol. DNA was eluted off of beads into 10mM Tris-HCl pH 8.0. and left at room temperature for 10 minutes prior to the next step.

1µg of genomic DNA was sonicated in 130µl nuclease free water to generate 500bp fragments using a Covaris S-2 sonicator (Covaris Inc.). Fragmented ends from sonication were repaired using End-It DNA End-Repair Kit (Epicentre) according to manufacturer's instructions. A-tails were added to blunt ended fragments during a 30-minute, 37ºC incubation using Klenow 3´-5´ exonuclease and dA-Tailing Buffer (New England Biolabs). Indexed oligonucleotides were ligated via 16 hour, 16ºC incubation with T4 DNA ligase and ligation buffer (New England Biolabs). Post-ligation, samples underwent two "clean-up" procedures. 5% volume of post-ligated DNA was amplified using Phusion HF Polymerase (New England Biolabs) according to manufacturer's instructions for 8 cycles of PCR. The libraries (93 haploid MA lines, 2 haploid parent samples, and 1 mated diploid ancestor sample) were then compiled into three pools of genomic libraries and run separately on an Illumina NextSeq 500 machine.

Quality Control, Mapping, and Identification of Mutations

Sequence reads from each library were quality controlled with the ea-utils and fastx toolkit software in order to remove low quality reads and residual adaptor sequence

(Commands in Appendix) (40, 41). Following a standard workflow (27), QCed reads were then mapped to the *S. cerevisiae* S288c reference genome R64 (42) with BWA v1.1.2 (23), sorted and indexed with SAMtools v1.0 (23), and assigned line identification numbers with Picard Tools v1.87. Duplicated reads were marked with Picard Tools and removed, and then the remaining sequence reads were locally realigned with GATK v3.2.2 (24). Individual files comparing SNM and indel variants for each line to the ancestor were identified using GATK's Unified Genotyper tool. In order to call a variant, a minimum coverage of 10 reads was required. VCF files comparing the ancestor to each gene conversion line were then examined to detect loss of heterozygosity suggesting a gene conversion event had occurred.

References

1. de Massy B (2013) Initiation of meiotic recombination: how and where? Conservation and specificities among eukaryotes. *Annual review of genetics* 47:563-599.
2. Petes TD, Malone RE, & Symington LS (1991) 8 Recombination in Yeast. *Cold Spring Harbor Monograph Archive* 21:407-521.
3. Haber JE (2000) Partners and pathways: repairing a double-strand break. *TRENDS in Genetics* 16(6):259-264.
4. Pâques F & Haber JE (1999) Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. *Microbiology and molecular biology reviews* 63(2):349-404.
5. Galtier N & Duret L (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *TRENDS in Genetics* 23(6):273-277.
6. Gresham D*, et al.* (2008) The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS genetics* 4(12):e1000303.
7. Gerstein AC, Kuzmin A, & Otto SP (2014) Loss-of-heterozygosity facilitates passage through Haldane's sieve for Saccharomyces cerevisiae undergoing adaptation. *Nature communications* 5.
8. Zeyl C, Vanderford T, & Carter M (2003) An evolutionary advantage of haploidy in large yeast populations. *Science* 299(5606):555-558.
9. Lamour KH*, et al.* (2012) Genome sequencing and mapping reveal loss of heterozygosity as a mechanism for rapid adaptation in the vegetable pathogen Phytophthora capsici. *Molecular Plant-Microbe Interactions* 25(10):1350-1360.
10. Forche A*, et al.* (2011) Stress alters rates and types of loss of heterozygosity in Candida albicans. *MBio* 2(4):e00129-00111.
11. Schoustra SE, Debets AJ, Slakhorst M, & Hoekstra RF (2007) Mitotic recombination accelerates adaptation in the fungus Aspergillus nidulans. *PLoS genetics* 3(4):e68.
12. Mitchel K, Zhang H, Welz-Voegele C, & Jinks-Robertson S (2010) Molecular structures of crossover and noncrossover intermediates during gap repair in yeast: implications for recombination. *Molecular cell* 38(2):211-222.
13. Lee PS*, et al.* (2009) A fine-structure map of spontaneous mitotic crossovers in the yeast Saccharomyces cerevisiae. *PLoS genetics* 5(3):e1000410.
14. Judd SR & Petes T (1988) Physical lengths of meiotic and mitotic gene conversion tracts in Saccharomyces cerevisiae. *Genetics* 118(3):401-410.
15. Yim E, O'Connell KE, Charles JS, & Petes TD (2014) High-Resolution Mapping of Two Types of Spontaneous Mitotic Gene Conversion Events in Saccharomyces cerevisiae. *Genetics* 198(1):181-192.
16. Llorente B, Smith CE, & Symington LS (2008) Break-induced replication: what is it and what is it for? *Cell cycle* 7(7):859-864.
17. Lesecque Y, Mouchiroud D, & Duret L (2013) GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. *Molecular biology and evolution* 30(6):1409-1419.
18. Marais G (2003) Biased gene conversion: implications for genome and sex

evolution. *TRENDS in Genetics* 19(6):330-338.
19. Marsolier-Kergoat M-C (2013) Models for the Evolution of GC Content in Asexual Fungi Candida albicans and C. dubliniensis. *Genome biology and evolution* 5(11):2205-2216.
20. Yin Y & Petes TD (2013) Genome-wide high-resolution mapping of UV-induced mitotic recombination events in Saccharomyces cerevisiae. *PLoS genetics* 9(10):e1003894.
21. Arbeithuber B, Betancourt AJ, Ebner T, & Tiemann-Boege I (2015) Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences* 112(7):2109-2114.
22. Liti G*, et al.* (2009) Population genomics of domestic and wild yeasts. *Nature* 458(7236):337-341.
23. Li H*, et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
24. McKenna A*, et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20(9):1297-1303.
25. Yang SS, Yeh E, Salmon E, & Bloom K (1997) Identification of a mid-anaphase checkpoint in budding yeast. *The Journal of cell biology* 136(2):345-354.
26. Chua P & Jinks-Robertson S (1991) Segregation of recombinant chromatids following mitotic crossing over in yeast. *Genetics* 129(2):359-369.
27. Zhu YO, Siegal ML, Hall DW, & Petrov DA (2014) Precise estimates of mutation rate and spectrum in yeast. *Proceedings of the National Academy of Sciences* 111(22):E2310-E2318.
28. Kvitek DJ, Will JL, & Gasch AP (2008) Variations in stress sensitivity and genomic expression in diverse S. cerevisiae isolates. *PLoS genetics* 4(10):e1000223.
29. Guijo S, Mauricio J, Salmon J, & Ortega J (Determination of the relative ploidy in different Saccharomyces cerevisiae strains used for fermentation and'flor'film ageing of dry sherry-type wines.
30. Bakalinsky AT & Snow R (1990) The chromosomal constitution of wine strains of Saccharomyces cerevisiae. *Yeast* 6(5):367-382.
31. Zierhut C & Diffley JF (2008) Break dosage, cell cycle stage and DNA replication influence DNA double strand break response. *The EMBO journal* 27(13):1875-1885.
32. Ira G*, et al.* (2004) DNA end resection, homologous recombination and DNA damage checkpoint activation require CDK1. *Nature* 431(7011):1011-1017.
33. Aylon Y, Liefshitz B, & Kupiec M (2004) The CDK regulates repair of double-strand breaks by homologous recombination during the cell cycle. *The EMBO journal* 23(24):4868-4875.
34. Orr-Weaver TL & Szostak JW (1983) Yeast recombination: the association between double-strand gap repair and crossing-over. *Proceedings of the National Academy of Sciences* 80(14):4417-4421.
35. Durkin SG & Glover TW (2007) Chromosome fragile sites. *Annu. Rev. Genet.* 41:169-192.
36. Song W, Dominska M, Greenwell PW, & Petes TD (2014) Genome-wide high-

resolution mapping of chromosome fragile sites in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*:201406847.

37. Takeuchi Y, Horiuchi T, & Kobayashi T (2003) Transcription-dependent recombination and the role of fork collision in yeast rDNA. *Genes & development* 17(12):1497-1506.
38. Pelechano V, Chávez S, & Pérez-Ortín JE (2010) A complete set of nascent transcription rates for yeast genes. *PLoS One* 5(11):e15442.
39. Joseph SB & Hall DW (2004) Spontaneous mutations in diploid Saccharomyces cerevisiae more beneficial than expected. *Genetics* 168(4):1817-1825.
40. Aronesty E (2011) ea-utils: Command-line tools for processing biological sequencing data.
41. Gordon A & Hannon G (2010) Fastx-toolkit. *FASTQ/A short-reads preprocessing tools (unpublished)* http://hannonlab. *cshl. edu/fastx_toolkit*.
42. Engel SR, *et al.* (2014) The reference genome sequence of saccharomyces cerevisiae: then and now. *G3: Genes| Genomes| Genetics* 4(3):389-398.

Table 3.1: **Aneuploidy observed in MA lines**

Chromosome	3n Lines	1n lines
I	7, 18, 21	11
II		
III		
IV		
V	4, 40, 49, 50, 82, 83	
VI		
VII	31, 59, 61, 66, 79, 93	
VIII		
IX	47, 76	
X	66	
XI		
XII	1, 13, 18, 27, 44, 53, 77	
XIII		
XIV	33, 76	
XV	11	
XVI	8, 15, 25, 31, 33, 38, 40, 47, 56, 69	

Table 3.2: **Summary of LOH events by class.**

Class	# of LOH Events	Conversion to PA	Conversion to IN	Bases Covered	Average Event Length (bp)	Median Event Length (bp)	Minimum Event Length (bp)	Maximum Event Length (bp)
A	39	20	19	4600757	117968.13	79296	1487	569061
B	219	110	109	7046447	32175.56	6189	27	551879
C	12	5	7	895622	74635.17	12190	1079	260301
D	4	3	1	879599	219899.75	211672.5	38099	418155
Total	274			13422425	48986.95	7949		

Figures

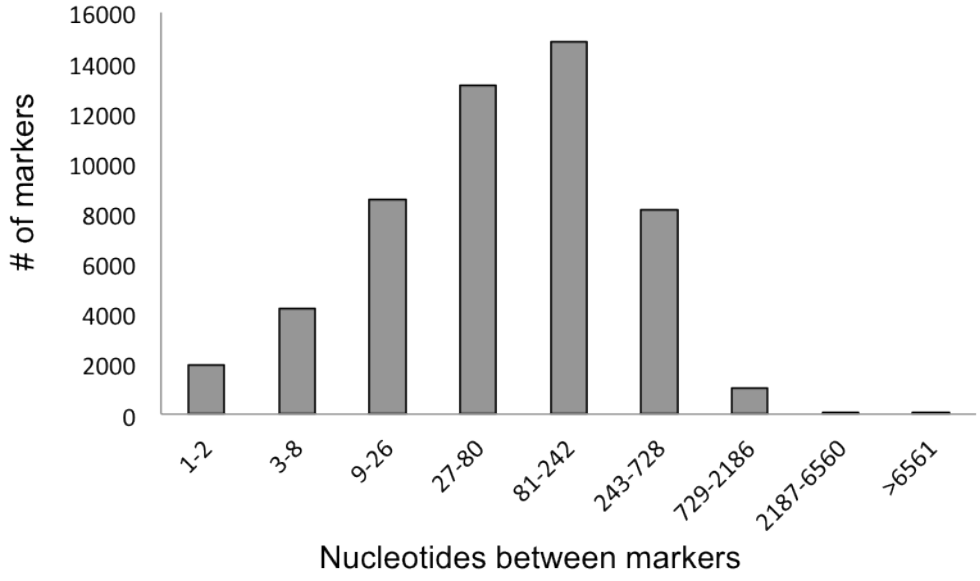

Figure 3.1: **Distribution of nucleotides between heterozygous markers in the diploid ancestor.** Total number of markers is 51895.

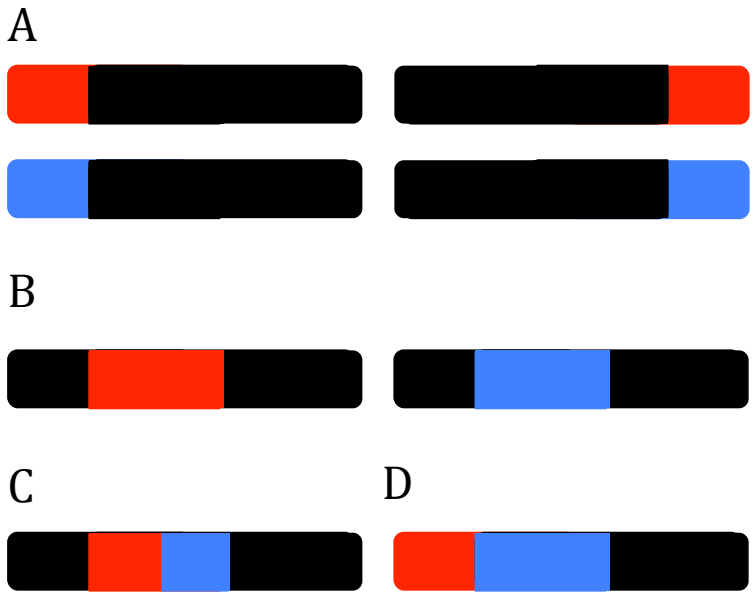

Figure 3.2: **Patterns of LOH may be created multiple ways.** Black represents heterozygous regions while blue and red represent regions converted from parent D or Y respectively. A) LOH events that appear to extend to the end of the chromosome are consistent with the pattern expected from crossover conversion or break induced repair events. B) LOH events contained by heterozygous regions are consistent with the pattern expected from synthesis-dependent strand annealing or double Holiday junction. C) Two adjacent LOH events seemingly from different parents surrounded by heterozygous regions are consistent with the pattern expected from a reciprocal crossover preceding synthesis-dependent strand annealing or double Holiday junction. D) Two adjacent LOH events seemingly from different parents extending to the end of the chromosome on one side and flanked by heterozygous markers are consistent with the pattern expected from a synthesis-dependent strand annealing or double Holiday junction preceding a crossover recombination or break induced repair event.

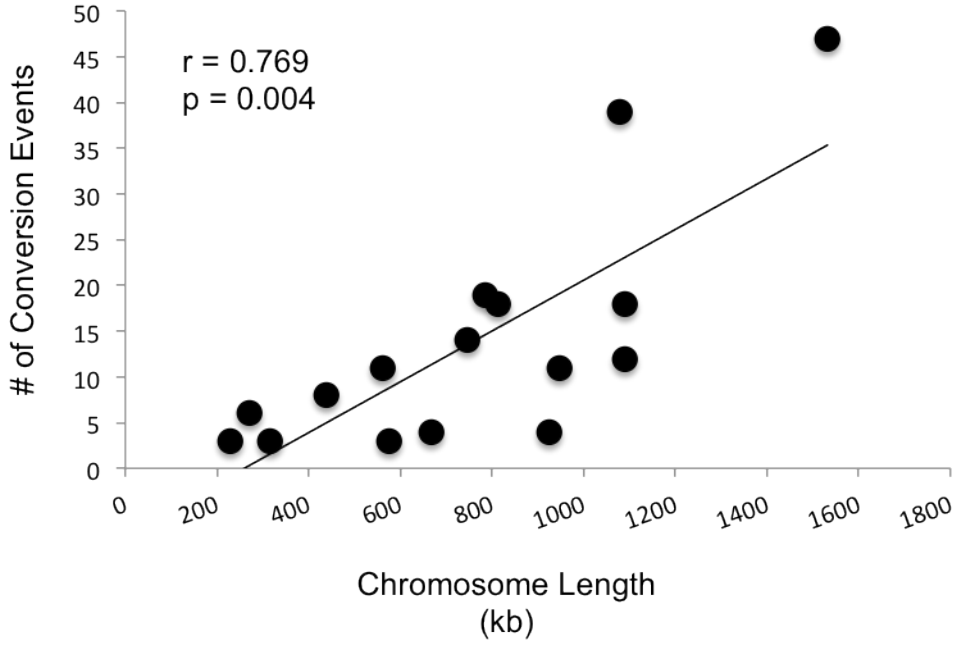

Figure 3.3: **Number of gene conversion events is positively correlated with chromosome length**.

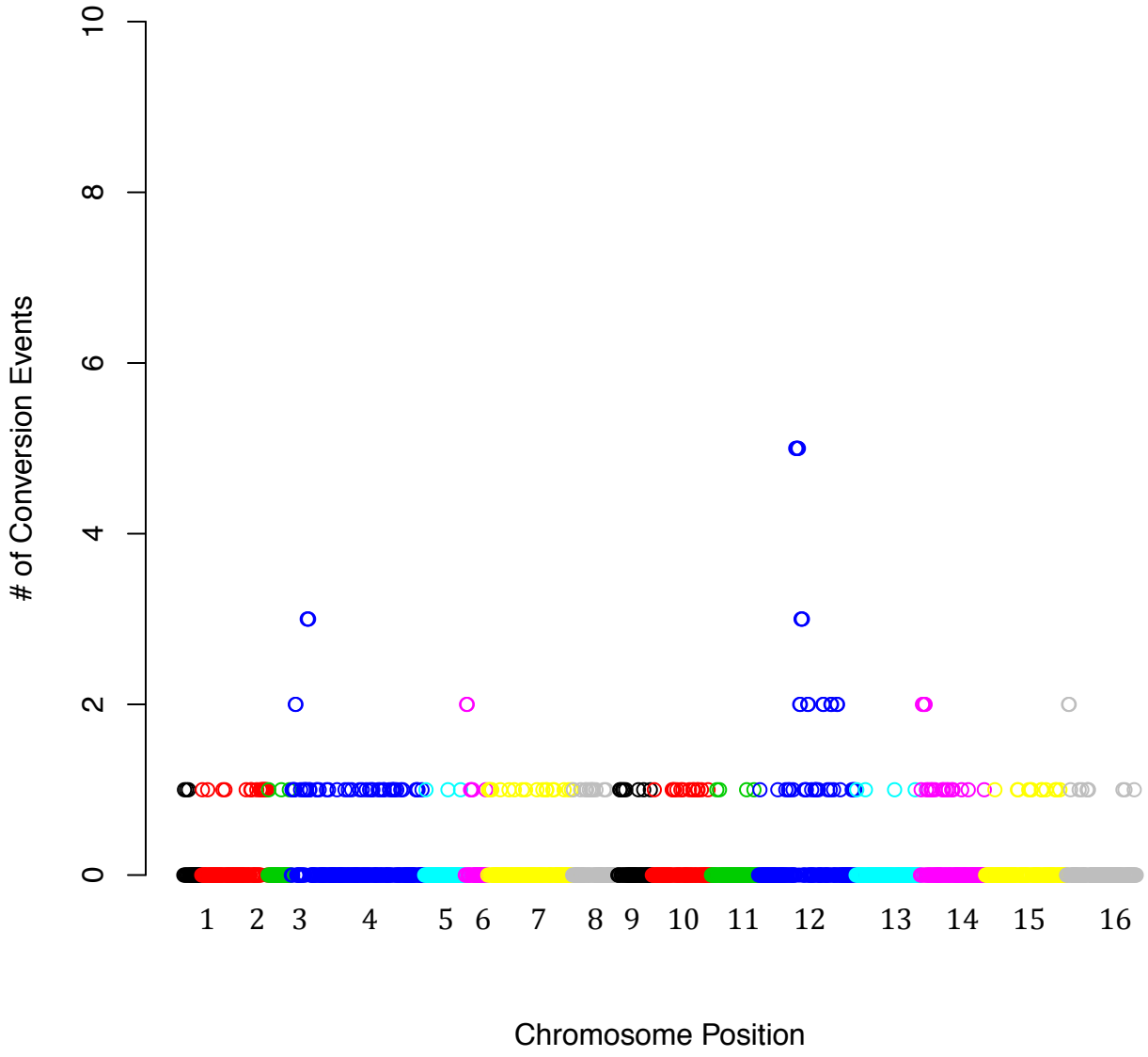

Figure 3.4: **Gene conversion hot spots exist mainly on chromosome IV and XII**. Number of gene conversion events mapped to any specific base. Colors indicate different chromosomes.

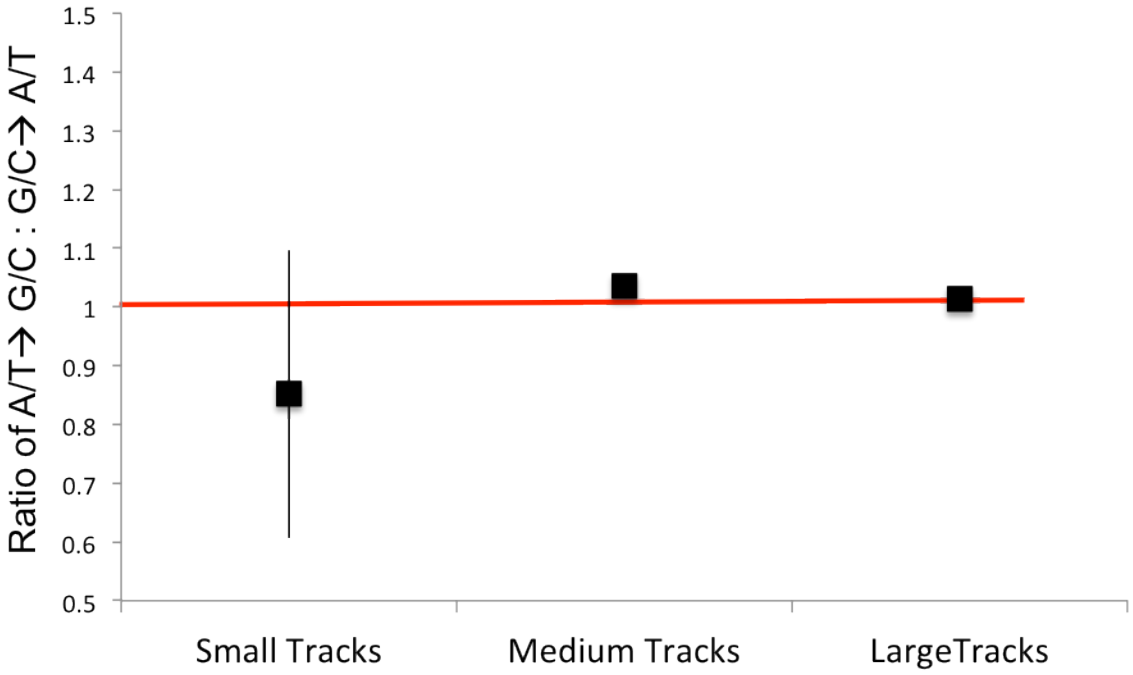

Figure 3.5: **Degree of G/C bias relative to length of gene conversion tracts**. Tract bin sizes were determined by distribution of equal number of conversion events.

CHAPTER 4

SELECTION ON NONSENSE CODONS IN INTRONS[1]

[1] Behringer, M.G. and D. W. Hall. To be submitted to *G3: Genes, Genomes, Genetics*.

Abstract

Introns occasionally remain in mature mRNAs due to splicing errors and the translated, aberrant proteins that result represent a metabolic cost and may also have deleterious functions. The nonsense mediated decay pathway degrades aberrant mRNAs, which it recognizes by the presence of an in-frame premature termination codon. We investigated whether selection has shaped the location and context of premature termination codons in introns to reduce waste and facilitate nonsense mediated decay in seven model organisms. We found that premature termination codons occur earlier in introns than expected by chance, suggesting that selection favors earlier position. This pattern is more pronounced in species with larger effective population sizes. The pattern does not hold in mammals, perhaps because of the different manner in which premature termination codons are used to detect aberrant mRNAs for decay. We also found evidence suggesting that selection acts on premature termination codon context. There is an elevated frequency of a purine nucleotide in mammals, immediately downstream of the first premature termination codon, similar to true termination codons. We conclude that both the location and context of premature termination codons are shaped by selection for reduced waste and efficient degradation of aberrant mRNAs.

Introduction

It is clear that selection can play a major role in shaping genome architecture (Lynch 2007). Much of the work to date has focused on exons in protein coding genes, where tests based on silent and replacement site substitutions can be employed. However, in noncoding regions, the role of selection in shaping nucleotide content is less easily investigated because of the difficulty identifying expected patterns. In introns, length,

phase and frequency of occurrence have been studied as a product of the processes of selection and drift (Castillo-Davis, et al. 2002; Lynch 2002; Whitney and Garland Jr 2010; Kelkar and Ochman 2012) but, apart from sequence-based splicing signals and GC content (Mount 1982; Deutsch and Long 1999; Amit, et al. 2012; Farlow, et al. 2012), few studies have examined the nucleotide composition of introns (Lim and Burge 2001; Halligan, et al. 2004; Andolfatto 2005). In this study, we investigate the role of selection in determining the position and nucleotide context of premature termination codons (PTCs) within introns.

During post-transcriptional processing, splicing errors can result in introns being present in mature mRNAs (Gilbert 1978). Translation of such mRNAs results in the production of proteins that are aberrant in amino acid sequence and usually shortened. A shorter protein results from the presence of in-frame, premature termination codons (PTCs) within the unspliced intron, or in a downstream exon due to a frame-shift. Aberrant proteins may reduce fitness because they have an activity that is damaging to the cell, and/or because they represent wasted resources, particularly amino acids and sequestered ribosomes (Drummond and Wilke 2009). For these reasons, we hypothesize that selection favors both efficient splicing and mechanisms that minimize the effects of splicing errors.

There is strong evidence for selection acting on both the efficiency of splicing and on the effects of splicing errors. Splicing site consensus sequences are highly conserved, indicating they are constrained by selection for efficient splicing (Sheth, et al. 2006). In addition, splicing efficiency has been shown to be a function of intron length. Transcripts containing large introns are more dependent on 5' and 3' splicing context to define exons,

while splice context in shorter introns appears to be more lax (Jaillon, et al. 2008; Farlow, et al. 2012). The effects of splicing errors are minimized by the nonsense mediated decay (NMD) pathway, which is believed to have evolved for this reason (Jaillon, et al. 2008). The NMD pathway exists in all eukaryotes and causes rapid decay of mature mRNAs that possess PTCs (Baker and Parker 2004). In addition to NMD, there is also evidence that the occurrence of PTCs can be selected. For example, in *Paramecium tetraurelia*, there is evidence that PTCs are more common in introns whose length is a multiple of three (Jaillon, et al. 2008).

In this study our goal is to test two hypotheses. First, we hypothesize that PTCs will be selected to occur early in introns to reduce both the time taken for a ribosome to translate an aberrant mRNA containing unspliced introns and the length of the resulting aberrant proteins. Second, we hypothesize that the RNA sequence immediately downstream of a PTC will be selected for efficient translation termination. It has been shown that the 3'UTR sequence immediately following a true termination codon (TTC) alters termination efficiency. Specifically, the presence of a purine nucleotide in the first position immediately following a TTC increases translational termination efficiency in eukaryotes (Brown, et al. 1990). We hypothesize that efficient termination at PTCs will lead to more efficient entry into the NMD pathway, which will be favored by natural selection. We thus expect selection to favor a purine nucleotide in the first position following a PTC, similar to TTCs. Since selection is caused by the deleterious effects of having aberrantly spliced mRNAs, we expect our predictions to be more likely to hold for genes that are expressed at high levels, which might be expected to produce more aberrant mRNAs, and short introns, which seem to be less-effectively spliced (Talerico

and Berget 1994). To test whether there is evidence that selection has acted to minimize the deleterious effects of failure to remove an intron from a mRNA by modifying the location of and sequence surrounding the first PTCs, we utilized data from seven model organisms. These species have well-annotated genomes, reliable splicing information and expression data, and represent all of the species with the necessary data to test our predictions. We examined both the distance between the 5' splice site and the first PTC, as well as the identity of the base immediately following the first PTC. Our analysis focused on a single intron in each gene so that each intron-containing gene contributed equally to the data set. We chose the first intron because it exists in all intron-containing genes. For some analyses, in species containing genes with more than one intron we also analyzed the last intron because NMD is not initiated from the last intron in mammals (Maquat 2005) and thus we did not expect to see an elevated purine frequency in the first position following the first PTC in the last intron in mammals. We looked for evidence for first PTCs to occur earlier than expected, and for the first base following a PTC to more likely be a purine, and determined whether there was an effect of level of expression or intron length on the patterns observed.

If incorrect splicing is varies in error rate (Wilhelm, et al. 2008; Drummond and Wilke 2009; Fox-Walsh and Hertel 2009) and NMD is reasonably effective (though not all genes are NMD sensitive in yeast (Sayani, et al. 2008)), then selection on PTC position and termination efficiency caused by the presence of aberrant mRNAs may not be too strong. We thus expect our predictions be less likely to hold in species with smaller effective population sizes, as has been observed for other genomic features (Lynch and Conery 2003).

Materials and methods

Data Collection

Genome data were collected for *Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Homo sapiens. Mus musculus, Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe* as GenBank files from NCBI (*Caenorhabditis elegans* Sequencing Consortium 1998; Lin, et al. 1999; Mayer, et al. 1999; Adams, et al. 2000; Erfle, et al. 2000; Tabata, et al. 2000; Theologis, et al. 2000; Lander, et al. 2001; Wood, et al. 2002; Engel, et al. 2013). Genomes were parsed using Feature Extract to collect all genes with annotated introns (Wernersson 2005)(Table 4.1). Files were sorted to remove duplicate genes and genes with alternative splicing, because sequence data for the first intron, terminal intron, and 3' UTRs of these genes were often identical. Our goal was to have each intron-containing gene contribute once to the data set. The effective population sizes for each species were obtained from previously published reports and are based on estimates from nuclear synonymous sites (Schoen and Brown 1991; Chen and Li 2001; Wright, et al. 2002; Sivasundar and Hey 2003; Cutter 2006; Shapiro, et al. 2007; Wright and Andolfatto 2008; Skelly, et al. 2009; Brown, et al. 2011; Phifer-Rixey, et al. 2012). Expression data were collected for each organism from previously published data, and genes not represented by expression data were removed from the study (*A. thaliana*,(Carviel, et al. 2009); *C. elegans*, Michael Smith Genome Sciences Centre (http://www.bcgsa.ca); *D. melanogaster*, FlyAtlas (http://flyatlas.org; (Chintapalli, et al. 2007)); *H. sapiens* (Dezső, et al. 2009); *S. cerevisiae* (Pelechano, et al. 2010); *Sc. pombe*, (Tanizawa, et al. 2010)). Remaining genes (Table 4.1) were then used for analysis. For most analyses, only the first intron was examined in each gene both to prevent pseudo-

replication across the same gene and because many genes contain only a single intron, especially in *S. cerevisiae* where 229 out of 241 intron-containing genes have only one intron.

Identification of first PTC, and other nonsense codon positions

Using the annotation data for each organism, we determined the length of the first intron for each gene. We then indexed the position of the first 5' splice site, the position of the first PTC (frame 0), and the first intronic nonsense codon (NC) in the +1 and +2 reading frames for each gene using custom Perl scripts. The distance between the 5'splice site and both the PTC and intronic NCs, was determined for each gene and then assigned to 30-nucleotide bins (equivalent to the coding sequence for 10 amino acids). Differences between the in-frame and out-of-frame distributions was determined using the Kolmogorov-Smirnoff test.

We also identified the last introns for the *H. sapiens*, *M. musculus*, *A. thaliana*, *C. elegans*, and *D. melanogaster* genes that contained at least two introns. This was done by reversing the annotation and sequences of all genes using custom Perl scripts and counting backwards to determine phase and extract the final introns and exons. Once extracted, the final introns and exons were returned to the proper orientation and PTC position and context was identified as before. We manually checked that the predicted and actual PTC positions were identical in a handful of genes with different phases of their last introns to make sure the scripts worked as expected.

In addition, we compared the position of the PTC to the length of the first intron, as well as the position of the PTC to gene expression for all species. Intron length was divided into two groups: short introns, consisting of the shortest quartile of introns in

each organism; and long introns, consisting of all other remaining introns. Gene expression was also divided into two groups: highly expressed genes, consisting of the top 100 expressed genes, similar to Castillo-Davis et al. (2002); and medium/low expressed genes, consisting of all other remaining genes, except in *Sc. pombe* where due to sample size, a natural break in the data were used to determine high/low gene expression. We tried other cut-offs for gene expression, including the upper 10% and upper quartile, but results were unaffected (data not shown). Position of PTCs and frame +1 and +2 intronic NCs was compared in short and long introns and high and low expressed genes using the Mann-Whitney U test.

Comparison of Termination Codon 3' Context

We first identified the nucleotide immediately following the first PTCs and frame +1 and +2 intronic NCs. To determine whether purines are more frequent than expected, we determined the frequency following nonsense trinucleotides that are present in any orientation or frame within first introns in each species. For comparison, we also determined the purine frequency after the second intronic PTCs in genes where they exist. Difference between purine percentages was then determined using z test for proportions.

Results

In seven model species, we analyzed protein-coding genes that met the following criteria: they (1) contained at least one intron, (2) could be matched to published expression data, (3) were not reported to have alternative spliceforms and (4) were not duplicates according to annotation data. For the seven species examined, there were between 161 (*Saccharomyces cerevisiae*) and 10,869 genes (*Caenorhabditis elegans*) that

met these criteria (Table 1). Throughout the rest of this study we distinguish in-frame (Frame 0) termination codons, PTCs, from out-of-frame (frame +1 and +2) nonsense codons, NCs (TAA, TAG or TGA), as the latter are unable to cause translation termination.

Location of in-frame premature termination codons

The position of the first PTC, and the first intronic frame +1 or +2 NC, relative to the 5' splice site was determined for the first intron of each gene. Since out-of-frame NCs do not terminate translation, they should not be visible to selection and their position thus controls for the pattern of nucleotide composition within introns. We determined whether the first PTCs occur earlier than the first out-of-frame NCs. For all five non-mammal species, first PTCs appear significantly earlier compared to first out-of-frame NCs. In contrast, PTC position in mammals is significantly later than first out-of-frame NCs (Table S1, Fig. 1). Additionally, in *S. cerevisiae* when we examine only the 27 genes in our dataset classified as NMD sensitive (Sayani, et al. 2008), the fold increase of PTCs that occur within the first 30 nucleotides of the intron is greater in Frame 0 (1.905) than the fold increase of PTCs for NMD insensitive (1.453) or all *S. cerevisiae* genes (1.512) datasets (Appendix C).

In mammals, entry into the NMD pathway is different than in non-mammals (see Discussion), and PTCs in the last intron generally are not able to trigger NMD. To determine if the position of the first PTCs is affected by recruitment of the NMD pathway, we performed the same position analysis in terminal introns. *Sc. pombe* and *S. cerevisiae* have very few genes containing multiple introns and were thus not included in this analysis. For the remaining species, the first PTC did not occur earlier that out-of-frame

NCs; instead NCs in Frame 1 occur earlier in the first 30 nucleotides than PTCs and Frame 2 NCs. (Appendix C).

It is possible that the conserved splice consensus at the 5' end of the intron affects PTC position. Except for *S. cerevisiae*, all of the species examined contain a stop-codon like trinucleotide (TGA or TAA) within the 5' splice consensus G**TRA**GT. Phase 2 introns, which are those in which the first base of the intron begins, if unspliced, in the third position of a codon, would have a PTC in the splice consensus sequence. Phase 1 (unspliced first intron base is second position in a codon) and phase 0 (unspliced first intron base is first position in a codon) introns would both have an out-of-frame NC in the splice consensus sequence. If a majority of introns are phase 2, this would make PTCs appear relatively earlier. We thus repeated our analysis discounting phase/frame combinations that introduce splice site PTCs and out-of-frame NCs. To do this we reduced the data set by examining the position of PTCs in phase 0 and phase 1 introns and compared them to the position of out-of-frame NCs in phase 0, frame +2 and phase 1, frame +1 introns respectively and found in non-mammals the pattern of earlier PTC position relative to out-of-frame NC position in first introns did not change (Figure 4.2, Appendix C). In contrast, in mammals the position of the first PTC was no longer significantly later than the first out-of-frame NC (Figure 4.2, Appendix C). In last introns, after controlling for stop codons introduced by the splice site consensus sequence, PTCs occurred significantly earlier in *C. elegans* and *D. melanogaster*, significantly later in *A. thaliana*, and was not different from NCs in *H. sapiens*, and *M. musculus* (Figure 4.3, Appendix C).

Premature termination codon position is affected by intron length but not gene expression

Previous work has shown that highly expressed genes tend to have short introns (Castillo-Davis, et al. 2002). In addition, longer introns appear to experience greater selective constraint than shorter introns (Farlow, et al. 2012). We examined PTC position as a function of intron length and found that across species PTCs occur closer to the 5' splice site in shorter introns (Figure 4.4; Mann-Whitney U test: for all species $p < 10^{-12}$). This relationship is not driven by gene expression: PTC position did not differ in highly expressed genes in any of the species (Appendix C; Mann-Whitney U test: *A. thaliana*, p = 0.756; *C. elegans*, p = 0.253; *H. sapiens*, p = 0.486; *M. musculus*, p = 0.680; *D. melanogaster*, p = 0.016; *Sc. pombe*, p = 0.819).

Evidence of selection on termination context cannot be separated from splice site sequences

The 3' nucleotide adjacent to a nonsense codon has been shown to be important for efficient termination in *S. cerevisiae* (Namy, et al. 2001). Additionally, in other eukaryotes, when a purine nucleotide immediately follows a nonsense codon, it results in high efficiency termination (Brown, et al. 1990). For all species except *A. thaliana* and *S. cerevisiae*, we found that in first introns, the frequency of purine nucleotides in the first position following the first PTCs is significantly higher than expected (z-test for proportions: *C. elegans*, *H. sapiens*, *M. musculus* and *D. melanogaster*: $p < 10^{-5}$; *A. thaliana*: p = 0.646, *S. cerevisiae*: p = 0.491; Table 4.2, Appendix C). In contrast, the frequency of purines following the second PTC is not significantly different from

expected (z-test for proportions: *C. elegans*: p = 0.077; *H. sapiens*: p = 0.114; *M. musculus*: p = 0.190; *D. melanogaster*: p = 0.119; *Sc. pombe*: p = 0.555; Table 4.2, Appendix C).

Although *S. cerevisiae* (Appendix C), has a purine frequency after the first PTC that is not significantly different than expected, when we examined only the 27 genes classified as NMD sensitive (Sayani, et al. 2008), the frequency of a purine in the first position after PTCs is greatly increased compared to the expectation, though it is still not significant (z = 1.55, p = 0.121; Appendix C). We also examined PTC context in the last intron in all species except *S. cerevisiae* and *Sc. pombe*, to determine if purine enrichment behind PTCs is caused by other functionally important features nested within first introns (Majewski and Ott 2002). For all species, we found that the frequency of purine nucleotides in the first position following the first PTCs of last introns is also significantly higher than expected (*A. thaliana*: z = 3.37, p = 7.6 x 10^{-4}; *C. elegans*: z = 2.67, p = 0.008; *H. sapiens*: z = 7.77, p < 10^{-5}; *M. musculus*: z = 6.44, p < 10^{-5}; *D. melanogaster*: z = 6.14, p < 10^{-5}; Table 4.2, Appendix C). The purine frequency following the second in-frame PTC in last introns is either significantly lower or not significantly different than expected in all species (*A. thaliana*: z = 0.71, p = 0.478; *C. elegans*: - z = 5.58, p < 10^{-5}; *H. sapiens*: z = 1.85, p = 0.064; *M. musculus*: z = 0.72, p = 0.472; *D. melanogaster*: z = -1.71, p = 0.087; Table 4.2, Appendix C).

The 5' splice consensus GTRAGT introduces a guanine nucleotide behind the TRA stop codons. To examine whether the splice site is driving the observed pattern of higher frequency of purine nucleotides following PTCs, we repeated our analysis discounting phase/frame combinations that introduce splice site PTCs, in the same

manner as our previous analysis of PTC position. In non-mammals, we found that the frequency of purine nucleotides in the first position following the first PTC is either significantly lower (z-test for proportions: *A. thaliana*: z = -3.84, p = 1.2 x 10^{-4}, *C. elegans*: z = -5.33, p < 10^{-5}) or not significantly different (*D. melanogaster*: z = -0.91, p = 0.363; *Sc. pombe*: z = 1.15, p = 0.250; *S. cerevisiae*: z = -0.888, p = 0.373; Table 4.3, Appendix C) than expected in the first intron. The frequency of purines following the second in-frame PTC in the first intron is not statistically different from the expectation (*A. thaliana*: z = -0.38, p = 0.704; *C. elegans*: z = -1.88, p = 0.060; *D. melanogaster*: z = -1.80, p = 0.072; *Sc. pombe*: z = 0.62, p = 0.535; Table 4.3, Appendix C).

In mammalian first introns, however, after correcting for the splice site, the frequency of purine nucleotides in the first position following the first PTC remains significantly greater than expected (z-test for proportions: *H. sapiens*: z = 6.05, p < 10^{-5}; *M. musculus*: z = 7.43, p < 10^{-5}; Figure 4.4), while the frequency following the second PTC is not different (*H. sapiens*: z = 1.85, p = 0.064; *M. musculus*: z = -1.94, p = 0.052; Table 4.3, Appendix C).

Again, we examined purine frequency within terminal introns. Similar to the first intron, when we removed the effect of the 5' splice sequence, we found that for all organisms, the frequency of purine nucleotides in the first position following the first PTC is either significantly lower or not significantly different than expected (*A. thaliana*: z = -1.95, p = 0.051; *C. elegans*: z = -5.24, p < 10^{-5}; *H. sapiens*: z = 1.31, p = 0.190; *M. musculus*: z = -0.99, p = 0.322; *D. melanogaster*: z = -3.20, p = 0.001; Table 4.3, Appendix C). This pattern is also observed following the second PTC (*A. thaliana*: z = 1.10, p = 0.271; *C. elegans*: z = -5.40, p < 10^{-5}; *H. sapiens*: z = 2.09, p = 0.037; *M.*

musculus: z = 0.83, p = 0.407; *D. melanogaster*: z = -1.28, p = 0.20; Table 4.3, Appendix C).

Correlations with effective population size

Several genomic patterns are often more apparent in species with larger effective population sizes, where selection is more effective (Lynch and Conery 2003). For those five species that exhibit significantly early first PTC position, the logarithm of the ratio of median PTC nucleotide position to median first NC nucleotide position is negatively correlated with the effective population size (df = 4, Pearson's r = -0.82, p = 0.089). This pattern becomes statistically significant after the effect of splice site is removed (df = 4, Pearson's r = -0.98, p = 0.004; Figure 4.5). Thus, for species that show early position of the first PTC, those with larger effective population sizes have earlier PTCs in their first intron.

A similar pattern is seen with effective population size and purine frequency in the 3' position following the first PTC across the five species that are significantly elevated. The logarithm of the ratio of purine frequency after the first PTC to the expectation of purine frequency after NCs was positively correlated with the log of the effective population size (df = 4, Pearson's r = 0.94, p = 0.017, Appendix C). When the effect of the nonsense codon in the splice site is removed, the correlation is no longer significant (df = 6, Pearson's r = 0.68, p = 0.091), though the slope is similar.

Discussion

We find mixed support for our hypotheses. The prediction that the first PTC occurs early holds in first introns for five of the seven species, even after removing the effects of splice site. However, the two mammalian species show either the opposite

pattern, such that first PTCs are significantly later than random, or no difference in position when the effect of splice site is removed. For the last intron, there is no consistent pattern for the position of the first PTC: it is either not significantly different than random (all five species), or, when the effect of splice site is removed, it is earlier (*C. elegans* and *D. melanogaster*), later (*A. thaliana*) or not different (*H. sapiens* and *M. musculus*) than random.

Our results suggest that selection may favor early translation termination in the five non-mammal species in incorrectly spliced mRNAs that retain the first intron. We postulate that selection favors the reduction of waste caused by splicing errors; PTCs that occur early reduce translation waste. The absence of the pattern in last introns suggests that selection is less effective in terminal introns.

If the cost of translation is driving first PTCs to be early, we expect them to be even earlier in highly expressed genes. However, highly expressed genes do not exhibit earlier first PTCs than other genes. In *Sc. pombe,* splicing efficiency is increased in highly expressed genes so the opportunity for selection on PTCs in introns for those genes is less (Wilhelm, et al. 2008).

We do observe that short introns exhibit earlier PTC position than long introns. This may be due to the difference in splice site paring strategies between short and long introns. While long introns use a exon definition strategy where splicing errors result in exon skipping, short introns use an intron definition strategy where splicing errors result in intron inclusion (Talerico and Berget 1994). Therefore, error in splicing shorter introns is more likely to result in translation of the intron, which explains the stronger selection signature observed in short introns.

Our analyses also indicate that the 3' nucleotide immediately downstream of the first PTC is enriched for purines relative to the second PTC in first introns in five of the seven species, with no significant difference in the other two, *A. thaliana* and *S. cerevisiae*, and in all five species in which the analysis could be performed in last introns. However, after removal of the effect of splice site, only first introns in the two mammalian species exhibit elevated purine frequency. The other species show either no difference or significantly lower frequency of purines (*A. thaliana* and *C. elegans* in first introns and *C. elegans* and *D. melanogaster* in last introns). This finding suggests either that a 3' downstream purine is generally less important for efficient translation termination at PTCs in non-mammals, or that the selection for it is essentially concealed by the presence of the 5' splice site. In mammals, the first PTC is generally further downstream in the intron, and so evidence for selection remains after removing the effects of the 5' splice site. The absence of an elevated 3' purine frequency in the first PTCs in last introns in mammals is not unexpected given that the NMD pathway is not initiated in the last introns in mammals (Maquat 2005).

When our predictions do hold, they are more pronounced in species with larger effective population sizes (Fig. 5 and Fig. S7). Because the frequency of aberrant splicing varies, the strength of selection acting on unspliced transcripts is likely weak, which explains the strong sensitivity to effective population size (Wilhelm, et al. 2008; Drummond and Wilke 2009; Fox-Walsh and Hertel 2009).

Nature of Selection on PTCs

If first PTCs are selected to occur earlier and the 3' downstream nucleotide is selected for efficient termination, as some of our results suggest, then there must be one

or more benefits to early, efficient termination of aberrant mRNAs containing introns. There are several possible costs that might cause selection to favor earlier PTC position. These include: 1) The opportunity cost of not creating enough functional proteins because ribosomes are occupied by aberrant mRNAs encoding non-functional protein products; 2) The energetic cost of elongating aberrant proteins; 3) The energetic cost of breaking down abnormal proteins; and 4) The functional cost of creating proteins that are deleterious to the cell. All of these costs would be reduced by more efficient NMD since aberrant mRNAs would be more rapidly degraded. In addition, earlier ribosome release would reduce both the opportunity cost and the elongation and break down energetic costs of the aberrant proteins, which would select for earlier position of first PTCs.

Of these four possibilities, we suspect that the opportunity cost of ribosomes translating non-functional products and the clean up costs of the abnormal proteins may be the greater forces contributing to early PTC position within introns. The cost to decay an aberrant protein is directly proportional to the amount of aberrant protein produced, and the occupation of aberrant mRNAs by ribosomes, producing non-functional product, represents an opportunity cost to the organism (Drummond and Wilke 2009). Ribosomes translating aberrant mRNAs are unavailable for production of other functional protein products. Selection for efficient translation termination as early as possible in the incorrectly spliced message would minimize the amount of aberrant protein the cell has to decay, as well as the duration that the aberrant transcript occupies the ribosome.

Elongation and sequestering of limited amino acid resources in aberrant product has been hypothesized to be a major energetic cost (Stoebel, et al. 2008). However, studies in *E. coli* show that supplementing amino acids does not change the cost of

protein expression, which suggests that amino acid limitation is therefore not likely a major selection pressure on PTC position or context (Stoebel, et al. 2008).

Finally, while it is clear that creating aberrant proteins that are toxic will be deleterious in terms of fitness, and this is likely the primary selection pressure associated with the evolution of the NMD pathway itself, we suggest that it is less likely to be explanation for the position of PTCs. The reason is that an earlier PTC position only moderately shortens the aberrant protein, which may generally have little or no effect on its toxicity.

Optimal PTC Position May Vary Based on Specifics of NMD Pathway

Given our evidence that selection appears to favor earlier PTCs that occur closer to the 5' splice site, the absence of early PTCs in *H. sapiens* and *M. musculus* requires explanation. One possibility is that the effective population size of these mammals is sufficiently small to render selection ineffectual at shaping PTC position. However, in both these species PTCs occur substantially earlier in short compared to long introns, suggesting selection can be effective. Another possibility is that PTCs are not selected to be early in all introns in mammals because of how mammalian NMD works. In mammals, when the ribosome stalls at the first PTC and the surveillance complex is assembled, instead of interacting with a faux 3' UTR as it does in other eukaryotes, the surveillance complex interacts with the next exon junction complex in order to flag the transcript for decay (Maquat 2005). Because the complex interacts with the next exon junction complex, selection in mammals might be expected to favor the PTC to occur closer to the exon junction complex, increasing the probability of interaction. However, the selective pressure of freeing ribosomes occupied by aberrant RNA transcripts remains, perhaps

creating a selective tug-of-war between freeing ribosomes and terminating translation closer to the next exon junction complex. This might explain a pattern of PTC position that shows no difference from the expectation when averaged across all introns, but is earlier in shorter introns (Figures 1, 3). The shortest quartile of introns in mammals are those less than 998 nucleotides and so the distance to the next exon junction complex may always be small enough that the NMD surveillance complex can interact with the next exon junction, which would allow the selective pressure for freeing ribosomes to predominate. In contrast, in larger introns, the distance to the next exon junction is greater and the selective pressure of the NMD surveillance complex becomes more important. Unfortunately, we do not have data on whether a change in PTC position on the order of tens to hundreds of nucleotides in introns whose length is in the hundreds to thousands of nucleotides would make a difference to NMD initiation. This hypothesis predicts that NMD causes the position of the first PTC to be relatively late in mammals except in genes that are not NMD sensitive, which should exhibit relatively earlier PTC position. While there are no data to distinguish NMD sensitive and nonsensitive genes in mammals, aberrant mRNAs in which only the last intron is unspliced cannot activate the NMD pathway because there is no downstream exon junction. In support of the importance of NMD for causing late PTC position, last introns in mammals do not show later PTC position than expected.

Future directions

The data suggest that both the position of, and the downstream nucleotide immediately following the first PTC may be shaped by selection. Data from additional species is clearly needed to confirm the patterns, especially in species with large effective

population sizes where the patterns are expected to be more robust. In addition, the fact that patterns are more apparent for NMD sensitive than nonsensitive genes in *S. cerevisia*e suggests that distinguishing these two classes of genes in other species would be useful for elucidating the predicted patterns.

One variable that we have not addressed is the frequency with which aberrant mRNAs are produced. The expectation is that genes that routinely produce aberrant mRNAs would experience selection on PTC position and termination efficiency more often. For example, in the extreme case in which a gene is never aberrantly spliced, it would never experience selection on PTCs. Tantalizing data that can be interpreted as supporting this prediction comes from *S. cerevisiae* in which NMD sensitive genes exhibit earlier, first PTC position and greater pruine frequency in the 3' downstream position than NMD insensitive genes (Appendix C). If NMD sensitivity is related to the probability of aberrant splicing, such that genes that produce aberrant mRNAs are more likely to be NMD sensitive, then NMD sensitive genes will also be those with earlier PTCs. Testing the prediction that genes with less efficient splicing will show greater PTC position effects is challenging. The problem is that steady-state levels of aberrant mRNAs, such as would be measured in a standard RNA-seq experiment, are affected by both the production of and the degradation of the aberrant mRNAs (Pelechano, et al. 2010) For this reason, it would be necessary to measure aberrant mRNA production directly, which is a more challenging, and expensive, undertaking.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF. 2000. The genome sequence of *Drosophila melanogaster*. Science 287:2185-2195.

Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, Lev-Maor G, Burstein D, Schwartz S, Postolsky B. 2012. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Reports 1:543-556.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437:1149-1152.

Baker KE, Parker R. 2004. Nonsense-mediated mRNA decay: terminating erroneous gene expression. Current Opinion in Cell Biology 16:293-299.

Brown CM, Stockwell P, Trotman C, Tate W. 1990. Sequence analysis suggests that tetra-nucleotides signal the termination of protein synthesis in eukaryotes. Nucleic Acids Research 18:6339-6345.

Brown WR, Liti G, Rosa C, James S, Roberts I, Robert V, Jolly N, Tang W, Baumann P, Green C. 2011. A Geographically Diverse Collection of *Schizosaccharomyces pombe* Isolates Shows Limited Phenotypic Variation but Extensive Karyotypic Diversity. G3: Genes, Genomes, Genetics 1:615-626.

Caenorhabditis elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282:2018.

Carviel JL, AL-DAOUD F, Neumann M, Mohammad A, Provart NJ, Moeder W, Yoshioka K, Cameron RK. 2009. Forward and reverse genetics to identify genes involved in the age-related resistance response in *Arabidopsis thaliana*. Molecular Plant Pathology 10:621-634.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. Nature Genetics 31:415-418.

Chen F-C, Li W-H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. American Journal of Human Genetics 68:444.

Chintapalli VR, Wang J, Dow JA. 2007. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nature Genetics 39:715-720.

Cutter AD. 2006. Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer *Caenorhabditis elegans*. Genetics 172:171-184.

Deutsch M, Long M. 1999. Intron-exon structures of eukaryotic model organisms. Nucleic Acids Research 27:3219.

Dezső Z, Nikolsky Y, Nikolskaya T, Miller J, Cherba D, Webb C, Bugrim A. 2009. Identifying disease-specific genes based on their topological significance in protein networks. BMC Systems Biology 3:36.

Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. Nature Reviews Genetics 10:715-724.

Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, Dwight SS, Hitz BC, Karra K, Nash RS. 2013. The Reference Genome Sequence of *Saccharomyces cerevisiae*: Then and Now. G3: Genes Genomes Genetics:g3. 113.008995.

Erfle H, Ventzki R, Voss H, Rechmann S, Benes V, Stegemann J, Ansorge W, Zheng L, Cornel A, Wang R. 2000. Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. Nature 408:820-822.

Farlow A, Dolezal M, Hua L, Schlötterer C. 2012. The genomic signature of splicing-coupled selection differs between long and short introns. Molecular Biology and Evolution 29:21-24.

Fox-Walsh KL, Hertel KJ. 2009. Splice-site pairing is an intrinsically high fidelity process. Proceedings of the National Academy of Sciences 106:1766-1771.

Gilbert W. 1978. Why genes in pieces? Nature 271:501.

Halligan DL, Eyre-Walker A, Andolfatto P, Keightley PD. 2004. Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. Genome Research 14:273-279.

Jaillon O, Bouhouche K, Gout J-F, Aury J-M, Noel B, Saudemont B, Nowacki M, Serrano V, Porcel BM, Ségurens B. 2008. Translational control of intron splicing in eukaryotes. Nature 451:359-362.

Kelkar YD, Ochman H. 2012. Causes and consequences of genome expansion in fungi. Genome Biology and Evolution 4:13-23.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Lim LP, Burge CB. 2001. A computational analysis of sequence features involved in recognition of short introns. Proceedings of the National Academy of Sciences 98:11193-11198.

Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M, et al. 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature 402:761-768.

Lynch M. 2002. Intron evolution as a population-genetic process. Proceedings of the National Academy of Sciences 99:6118-6123.

Lynch M. 2007. The Origins of Genome Architecture. Sunderland, MA: Sinauer Associates.

Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302:1401-1404.

Majewski J, Ott J. 2002. Distribution and characterization of regulatory elements in the human genome. Genome Research 12:1827-1836.

Maquat LE. 2005. Nonsense-mediated mRNA decay in mammals. Journal of cell science 118:1773-1776.

Mayer K, Schüller C, Wambutt R, Murphy G, Volckaert G, Pohl T, Düsterhöft A, Stiekema W, Entian K-D, Terryn N. 1999. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature 402:769-777.

Mount SM. 1982. A catalogue of splice junction sequences. Nucleic Acids Research 10:459-472.

Namy O, Hatin I, Rousset J-P. 2001. Impact of the six nucleotides downstream of the stop codon on translation termination. EMBO Reports 2:787-793.

Pelechano V, Chávez S, Pérez-Ortín JE. 2010. A complete set of nascent transcription rates for yeast genes. PLoS One 5:e15442.

Phifer-Rixey M, Bonhomme F, Boursot P, Churchill GA, Piálek J, Tucker PK, Nachman MW. 2012. Adaptive evolution and effective population size in wild house mice. Molecular Biology and Evolution 29:2949-2955.

Sayani S, Janis M, Lee CY, Toesca I, Chanfreau GF. 2008. Widespread impact of nonsense-mediated mRNA decay on the yeast intronome. Molecular Cell 31:360-370.

Schoen DJ, Brown A. 1991. Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. Proceedings of the National Academy of Sciences 88:4494-4497.

Shapiro JA, Huang W, Zhang C, Hubisz MJ, Lu J, Turissini DA, Fang S, Wang H-Y, Hudson RR, Nielsen R. 2007. Adaptive genic evolution in the *Drosophila* genomes. Proceedings of the National Academy of Sciences 104:2271-2276.

Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. 2006. Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Research 34:3955-3967.

Sivasundar A, Hey J. 2003. Population genetics of *Caenorhabditis elegans*: the paradox of low polymorphism in a widespread species. Genetics 163:147-157.

Skelly DA, Ronald J, Connelly CF, Akey JM. 2009. Population genomics of intron splicing in 38 *Saccharomyces cerevisiae* genome sequences. Genome Biology and Evolution 1:466.

Stoebel DM, Dean AM, Dykhuizen DE. 2008. The cost of expression of *Escherichia coli* lac operon proteins is in the process, not in the products. Genetics 178:1653-1660.

Tabata S, Kaneko T, Nakamura Y, Kotani H, Kato T, Asamizu E, Miyajima N, Sasamoto S, Kimura T, Hosouchi T. 2000. Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. Nature 408:823.

Talerico M, Berget SM. 1994. Intron definition in splicing of small Drosophila introns. Molecular and Cellular Biology 14:3434-3445.

Tanizawa H, Iwasaki O, Tanaka A, Capizzi JR, Wickramasinghe P, Lee M, Fu Z, Noma K-i. 2010. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. Nucleic Acids Research 38:8164-8177.

Theologis A, Ecker JR, Palm CJ, Federspiel NA, Kaul S, White O, Alonso J, Altafi H, Araujo R, Bowman CL. 2000. Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. Nature 408:816-820.

Wernersson R. 2005. FeatureExtract—extraction of sequence annotation made easy. Nucleic Acids Research 33:W567-W569.

Whitney KD, Garland Jr T. 2010. Did genetic drift drive increases in genome complexity? PLoS Genetics 6:e1001080.

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bähler J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453:1239-1243.

Wood V, Gwilliam R, Rajandream M-A, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S. 2002. The genome sequence of *Schizosaccharomyces pombe*. Nature 415:871-880.

Wright SI, Andolfatto P. 2008. The impact of natural selection on the genome: emerging patterns in *Drosophila* and *Arabidopsis*. Annual Review of Ecology, Evolution,

and Systematics 39:193.

Wright SI, Lauga B, Charlesworth D. 2002. Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. Molecular Biology and Evolution 19:1407-1420.

Tables

Table 4.1: Species used in this study

Species	Assembly	N_e (x 10^3)	Total Genes	Genes Analyzed
A. thaliana	TAIR10	40	33583	10141
C. elegans	WBcel215	80	21187	10869
H. sapiens	GRCh27.p10	90*	37150	6075
M. musculus	GRCm38.p2	120	34293	8373
D. melanogaster	Release 5	1150	15581	5260
Sc. pombe	ASM294v2	10000	5883	652
S. cerevisiae	R64-1-1	26000	6352	161

*Ancestral estimate used since analysis relies upon genome-wide data.

Species, genome assemblies, effective population size estimates, total genes in each assembly and number of genes meeting the criteria used in the study. Effective population sizes were obtained from the primary literature (Schoen and Brown 1991; Chen and Li 2001; Wright, et al. 2002; Sivasundar and Hey 2003; Cutter 2006; Shapiro, et al. 2007; Wright and Andolfatto 2008; Skelly, et al. 2009; Brown, et al. 2011; Phifer-Rixey, et al. 2012). Species are listed in order of increasing effective population size.

Table 4.2: Purine frequency is elevated following PTCs.

Species	First Intron Expectation	PTC1	PTC2	Last Intron Expectation	PTC1	PTC2
A. thaliana	0.484	0.481	0.480	0.493	0.516	0.498
C. elegans	0.603	0.631*	0.591	0.606	0.624*	0.567
H. sapiens	0.592	0.651*	0.606	0.573	0.642*	0.590
M. musculus	0.582	0.644*	0.592	0.574	0.623*	0.579
D. melanogaster	0.529	0.605*	0.513	0.525	0.585*	0.508
Sc. pombe	0.473	0.594*	0.489	–	–	–
S. cerevisiae NMD	0.497	0.704*	0.593	–	–	–

* observed different than expected, $p < 0.05$

Table 4.3: Purine frequency following PTCs shows no consistent pattern when splice context is removed.

	First Intron			Last Intron		
Species	Expectation	PTC1	PTC2	Expectation	PTC1	PTC2
A. thaliana	0.484	0.453[#]	0.481	0.493	0.478[#]	0.502
C. elegans	0.601	0.560[#]	0.588[#]	0.606	0.565[#]	0.561[#]
H. sapiens	0.592	0.651[*]	0.610	0.573	0.586	0.594[*]
M. musculus	0.582	0.644[*]	0.592	0.574	0.565	0.581
D. melanogaster	0.529	0.518	0.508	0.525	0.489[#]	0.510
Sc. pombe	0.473	0.511	0.493	–	–	–
S. cerevisiae NMD	0.497	0.704[*]	0.593	–	–	–

[*] observed higher than expected, $p < 0.05$

[#] observed lower than expected, $p < 0.05$

Figures:

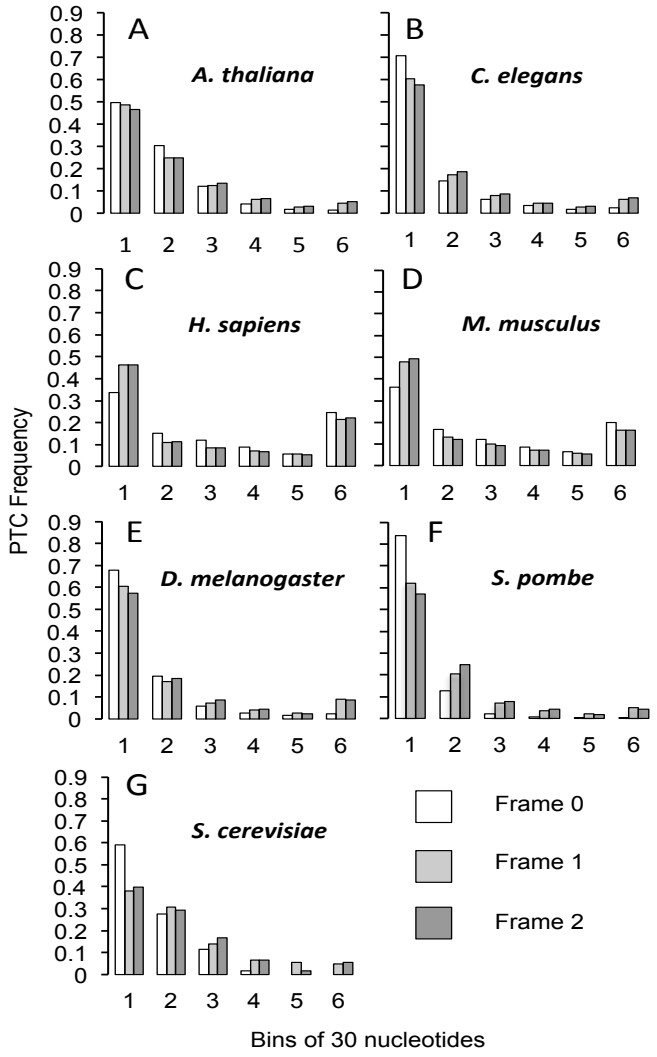

Figure 4.1: **First PTCs occur earlier than expected in the first intron for all non-mammals**. Observed distances in base pairs between the 5' splice site and first PTCs (white bar) and out-of-frame NCs (gray bars). Distances are separated into 6 bins, each of which is 30 nucleotides in length, except for bin 6 which contains all distances longer than 150 nucleotides. Panels A-F are in order of increasing effective population size (Table 4.1).

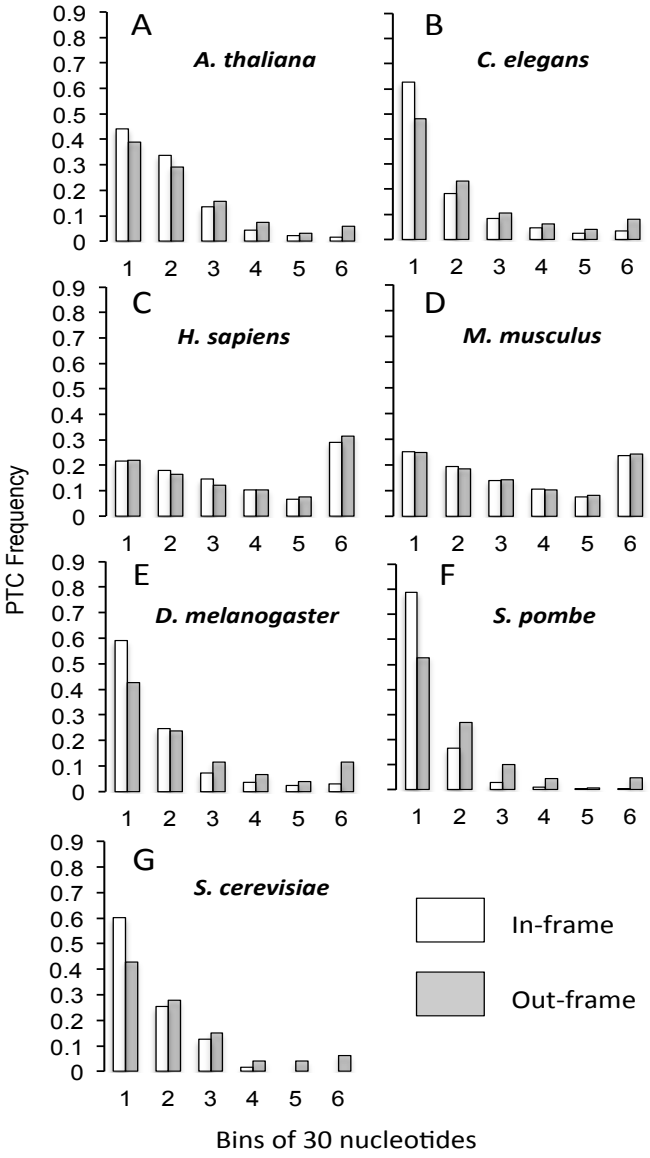

Figure 4.2: **After removing the effect of the splice site, the first PTC still occurs earlier than expected in the first intron for non-mammals**. Observed distances in base pairs between the 5' splice site and first PTCs (white bar) and out-of-frame NCs (gray bars). Distances are separated into 6 bins, each of which is 30 nucleotides in length, except for bin 6 which contains all distances longer than 150 nucleotides. Panels A-F are in order of increasing effective population size (Table 4.1).

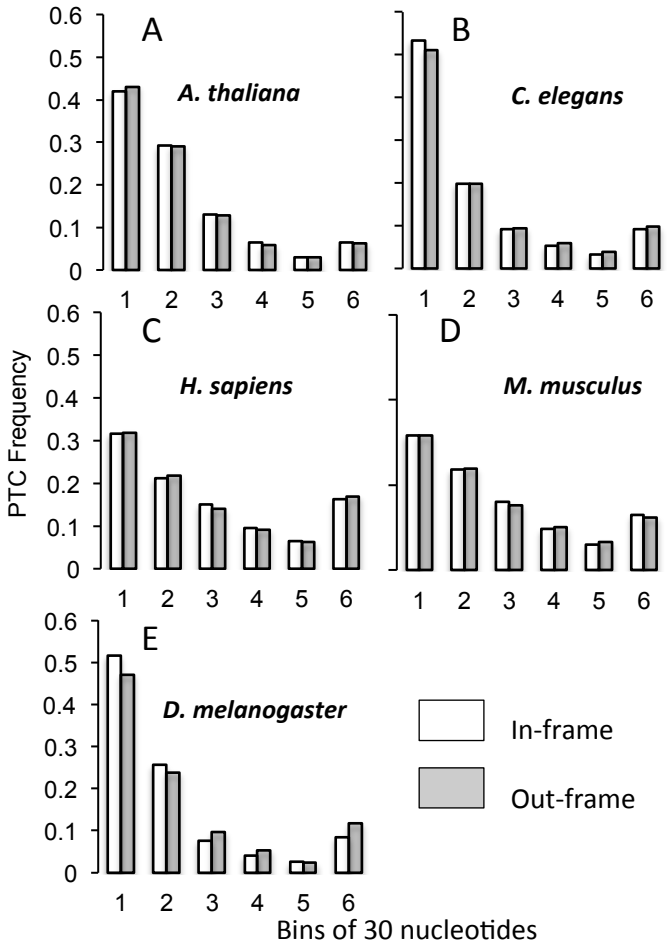

Figure 4.3: **After removing the effect of splice site on PTC and NC frequencies, in-frame PTCs occur earlier than expected in the last intron for *C. elegans* and *D. melanogaster* and significantly later for *A. thaliana***. Observed distances in base pairs between the 5' splice site and first PTCs (white bar) and out-of-frame NCs (gray bars). Distances are separated into 6 bins, each of which is 30 nucleotides in length, except for bin 6 which contains all distances longer than 150 nucleotides. Panels A-F are in order of increasing effective population size (Table 4.1).

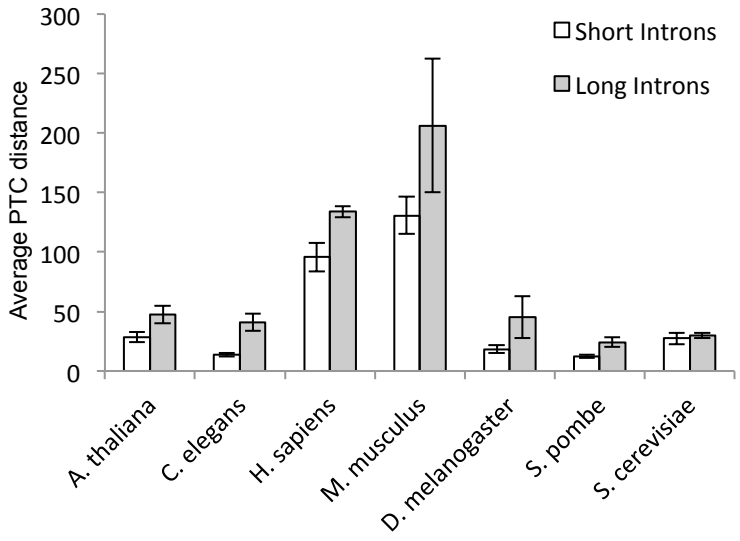

Figure 4.4: **PTCs occur earlier in short introns.** Average position of the first PTC in first introns, measured in nucleotides downstream of the 5' splice site, for short and long introns. Short introns are those in the first quartile and long are those in the second through fourth quartiles of intron length. Error bars are calculated as plus/minus one standard error of the mean.

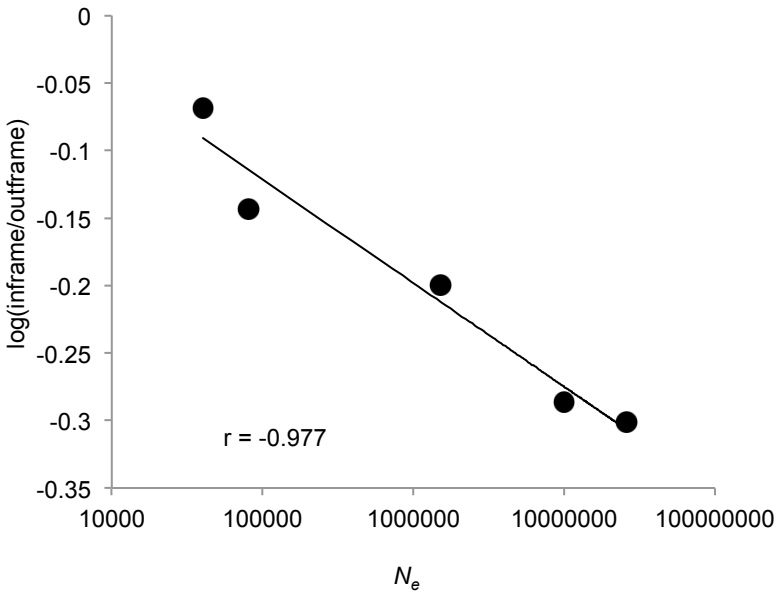

Figure 4.5: **Median position of in frame PTCs correlates with effective population size in non-mammals.** Data points represent log ratios of median in-frame PTC distance and out-frame PTC distance from 5' splice site. (df = 4, Pearson's r = -0.978, p = 0.003). Organisms in order of increasing N_e (i.e. points going left to right): *A. thaliana*, *C. elegans*, *D. melanogaster*, *Sc. pombe,* and *S. cerevisiae*.

CHAPTER 5

CONCLUSION

Throughout this dissertation, I presented three studies that examine the role of four major evolutionary forces: mutation, recombination, natural selection, and random genetic drift on genetic variation and genome architecture. In chapter 2, I presented the first study of genome-wide parameters of mutation in the fission yeast *Schizosaccharomyces pombe*, a species that is about 600 million years diverged from the model budding yeast *Saccharomyces cerevisiae*. Through this study I show that the mutation rate of 0.17 x 10-10 for *Sc. pombe* is almost identical to the mutation rate described for *S. cerevisiae.* Thus, suggesting that the published effective population size of 10 million may be an underestimate for *Sc. pombe.* In addition to mutation rate, there are various differences between the mutation spectrums for both yeast species, however the most striking may be the increase of C:G → A:T mutations in *Sc. pombe* compared to *S. cerevisiae*. Additionally, when looking at the effect of neighboring nucleotides, similar to *S. cerevisiae*, CpG sites have a higher relative mutation rate. This increased relative mutation rate exists independent of methylation since CpG sites are not enriched for C:G→T:A mutations. Lastly, complex mutations and double single nucleotide mutations occur much more frequently than expected. I hypothesized that these mutations may be footprints of DNA repair. Throughout the experiment, *Sc. pombe* was maintained as a haploid, thus reducing the possibility of repair by homologous

recombination. These complex mutations may have been created by repair of double-stranded breaks by non-homologous end joining, a more mutagenic repair pathway.

In chapter 3, I presented the first long-term, genome-wide study of mitotic gene conversion in *S. cerevisiae*, which elucidates the rates and biases of gene conversion by examining loss of heterozygosity due to mitotic recombination. Through this study we observed tract lengths for four different classes of mitotic recombination events. While our median tract length of ~ 7kb was similar to what has been previously reported, we observed a ten fold increase in class B gene conversions, or gene conversions known to be caused by heteroduplex repair. Additionally, 13% of class b gene conversions resulted in tract lengths > 50 kb. Since these tracts did not run to the end of the chromosome we concluded that they were not created by break induced replication and may be the result of double stranded gap repair or a unknown homologous recombination pathway. Multiple gene conversion hot spots were identified. One region, located near a Ty1 elements and the rDNA array was previously described in a study mapping fragile sites within the *S. cerevisiae* genome. Finally, across 85 mutation accumulation lines we observed 38 trisomy events. This is a 2.5 fold increase compared to what was observed in isogenic *S. cerevisiae* mutation accumulation lines. This increase in trisomy may be due to the mating of two diverged strains introducing incompatibilities between segregating chromosomes, or it may be due to aneuploidy commonly associated with wild isolates of *S. cerevisiae*.

In my final chapter, chapter 4, I describe a specific example of genome-wide selection. Here, I present a comparative study of genome-wide selection on in-frame premature termination codons within introns across seven major model organisms.

Through this study, we observe within non-mammalian first introns, premature termination codons occur earlier than expected based on out of frame trinucleotides. This observation holds when the context from the splice site is removed. Additionally, the average distance between non-mammalian premature termination codons decreases with respect to effective population size. This result is independent of gene expression, however premature termination codons within short introns are decidedly closer to the 5' splice site than premature termination codons within long introns. I hypothesize this is because errors in splicing short introns more commonly results in intron inclusion, allowing premature termination codons within introns to be acted upon by selection. I also examined 3' context of premature termination codons to determine if there is selection on premature termination codons to resemble true termination codons. However, once the effect of the 5' splice site was removed, I observed conflicting results. We concluded that the selective pressure on premature termination codon position within introns might be driven by the opportunity cost of ribosomes translating non-functional products and the clean up costs of the abnormal proteins.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR MUTATION RATE AND SPECTRUM OF THE FISSION YEAST *SCHIZOSACCAROMYCES POMBE*

Order of Appendix A:

I. Commands

QC Reads

```
1. /usr/local/ea-utils/latest/bin/fastq-mcf ../../Adaptors.fasta
../Sample${i}_R1.fastq
../Sample${i}_R2.fastq -o Sample${i}_R1.filtered.fastq -o
Sample${i}_R2.filtered.fastq
-C 1000000 -q 20 -p 10 -u -x 0.01

2. /usr/local/fastx_toolkit/latest/bin/fastx_trimmer -Q33 -f 10 -i
Sample${i}_R1.filtered.fastq -o Sample${i}_R1.trimmed.fastq

3. /usr/local/fastx_toolkit/latest/bin/fastx_trimmer -Q33 -f 10 -i
Sample${i}_R2.filtered.fastq -o Sample${i}_R2.trimmed.fastq
```

BWA

```
4. /usr/local/bwa/latest/bwa aln ../../ReferenceGenome/SPombe.RefGenome.fa
Sample${i}_R1.trimmed.fastq > Sample${i}_1.sai

5. /usr/local/bwa/latest/bwa aln ../../ReferenceGenome/SPombe.RefGenome.fa
Sample${i}_R2.trimmed.fastq > Sample${i}_2.sai

6. /usr/local/bwa/latest/bwa sampe ../../ReferenceGenome/SPombe.RefGenome.fa
Sample${i}_1.sai Sample${i}_2.sai Sample${i}_R1.trimmed.fastq
Sample${i}_R2.trimmed.fastq
> Sample${i}.sam
```

```
SAMTOOLS

7. /usr/local/samtools/latest/samtools view -bS -T
../../ReferenceGenome/SPombe.RefGenome.fa Sample${i}.sam > Sample${i}.bam

8. /usr/local/samtools/latest/samtools sort Sample${i}.bam Sample${i}.sorted

9. /usr/local/samtools/latest/samtools index Sample${i}.sorted.bam

PICARD

10. java -Xmx2g -classpath "/usr/local/picard/latest/" -jar
/usr/local/picard/latest/AddOrReplaceReadGroups.jar I= Sample${i}.sorted.bam
O= Sample${i}.sorted.fixed.bam SORT_ORDER=coordinate RGID=PombeLane1 RGLB=bar
RGPL=illumina RGSM=Sample${i} RGPU=6 CREATE_INDEX=True
VALIDATION_STRINGENCY=LENIENT

11. java -Xmx2g -classpath "/usr/local/picard/latest/" -jar
/usr/local/picard/latest/MarkDuplicates.jar I=Sample${i}.sorted.fixed.bam
O=Sample${i}.sorted.fixed.marked.bam M=Sample${i}.metrics CREATE_INDEX=True
VALIDATION_STRINGENCY=LENIENT

GATK

12. java -Xmx2g -classpath "/usr/local/picard/latest/" -jar
/usr/local/gatk/latest/GenomeAnalysisTK.jar -R
../../ReferenceGenome/SPombe.RefGenome.fa
-I Sample${i}.sorted.fixed.marked.bam -T RealignerTargetCreator -o
Sample${i}.intervals

13. java -Xmx2g -classpath "/usr/local/picard/latest/" -jar
/usr/local/gatk/latest/GenomeAnalysisTK.jar -R
../../ReferenceGenome/SPombe.RefGenome.fa
-I Sample${i}.sorted.fixed.marked.bam -T IndelRealigner -targetIntervals
Sample${i}.intervals -o Sample${i}.sorted.fixed.marked.realigned.bam

14. java -Xmx2g -classpath "/usr/local/picard/latest/" -jar
/usr/local/gatk/latest/GenomeAnalysisTK.jar -R
../../ReferenceGenome/SPombe.RefGenome.fa
-I Sample${i}.sorted.fixed.marked.realigned.bam -T UnifiedGenotyper -rf BadCigar
-ploidy 1 -glm BOTH -o Sample${i}.sorted.fixed.marked.realigned.vcf

15. java -Xmx2g -classpath "/usr/local/picard/latest/" -jar
/usr/local/gatk/latest/GenomeAnalysisTK.jar -T BaseRecalibrator
-I Sample${i}.sorted.fixed.marked.realigned.bam -R
../../ReferenceGenome/SPombe.RefGenome.fa -rf BadCigar -knownSites
Sample${i}.sorted.fixed.marked.realigned.vcf -o recal_data.grp

16. java -Xmx2g -classpath "/usr/local/picard/latest/" -jar
/usr/local/gatk/latest/GenomeAnalysisTK.jar -R
../../ReferenceGenome/SPombe.RefGenome.fa
-I Sample${i}.sorted.fixed.marked.realigned.bam -T PrintReads-rf BadCigar
-o Sample${i}.mapped.bam -BQSR recal_data.grp
```

```
COMPARE SAMPLES
1. java -Xmx2g -classpath "/usr/local/picard/latest/" -jar
/usr/local/gatk/latest/GenomeAnalysisTK.jar -R
../ReferenceGenome/SPombe.RefGenome.fa
-I Sample1.sorted.fixed.marked.realigned.bam  -I
Sample2.sorted.fixed.marked.realigned.bam
-I <All SAMPLES> -I Ancestor1.sorted.fixed.marked.realigned.bam
-I Ancestor2.sorted.fixed.marked.realigned.bam   -T UnifiedGenotyper -rf
BadCigar
-ploidy 1 -glm BOTH -o ComparedVariants.vcf

DELLY
1. export OMP_NUM_THREADS=4

2. export LD_LIBRARY_PATH=/usr/local/boost/1.54.0/gcc447/lib/:${LD_LIBRARY_PATH}

3. time /usr/local/openmpi/1.4.4/gcc412/bin/mpirun -np $NSLOTS
/usr/local/delly/latest/delly -t DEL -o Delly/delly.deletion.vcf -q 10
../Sample${i}.sorted.bam

4. time /usr/local/openmpi/1.4.4/gcc412/bin/mpirun -np $NSLOTS
/usr/local/delly/latest/delly -t DUP -o Delly/delly.duplication.vcf -q 10
../Sample${i}.sorted.bam

5. time /usr/local/openmpi/1.4.4/gcc412/bin/mpirun -np $NSLOTS
/usr/local/delly/latest/delly -t INV -o Delly/delly.inversion.vcf -q 10
../Sample${i}.sorted.bam

6. time /usr/local/openmpi/1.4.4/gcc412/bin/mpirun -np $NSLOTS
/usr/local/delly/latest/delly -t TRA -o Delly/delly.translocation.vcf -q 10
../Sample${i}.sorted.bam

VEP

1. time perl /usr/local/ensembl-tools/latest/scripts/variant_effect_predictor/
variant_effect_predictor.pl -i ./NewVCF/VCF_File${i}.txt  --database --genomes
--species schizosaccharomyces_pombe -fork 4 -o
VEPfiles/VEP_File${i}.txt

TRANSCRIPTOME ASSEMBLY

      TopHat
1. export PATH=/usr/local/samtools/0.1.19/:${PATH}

2. time /usr/local/tophat/latest/bin/tophat -i 30 -I 20000 -G ../SPref/SPref.gtf
--transcriptome-index ../SP_transcriptome/SP ../SP_bowtie/SP RNAseq_SP${i}.fastq

      Cufflinks
3. time /usr/local/cufflinks/latest/bin/cufflinks -q -o ./cufflinks_out -g
../SPref/SPref.gtf ./tophat_out/hits_SP${i}.bam
```

II. List of SNMs

Sample Number	Chromosome	Position	Ref	Alt
1	III	580155	C	T
1	III	2076980	A	T
1	II	1494174	C	A
2	II	1389165	A	T
2	II	3017193	G	T
2	II	3146138	A	G
2	I	669655	C	T
2	I	880640	A	C
2	I	4996641	C	A
3	III	55812 C	A	
3	II	2141968	A	G
3	I	3994557	C	G
4	II	1162297	C	G
4	I	473457	A	T
5	II	2340641	A	G
5	II	3747120	C	A
6	II	898368	T	G
6	I	111563	G	A
6	I	1955853	T	A
6	I	2517445	G	C
7	III	381706	C	T
7	II	3718344	T	C
7	II	4225274	T	G
7	I	3902041	G	A
7	I	3960101	C	A
8	III	2423060	T	A
8	II	157822	G	T
8	II	1295544	T	G
8	II	1647382	T	G
8	II	1700167	C	T
8	II	2115836	A	C
8	II	3708095	G	T
8	II	4486513	G	C
8	I	1133813	C	A
8	I	1658502	C	A
8	I	3983911	G	A
8	I	3984187	G	A
8	I	3984205	G	C
10	III	1499597	C	T
10	III	1562543	C	T
10	III	2308505	A	T
10	II	696259	T	A
10	II	1986655	T	C
10	I	2513787	T	G
10	I	3393750	G	T
11	III	1509028	A	G
11	III	1642516	C	A
11	I	1644476	G	T
11	I	3028699	T	C
11	I	3198246	G	C
11	I	5492707	T	G
13	II	2680220	C	A

13	I	1960047	C	T
13	I	3854681	A	G
14	III	1690952	G	A
14	II	1818125	T	C
14	II	4060123	A	G
15	III	2180705	T	C
15	I	2463190	G	T
15	I	2856193	T	C
15	I	4173337	C	T
15	I	5509629	C	G
16	III	242246	G	T
16	III	2367774	T	C
16	II	1906273	T	G
16	I	4754440	G	A
18	I	1341863	G	T
19	III	2263561	C	A
19	II	2474145	C	A
19	II	2490679	G	T
19	II	3311797	G	T
19	II	4453075	G	T
19	I	5555745	C	T
21	III	1474621	A	G
22	III	376536	G	T
22	I	3368319	T	C
23	III	242636	T	C
23	III	1183665	G	T
23	III	1548026	G	A
23	III	2046831	G	A
23	I	3262756	G	T
28	III	986220	G	A
28	III	986259	G	A
30	III	139222	T	C
31	III	680111	T	G
31	III	1987347	A	G
31	II	4150004	G	C
31	I	957219	A	C
31	I	1624249	A	T
31	I	2802787	C	A
31	I	3860729	G	A
31	I	4884069	T	C
32	III	899900	A	G
32	III	2412296	T	A
32	II	1825364	C	A
32	II	2768284	G	A
32	I	5542196	T	C
35	III	32710 A	T	
35	III	2169677	G	T
35	I	2196349	G	T
35	I	4897573	G	A
35	I	4897638	A	G
35	I	5539765	A	C
37	III	721040	A	G
37	II	2252616	G	T
37	II	4474478	T	C
37	I	3214826	C	A

37	I	3264603	A	C
37	I	5175969	G	T
38	III	897531	C	A
38	II	3976165	A	G
38	II	4231234	G	T
38	II	4379038	T	G
38	I	5554419	G	C
39	II	690700	A	G
39	II	690844	A	G
39	II	2890229	G	T
39	II	3538201	A	C
39	II	3889936	G	T
39	I	1376330	C	A
39	I	3799259	C	A
39	I	5263344	A	G
40	III	200150	C	A
40	III	1475102	G	C
40	II	1517581	G	T
40	II	2623813	G	A
40	II	2876325	G	A
40	I	1537016	T	C
40	I	1849301	T	G
42	II	661662	A	T
42	II	3665485	A	C
42	II	4476439	A	C
42	I	146438	C	G
45	III	954825	G	C
45	III	2037436	T	A
45	II	489249	C	A
45	II	2062318	G	C
45	I	1361637	T	G
46	III	2272623	C	T
46	I	3575735	A	C
46	I	4934672	C	A
49	II	461209	C	A
49	II	1976875	G	C
49	II	2902706	T	A
49	II	4472561	G	C
49	I	4559857	G	T
50	III	461020	C	T
50	II	1174795	C	A
50	II	2786116	T	G
50	II	3034421	T	G
50	II	3875319	G	T
50	I	1908892	C	A
50	I	5558350	T	C
50	I	5568679	A	T
51	III	2447848	T	A
51	II	320347	C	A
51	I	2635912	T	C
52	III	125988	G	T
52	III	233473	A	G
52	III	234148	G	A
52	III	234181	G	A
52	III	1208004	C	T

52	II	949551	C	T
52	II	3209429	G	A
52	I	96507 A	G	
52	I	2659037	T	C
52	I	4346183	C	T
53	III	1067179	T	C
53	II	745479	T	C
53	II	1358067	A	G
53	II	1480412	C	T
53	II	1598125	G	T
53	II	2032247	G	A
53	I	1227230	C	T
53	I	1818049	C	A
53	I	2281445	A	C
53	I	3576582	G	T
53	I	5573770	A	G
54	II	441378	C	A
54	I	2597950	C	T
54	I	3077913	C	A
55	II	46906 T	G	
55	II	1127921	A	G
55	II	3360518	C	T
55	I	1531903	C	A
55	I	2455028	T	G
55	I	5395534	A	G
56	III	2197590	A	T
56	III	2365873	C	A
56	I	3634643	A	T
58	III	2268438	T	C
58	II	2570463	T	C
58	II	3283184	G	A
58	I	3296352	G	T
58	I	3340501	C	G
58	I	4460676	C	T
59	II	340654	C	A
59	II	2509945	G	T
59	I	1690626	G	T
59	I	2083120	C	T
59	I	2422509	C	T
60	III	1701925	G	T
60	III	1925626	C	A
60	II	913456	C	A
60	II	1312254	C	T
60	II	2330091	T	G
60	II	3427790	C	T
60	II	4276793	G	C
60	II	4346278	A	C
60	I	4537688	G	T
62	II	1173997	C	A
62	I	1259685	C	T
63	II	4341964	G	A
63	I	2584354	A	C
64	II	1185317	A	G
64	II	1863084	C	T
65	III	858990	C	A

65	I	2578709	C	A
65	I	5074604	A	C
67	II	3089439	T	C
67	I	1195626	C	T
67	I	3751598	T	C
67	I	3821837	A	T
67	I	4063301	G	A
67	I	4648161	G	T
68	III	1549009	A	G
68	II	1281036	C	T
68	II	1546664	G	A
68	II	2388613	G	T
68	I	133718	T	C
68	I	832207	G	T
69	II	1018883	T	A
69	II	2141993	T	C
69	II	3868645	T	A
69	I	3879455	C	A
69	I	4079703	A	G
70	III	1251637	A	C
70	II	2552702	G	T
70	I	2795303	C	T
70	I	3700385	T	A
70	I	5219833	T	C
71	III	1033329	C	T
71	II	480932	C	T
71	I	4810076	T	G
72	II	4333598	C	A
72	I	482170	A	T
72	I	514978	A	G
74	III	2180746	C	T
74	II	676823	T	G
74	II	2701154	G	A
74	II	4375489	T	C
74	I	514243	G	A
74	I	1871200	T	C
74	I	2022094	G	C
75	I	3291105	G	A
76	III	208140	C	A
76	II	1291129	T	G
76	II	2120863	C	A
76	II	2827649	A	T
77	III	243761	A	G
77	III	244643	G	A
77	III	1773567	G	A
77	III	1823584	G	T
77	III	1996086	C	T
77	II	1730758	C	T
77	I	314182	C	A
78	II	543861	C	A
78	II	3587308	C	T
79	II	1277511	A	T
79	I	1048635	G	A
81	II	3671308	C	G
81	II	4066663	C	A

81	I	3083202	G	A
82	III	1405215	T	G
82	II	393560	C	A
82	II	3271190	G	T
82	II	4273835	A	G
83	I	4277752	T	C
83	I	4307181	C	A
84	II	126034	C	G
84	II	628939	G	T
84	I	2047266	C	A
84	I	3714966	C	T
84	I	4309259	G	A
85	III	1560301	A	G
85	II	1659366	G	T
85	I	2296088	G	C
85	I	5309091	A	G
87	I	5562193	A	G
88	III	1970506	A	T
88	II	58468 A	G	
88	II	2016248	A	G
88	II	2115437	T	G
88	II	3126441	C	A
88	II	4107252	A	T
88	II	4250399	C	T
88	I	934940	C	A
88	I	3387145	G	T
88	I	4911431	G	T
89	III	1068022	G	C
89	II	66679 A	G	
89	II	4253853	T	G
89	II	4533598	G	A
89	II	4533922	G	A
89	II	4534498	A	C
89	I	1225598	G	A
89	I	3703023	C	A
89	I	5559293	A	G
90	III	2017598	A	G
90	II	907659	G	T
90	II	2356392	C	A
91	III	1264085	C	A
91	III	1517590	C	A
91	II	822484	G	T
91	II	2443898	C	A
91	I	2723302	T	G
92	II	1196504	G	T
92	I	1613171	G	A
92	I	2582088	C	A
92	I	4675905	G	T
93	II	3171809	C	A
94	III	507990	A	C
94	II	626278	G	C
94	II	816429	C	T
94	II	3814135	C	T
94	II	4502746	C	G

III. List of Indels

LINE	#CHROM	POS	5'	REF/ALT	3'	INS/DEL
1	I	216402	CTACAAAAAGTATTGACTTA	-/T	TTTTTTTTATCTCCTGTGAG	INS
1	I	887480	TTATCAACGTATTTTTCACT	-/A	AAAAAAAAAACAATCATTTC	INS
1	I	978504	GATAAAAACTACGATTTTGC	-/A	AAAAAAAGGATTCTCAAACA	INS
1	I	1877723	TTTCTTTTTGTTGTATATGA	-/T	TTTTTTTTTTAAAGGAAATG	INS
1	I	5310065	GTTTTTATGATCAATTCCTC	-/T	TTTTTTTTGTCTTATTACTC	INS
1	II	1122580	TAAATAACGGAAGCTTTCTC	-/T	TTTTTTTTTTTTCGAACATTC	INS
1	II	1226613	AGAAACGGTTGCAGGCACGG	-/TGGCATCAT	TGGCATCATTGGCATTTGTG	INS
1	III	1787927	CAGTGAACAAAATATGCATC	-/T	TTTTTTTTTTCTAAAAAACC	INS
2	II	3248398	CAGTTCCCAAGGTATGTTTA	-/T	TTTTTTTTACTGGATATGCT	INS
2	III	1351690	TTAACTCCATTCCTGTATTA	-/T	TTTTTTTAACAAAGGGGTTG	INS
3	I	2556743	CCAATAAGTTGGACAAATTG	-/A	AAAAAAAAAAAAAAACCACT	INS
3	II	2468409	CCCAGCTAAGTAAGTCTTTG	-/T	TTTTTTTTTTTTATTCATTT	INS
3	II	2969810	TGAGGAGCTTATTAAGCCAA	T/-	TTTTTTTTTTCATATCTTTA	DEL
3	II	3380519	CATACCCAACATTACCTCTT	C/-	CCCCCCCGTAGTACTCATTT	DEL
3	II	4480351	AACATGAATTTACATCATCC	-/T	TTTTTTTTAGAAGATATTTC	INS
3	III	234763	TTGGAATTATTTCTTTGGTA	-/TTGTT	TTGTTTTGTTTTGTTTTGTT	INS
4	I	2110988	TAATAAATAGTATTTTAAAC	-/AT	ATATATATATATATATATAT	INS
4	II	2327057	TATGTGAGGAATAAAGGAAG	-/TA	TATATATATATATATATATT	INS
5	II	4290954	ATAGTACTAAGAAAATTCTG	-/A	AAAAAAAAAAAAACAAAACA	INS
5	III	339602	GTGTGATAAAAAGTATAATC	-/T	TTTTTTTTTTGAAATCCTCT	INS
5	III	1715328	TCTATTAGACAACCTTTTTC	-/A	AAAAAAAAAAAAAATTCCGC	INS
6	I	2669515	ACTCCGAATTATAATCTATA	-/T	TTTTTTTTATCAGTGTGTGA	INS
7	I	3395628	GCTACCCTAAGAGAGTATAT	-/A	AAAAAAAACCTTTCGCTCCT	INS
7	II	861179	TTTAAAAGGGTCATTCATTC	-/TT	TTTTTTTTTTACTTAGTTCA	INS
7	III	947289	AGCATAGTTGGTTATTGCGC	-/A	ACGACTGTTAATCGTGAGGT	INS
7	III	2382418	GATTTTTGATTGACTGTTTG	CTCTTACCTAATTAATC/-	CTCTTACCTATATTGTCTAC	DEL
8	I	443299	TTTTACACATATTTTTGTTG	-/T	TTTTTTTTTTATCAAGGAAG	INS
8	I	1453569	CTTTCATTCCATTCAAATCG	-/T	TTTTTTATCGCGTTGTGCAT	INS
8	I	3966159	TTATATTTATGAAATTACCG	-/C	CCCTTTTTCAACATTTTATT	INS
8	II	524666	ATGCAAAGGGAAGACAGAGA	-/CAAGGCAAGG	CAAGGCAAGGCAAGGCAAGG	INS
8	II	948905	AATGAAAGGAGTAAAAATTT	-/A	AAAAAAAAAAAATAAAAAAT	INS
8	II	1149909	CAAAAATGAAATGGATAAAA	-/C	CCCCCCCCATTTCTTAAACA	INS
8	II	1738425	ACATTTGTTTTTCGACGTAT	-/A	AAAAAAAAACAGAAAGACAT	INS
8	II	2808906	CTCATTCATTGTTGCATATA	-/T	TTTTTTTTAATTTTGTTGCA	INS

8	III	490008	TTCTTTTTTCGTTAAAAAAC	-/A	AAATTGTTCTGAAAAAAAGG	INS	
10	I	1666321	ACTTATTCTAATTTATTGGA	-/T	TTTTTTTTTTATTAATTTAT	INS	
10	I	2232167	CTTGAAAATTGCAAACATTC	-/A	AAAAAAAAAAAAGCATAAAG	INS	
10	I	4177950	ATCTTTAAAAATGTTTATTT	A/-	AAAAAAAAAAAAAAAAAGAAA	DEL	
10	II	1233181	CTTCTTTCTCATTTGCTTCG	T/-	TTTTTTTTTTATATGATATT	DEL	
10	II	1267656	GCTGTCCATGCTTTGTTTAC	-/T	TTTTTTTTTTATTTCTTTTC	INS	
10	II	1702025	AAGGTCCACAAAGTACTTTC	-/A	AAAAAAATTATCCGTATAAA	INS	
10	II	2963988	TATATATATTTTTGATATTA	-/T	TTTTTTTTACCGCTAGCGAT	INS	
10	II	3974259	ATATAAAAAGTAAAAAAAAT	-/A	AAAAAAAAAAAACTCAAAAA	INS	
11	I	216147	CTAGTTCTTGTGTCTCTTGC	-/T	TTTTTTTTTTTTTTTCGGTT	INS	
11	I	1154564	GAACCTTTCACCTTCAAACC	-/T	TTTTTTTTTATCTATTTTAC	INS	
11	I	4934297	GGAATATTTCGAGTGGTACC	-/TA	TATATATATATATATATATA	INS	
11	III	974746	CACGTAACCCAATAAAAATA	T/-	TTTTTTAGGAAATTTACCAC	DEL	
11	III	1707997	CCATTAATAGAGATTTTTCT	-/A	AAAAAAAAAAAACTTGACAC	INS	
13	I	543809	AATGACACATTAAAAGGAAT	-/A	AAAAAAAATGTTAATAATAC	INS	
13	I	1152144	ACTTTAGTAATTTCTGTACC	-/T	TTTTTTTTTTTTGCAAATGA	INS	
13	I	4474733	AATCATTGCAGTTTTAATAT	-/A	AAAAAAAAAATGAAATGTGA	INS	
13	II	577240	CTGTAAACAATTGTTCTATA	TCTTCTCCTTGTTATTCTATG/-		TCTTCTCCTTCATGTTCCAC	DEL
13	II	2373654	GATGCAATACAGACTTGGTC	-/T	TTTTTTTTTAAAAAGTTTTG	INS	
13	II	3622335	TTACATGGCTAATATTTAGT	-/A	AAAAAAAAATAACATTTAAT	INS	
13	III	2143868	GCCAACAATTTTCTACTTAA	T/-	TTTTTTTTTTGAAGATAATA	DEL	
14	III	369347	ATGTGTTTTGAATTATCCCT	-/A	AAAAAAAAAAAACATTACCT	INS	
15	I	1488892	GGTTTGGGTAATTCTTTTTC	A/-	ACCGTTGTTTCTGCATTCAA	DEL	
15	I	2088859	CAATCACAATATTTAATATG	-/T	TTTTTTTTTTAAATAGTTTA	INS	
15	I	3844832	ATATTAAAATTGAAATTGCC	-/A	AAAAAAAAGACAAAAAAAGA	INS	
15	I	4233098	TTGAAAGTTAGGACCCAAAT	-/A	AAAAAAAAAAGAAAATCAGG	INS	
15	II	2768245	TGTTTTGCTATAAAATGCAA	-/T	TAAAAAATTACTGAAAGGGG	INS	
15	II	3333440	AGCAAAAGTCTTATATTAAT	-/A	AAAAAAAAGGAAAAGTTTTT	INS	
15	II	3380358	TCAAAATTTATTTTATTTTA	T/-	TTTTTTTTTATTTACTATAT	DEL	
15	III	1699221	TCGAATCCCTGCATAAGCGC	-/T	TTTTTTTATCACAGGATTAG	INS	
16	I	2804352	TTTTCTTCTACTTTTTTTTA	-/T	TTTTTTTTTTCAAATTTCTA	INS	
16	I	4476841	AATGAAAATTTGGAACTGCC	-/A	AAAAAAAAATAATAAAGTGC	INS	
16	II	31713 AATTATTCACTTAGATTATC	-/T	TTTTTTTAGCAATATTTCTA	INS		
16	II	2434556	ACAAAAGAAGTATAATCTAT	-/A	AAAAAAATATGTCGTATCAT	INS	
16	II	3925634	GTATACGGACGAAATTATGT	-/A	AAAAAAAAGTCTTTGGCTGT	INS	
17	I	4882281	CATGATATAATTTTTCTACC	-/TT	TTTTTTTTTTTTCATACTTA	INS	
19	I	4603496	AGTATTTTATTAAAAGAAAG	-/ATA	ATAATAATAATAACAATCAA	INS	

19	II	1218928	AGAGGACGTGCAAAAGCTAC	-/T	TTTTTTTTTGTTGCACAAGA	INS
19	II	2801607	CTGCCCTGAGCCATTGCAAC	-/T	TTTTTTTAGCAACCTTTTAC	INS
19	II	3578064	TAAATTCATTACTATAAATT	A/-	AAAAAAAAAAATATGTGAAA	DEL
19	III	96427	TCTTTTCCGCCATCAGTAGG	T/-	TTTTTTTAAGACGATCATAT	DEL
22	III	623537	ACAAATAATGATTAAAAAAC	A/-	AAAAAAAAAAAAACGGAAGG	DEL
23	I	3009971	ACTTTTACTATTAAATTATA	-/TTTAT	TTTATTTTATTTTATTTTTA	INS
23	I	3034071	TTTTTATTTTTATTATTATA	-/T	TTTTTTTTTAACTTGTTTCC	INS
23	I	3877894	AAGAAATACACAAAAAAAAT	-/A	AAAAAAACAATCAATTCAGG	INS
23	II	1225665	AGGAGGGATATGAGTTAAAT	-/A	AAAAAAAAAAAATTGGGCAG	INS
23	II	3541648	AGGACTATGAATGTCAGGAA	-/T	TTTTTTTTTTGGAGAAACAA	INS
26	I	2977255	CTTTCGCTTTTTTTTTTTTA	-/T	TTTTTTTTTTTTAAGAATAA	INS
26	I	4438285	ATAAACTTGGTTTATCTACC	-/T	TTTTTTTTATCTCGTACATG	INS
27	I	3381391	TCGCTAGTCATTGCATTTTC	-/A	AAAAAAAAAGTGAAATGGTT	INS
31	I	1333654	GTATCGAAGTACCTTGACAT	-/G	GGGGGGGGGTTCAATCACCA	INS
31	I	4799348	ATCTGCCTCATATCAAACTT	-/A	AAAAAAAAAAAAGAATAAAT	INS
31	II	505255	TCATTTTCGTTTGATTAATT	-/A	AAAAAAAAAGATATAAAGGC	INS
31	II	2468645	ATTTTTTTGCTAGATTGATC	-/TTT	TTTTTTTTTTTGAATTTTAG	INS
31	II	3028585	ATACTCGTAATTGAAATTAC	-/AT	ATATATATATATATATATAT	INS
31	III	1193734	ATCAGTAGAAAAATATGCTG	-/T	TTTTTTTTTTTTTAAAATAA	INS
32	I	1691802	TCCTTCAATCCACTGTAATT	-/A	AAAAAAAAAGATTAAATTAA	INS
32	II	2137084	TTCTGATAGTTTCATCTTTC	TTTCTTTTCTT/-	TTTCTTTTTCTTTTTTAAAT	DEL
32	II	2768239	CTCCCCTGTTTTGCTATAAA	-/C	ATGCAATAAAAAATTACTGA	INS
32	III	1590141	CGCGGATCTCCATATCAGAT	-/A	AAAAAAAAATACTACGACAA	INS
35	I	1475975	CCACCTCCTCCACCGGTAGG	-/A	GGTGATCAATTCACGTTTCA	INS
35	I	4742680	TTATATGCCCACATTTACAC	-/T	TTTTTTTTTTATCCAAACTT	INS
35	II	3404020	CGCGATCAGTAGAGTATTTT	-/A	AAAAAAAAATAATAGAAAAT	INS
35	III	1928744	TAAACATAATTTATACATTG	-/T	TTTTTTTTTTTTGAGAGTAA	INS
37	I	1360608	CTACCACATGTCCTTCATAT	-/C	CCCCCCCCAAACATCATTCA	INS
37	I	3574738	TTAGAATATACTGGATAGGC	-/AA	AAAAAAAAAAGAAAAACTCC	INS
37	I	4212540	GTATTAAGAGAATTTATTGT	A/-	AAAAAAAAAACATATAGATA	DEL
37	I	5009024	GATAGCGGTCCTAATTCAGC	-/T	TTTTTTTTTTGAGTAGAAAA	INS
37	I	5250486	GCTGGATTTGACCATATATT	-/CATTA	CATTACATTACAACACTGGG	INS
37	III	1206858	ATAGTTCTATGTATAACTAA	T/-	TTTTTTTAGTTAGAGAACAG	DEL
37	III	1674267	GTCTGCAATGTTTTTTACCA	T/-	TTTTTTCGTTAAAATACTTT	DEL
37	III	1744661	CATGATGAAACGACTTTCTA	-/T	TTTTTTTTTTACTTCTATTA	INS
38	I	3700425	AAAAAGTTTATAATAGTAAA	T/-	TTTTTTTTATGACACGTGCG	DEL
38	II	792174	AAGTCCGAAAAATGATAATT	-/A	AAAAAAAAAAAAAGCTGTTT	INS

39	II	1285834	TTCTAACAACGCTAAAAAAC	-/T	TTTTTTTTTTACTCATTTAA	INS
39	II	1345067	GCAAAGAGAGAGAGAAGAGT	-/A	AAAAAAAAAGTAATAAATGG	INS
39	III	518843	TATGTAAATTTCAATTATGC	-/TT	TTTTTTTTTTTTCTTTTTTG	INS
39	III	751262	AGTTCATGATAGATAGAACT	-/A	AAAAAAATTCCACTTCTCAA	INS
40	I	4948470	AAATAACAAAACTTAATTGG	-/T	TTTTTTTTAAAAAGATTATG	INS
40	II	513899	GATAATTGGGTTTTTATCAA	-/T	TTTTTTTTAATTAACTATCA	INS
40	II	661659	TATGAGTTTCTCTATTTTGT	-/A	AAAAAAAAACATTTATGAAA	INS
40	II	3268814	TCTAAAAGACGAAAATACAG	-/ATAT	ATATATATATATATATTTCG	INS
42	I	301305	GAAGACAGATGTTATAAAGG	-/T	TTTTTTTTTTTTTACTTCAA	INS
42	I	3798874	CATTGAAGCAACAGTCAGTA	-/T	TTTTTTTTTGTTAATCTCAA	INS
42	I	3992107	AAAAGAGAACAAAGAAAAAT	-/A	AAAAAAAGAAACAAAAGAAA	INS
42	I	5016002	AAACTAAAAAACCTAAATCC	-/T	TTTTTTTTTTGTAAAATTTA	INS
42	III	988227	ATTTTTTAATTTTATGAAAG	-/A	AAAAAAAAAATCTTTTTTTT	INS
45	I	2578547	TTTTTGCAAAGAATTTTTAC	-/A	ACAGTCAAGATTGTTGGCTT	INS
45	II	843434	TTAAACTCTGAATTTCTGCC	-/A	AAAAAAAAAAGAGAAAGAAA	INS
45	II	1406883	AAAAAGATAGTACCTACTTC	-/T	TTTTTTTTATAGAAGAATGC	INS
45	II	2084731	AAAAAATGAAAGCATATTTA	-/T	TTTTTTTATTAGAGCGCTTC	INS
45	II	3010975	GGCTTGTGTCTATGGGTATG	T/-	TTTTTTTTTTTATGTGTATG	DEL
46	I	3460476	AAGTTAAAAGCCGTAAATTG	-/A	AAAAAAAAAAAAAAACTAAA	INS
46	II	819795	CAGCGGCTTCTACCTACTTT	-/C	CCCCCCCCAACTTGCTACTC	INS
46	II	962601	AAAAACCAATGAATGAGCCA	AGCTGAGGCG/-	AGCTAGTATTGGATCCTATT	DEL
46	III	1717327	ATGTAGTGAGTGAATAGTCG	T/-	TTTTTTTTTTGGCTAATTCA	DEL
46	III	1794765	GATATTCGTAACAATTGGTT	-/A	AAAAAAAAAGGACTTACAGC	INS
46	III	2087758	CTTTGTCATCGTTACGATGC	-/T	TTTTTTTTTTTAAAAAATGC	INS
49	I	4458748	TCATTCATCATCTATAGCAT	-/TTAAA	TTAAATTAAATTAAATTAAA	INS
49	I	4471990	TTTGCATTCCCCAATGAAAT	-/A	AAAAAATCTTCAATATTTAA	INS
49	III	1753926	ATTCTCCCGCAAAGTTATAT	-/A	AAAAAAAAATGGAGCAAGCA	INS
49	III	1857250	AGGCGCATAAAGGAGCAGTT	AAAAACAAAAA/-	AAAAACAAAAAAAACTATTG	DEL
49	III	2256051	TTTTTCTTTGATTTAGATAT	-/A	AAAAAAACTTCAAGGGCACT	INS
50	II	472031	ACAATAAAATAAATTATGCT	-/A	AAAAAAAAAATATTATAAAT	INS
50	II	1653566	CCAATAAAATAGTAAAAGAT	-/A	AAAAAAAAACCAGTGTTCAA	INS
50	II	2579458	GGTTGCAAAAATATAAGAGC	A/-	AAAAAAAAAAAAAAACAGAA	DEL
50	III	2159196	TAGCTTTATTTGTGTTTGAT	-/A	AAAAAAAATGCCATTTGTAA	INS
51	I	997233	TTTAAGCTTTGTTGTTGCTG	-/T	TTTTTTTTTTTAGAGAAAAG	INS
51	I	1421034	AAATTCTCGCATTAAAACCT	-/A	AAAAAAAAAAAAAATAAACA	INS
51	I	3574812	TTGTATGTATGTAGTGAAAC	-/A	AAAAAAAAAAAAGGAAGAAA	INS
51	I	4422305	AAAAAAAATAATTTAAATTA	-/T	TTTTTTTTATTCAAATTGTA	INS

51	I	5058829	CTCGAATTCTACTTTAATCC	-/T	TTTTTTTTTTTTCTTTTGTA	INS
51	II	1165797	AATTCAAAAGTTCTAATCAA	-/T	TTTTTTTTATAGGGACCAGA	INS
51	II	1364192	TTTGTGTGTTTTCTTGTCCT	-/C	CCCCCCCATAAATGTTAGAT	INS
51	III	1351852	AATAAAAACAAAAATAGAGT	-/G	GAAAAAAAAATAGATATCGG	INS
51	III	1778791	CTCTCATCTTCATTATTCTG	-/T	TTTTTTTTTTACTCTACTTT	INS
51	III	2137868	TTAGCTGTATTAGGATAGAT	-/A	AAAAAAATATCGGTAAAATA	INS
52	I	275091	ATTACTAAAGGTTCTATATG	-/T	TTTTTTTTTAGCACCCATTT	INS
52	I	474596	TACTCGAAGCCGATTTCTTT	-/A	AAACTTGATGTGGAAGACGT	INS
52	I	609849	AAATGTTGAAATTTTTGTTC	-/T	TTTTTTTTTTTTTTCAAGTA	INS
52	I	4346182	CATGTTACCGTTTTTTTTTT	C/-	GTTCTGTCCAAAAGTATAAC	DEL
52	I	5095813	GTAGGTGCAGCAGACATTTT	-/A	AAAAAAAAAACAAGGAAATA	INS
52	II	86225	CGAATTTCATACAATTTAAT	-/G	GGGGGTAAACTGCGATTAAA	INS
52	II	1134480	AAGCTCTTTTAATATTTTTG	-/T	TTTTTTTAAAAAAAGCAAAT	INS
52	II	2871377	TCAATATACCAAGGAAATGT	-/A	AAAAAAAAAAAATAAGAGAA	INS
53	I	2819656	GAAAATCTGGAACGTAAAGA	-/T	TTTTTTATAAAAACGACTTT	INS
53	I	3300621	CATCAATGATGGTAAACAAA	-/AAAATAAAAT	AAAATAAAATAAAATAAAAT	INS
53	I	4089646	TCTTTAGATTCTATTTTTAC	-/T	TTTTTTTTTTACGAATGCCT	INS
53	I	4923830	TCAATATATATATATATATA	T/-	TTTTTTTTTTATTCGTTTTT	DEL
53	II	1170748	ACATACAAACATAGTTTCTC	-/T	TTTTTTTTTTCTTTTTTTGA	INS
53	II	1748761	TGAAGAGTGTATTGGCAGTT	-/G	GGGGGGGAAGCCGTAGAAAG	INS
53	III	1429860	AGGCGGAGCTGCAGAAGGTA	-/AAGGGGAGAAAGGT	AAGGGGAGAAAGGTAAGGGA	INS
54	I	3993180	GAATCACAAATATCTGCATC	T/-	TTTTTTTTTCTTCTCTTTCT	DEL
54	I	5555289	AGAAAAGAAAGTGTATATTT	A/-	AAAAAAAAAAAAAGGATTGT	DEL
54	II	1389008	CAATACTTTTATTGATAGGT	-/A	AAAAAAAAAAATCCTGTACT	INS
54	II	2011854	GTTTTGGGGGAAGAGTTAAT	-/A	AAAAAAACTTATATAGAAAA	INS
54	II	4310184	AATTTAAATATTATTATCCG	-/TA	TATATATATATACCTTAAAC	INS
55	I	3654923	ATATATACAATTCTTAAATA	-/T	TTTTTTTTTATAATTATACT	INS
55	I	4141185	AATTATATCGCATTGCCTAT	-/ATAC	ATACATACATACATACATAT	INS
55	I	4235585	CATTTCTTTTTTTTCTGTTC	T/-	TTTTTTTTTTTAATATCTTT	DEL
55	II	1214213	TTCAAGCAATAGTCCTTTTA	T/-	TTTTTTTAAGCTATCAAGAC	DEL
55	II	3561779	TTTTAAGTGATTGTCGTTCT	A/-	AAAAAAAAACATTATTTATT	DEL
55	III	672298	TTTTTGTTTTATTCTTTTTC	-/T	TTTTTTTTTTCTCTTTCTTC	INS
55	III	1676609	CAGGATCGAAAAGACCACTC	-/TTTAAA	TTTAAATTTAAATTTTTCAC	INS
56	I	2196174	AGCAACCATCTTACATTAAA	-/G	GGGGGGGAAATCCGTGATTT	INS
56	I	2582625	TCTTCAGATTCAGATTCCTC	-/CTCAGACTCAGA	CTCAGACTCAGACTCGGACT	INS
56	I	3990895	CCCAAAAAAACAAAATTTAT	-/AA	AAAAAAAAATATACCGGACG	INS
56	II	3546214	AGTAAGCAACGCGTTTCTTT	-/A	AAAAAAAACCTCCGAATATT	INS

56	III	2348394	TGGATTGCTACAATTACCGG	-/T	TTTTTTTTTTATAATTCTCA INS
58	I	1072991	GTTGAAAGTGATTCAGCCTA	-/TATATATATG	TATATATATGTATATATATG INS
58	I	1115695	ACAAACAAATGCGCCCCCGG	-/C	GCCGTAAAATTGCTGTGCCC INS
58	II	2805811	TGCATATTTTATTTGTATTG	-/TGTATA	TGTATATGTATATGTATGTG INS
58	II	3976264	GATTAAGAGATAAGTTAGAT	-/A	AAAAAAATATAAAGTAAAAT INS
58	II	4357961	TTTCAATCTAGTTCATTTAC	-/T	TTTTTTTTTTGCTAAGGGAT INS
58	III	763965	TTTTATTTTATTTTATTTCG	TTTAA/-	TTTAATTTAATTTAATTTAA DEL
59	I	1787682	GTTTTGGCTTAAAATTTACC	-/T	TTTTTTTGTTGTACAACTAA INS
59	I	3009880	ATTGTACATATATATACATA	-/T	TTTTTTTTAAATTGAGCCCA INS
59	II	1520498	TAATGAAGCCATTATAAGTT	-/A	AAAAAAAAAGTGAAACTGGG INS
59	III	1790725	CAAATATGTTTAGAATGCTC	-/T	TTTTTTTTTATTTTGTTTAT INS
59	III	2068246	TCATATATTATACTACCACA	-/TACCC	TACCCTACCCTACCCTACCC INS
60	I	2254403	AATTGTTTGTCTGGCTTTTC	-/T	TTTTTTTTTTATTTACATCG INS
60	I	4241118	GTAAAATTATATCTTGTTTG	A/-	AAAAAAAAAAAAAAAAAAAT DEL
60	I	4566138	ACATAGATAGGTTTTAGCTC	-/TT	TTTTTTTTTTCCTACTCAAC INS
60	I	5270735	GTAATATATATATGTATATA	-/TATATACATATT	TATATACATATTTATATACA INS
60	II	1451705	TTATCAACAATGGAAAAATG	-/A	AAAAAAAAAAAGGAAAAACG INS
60	II	1755439	GTTCCATTGTTCAGTTTTTC	-/T	TTTTTTTTTTTGTTTTAAAT INS
60	II	4372634	TTTGTATGCTGTGCATTTAA	-/T	TTTTTTTTTTTTATTTATAAA INS
60	II	2720152	AGACATTGTCTAAAATATAT	-/A	AAAAAAAAAAAACAGAACTG INS
62	I	719597	CTTAGTAAGCATAGTAGCTG	-/TA	TATATATATATATATATATA INS
62	I	4299896	ATTGCGAACGGTTGAGTGTG	-/T	TTTTTTTTACCTTTGTGAGA INS
62	I	5177141	AACTTTAGAAAATTTAGTAT	-/A	AAAAAAATAATGCTTAATTA INS
62	II	1679417	ATTCATTTCATTGGCTGAAT	-/AA	AAAAAAAAAGCATCCGCCAA INS
62	II	3205112	CAGGGGTTTGAAGCAGCTGA	-/T	TTTTTTTATCACGGGCTATT INS
62	II	3538216	GCACATATGTTTATTCAATT	-/A	AAAAAAAAATCGGTCTTGAG INS
63	I	3570217	GAATAAAAACAAAACAAAAT	-/A	AAAAAAAATCACGAGAGCAC INS
63	I	3748983	TTTTTTTTTACATACTCTGT	-/A	AAAAAAAAAAAAAGAATACC INS
63	I	4916245	GAATACGGTTCAATGATGAT	-/A	AAAAAAATGTCAAAATGAAT INS
63	II	631718	GCAAGAAAATATAAGATATT	-/A	AAAAAAAAAAAGATATATTT INS
63	II	862301	CAAGATTGGACATCCACTCC	-/T	TTTTTTTTTTTTAAAATTTA INS
63	II	2430041	TCCTATCACAAGATGCAGAC	-/T	TTTTTTTGGAAAATATTACC INS
63	II	4209406	ATTGAGCATTACTGGCGTTC	-/T	TTTTTTTTTTTTTTAAATAA INS
63	II	4446996	AGCTATGTTCTTTAAAAAAA	-/C	ATGGTGGTCAGCAACAAATA INS
63	III	1244495	CTTCCCTTTTTATATAAGCG	T/-	TTTTTTTTTTAGTGGAAAGT DEL
63	III	2391117	CATAAAAGAAAGAATATATG	-/TA	TATATATATATATATTGACT INS
64	II	740358	GTCGTTTGAACTTAGAAGAT	-/G	GAAAAAAATGAGCTCAAATT INS

64	III	1810978	GGAGAAACCTAAAAGAAATT	-/A	AAAAAAAAAAAATAAAAATA	INS
65	I	5548232	TGTGTATCTACTATTTTTTT	-/G	GGGGGGGGTTTTTTATGTAC	INS
65	II	962149	AGATTAAACAGGAGAAAATG	A/-	AAAAAAAAACCGAGATAGAG	DEL
65	II	1304055	TACAAGTCAAAAAATATGAT	-/A	AAAAAAAATAATATATAATT	INS
65	II	3153754	TTTGAGTGCATTTAATTTAA	T/-	TTTTTTTTTTTTAATTTTGT	DEL
67	I	2249839	AATTATTCTAAAAAAAAATA	-/T	TTTTTTTTTTGAAATTTATT	INS
67	II	114287	AGGTAGATAATTTTTTTTCT	-/A	AAAAAAAAAAGGAATATTGT	INS
67	II	3702248	CCCGTTCGATCATCTACAGC	-/T	TTTTTTTTTATGAAAATCAC	INS
67	II	4008073	CCCATTCAACCAACGTACTG	-/CTA	CTACTACTACTACTACTACT	INS
67	II	4201005	TACACAAGTTCCCAATAAGT	-/A	AAAAAAAAATAGGAGGAGCA	INS
67	III	348705	ATATATATATATATATAATC	-/T	TTTTTTTTTTTTATATATAC	INS
68	I	1526409	ATCCGATATTTGCGAGGGAT	-/A	AAAAAAAATTAATACAATTT	INS
68	II	590352	ATCTACTCCATCCTCTCCAA	CAAAAGGATTC/-	CAAAAGGATGACGAAGAAGA	DEL
68	II	3268664	AAATGCATTTTCTTTTTCCT	-/A	AAAAAAAAAAAAAGAAAACA	INS
68	II	3426037	TATATCGCTTGAGAATTCTG	-/T	TTTTTTTTTGAAAAGGTTCA	INS
69	I	2071188	TGAGCATATCGAGTTTTTGG	T/-	TTTTTTTTTTTTCTACTAGAA	DEL
69	II	677175	ATCTCACTTACGTATGTATG	TA/-	TATATATATATATATATATA	DEL
69	II	1926356	TATAAATTCATTGCCATACA	-/T	TTTTTTTATAAATTCGGATT	INS
69	II	1972068	TCGTTTTCATTAGTATCAAC	-/T	TTTTTTTTATATTTTTTTGC	INS
69	II	3515404	AAGCTGCCTCTCAATTAGCT	-/GCTCATTATA	GCTCATTATAGCTCATTCTA	INS
70	I	1909718	CGTACTTTGAGAAAAAGCAC	A/-	AAAAACTATACGTTTGGCTC	DEL
70	I	3040619	TTTAAGCTTAGTGGTTTAAT	-/A	AAAAAAAATATATACCGCAA	INS
70	I	4281325	ATATTTGAAGCGTCTTTAAA	-/T	TTTTTTTATCCTTTTACTTA	INS
70	III	2049195	TTTCAATAGGGTTCTTTTTC	-/TT	TTTTTTTTTAATATATATAT	INS
71	I	2214725	GTAAATAATTACTACGATAG	-/T	AAAAAAATGATCGTGCCAAG	INS
71	I	2236977	GTCACATAGAAAGGGTATTT	-/A	AAAAAAAAATAAAGGAATTC	INS
71	I	4969785	TTCCGAGACGATTCAAGGAC	-/T	TTTTTTTATTTTTTTTCATA	INS
71	II	2339553	TAAGTAAGAACAAATTAAAT	-/A	AAAAAAAAAATAAGAACAAA	INS
71	II	2701957	ATCTTATTTGTATTTCTACA	-/T	TTTTTTTTGGAAAGAATGAT	INS
71	III	896018	TCATCGTTACATAGATTCAT	-/A	AAAAAAAAAACAAGAGAAAA	INS
71	III	1154189	TGAATAGTGCAAGATTTTAG	-/T	TTTTTTTTTTTTAAATAAAT	INS
71	III	1501758	AGTTTGACTCGTTATTAAAT	-/A	AAAAAAAATAGAGGCAAATT	INS
72	I	442786	TCTCCAAGGTTTCTTTACTC	-/TT	TTTTTTTTTTTTACATTTTT	INS
72	I	1041836	ATAACTCAATATGAAGCATC	-/T	TTTTTTTGCAAACCTTTTAT	INS
72	I	2876870	CCCCCATTCGTTCTTTTTTC	-/T	TTTTTTTTTTTGCTTCCGAC	INS
72	II	842330	AACTGTTGTTCACATAGATG	-/A	AAAAAAAAAAAGATCATTTA	INS
72	II	2969664	TAGGGTGATTAAAGGTCGAC	-/TGGTA	TGGTATGGTATGGTATGGTA	INS

72	III	817773	ACTGATTCACTCATTTGATG	-/A	AAAAAAAATGAGGCATTTTT	INS
74	II	673772	CAACGGTGTTGTATGGTATT	-/AA	AAAAAAAAAAGATACCTTGC	INS
74	II	1454505	AGAGAGTAGAGAAGAGAGTT	-/A	AAAAAAAAAGTGTGTAGTAG	INS
74	II	1779552	AAATTTTAAGTGCATATACC	-/GATT	GATTGATTGATTGATTTGTT	INS
74	II	1854438	TTTAAAAAAGAAAACGATTA	-/AT	ATATATATAATCGATGAAAT	INS
75	I	3841563	TAAATAACTGCGACTTTATT	-/A	AAAAAAACATTAATAAAATG	INS
75	II	4462508	TAGAAATTTTTTATTGATGC	-/AA	AAAAAAAAAAAAAAAACTTA	INS
76	I	1563429	GGGTATAGTCATCCATAGGC	-/A	AAAAAAAGCATATATATTCA	INS
76	I	2782985	TTTCACTGTTATGTCGATAC	-/T	TTTTTTTATTTTTTTCCCAT	INS
76	II	2207093	TTCCCTAATGATCGATCTTC	-/T	TTTTTTTTTTTGTTTCTTTT	INS
76	III	207016	ATAAATTAATTAATGAAATT	-/AAA	AAAAAAAAAAAAAATTGAAT	INS
77	I	3501055	CCGTGAGAGCAAGTCCGCTG	-/AGAAGGACA	AGAAGGACAAGAAGGACAAG	INS
77	III	2067552	AAAATAAAAAATCGAAAACC	-/A	AAAAAAAAAAAAATTTATTA	INS
78	I	598679	TATATAATTAGCAATTTTAT	-/A	AAAAAAAATAAAAAAATAGC	INS
79	III	563214	GGGAAGAGTGGTGTATAGTT	A/-	AAAAAAAAAAAGAGAAAAGG	DEL
80	I	2289760	AGAAGAAGGAAAAATATAAT	-/A	AAAAAAAAAAGAAACAAACA	INS
80	II	846003	AGAAAGAAAGAAAATCTCTG	-/T	TTTTTTTTTTTGTGGGGAGT	INS
81	III	447208	TTCTTTTTTTTCTTTTACTC	-/TTT	TTTTTTTTTTGTTTTCTTCG	INS
81	III	627754	GTGAATGAAGTAAATTGAAT	-/A	AAAAAAAAATAAACGAATTC	INS
81	III	2028192	AAGGGAAACCATATTCTCAT	-/A	AAAAAAAAAACGGCACTAGA	INS
82	I	1359503	GATAATAGGAAATTAGAAAT	-/A	AAAAAAAAAAAAAAAATTTGG	INS
82	I	1830464	ACAAAACAAGATTAAACCGC	-/T	TTTTTTTTTTAAAAGGTTTT	INS
82	I	3014393	TTCCACTACGGGAAATTTTG	-/T	TTTTTTTTTTTTTAATCTTA	INS
82	I	3993829	TAAAAAAAAAAAAAAAAAGG	A/-	AAAAAAAAAAAATCCCCCAC	DEL
82	II	627565	GCCTACATATATACTTGCGC	-/A	AAAAAAAGCAGTCACATAAC	INS
82	II	1088167	AAACACTCAGTGAGCTTGAT	-/A	AAAAAAAAGCAGCGTATTGA	INS
82	II	3473432	GTAAAGAAAAGCAATTGGTT	-/A	AAAAAAAAACATTTCAGGGA	INS
82	II	3607662	CCCCGCTTTCTCATTTGTTA	-/T	TTTTTTTTTTTAATTTTTTT	INS
82	III	423791	AAAAAATATTTTGTATATTA	-/TATATATG	TATATATGTATATATGTATA	INS
82	III	1233349	ATTATGTACAGCTGTTAAAT	-/A	AAAAAAAATAAAAAAATAGC	INS
83	I	460756	TGAGAGAGCATGGTGTAAGT	-/AA	AAAAAAAAAAAGGAATTTAG	INS
83	II	400900	AAGCACTTATAAACATAAAT	-/A	AAAAAAAATAAAATTAATAA	INS
83	III	704878	AAGGTTTATTCATATTTTTT	-/A	AAAAAAAAAAGTGGATTTCG	INS
83	III	1058588	TTATAATCTAACGGAAAGCT	-/A	AAAAAAACGACTGCATTAAC	INS
83	III	2220935	TTCTTTTCTTCTTCTCTTCC	-/TT	TTTTTTTTTTGTAAAGACTG	INS
84	I	553888	CTCTGCTAACCCGAGTTTCT	-/A	AAAAAAAAAAACGAAAGCCG	INS
84	I	1154677	ATTTTCCTTATTTGTTACTA	T/-	TTTTTTTCGTTAATATTCCC	DEL

84	I	3237470	TCGTTCAAAGATATAGAATA	T/-	TTTTTTTTTTTTAATATAAA	DEL
84	I	3798333	CAAAAATAAGACGAGTCACT	-/A	AAAAAAAAACATACCTGGCT	INS
84	I	4008950	TCATTTCCACTACTTTCGAT	-/A	AAAAAAAGATAATGTTCAAT	INS
84	I	4294954	ATCGAGTATTGTTGAATTAT	-/G	GGGGGGGTTCTTTCTTACAT	INS
84	II	363279	TTCCTTTGCTACAGCTTTTA	-/T	TTTTTTTTTCTTTACTGTCT	INS
84	II	1047742	AATAAGCTCTTTAATGATGC	-/T	TTTTTTTTATTAGCTGATTA	INS
84	III	1675050	TAATTACGAAAAGATTAAAT	-/A	AAAAAAAATGAAAATATTTG	INS
85	I	3032617	ATAATTTTATTTATTGAATG	GTTTTATTTGTTATATGAATA/-	GTTTTATTTGTTGAATGAAT	DEL
85	I	3751488	GATTTCCAATCAACTTATTT	-/AAA	AAAAAAAAAAAAGAAAAAAA	INS
85	I	4198907	CAATGACCACCATTTCCTTC	-/T	TTTTTTTAGTGAAATAATTG	INS
85	I	4232151	GGAAATCGCAATCCATTCTA	-/T	TTTTTTTTTTACCTATTTAT	INS
85	I	4251918	CAAATTAATAGCTTAAGATC	-/T	TTTTTTTCAGTTTGCCAAAT	INS
85	I	4723298	TTTTCCTTCCTTTTTATTGG	A/-	AAAAAAAAAAGAAAACAAAA	DEL
85	II	109227	TCTTTAATCATTTTGTTTTG	-/T	TTTTTTTTTTTAAAGCAAAA	INS
85	II	1187402	AAAGTTTCGCCAAAAATTGC	-/T	TTTTTTTTTTATCTCAAAAT	INS
85	III	2044344	CTTTGCTTTGCTTGGCTTGG	-/CTCGACTTGACTTGACTCGACTTGACTTGACTCTT		INS
87	III	1239214	TAAAATGGGTAGGGAAATGT	-/A	AAAAAAAAAAAAAAAGAAGA	INS
88	I	919796	AAAATACTGTTCGATGTAGA	-/T	TTTTTTTTTTAAAAAGACAT	INS
88	I	1699805	GGTTAATATAGTTGATCCTG	-/T	TTTTTTTTCTTGTTTTTTCT	INS
88	I	4796843	TGTATGAGTTATTGACTGAT	-/A	AAAAAAACTAGATGCAGGCA	INS
88	III	1044612	AAAAAGATTAACTTATTTCC	-/T	TTTTTTTTTTTTAATTTTTCT	INS
89	I	110660	TTTAATCAAGGATAAGAATA	-/T	TTTTTTTTGAATGAGTAATA	INS
89	I	1658327	GTTTTGGATAGTCTCAAATA	-/ATTAT	ATTATATTATTTAAAGGCTT	INS
89	I	5534399	GATATATAGAATAAATTGTG	A/-	AAAAAAAAAAAGTTAATATT	DEL
89	III	1556350	AGTCTTTTTTTGCTCTACAG	-/T	TTTTTTTTTTTTTAAACTTT	INS
90	I	588541	CCTCTGAGGAATTTTTTATC	-/TT	TTTTTTTTTAAAATACCAAC	INS
90	I	954072	TGCTGTTCAACATTCTTGGT	-/A	AAAAAAAAAAAAAAGGTTTT	INS
91	I	101454	AATAGATTCAGATGCGATAT	-/A	AAAAAAAAGTAGTATCCATT	INS
91	I	1158207	TAAGCAAATGCGTCAACGGC	-/A	AAAAGGAATATTATAAGGCA	INS
91	I	1509738	GTTATTCCTCGAGTAATTAG	A/-	AAAAAAAAAAAAGCAGCTTT	DEL
91	I	5522135	TAAAGATGTTTTGTTGGATA	-/C	AAAATTTCATGTAATCACCA	INS
91	II	1081008	CGAAGTGCATTTTTATTCAA	-/T	TTTTTTTTTACGTTTTTCTC	INS
91	II	1747739	GTAAAAAGCTCGACTTCTGC	-/A	AAAAAAACTGCTAATATAGA	INS
92	I	286307	CTGGGGTTATGCCTGCTTTT	-/CCACCA	CCACCACCACCACCACCTCC	INS
92	I	380710	AATCGTACCAGAAAAAATCG	-/ATCA	ATCAATCAATCAATCAATGG	INS
92	I	1206603	TGGGTCTAGCCGGTGATCAG	-/T	TTTTTTTTTTTTCCTTTCCA	INS
92	III	818071	CATATTGGTCGAAAAATTGG	-/T	TTTTTTTTAGCAAAATATAG	INS

92	III	1370083	CTCCAACCATAGTACTAACG	-/T	AGGCCCTCAGACGCTTAACT	INS
93	III	1780525	TAAACAATTGAGCTGAATAT	-/A	AAAAAAAAAACAAAGATATT	INS
94	I	1834072	TGGATTTATTATTATTTAAA	-/T	TTTTTTTTATCTAATTTACT	INS
94	I	3623774	AGAACAAAAATTTACAACCC	-/T	TTTTTTTTTTAATTAAGAGC	INS
94	II	3862802	CAAACTTGACATAGCTTAGC	-/TTT	TTTTTTTTTTCTATTTTGCT	INS

IV. List of Double SNMs

Line	Chromosome	Position	REF	ALT
10	I	4059081	C	T
10	I	4059082	T	G
15	I	3830474	A	C
15	I	3830475	T	C
16	II	2790354	A	G
16	II	2790355	A	G
23	II	2023222	G	T
23	II	2023223	A	T
26	III	174648	G	A
26	III	174652	T	C
32	I	3571186	C	A
32	I	3571187	C	A
45	II	326319	T	C
45	II	326320	G	A
50	I	4494400	A	G
50	I	4494403	C	T
52	II	2768291	C	T
52	II	2768292	A	G
56	I	3984777	A	T
56	I	3984778	G	C
56	I	4059081	C	T
56	I	4059082	T	G
59	II	3914086	T	C
59	II	3914087	G	A
70	II	4423099	A	G
70	II	4423104	A	T
76	II	4031823	T	C
76	II	4031824	T	A
82	II	2768310	C	T
82	II	2768311	T	G
87	II	2023192	A	T
87	II	2023193	G	T
91	I	4059062	C	T
91	I	4059063	A	G

V. List of Complex Mutations

Line	Chromosome	Position	Ref	Alt
2	III	2037436	T	TC
2	III	2037437	T	A
2	II	689218	T	C
2	II	689241	C	T
2	II	689243	T	G
8	II	1774914	T	C
8	II	1774915	C	CT
10	II	690842	C	G
10	II	690861	T	C
10	II	690863	G	T
17	I	4268844	G	GC
17	I	4268845	T	A
37	I	5556704	A	G
37	I	5556708	A	G
37	I	5556711	A	G
38	III	615646	G	A
38	III	615666	A	G
38	III	615667	T	A
38	III	615676	C	T
38	III	615691	G	A
38	III	615693	T	C
38	III	615712	A	G
38	III	615713	A	G
38	III	615721	A	G
38	III	615735	T	C
39	I	5263416	G	C
39	I	5263417	G	T
39	I	5263418	T	C
39	I	5263424	A	G
39	I	5263425	A	G
45	I	2942976	C	T
45	I	2942979	A	C
45	I	2942982	G	A
46	II	1751125	C	T
46	II	1751126	T	C
46	II	1751127	C	T
46	II	1751130	C	T
54	II	154443	T	C
54	II	154446	G	A
54	II	154464	T	G
54	II	154533	A	G
54	II	154545	G	A
54	II	154550	C	T
54	II	154551	A	G
54	II	154554	A	T
54	II	154560	C	T
56	I	3988202	A	T
56	I	3988205	G	T
56	I	3988223	A	G
56	I	3988247	A	T
56	I	3988361	A	G
56	I	3988364	C	T

56	I	3988373	A	T
56	I	3988379	G	A
56	I	3988388	A	T
56	I	3988391	G	T
56	I	3988409	A	G
56	I	3988454	T	G
56	I	3988457	C	T
56	I	3988466	A	T
56	I	3988472	G	A
56	I	3988667	G	A
56	I	3988670	A	G
56	I	3988676	G	A
56	I	3988679	A	G
56	I	3988683	G	C
56	I	3988687	A	G
56	I	3988940	A	G
56	I	3988943	G	A
56	I	3988952	G	A
56	I	3989000	G	A
58	II	4273754	G	A
58	II	4273772	G	A
58	II	4273787	G	A
58	II	4273790	A	G
58	II	4273794	T	C
65	II	4422994	A	G
65	II	4422996	C	T
65	II	4423624	A	G
65	II	4423629	A	T
91	III	228029	A	G
91	III	228032	G	T
91	III	228033	G	C
91	III	228035	A	T
91	I	3988427	G	A
91	I	3988430	C	T
91	I	3988431	A	G
91	I	3988433	T	A
92	II	696922	T	C
92	II	696926	C	G
92	II	696988	C	T
92	II	697019	G	C

VI. List of Medium Deletions

Line	Chromosome	Position	Deletion Size
13	I	2942921	190
56	I	3988243	205
65	II	4423006	628

VII. Primers for Sanger Sequencing

Name	Sequence	Notes	Variant Type
SpChr3_2143380F	TGCTGCTCAACCGTATGTAG	L13_Chro3_ 2143868_Del_T	Del
SpChr3_2144137R	TTCGGAAACCGTGAACTAAT		
SpChr2_2576814F	TGGTTGATTCGTCAGCTCAT	L13_Chro2_ 577240_Del_ATCTTCTCCTTGTTATTCTATG	Del
SpChr2_577577R	CCCTTCCATCATCCATCAAT		
SpChr2_2373318F	TGGCTGGTAAGCACTTGTGA	L13_Chro2_2373654_Ins_T	Ins
SpChr2_2373986R	TGCAGGCAAAGTAACAACCA		
SpChr2_2679974F	CGCCCAGGAATGACTTTTAC	L13_Chro2_2680220_bp_C-A	SNM
SpChr2_2680635R	GAAAGGAACGCGTTTTGAAG		
SpChr2_3621917F	GCTTTTCATCAACCCGAAAA	L13_Chro2_3622335_Ins_A	Ins
SpChr2_3622618R	GCACATGGCATATCAATCCA		
SpChr1_543479F	TTTGATCGCGCATCTTCATA	L13_Chro1_543809_Ins_A	Ins
SpChr1_544090R	CAGAAATAAGCCAAATTTTATCACG		
SpChr1_1151913F	CTGACGACTCACTGGGGAAT	L13_Chro1_1152144_Ins_T	Ins
SpChr1_1152369R	ATGGCATCAAAGGGAGAGTG		
SpChr1_1959555F	CCAAGTAGCGCACAATCTCA	L13_Chro1_1960047_bp_C-T	SNM
SpChr1_1960278R	TAACGGACGGACATGAACAA		
SpChr1_3854412F	CCGCTGTTTCCGATTTACC	L13_Chro1_3854681_bp_A-G	SNM
SpChr1_3854923R	TCCTTTGCAATGAAGCAAGTC		
SpChr1_4474284F	CCTCTTTACTGGTCCCTCACC	L13_Chro1_4474733_Ins_A	Ins
SpChr1_4475076R	TACTTGTCGCGTTGCTTACG		
SpChr3_174357F	TCCTATGGAGGAAATAACAACGA	"L26_Chro3_174648_bp_G-A, 174652_bp_T-C"	Double
SpChr3_174988R	TTGCTGCCGTAGGAAGAGTT		
SpChr1_2977016F	CGGTCCAAATTCGACTATTCTT	L26_Chro1_2977255_Ins_T	Ins
SpChr1_2977646R	CTCGGAGTTTAGCGATACGG		
SpChr1_4437951F	AGTGCGCTGGAGTACAGACA	L26_Chro1_4438285_Ins_T	Ins
SpChr1_4438513R	ATACTGTCAAAGCGCAAGCA		
SpChr3_764080F	AGAGCGGCAGTAAGCAAGAG	58_Chro3_763965_GTTTAA-G_Del	Del
SpChr3_764420R	CAAACAAAGCCACAAAAGGA		
SpChr3_2268000F	ACGCCGTGGTATTTCGTATC	58_Chro3_2268438_T-C_SNP	SNM
SpChr3_2268856R	TCGCCAAAACCCAATCTTTA		
SpChr2_2570066F	AGAGGAGCTTCTTCGGCAAA	58_Chro2_2570463_T-C_SNP	SNM
SpChr3_2570941R	TCCGCTAGCATTATCATTCATT		
SpChr2_2805497F	TGCGAACCAAGTGTATGGAA	58_Chro2_2805811_G-GTGTATA_In	Ins
SpChr2_2806300R	CATTCCCAAAACCAACGAAC		

SpChr2_3282851F	GTTGCGCAATGTAAGCTACG	58_Chro2_3283184_G-A_SNP	SNM
SpChr2_3283630R	TCCAGATAAATAGGGGCATTG		
SpChr2_3975812F	CAGAGCAAACCTCCAACCTT	58_Chro2_3976264_T-TA_In	Ins
SpChr2_3976598R	TTGGTTTCCCTTTTCCGTTA		
SpChr2_4273519F	GTCCTCGAGAGCCCCTTTAG	58_Chro2_4273754_4273772_4273787_4273790_4273794_G_A_Complex	
Complex			
SpChr2_4274319R	CATGCCATCACGACAAGTTC		
SpChr2_4357594F	GCTTAATAGTTGCGCCGTTT	58_Chro2_4357961_C-CT_In	Ins
SpChr2_4358425R	CCCAAAGCTCAGCGTTATTC		
SpChr1_237376F	CGACTGTGCAAGAATGAGGA	58_Chro1_237839_T-TA_In	Ins
SpChr1_238181R	CTGTTTTGCTAGCCGTCCA		
SpChr1_1072625F	CTCTTGAATTTGGCGAAACG	58_Chro1_1072991_A-ATATATATATG_In	Ins
SpChr1_1073426R	AAAGGTCCGGTTTCGAAGAG		
SpChr1_1115334F	GCCTCAGCCTGCTTCTATGT	58_Chro1_1115695_G-GC_In	Ins
SpChr1_1116137R	GCGCAAATAGGTGAATACGTG		
SpChr1_3295873F	ATGAAATTGCCGACGACTCT	58_Chro1_3296352_G-T_SNP	SNM
SpChr1_3296657R	AGGCCCTAGGACGTCTCTTC		
SpChr1_3340110F	GATTTGGAAACCGCGAAAT	58_Chro1_3340501_C-G_SNP	SNM
SpChr1_3340943R	TTGTCATCCTTAAGAGCATCCA		
SpChr1_4460325F	ATGCCTGATTTACCGCACTT	58_Chro1_4460676_C-T_SNP	SNM
SpChr1_4461127	GCGTTCATGTCGTTCTTTCTC		
SpChr3_446909F	CGTCTGTCTTTGCAGCAATC	81_Chro3_447208_C-CTTT_In_	Ins
SpChr3_447691R	GACGGACGAGGATGAAGAAA		
SpChr3_2027747F	CGGATCCAGCGTACCTTTAC	81_Chro3_2028192_T-TA_In	Ins
SpChr3_2028547R	CGTGAAGCTGGTCTTGATGA		
SpChr2_3670888F	GCTCCAGAATTTGTGCCACT	81_Chro2_3671308_C-G_SNP	SNM
SpChr2_3671747R	TGGTTTCAGCCAATTTCGTT		
SpChr2_4066286F	CAATCAACATGCCCAAATCA	81_Chro2_4066663_C-A_SNP	SNM
SpChr2_4067075R	AGCCATGGACTCTACGCATT		
SpChr1_3082845F	CCTTAGCAGAAAGTGCTTCCA	81_Chro1_3083202_G-A_SNP	SNM
SpChr1_3083636R	CGCGTTGAAGGTGATTTTGT		
SpChr3_817740F	GATGCTGGATTCCGACTGAT	92_Chro3_818071_G-GT_In	Ins
SpChr3_818519R	TTGGGAATTTTGAGCCATAGA		
SpChr3_1369691F	TCAGGGCTGCTTTTTAATGA	92_Chro3_1370083_G-GT_In	Ins
SpChr3_1370405R	CGGGAATAGACGCTGTTGAG		
SpChr2_1196060F	CCGGAGTCGATTATCCAAAA	92_Chro2_1196504_G-T_SNP	SNM
SpChr2_1196931R	CGACCACGAAGTTCTCCATT		

SpChr1_285813F	CGGGATCTAAGGGAGCAAAT	92_Chro1_286307_T-TCCACCA_In	Ins
SpChr1_286591R	GCTTCCTTCATCGCAAAACT		
SpChr1_380275F	GCCTCAAAGAGCAAGAGTGC	92_Chro1_380710_G-GATCA_In	Ins
SpChr1_381063R	TTGAAAAAGTGCGACCCAT		
SpChr1_1206242F	TACGATGAGTTCCCCAGCTT	92_Chro1_1206603_G-GT_In	Ins
SpChr1_1207023R	TGCCAATGCAAATCAAGTTT		
SpChr1_1612806F	ATGCGGCCATTGCTACTATC	92_Chro1_1613171_G-A_SNP	SNM
SpChr1_1613620R	TGAACCATCTGCATTATCGAA		
SpChr1_2581721	ACAATTTTGGGGTTTGCATC	92_Chro1_2582088_C-A_SNP	SNM
SpChr1_2582481	GAAGAGGGTCGTCGTGTTGT		
SpChr1_4675475F	CTATGCCATGCACTTTGGTG	92_Chro1_4675905_G-T_SNP	SNM
SpChr1_4676251R	GGGTGAAAAGTAGATCAAGGAA		

Supplemental Figures

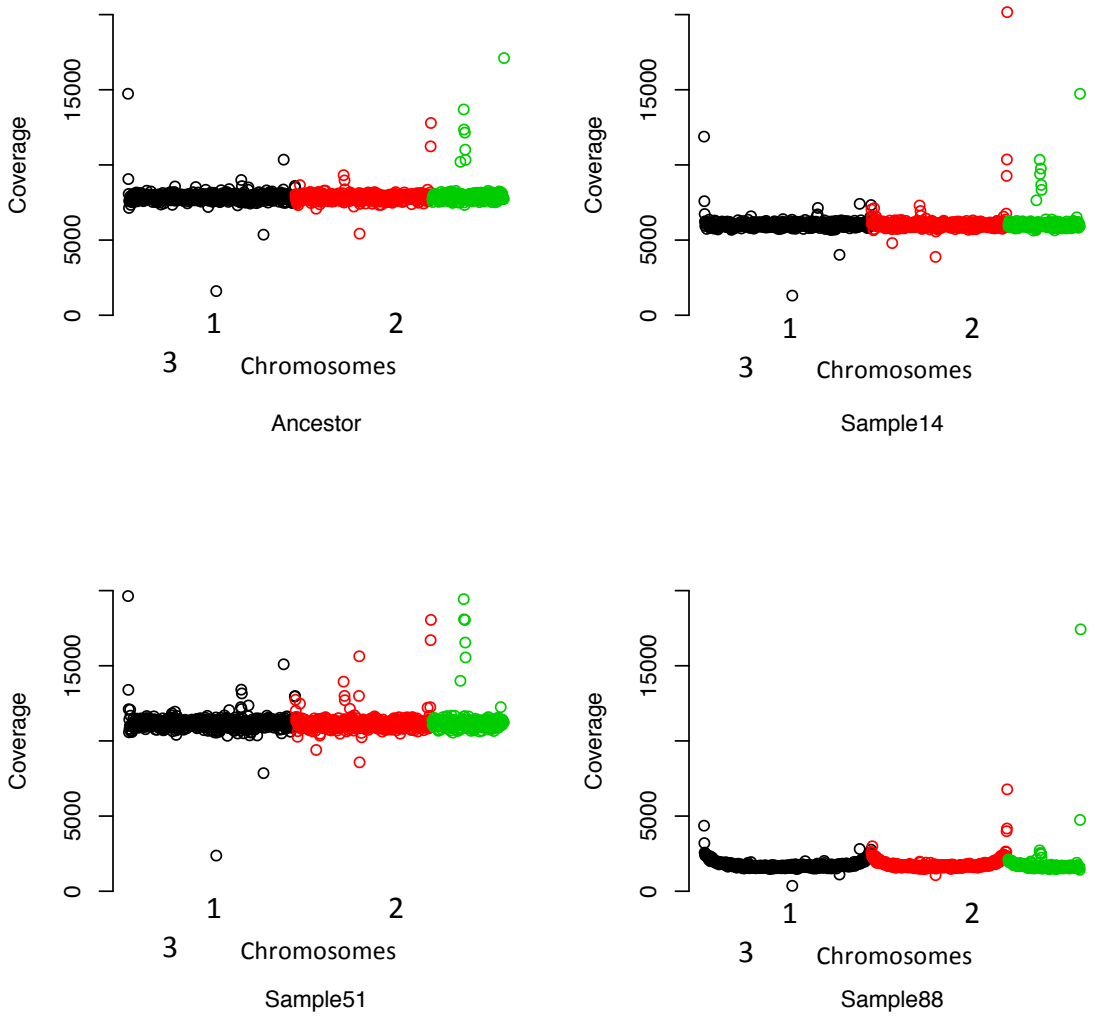

Figure S1. For all lines, uniform coverage was achieved across the genome except for repetitive elements. Black, red, and green represent chromosomes I, II, and III, respectively. Coverage (y-axis) was determined by number of reads mapping to 10 Kb windows across the genome (x-axis).

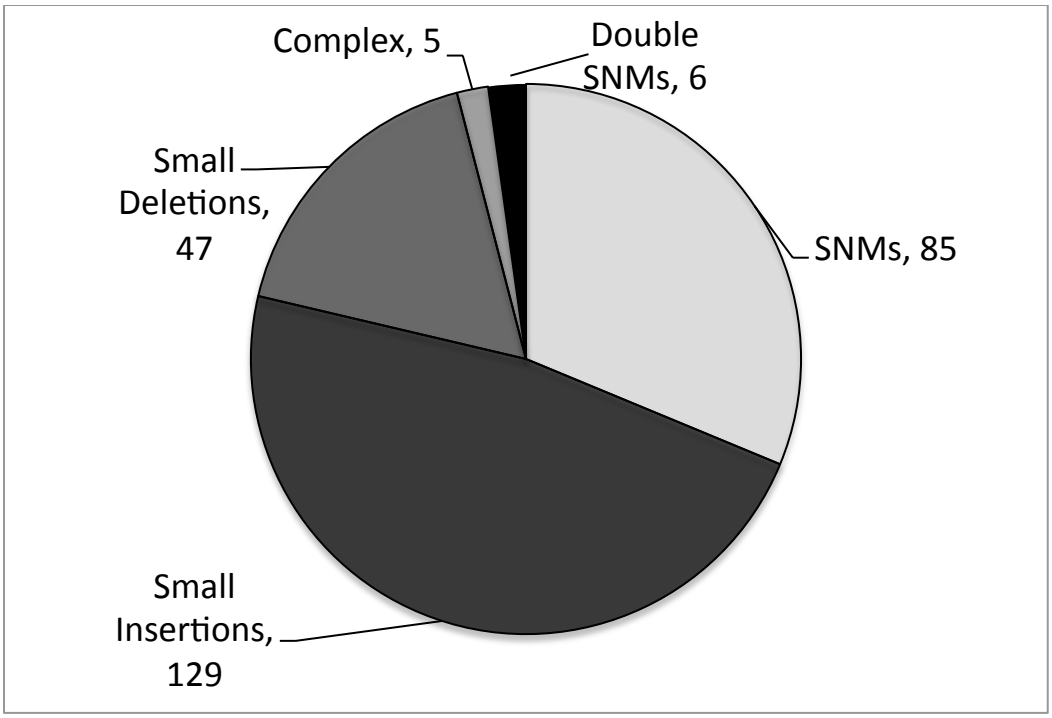

Figure S2. Summary of differences identified between MA ancestor and *Sc. pombe* reference genome.

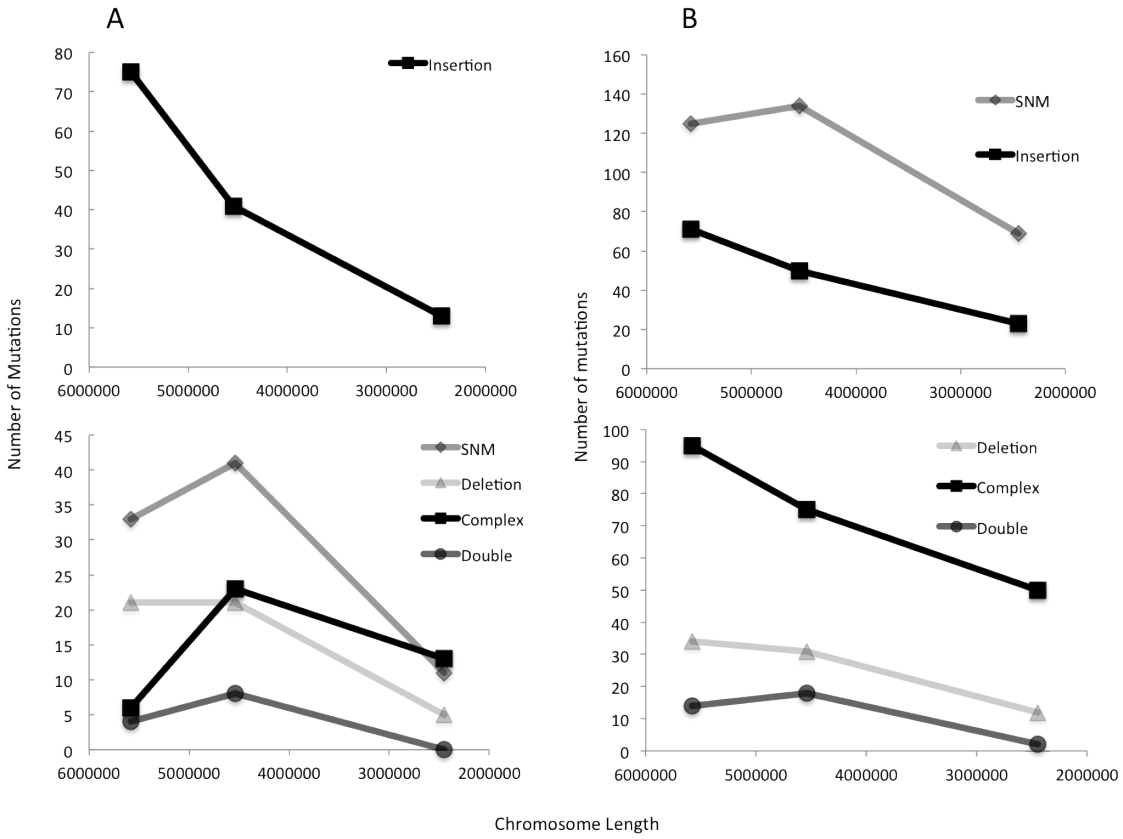

Figure S3. Summary of accumulation for each mutation type across each chromosome in Ancestor (A) and MA lines (B).

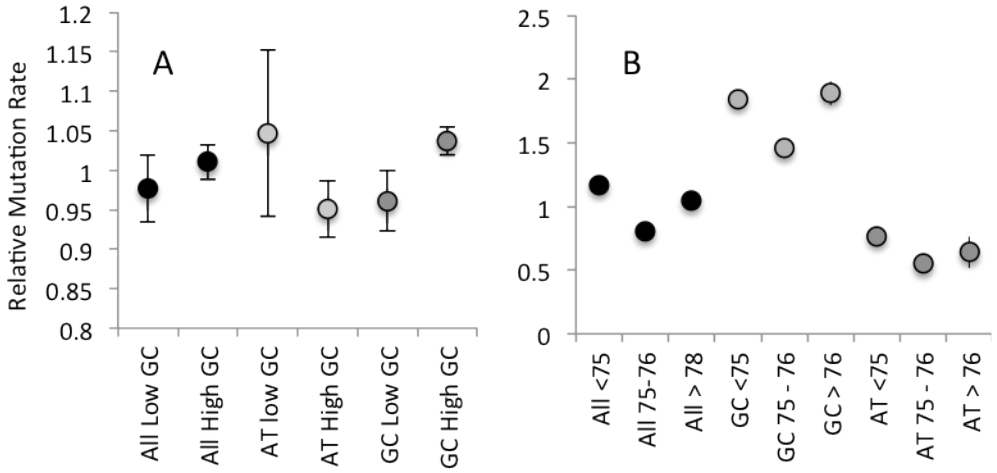

Figure S4. Relative mutation rates at all bases, only G/C bases, and only A/T bases with respect to local (A) 1-kb GC content and (B) replication time during cell cycle. Categories were defined by obtaining bins that contain roughly equal numbers of SNMs.

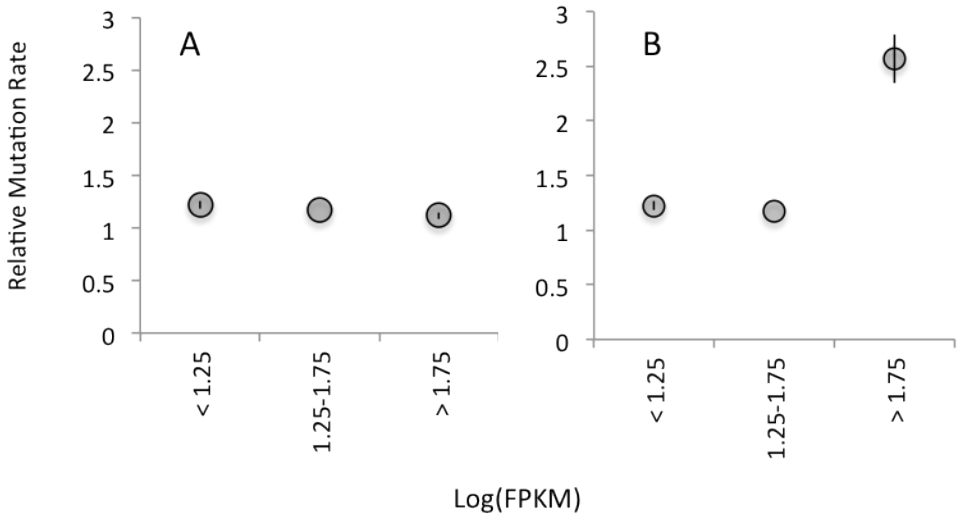

Figure S5. Relative mutation rates at all bases respect to local transcription rate. Categories were defined by obtaining bins that contain roughly equal numbers of SNMs. A) All SNMs excluding SNMs that occur in rRNA and tRNA genes which are not captured in mRNA libraries. B) All SNMs including SNMs that occur in rRNA and tRNA genes assuming high expression.

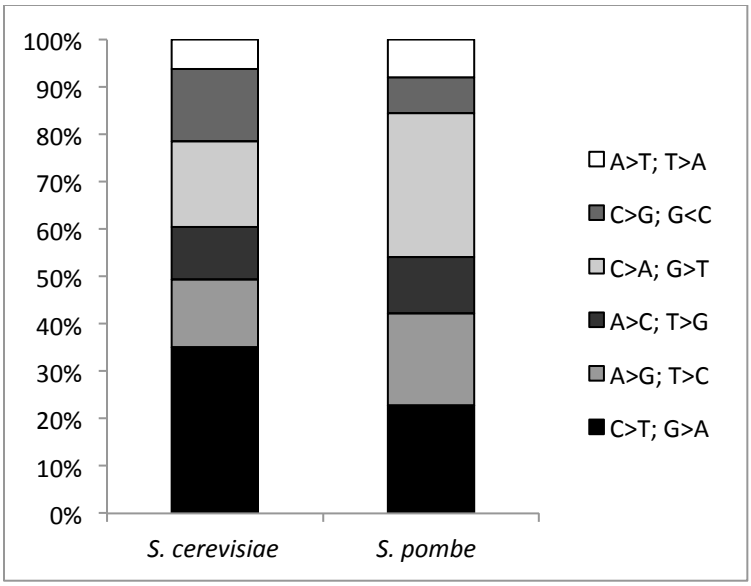

Figure S6. Summary of mutations for each of six possible nucleotide changes for *S. cerevisiae* MA lines and *Sc. pombe* MA lines.

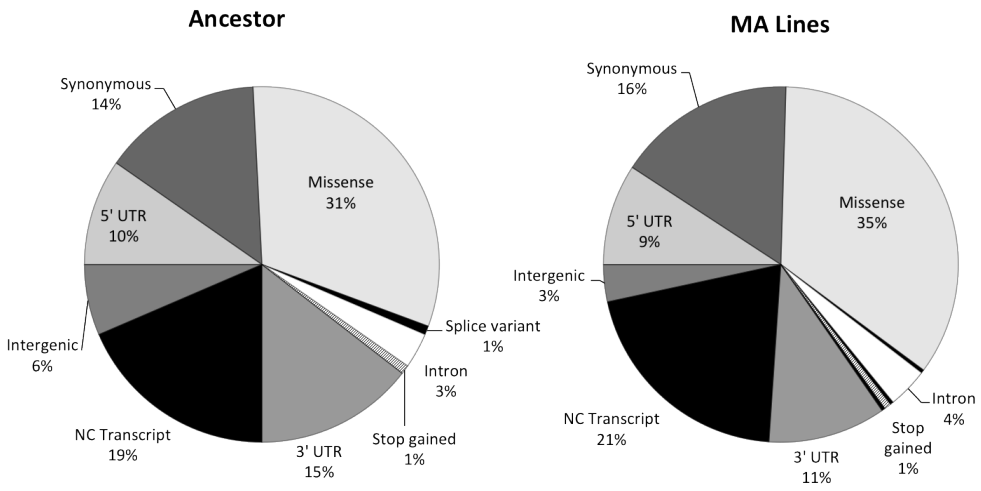

Figure S7. Summary of all effects of SNMs identified in the Ancestor and MA lines. Numbers for MA lines represent numbers of SNMs identified in each genomic region across all 79 lines. Numbers of effects does not sum to number of mutations since some mutations had multiple effects.

APPENDIX B:

SUPPLEMENTARY MATERIAL FOR RATES AND BIASES OF MITOTIC GENE CONVERSION IN *SACCHAROMYCES CEREVISIAE*

Order of Appendix B:
I. Commands for Assembly of Gene Conversion lines
II. List of Gene Conversion Events
III. Tables
IV. Figures

I. Commands for Assembly of Gene Conversion lines

```
#!/bin/bash

#$ -q rcc-30d

# S.c. Gene Conversion
date

AR=(<SAMPLES>)

for i in "${AR[@]}"
do
      cd Sample${i}
      gunzip *
      cat *R1_001.fastq > Sample${i}_R1.fastq
      cat *R2_001.fastq > Sample${i}_R2.fastq

      # clean and trim
      echo "#!/bin/bash" > GC${i}_qc_clean.sh
      echo "" >> GC${i}_qc_clean.sh
      echo "#$ -q rcc-30d" >> GC${i}_qc_clean.sh
      echo "" >> GC${i}_qc_clean.sh
      echo "time /usr/local/ea-utils/latest/bin/fastq-mcf
/panfs/pstor.storage/scratch/dwhlab/Megan/Pombe_MA/Adaptors.fasta
Sample${i}_R1.fastq Sample${i}_R2.fastq -o
Sample${i}_R1_filtered.fastq -o Sample${i}_R2_filtered.fastq -C
1000000 -q 20 -p 10 -u -x 0.01" >> GC${i}_qc_clean.sh
      echo "" >> GC${i}_qc_clean.sh
      echo "time /usr/local/fastx_toolkit/latest/bin/fastx_trimmer -
Q33 -f 10 -l 110 -i Sample${i}_R1_filtered.fastq -o
Sample${i}_R1_trimmed.fastq" >> GC${i}_qc_clean.sh
      echo "" >> GC${i}_qc_clean.sh
```

```
      echo "time /usr/local/fastx_toolkit/latest/bin/fastx_trimmer -
Q33 -f 10 -l 110 -i Sample${i}_R2_filtered.fastq -o
Sample${i}_R2_trimmed.fastq" >> GC${i}_qc_clean.sh
      echo "" >> GC${i}_qc_clean.sh
      echo "qsub GC${i}_Assemble.sh" >> GC${i}_qc_clean.sh
      echo "" >> GC${i}_qc_clean.sh
      echo "/usr/local/fastqc/latest/fastqc -o
/panfs/pstor.storage/scratch/dwhlab/Megan/GeneConv_2/fastq_out2
Sample${i}_R1_trimmed.fastq" >> GC${i}_qc_clean.sh
      echo "/usr/local/fastqc/latest/fastqc -o
/panfs/pstor.storage/scratch/dwhlab/Megan/GeneConv_2/fastq_out2
Sample${i}_R2_trimmed.fastq" >> GC${i}_qc_clean.sh
      echo "" >> GC${i}_qc_clean.sh
      echo "exit" >> GC${i}_qc_clean.sh

      # BWA and GATK
      echo "#!/bin/bash" > GC${i}_Assemble.sh
      echo "" >> GC${i}_Assemble.sh
      echo "#$ -q rcc-30d " >> GC${i}_Assemble.sh
      echo "" >> GC${i}_Assemble.sh
      echo "export PATH=/usr/local/samtools/0.1.19/:${PATH}" >>
GC${i}_Assemble.sh
      echo "echo \"BWA\" >&2" >> GC${i}_Assemble.sh
      echo "time /usr/local/bwa/latest/bwa aln
../ReferenceGenome/SCerevisiae.RefGenome.fa
Sample${i}_R1_trimmed.fastq >Sample${i}_1.sai" >> GC${i}_Assemble.sh
      echo "" >> GC${i}_Assemble.sh
      echo "time /usr/local/bwa/latest/bwa aln
../ReferenceGenome/SCerevisiae.RefGenome.fa
Sample${i}_R2_trimmed.fastq >Sample${i}_2.sai" >> GC${i}_Assemble.sh
      echo "" >> GC${i}_Assemble.sh
      echo "time /usr/local/bwa/latest/bwa sampe
../ReferenceGenome/SCerevisiae.RefGenome.fa Sample${i}_1.sai
Sample${i}_2.sai Sample${i}_R1_trimmed.fastq
Sample${i}_R2_trimmed.fastq > Sample${i}.sam" >> GC${i}_Assemble.sh
      echo "qsub GC${i}_compress_fastq.sh" >> GC${i}_Assemble.sh
      echo "" >> GC${i}_Assemble.sh
      echo "echo \"SAMTOOLS\" >&2" >> GC${i}_Assemble.sh
      echo "/usr/local/samtools/latest/samtools view -bS -T
../ReferenceGenome/SCerevisiae.RefGenome.fa Sample${i}.sam >
Sample${i}.bam" >> GC${i}_Assemble.sh
      echo "" >> GC${i}_Assemble.sh
      echo "/usr/local/samtools/latest/samtools sort Sample${i}.bam
Sample${i}.sorted" >> GC${i}_Assemble.sh
      echo "" >> GC${i}_Assemble.sh
      echo "/usr/local/samtools/latest/samtools index
Sample${i}.sorted.bam" >> GC${i}_Assemble.sh
      echo "" >> GC${i}_Assemble.sh
      echo "echo \"PICARD\" >&2" >> GC${i}_Assemble.sh
      echo "java -Xmx2g -classpath \"/usr/local/picard/latest/\" -
jar /usr/local/picard/latest/AddOrReplaceReadGroups.jar
I=Sample${i}.sorted.bam O=Sample${i}.sorted.fixed.bam
SORT_ORDER=coordinate RGID=GeneConv RGLB=bar RGPL=illumina
RGSM=Sample${i} RGPU=6 CREATE_INDEX=True
VALIDATION_STRINGENCY=LENIENT" >> GC${i}_Assemble.sh
```

```
      echo "" >> GC${i}_Assemble.sh
      echo "java -Xmx2g -classpath \"/usr/local/picard/latest/\" -
jar /usr/local/picard/latest/MarkDuplicates.jar
I=Sample${i}.sorted.fixed.bam O=Sample${i}.sorted.fixed.marked.bam
M=Sample${i}.metrics CREATE_INDEX=True
VALIDATION_STRINGENCY=LENIENT" >> GC${i}_Assemble.sh
      echo "" >> GC${i}_Assemble.sh
      echo "echo \"GATK\" >&2" >> GC${i}_Assemble.sh
      echo "" >> GC${i}_Assemble.sh
      echo "java -Xmx2g -classpath \"/usr/local/picard/latest/\" -
jar /usr/local/gatk/latest/GenomeAnalysisTK.jar -R
../ReferenceGenome/SCerevisiae.RefGenome.fa -I
Sample${i}.sorted.fixed.marked.bam -T RealignerTargetCreator -o
Sample${i}.intervals" >> GC${i}_Assemble.sh
      echo "" >> GC${i}_Assemble.sh
      echo "java -Xmx2g -classpath \"/usr/local/picard/latest/\" -
jar /usr/local/gatk/latest/GenomeAnalysisTK.jar -R
../ReferenceGenome/SCerevisiae.RefGenome.fa -I
Sample${i}.sorted.fixed.marked.bam -T IndelRealigner -
targetIntervals Sample${i}.intervals -o
Sample${i}.sorted.fixed.marked.realigned.bam" >> GC${i}_Assemble.sh
      echo "" >> GC${i}_Assemble.sh
      echo "java -Xmx2g -classpath \"/usr/local/picard/latest/\" -
jar /usr/local/gatk/latest/GenomeAnalysisTK.jar -T BaseRecalibrator
-I Sample${i}.sorted.fixed.marked.realigned.bam -R
../ReferenceGenome/SCerevisiae.RefGenome.fa -rf BadCigar -knownSites
Sample${i}.sorted.fixed.marked.realigned.vcf -o
Sample${i}.recal_data.grp" >> GC${i}_Assemble.sh
      echo "" >> GC${i}_Assemble.sh
      echo "java -Xmx2g -classpath \"/usr/local/picard/latest/\" -
jar /usr/local/gatk/latest/GenomeAnalysisTK.jar -R
../ReferenceGenome/SCerevisiae.RefGenome.fa -I
Sample${i}.sorted.fixed.marked.realigned.bam -T PrintReads -rf
BadCigar  -o Sample${i}.mapped.bam -BQSR Sample${i}.recal_data.grp"
>> GC${i}_Assemble.sh
      echo "" >> GC${i}_Assemble.sh
      echo "exit" >> GC${i}_Assemble.sh

      # compress fastq again
      echo "#!/bin/bash" > GC${i}_compress_fastq.sh
      echo "" >> GC${i}_compress_fastq.sh
      echo "#$ -q rcc-30d" >> GC${i}_compress_fastq.sh
      echo "" >> GC${i}_compress_fastq.sh
      echo "bzip2 Sample${i}_R*.fastq" >> GC${i}_compress_fastq.sh
      echo "rm Sample${i}_R*_filtered.fastq" >>
GC${i}_compress_fastq.sh
      echo "bzip2 Sample${i}_R*_trimmed.fastq" >>
GC${i}_compress_fastq.sh
      echo "" >> GC${i}_compress_fastq.sh
      echo "exit" >> GC${i}_compress_fastq.sh


      chmod u+x GC${i}_qc_clean.sh
      chmod u+x GC${i}_Assemble.sh
      chmod u+x GC${i}_compress_fastq.sh
```

```
        qsub GC${i}_qc_clean.sh
        cd ..
done
```

II. List of Gene Conversion Events

Sample	Start Chromosome	Left	Genotype	Right	Genotype	End Left	Right	Genotype	TractLength
5	ref|NC_001136|	125130	Het	125401	D	136920	D	136940	Het 10801
	ref|NC_001140|	6569	Het 6947	D	8441	D	9035	Het 1494	
	ref|NC_001141|	20218	Het 24816	Y	28464	Y	28539	Het 3648	
	ref|NC_001141|	29044	Het 29410	Y	40512	Y	40928	Het 11102	
	ref|NC_001141|	41898	Het 41905	Y	50712	Y	51552	Het 8807	
	ref|NC_001141|	51756	Het 51897	Y	57794	Y	58256	Het 5897	
	ref|NC_001144|	419474	Het	419667	Y	421733	Y	422064	Het 2013
	ref|NC_001144|	423854	Het	424103	D	451216	D	451337	Het 27113
	ref|NC_001144|	460694	Het	490579	Y	507213	Y	507450	Het 16634
	ref|NC_001144|	507483	Het	507486	Y	527873	Y	527967	Het 20387
	ref|NC_001144|	528191	Het	528220	Y	541310	Y	541474	Het 13090
	ref|NC_001144|	541474	Het	541508	Y	708595	Y	708955	Het 167087
	ref|NC_001144|	709138	Het	709175	Y	757400	Y	757472	Het 48225
	ref|NC_001144|	757478	Het	757576	Y	789072	Y	789096	Het 31496
	ref|NC_001144|	789269	Het	789615	Y	1059505	Y	1059564	Het 269890
	ref|NC_001145|	8262	Het 8279	Y	120376	Y	120604	Het 112097	
	ref|NC_001145|	120628	Het	120749	Y	133610	Y	133858	Het 12861
4	ref|NC_001136|	626028	Het	626607	Y	668861	Y	668949	Het 42254
	ref|NC_001136|	668967	Het	669064	Y	778722	Y	778775	Het 109658
	ref|NC_001136|	778811	Het	778856	Y	801939	Y	801948	Het 23083
	ref|NC_001136|	801948	Het	801982	Y	851108	Y	851114	Het 49089
	ref|NC_001136|	851116	Het	851208	Y	891443	Y	891837	Het 40235
	ref|NC_001136|	891837	Het	892057	Y	909656	Y	909727	Het 17599
	ref|NC_001136|	910688	Het	910905	Y	921813	Y	921827	Het 10908
	ref|NC_001136|	922174	Het	922223	Y	993018	Y	993020	Het 70795
	ref|NC_001136|	993020	Het	993025	Y	1010799	Y	1010858	Het 17774
	ref|NC_001136|	1010886	Het	1010910	Y	1055183	Y	1055184	Het 44273
	ref|NC_001136|	1055185	Het	1055277	Y	1059600	Y	1059923	Het 4323
	ref|NC_001136|	1059932	Het	1059944	Y	1120367	Y	1120541	Het 60597
	ref|NC_001136|	1120541	Het	1120669	Y	1151169	Y	1151209	Het 30500
	ref|NC_001136|	1151314	Het	1151324	Y	1159396	Y	1159464	Het 8072
	ref|NC_001136|	1159479	Het	1159689	Y	1163806	Y	1163953	Het 4117
	ref|NC_001136|	1164350	Het	1164448	Y	1233618	Y	1233625	Het 69170
	ref|NC_001136|	1233634	Y	1233967	Y	1252244	Y	1252248	Het 18149

```
      ref|NC_001136|   1252248    Het   1252273    Y     1443536    Y     1443538    Het   191263
      ref|NC_001136|   1443643    Het   1443800    Y     1523096    Y     1523114    Het   79296
      ref|NC_001142|   523980     Het   524213     D     532027     D     532743     Het   7814
      ref|NC_001144|   13884 Het  13986 Y     16757 Y     16960 Het  2771
      ref|NC_001144|   804978     Het   805152     Y     806179     Y     806728     Het   1027
8     ref|NC_001136|   212036     Het   212179     D     213154     D     213530     Het   975
      ref|NC_001136|   213539     Het   214436     D     216493     D     217351     Het   2057
      ref|NC_001142|   709084     Het   709147     D     712026     D     712319     Het   2879
      ref|NC_001146|   221673     Het   222696     D     233498     D     233507     Het   10802
      ref|NC_001147|   131858     Het   131978     D     132419     D     132501     Het   441
29    ref|NC_001133|   28352 Het  28402 Y     42684 Y     44675 D    14282
      ref|NC_001133|   42684 Y    44675 D     47908 D     49458 Het  3221
      ref|NC_001136|   125130     Het   125401     D     126343     D     126531     Het   942
      ref|NC_001139|   22309 Het  22900 Y     26801 Y     27002 Het  3901
      ref|NC_001140|   101494     Het   102263     Y     129478     Y     129863     Het   27215
      ref|NC_001140|   262100     D     262417     Y     265246     Y     266203     Het   2829
      ref|NC_001143|   100670     Het   102797     Y     114199     Y     114774     Het   11402
      ref|NC_001146|   452401     Het   453168     Y     454378     Y     454997     Het   1210
30    ref|NC_001136|   2346  het  4096  ?     18141 Y     18195 D    14045
      ref|NC_001136|   18141 Y    18195 D     18345 ?     18351 Het  456
      ref|NC_001136|   18651 D    18805 Y     116332     Y     116482     Het   97677
      ref|NC_001136|   121676     Het   121939     D     137511     D     137704     Het   15572
      ref|NC_001138|   11160 Het  11305 Y     11941 ?     12049 Het  1148
      ref|NC_001139|   446399     Het   446571     Y     450372     Y     450908     Het   3801
      ref|NC_001139|   724581     Het   724903     D     728685     D     729257     Het   3782
      ref|NC_001142|   570822     Het   571066     Y     744072     Y     744255     het   173006
      ref|NC_001146|   9920  Het  9930  ?     325628     Y     326003     D     316073
      ref|NC_001146|   325628     Y     326003     D     331884     D     332278     Het   5881
32    ref|NC_001134|   293504     Het   294230     D     294623     D     295631     Het   393
      ref|NC_001135|   4529  Het  4584  D     4794  D     5235  Het  210
      ref|NC_001139|   761877     Het   761973     D     763747     D     764155     Het   1774
      ref|NC_001142|   484372     Het   484566     D     485226     D     485447     Het   660
      ref|NC_001142|   538038     Het   538483     Y     541613     Y     541712     Het   3130
      ref|NC_001144|   627007     Het   627274     D     627543     D     627774     Het   269
35    ref|NC_001134|   801782     Het   801934     D     802198     D     802200     Het   264
      ref|NC_001134|   802200     Het   802237     D     802428     D     802464     Het   191
      ref|NC_001134|   802464     Het   802500     D     802527     D     802535     Het   27
```

```
      ref|NC_001134|   802607     Het   802634     D     803138     D     803212     Het   504
      ref|NC_001134|   803233     Het   803261     D     803398     D     803493     Het   137
      ref|NC_001134|   803493     Het   803495     D     803609     ?     803737     Het   114
      ref|NC_001134|   803737     Het   803993     D     804311     D     804420     Het   318
      ref|NC_001134|   804420     Het   804467     D     804646     D     804707     Het   179
      ref|NC_001134|   804646     D     804775     D     806262     ?     807206     Het   1487
      ref|NC_001136|   472593     Het   472644     Y     473555     Y     473912     Het   911
      ref|NC_001137|   10058 Het   10202 D     11032 D     11186 Het   984
      ref|NC_001139|   954   Het   978   Y     6182  ?     6368  Het   305155
      ref|NC_001139|   8613  Het   8879  Y     10518 ?     10527 Het   1639
      ref|NC_001139|   11102 Het   11149 Y     306133      Y     306650      Het   294984
      ref|NC_001139|   648206     Het   648788     D     707103     D     707633     Het   432412
      ref|NC_001139|   712317     Het   712615     D     763747     D     764155     Het   51132
      ref|NC_001139|   764998     Het   765005     D     1081200    ?     1081207    Het   316183
      ref|NC_001140|   201398     Het   201645     D     209176     D     210269     Het   7531
36    ref|NC_001136|   110696     Het   111192     D     126343     D     126524     Het   15151
      ref|NC_001136|   126538     Het   127233     D     146675     D     146729     Het   19442
      ref|NC_001137|   518862     Het   518973     Y     527960     Y     529325     Het   8987
      ref|NC_001138|   13696 Het   16759 D     57870 D     57967 Het   41111
      ref|NC_001138|   58170 Het   58233 D     62177 D     63471 Het   3944
      ref|NC_001138|   63471 Het   64651 Y     80524 Y     81673 Het   15873
      ref|NC_001139|   482781     Het   482923     D     485313     D     485316     Het   2390
      ref|NC_001141|   439600     Het   20606 D     X     X     End   End   268003
      ref|NC_001142|   X     X     20606 D     288609      D     289605      Het   268003
      ref|NC_001142|   289605     Het   289722     D     289974     D     290156     Het   252
      ref|NC_001142|   290156     Het   290709     D     293070     D     293102     Het   2393
      ref|NC_001142|   293133     Het   293436     D     374406     D     374407     Het   80970
      ref|NC_001142|   374413     Het   374592     D     396640     D     396723     Het   22048
      ref|NC_001142|   399390     Het   399489     Y     419397     Y     419759     Het   19908
      ref|NC_001146|   14268 Het   15990 Y     26378 Y     26408 Het   10388
      ref|NC_001146|   26953 Het   27007 Y     93892 Y     93966 Het   66885
      ref|NC_001146|   93966 Het   94390 Y     218946      Y     219260      Het   124556
      ref|NC_001146|   229899     Het   230095     D     231538     D     232094     Het   1443
37    ref|NC_001134|   76212 Het   76213 Y     83470 Y     83944 Het   7257
      ref|NC_001139|   846990     Het   847197     D     856232     D     856944     Het   9035
      ref|NC_001139|   997840     Het   997972     D     998284     D     999229     Het   312
      ref|NC_001140|   162628     Het   162837     D     172608     D     172701     Het   9771
```

```
      ref|NC_001141|   77546 Het   77585 Y     82457 Y     82538 Het   4872
      ref|NC_001141|   320609      Het   320712      Y     322620      Y     323727      Het   1908
      ref|NC_001144|   622999      Het   623059      Y     626688      Y     627007      Het   3629
      ref|NC_001145|   556209      Het   556437      D     560317      D     560337      Het   3880
      ref|NC_001146|   14268 Het   15990 Y     26378 Y     26408 Het   10388
      ref|NC_001146|   26953 Het   27007 Y     129262      Y     129427      Het   102255
      ref|NC_001146|   129427      Het   129503      Y     282026      Y     282032      Het   152523
      ref|NC_001146|   282704      Het   282868      Y     442868      Y     442912      Het   160000
40    ref|NC_001134|   317687      Het   318192      Y     322797      Y     322869      Het   4605
      ref|NC_001134|   622981      Het   623027      D     624215      D     624219      Het   1188
      ref|NC_001134|   757513      Het   758401      D     758975      D     759048      Het   574
      ref|NC_001140|   233885      Het   234310      D     234687      D     235180      Het   377
      ref|NC_001144|   460694      Het   490579      D     789230      ?     789247      Het   298651
      ref|NC_001144|   789269      Het   789615      D     849645      D     849651      Het   60030
      ref|NC_001144|   1053367     Het   1053416     D     1059505     ?     1060287     Het   6089
      ref|NC_001144|   849654      Het   849924      D     873372      D     873649      Het   23448
      ref|NC_001144|   873725      Het   873872      D     1053107     D     1053284     Het   179235
      ref|NC_001146|   39282 Het   39652 Y     41680 Y     42185 Het   2028
      ref|NC_001146|   542052      Het   542236      Y     551869      Y     552035      Het   9633
      ref|NC_001147|   459840      Het   460430      D     472579      D     472668      Het   12149
41    ref|NC_001135|   296689      Het   296914      D     304117      D     304284      Het   7203
      ref|NC_001136|   232597      Het   232831      D     238389      D     239462      Het   5558
      ref|NC_001136|   1456284     Het   1456594     Y     1485742     Y     1485812     Het   29148
      ref|NC_001136|   1504290     Het   1504297     Y     1510882     Y     1511078     Het   6585
      ref|NC_001139|   855250      Het   855821      Y     859422      Y     859580      Het   3601
      ref|NC_001142|   575252      Het   575801      D     576323      D     576551      Het   522
      ref|NC_001142|   619267      Het   619858      Y     648002      Y     648251      Het   28144
      ref|NC_001144|   460694      Het   490579      D     507213      D     507450      Het   16634
      ref|NC_001144|   507486      Het   507626      D     849645      D     849651      Het   342019
      ref|NC_001144|   849654      Het   849924      D     1059505     D     1059564     Het   209581
      ref|NC_001144|   1061678     Het   1061688     Y     1061752     Y     1061759     Het   64
      ref|NC_001145|   856994      Het   857195      D     857568      D     858326      Het   373
48    ref|NC_001139|   8613  Het   8879  D     269180      D     269337      Y     260301
      ref|NC_001139|   269180      D     269337      Y     279435      Y     279936      Het   10098
      ref|NC_001139|   1053627     Het   1053897     D     1061803     D     1062031     Het   7906
      ref|NC_001140|   222295      Het   222462      D     391827      ?     391848      Het   169365
      ref|NC_001140|   392676      Het   393351      D     526329      D     526348      Het   132978
```

```
      ref|NC_001141|   418607     Het   418859     Y     431747     Y     431755     Het   12888
      ref|NC_001143|   78837 Het   79181 Y     82281 Y     82621 Het   3100
      ref|NC_001144|   460694     Het   490579     D     507213     D     507450     Het   16634
      ref|NC_001144|   507486     Het   507626     D     789230     D     789247     Het   551879
      ref|NC_001144|   789269     Het   789615     D     1059505    D     1059564    Het   269890
      ref|NC_001146|   57896 Het   57930 D     75987 D     76810 Het   18057
      ref|NC_001146|   79834 Het   80375 Y     95495 Y     95510 Het   15120
49    ref|NC_001134|   792576     Het   792961     D     800911     D     801342     Het   7950
      ref|NC_001135|   185372     Het   185655     D     187853     D     188202     Het   2198
      ref|NC_001136|   336943     Het   337570     D     357632     D     357802     Het   20062
      ref|NC_001136|   357802     Het   358050     D     368092     D     368207     Het   10042
      ref|NC_001138|   13696 Het   16759 D     54858 D     54904 Y     38099
      ref|NC_001138|   54858 D     54904 Y     62177 Y     63471 Het   7273
      ref|NC_001139|   148118     Het   148376     Y     150335     Y     150399     Het   1959
72    ref|NC_001134|   683268     Het   684214     D     706067     D     706728     Het   21853
      ref|NC_001136|   122724     Het   123193     D     131141     D     131453     Het   7948
      ref|NC_001137|   337672     Het   337696     Y     338516     Y     338773     Het   820
      ref|NC_001138|   255242     Het   256052     D     261802     D     262052     Het   5750
      ref|NC_001141|   254748     Het   254828     D     257945     D     258755     Het   3117
      ref|NC_001143|   501392     Het   501710     D     505031     D     505177     Het   3321
      ref|NC_001144|   274432     Het   274976     Y     278845     Y     279188     Het   3869
      ref|NC_001146|   262310     Het   262400     Y     268643     Y     269266     Het   6243
      ref|NC_001147|   460430     Het   461250     Y     461434     Y     461760     Het   184
85    ref|NC_001134|   7683  Het   8161  Y     20520 Y     21325 Het   12359
      ref|NC_001136|   125130     Het   125401     D     147635     D     148054     Het   22234
      ref|NC_001136|   1179958    Het   1180090    Y     1186811    Y     1187273    Het   6721
      ref|NC_001144|   460694     Het   490579     D     507213     D     507450     Het   16634
      ref|NC_001144|   507486     Het   507626     D     641709     D     642132     Y     134083
      ref|NC_001144|   641709     D     642132     Y     1060287    Y     1060403    Het   418155
      ref|NC_001146|   778589     Het   778804     D     779265     D     779375     Het   461
87    ref|NC_001134|   686624     Het   686767     Y     689309     Y     690123     Het   2542
      ref|NC_001136|   5608  Het   17231 D     17461 D     17538 Het   230
      ref|NC_001136|   17772 Het   17955 D     18141 D     18195 Het   186
      ref|NC_001136|   18651 Het   18805 D     135409     D     136069     Het   116604
      ref|NC_001136|   1441973    Het   1442387    Y     1447378    Y     1447428    Het   4991
      ref|NC_001144|   460191     Het   460251     Y     507213     Y     507450     Het   46962
      ref|NC_001144|   507483     Het   507486     Y     530404     Y     530423     Het   22918
```

```
      ref|NC_001144|   530423     Het   530531     Y     541310     Y     541474     Het   10779
      ref|NC_001144|   541474     Het   541508     Y     585709     Y     585942     Het   538891
      ref|NC_001144|   585942     Het   586052     Y     625055     Y     625064     Het   39003
      ref|NC_001144|   625064     Het   626279     Y     709102     Y     709123     Het   82823
      ref|NC_001144|   709138     Het   709175     Y     789096     Y     789104     Het   79921
      ref|NC_001144|   789230     Het   789251     Y     1059505    Y     1059564    Het   270254
95    ref|NC_001136|   78264 Het  78875 D     88907 D     89384 Het  10032
      ref|NC_001136|   99929 Het  100476     D     126538     D     127233     Het   26062
      ref|NC_001136|   128619     Het   129117     D     155089     D     155368     Het   25972
      ref|NC_001136|   1155625    Het   1155661    Y     1159797    Y     1161206    Het   1145025
      ref|NC_001139|   10736 Het  10804 Y     10982 Y     10985 Het  178
      ref|NC_001139|   336388     Het   336790     Y     337618     Y     338283     Het   828
      ref|NC_001140|   177049     Het   177113     D     220155     D     220635     Y     43042
      ref|NC_001140|   220155     D     220635     Y     378228     Y     378234     Het   157593
      ref|NC_001140|   378234     Het   378273     Y     391827     Y     391847     Het   13554
      ref|NC_001140|   391898     Het   391898     Y     525353     Y     526329     Het   133455
      ref|NC_001143|   608711     Het   608829     Y     639414     Y     639971     Het   30585
      ref|NC_001144|   388762     Het   389118     D     389851     D     390142     Het   733
      ref|NC_001144|   533245     Het   533365     D     533966     D     534385     Het   601
      ref|NC_001144|   1046001    Het   1046311    Y     1052500    Y     1052529    Het   6189
      ref|NC_001146|   333907     Het   334339     D     337166     D     337563     Het   2827
```

Chromosome	Chromosome Length	Events	Bases Converted	Average Tract Length
I	230218	3	110889	36963
II	813184	18	61942	3441.222222
III	316620	3	9611	3203.666667
IV	1531933	47	2530434	53839.02128
V	576874	3	10791	3597
VI	270161	6	112050	18675
VII	1090940	18	1363957	75775.38889
VIII	562643	11	566226	51475.09091
IX	439888	8	52239	6529.875
X	745751	14	611229	43659.21429
XI	666816	4	48408	12102
XII	1078177	39	3979710	102043.8462
XIII	924431	4	129211	32302.75
XIV	784333	19	1016773	53514.36842
XV	1091291	12	171493	14291.08333
XVI	948066	11	798551	72595.54545

Table S1: Summary of gene conversion events by chromosome.

Lines Analyzed	18
Conversion Events	220
Bases Converted	11573514
A>T;T>A	3944
C>G;G>C	2535
A>C;T>G	2800
A>G;T>C	18423
C>A;G>T	2761
C>T;G>A	18395
A>C/G;T>G/C	21223
C>A/T;G>T/A	21156
Bases Lost	314
Conversion Rate	5.0625E-10
Bases Converted Rate	2.66322E-05

Table 2: Gene conversion summary across all events.

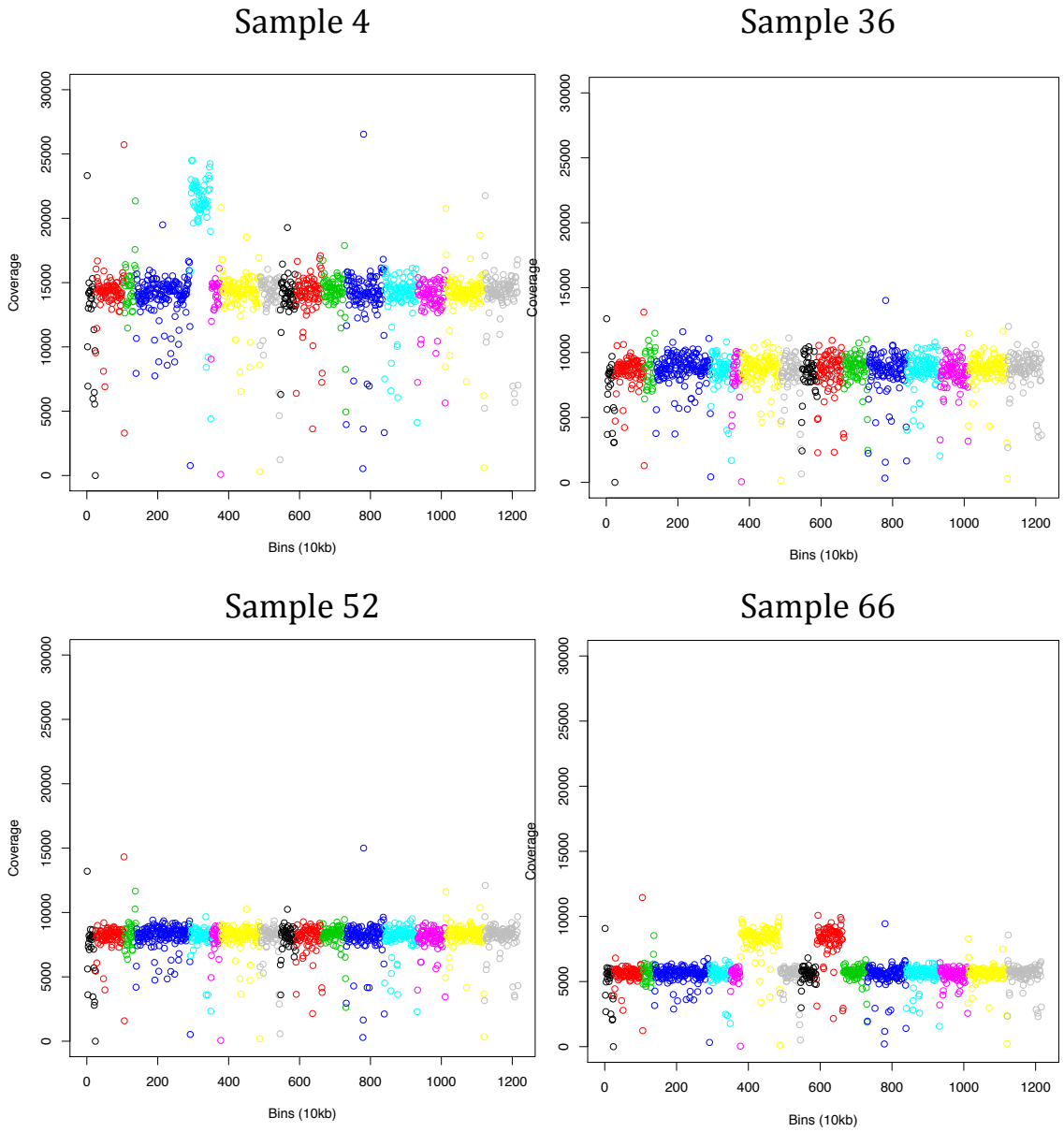

Figure S1: Coverage graphs for Samples 4, 36, 52,and 66. Chromosomes I-XVI are denoted by different colors. Trisomy for chromosome V can be observed in Sample 4 and trisomy for chromosomes VII and X can be observed in Sample 66.

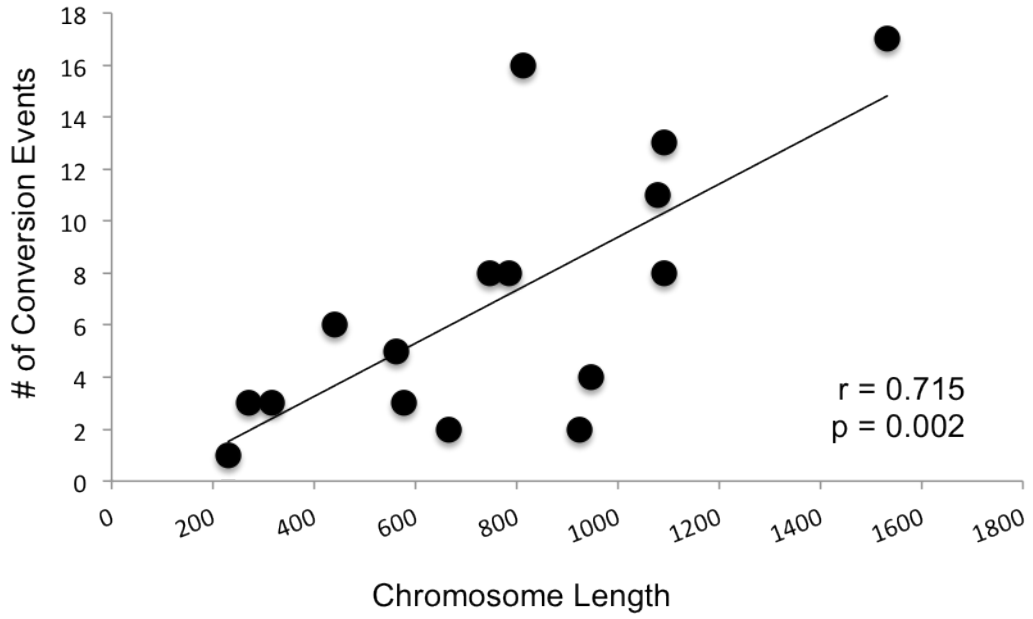

Figure S2: Number of chromosome events that are shorter than the median tract length is positively correlated with chromosome length(kb).

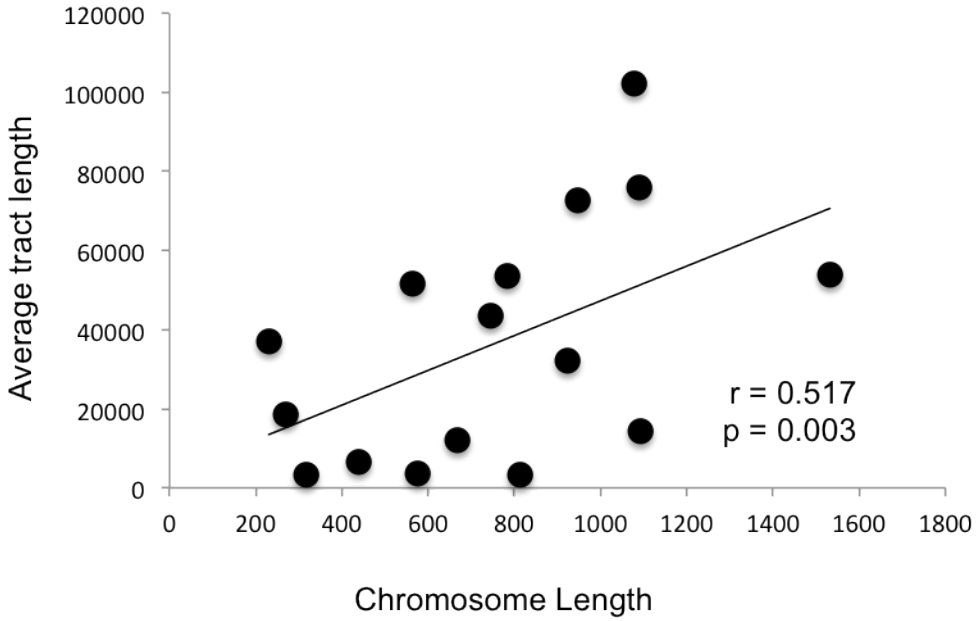

Figure S3: Average gene conversion tract length is positively correlated with chromosome length(kb).

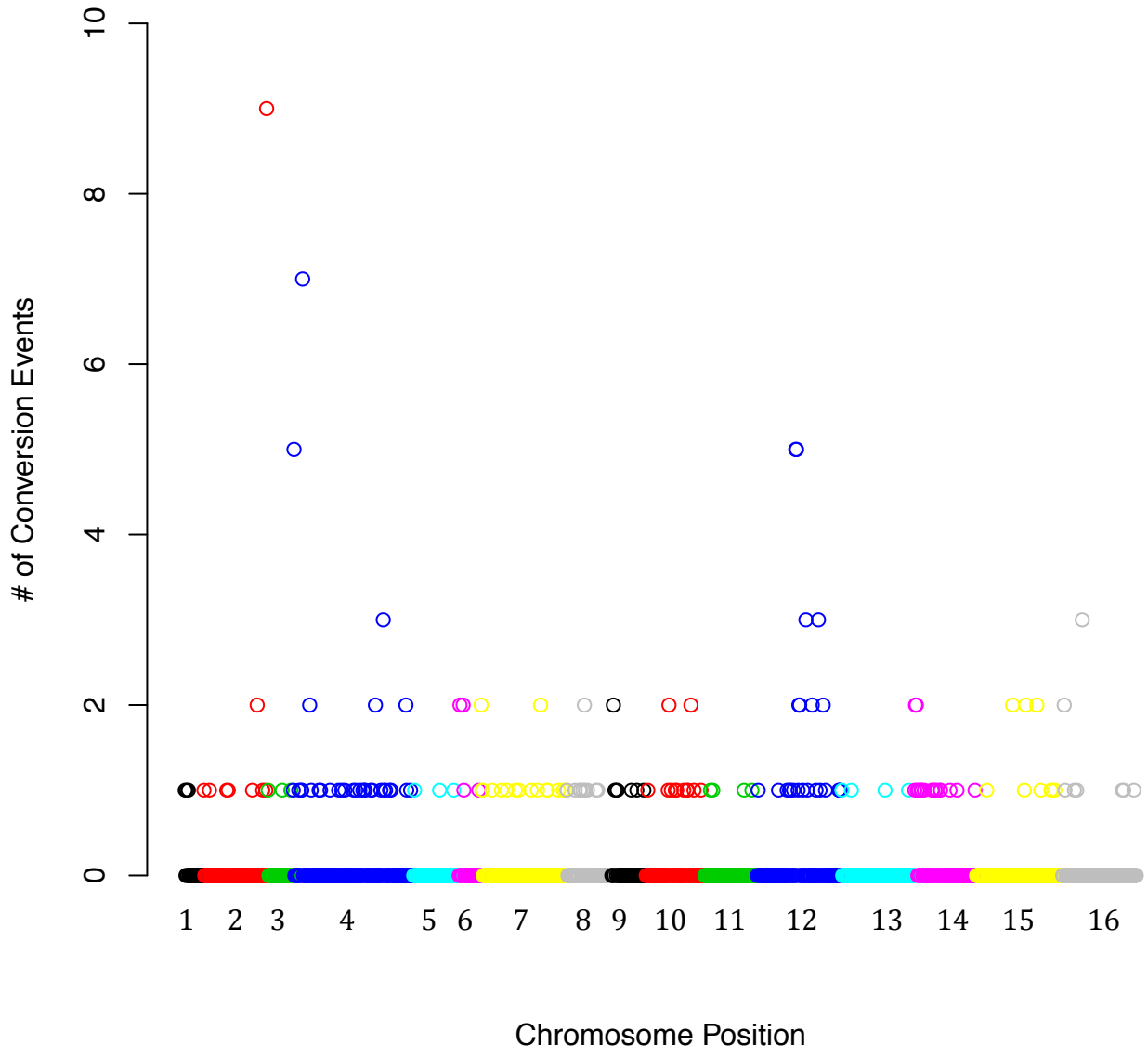

Figure S4: Gene conversion events mapped to 10kb windows across all chromosomes. Chromosome II, IV and XII are enriched for gene conversion events.

APPENDIX C:

SUPPLEMENTARY MATERIALS FOR SELECTION ON NONSENSE CODONS IN INTRONS

Order of Appendix:

I. Supplementary Tables

II. Supplementary Figures

First Intron distance from 5' splice site to first PTC (in nucleotides)

	A. thaliana			*C. elegans*			*H. sapiens*			*M. musculus*		
	Frame 0	Frame 1	Frame 2	Frame 0	Frame 1	Frame 2	Frame 0	Frame 1	Frame2	Frame 0	Frame 1	Frame 2
Average	38.27	51.04	52.56	30.55	46.83	48.94	104.45	91.85	91.75	90.21	73.88	72.59
Q1	11	6	9	3	1	6	13	1	1	14	1	1
Q2	31	32	34	15	21	24	63	39	41	54	35	32
Q3	54	63	66	36	53	57	147	132	135	126	108	106
	Frame 0 v. 1	Frame 0 v. 2	Frame 1 v. 2	Frame 0 v. 1	Frame 0 v. 2	Frame 1 v. 2	Frame 0 v. 1	Frame 0 v. 2	Frame 1 v. 2	Frame 0 v. 1	Frame 0 v. 2	Frame 1 v. 2
KS D value	0.091	0.088	0.057	0.113	0.142	0.0715	0.167	0.149	0.018	0.16	0.174	0.029
P value	< 2.2E-16	< 2.2E-16	8.90E-15	< 2.2E-16	< 2.2E-16	< 2.2E-16	< 2.2E-16	< 2.2E-16	0.256	< 2.2E-16	< 2.2E-16	0.002

	D. melanogaster			*S. pombe*			*S. cerevisiae*		
	Frame 0	Frame 1	Frame2	Frame 0	Frame 1	Frame2	Frame 0	Frame 1	Frame2
Average	29.74	61.65	61.66	18.50	53.63	44.68	29.53	57.79	64.63
Q1	5	1	1	3	5.75	6	8	20	18.5
Q2	17	19	22	11	21	24	20	40	43
Q3	36	55	59	24	44	49	48	68	73
	Frame 0 v. 1	Frame 0 v. 2	Frame 1 v. 2	Frame 0 v. 1	Frame 0 v. 2	Frame 1 v. 2	Frame 0 v. 1	Frame 0 v. 2	Frame 1 v. 2
KS D value	0.109	0.134	0.033	0.221	0.264	0.054	0.264	0.2699	0.0736
P value	< 2.2E-16	< 2.2E-16	0.008	3.08E-14	< 2.2E-16	0.305	2.37E-05	1.39E-05	0.769

Table S1: Statistics for distance from 5’ Splice site to first PTC/NC within first introns for all reading frames.

Last Intron distance from 5' splice site to first PTC (in nucleotides)

	A. thaliana			*C. elegans*			*H. sapiens*		
	Frame 0	Frame 1	Frame2	Frame 0	Frame 1	Frame2	Frame 0	Frame 1	Frame2
Average	38.27	51.04	52.56	30.55	46.83	48.94	69.36	56.89	59.04
Q1	11	6	9	3	1	6	9	1	1
Q2	31	32	34	15	21	24	42	24	29
Q3	54	63	66	36	53	57	96	78	83
	Frame 0 v. 1	Frame 0 v. 2	Frame 1 v. 2	Frame 0 v. 1	Frame 0 v. 2	Frame 1 v. 2	Frame 0 v. 1	Frame 0 v. 2	Frame 1 v. 2
KS D value	0.167	0.031	0.165	0.136	0.039	0.169	0.029	0.023	0.021
P value	< 2.2E-16	3.95E-05	< 2.2E-16	< 2.2E-16	< 2.2E-16	< 2.2E-16	0.013	0.077	0.156

	M. musculus			*D. melanogaster*		
	Frame 0	Frame 1	Frame2	Frame 0	Frame 1	Frame2
Average	62.77	49.93	56.06	29.74	61.65	61.66
Q1	9	1	1	5	1	1
Q2	39	24	31	17	19	22
Q3	86	71.75	78	36	55	59
	Frame 0 v. 1	Frame 0 v. 2	Frame 1 v. 2	Frame 0 v. 1	Frame 0 v. 2	Frame 1 v. 2
KS D value	0.1578	0.091	0.069	0.109	0.049	0.077
P value	< 2.2E-16	< 2.2E-16	< 2.2E-16	< 2.2E-16	8.19E-06	6.97E-14

Table S2: Statistics for distance from 5’ Splice site to first PTC/NC within terminal introns for all reading frames.

First Intron distance from 5' splice site to first PTC (in nucleotides)

	A. thaliana		*C. elegans*		*H. sapiens*		*M. musculus*		*D. melanogaster*		*S. pombe*		*S. cerevisiae*	
	Inframe	Outframe	Inframe	Outframe	Inframe	Outframe	Inframe	Outframe	Inframe	Outframe	Inframe	Outframe	Inframe	Outframe
Average	42.55	60.65	39.84	58.12	124.33	132.7	106.35	107.21	37.98	86.21	21.24	55.13	28.89	60.42
Q1	17	18	11	14	36	36	30	32	11	17	8	12	8	20.25
Q2	35	41	23	32	81	89	72	74	24	38	15	29	20	35
Q3	57	74	48	68	168	183	144	149	44	83	27	51	47	67.5
KS	0.105		0.145		0.034		0.015		0.194		0.272		0.264	
P value	< 2.2E-16		< 2.2E-16		0.006		0.437		< 2.2E-16		5.32E-15		2.37E-05	

Table S3: Statistics for distance from 5’ Splice site to first PTC/NC within first introns with genes/frames introducing PTC/NCs due to splice site consensus sequences removed.

Last Intron distance in nucleotides from 5' splice site to first PTC

	A. thaliana		*C. elegans*		*H. sapiens*		*M. musculus*		*D. melanogaster*	
	Inframe	Outframe	Inframe	Outframe	Inframe	Outframe	Inframe	Outframe	Inframe	Outframe
Average	59.12	59.26	52.93	55.29	82.16	83.38	75.26	75.44	62.86	78.57
Q1	18	17	14	14	24	23	24	24	14	14
Q2	39	36	27	29	56	54	53	53	29	32
Q3	66	65	62	68	110	111	99	99	54	68
KS D value	0.04		0.047		0.021		0.008		0.067	
P value	1.05E-06		3.52E-08		0.249		0.977		7.85E-08	

Table S4: Statistics for distance from 5’ Splice site to first PTC/NC within terminal introns with genes/frames introducing PTC/NCs due to splice site consensus sequences removed.

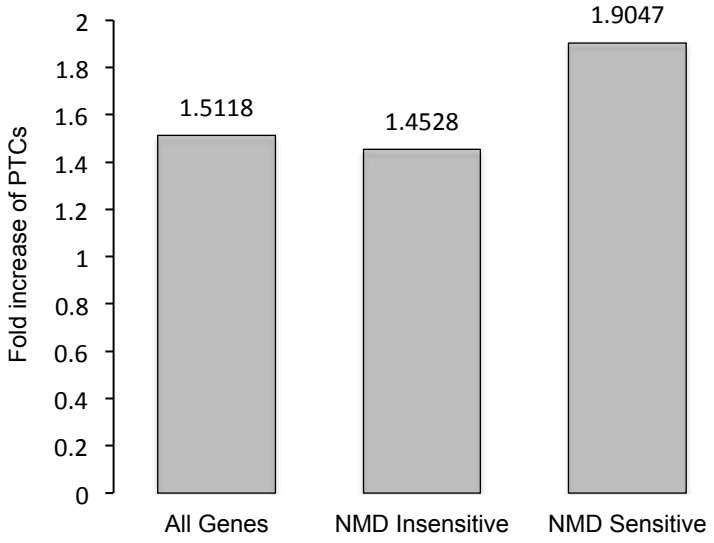

Figure S1: Enrichment of PTCs at the 5' end of introns is greatest for NMD sensitive genes in *S. cerevisiae*. Ratios of PTC frequency within the first 30 nucleotides of the 5' splice site in frame 0 vs. the average nonsense codon frequency within the first 30 nucleotides of the 5' splice site in frames 1 and 2 for all (n = 161), only NMD insensitive (n = 134), and only NMD sensitive (n = 27) genes in *S. cerevisiae*. Other species may show similar increased PTC enrichment in the first 30 nucleotides for NMD sensitive genes, but data are not available to distinguish sensitive and non-sensitive genes.

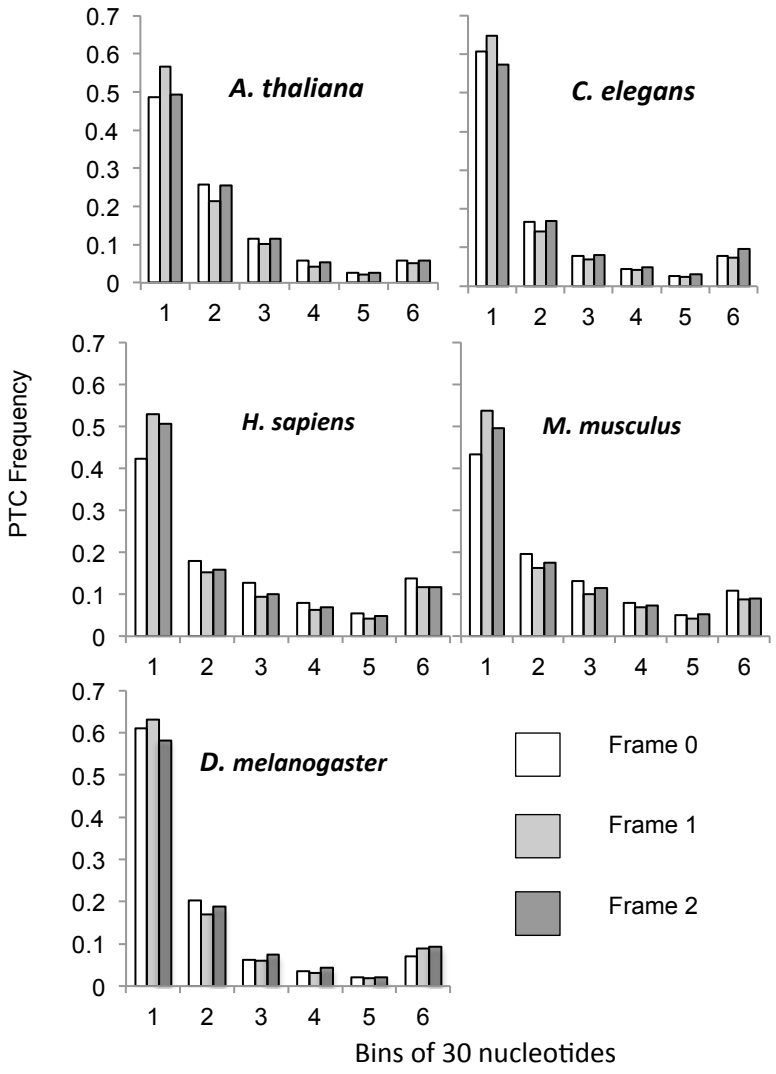

Figure S2: In-frame PTCs occur earlier than expected in the last intron for non-mammals. Observed (white/gray bars) distances in base pairs between the 5' splice site and intronic termination codons for in-frame PTCs, frame 1, and frame2 first out-of-frame NCs. Distances are separated into 6 bins, each of which is 30 nucleotides in length, except for bin 6 which contains all distances longer than 150 nucleotides. Panels A-F are in order of increasing effective population size (Table 1).

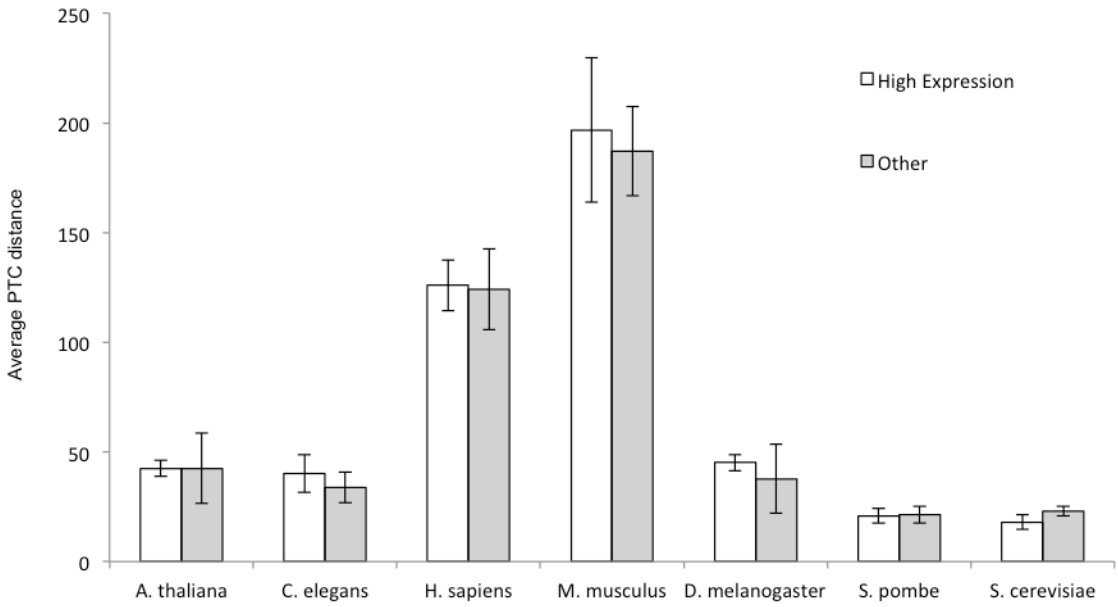

Figure S3: PTC position is generally unaffected by level of gene expression. Average PTC position for highly expressed genes and for all other genes. Highly expressed genes are the 100 genes showing highest levels of gene expression. Confidence intervals are calculated as plus/minus one standard error of the mean. There are no significant differences in any species.

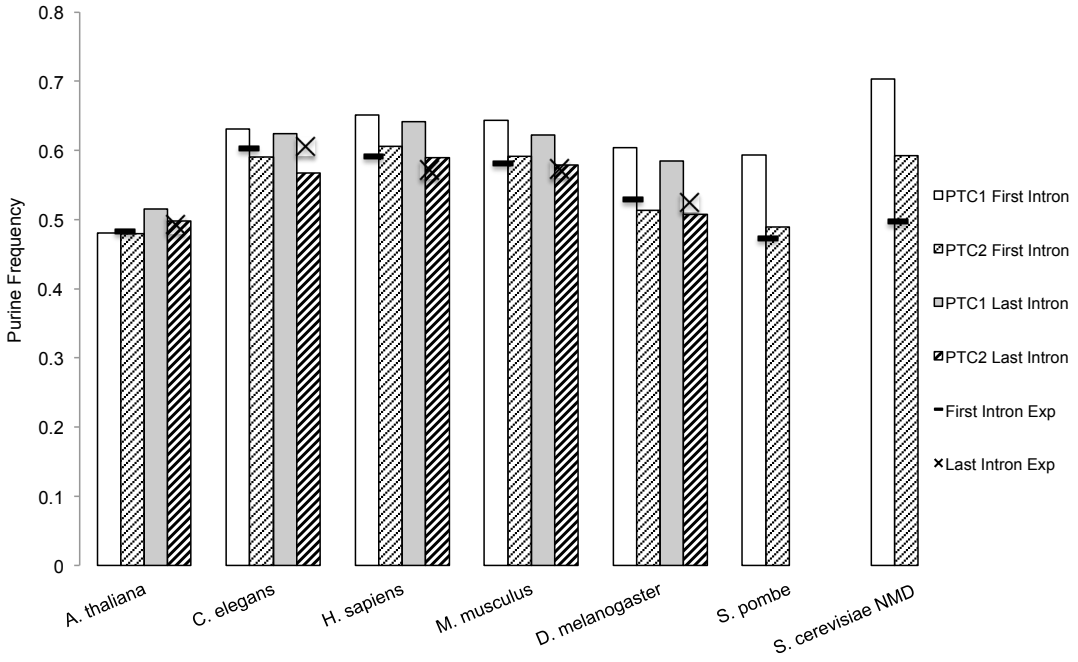

Figure S4: The first in-frame PTC is enriched for a purine nucleotide in first downstream position for both first introns and last introns. The frequency of a purine nucleotide in the first position following the first and second in-frame PTCs solid/striped bars). The expected frequency was calculated from the purine frequency of the first nucleotide after all termination codon-like trinucleotides (TAA, TAG, TGA) within introns, regardless of frame.

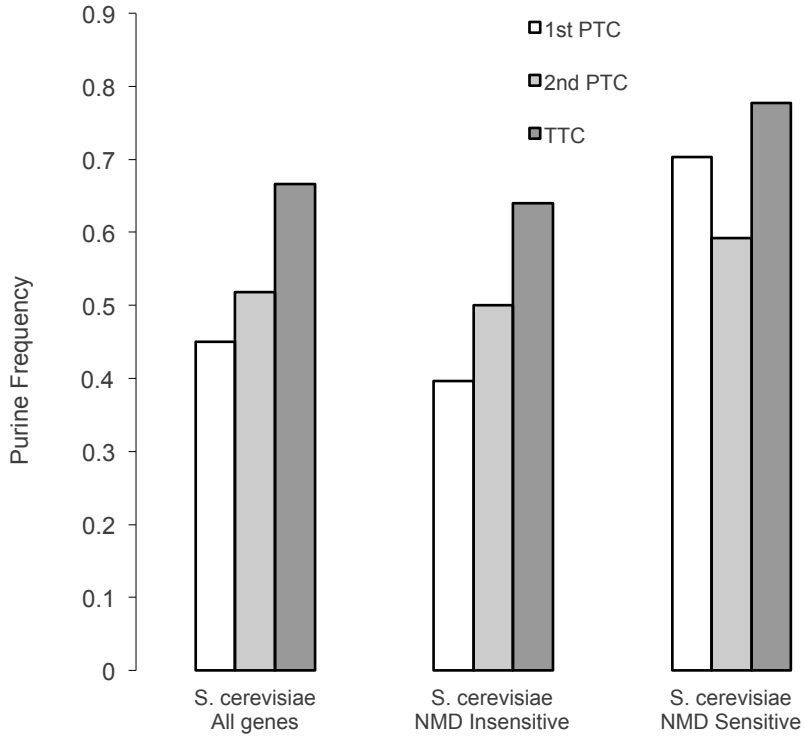

Figure S5: NMD sensitive genes show elevated purine frequency following first PTC. Purine frequencies of the first nucleotide following: the first in-frame PTCs, second in-frame PTCs, and TTCs among all (n = 161), only NMD insensitive (n = 134) and only NMD sensitive (n = 27) genes of *S. cerevisiae*.

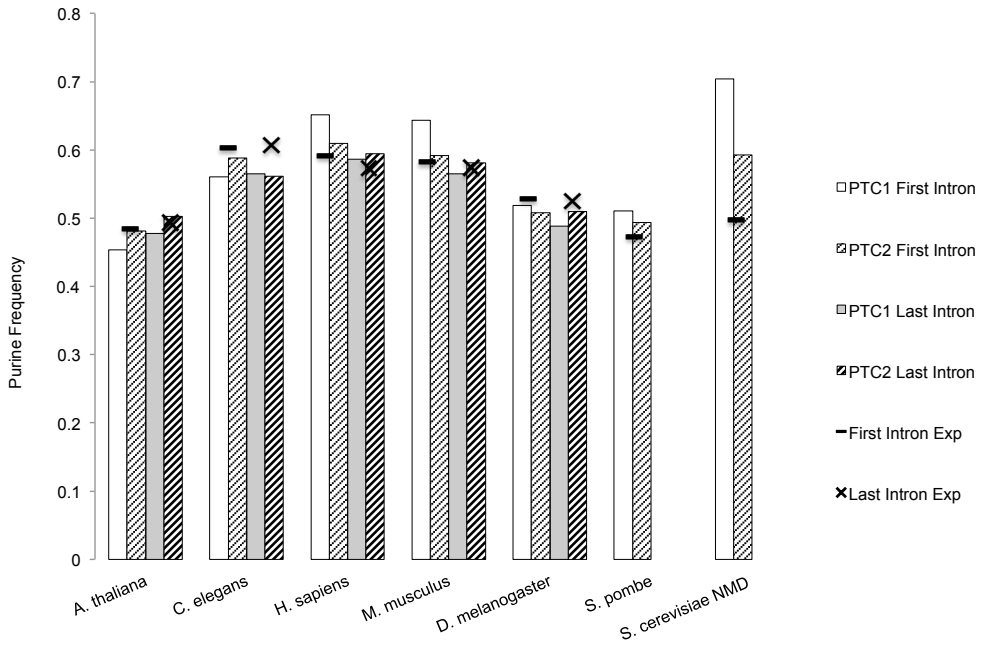

Figure S6: The nucleotide following the first in-frame PTC of the first intron is enriched for a purine in mammals compared to the 2nd PTC and the expectation. The nucleotide following the first in-frame PTC of the last intron however, is not enriched for a purine in any organism. The expected frequency was calculated from the purine frequency of the first nucleotide after all termination codon-like trinucleotides (TAA, TAG, TGA) within the first and last introns, regardless of frame.

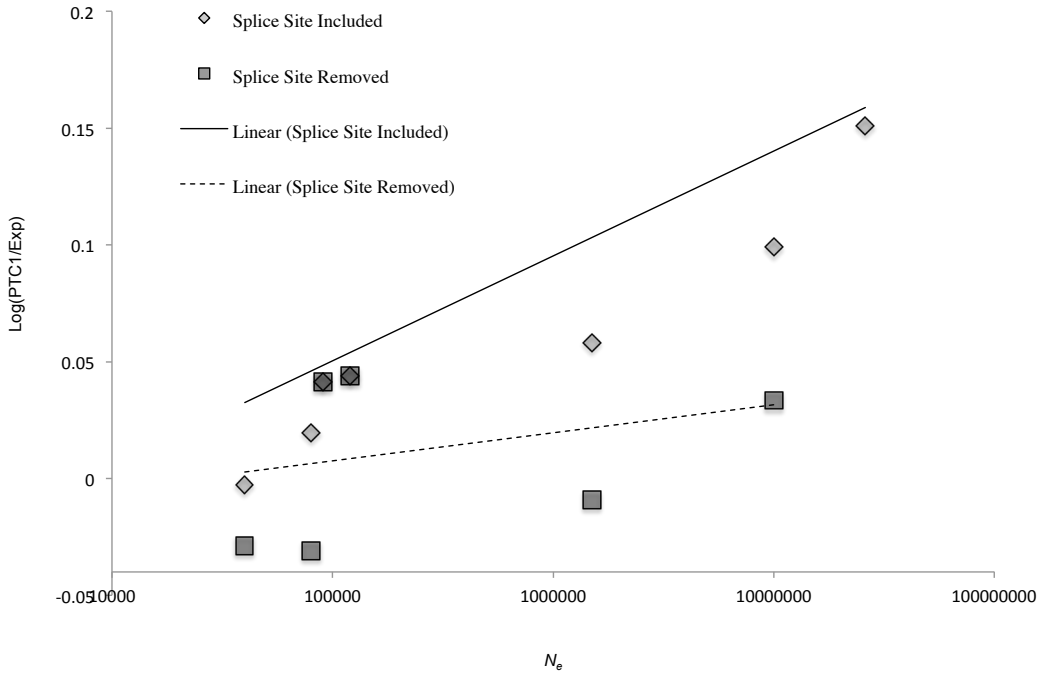

Figure S7: Purine frequency of in frame PTCs correlates with effective population size when PTCs due to splice site consensus sequences are not removed. Data points represent log ratios of in-frame PTC purine frequency and expected purine frequency. (With splice site: df = 6, Pearson's r = 0.924, p = 0.0029; excluding *S. cerevisiae*: df = 5, Pearson's r = 0.837, p = 0.038; Without splice site: df = 5, Pearson's r = 0.326, p = 0.528). Organisms in order of increasing *Ne* (i.e. points going left to right): *A. thaliana*, *C. elegans*, *H. sapiens*, *M. musculus*, *D. melanogaster*, *S. pombe,* Since *S. cerevisiae* does not introduce PTCs due to its splice site, it was not included in the splice site removed.