# STATISTICAL METHODS FOR GENOME VARIANT CALLING AND POPULATION GENETIC INFERENCE FROM NEXT-GENERATION SEQUENCING DATA

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Xin Ma

January 2012

# STATISTICAL METHODS FOR GENOME VARIANT CALLING AND POPULATION GENETIC INFERENCE FROM NEXT-GENERATION SEQUENCING DATA

Xin Ma, Ph.D.

Cornell University 2012

Next Generation Sequencing (NGS) technology has been widely adopted as a platform for DNA sequence variation detection and hence, accurate and rapid detection of genome variations using NGS data is critical for population genetics analyses. In my dissertation, I present three models that I developed to detect genome variation with high accuracy.

In Chapter 2, I analyzed sequence data in orang-utan. The orang-utan species, *Pongo pygmaeus* (Bornean) and *Pongo abelii* (Sumatran),are great apes found on the islands of Borneo and Sumatran. Populations on both islands are from the same ancestry but were subsequently isolated after the split. Due to recent deforestation to both islands, these species are critically endangered. Knowing their demographical history will not only help us better protect them, but it will provide us with a higher resolution evolutionary map for primates. It will also give us a powerful perspective on hominid biology because orang-utans are the most phytogenetically distant great apes from humans. In this study, we have sampled five wild-caught orang-utans from each of the two populations. One individual was sequenced to 20X coverage; the rest have median coverages between 6-8X. I developed a Bayesian population genomic variation detection tool which not only captures the population structure between these two populations but also pools all the allele frequency information among all in-

dividuals within the same population to boost the power of the variation detection in low coverage individuals.Our analysis revealed that, compared to other primates, the orang-utan genome has many unique features. From the population perspective, both Pongo species are deeply diverse; however, Sumatran individuals possess greater diversity than their Bornean counterparts, and more species-specific variation. Our estimate of Bornean/Sumatran speciation time, 400k years ago (ya), is more recent than most previous studies and underscores the complexity of the orang-utan speciation process. Despite a smaller modern census population size, the Sumatran effective population size (Ne) expanded exponentially relative to the ancestral Ne after the split, while Bornean Ne declined over the same period with more deleterious mutation accumulation. Despite some evidence for stronger negative selection in Sumatran orang-utans, detecting patterns of selection by fitting different selection models upon the baseline demographical model with nonsynonmous SNPs using ∂a∂i showed that the distribution of selection forces is actually similar to that in human with roughly 80% of mutations having a selection coefficient more negative than $s \approx 3 \times 10^{-5}$.

In Chapter 3, I undertook a second project aimed at understanding the molecular mechanisms that lead to mutation variation in yeast. This work is likely to provide insights not only in molecular evolution but also in understanding human disease progression. To analyze with limited bias genomic features associated with DNA polymerase errors, we performed a genome-wide analysis of mutations that accumulate in mismatch repair (MMR) deficient diploid lines of *Saccharomyces cerevisiae*. These lines were derived from a common ancestor and were grown for 160 generations, with bottlenecks reducing the population to one cell every twenty generations. We sequenced one wild-

type and three mutator lines at coverages from eight and twenty-fold using Illumina Solexa 36-bp single reads. Using an experimentally aware Bayesian genotype caller developed to pool experimental data across sequencing runs for all strains, we detected 28 heterozygous single-nucleotide polymorphisms (SNPs) and 48 single nucleotide (nt) insertion/deletions (indels) from the data set. This method was evaluated on simulated data sets and found to have a very low false positive rate ($\sim 6 \times 10^{-5}$) and a false negative rate of 0.08 within the unique (i.e., non-repetitive) mapping regions of the genome that contained at least seven-fold coverage. The heterozygous mutations identified by the Bayesian genotype caller were confirmed by Sanger sequencing. Our findings is interesting because frameshift mutations in homopolymer (HP) tracts, which are present at high levels in the yeast genome (> 77,400 for five to twenty nt HP tracts), are likely to disrupt gene function and further demonstrate that the mutation pattern seen previously in mismatch repair defective strains using a limited number of reporters holds true for the entire genome.

In Chapter 4, I presented an analysis of mutation hotspots in yeast deficient in DNA mismatch repair (MMR). Classical evolutionary theory assumes that mutations occur randomly in the genome; however studies performed in a variety of organisms indicate existence of context-dependent mutational biases. All of these biases involve local sequence context (e.g., increased rate of cytosine deamination at methylated CpG's in mammals), but the source of mutagenesis variation across larger genomic contexts (e.g., tens or hundreds of bases) have not been identified. Therefore, we use high-coverage whole genome sequencing (>200X coverage) of progenitor and derived conditional MMR mutant line of diploid yeast to confidently identify 92 mutations that accumulated after 160 generations of vegetative growth by using log-likelihood ratio test. We found

that the 73 single and double bp insert/deletion mutations accumulate much more frequently in homopolymeric poly-A and poly-T tracts with all mutations occurring at sites with at least 5 hp runs. Surprisingly, we demonstrated that the the likelihood of an indel mutation in a given poly (dA:dT) homopolymeric tract is increased by the presence of nearby poly (dA:dT) tracts in up to a 1000 bp region centered on the given tract. Furthermore, we identified nine positions that were mutated independently in at least two replicate lines and these all occurred at sites with at least 8 homopolymeric runs, suggesting greater instability for higher poly $A_n$ or poly $T_n$ sites. Our work suggests that specific mutation hotspots can contribute disproportionately to the genetic variation that is introduced into populations, and provides the first long-range genomic sequence context that contributes to mutagenesis.

## BIOGRAPHICAL SKETCH

Xin Ma was born in Chengdu, Sichuan of China and grew up in a family full of love. Under the influence from all her family members for of whom are professors in different fields, she was not only addict to reading a wide range of books but also showed early interest in numbers and formulas. In 1989, she was one of the five students chosen to attend the best middle school in the province from her county. During her high school study, she was particularly good at articulating complex math and physics problems. In addition, she also participated in the Chemistry Olympiad and ranked first (2000). In 2001, she came to University of Maryland at College Park (UMCP) at Maryland in United States on her own for undergraduate study in Computer Science with a provost scholarship from the university. She then attended State University of New York at Stony Brook (SUNY) in New York for additional training in Math, Applied Math and Chemistry. She was on the dean's list for every year in school and she further did a Honor Thesis with Professor Scott Sutherland, the chair of department of Mathematics of SUNY. She graduated with honor with three BS degrees in 2005 and she got a MS degree in Applied Math from SUNY in 2006. In the same year, she was admitted as a graduate student in department of Statistics of Cornell university in New York where she was very lucky to meet with her advisor, Carlos Bustamante.

Apart of her main focus in science, she also had a deep interest in piano. Her systematical and continuous practicing has well paid off and her playing skill is almost equivalent to professional. She held a few recitals and won numerous prize in different piano contests.

To my family

## ACKNOWLEDGEMENTS

First and foremost, I would like to use this precious opportunity to offer my sincerest gratitude to my advisor, Professor Carlos Bustamante in the Department of Genetics of Stanford University. He has been so helpful, generous and supportive since my joining of the lab. His broad knowledge of theoretical population genetics and ecological evolutionary theory has provided me with numerous headways on how to tackle specific problems that I encountered in projects while giving me room to express my own idea and work in my own way. His expertise in genetics has vastly improved my research skills and prepared me for future challenges. My graduate life would not have been so fruitful or enjoyable without him.

I also express my gratitude to Professor Eric Alani in the Molecular Biology department of Cornell University, who is always willing to invest his own time to share suggestions based on his own experience and understanding. He has always been so patient in providing detailed explanations on collaborative projects. His broad knowledge and sharp expertise in molecular biology has broadened my horizon in genetics. My projects would never have progressed so efficiently and productively without his great instinct and enthusiasm in molecular genetics.

My appreciation also goes to Professor Martin Wells from the department of Biological Statistics and Computational Statistics of Cornell University, for his friendship, encouragement, and numerous fruitful discussions. I would love to thank him for many useful comments on my work, which were crucial to articulate complex problems clearly. I am particular grateful for his advice, comments, and willingness to discuss any questions or ideas that I have had. I feel privileged to have him as one of my committee members and been able to work

with him. His serious and impartial working attitude, factual and realistic work style have greatly influenced me and set up an example on how to have success developing interdisciplinary career in future.

I also would like to express my appreciation to Devin Locke at University of Washington, for providing early access to the raw sequencing data for the orang-utan project and also the extraordinary organization of the consortium study with a stimulating environment since the very beginning.

A special thank-you goes out to my colleagues Sarah E. Zanders, Nishant K.T and Maria V. Rogacheva from Eric Alani's group at Cornell University. I am grateful for their fantastic experimental works which were particularly helpful and strongly helped out the progression of projects.

Another thank-you note goes out to Adam Siepel and Andre Martins at Cornell University, Ryan Gutenkunst,Jeremiah Degenhardt, Joanna Kelly, Kirsten Eilertson, Shaila Musharoff, Carolin Kosiol in Carlos Bustamante's group. I am indebted to them for the assistantship they offered and significant contributions they did on my projects.

I also would like to thank Simon Gravel who provided useful comments on my thesis.

Finally, I thank my family for supporting me throughout my studies in the United States for the past 10 years with tremendous love.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

CHAPTER 1

# INTRODUCTION

## 1.1 Genomic variation

Understanding the genetic variation of various species is a major challenge facing modern biology. The central theme of population genetics is to make sense of these variations, which includes the inference of past evolutionary histories as well as the dissection of molecular mechanisms leading to the variation and patterns of mutations we observe today. Understanding how genotypic variation translates into phenotypic variation and how it is shaped in populations and species, is fundamental to our understanding of evolution. Understanding the basis of genetic variation is of great importance to finding the disease susceptibility loci, facilitating better preservation of endangered species and inferring genes essential for environmental adaptions.

## 1.2 Next-generation sequencing technology

Early stages of population genetics studies in the last century were usually concentrated on only dozens to hundreds of variation markers due to technical limitations of obtaining accurate genomic variation information. Technical advances in molecular biology in the last two decades have made it possible to survey variation in natural populations on an enormous scale. The most dramatic examples include sequencing and assembling a single individual's genome [98] as well as hundreds of individuals from multiple populations as in the 1000

genomes project [25]. There have also been a large number of genome variation survey studies performed species other than humans in the last few years. Five years ago, the main technology for genomic variation discovery was automated Sanger sequencing [70], which was applied to finish the human genome [14]. However, even a shot-gun approach [98], the cost of obtaining DNA sequence for a single genome using ”first-generation” technology is beyond the reach of most research labs.

Next-generation sequencing (NGS) technologies have emerged as sequencing technologies to sequence entire genomes.These have involved HiSeq 2000 from Illumina, SOLiD from Applied Biosystems, and 454 from Roche. The NGS technologies have contributed tremendously to population genetics advances by serving as major platforms for detecting genomic variation on both the DNA and RNA levels. As a result, we have seen a sharp increase in throughput of DNA sequences and sharp decrease in price in price of sequencing during the last two years [79]. We have the capacity to obtain sequence information at unprecedented speed, thereby enabling previously unimaginable scientific achievements and novel biological applications. For example, it's now possible to perform de novo sequencing to reveal the complete DNA genomic sequence of endangered species [53], to quantify expression levels through RNA sequencing [73, 31, 95], and to couple with other technologies for the genome-wide profiling of epigenetic marks and chromatin structural features (ChIP-seq, methyl-seq and DNase-seq) [58]. Such technologies also allow deep large scale exon sequencing to identify variants responsible for both rare mendelian diseases and common diseases [75], as well as sequencing of cancer genomes to identify somatic mutations involved in tumorigenesis [49] .

## 1.3 Challenges in deciphering the NGS data

NGS technology provides affordable, efficient and reliable large-scale DNA/RNA sequencing. However, the process from obtaining raw reads from the sequencing machine to the final inferred variation calls involves a series of steps that introduces errors. First, the samples used for DNA extraction can be contaminated. In studies to detect cancer genome alternations, cancer samples are usually mixed with the peripheral blood samples,which can result in allelic biases in sequencing reads and lead to artefactual variant calling [71]. In NGS experiments, genomic sequences are first fragmented and amplified using PCR. This could lead to chimeric DNA sequences, new bases from the amplification, as well as biases of depths of different sequences, which are a major concern when quantification of different genes and alleles are the main task. The base reads from the sequencing machines are not error free and usually have higher error rate than traditional Sanger sequencing techniques. The typical error rate of NGS data ranges from a few tenths of a per cent to several per cent, depending on the platform. Once sequences are obtained, the first step is either de novo assembly or mapping to the reference genome. Various algorithms are developed to align short reads to reference genome, however, it's important to recognize that reads can be mapped to wrong genomic locations [51]. The last step for calling the variants after the reads are aligned to a reference genome involves dealing with the allelic imbalance due to low coverage or artificial amplification in high coverage data. Errors from each of the steps mentioned above will accumulate and bias the accuracy of the final detection of SNPs and hence, affect downstream population genetics analyses [76].

SNP calling in early NGS studies is usually based on the simple filtering of

quality scores and making cut-offs to call heterozygotes or non-reference alleles based on the empirically distribution of quality scores [38]. A commonly used cutoff is a Phred-type quality score of 20. Such simple filtering based methods will lead to false negatives for heterozygosity calling in low to intermediate coverage datasets. Another drawback is these methods provide no statistical confidence/probabilities for the genotype calls. It's reasonable to utilize the quality score and empirical sequencing error rate and use bayesian theory to obtain the posterior probabilities of genotypes of each individual [52, 54, 55]. These probabilistic methods provide measurement of statistical uncertainty of genotype calls and provide a general framework to incorporate any prior information that could be utilized to make genotype calls, such as allele frequencies and neighboring LD information.

The data in NGS studies varies dramatically due to different platforms, and in particular, different experimental designs. Some studies are designed to have individuals sequenced to low to intermediate coverages ($< 5X$ per site per individual, on average) in order to attain an optimistic sample size to enable population level inferences (i.e., orang-utan sequencing project described in Chapter2; 1000 genomes project [25]). Other NGS studies choose to sequence either part of the large genome or the entire small genomes at high coverage ($> 40X$ per site per individual, on average) to identify all possible variants in the specific individuals. However, such an approach can sacrifice population level information (i.e., yeast project described in Chapter4; [57, 49]). The different coverage and experimental designs lead to unique features for each study. My PhD research was focused on developing statistical method to convert base calls and quality scores into genotypes(SNPs or short indels). In the next two sections, I will explain the potential challenges identified when analyzing these datasets

of various coverages.

### 1.3.1 Feature of the data with low coverage

In the pilot one phase of the 1000 genomes project, 176 individuals were sequenced to $4X$ coverage per individual genome wide before filtering. Hence, given true segregating heterozygote loci and using a simple binomial model with the probability of observing either allele to be 0.5, the chance of not observing a single copy of either one of two segregating alleles is 0.0625 assuming there are no errors associated with any steps and no filter steps are needed. Therefore, there is a reasonable chance that we miss many true variation sites when coverage is low because probability is high that both chromosomes of a diploid individual are not sampled to enough number of copies in order to make the right statistical inference in a small number of trials. This is an even greater concern when the sequencing is not error-free. High false negative rates are problematic, especially for disease mapping studies where we are trying to find rare causal variants.

Previous bayesian SNP calling algorithms usually assume genotype prior probabilities at an individual level based on knowledge of error rate from known sequenced genomes [54, 55]. When the coverage is low, heterozygote calls are usually missed when controlling for false polymorphisms. Therefore, we need to come up with a better variation detection method which accounts for uncertainties obtained from low coverage. One way is to leverage the information across individuals in the population. When assigning prior probabilities of genotypes to individuals, we can use the allele frequency information

in the population and utilize the Hardy-Weinberg Equilibrium (HWE) to assist genotype probability calculations. Uncertainties in the allele frequencies and complex subpopulation structures ( as in the Orang-utan project in Chapter 2 ) can be modeled jointly. And the joint likelihood of all individuals would greatly enhance the power and reduce the false positives in genotype calling.

### 1.3.2 Feature of the data with extremely high coverage

In an exome resequencing project [57], the exomes in a single individual were sequenced to above $200X$ coverage. Hence, for each of the non-segregating loci, we might observe a fair amount of the null allele along with copies of the second allele that are attributed to the combination of sequencing error, mapping error and sample contaminations. This would lead to spurious signals of a variant site. Therefore, to deal with the problem, a bayesian type of statistics, might not be appropriate here because it can give us a high false positive rate by allowing a skew in terms of the ratio for the two alleles.

## 1.4 Outline of the dissertation

In this thesis I have developed advanced statistics calling algorithms that are aware of the unique features of DNA sequencing experiments and take into account of the potential uncertainties of a specific technology and the statistical limitations. Such algorithms are bound to play key roles for the successful applications of the NGS technology. With accurate and rapid genotyping and SNP calling tools developed for NGS studies, NGS technologies will have a striking

impact on genomic research and the entire biological field along with its ability to tackle the unsolved challenges unconquered by previous genomic technologies [108].

In the next three chapters, I present different tools that were developed to account for the different features of the dataset. To be more specific, I developed a population model aware bayesian SNP caller for orang-utan project where most individuals were sequenced at intermediate coverage so that we can leverage information from all individuals within the same population to boost the power for variation detection at individual level. In the two yeast projects, we are not sequencing the wild population, but carefully designed experimental strains to understand mutation mechanisms in the yeast genome. The mutated strains are all derived from a single clone of yeast, which leads to the propagation of alleles from the common ancestral strain and form a unique segregating SNP inheritance pattern. To account for this, I developed another bayesian SNP caller which utilized the feature from the experimental design for the first yeast sequencing project where each of the 5 strains were sequenced to intermediate coverage. Lastly, I used a likelihood ratio test for the second yeast sequencing project where two strains were sequenced to extreme high coverage so that the test is more stringent in terms of the ratio of the observed copies of each allele.

In all of these chapters, we can see that accurate detection of variation is critical in order to perform any of the downstream analyses. As technologies move forward, the next wave of new sequencing technologies involving single molecule sequencing (e.g., Helicos and Pacific Biosciences) will emerge to give researchers new powerful tools to attack biological questions. This will again raise new challenges in terms of developing statistical and computational meth-

ods to handle data. NGS as well as the 3rd generation techniques are providing formidable throughput increases, but also introduce tradeoff between sample size, accuracy of sequencing, and the fraction of the genome sequenced. The algorithms presented here show that data-feature-aware variant calling will serve as an important and necessary tool in modern statistical genomics.

CHAPTER 2

# COMPARATIVE AND DEMOGRAPHIC ANALYSIS OF ORANG-UTAN GENOMES WITH THE SELECTION STRENGTH INFERRING

The work described in this chapter comes from a mixture of published work including Nature Orang-utan Genome Consortium Paper,(Locke *et al.*, 2011 [60]), and a manuscript under preparation in which I am the primary author. I would like to thank Ryan Gutenkust, Joanna Kelley, Andre L. Martins, Jeremiah D. Degenhardt, Kirsten Eilertson, Shaila Musharoff, Adam Siepel, Tomas Vinar, Carolin Kosiol, and Carlos D. Bustamante, the co-authors in the work unpublished from this chapter, for their permission to use this information in my dissertation.

## 2.1 Introduction

"Orang-utan" is derived from the Malay term "man of the forest" and aptly describes the Southeast Asian great apes native to Sumatra and Borneo. The orang-utan species, Pongo abelii (Sumatran) and Pongo pygmaeus (Bornean), are the most phylogenetically distant great apes from humans featured with an ancient split in terms of their divergence time, thereby providing an informative perspective on hominid evolution. They are the only primarily arboreal great apes, characterized by strong sexual dimorphism and delayed development of mature male features, a long lifespan (35-45 years in the wild, over 55 years in captivity), and the longest interbred interval among mammals (8 years on average) [96]. Orang-utans create and adeptly use tools in the wild, and while long presumed socially solitary, dense populations of Sumatran orang-utans show complex social structure and geographic variability in tool use indicative of cul-

tural learning [97]. Therefore, knowing their demographical history and quantifying their evolutionary force will gain us insights on the evolution for the primate branch and give us an powerful perspective on hominid biology, which will enable us to look at the whole primates evolutionary selection process with higher resolution in future.

Furthermore, widespread deforestation on these Indonesian islands is rapidly changing and limiting the habitat and resources available to orang-utans [2]. These two endangered species are particular vulnerable now due to both continued deforestation of the two islands and the environmental threatens recently on different aspects of life includes climates, vegetation and the like [35]. A 2004 study estimated 7,000-7,500 Sumatran individuals and 40,000-50,000 Bornean individuals remained in the wild in fragmented subpopulations [86, 69] and The International Union for Conservation of Nature lists Sumatran orang-utans as critically endangered and Bornean orang-utans as endangered. Hence, a detailed study of the population structure, demographic history and inference of the distribution of selective effects will also provide insight into orang-utan population history and how additional environmental changes will affect the populations. This knowledge may guide the development of specific human-wildlife mitigation strategies to help prioritize regions and populations, which will contribute to conservation of these critically endangered species as well.

Ancestral orang-utan species once ranged broadly across Southeast Asia, including the mainland. Modern orang-utan species, in comparison, are geographically restricted and their island distribution reflects the impact of environmental factors such as variation in sea level over the Sunda Shelf and

more recently regional human population expansion. These factors resulted in a complex population genetic history, which includes one of the most recent speciation events among hominid, along with that of chimpanzees and bonobos. Historically, protein markers, restriction fragment length polymorphisms, and small sets of mitochondrial and nuclear markers have been used to estimate orang-utan divergence. We employed short read sequencing to address this question from a genome-wide perspective. Here we present a Sumatran orang-utan draft genome assembly and short read sequence data from five Sumatran and five Bornean orang-utan genomes using a whole-genome shotgun strategy. We accurately detected SNPs on both autosomal and mtDNA region and further constructed a phylogenetic tree combining samples from our study and other studies [99, 6] on mtDNA hypervariable region I (HRVI) to compare with the result from the analysis of autosomal data. We also classified detected the SNPs into functional categories and carried out a detailed site frequency spectrum analysis including damaging allele predictions based on PolyPhen2 program. On top of that, we used $\partial$a$\partial$i to infer the selection strength under the baseline demographical history aiming to quantify potential difference in terms of the evolutionary selection in Bornean and Suamtran populations.

Overall, our study shows the potential complex speciation, demographic history of Orang-utan, and shed light on the selection patterns in primate evolution. The resources and analysis presented here offer new opportunities in evolutionary genomics, insights into hominid biology, and an extensive database of variation for conservation efforts.

Table 2.1: Next generation sequencing data summary

| Sample ID | Studbook# | Name | Instit. When Sampled; Local ID | Origin | Sex | RawData(GB) | Coverage | Major Read Type |
|---|---|---|---|---|---|---|---|---|
| KB5404 | 590 | Billy | Lincoln Park Zoo, Chicago; 000362 | Borneo | F | 61 | 20.3 | 50 bp Paired Ends |
| KB5406 | 356 | Dinah | Dallas Zoo; 001036 | Borneo | F | 23 | 7.7 | 50 bp Paired Ends |
| KB5405 | 360 | Dennis | Dallas Zoo; na | Borneo | M | 26 | 8.7 | 75 bp Paired Ends |
| KB4204 | 364 | Dolly | Dallas Zoo; 001041 | Borneo | M | 25 | 8.3 | 50 bp Paired Ends |
| KB5543 | 990 | Louis | Los Angeles Zoo; 001929 | Borneo | M | 29 | 9.7 | 50 bp Paired Ends |
| SB550 | 53 | Doris | San Diego Zoo; 148001 | Sumatra | F | 21 | 7 | 50 bp Paired Ends |
| KB9528 | 732 | Baldy | Sacramento Zoo; 100083 | Sumatra | M | 28 | 9.3 | 50 bp Paired Ends |
| KB4361 | 1600 | Likoe | Miami Metro Zoo; M00176 | Sumatra | F | 21 | 7 | 50 bp Paired Ends |
| KB4661 | 695 | Bubbles | San Diego Zoo; 177257 | Sumatra | M | 20 | 6.7 | 50 bp Paired Ends |
| KB5883 | 550 | Sibu | Atlanta Zoo; 681456 | Sumatra | M | 25 | 8.3 | 50 bp Paired Ends |

## 2.2 Bayesian population genomic SNP caller

The orang-utan population diversity survey utilized DNA from 5 Sumatran and 5 Bornean wild-caught orang-utans, provided by Dr. Oliver Ryder and the San Diego Zoo's Institute for Conservation Research, San Diego, California [60]. DNA from each orang-utan was individually fragmented and ligated with adapters suitable for PCR amplification and sequencing on the Illumina GA/GAII platform in accordance with the manufacturer's protocols. For a thorough Bornean vs Sumatran comparison, one female Bornean individual (KB5404) was selected for deep (20x) coverage, and the remaining individuals were targeted for $\sim$ 8x coverage (Table 2.1). A mix of paired end and fragment reads (36 bp, 50 bp, and 75 bp read length) was used to reach coverage targets. Pair spans for the 9 $\sim$ 8x coverage individuals are approximately 180-280 bp. Pair spans of 180-280 bp, 280-380 bp and 380-480 bp were used for KB5404.

Illumina sequence reads from all 10 donor individuals were aligned against the Sumatran reference genome (v2.0.2) using Novoalign (www.novocraft.com). Only reads with less than 1 mismatch per 17 bp of "effective sequence" (i.e., ex-

cluding ambiguous base calls and 2 bp 5′ and 3′ of the read ends) were retained for SNP calling. Furthermore, we required that both reads from a mate pair align to this threshold and fall within the bounds of a log-normal estimated distribution for insert size in order to retain either read from the pair. As we see in Figure 2.1, for three libraries from individual KB5404, a Bornean donor individual sequenced to ~ 20X coverage, the log-normal distribution did an excellent job of modeling the dispersion around the modal insert size.

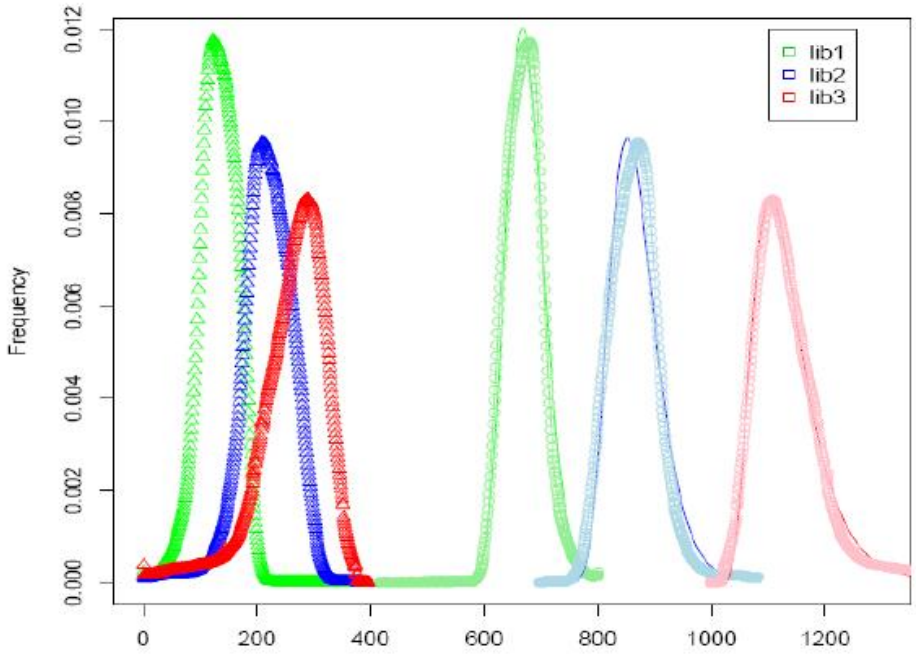



Figure 2.1: Distribution of pair span for three libraries sequenced for the high coverage (20x) Bornean individual (KB5404). Left is real distribution and the right is shifted+flipped distribution fitted by lognormal.

To account for potential biases in the assignment of Illumina quality scores, we first used Dohm *et al.*'s recalibration method [22]. Specifically, for reads of

length L (e.g., L = 36 or L = 72), at a given position $i$ ($1 \leq i \leq n$) for individual j ($1 \leq j \leq 10$), the empirical quality score function $e_{ijL}$ is:

$e_{ijL}$ = (# of non-reference alleles at position $i$ for reads from individual $j$ of length $n$) /(total # of reads that map to position $i$ for reads from individual $j$ of length $n$).

We then used the lower of the two scores (raw and recalibrated) in the SNP calling algorithm to reduce the influence of sequencing error on SNP calls, which potentially increased the false negative rate. We therefore assessed the impact of this scheme on the false negative rate in our simulation study (see next chapter for simulation).

SNPs for each population were called separately using a Bayesian population genomic approach that pools information regarding allele frequency among individuals within the same species when calling genotypes, i.e. it leverages the short-read sequence data for all individuals of given species aligned to the reference genome. The model utilizes a prior distribution on allele frequencies for variable sites as well as Hardy-Weinberg assumptions at the species level within species to derive a posterior distribution on genotype for each individual at each SNP based on the sequence data. We also applied a series of post- and pre-calling filters to reduce the possibility of errors including the following. First, we removed positions with an ambiguous base (N) in the reference genome, along with 5 bp 5′ and 3′ of that position, since the presence of even a single ambiguous base is an effective indicator of low-quality sequence [40]. SNPs were only called where the reference genome had a consensus quality score greater than 90 (on a scale of 1-97, based on the phred scores of underlying whole-genome shotgun reads). Regions of known segmental duplication

were also excluded. Furthermore, we required all potential SNP sites to have at least 7 individuals with greater than 2X coverage at that locus in order to be considered in the population genetic analysis. Finally, we did not allow for SNPs within 5 bp of each other, indels within 10 bp, or more than 8 individuals to be classified as heterozygous, in order to minimize the rate of false positives caused by recent segmental duplications.

The details of our calling algorithm are as follows. We wish to estimate the genotype for a given individual by jointly considering the reads for that individual and the estimate of the frequency for the allele in the population given the genotype calls. Under standard population genetic theory the allele frequency distribution for a single allele within a population follows a beta distribution [16, 27]. Therefore, our model assumes a vague prior distribution with a skew towards rare alleles for the minor allele frequency within each of the two populations by utilizing a beta distribution with parameters $\alpha = 0.01$,$\beta = 0.09$. Denote the 10 individuals as $i = 1, 2, \cdots 10$. For a particular site on the genome, let $A$ and $a$ be, respectively, the major and minor allele. Let $\hat{p}$ represent the minor allele frequency for a specific population at this site, where the prior is:

$$\mathbb{P}(\hat{p}) \sim Beta(\alpha = 0.01, \beta = 0.09) \tag{2.1}$$

let $N_i$ represent the total number of alleles observed for individual $i$ ; let $r_{ij}$ be the type of the $j^{th}$ allele copy among these $N_i$ allele copies where $j = 1, \cdots N_i$ ; let $e_{ij}$ be the corresponding error probability determine as above by either the reclibrated or raw quality score. For a particular site on the reference genome, we have that:

$$\mathbb{P}(Genotype|Data) \sim \mathbb{P}(Data|Genotype)\cdot \mathbb{P}(Genotype) \quad (2.2)$$

$$\sim \mathbb{P}(Data|Genotype)\cdot \int \mathbb{P}(Genotype|\hat{p})\cdot p(\hat{p})d\hat{p} \quad (2.3)$$

$$\sim \mathbb{P}(Data|Genotype)\cdot \int \mathbb{P}(Genotype|\hat{p})\cdot Beta(\alpha,\beta)d\hat{p} \quad (2.4)$$

Therefore, for individual $i$ , the posterior distribution on the three possible genotypes (AA, Aa, aa) are:

$$\mathbb{P}_i(AA|Data) \sim \mathbb{P}(Data|AA)\cdot \int (1-\hat{p})^2\cdot Beta(\alpha,\beta)d\hat{p} \quad (2.5)$$

$$\sim \frac{\prod_j^{N_i}(1-e_{ij})^{1_{(r_{ij}=A)}}\cdot {e_{ij}}^{1_{(r_{ij}=a)}}\cdot \beta(\beta+1)}{(\alpha+\beta)(\alpha+\beta+1)} \quad (2.6)$$

$$\mathbb{P}_i(Aa|Data) \sim \mathbb{P}(Data|Aa)\cdot \int 2\hat{p}\cdot(1-\hat{p})\cdot Beta(\alpha,\beta)d\hat{p} \quad (2.7)$$

$$\sim \frac{0.5^{N_i}\cdot 2\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)} \quad (2.8)$$

$$\mathbb{P}_i(aa|Data) \sim \mathbb{P}(Data|aa)\cdot \int \hat{p}^2\cdot Beta(\alpha,\beta)d\hat{p} \quad (2.9)$$

$$\sim \frac{\prod_j^{N_i}(1-e_{ij})^{1_{(r_{ij}=a)}}\cdot {e_{ij}}^{1_{(r_{ij}=A)}}\cdot \alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \quad (2.10)$$

Based on the posterior probability for each one of the three potential genotypes for individuals within the same population, the genotype frequencies can be estimated for the population and the parameter of the allele frequency distribution, alpha and beta, can be updated with respect to the population level including individuals which belongs to the same population and the iteration will

Table 2.2: Simulation result for each of the 10 individuals

| Individual | Coverage | Correctly called | Incorrectly called | Missed | Sensitivity | FDR |
|---|---|---|---|---|---|---|
| 1. 5504 | 20X | 33819 | 175 | 194 | 98.9% | 0.6% |
| 2. 5503 | 10X | 33640 | 589 | 904 | 95.8% | 2.6% |
| 3. 9258 | 9X | 23240 | 984 | 1204 | 91.4% | 4.7% |
| 4. 5406 | 7X | 21693 | 1036 | 1537 | 89.4% | 6.3% |
| 5. 5405 | 8X | 23161 | 1154 | 1342 | 90.3% | 5.2% |
| 6. 4202 | 8X | 33267 | 740 | 1318 | 94.2% | 3.7% |
| 7. 550 | 7X | 33024 | 847 | 1515 | 93.3% | 4.3% |
| 8. 5883 | 8X | 33074 | 748 | 1363 | 94.0% | 3.9% |
| 9. 4361 | 6X | 22532 | 1666 | 1806 | 86.6% | 6.9% |
| 10 .4661 | 6X | 22211 | 1721 | 1831 | 86.2% | 7.1% |

continue until the two consecutive updated parameters differ less than 0.001. In the above posterior probability formulae, $1_{(r_{ij}=A)}$ denotes the indicator function, which is 1 when $r_{ij} = A$ and 0 otherwise. Based on the posterior probabilities we classified the site as a variate site or nonvariate site exclusively.

## 2.3 Simulation

To estimate our SNP calling sensitivity and false discovery rate, we adapted a modified simulation protocol from the 1000 Genomes project. The dataset, which consists of 500 individuals, was simulated under a split model with the time of separation for Bornean and Sumatran populations of approximately 1M years ago. 10 individuals were randomly chosen and 36 basepair paired-end reads were generated by ART (http://biomedempire.org/) at the exact observed coverage in the real data for each individual, in order to be comparable to the real dataset. A single individual in the Sumatran group was randomly chosen as the reference genome and all the short reads were aligned against it. The simulation results are listed below in Table 2.2:

## 2.4 Validation

We validated a subset of SNPs by PCR-based re-sequencing on the 3730 platform. SNPs were selected from arbitrarily chosen regions of the orang-utan genome, sampling from chromosomes 1 and 3-11, with the sole requirement that predicted genotypes were available for all 10 sequenced orang-utan individuals at each site. Several categories of SNP were selected, including singletons, doubletons and higher frequency SNPs with three or more alleles observed among the 10 individuals we sequenced. Among doubletons and higher frequency SNPs, both heterozygous and homozygous sites were selected, as well as sites with a combination of heterozygous and homozygous alleles in the higher frequency category. Overall, the set was biased toward singleton SNPs (63 out of 108 sites, see below) to assess the ability of the SNP caller to successfully detect such sites with 8-10X coverage of short read sequence alignments.

From an initial set of 114 sites, 108 amplicons were successfully designed and sequenced. Manual genotype calls were then made at sites with sufficient Sanger data quality, which allowed 87.0% (940/1,080) of all possible genotypes to be called (Table 2.3). Overall genotyping accuracy, defined as the concordance between the predicted genotype and the Sanger data was very high, with 99.0% (931/940) of sites confirming computational predictions.

Genotyping accuracy for the singleton pool was high (98.9%), but this figure includes validated sites that were homozygous with respect to the reference genome for 9/10 individuals. Specifically among the individuals bearing a heterozygous singleton SNP we found 4 false positives out of 51 sites where validation data was available (a 7.8% false positive rate). Of the four false pos-

itive heterozygous singleton calls, one was a mis-called homozygous variant; the other three were homozygous wildtype alleles according to the Sanger data. We also found 2 false negative calls among the singleton pool (489 calls total) for a false negative rate of 0.4%. For the 11 doubleton SNPs in our validation set, a full 100% (110/110) were concordant between the Sanger data and the predicted genotypes. For higher frequency SNPs the overall concordance was 99.0% (291/294), and the validation rate of non-reference allele genotypes was 98.0% (149/152) with one false negative (0.7%). Overall, the high rate of concordance between the genotypes predicted by the SNP caller and the Sanger-based sequence data suggests a high level of accuracy among the large pool of SNPs detected across the orang-utan genome using this methodology. These results should provide confidence in the use of these SNPs in downstream analysis and applications.

## 2.5 Bioinformatics

### 2.5.1 Principal component analysis

To quantify patterns of population substructure using the SNPs discovered using our algorithm, we used a modified Principal Component Approach(Figure 2.2).

Specifically, we encoded each SNP for each individual as the posterior expected # of copies of the alternate (i.e., non-reference) allele for sites with coverage of at least 2X per individual:

Table 2.3: Validation result for each of the 10 individuals

| SNP Category | Sites Assessed | Resequencing Success Rate Across 10 Individuals | Overall Genotype Accuracy | Non-reference Allele Frequency | False Negative |
|---|---|---|---|---|---|
| Singletons (1 allele observed) | 63 | 540/630 (85.7%) | 534/540 (98.9%) | 47/51 (92.2%) | 2/489 (0.4%) |
| Doubletons (2 alleles) | 11 | 106/110 (96.4%) | 106/106 (100%) | 15/15 (100%) | 0/91 (0.0%) |
| High Freq. (3+ alleles) | 34 | 294/340 (86.5%) | 291/294 (99.0%) | 149/152 (98%) | 1/142 (0.7%) |
| Total | 108 | 940/1080 (87.0%) | 931/940 (99.0%) | 221/218 (96.8%) | 3/722 (0.4%) |

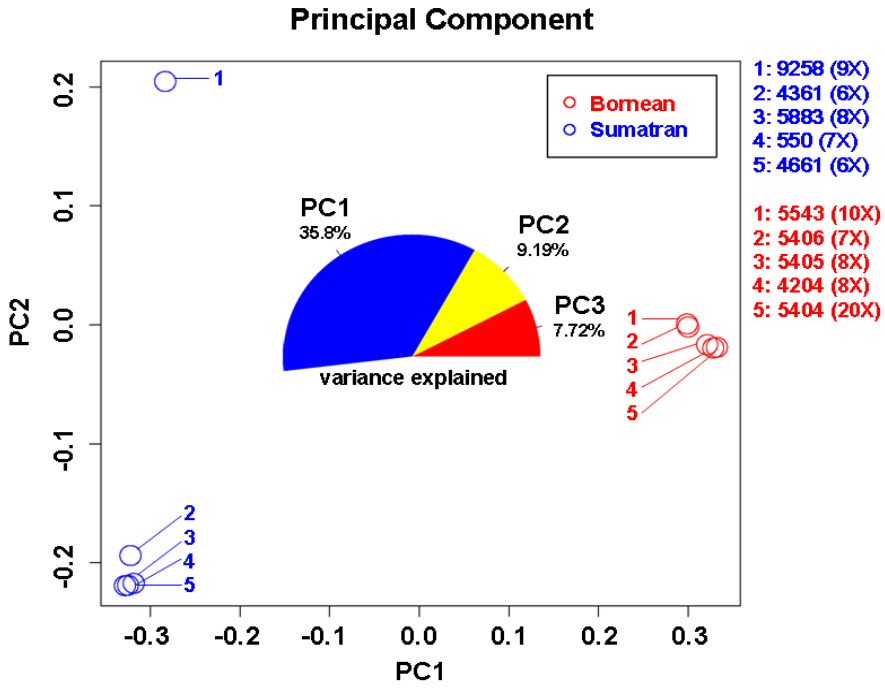



Figure 2.2: Principal component analysis of the genomic SNP data found by resequencing 10 Bornean and Sumatran donor individuals to at least 6X coverage.

E(# of copies of ″a″ — Data) = 2*Pr(aa — Data) + 1*Pr(Aa — Data)

This approach integrates out over uncertainty in assignment of the heterozygous vs. homozygous state. We also utilized PCA on the genotype matrix based on a maximum a posteriori approach that assigned each individual the genotype with highest posterior probability. The results are qualitatively very similar and show the extremely high quality of the data. Namely, the first principal component (PC1) separates the Bornean from Sumatran samples and explains approximately 35.8% of the variance. The second PC identifying one Sumatran individual (9258) as distinct from the other four shows more complex Suma-

tran population structure. This observation is consistent with the higher overall genetic diversity of Sumatran orang-utans.

### 2.5.2 Site frequency spectrum

In total, we identified 13.2 M putative SNPs across 1.96 Gb of the genome, equivalent to 1 SNP every 149 bp on average. Within the Bornean and Sumatran groups we detected 6.69 M (3.80 M Bornean-exclusive) and 8.96 M (5.19 M Sumatran-exclusive) SNPs, respectively (Figure 2.3a). Observing 36% more SNPs among Sumatran individuals strongly supports a larger Ne. Using Watterson's approach [101] we estimated nucleotide diversity from the SNP data as $\theta_W$= 1.21 and $\theta_W$ = 1.62 per kb for the Bornean and Sumatran species, separately, and $\theta_W$ = 1.89 per kb for the orang-utan species combined, roughly twice the diversity of modern humans [56].

The modal category of SNPs were singletons, with 2.0 M and 3.7 M SNPs observed as single heterozygous sites in a Bornean or Sumatran individual, consistent with the expectation that most genetic variation for an outcrossing population ought to be rare due to mutation drift equilibrium. We observed little correlation between Bornean and Sumatran SNPs in the AFS (i.e., the "heat" of the map is not along the diagonal as expected for populations with similar allele frequencies, but rather along the edges)(Fig 2.3b). This was further supported by Principal Component Analysis, in which PC1 corresponded to the Bornean/Sumatran population label and explained 36% of the variance.

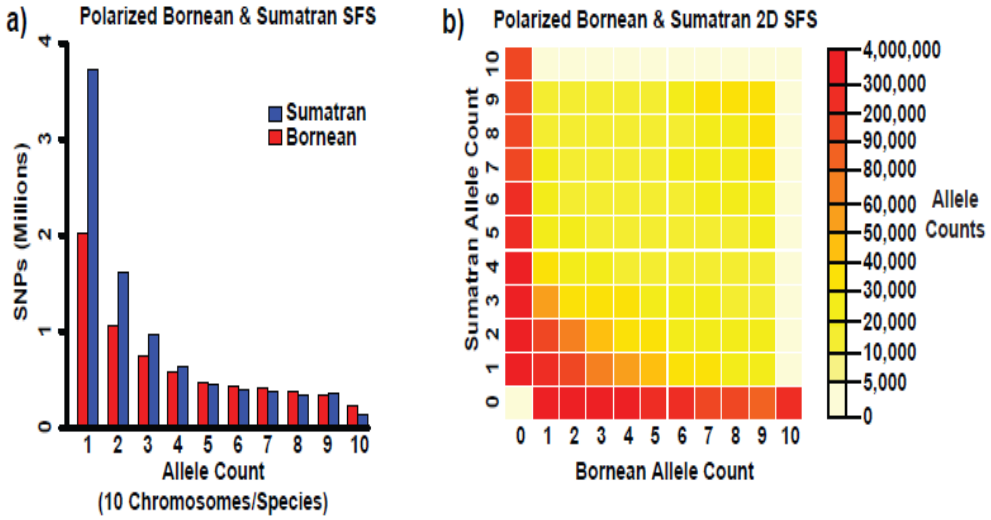



Figure 2.3: Orang-utan population genetics. a, Polarized site frequency spectra (SFS) for 13.2 million Bornean (blue) and Sumatran (red) SNPs using outgroup of human, chimpanzee and macaque are shown, note the enrichment of low-frequency SNPs among Sumatran individuals; b,The majority of SNPs were restricted to their respective island populations as the heat of the 2D SFS, representing high allele counts, lay along the axes.

## 2.6 Demographical history of orang-utan

### 2.6.1 Using autosomal variation markers

The data consists of SNPs detected in 1.96 Gb of genome sequenced in each of 5 Bornean and 5 Sumatran individuals. To account for occasional missing data, the frequency spectrum was projected down to 8 chromosome samples per population. The resulting cutoff of having 4 or more individuals called in

each population yielded 12.74 million usable SNPs in the frequency spectrum.

Two analysis were performed, differing in how SNPs were polarized. We first worked with the folded spectrum, which ignores ancestral state information and considers only minor allele frequencies. However, we found that this spectrum very poorly constrained the split time in models with migration, so we also worked with a polarized spectrum. For the polarized spectrum, we used the ancestral state inferred using the outgroup information from human, chimpanzee and macaque which assigns a probability for each possible ancestral state for each SNP, and the renormalized probabilities of the two segregating states were used for each SNP in the data set. One limitation of the current ancestral state algorithm is that if allele A is fixed in one population sample and the second sample has alleles A and G segregating, the ancestral state is always called as A. To compensate for this limitation, the data and model spectra were partially folded, ignoring ancestral state information for entries in which one allele is fixed in a population.

Based on these data, our best fitted demographic model consists of a two-population model with divergence and potential migration, growth, and difference in population size. Among several models tested we found very strong statistical support ($10^5$ log-likelihood units) for the most complex model, which included a split with growth and subsequent low-level migration. We estimated a relative Ne of 210% for Sumatran orang-utans relative to the ancestral and 49% for Bornean orang-utans, noting a four-fold difference for the derived populations (Figure2.4). Assuming a mutation rate of $2.0 \times 10^{-8}$ and 20 years per generation, we estimated an ancestral Ne of 17,900 and a split time of 400k ya.

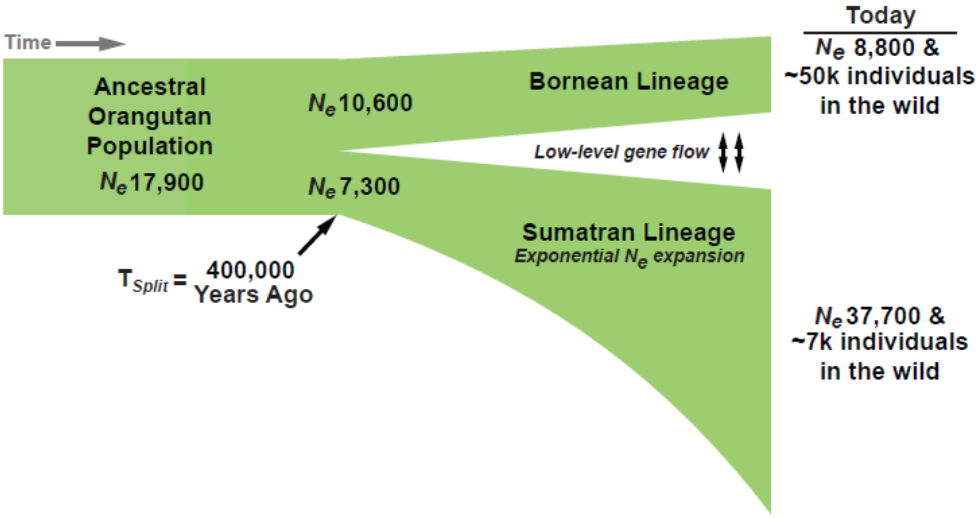



Figure 2.4: Orang-utan demographics. Our demographic model estimated the ancestral orang-utan population (Ne = 17,900) split approximately 400,000 years ago, followed by exponential expansion of Sumatran Ne and a decline of Bornean Ne, culminating in higher diversity among modern Sumatran orang-utans despite a lower census population size. The model also supported low-level gene flow (<1 individual/generation) indicated by arrows.

### 2.6.2 Using mtDNA variation markers

The inferred demographical history by looking at the genomic region of a total of 10 wild-caught individuals from both populations (i.e.,Bornean and Sumatran) with unknown provenance indicates that these are fairly distinct two populations. In particular, for these 10 individuals, the PCA plot utilizing the 12M

genomic variation locus agreed on the result suggesting that there is one outlier Sumatran individual that has a large amount of the Bornean population specific variations, which could be an indicator of either a history of complex speciation between Bornean and Sumatran population or a deep population structure in the Sumatran population. In order to further study the phylogenetic divisions of these 10 wild-caught orang-utan individuals closely and elucidate the evolutionary relationships among these apes, we looked at mtDNA region because it is inherited solely from the mother and it enables us to trace maternal lineage far back in time and detect any sex biased immigration which is not detectable in the autosomes due to the recombination.

We have first detected 1090 SNPs among a total of 13648 callable sites spanning about 83% of the whole mitochondrial utilizing the short reads from each of the individuals. Given the enormous coverage we have at each site, the estimated false positive (FP) rate and false negative (FN) rate are both negligible because we can easily distinguish the genuine heterozygous mutations from any residual background noises, i.e., sequencing errors, especially for mitochondrial genome where there isn't any recombination and only one haplotype needs to be determined. Therefore, we definitely expect to have discovered the majority of the mutations in the callable region with high accuracy. The estimated mutation rate is equivalent to one SNP every 12.5 bp, and is much more frequent than that on autosomal region (one SNP every 148.2bp). The higher mutation rate on mtDNA compared with autosomes is attributed to the fact that oxidative reduction primarily happened on mtDNA which is mutagenic and mutations on mtDNA tend not to get lost from the population because of the lack of recombination.

In order to construct the phylogenetic tree of all 10 orang-utan individuals along with the outgroup of human and chimp for the fossil calibration purpose and further obtain an estimate of the coalescent time of the common ancestor (TMRCA), we incorporated the collapsed hypervariable regionI(HVRI) haplotypes from 59/7 distinct Bornean/Sumatran haplotypes based on the previous studies [99, 6] along with the HVRI haplotpyes from our 10 individuals where they are callable. jModelTest [30, 80] was used to select the appropriate mutation model based on the Akaike information criterior(AIC). We used BEAST1.6.1 [24] to construct phylogenetic tree where Tracer1.5 (http://beast.bio.ed.ac.uk/Tracer) was further used to make sure the adequate burn-in period is achieved. In order to estimate TMRCA, the two calibration points were chosen to be the Pan-Homo which divergent at approximately 4.5-6 Ma, and Ponginae-Hominidae which divergent at approximately 12-16 Ma [60].The prior for mutation rate for was set to follow a Normal distribution with mean of 0.29 with a 95% interval of 0.22 to 0.36 mutations/bp/My for HVRI region FigTree1.3.1(http://tree.bio.ed.ac.uk/software/figtree/) was used to display and visualized summarized and annotated trees produced by BEAST.

Based on the phylogenetic tree (Figure 2.5A), our 10 wild-caught samples fall all over the tree, indicating an unbiased sampling. We were also able to identify the geographic origins for all of the wild-caught Bornean individuals. Using the two fossil calibration point of Pan-Homo which divergent at approximately 4.5-6 Ma, and Ponginae-Hominidae which divergent at approximately 12-16 Ma [60], we were able to estimate the TMRCA to be 3.67 Ma, which is quite different from the estimated split time of 400K years ago using the whole genomic variation data. We propose that the older estimated divergence time

here is likely due to the female philoatry and male dispersals since the split of the two populations [74]. Therefore, every migration event will reduce the coalescent time of the nuclear DNA but the mtDNA will be unaffected because it is maternally inherited and does not recombine with the other population. Hence, this observation is compatible with a moderate level of migration in the past.

Notably, not only is there a deep split in Sumatran populations, which corresponds well with the fact that Sumatran individuals possess greater diversity than their Bornean counterparts and more species-specific variation [60], but also the 3 Sumatran individuals with cyan highlighting clad color clustered together far from the rest of the Sumatran individuals with low branch support on the tree. The split time between this particular branch and the rest of the Sumatran cluster is estimated to be 3.12 Ma. This particular mitochondrial phylogeny that we estimate could either be attributed to old population structure among Sumatran orang-utans or complex speciation. These two explanations differ in that the first would be consistent with multiple source populations on the island of Sumatra with little migration between these populations. The second would be consistent with a history of climatic fluctuations leading a land bridge, which allowed gene flow between species with subsequent isolation [6]. The outlier Sumatran orang-utan, would then be due to a rare female migration from the Bornean population and persistence of the Bornean mtDNA type in Sumatra with subsequent divergence.

Another phylogenetic tree based on the mtDNA coding region was also constructed using the 10 individuals from our sample including the outgroups (Figure not shown here) to make sure the estimates of the divergence time attended from the first tree was not severely biased by nor the noisy in HVR [9] or po-

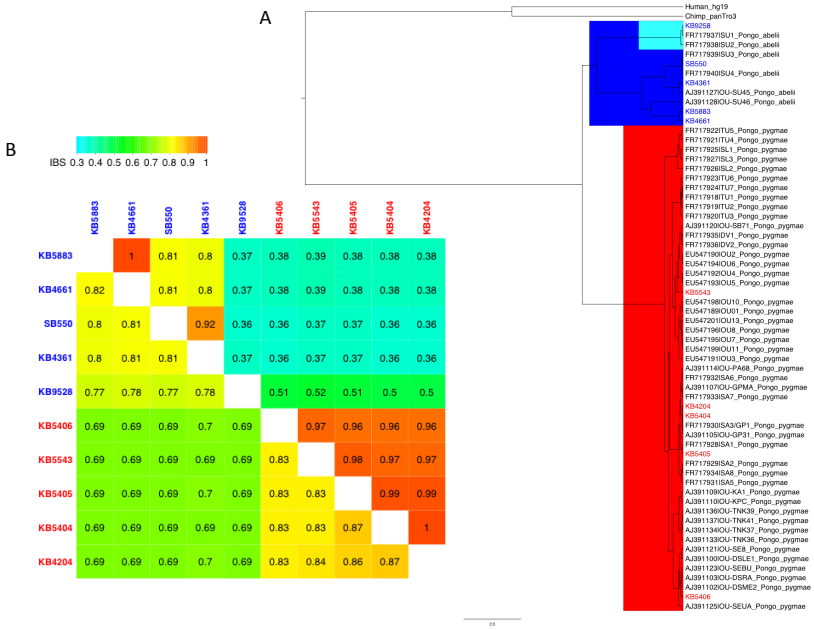



Figure 2.5: Demographical history of Orang-utan using mtDNA. A,phylogenetic tree among human, chimpanzee and orang-utan species. The clade that been highlighed with red color is the Bornean branch where as the blue color represented the Sumatran branch. The nodes that have been colored in cyan represented the grouping of the three Sumatran individuals that has the furthest distance to the rest clustering of the Sumatran group and taxas of the 10 individuals from our study has been colored with red or blue depends on whether it's a Bornean or Sumatran sample respectively; B,IBS sharing matrix among the 10 individuals from our study. Red names represent Bornean sample where blue names represented Sumatran samples. The colors within the matrix gradient from light blue to orange along with the increasing of the IBS sharing percentage. The upper triangular matrix is the IBS sharing based on mtDNA locus with variation (n=1084) whereas the lower triangular matrix is based on all of the autosomal variation locus (n=11866619).

tential multiple hits which could increase the chance for homoplasy. The estimated split time between two populations were slightly shorted compared to the pervious one (TMRCA=3.1Ma) with one outlier Sumatran individual clustered with the Bornean group. This is expected because we have included less

number of samples here hence less variation sampled from each of the population. Therefore, the divergence time between these two groups would be smaller and given that this particular Sumatran individual had a fair amount of Bornean-like haplotypes on mtDNA, it fails to cluster with other Sumatran due to the lack of similar haplotypes in its own kind, but ends up in the Bornean group.

To make sure the pattern that observed on the HRVI phylogenetic tree is not a local effect retained in that small region, we also examined allele sharing across the 10 orang-utan individuals by calculating identity by state (IBS) coefficients (i.e., proportion of times a given pair of individuals had the exact same genotype across SNPs with alleles ”A” and ”a”) among all pairs of individuals for autosomal genomic region and mtDNA coding region, respectively (Figure 2.5B). Despite the Sumatran individual KB9528, we find that allele sharing clearly tracks subpopulation on both regions, as identified by the PCA analysis in previous study [60], suggesting that vast majority of the genetic variations occur within subspecies. Moreover, Bornean population had a higher IBS sharing percentage on both autosomes and mtDNA region than that of Sumatran is well consistent with the fact that there is a higher level of nucleotide diversity in Sumatran [60]. We also identified an Sumatran individual (KB9528) had a higher IBS sharing with its own group on genomic region but substantial allele sharing with the other group(i.e., Bornean population) on mtDNA. This collapses well with the phylogenetic tree highlighting the potential of complex histories which can not be detected by using nuclear DNA because it homogenized due to males migrating in additional with the genomes recombining.

## 2.7 Functional annotation of autosomal variations

Given the demographic history that we posit using autosomal and mtDNA genomic region, there will be certain expectations about the pattern we will see in the site frequency spectrum. In order to test if the empirical site frequency spectrum reflects the expectation, we annotated all the genomic SNPs found by reseuqncing 5 Bornean and 5 Sumatran orang-utans based on their structure genomic elements, resulting a total of 20864 exonic SNPs of which 12265 synonymous and 8600 nonsynonymou SNPs. The raw counts distribution of the number of synonymous and non synonymous SNPs for two subspecies were plotted separately and presented in Figure2.6A along with the sites frequency spectrum for genomic, synonymous and non-synonymous SNPs (Figure2.6B). We further did the comparison of the two-dimensional SFS of SNPs among these three categories (Figure2.6C).

Methylation of Cs in most CpGs, which enhances the deamination of C and introduces the recurrent mutation at the same site, can cause the most CpGs in mammals uniquely hypermutable [41, 57]. Therefore, it will not only cause bias in Site frequency spectrum(SFS) but also affect all of the downstream analysis including selection inferring which will be based on the SFS of the nonsynonymous sites. In order to rule out this possibility, for each of the two subspecies, we compared the SFS of the whole genomic segregated sites, the sites may potentially be affected by CpG methylation, and the sites that are free of this bias (Figure2.7). We applied two-sided Mann-Whitney U test on data and the test statistics is not significant at all, (p-value = 0.9705 for Bornean group and 0.6842 for Sumatran group), indicating that CpG methylation hadn't have a strong effect on shaping SFS.

In SFS, the observed proportionally more non-synonymous singletons than synonymous singleton is suggestive of negative/purifying selection in both sub-species [23]. There is a weak trend that synonymous SNPs are more likely to be private to a population and to have high frequency within that population (Figure 2.6C: Bottom-left) and there is a strong trend for nonsynonymous SNPs to be lower in frequency that synonymous (Figure 2.6C: Bottom-right). The finding of a proportionally large number of non-synonymous changes in the singleton class is consistent with either selection against these potentially deleterious sites driving down the frequency or insufficient time for selection to act on recent, and therefore low frequency sites [15].

Additionally, based on the constructed complete gene orthologs alignments between human, orang-utan and chimp, using human (hg18) as an outgroup, we further classified each of the non-synonymous and synonymous SNPs into categories of polymorphism vs. divergent for two species (table 2.4). We found 51171/50603 fixed synonymous difference over a 6.76M aligned coding region yield a genomic average synonymous divergence rate of 3.36%/3.32% in Bornean/Sumatran. Correspondingly, we found another 27782/27454 fixed non-synonymous difference which gave an estimated non-synonymous divergence rate of 0.53%/0.52% respectively for each subspecies. The values in the divergence non-syn and syn are very similar in both groups because the outgroup is too far away and the split time is relatively resent. We also discovered 5302 synonymous and 4042 non-synonymous SNPs among the 5 Bornean individuals yielding a 0.35% for synonymous substitution rate and 0.08% for non-synonymous substitution rate. Likewise, 7076 synonymous and 4847 non-synonymous SNPs among the 5 Sumatran individuals gave us a 0.46% for synonymous substitution rate and 0.09% for non-synonymous substitution rate

within Sumatran group. For both Synonymous and non-synonymous SNPs, the larger numbers from Sumatran groups are consistent with previous studies finding higher levels of genetics variability in Sumatran population. The excess of synonymous variation in the Sumatran group (7k vs. 5.3k) is primarily just due to the difference in the Ne values in two different populations and this actually fits quite well to the demographic model where larger Ne corresponds to more neutral variations overall. Therefore, our Ne estimate is telling us the same thing as the MK table: more neutral in Sumatran, same as the SFS as well.

Notably, the ratio of non-synonymous to synonymous differences(15.78%) is smaller than the ratio of non-synonymous to synonymous polymorphisms(21.16%/19.91%) in both groups, indicating a highly significant excess of amino acid variation relative to divergence (Bornean: $\chi^2$ = 235.7, p-value < 2.2e-16; Sumatran: $\chi^2$ = 134.8, p-value < 2.2e-16) [12]. Moreover, the fact that the ratio of non-synonymous to synonymous divergent and non-synonymous to synonymous polymorphism is slightly smaller within Bornean (0.75) than Sumatran(0.79) showed that there are more deleterious mutations within the group of Bornean($\chi^2$ = 14.51, p-value < 0.0001396), which is expected giving that Bornean had a population size decline while Sumatran had a population size expansion based on the demographic model using the whole genomic data and hence Bornean group would cumulated more deleterious mutations.

## 2.8 Polyphen analysis

To estimate the number of deleterious mutations carried within Bornean and Sumatran group, we looked at 2039 unique transcripts in the Orang-utan

Table 2.4: Summary statistics of McDonald-Kreitman cell entries for Bornean and Sumatran population using Human (hg19) as an outgroup

Bornean(Modest population size decline)

| | Divergent | Divergent% | Polymorphism | Polymorphism% | Total aligned length |
|---|---|---|---|---|---|
| Synonymous | 51171 | 3.36% | 5302 | 0.35% | 1.52M |
| Nonsynonymous | 27782 | 0.53% | 4042 | 0.08% | 5.24M |
| Nonsyn/syn diference: 15.78% Nonsyn/syn polymorphism: 22.16% | | | | | |

Sumatran(5X population size expansion)

| | Divergent | Divergent% | Polymorphism | Polymorphism% | Total aligned length |
|---|---|---|---|---|---|
| Synonymous | 50603 | 3.32% | 7076 | 0.46% | 1.52M |
| Nonsynonymous | 27454 | 0.52% | 4847 | 0.09% | 5.24M |
| Nonsyn/syn diference: 15.78% Nonsyn/syn polymorphism: 19.91% | | | | | |

genome comprising 1711 non-synonymous SNPs, and then used PolyPhen-2 algorithm [3] which predicts whether a SNP is ”benign”, ”possibly damaging” or ”probably damaging” on the basis of evolutionary conservation and structural data [62]. The number of sites we considered here is very conservative and underestimated because we have applied a series of stringent filtering steps to remove any ambiguities or ascertainment bias due to the low quality region of the genome assembly. Figure2.8 compared the proportional number of sites within each of the three categories for the two populations along with the raw counts for each of the bins. We found that there are slightly more ”benign” mutations within Sumatran group whereas Bornean group has more ”possibly damaging” and ”probably damaging” variation. Despite the lack of statistical difference which is likely due to the small number of transcripts that passed the filter, this observed trend coincide with the estimate demographical history very well and this effect could be more exaggerated if we would be able to surveyed more

genes.

## 2.9 Detecting patterns of Selection in orang-utans using ∂a∂i

In order to differentiate if the pattern of SFS and the varying level of deleterious mutation accumulation can be simply attributed to the different demographic history within the Bornean and Sumatran population or if there is further evidence of stronger selection in one of the populations instead, we fitted the nonsynonymous spectrum using multiple models that accounts for both demographic and selective forces to access the strength of the selection for the two orang-utan groups. Our baseline demographic model is an isolation-with-migration model in which the two populations change size exponentially after divergence. It is thus be important to evaluate the uncertainty introduced into our selection inferences by uncertainty in the demographic history of these populations. Given the huge non-coding data set fit previously, we expect that systematic uncertainties due to model choice dominate statistical uncertainties in the demographic parameters. Thus we evaluate several demographic models: the full model, a model with no migration, a model with no migration, and a model with no population growth. The parameters for these models are detailed in supplementary table1, along with the log-likelihoods for the log-likelihood of each of these models to the synonymous and non-coding data. The full model fits substantially better than the other models. When fitting the selection models to the nonsynonymous data, the demographic parameters were held fixed with a fixed population-scaled rate of mutation influx $\theta_{non}$, which is set to be 2.5 times the influx for synonymous mutation rate $\theta_{syn}$.

Among all the fitting models with a single selection coefficient to all populations, the best fitting model is simply that 36% of mutations are moderately deleterious, with $\gamma$ = -0.85, and the reminder lethal(pt mass+lethal). The next best-fitting model is the normal + lethal model (although it drives to $\sigma$= 0, becoming identical to the pt mass + lethal model. The third-best model is the exponential + lethal model (Table 2.5). If we relax the assumption to allowed different selection coefficients between populations, then these models improve somewhat on the fits achievable with uniform selection across populations (Table 2.6).

Overall, It indicates the distribution of selection forces is fairly similar to that in human [10](Figure2.9) and interestingly, all our results point to roughly 80% of mutations having a selection coefficient more negative than $s \approx 3 \times 10^{-5}$ (Table 2.7). Even though the full model has the best fitting, the models with and without migration give similar results for the distribution of fitness effects, suggesting the uncertainty of the migration parameter will unlikely to cause any huge bias in terms of the selection strength inference and there may not be that much information about $\gamma$ in the shared polymorphisms. However, not modeling population growth does change the results more substantially.

## 2.10 Discussion

The orang-utan story is thus a tale of two islands with distinct evolutionary histories. Our high-resolution population studies explored the counter-intuitive nature of orang-utan diversity - greater variation among Sumatran orang-utans than their Bornean counterparts despite a smaller population size (approxi-

Table 2.5: Inferred parameters for single-$\gamma$ selection models

| selection | | | demographic model | | | no 9528 |
|---|---|---|---|---|---|---|
| model | distribution | param | full | no-mig | no-growth | full |
| neutral | $\Pr(\gamma = 0) = 1$ | $LL$ | -6170.5 | -6221.86 | -6328.4 | -2632.6 |
| fixed (pt mass) | $\Pr(\gamma = k) = 1$ | $LL$ | -1396.8 | -1510.8 | -1073.5 | -668.9 |
| | | $k$ | -3.1 | -3.1 | -3.5 | -3.1 |
| exponential | $\Pr(\gamma = -x) = EXP(\lambda)$ | $LL$ | -648.7 | -739.5 | -521.7 | -367.9 |
| | | $\lambda$ | 10.2 | 10.6 | 11.5 | 10.1 |
| neutral + lethal | $\Pr(\gamma = 0) = p^0; \Pr(\gamma = -\infty) = 1 - p^0$ | $LL$ | -313.3 | -365.0 | -471.6 | -216.9 |
| | | $p^0$ | 0.28 | 0.28 | 0.28 | 0.28 |
| normal | $\Pr(\gamma = x) = NORM(\mu, \sigma)$ | $LL$ | -389.8 | -402.9 | -318.5 | -253.1 |
| | | $\mu$ | -16.4 | -17.2 | -19.4 | -16.9 |
| | | $\sigma$ | 12.0 | 13.1 | 14.8 | 12.5 |
| pt mass + lethal | $\Pr(\gamma = k) = p; \Pr(\gamma = -\infty) = 1 - p$ | $LL$ | -257.05 | -313.2 | -307.9 | -195.9 |
| | | $p$ | 0.36 (0.34,0.37) | 0.35 | 0.43 | 0.35 |
| | | $k$ | -0.85 (-0.99,-0.72) | -0.83 | -1.46 | -0.82 |
| exponential + lethal | $\Pr(\gamma = -x) = pEXP(\lambda); \Pr(\gamma = -\infty) = 1 - p$ | $LL$ | -259.0 | -314.3 | -307.9 | -197.0 |
| | | $\lambda$ | 1.05 | 1.00 | 2.5 | 0.96 |
| | | $p$ | 0.36 | 0.35 | 0.46 | 0.35 |
| exponential + neutral | $\Pr(\gamma = -x) = pEXP(\lambda); \Pr(\gamma = 0) = 1 - p$ | $LL$ | -279.9 | -326.2 | -321.6 | -208.7 |
| | | $\lambda$ | 126.4 | 133.7 | | 172.0 |
| | | $p$ | 0.235 | 0.233 | 0.180 | 0.244 |
| gamma | $\Pr(\gamma = -x) = GAMMA(\alpha, \beta)$ | $LL$ | -268.7 | -319.2 | -313.0 | -202.8 |
| | | $\alpha$ | 0.16 | 0.16 | 0.28 | 0.15 |
| | | $\beta$ | 14392 | 18725 | 547 | 27515 |
| neutral + pt mass + lethal | $\Pr(\gamma = 0) = p^0;$ | $LL$ | -257.1 | -313.2 | -308.0 | -195.9 |
| | $\Pr(\gamma = k) = p;$ | $p^0$ | 0.02 | 0 | 0.03 | 0 |
| | $\Pr(\gamma = -\infty) = 1 - p^0 - p$ | $p$ | 0.34 | 0.35 | 0.40 | 0.35 |
| | | $k$ | -0.91 | -0.84 | -1.61 | -0.81 |
| neutral + gamma | $\Pr(\gamma = 0) = p^0;$ | $LL$ | -269.0 | -319.3 | -313.3 | -203.1 |
| | $\Pr(\gamma = -x) = (1 - p^0)GAMMA(\alpha, \beta)$ | $p^0$ | 0.01 | 0.01 | 0.00 | 0 |
| | | $\alpha$ | 0.16 | 0.17 | 0.31 | 0.15 |
| | | $\beta$ | 13988 | 13662 | 316 | 18285 |
| neutral + exponential + lethal | $\Pr(\gamma = 0) = p^0;$ | $LL$ | -259.0 | -314.5 | -707.9 | -197.1 |
| | $\Pr(\gamma = -x) = (1 - p^0 - p)EXP(\lambda);$ | $p^0$ | 0.00 | 0.02 | 0 | 0 |
| | $\Pr(\gamma = -\infty) = p$ | $\lambda$ | 1.08 | 1.11 | 2.54 | 0.99 |
| | | $p$ | 0.65 | 0.64 | 0.54 | 0.64 |
| normal + lethal | $\Pr(\gamma = x) = pNORM(\mu, \sigma)$ | $LL$ | -257.4 | -293.2 | -307.8 | -196.1 |
| | $\Pr(\gamma = -\infty) = 1 - p$ | $p$ | 0.36 | 0.43 | 0.44 | 0.35 |
| | | $\mu$ | -0.84 | -4.11 | -1.81 | -0.81 |
| | | $\sigma$ | 0.01 | 6.45 | 1.15 | 0.01 |

Table 2.6: Inferred parameters for multiple-$\gamma$ models

| selection | | | demographic model | | | no 9258 |
|---|---|---|---|---|---|---|
| model | distribution | param | full | no-mig | no-growth | full |
| 3 fixed | $\Pr(\gamma_A = k_A, \gamma_B = k_B, \gamma_S = k_S) = 1$ | $LL$ | -836.2 | | | |
| | | $k_A$ | -0.23 | | | |
| | | $k_B$ | -5.64 | | | |
| | | $k_S$ | -6.78 | | | |
| 2 fixed + lethal | $\Pr(\gamma_A = (k_B + k_S)/2, \gamma_B = k_B, \gamma_S = k_S) = p$; | $LL$ | -241.7 | -286.4 | -279.5 | -179.1 |
| | $\Pr(\gamma = -\infty) = 1 - p$ | $k_B$ | -0.32 | 0.03 | -0.33 | 0.00 |
| | | $k_S$ | -1.35 | -1.69 | -2.52 | -1.61 |
| | | $p$ | 0.36 | 0.36 | 0.43 | 0.36 |
| 3 fixed + lethal | $\Pr(\gamma_A = k_A, \gamma_B = k_B, \gamma_S = k_S) = p$; | $LL$ | -227.4 | -185.6 | -168.3 | -173.7 |
| | $\Pr(\gamma = -\infty) = 1 - p$ | $k_A$ | 0.89 (0.40,1.32) | 1.29 | 0.611 | 1.08 |
| | | $k_B$ | -3.97 (-4.57,-2.95) | -8.15 | -11.71 | -3.91 |
| | | $k_S$ | -4.32 (-4.85,-3.51) | -8.43 | -12.09 | -4.74 |
| | | $p$ | 0.42 (0.39,0.44) | 0.44 | 0.52 | 0.42 |

Table 2.7: Cumulative selection coefficient probabilities

| demographic scenario | selection model | $s < -10^{-2}$ | $-10^{-2} < s < -10^{-3}$ | $-10^{-3} < s < -10^{-4}$ | $-10^{-4} < s$ |
|---|---|---|---|---|---|
| full model | pt + lethal | 64 | 0 | 0 | 36 |
| | expon + lethal | 64 | 0 | 7 | 29 |
| | norm + lethal | 64 | 0 | 0 | 36 |
| | gamma | 47 | 16 | 11 | 26 |
| no migration | pt + lethal | 65 | 0 | 0 | 35 |
| | expon + lethal | 65 | 0 | 5 | 30 |
| | norm + lethal | 57 | 0 | 27 | 16 |
| | gamma | 48 | 16 | 11 | 25 |
| no growth | pt + lethal | 57 | 0 | 0 | 43 |
| | expon + lethal | 54 | 0 | 19 | 27 |
| | norm + lethal | 56 | 0 | 16 | 28 |
| | gamma | 20 | 35 | 21 | 24 |

mately 7-fold lower by recent estimates). Further dissection of the orang-utan speciation process will require a broader survey, incorporating representatives from additional orang-utan subpopulations.

Finally, even though we found deep diversity in both Bornean and Sumatran populations, it is not clear whether this diversity will be maintained with continued habitat loss and population fragmentation. Evidence from other species suggests fragmentation is not the death knell of diversity [4], but their slow reproduction rate and arboreal lifestyle may leave orang-utan species especially vulnerable to rapid dramatic environmental change. It is our hope that the genome assembly and population variation data presented here provide a valuable resource to the community to aid the preservation of these precious species.

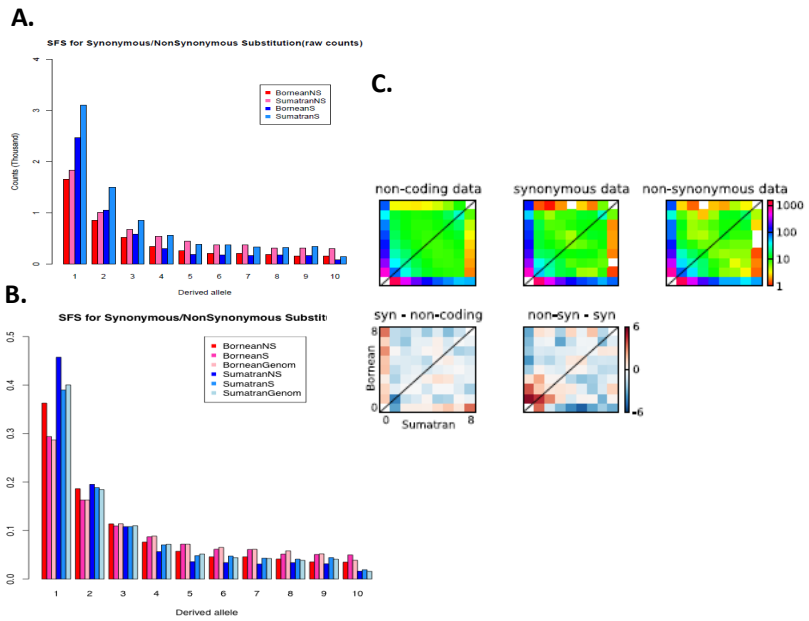



Figure 2.6: Summary of site frequency distribution of functional annotated variations. A,Raw counts distribution for the synonymous and non synonymous SNPs within Bornean and Sumatran populations respectively; B,site frequency spectrum allocated all SNPs (n=11866619) into three different groupings include Synonymous, Nonsynonymous and Genomic region for Bornean(Red color scheme) and Sumatran(Blue color scheme) population separately based on the SNP functional annotation; C,Two-dimensional SFS of SNPs in different functional categories along with the residuals between them. The three data sets studied: non-coding spectrum, synonymous data, nonsynonymous data (Top). For purposes of comparison, the non-coding and synonymous spectra have been re-scaled to represent the same number of segregating SNPs as the non-synonymous spectrum. Bottom-left is the residuals between the synonymous and non-coding spectra. Red and blue indicate, respectively, that the synonymous spectrum has greater or fewer SNPs in that entry than the non-coding data whereas the Bottom-right plot is residuals between the synonymous and nonsynonymous spectrum.

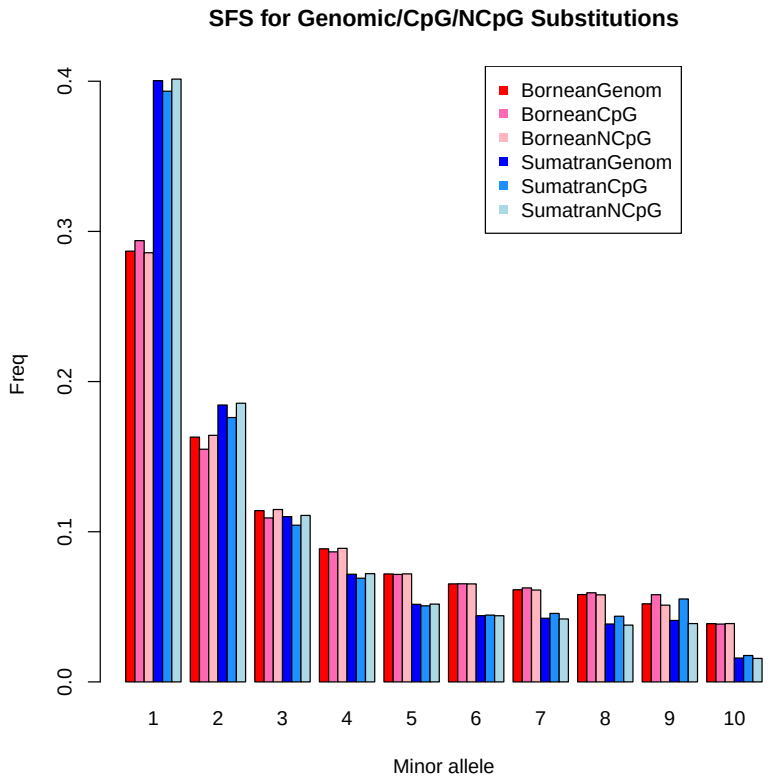



Figure 2.7: SFS for SNP with groupings of CpG sites, nonCpG sites, genomic regions

| | Benign | Possibly | Probably | Total |
|---|---|---|---|---|
| Private to Bornean | 413 | 130 | 155 | 698 |
| Private to Sumatran | 416 | 108 | 142 | 666 |
| Shared | 214 | 45 | 88 | 347 |
| Total | 1043 | 283 | 385 | 1711 |

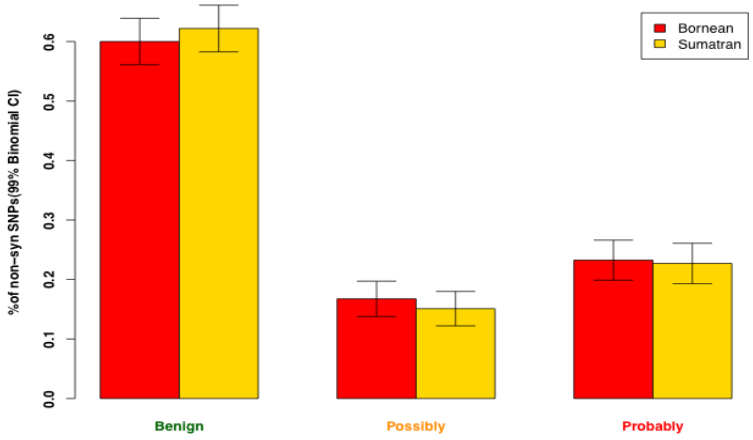



Figure 2.8: Distribution of the proportion SNPs within each of the Polyphen categories based on Polyphen2 result with bars representing the 95% CIs along with the raw counts break ups.

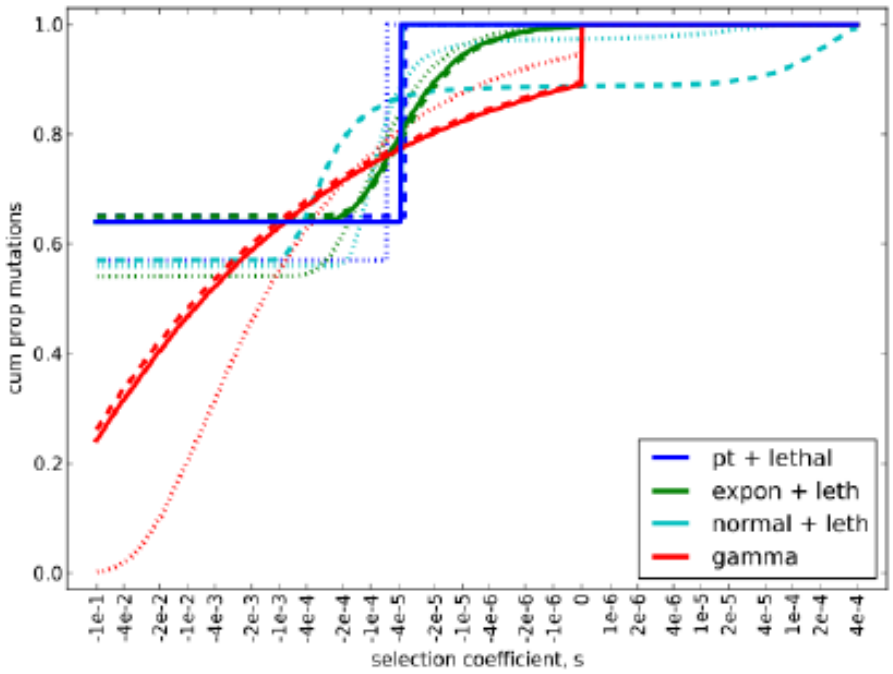



Figure 2.9: Cumulative $s$ distributions. Solid lines are for the full demographic model, dashed are for the model with no migration, and dotted are for the model with no growth.

CHAPTER 3

# DETECTION OF HETEROZYGOUS MUTATIONS IN THE GENOME OF MISMATCH REPAIR DEFECTIVE DIPLOID YEAST USING A BAYESIAN APPROACH

The material described within this chapter comes from a mixture of the Genetics paper with first author of Sarah Zanders, myself, Arindam RoyChoudhury, and PLoS-Genetics papers with first author of Julie Heck [106, 36]. I am deeply grateful for all the wet lab experimental work that was done by Eric Alani's lab in Cornell including Julie Heck, Sarah Zanders and Eric Alani himself. I am also appreciative of the advices that provided by Arindam RoyChoudhury and Amit Indap who was a former post-doc and a former technician, respectively, in Carlos bustamante's lab .

## 3.1 Introduction

Mutation rates in prokaryotic and eukaryotic organisms are typically determined by measuring reversion or forward mutation for specific marker alleles. These values are then extrapolated to obtain genome-wide estimates. Mutation rates in higher eukaryotes are also estimated by analyzing sequence divergence between different strains or species, followed by reconstructing the accumulation of mutations since divergence (reviewed in [77]). These approaches suffer from two main limitations. First, recent studies have shown that mutation rate and repair efficiency vary across the genome and are affected by parameters that include base composition, local recombination rate, gene density, transcriptional activity, repair efficiency, chromatin structure, nucleosome position, and

replication timing [34, 103, 67, 5, 32, 17, 92, 100, 88]. Second, genomic comparisons can yield inaccurate rate measurements because DNA repair and subsequent purifying natural selection can bias the number and type of mutations that remain in the population, especially for mutations that occur in coding regions (reviewed in [77]).

The DNA mismatch repair system improves the fidelity of DNA replication by about 1000-fold by excising DNA mismatches in the newly replicated strand that arise from polymerase misincorporation and slippage (reviewed in [72, 48, 68]. Eukaryotes contain multiple MutS (MSH) and MutL (MLH) homologs (reviewed in [48]). In *S. cerevisiae*, two heterodimeric MutS homolog complexes, MSH2-MSH3 and MSH2-MSH6, act in mismatch recognition. MSH2-MSH6 is primarily involved in repairing base-base and small insertion/deletion loop mismatches. MSH2-MSH3 acts primarily on insertion/deletion loop mismatches up to seventeen nt in length. In the presence of ATP, both MSH complexes interact primarily with MLH1-PMS1 to form a mismatch-MSH-MLH complex that interacts with downstream repair components. Recent work in humans and yeast suggests that MLH1-PMS1 contains an ATP-$Mn^{2+}$-dependent latent endonuclease activity that acts near the mismatch and is essential for MMR, most likely in excision steps [45, 44]. Null mutations in MSH2 and MLH1, the key partners in the MSH and MLH complexes, confer severe defects in MMR; reporter assays have shown that strains bearing these mutations display high rates of base substitutions and DNA slippages. For example, in an assay that measures frameshift mutations in homopolymeric runs, *msh2*$\Delta$ and *msh1*$\Delta$ mutations confer mutation rates that are ~ 10,000-fold higher than wild-type [93, 94, 66, 29].

Our goal in this study was to analyze with limited bias the rate at which mutations occur in MMR defective lines due to DNA polymerase errors during DNA replication, and to identify novel genomic features associated with these errors. The baker's yeast *S. cerevisiae* is an ideal model system to perform these studies because genetic analysis of many of the key MMR factors has been performed; more importantly the effect of null mutations in these factors has been extensively characterized using a variety of mutator assays [48]. Previously, Heck et al. [36] grew wild-type and conditional *mlh1* (*mlh1*-$7^{ts}$) diploid strains of *S. cerevisiae* for 160 generations with bottlenecks that reduced the population size to one cell every 20 generations. These lines were grown at $35^oC$, the non-permissive temperature for *mlh1*-$7^{ts}$. A conditional *mlh1* allele was chosen instead of a null so that mutation accumulation in the absence of MMR could be limited to 160 generations by shifting cells at generation 160 to the permissive temperature for MMR function. The *mlh1*-$7^{ts}$ mutation contains two mutations within the ATP binding domain of MLH1 (*K67A*, *D69A*). Unlike *mlh1*$\Delta$ strains that display poor spore viability due to defects in meiotic crossing over, *mlh1*-$7^{ts}$ lines display wild-type spore viability at the permissive temperature. Such a phenotype allowed us to easily identify recessive lethal mutations [36]. At the non-permissive temperature, the *mlh1*-$7^{ts}$ mutation conferred a phenotype similar to the null in the canavanine resistance mutation assay and a mutator phenotype in the *lys2*-$A_{14}$ reversion assay that was 1000-fold higher than *MLH1* but four-fold lower than the null ([36]; Julie Heck, and E. A., unpublished observations).

Tetrad analysis showed that the *mlh1*-$7^{ts}$ bottleneck lines would be ideally suited for a high-throughput DNA sequencing approach that would identify mutagenesis patterns. First, the wild-type lines maintained high spore viability

(~94%) at generation 160. In contrast, *mlh1*-7$^{ts}$ lines displayed spore viabilities that ranged from 1.1 to 77%, demonstrating that the lines had accumulated recessive lethal mutations. Second, comparative genome hybridization (CGH) and pulse-field (PFGE) analysis of the *mlh1*-7$^{ts}$ strains indicated that they did not undergo major genome rearrangements [36]. Third, because the lines were grown as diploids for a limited number of generations, secondary mutations (dominant or recessive) should rarely occur that alter the rate or type of mutagenesis. Also, because there is no sexual reproduction and mutations should clonally propagate after escaping the initial bottleneck, newly arising mutations should appear as heterozygous sites. Finally, the above strategy should limit biases in mutation accumulation because the diploid cells were grown in rich media under minimal selection pressure where deleterious mutations could accumulate [36].

As described below, a Bayesian method was developed to detect heterozygous mutations in one wild-type and three *mlh1*-7$^{ts}$ lines using whole-genome sequencing. We detected 28 heterozygous single-nucleotide polymorphisms (SNPs) and 48 single nt insertion/deletion (indels) in the mutator lines, all of which mapped to homopolymeric runs of nucleotides (HP tracts).

The mutation spectra match closely with that seen in MMR defective strains using different reporter constructs [93, 94, 66]. This demonstrates that the mutation pattern seen previously using a limited number of reporters holds true for the entire genome. In addition, we were able to correlate genotype to phenotype for one locus in one mutator line. Together this work provides new insights into how mismatch repair can shape genome stability and dynamics, mutation mechanisms and evolution.

## 3.2 Strain selection, sequencing and reads mapping

Bottleneck experiments involving ten independent wild type (*MATa/MATalpha*, *his3/HIS3*, *LEU2/leu2*, *cyhr/cyhs*, *ade2/ADE2*, *ura3/ura3*, *trp1/trp1*) and *mlh1-7*$^{ts}$ (*MATa/MATalpha,mlh1-7::KanMX4/mlh1-7::KanMX4*, *his3/HIS3*, *LEU2/leu2*, *cyh*$^{r}$*/cyh*$^{s}$, *ade2/ADE2*, *ura3/ura3*, *trp1/trp1*) lines were performed previously by J.A.Heck [36].

One wild-type and three *mlh1-7*$^{ts}$ lines (Mut2, Mut3, and Mut4) allowed to accumulate mutations for 160 generations were sequenced at the Cornell University Life Sciences Core Laboratory Center (CLC) using the Illumina Genome Analyzer technology (http://www.illumina.com) where these three mutator lines were chosen to ensure a reasonable sample set of mutations, and displayed a lower range of spore viabilities (2.5-15.6%) following tetrad dissection compared to the entire set (1.1-77%). The wild-type progenitor of all the strains was also sequenced. The analysis was performed with three independent *mlh1-7*$^{ts}$ lines to control for chance associations within an individual line and for mutations that could alter the mutation rate of a given line. The Mut2, Mut3, and Mut4 lines at generation 160 displayed 15.6, 7.1, and 2.5 % spore viability, respectively [36].

Yeast genomic DNA for whole genome sequencing was prepared using a Qiagen genomic DNA preparation kit (www.qiagen.com). Sequencing was performed using the Illumina pipeline for 36 bp single-end reads. The wild-type and Mut2 generation 160 strains were sequenced to 9X and 8X average genome coverage depth, respectively. Mut3 (160) and Mut4 (160) were sequenced to average depths of 16X and 20X, respectively. Reads were aligned onto the

S288c genome (http://genome.ucsc.edu/cgi-bin/hgGateway) using Novoalign (www.novocraft.com), a program that performs a gapped alignment with high specificity and sensitivity. In total, 25 million, out of 35 million sequenced, 36 nt sequence reads were uniquely mapped to the yeast genome, allowing up to two mismatches per read.

## 3.3 Experimentally-aware Bayesian genotype caller

We analyzed five diploid strains in this study: a wild-type strain at generations 0 and 160 (Wt0, Wt160) and three derived *mlh1*-7$^{ts}$ mutator lines grown vegetatively (i.e., no meiosis) and bottlenecked to one cell every twenty generations until generation 160 (Mut2, Mut3, Mut4). Several aspects of the experiment required us to develop a novel approach for calling genotypes from the sequencing data. First, the initial wild-type strain (Wt0) likely contained SNPs and indels that distinguish it from the reference yeast genome. Because all lines were grown vegetatively, they were all expected to have the "propagated" SNPs and indels. Thus reads from the five sequenced lines were used to identify these variants. Furthermore, we expect new mutations (i.e., those occurring in Wt160, Mut2, Mut3, or Mut4 during generations 1-160) to be heterozygous at the end of the experiment and few, if any, variants expected to be shared (i.e., would require independent hits in replicate lines). Lastly, the sequencing depth ($\sim 8-16X$) suggests moderate but not exceptional power to detect heterozygous mutations from the sequence of a single line on its own. Therefore, we developed a Bayesian SNP caller that (1) aligns all reads to the genome, (2) uses read depth and quality scores at a given position to call genotypes for all five lines simultaneously.

Importantly, our Bayesian model allows us to distinguish between a propagated mutation, (defined as a variant seen in all five strains in either heterozygous or homozygous state from Wt0) and a derived mutation, defined as a DNA sequence variant that arose in only a single line. First, we indexed the five diploid strains as s = 1, 2, 3, 4, 5 for Wt0, Wt160, Mut2, Mut3, Mut4 respectively. We set the prior probability of strain s being heterozygous as Priors = $10^{-7}, 10^{-8}, 10^{-5}, 10^{-5}, 10^{-5}$ for s = 1, 2, 3, 4, and 5, respectively, according to mutation rates previously determined in wild-type and mismatch repair defective organisms [77, 42, 21]. It is important to note that Wt160 was assigned a lower prior probability of being heterozygous relative to Wt0. This is because a heterozygosity in Wt0 is defined as the difference between the Wt0 strain [36] and the S288c reference genome (http://genome.ucsc.edu/cgi-bin/hgGateway). There were a significant number of differences between the two strains. On the other hand, a heterozygosity in Wt160 was defined as one that occurred during the bottleneck experiment (propagated). Because there were only 160 generations between Wt0 and Wt160, we expected the number of differences between the lines to be small; in fact, none were detected.

At a given locus, let A and a be the major and minor allele types, respectively, based on the allele counts from all the strains. Let $N_s$ be the total number of alleles observed for strain s; let $A_j$,s be the type of the $j^{th}$ allele copy among these $N_s$ alleles, j = 0, 1, . . .$N_s$. Let $e_j$ be the probability that the $j^{th}$ allele has been assigned the wrong allele-type. We estimated $e_j$ from the error rates given by DOHM et al. [22] for 36 bp Solexa reads as a function of read position.

In order to call SNPs and indels in Wt0, we used the allele count data from Wt0 along with that from the other four strains. The posterior probabilities of a

given genomic position being homozygous or heterozygous in Wt0 are:

$$\mathbb{P}_1(Heter.|Data) = \frac{\mathbb{P}_1(Data|Heter.) \times \mathbb{P}_1(Heter.)}{\mathbb{P}(Data)} \tag{3.1}$$

$$\propto Prior_1 \times (0.5)^{\sum_{s=1}^{5} N_s} \tag{3.2}$$

$$\mathbb{P}_1(Homo.|Data) = \frac{\mathbb{P}_1(Data|Homo.) \times \mathbb{P}_1(Homo.)}{\mathbb{P}(Data)} \tag{3.3}$$

$$\propto (1 - Prior_1) \times \prod_{j=1}^{N_1} (1 - e_j)^{1_{(A_j=A)}} \times {e_j}^{1_{(A_j=a)}} \tag{3.4}$$

$$\times \prod_{s=2}^{5} ((1 - Prior_s) \times \prod_{j=1}^{N_s} (1 - e_j)^{1_{(A_j=A)}} \times {e_j}^{1_{(A_j=a)}} + Prior_s \times (0.5)^{N_s}) \tag{3.5}$$

where $P_s(.)$ denotes the probability in the context of strain s. Based on the posterior probabilities above, we classified each locus as homozygous or heterozygous for Wt0. If a locus was classified as heterozygous for Wt0, then it was assumed to have a propagated mutation in the rest of the strains. To call derived mutation in strains s = 2, 3, 4, 5, we used similar logic:

$$\mathbb{P}_s(Heter.|Data) = \frac{\mathbb{P}_s(Data|Heter.) \times \mathbb{P}_s(Heter.)}{\mathbb{P}(Data)} \tag{3.6}$$

$$\propto Prior_s \times (0.5)^{N_s} \tag{3.7}$$

$$\mathbb{P}_s(Homo.|Data) = \frac{\mathbb{P}_s(Data|Homo.) \times \mathbb{P}_s(Homo.)}{\mathbb{P}(Data)} \tag{3.8}$$

$$\propto (1 - Prior_s) \times \prod_{j=1}^{N_s} (1 - e_j)^{1_{(A_j=A)}} \times {e_j}^{1_{(A_j=a)}} \tag{3.9}$$

We used the posterior probabilities calculated above, to make a decision as to whether a site is called as heterozygous for a new mutation, heterozygous for a propagated mutation, or invariant for the four evolved strains: s = 2, 3, 4, 5. Specifically, if the posterior probability of heterozygosity was greater than 50% at a given position, then we classified the site as containing a SNP or indel. Visual inspection of the alignments for some of the inferred indel positions

revealed that pairwise alignment of reads could induce false positives across multiple lines due to variations on how the alignment software interprets the alignment of different reads around a given position. These were characterized by one allele count being much smaller (but non-zero) compared to the other, across multiple strains. To bioinformatically call such sites from our data set, we carried out an additional test by asking the question of whether if the new allele observed outside the strain which has the mutation. If observed, we flagged it as low-confidence call because under this particular experimental design, the probability that a mutation occurred simultaneously at two or more strains is negligible. (Figure 3.1).

We expected, based on previous estimates of mutation rate in MMR defective strains, to find $\sim$ 125 mutations for each of the MMR deficient strains ($\sim$ one mutation per line generation). This corresponds to a prior mutation rate of $10^{-5}$ mutations per site per generation. However, we detected 12, 24, 40 mutations for each of the MMR deficient strains, which yield mutation rates of $1 \times 10^{-6}, 2 \times 10^{-6}$ and $3 \times 10^{-6}$ in each line, respectively. Although our estimated prior values differ somewhat from the real data, the alignment analysis allowed us to calculate very accurate posterior subjective probabilities. This accuracy is due to the large number of observations and has in practice made the influence of the prior negligible. Thus given the high coverage for the Mut lines, the difference in our prior estimates does not influence our analysis. Even with low coverage data where accurate estimates of prior are critical, a higher prior value would yield a larger number of false positives. The majority of mutations (and all low confidence mutations) were verified by Sanger sequencing, suggesting that false positives were rare, but we may have false negatives (i.e., missed variants) due to the medium coverage ($\sim 8 - 16X$) of the lines.

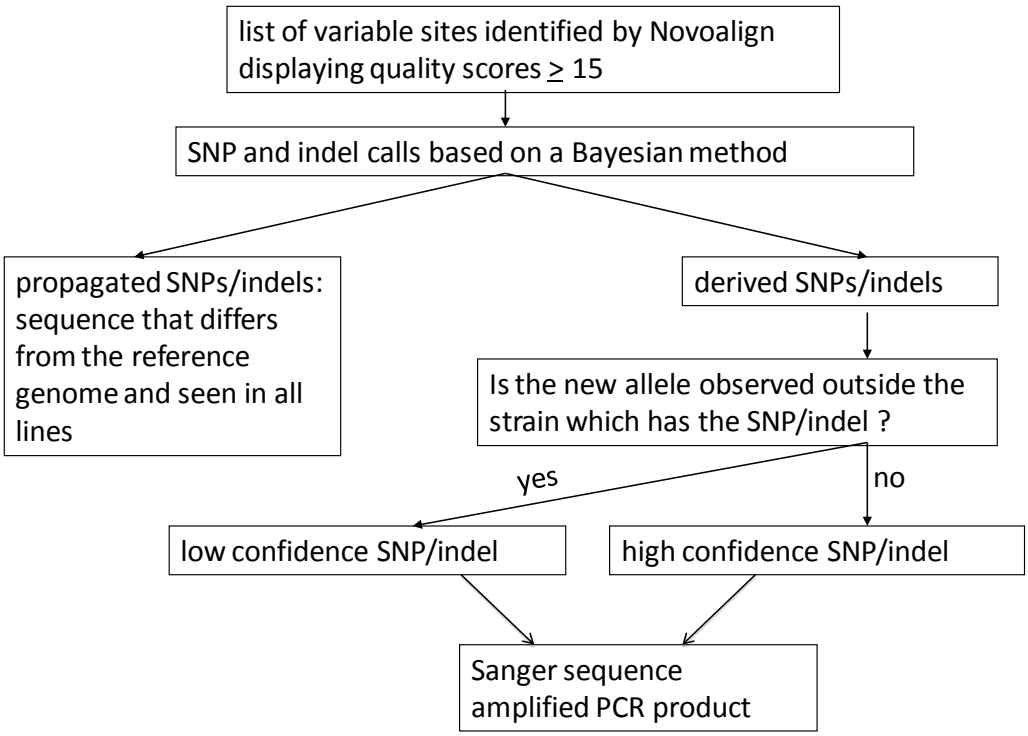



Figure 3.1: Flow chart describing bioinformatic methods used to identify heterozygous mutations from Illumina GA whole-genome sequencing

## 3.4 Simulation

To estimate the False Positive (FP) and False Negative (FN) rates (as well as to check our bioinformatics and SNP/indel calling pipelines), we set up a simulation to test the accuracy of our Bayesian approach. We started with a complete genome of a yeast S288c strain (http://genome.ucsc.edu/cgi-bin/hgGateway; June 2008 assembly from the SGD (http://www.yeastgenome.org/) and introduced SNPs and indels to simulate five strains: Wt0, Wt160, Mut2, Mut3 and Mut4. To simulate Wt0, we duplicated the S288c genome to create a diploid.

We then randomly selected $n_\mu$ and $n_d$ positions for SNPs and indels respectively. ($n_\mu = 2, n_d = 8$; the values of $n_\mu$ and $n_d$ were chosen to mimic changes between S288c and the Wt0 strain used in the bottleneck experiment). One of the two copies of S288c was randomly selected to incur each SNP or indel. For an indel mutation, the nt in that copy was deleted, or a new randomly chosen allele was inserted after it. For a SNP position, the nt was randomly changed to another nt. The resulting two copies of the genome were defined as the Wt0 diploid. The other four strains were all simulated directly from Wt0 by introducing SNPs and indels in the two copies of Wt0. The mechanism of adding SNPs and indels was exactly as described above. The values of $n_\mu$ and $n_d$ for each of the simulations are given in Table 3.1. These values mimic the number of mutations that were expected in the bottleneck experiments. One distinction between the simulations and the real data is that the indels in the simulations were not all introduced into HP tracts.

Table 3.1: SNP and INDEL introduction table for simulation. This table illustrated how Wt0 strain was created from the regular isogenic reference diploid cell along with the procedure of how the rest four strains were created independently upon the creation of Wt0

| | $0 \rightarrow Wt0$ | $WT0 \rightarrow Wt160$ | $WT0 \rightarrow Mut2$ | $WT0 \rightarrow Mut3$ | $WT0 \rightarrow Mut4$ |
|---|---|---|---|---|---|
| $n_\mu$ | 2 | 1 | 25 | 25 | 25 |
| $n_d$ | 1 | 1 | 100 | 100 | 100 |

Next, we simulated 32 nt Illumina GA reads from each of the five strains by randomly choosing read-start positions and copying 32 nt of strain s starting from that position. For each strain, the number of reads simulated matches

the coverage achieved in the real sequencing experiment. We also simulated a quality score for each position of each read, following the error rate distribution given in DOHM et al. [22]. The reads were aligned with S288c using Novoalign (www.novocraft.com). Based on the alignment, we listed the allele-counts and associated quality scores in each of the variable, potentially heterozygous, positions. We used this list as the input to a computer-program created based on our method of heterozygosity detection, which went through all the steps described in the last section. The method which was evaluated on this simulated data sets gave very low false positive rate ($\sim 6 \times 10^{-5}$) and a false negative rate of 0.08 within the unique mapping regions of the genome that contained at least seven-fold coverage. The rates of false positives and negatives (based on the output of the program) are given in Table 3.2.

Table 3.2: False-positive and -negative rates based on the simulation analysis

| | False-positive rate (in units of no. of SNP calls) | False-negative rate (in units of no. of SNP calls) |
|---|---|---|
| Mutant | $6 \times 10^{-5}$ | 0.030 |
| Indel | 0 | 0.089 |
| Total | $6 \times 10^{-5}$ | 0.078 |
| | | |
| Propagated | 0 | 0 |
| Derived | $6 \times 10^{-5}$ | 0.091 |
| Total | $6 \times 10^{-5}$ | 0.078 |

We believe that these rates are similar to those seen in the bottleneck exper-

iment and the negligible false positive and false negative rate for Propagated variation is attributed to the fact that we have boost the calling power for propagated variation by leveraging informations from the other 4 strains during the classification of WT0. Based on simulations, we estimated that the method, as applied to regions with at least seven-fold sequencing coverage, allowed us to detect heterozygous mutations in 60%, 41%, 69%, and 84% of the total genome for the generation 160 wild-type, Mut2, Mut3, and Mut4 lines, respectively. We also found that our ability to detect indels in HP tracts is lower because indels in HP tracts can be identified only if the entire tract and sequence flanking both sides are present in a 36 nt read.

## 3.5 Validation

Our method for heterozygous mutation calling from the whole-genome sequencing data yielded both low and high confidence predictions. All low confidence predictions (ten in total) were verified and either validated (n = 4) or disproved (n = 6) using Sanger sequencing. Briefly, to assay heterozygous mutations predicted from the whole-genome sequence data, genomic DNA was prepared from wild-type generation zero, and mutation accumulation lines Mut2, Mut3, and Mut4 using standard techniques by Sarah Zanders who was a former graduate student of Eric Alani's group in Cornell University. Approximately 400 bases-pairs of DNA flanking the predicted mutated site was amplified in all lines using PCR and Sanger sequenced at the Cornell CLC using an Applied Biosystems Automated 3730 DNA Analyzer. The sequencing traces were all analyzed visually. A heterozygous base change mutation was confirmed if a doublet representing both alleles was observed only in the sequencing trace of

the predicted Mut line, but all other lines showed only a singlet representing the parental allele. A heterozygous indel mutation was confirmed if the sequencing reaction failed (i.e. tall singlet peaks fall to small doublet peaks or random noise) at the predicted location only in the predicted Mut line, but the sequencing reactions in all other lines were able to successfully sequence past the site.

For the high confidence predictions, 31 (out of 65) were sequenced and verified using the methods described above. Of those 31 mutations, ten were further verified by genotyping the haploid progeny of the diploid containing the heterozygous mutation via Sanger sequencing. Both alleles comprising the heterozygote were observed in the haploid progeny with the exception of the frameshift mutation in the essential *MDN1* gene. Six additional high confidence predictions were also verified by genotyping the haploid progeny of the heterozygous diploid.

We also found and verified by Sanger sequencing of the diploid lines four heterozygous mutations that were detected in earlier, less accurate prediction protocols that were not found using the final more stringent prediction method.

## 3.6 Bioinformatics of detected variations and discussion

We developed and employed an ”experiment aware” probabilistic framework using maximum likelihood and Bayesian methods that utilized sequence coverage of the entire data set ( 70-fold; [22]). Briefly, the approach classifies each site in the yeast genome with uniquely mapping reads into one of three categories: (1) invariant across all strains, (2) heterozygous in the wild-type (and all derived strains) which we term ”propagated” SNPs or indels, or (3) heterozy-

gous in one of the mutant strains which we term "derived" SNPs or indels. This method allowed us to pool experimental data across sequencing runs for all strains and detect with high reliability heterozygous single-nucleotide polymorphisms (SNPs; 28 identified) and single nt indels (48 identified) from the 36 nt read data set.

Overall, we did not detect any mutations in the wild-type generation 160 line, which was predicted based on the previously calculated mutation rate of $3.3 \times 10^{-10}$ mutations per base per generation (< 1 expected; [65]). As showed in Table 3.3, only heterozygous mutations, comprised of 28 base substitution and 48 single nt indel mutations, were detected in the three MMR-defective lines. All of the mutations were unique between lines except for a single nt deletion mutation between Saccharomyces Genome Database (SGD; http://www.yeastgenome.org) coordinates 92,271-92,279 on chromosome II, which occurred independently in both Mut2 and Mut3 (Table 3.3). All 48 indels, comprised of 46 deletions and 2 insertions, occurred in HP tracts (47 poly A or T tracts, 1 poly G or C tract) between 5-13 base pairs long (Table 3.3 and 3.4).

Table 3.3: Genome location of mutations detected in the Mut2, 3, and 4 lines

| | | | | | | | Distribution of Sequence reads | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Chromosome | SGD Position | HP tract | strain | Gene | aa change | Mutation | A | G | T | C | Indel |
| 1 | 139,349-139,358 | 10 | 4 | | | del | 1 | | 11 | | 8 |
| 2 | 275,549-275,557 | 9 | 2 | | | del | | | 3 | | 3 |
| 2 | 92,271-92,279 | 9 | 3 | | | del | 1 | | 10 | | 8 |
| 2 | 92,271-92,279 | 9 | 2 | | | del | | | 3 | | 0 |
| 2 | 423,462-423,469 | 8 | 3 | | | ins | 8 | | | | 16 |
| 2 | 662,560-662,569 | 10 | 4 | *YBR219C* | fs;103-106 | del | 6 | | | | 5 |
| 2 | 653,035-653,045 | 11 | 4 | | | del | 13 | | | | 9 |
| 3 | 212,451-212,457 | 7C | 3 | *YCR048W/ARE1* | fs;176-178 | del | | | | 15 | 11 |
| 3 | 275,289 | n/a | 4 | *YCR091W/KIN82* | 297;V to I | G to A | 12 | 15 | | | |
| 4 | 512,796 | n/a | 2 | | | A to T | 8 | | 7 | | |

Continued on next page

**Table 3.3 – continued from previous page**

| | | | | | | | Distribution of Sequence reads | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Chromosome | SGD Position | HP tract | strain | Gene | aa change | Mutation | A | G | T | C | Indel |
| 4 | 814,336 | n/a | 2 | | | G to A | 10 | 6 | | | |
| 4 | 929,182-929,193 | 12 | 3 | | | del | 1 | | 5 | | 4 |
| 4 | 963,768 | n/a | 3 | *YDR252W/BTT1* | 120:G to V | T to G | | 12 | 18 | | |
| 4 | 231,908-231,914 | 7 | 4 | | | del | 12 | | | | 12 |
| 4 | 1,386,657-1,386,664 | 8 | 4 | | | del | | | 17 | | 11 |
| 4 | 470,576-470,584 | 9 | 4 | | | del | 19 | | | | 12 |
| 4 | 832,716-832,726 | 11 | 4 | | | del | 15 | | | | 7 |
| 4 | 50,592-50,603 | 12 | 4 | | | del | | | 10 | | 9 |
| 4 | 1,054,759 | n/a | 4 | | | A to T | 9 | | 11 | | |
| 5 | 305,972 | n/a | 2 | | | C to A | 5 | | | 5 | |
| 5 | 479,369 | n/a | 2 | *YER155C/BEM2* | 1159; S to C | T to A | 8 | | 3 | | |
| 5 | 225,319-225,327 | 9 | 3 | | | del | | | 9 | | 7 |
| 5 | 403,576 | n/a | 3 | *YER122C/GLO3* | 258;A to E | G to T | | 8 | 14 | | |
| 5 | 34,325-34,333 | 9 | 4 | | | del | 11 | | | | 7 |
| 5 | 402,832-402,843 | 12 | 4 | | | del | 13 | | | | 7 |
| 6 | 223,108-223,118 | 11 | 3 | | | del | 10 | | | | 7 |
| 6 | 114,200-114,210 | 11 | 4 | | | del | 6 | | | | 7 |
| 6 | 88,832 | n/a | 4 | *YFL024C/EPL1* | 504;D to E | G to T | | 15 | 13 | | |
| 6 | 225,229 | n/a | 4 | *YFR034C/PHO4* | 240;R to G | G to C | | 9 | | 8 | |
| 7 | 194,092-194,098 | 7 | 3 | *YGL163C/RAD54* | fs;771-773 | del | | | 13 | | 13 |
| 7 | 878,690-878,701 | 12 | 3 | | | del | 6 | | | | 4 |
| 7 | 653,363-653,369 | 7 | 4 | | | del | 14 | | | | 10 |
| 7 | 882,549-882,558 | 10 | 4 | | | del | 10 | | | | 6 |
| 7 | 20,017-20,027 | 11 | 4 | | | del | 10 | | | | 5 |
| 7 | 678,172-678,182 | 11 | 4 | | | del | 11 | | | | 9 |
| 8 | 150,380-150,386 | 7 | 3 | | | del | 11 | | 1 | | 12 |
| 8 | 472,612-472,624 | 13 | 3 | | | del | | | 10 | | 8 |
| 8 | 288,299 | n/a | 3 | *YHR092C/HXT4* | 172;K to K | C to T | | | 5 | 6 | 1 |
| 8 | 370,253 | n/a | 3 | *YHR132W-A/IGO2* | 46:Y to F | A to T | 15 | | 12 | 1 | 1 |
| 9 | 270,327 | n/a | 2 | *YIL046W/MET30* | 560;A to S | G to T | | 5 | 4 | | |
| 9 | 375,856 | n/a | 3 | *YIR010W/DSN1* | 143;M to I | G to A | 7 | 10 | | | |
| 9 | 199,995 | n/a | 4 | *YIL087C/AIM19* | 41;T to I | G to A | 12 | 18 | | | |
| 10 | 445,012-445,020 | 9 | 3 | | | del | | | 10 | | 6 |
| 10 | 131,051-131,059 | 9 | 4 | | | del | | | 11 | | 15 |
| 10 | 469,684-469,694 | 11 | 4 | | | del | 19 | | | | 6 |
| 11 | 162,688-162,695 | 8 | 4 | | | del | | | 9 | 1 | 7 |
| 11 | 403,466 | n/a | 4 | *YKL018C-A* | 19;S to S | C to T | | | 13 | 13 | |
| 12 | 405,712-405,719 | 8 | 2 | *YLR131C/ACE2* | fs;369-371 | del | | | 4 | | 3 |
| 12 | 32,320-32,330 | 11 | 3 | | | del | 9 | | | | 5 |
| 12 | 964,065 | n/a | 3 | *YLR420W/URA4* | 95:R to H | G to A | 11 | 12 | | | |
| 12 | 1,009,007 | n/a | 3 | *YLR436C/ECM30* | 746;I to V | T to C | | | 12 | 12 | |

Continued on next page

**Table 3.3 – continued from previous page**

| | | | | | | | Distribution of Sequence reads | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Chromosome | SGD Position | HP tract | strain | Gene | aa change | Mutation | A | G | T | C | Indel |
| 12 | 363,531-363,537 | 7 | 4 | *YLR106C/MDN1* | fs;68-70 | del | 11 | | | | 13 |
| 12 | 201,846-201,856 | 11 | 4 | | | del | 14 | | | | 9 |
| 12 | 1,047,741 | n/a | 4 | *YLR454W/FMP27* | 1249:D to G | A to G | 10 | 12 | | | 1 |
| 13 | 763,010-763,016 | 7 | 2 | *SNR86* | (small nucleolar RNA) | ins | 5 | | | | 5 |
| 13 | 241,855-241,867 | 13 | 3 | | | del | | | 12 | | 6 |
| 13 | 311,843 | n/a | 3 | | | C to T | | | 15 | 25 | |
| 13 | 139,705-139,709 | 5 | 4 | *YML067C/ERV41* | fs;138-139 | del | | | 15 | | 13 |
| 13 | 816,457-816,463 | 7 | 4 | *YMR275C/BUL1* | fs;706-708 | del | | | 11 | | 9 |
| 14 | 761,792 | n/a | 2 | *YNR069C/BSC5* | 267;V to V | C to T | | | 8 | 6 | 1 |
| 14 | 222,733 | n/a | 3 | *YNL225C/CMN67* | 580;V to M | C to T | 1 | | 17 | 10 | |
| 14 | 435,595-435,601 | 7 | 4 | *YNL101W/AVT4* | fs;199-201 | del | 9 | | | | 10 |
| 14 | 481,123-481,129 | 7 | 4 | | | del | 15 | | | | 8 |
| 14 | 685,574-685,582 | 9 | 4 | | | del | | | 12 | 1 | 9 |
| 14 | 575,616-575,626 | 11 | 4 | | | del | 12 | | | | 8 |
| 14 | 400,002 | n/a | 4 | *YNL121C/TOM70* | 180;G to STOP | C to A | 15 | | | 15 | |
| 14 | 734,521 | n/a | 4 | *YNR058W/BIO3* | 77;L to L | A to G | 11 | 10 | | | |
| 15 | 854,146-854,153 | 8 | 2 | | | del | 4 | | | | 5 |
| 15 | 874,052-874,057 | 6 | 3 | *YOR296W* | fs;1284-1286 | del | 10 | | | | 7 |
| 15 | 767,667-767,673 | 7 | 3 | *YOR228C* | fs;36-38 | del | | | 10 | | 7 |
| 15 | 822,829-822,835 | 7 | 3 | *YOR267C/HRK1* | fs;678-680 | del | | | 11 | | 8 |
| 16 | 146,421-146,427 | 7 | 2 | *YPL216W* | fs;868-870 | del | 7 | | | | 5 |
| 16 | 22,677 | n/a | 4 | | | C to T | | | 20 | 14 | |
| 16 | 131,583 | n/a | 4 | *YPL222W/FMP40* | 475;A to T | G to A | 11 | 15 | 1 | | |
| 16 | 509,632 | n/a | 4 | *YPL022W/RAD1* | 980;A to S | G to T | | 19 | 18 | 1 | |
| 16 | 570,131 | n/a | 4 | *YPR007C/REC8* | 415;S to M | G to A | 11 | 11 | | | |

Table 3.3: The type of mutation (base substitution, single nt insertion (ins), single nt deletion (del)) is shown, as well as the length of the HP tract that contains an indel. The specific Mut line (2, 3, or 4) is indicated under "strain". All HP tracts were polyA or polyT except for the mutation in Chromosome 3 at 212,451-212,457, which involved a polyC tract. For mutations that occurred within an open reading frame, both the gene name and predicted amino acid (aa) changes (fs; frameshift) are provided. n/a; not applicable. Coordinates are presented as shown in the SGD (http://www.yeastgenome.org/). The number and distribution of the sequence reads are presented for each mutation. The frameshift mutation in *YLR106C/MDN1* conferred a recessive lethal phenotype (data not shown).

Due to the constraints of using 36 nt Illumina GA reads, we do not have the power to detect mutations in HP tracts larger than 13 nt, but less than 400 such

tracts are present in the yeast genome. Visual inspection of the DNA sequences surrounding the indel mutations ( 400 bp; Figure 3.2) suggested that they were enriched for HP runs. These are primarily poly (dA:dT) tracts that are present in the yeast genome at a 20-fold higher frequency than poly (dG:dC) tracts. Consistent with this, the AT content of the genomic regions surrounding the indel mutations was significantly higher than for unmutated HP regions (windows up to 500 bp; data not shown). Detailed bioinformatic and genetic analysis will be required to determine if this pattern is significant; however, a previous study [33] showed that DNA polymerase slippage was not greatly influenced by sequence context, including nearby HP tracts.

Our analysis permitted the detection of up to two single nt indels in a 36 nt reads; these indels can be right next to each other to create a two nt indel or separated from each other. We assigned this limit because creating high quality and unique alignments became very difficult when allowing indels larger than two nt. We were unable to detect indels of two nt in any of the lines. Such a result is not surprising based on previous studies of wild-type and MMR mutants analyzed for reversion of frameshift mutations in HP runs. In these studies the overwhelming majority of mutations involved single nt deletions. For example TRAN et al. ([93]) found that 225 of 227 reversions in +1 HP tracts in wild type, polymerase proofreading, and mismatch repair mutants were due to deletions of a single nt. For -1 HP tracts, they found that 206 of 218 reversions were due to additions of a single nt. The remaining revertants in both HP tracts involved expansions or contractions of no greater than two nt in size.

The predominance of single nt deletions over single nt insertions and base substitutions was similar to previous reports for the mutational spectra in re-

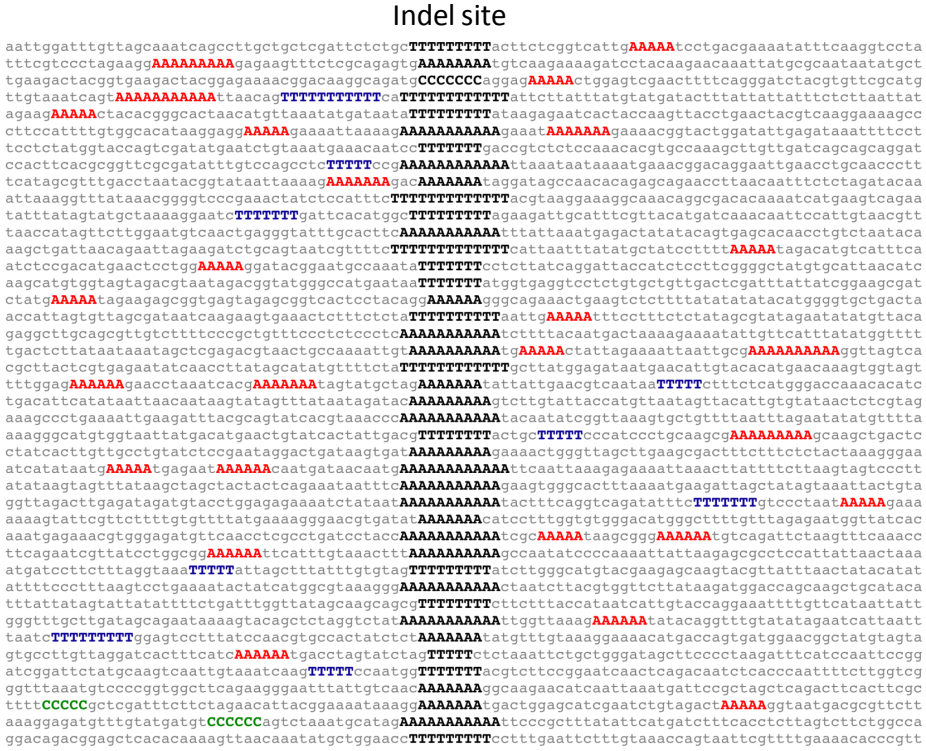



Figure 3.2: The 100-bp region surrounding indel mutations in the Mut3 and Mut4 lines. The locations of the indel mutations are indicated in black boldface type. HP runs of ≥ 5 in this window are color coded as shown: red, $A_n$; blue, $T_n$; green, $C_n$

porter genes in MMR null mutants [93, 94, 66, 21]. The average mutation rate in the 5 to 13 bp HP tracts was $1.0 \times 10^{-6}$ per HP tract per generation (Table 3.4). The rate was an order of magnitude greater ($1.1 \times 10^{-5}$) if only runs between 8 to 13 base pairs long were considered (Table 3.4). These values approach the rates seen in MMR-defective yeast (*mlh1*, *msh2*) containing reporters bearing 10 bp polyT ($2.8 \times 10^{-4}$; [93]) and 10 bp polyA ($7.3 \times 10^{-5}$; [29]) tracts. Low sequence coverage provides one explanation for why the rate is lower than seen previously in reporter assays. In our analysis, indels in HP tracts can be identified only if the entire tract and sequence flanking both sides are present in a 36 nt

read; the longer the HP tract, the less likely it is to obtain reads that cover the entire tract. Thus higher sequence coverages are required to identify indels in HP tracts. Consistent with this, a higher indel mutation rate was seen in lines that had higher sequencing coverage (Table 3.4). In contrast, SNPs that occur outside of an HP tract should not be as affected by sequence coverage (aside from the relationship between coverage and probability of detecting sufficient copies of the alternate base in order to reliably make a call). This was seen for the analysis of base substitutions (Table 3.4).

The average rate of base substitution mutations was $3.7 \times 10^{-9}$ mutations per base per generation (Table 3.4), which is eleven-fold higher than the base substitution rate observed in wild-type haploid strains [65]. Of the 28 base substitution mutations detected in the Mut2-4 lines, sixteen were transitions and twelve were transversions (Table 2.3). Nineteen of these mutations resulted in a change from a G-C to an A-T base pair, whereas only four were in the opposite direction. This overall mutational bias towards A-T base pairs was seen and discussed previously (e.g. [65, 46, 19]). The modest increase that we observed in the base substitution rate in MMR defective strains is significantly lower than predicted ($\sim$ 100 fold increase for base substitutions and frameshifts; [43, 21]. We suggest two reasons for these differences. First, our measurements were determined from a genome-wide measurement rather than by extrapolation from a few marker loci. Second, the *mlh1*-7$^{ts}$ allele is not a complete null mutation. It phenocopies the *mlh1*-7$^{st}$ phenotype in the *CAN1* mutational assay, but has a four-fold lower mutation rate than mlh1 in the *lys2*-$A_{14}$ reversion assay ([36]; data not shown). Because *mlh1*-7$^{ts}$ strains display residual DNA repair, it is possible that there is a bias towards the repair of specific mismatches in these strains. While we cannot rule this out, the fact that the mutation signature seen

| Base Substitution Mutations | | | | |
|---|---|---|---|---|
| Strain | No.mutations | %genome≥ 7X coverage | Genome Size (bp) adjusted | Mutation rate (per base per gen $\times 10^{-9}$) |
| Mut2 | 6 | 41 | 9,898.136 | 3.8 |
| Mut3 | 9 | 69 | 16,657838 | 3.4 |
| Mut4 | 13 | 84 | 20,279,107 | 4.0 |
| Average | | | | 3.7 |
| **Single-nucleotide indel mutations in 5- to 13-nt HP tracts** | | | | |
| Strain | No.mutations | No. HP tracts ≥ 7X covergae | | Mutation rate (per base per gen $\times 10^{-9}$) |
| Mut2 | 6 | 57,502 | | 5.5 |
| Mut3 | 15 | 99,714 | | 9.4 |
| Mut4 | 27 | 122.816 | | 14 |
| Average | | | | 10 |
| **Single-nucleotide indel mutations in 8- to 13-nt HP tracts** | | | | |
| Strain | No.mutations | No. HP tracts ≥ 7X coverage | | Mutation rate (per base per gen $\times 10^{-9}$) |
| Mut2 | 4 | 2,820 | | 89 |
| Mut3 | 10 | 7,054 | | 89 |
| Mut4 | 19 | 8,696 | | 140 |
| Average | | | | 110 |

Table 3.4: Mutation rates for Mut2, Mut3 and Mut4 lines grown in bottlenecks for 160 generation: The base substitution mutation rate was determined by calculating the percentage of the genome in which at least sevenfold DNA sequencing coverage to unique regions was obtained. This was done because our statistical analysis did not have sufficient power to reliably detect heterozygous mutations in regions with lower coverage. This information was used to calculate the mutation rate on the basis of the following formula: (number of mutations)/(160 generations)/(adjusted genome size) with the diploid *S.cerevisiae* genome size determined as 24,141,794 bp(http:/www.yeastgenome.org/). To obtain indel mutation rates, we first determined the number of HP tracts of a given length in unique regions of the genome which had ≥ sevenfold sequence coverage. We then used the following equation to calculate mutation rate:(number of indels)/(160 generations)/(number of HP tracts with ≥ sevenfold coverage)

in *mlh1*-7$^{ts}$ appeared indistinguishable from mlh1 null strains argues against such a possibility [93, 94, 66]. Lastly, we cannot rule out the possibility that mutation rates in MMR-defective strains are different in haploid vs. diploid yeast, though a recent analysis of mutation rates in diploid bottleneck lines showed

that wild-type diploid yeast displayed an estimated base substitution rate that was very similar to that reported previously for haploid yeast ([65], Nishant, Wei, Mancera, *et al*., unpublished data).

Because the three lines showed viability that ranged from 2.5 to 15.6%, we expected to identify mutations that conferred a lethal phenotype. We examined whether any of the mutations that mapped to open reading frames in the Mut4 line (2.5% viability) were not detected in haploid progeny. This was done by sequencing DNA surrounding a particular mutation in twenty viable spore clones obtained by sporulating the Mut4 generation 160 line. Of these fourteen mutations, only the frameshift mutation in *MDN1* was not detected, consistent with previous work showing that *mdn1*Δ mutants are inviable [28]. While it is unclear how many mutations would confer lethality in the absence of other mutations, the assortment of five independent lethal mutations would result in 3% spore viability, similar to that seen in the Mut4 line. We hypothesize that other lethal mutations were not identified in Mut4 and other lines because: 1. A large number of frameshift mutations in HP tracts may not have been detected because indels can be identified only if the entire tract and sequence flanking both sides are present in a 36 nt read. Identifying indels in HP tracts is very challenging using short read sequencing. However, increasing sequence coverage and using paired-end reads of a larger size (~180 bp) should provide a good test of this idea. 2. Our sequence analysis did not cover the entire genome (84% for Mut4). 3. While previous CGH and PFGE analysis (~ 1 KB resolution; [36]) did not reveal rearrangements, it is possible that mutations occurred that involved indels larger than two nt and smaller than 1 KB. However, we found this to be less likely because a previous analysis of mutation spectra in MMR mutants indicate that indels greater than two nt are extremely rare [93].

In the *S. cerevisiae* S288c haploid genome there are over 77,425 HP tracts five nt or greater. Frameshift mutations in coding regions that disrupt protein function are likely to have significant effects on organism fitness. In wild-type yeast, insertion/deletion mutations appear to be relatively rare compared to base substitutions; comparative analysis of multiple domestic and wild yeast strains identified $\sim 14,000$ indels compared to $\sim 235,000$ SNPs [102, 59]. In contrast, MMR mutants display a strong bias towards frameshifts over base substitutions in the genome. Thus our data, together with previous work, illustrate the critical role that MMR plays in preventing frameshifts in HP tracts across the genome.

CHAPTER 4

# MUTATIONAL HOTSPOTS IN THE YEAST GENOME ARE CAUSED BY LONG-RANGE CLUSTERING OF $A_N$ AND $T_N$ HOMOPOLYMERIC RUNS

This project was carried out as a continuous study from Chapter 3. The content that described within this chapter came from a manuscript in preparation with both myself and Maria Rogacheva as primary authors. I sincerely appreciate the outstanding experimental work that Maria Rogacheva did for this project, and I am also thankful for the contributions from all of the other co-authors including K. T. Nishant, Sarah Zanders, Carlos D. Bustamante and Eric Alani.

## 4.1 Introduction

Mutations arising from cellular metabolism and environmental insults confer fitness defects that are either removed by natural selection, drift neutrally in the population, or provide the raw fuel of adaptive evolution (reviewed in [77]). A corner stone of classical evolutionary theory is that mutations occur randomly throughout the genome and that biases in mutation contribute little to the ultimate outcome of the evolutionary process. However, experiments performed over many years suggest that not all sites in the genome have an equal probability of acquiring a mutation (reviewed in [104]). Both experimental and indirect methods have been used to infer mutation rates (reviewed in [7, 77]). Through such work mutation rates have been shown to vary with respect to base composition, local recombination rate, gene density, transcription, nucleosome location, and replication timing [34, 103, 67, 5, 32, 17, 92, 100, 88]. In addition, studies have suggested that larger genomic contexts exist that can af-

fect mutation patterns but specific sequences within such contexts have not been identified [8, 34]. For example, Bailey *et al*. [8] obtained evidence for hotspots in mammalian chromosomal evolution by observing conserved chromosome breakpoints and argue against a random-breakage model for chromosome evolution [26]. Understanding the molecular mechanisms that lead to mutation variation is a major challenge that is likely to provide insights into human disease progression (e.g., mutation accumulation in cancer tumors) and molecular evolution.

A goal of the current work is to determine whether broad DNA sequence contexts underlie variability in mutagenesis across the genome. To test for such a context, we focused on identifying mutations that arise during DNA replication. The rate of such errors is low, ranging from $3 \times 10^{-10}$ to $2 \times 10^{-8}$ mutations per base pair per generation (reviewed in [77]). To accelerate the accumulation of mutations in a population, we employed conditional mismatch repair (MMR) mutants. MMR is a highly conserved pathway that excises DNA replication errors arising primarily from polymerase misincorporation and slippage events [93, 20, 21, 29, 89, 85]. In eukaryotes, two heterodimeric MutS homolog complexes, MSH2-MSH3 and MSH2-MSH6, act in mismatch recognition. Both MSH complexes interact primarily with MLH1-PMS1 to form a mismatch-MSH-MLH complex that activates downstream repair steps including strand discrimination, excision, and resynthesis [48].

In contrast of the previous study describe in Chapter 3 which are embedded with limitation due to short length of the single-end reads, in this study we used a Likelihood Ratio Test along and paired-end sequencing technologies with longer reads to identify mutations that accumulated in MMR-deficient lines of

baker's yeast. We identified novel mutational hotspots in the genome as determined by their identification in independent lines. Importantly, we identified broad sequence contexts that contribute to mutational hotspots: the likelihood of a mutation in a given poly(dA:dT) homopolymeric (HP) tract is increased by the presence of poly (dA:dT) tracts in a 1000 bp region centered on the given tract. Such work is of interest because the presence of mutational hotspots is expected to contribute disproportionately to the genetic variation available to natural selection and to causative mutations in genetic diseases.

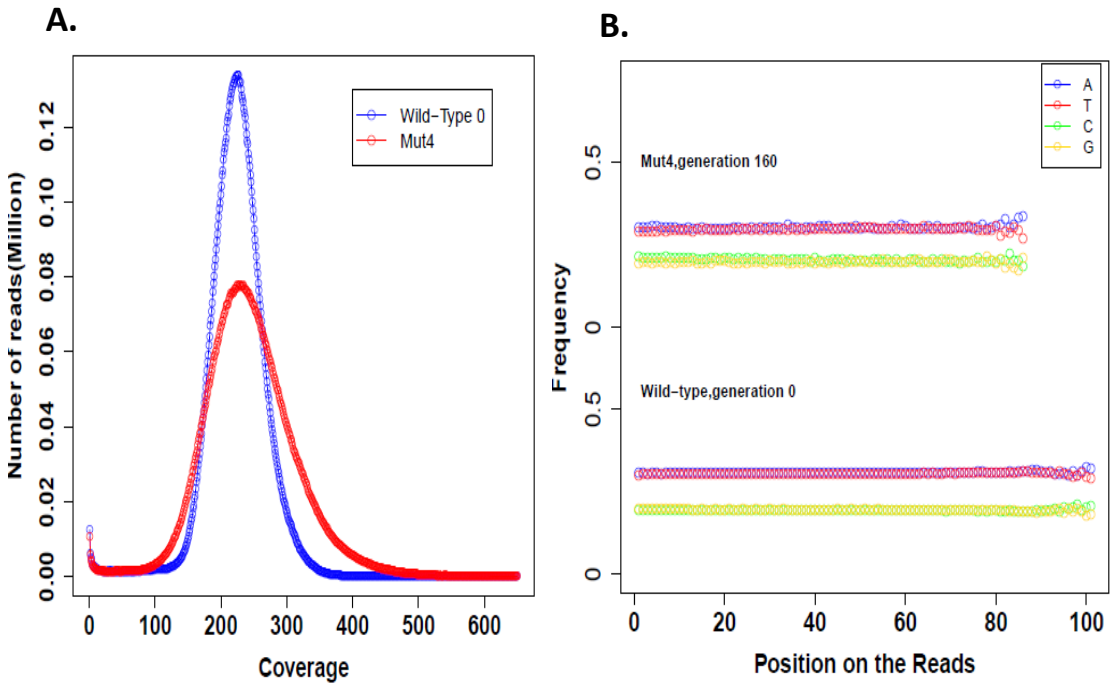



Figure 4.1: Statistics for pair-end reads data: A. Frequency plot showing Mut4 genome coverage after aligning all Mut4 pair-end reads to the S288c reference genome; B. Frequency plot of nucleotide distribution at each position along the DNA sequence reads of wild-type generation 0 and Mut4 generation 160

## 4.2 Whole-genome sequencing analysis of Mut lines and variation detection using Loglikelihood Ratio Test

Wild-type at generation 0 and the Mut4 line at generation 160 were paired-end sequenced (spanning 300 to 400 bp) to 220- and 240-fold coverage, respectively, using an Illumina Genome Analyzer. Read lengths were 101 nt for wild-type and 86 nt for Mut4 (Figure 4.1A). The AT/GC ratio was similar across the read sequence, with some fluctuation seen at the end of the sequence, which is most likely due to the presence of adapters. This information is supportive of high quality sequence data (Figure 4.1B).

Table 4.1: Counts and mapping percentage statistics for pair-end read analysis before and after alignment

| | Wild-type Generation 0 | *Mut4* Generation 160 |
|---|---|---|
| Raw Read | 31,952,368 | 42,572,646 |
| Unpaired Reads | 177,108 | 187,455 |
| Read Pairs | 15,301,769 | 19,178,204 |
| Unmapped Reads | 177,185 | 187,674 |
| Mapping (%) | 96.33 | 90.54 |
| Unpaired Read Duplications | 146,690 | 157,184 |
| Paired Read Duplications | 1,534,481 | 778,034 |
| Duplication (%) | 0.10 | 0.04 |
| Mean Coverage | 224.1 | 244.0 |
| SD of Coverage | 44.4 | 69.9 |

To detect mutations, all paired-end reads were mapped against the S288c

reference genome using NOVOALIGN (http://www.novocraft.com), whose paired-end mode shows better sensitivity and specificity compared to other aligners [50]. Our analysis permitted the detection of up to two single nt in-dels; these in-dels can be right next to each other to create a two nt in-del or separated from each other. Those reads that contained too many mismatches, mapped to multiple genome positions, or had identical start and end coordinates (PCR errors) were eliminated to minimize the false positive rate for calling SNPs. The mapping statistics are shown in Table 4.1 and 4.2. We mapped mutations in 92% of the yeast genome. The remaining 8% consisted of repetitive DNA (e.g. regions in Chromosome IX (~20 KB) and Chromosome X), DNA sequenced at low coverage, and difficult to amplify GC-rich sequences.

Table 4.2: Total length with sufficient coverage (mean coverage ± 3 × SD for each chromosome)

| Chromosome Name | Total reference length | Total length with sufficient coverage | Percentage |
|---|---|---|---|
| Chr1 | 230208 | 198537 | 0.86 |
| Chr2 | 813178 | 768333 | 0.94 |
| Chr3 | 316617 | 284181 | 0.89 |
| Chr4 | 1531919 | 1420463 | 0.93 |
| Chr5 | 576869 | 543476 | 0.94 |
| Chr6 | 270148 | 248560 | 0.92 |
| Chr7 | 1090947 | 1019541 | 0.93 |
| Chr8 | 562643 | 508976 | 0.90 |
| Chr9 | 439885 | 406441 | 0.92 |
| Chr10 | 745742 | 681812 | 0.91 |
| Chr11 | 666454 | 648807 | 0.97 |
| Chr12 | 1078175 | 960619 | 0.89 |
| Chr13 | 924429 | 868987 | 0.94 |
| Chr14 | 784333 | 742785 | 0.95 |
| Chr15 | 1091289 | 1029697 | 0.94 |
| Chr16 | 948062 | 860512 | 0.91 |
| **Total** | **12070898** | **11106178** | **0.92** |

By looking at k-mer sizes (Table 4.3), we were able to identify opti-

mal read assemblies for each line-those with the largest N50 generated with Velvet software [107]. We further sorted assembly contigs according to their genomic position after aligning them onto the S288c reference genome. We visualized all sorted contigs of wild-type and Mut line with Mauve (http://gel.ahabs.wisc.edu/mauve; Figure 4.2,4.3 and 4.4). Our data indicate that the wild-type and Mut4 lines did not undergo major genome rearrangements, consistent with previous comparative genome hybridization and pulse-field gel electrophoresis analysis [37].

Figure 4.2: Frequency plot of the length distribution of the assembled contigs generate by software Velvet based on resequencing pair-end reads for wild-type generation 0 and Mut4 generation 160

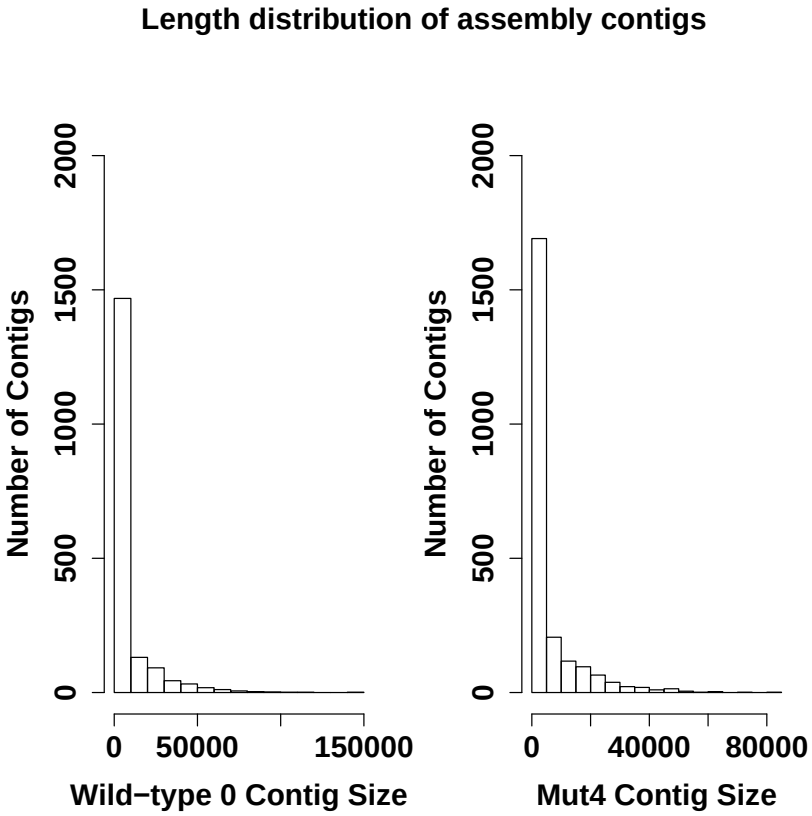


Table 4.3: Assembly information: k-mer size and N50 statistics (* indicates the best assembly that used to create a genome assembly to search for large genome rearrangements)

Wild-type generation 0 assembly information (total reads= 31,952,368)

| K-mer Size | Nodes | N50 | Max contig size | Total contig length | Reads used |
|---|---|---|---|---|---|
| 61 | 4352 | 19443 | 71479 | 11507459 | 30351625 |
| 63 | 4344 | 19794 | 71477 | 11512221 | 30210624 |
| 65 | 4253 | 20006 | 71475 | 11515985 | 30163516 |
| 67 | 4223 | 20972 | 71473 | 11523871 | 30021693 |
| 69 | 4134 | 21223 | 72852 | 11525043 | 29864178 |
| 71 | 4110 | 21489 | 72850 | 11529371 | 29771283 |
| 73 | 3999 | 22355 | 72848 | 11534338 | 29653047 |
| 75 | 3833 | 23524 | 112684 | 11537812 | 29536615 |
| 77 | 3765 | 24131 | 112687 | 11540721 | 29360772 |
| 79 | 3653 | 26620 | 112687 | 11549872 | 29271486 |
| 81 | 3563 | 27799 | 112687 | 11552278 | 29125914 |
| 83 | 3455 | 28883 | 118630 | 11553216 | 29000435 |
| 85* | 3421 | 30253 | 146630 | 11557688 | 28850310 |
| 87 | 3317 | 30242 | 146631 | 11558810 | 28670731 |

Mut4 generation 160 assembly information (total reads= 31,952,368)

| K-mer Size | Nodes | N50 | Max contig size | Total contig length | Reads used |
|---|---|---|---|---|---|
| 61 | 4617 | 15538 | 61292 | 11549082 | 38757107 |
| 63 | 4524 | 16136 | 61292 | 11555900 | 38458275 |
| 65 | 4424 | 16700 | 72502 | 11561885 | 38254451 |
| 67 | 4368 | 17996 | 72502 | 11569829 | 37988609 |
| 69 | 4217 | 18248 | 71480 | 11574427 | 37695869 |
| 71 | 4184 | 19124 | 70055 | 11580436 | 37433090 |
| 73* | 4042 | 19590 | 82618 | 11581308 | 37004284 |
| 75 | 4009 | 19415 | 82618 | 11585715 | 36726837 |
| 77 | 3969 | 18620 | 94518 | 11586272 | 36406509 |
| 79 | 4129 | 16846 | 94444 | 11592198 | 36050620 |
| 81 | 4923 | 8896 | 49264 | 11553665 | 35679229 |
| 83 | 13897 | 1298 | 8845 | 10932964 | 34234849 |

Because Mut bottleneck lines were derived from a common ancestor [37], we expected to see identify propagated mutations in which wild-type differed from the S288c reference genome, and derived heterozygous mutations resulting from defects in DNA mismatch repair. In order to detect derived mutations

Figure 4.3: Multiple contig alignment plot (Mauve Aligner) of wild-type generation 0 and Mut4 generation 160 with the reference S288c genome of genomic location between 6.07M to 6.24M after concatenating the 16 chromosomes.The alignment is organized into one horizontal "panel" per input genome sequence. Each genome panel contains the name of the genome sequence (i.e., the name of the fasta files), a scale showing the sequence coordinates for that genome, and a single black horizontal center line. Blocks that lie above the center line represent the aligned region that is in the forward orientation relative to the first genome sequence. Blocks below the center line indicate regions that align in the reverse complement(inverse) orientation. Each of colored blocks above and possibly below the center line surround a region of the genome sequence that aligned to part of another genome, and is presumably homologous and internally free from genomic rearrangement. A genomic rearrangement would show a cross of lines between a pair of contigs from a pair of genomes. Regions outside blocks lack detectable homology among the input genomes. Inside each block Mauve draws a similarity profile of the genome sequence. The height of the similarity profile corresponds to the average level of conservation in that region of the genome sequence. Areas completely white were not aligned and probably contain sequence elements specific to a particular genome. The height of the similarity profile is calculated to be inversely proportional to the average alignment column entropy over a region of the alignment

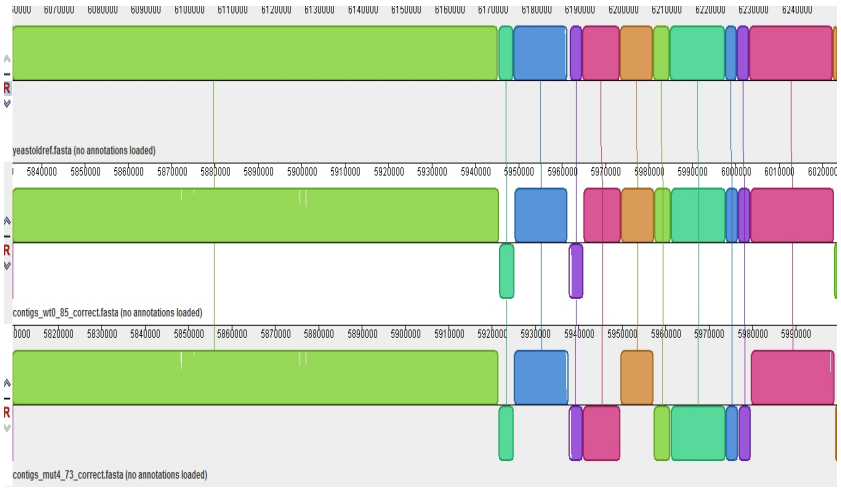
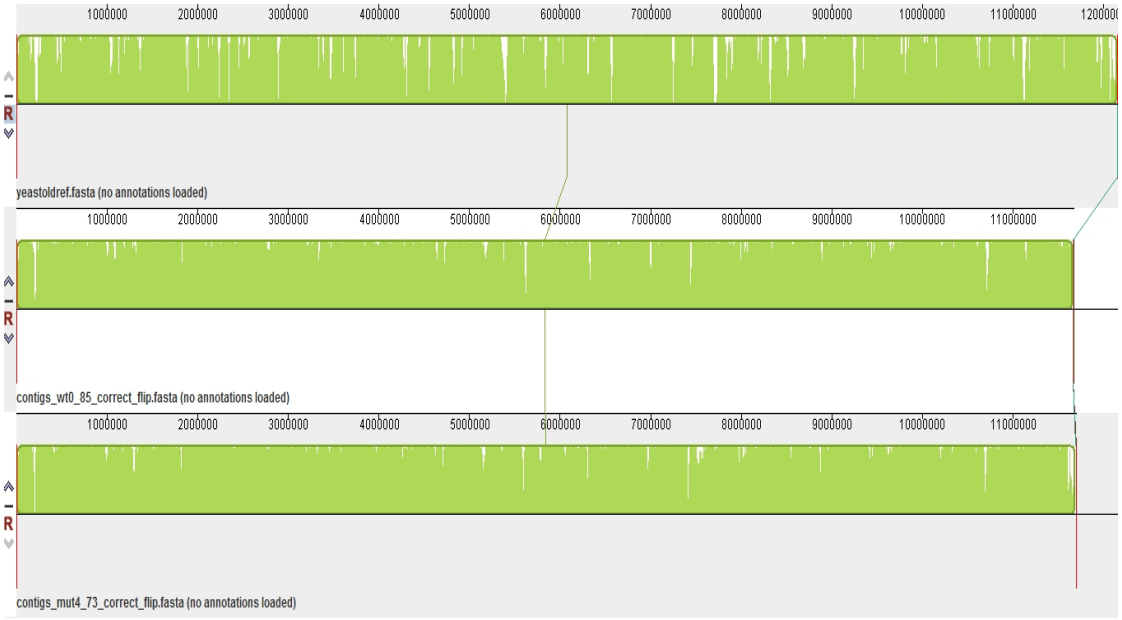


Figure 4.4: Multiple contig alignment plot (Mauve Aligner) of wild-type generation 0 and Mut4 generation 160 with the reference S288c genome for the whole genomic region (total of 12M bp. Contigs in the forward orientation of strain wild-type generation 0 and Mut4 generation 160 were kept unchanged and the region in the reverse complement orientation was reversed and complemented so that it also has the forward orientation compared with the reference S288c genome. As shown we do not detect any genome rearrangements




that arose in the Mut4 line, we developed a modified two-step log likelihood ratio test described below. This test does not bias classification by imposing a prior, yet is able to separate propagated from derived mutations.

At any given locus with sufficient coverage, we first tested whether a SNP found in this locus is a propagated mutation by pooling loci information from the wild-type-0 and Mut4-160 lines. Let $N_{w+m}$ equal the total coverage for wild-

type-0 plus Mut4-160 at this locus. Let A and a be two segregating allele types, and let $X_i$ be the true allele type we should observe for this site at the $i^{th}$ read. $Y_i$ is defined as the observed allele type from the $i_{th}$ short sequence data where $i = 1, \cdot, N_{w+m}$ , and $e_i$ is defined as the corresponding observed error probability.

Therefore, we have that:

$$\mathbb{P}(X_i = A) = p \tag{4.1}$$

$$\mathbb{P}(X_i = a) = 1 - p \tag{4.2}$$

$$\mathbb{P}(Y_i = A|X_i = A) = 1 - e_i \tag{4.3}$$

$$\mathbb{P}(Y_i = A|X_i = a) = e_i \tag{4.4}$$

$$\mathbb{P}(Y_i = a|X_i = A) = e_i \tag{4.5}$$

$$\mathbb{P}(Y_i = a|X_i = a) = 1 - e_i \tag{4.6}$$

Using the Bayes' rule we can get:

$$\begin{aligned} \mathbb{P}(Y_i = A) &= \mathbb{P}(Y_i = A|X_i = A) \times \mathbb{P}(X_i = A) + \mathbb{P}(Y_i = A|X_i = a) \times \mathbb{P}(X_i = a) && (4.7)\\ &= (1 - e_i) \times p + e_i \times (1 - p) && (4.8)\\ \mathbb{P}(Y_i = a) &= \mathbb{P}(Y_i = a|X_i = A) \times \mathbb{P}(X_i = A) + \mathbb{P}(Y_i = a|X_i = a) \times \mathbb{P}(X_i = a) && (4.9)\\ &= e_i \times p + (1 - e_i) \times (1 - p) && (4.10) \end{aligned}$$

Where

$$Y_i \sim Bernoulli(1, (1 - e_i) \times p + e_i \times (1 - p)) \tag{4.11}$$

From all above, we can get the likelihood as:

$$\begin{aligned} \mathbb{L} &= \prod_{i=1}^{N_{w+m}} \{(1 - e_i) \times p + e_i \times (1 - p)^{I_i}\}\{e_i \times p + (1 - e_i) \times (1 - p)^{I_i}\} && (4.12)\\ &\quad \text{where } I_i = 1 \text{ if } Y_i = \text{A; and } I_i = 0 \text{ if } Y_i = \text{a} && (4.13) \end{aligned}$$

The LRT is performed with hypothesis of $H_o : p = 0.5$ versus $H_a : p \neq 0.5$

If a specific site was classified as a propagated mutation, the analysis was discontinued. Otherwise, we carried out a second step where we used the same method but only applied it with the data from Mut4 line to determine the presence of a Mut4-specific derived mutation. We further grouped all potential derived mutations in Mut4 line by confidence level. The high confident group displayed P-values less than 0.1; for the low confidence group P-values were chosen to be greater than 0.15 but less than 0.3. No sites were observed with P-values between 0.1 and 0.15.

Given the very high sequencing coverage, we believe we can ignore the estimated false positive rate; this is because we can easily distinguish true heterozygous mutations from low frequency sequencing errors. We also believe that we can ignore the false negative rate in regions with sufficient coverage because we employed a very conservative P-value during the classification step. As a result we expected to identify a comprehensive list of mutations spanning 92% of the Mut4 genome.

## 4.3 Confirming DNA sequence heterozygosity in the Mut4 line by Sanger Sequencing

39 heterozygous mutations were previously found in the diploid Mut4 line using single-end 36 nt sequencing [106]. The new paired-end sequencing of the Mut4 line allowed us to detect 53 new high confidence mutations (Table 4.4). We randomly chose 12 of the new calls to be confirmed by Sanger sequencing

(performed at the Cornell CLC using an Applied Biosystems Automated 3730 DNA analyzer) using PCR amplification methods described in Zanders *et al*. [106]; all 12 high confidence predictions were verified. As described in [106], a heterozygous base substitution was confirmed if a doublet representing both alleles were observed in the sequencing trace of the Mut4 line but only a singlet peak representing the parental allele was seen in the wild-type line. A heterozygous in-del mutation was confirmed if the sequencing reaction failed (i.e., tall singlet peaks become small doublet peaks or to random noise) at the predicted location in the Mut4 line, but the sequencing reactions in a wild-type line could successfully sequence past the site. Of the 17 low confidence alleles identified, all were analyzed by Sanger sequencing, but only three were verified and included in Table 4.4. The false negative rate here is negligible because we validated the entire low confidence group and found that most of these predictions were false positives. Overall, the high rate of agreement between confidence levels predicted from the model and the Sanger sequence validation suggests that our method has a high level of predictive accuracy.

Table 4.4: Genome location of mutations detected in the *Mut4* line

| Chromosome | SGD Position | HP tract | Gene | aa change | Mutation |
|---|---|---|---|---|---|
| 1 | 17,930-17,933 | 4 | | | del* |
| 1 | 139,349-139,358 | 10 | | | del |
| 2 | 487,049-487,059 | 11 | | | del* |
| 2 | 662,560-662,569 | 10 | *YBR219C* | fs;103-106 | del |
| 2 | 653,035-653,045 | 11 | | | del |
| 3 | 22,747-22,751 | 5 | *YCL059C/KRR1* | fs;210-211 | del* |
| 3 | 94,664 | n/a | *YCL016C/DCC1* | 367:R to H | G to A* |

Continued on next page

**Table 4.4 – continued from previous page**

| Chromosome | SGD Position | HP tract | Gene | aa change | Mutation |
|---|---|---|---|---|---|
| 3 | 275,289 | n/a | *YCR091W/KIN82* | 297;V to I | G to A |
| 4 | 50,592-50,603 | 12 | | | del |
| 4 | 157,979-157,990 | 12 | | | del* |
| 4 | 216,494-216,503 | 10 | | | del* |
| 4 | 231,908-231,914 | 7 | | | del |
| 4 | 283,286-283,297 | 12 | | | del* |
| 4 | 314,305-314,316 | 12 | *YDL079C/MRK1* | fs;145-148 | del* |
| 4 | 424,586 | n/a | *YDL017W/CDC7* | 126;E to D | G to T* |
| 4 | 470,576-470,584 | 9 | | | del |
| 4 | 525,614-525,623 | 10 | *YDR037W/KRS1* | fs:59-61 | del* |
| 4 | 567,580-567,590 | 11 | | | del* |
| 4 | 688,006-688,017 | 12 | | | del* |
| 4 | 832,716-832,726 | 11 | | | del |
| 4 | 838,276-838,285 | 10 | | | del* |
| 4 | 894,994-895,005 | 12 | | | ins* |
| 4 | 900,854-900,864 | 11 | *YDR217C/RAD9* | fs;873-876 | del* |
| 4 | 1,054,759 | n/a | | | A to T |
| 4 | 1,386,657-1,386,664 | 8 | | | del |
| 4 | 1,428,260-1,428,265 | 6 | | | del* |
| 4 | 1,487,385-1,487,396 | 12 | | | del* |
| 4 | 1,521,819-1,521,829 | 11 | | | ins* |
| 5 | 34,325-34,333 | 9 | | | del |
| 5 | 162,568-162,577 | 10 | | | del* |
| 5 | 189,385-189,396 | $(AT)_6$ | | | di-ins* |
| 5 | 402,832-402,843 | 12 | | | del |
| 5 | 528,697-528,703 | 7 | *YER171W/RAD3* | fs;539-541 | del* |
| 6 | 88,832 | n/a | *YFL024C/EPL1* | 504;D to E | G to T |
| 6 | 114,200-114,210 | 11 | | | del |
| 6 | 225,229 | n/a | *YFR034C/PHO4* | 240;R to G | G to C |
| 7 | 20,017-20,027 | 11 | | | del |
| 7 | 145,563-145,570 | 8 | | | ins* |
| 7 | 172,140-172,147 | 8 | *YGL176C* | fs;311-314 | del* |
| 7 | 185,969-185,978 | 10 | | | del* |
| 7 | 533,997-534,006 | 10 | | | ins* |
| 7 | 653,363-653,369 | 7 | | | del |
| 7 | 678,172-678,182 | 11 | | | del |
| 7 | 805,051-805,062 | 12 | *YGR157W/CHO2* | two nt after stop | del* |
| 7 | 808,486-808,494 | 9 | | | del* |
| 7 | 882,549-882,558 | 10 | | | del |
| 8 | 371619-371628 | 10 | | | del* |
| 8 | 402,061-402070 | 10 | | | del* |
| 9 | 197,677-197,686 | 10 | | | del* |

Continued on next page

**Table 4.4 – continued from previous page**

| Chromosome | SGD Position | HP tract | Gene | aa change | Mutation |
|---|---|---|---|---|---|
| 9 | 199,995 | n/a | *YIL087C/AIM19* | 41;T to I | G to A |
| 9 | 321,265-321,269 | 5 | | | del* |
| 10 | 131,051-131,059 | 9 | | | del |
| 10 | 460,308-460,316 | 9 | *YJR012C* | fs;35-37 | ins* |
| 10 | 469,684-469,694 | 11 | | | del |
| 10 | 657,120 | n/a | *YJR126C/VPS70* | 524;A to T | G to A* |
| 11 | 162,688-162,695 | 8 | | | del |
| 11 | 371527-371,535 | 9 | | | del* |
| 11 | 403,466 | n/a | *YKL018C-A* | 19;S to S | C to T |
| 11 | 465,694 | n/a | *YKR014C/YPT52* | 126;K to K | A to G* |
| 11 | 615,684-615,694 | 11 | | | del* |
| 12 | 201,846-201,856 | 11 | | | del |
| 12 | 363,531-363,537 | 7 | *YLR106C/MDN1* | fs;68-70 | del |
| 12 | 447,845-447,855 | 11 | | | del* |
| 12 | 1,047,741 | n/a | *YLR454W/FMP27* | 1249:D to G | A to G |
| 13 | 139,705-139,709 | 5 | *YML067C/ERV41* | fs;138-139 | del |
| 13 | 141,724-141,731 | 8 | | | del* |
| 13 | 168,876 | n/a | *SUP5*(tRNA-Tyr) | | C to G* |
| 13 | 169,482 | n/a | *YML053C* | 91;N to N | C to T* |
| 13 | 353,769-353,777 | 9 | | | del* |
| 13 | 468,259-468,272 | 14 | *ARS1316* | | del* |
| 13 | 816,457-816,463 | 7 | *YMR275C/BUL1* | fs;706-708 | del |
| 14 | 400,002 | n/a | *YNL121C/TOM70* | 180;G to STOP | C to A |
| 14 | 435,595-435,601 | 7 | *YNL101W/AVT4* | fs;199-201 | del |
| 14 | 481,123-481,129 | 7 | | | del |
| 14 | 575,616-575,626 | 11 | | | del |
| 14 | 685,574-685,582 | 9 | | | del |
| 14 | 734,521 | n/a | *YNR058W/BIO3* | 77;L to L | A to G |
| 15 | 151,094-151,100 | 7 | *YOL089C/HAL9* | fs;797-799 | del* |
| 15 | 201,885-201,893 | 9 | *YOL068C/HST1* | six nt upstream ATG | del* |
| 15 | 227543-227552 | 10 | | | del* |
| 15 | 277,878-277,891 | $(TA)_7$ | | | di-ins* |
| 15 | 1,061,340-1,061,348 | 9 | | | del* |
| 16 | 22,677 | n/a | | | C to T |
| 16 | 99,457-99,466 | 10 | | | del* |
| 16 | 131,583 | n/a | *YPL222W/FMP40* | 475;A to T | G to A |
| 16 | 276,080-276,093 | $(AT)_7$ | | | di-ins* |
| 16 | 445,449-445,456 | 8 | | | del* |
| 16 | 485,927-485,952 | $(CA)_13$ | | | di-del* |
| 16 | 509,632 | n/a | *YPL022W/RAD1* | 980;A to S | G to T |
| 16 | 570,131 | n/a | *YPR007C/REC8* | 415;S to S | C to T |
| 16 | 639,362-639,372 | 11 | | | del* |

Continued on next page

**Table 4.4 – continued from previous page**

| Chromosome | SGD Position | HP tract | Gene | aa change | Mutation |
|---|---|---|---|---|---|
| 16 | 818,771-818,780 | 10 | *YPR143W/RRP15* | fs;150-153 | del* |

Table 4.4: The type of mutation (base substitution, single nt insertion (ins), single nt deletion (del), single di-nt insertion (di-ins), single di-nt deletion (di-del)) is shown, as well as the length of the HP/di-nucleotide tract that contains an in-del. All HP tracts were poly(dA:dT). Di-nucleotide tracts are shown by sequence and repeat size. For mutations that occurred within an open reading frame, both the gene name and predicted amino acid (aa) changes (fs; frameshift) are provided. n/a; not applicable. Coordinates are presented as shown in the SGD (http://www.yeastgenome.org/). * Mutations identified in the paired-end sequencing run.

## 4.4 Bioinformatics of the detected variation and Estimates of mutation rates in *mlh1*-$7^{ts}$

We examined four independent conditional MMR-defective (*mlh1*-$7^{ts}$) diploid lines (referred to as Mut) of yeast derived from a common ancestor and grown for 160 generations at the non-permissive temperature, with bottlenecks reducing the population to one cell every 20 generations. At the non-permissive temperature, the *mlh1*-$7^{ts}$ mutants show a null-like phenotype in the canavanine resistance mutation assay and a nearly null phenotype in the $lys2_{A14}$ reversion assay (Heck *et al*. 2006b).

To identify mutations present in Mut lines, we performed paired-end whole genome sequencing of three lines to moderate coverage and on one line, Mut4, to very high coverage. As shown in Table 4.4, we detected 19 base substitutions and 73 single- and di-nucleotide insertion/deletion (in-del) mutations in Mut4, all of which were heterozygous. The 73 in-dels, representing nearly 80% of all of the mutations detected in Mut4, consisted of 65 deletions and 8 insertions, and

occurred in 4- to 13-nt long HP tracts or in 6- to 13-repeat dinucleotide (di-nt) tracts (Table 4.4). The mutations in the HP tracts were all in $A_n$ or $T_n$ sequences, consistent with these repeats representing ~ 95% of the 5 to 20 nt HP tracts in the genome and greater than 99% of the HP tracts 8 nt or larger. The predominance of nucleotide deletions over insertions and base substitutions in MMR defective strains was similar to that seen previously in a genome wide analysis [106] and in reporter constructs [93, 29].

As shown in Table 4.5, the mutation rate for 5- to 13-bp poly(dA:dT) tracts was $2.8 \times 10^{-6}$/tract/generation. The rate was an order of magnitude higher ($3.1 \times 10^{-5}$) when only runs between 8- to 13-bp long were considered and was similar to that seen using reporter assays in MMR null mutants (T10 tract, $2.8 \times 10^{-4}$ [93]; A10 tract, $7.3 \times 10^{-5}$ [29]). The mutation rate in the 6- to 17-bp di-nucleotide tracts was $6.8 \times 10^{-5}$/di-nt tract/generation, which is also similar to that seen using reporter assays in MMR null mutants [85]. The base substitution rate in Mut4 was of $5.3 \times 10^{-9}$ mutations/base/generation, which is 1.3-fold higher than the genomic base substitution rate previously calculated for Mut4 ($4.0 \times 10^{-9}$; [106]). The similar rate for base substitutions seen in Zanders et al. [106] and the current study indicates that our experimentally aware Bayesian caller, which pooled information from across bottleneck lines, was effective at mutation detection even under conditions of low coverage and short read length.

To estimate the efficiency of detection, we took advantage of the fact that the Mut4 line at generation 160 showed 3% spore viability [37]. In Mut4, 34 of the heterozygous mutations map to open reading frames, five of which are frameshifts in HP tracts in genes (*KRR1, KRS1, RAD3, MDN1, RRP15*) in which null mutations confer lethality. Genotyping analysis (Table

Table 4.5: Mutation rates for Mut4 line grown in bottlenecks for 160 generations. The base substitution mutation rate was determined by calculating the percentage of the genome in which at least 34-fold DNA sequencing coverage to unique regions was obtained. This was done because our statistical analysis did not have sufficient power to reliably detect heterozygous mutations in regions with lower coverage. This information was used to calculate the mutation rate based on the following formula: (number of mutations)/(160 generations)/(adjusted genome size), with the diploid *S. cerevisiae* genome size determined as 24,141,794 bp (http://www.yeastgenome.org/). To obtain in-del mutation rates, we first determined the number of HP and di-nucleotide tracts of a given length in unique regions of the genome which had $\geq$ 34-fold sequence coverage. We then used the following equation to calculate mutation rate: (number of in-dels)/(160 generations)/(number of HP/di-nt tracts with $\geq$ 34-fold coverage).

| # base mutation | % genome sufficient coverage | Genome Size(bp) adjusted | Mutation rate(per base/gen $\times 10^{-9}$) |
|---|---|---|---|
| 19 | 92% | 22,212,356 | 5.3 |
| | # mutations | # HP tracts 34X coverage | Mutation rate (per base/gen $\times 10^{-7}$) |
| 5-13nt HP tract in-dels | 68 | 147,422 | 28 |
| 8-13nt HP tracts in-dels | 55 | 10,548 | 314 |
| 6-17di-nucleotide tracts | 4 | 368 | 679 |

Table 4.6: Segregation of heterozygous mutations in the Mut4 line: Genotyping was done by Sanger sequencing as described in the previous section and * represents non-synonymous mutation.

| locus | location | knockout recessive lethal? | WT:mut spore seg. |
|---|---|---|---|
| *KIN82* | Chr3 275085 (SNP22) | no | 3:3 |
| *EPLI* | Chr6 88632 (SNP29) | no | 5:11 |
| *PHO4* | Chr6 225029 (SNP32) | no | 4:2 |
| *YBR219C* | Chr2 662320 (SNP38) | no | 3:3 |
| *AIM19* | Chr9 199795 (SNP45) | no | 2:4 |
| *TOM70* | Chr14 399797 (SNP64) | no | 2:4 |
| *AVT4* | Chr14 435396 (SNP65) | no | 2:4 |
| *BIO3* | Chr14 734316 (SNP69) | no | 2:4 |
| *FMP27* | Chr12 1047541 (SNP79) | no | 2:4 |
| *ERV41* | Chr13 139505 (SNP80) | no | 4:7 |
| *BUL1* | Chr13 816264 (SNP81) | no | 4:2 |
| *YJRO12C* | Chr10 460308 | no | 3:5 |
| *KRR1* | Chr3 22745 | yes | 12:0 |
| *KRS1* | Chr4 525612 | yes | 12:0 |
| *CDC7** | Chr4 424584 | yes | 10:2 |
| *RAD3* | Chr5 528691 | yes | 12:0 |
| *MDN1* | Chr12 363531 | yes | 16:0 |
| *FMP40* | Chr16 131383 (SNP82) | no | 4:2 |
| *RAD1* | Chr16 509432 (SNP83) | no | 2:9 |
| *REC8* | Chr16 569931 (SNP84) | no | 1:10 |
| | Chr16 639362 | no | 1:11 |
| *RRP15* | Chr16 818766 | yes | 12:0 |

4.6) showed that Mut4 viable spores contained only wild-type alleles of *KRR1, KRS1, RAD3, MDN1, RRP15*; both wild-type and mutant alleles were detected in spores for other heterozygous Mut4 mutations. Importantly, the low spore viability (3%) seen in Mut4 is consistent with five recessive lethal mutations. Thus we are confident that these mutations encompass most, if not all, mutations present in coding regions in this line.

## 4.5 Detection of mutations at a single position in independent lines and identification of mutational hotspots in the genome

Notably that we have detected one mutation which occured at a single position in two independent lines in previous study described in Chapter 3, therefore, we performed whole genome scans to search for potentially more independent mutations in the same genome position. This was performed by utilizing the 36 nt single-end data set (Chapter 3, [106]) and combining it with the current paired-end data. We used a Bayesian population genomic approach that pools the information regarding allele frequency among individuals within the same species when calling genotypes (chapter 2, [60]). This model leverages all the short-read sequence data for all wild-type/Mut lines that are aligned to the reference genome. The fact that we can confidently call genotypes in wild-type-0 and Mut4-160 increased our confidence of calling mutations in other lines (i.e., Mut2/3). By pooling data, this model is designed to call genotypes in lines with low coverage by borrowing power from lines with high coverage. However, there is a limit on how much information can be obtained because in many cases the short read sequence data cannot detect mutations in homopolymeric tracts because reads often do not encompass the entire tract. As shown in Table 4.7, nine such mutations were found. All of them were single nucleotide in-dels in poly(dA:dT) tracts of 9-14 nt.

Because each of the Mut lines was derived from a common ancestor and the power of in-del mutation detection depends heavily on the read length and coverage, it is reasonable to assume that each of the Mut lines involves the same

Table 4.7: Mutations observed in multiple lines: Listed are mutations that occurred independently in two of the three mutator generation 160 lines that were sequenced

| Chromosome | SGD Position | poly(dA:dT) tract | lines:type of mutation |
|---|---|---|---|
| 2 | 92,273-92,281 | 9 | 2:del, 3:del |
| 4 | 216,494-216,503 | 10 | 3:del, 4:del |
| 4 | 314,305-314,316 | 12 | 3:del, 4:del |
| 7 | 533,997-534,006 | 10 | 3:del, 4:ins |
| 7 | 394,901-394,911 | 11 | 2: del, 3:del |
| 8 | 519,049-519,060 | 12 | 2: del, 3:del |
| 9 | 169,789-169,797 | 9 | 2:del, 3:del |
| 9 | 406,049-406,058 | 10 | 2:del, 3:del |
| 13 | 468,259-468,272 | 14 | 3:del, 4:del |

number of single-nt in-dels in HP tracts of 9- and 14-nt for a total of 50 per line (the amount seen in the Mut4 line using high coverage data). Based on a total of 3237 HP tracts of 9-14 nt in the S288c diploid *Saccharomyces* reference genome (http://www.yeastgenome.org), we can assume that the total number of HP tracts of 9-14 nt in the genome that we can have potential double-hit mutations is 2005. This number is based on the coverage of the Mut2 line (8-10X), and the fact that HP tract lengths are long. This is a conservative calculation because we require that all lines have at least 5X coverage in order to be considered as a site for independent mutations in multiple lines. We can often identify mutations in the Mut2 line at positions of low coverage (i.e. 4X) when the Mut3 (10X) and Mut4 (200X) lines show high coverage.

Given that each of the three lines has 50 mutations in homopolymeric (HP)

tracts of size 9-14 nt and a total of 2005 possible target sites where mutations could occur, the tail probability of observing at least five double-hits in lines Mut2 and 3 or at least four double-hits in lines Mut3 and 4 can be calculated:

$$\text{Probability} = \sum_{k}^{50} \frac{\binom{2005}{k} \times \binom{2005-k}{50-k} \times \binom{(2005-k)-(50-k)}{(50-k)}}{\binom{2005}{50} \times \binom{2005}{50}} \tag{4.14}$$

using $k = 5$ as an example, we obtain $P = 7.15 \times 10^{-3}$

We employed the most conservative way of calculating the probability of double hits. Even with this method, the probability is still significantly higher than expected by chance, indicating that the coincident mutations likely result from mutagenesis hotspots.

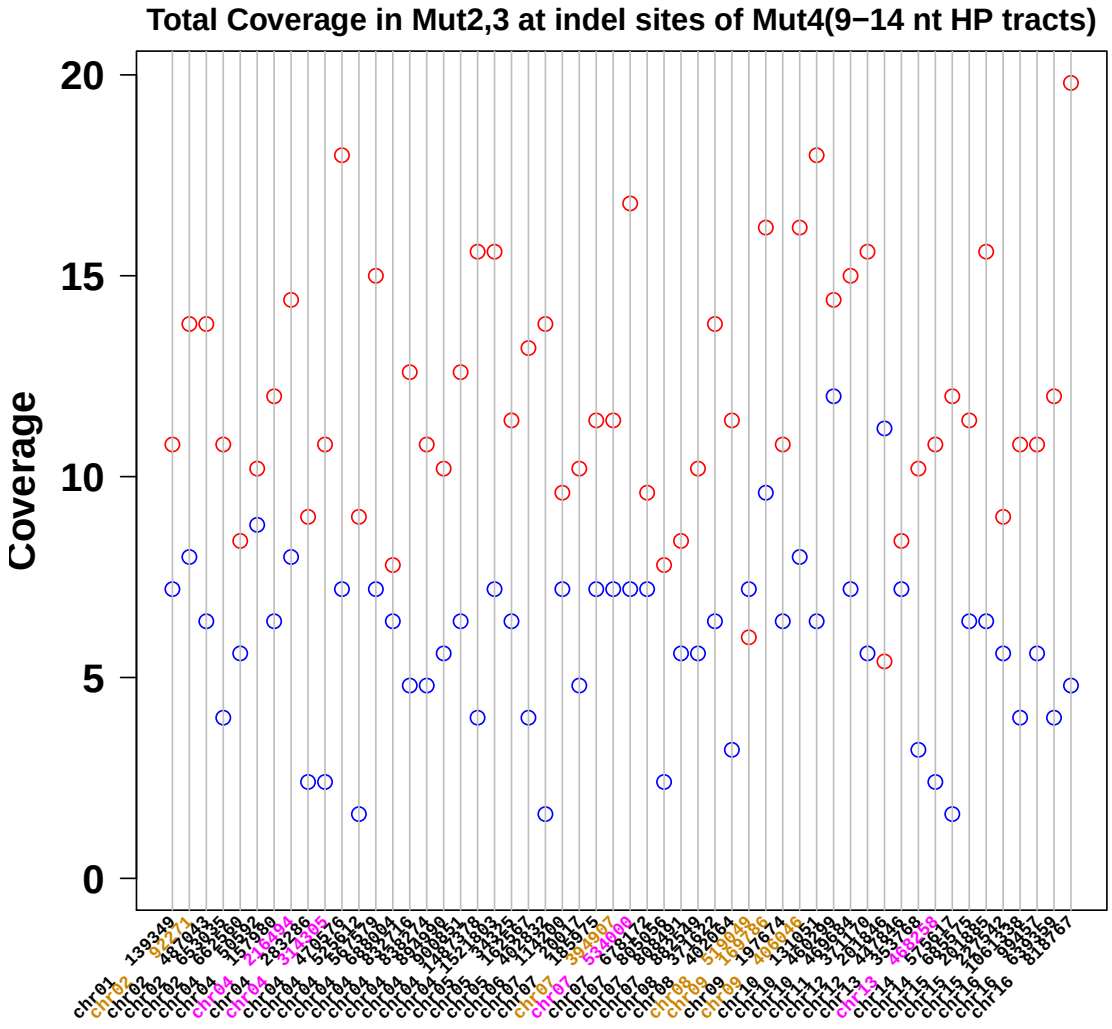
All of the double-hit mutations involved either Mut2/3 or Mut3/4; no mutations were detected in both Mut2 and Mut4. This is likely due to Mut2 having the lowest coverage among the three Mut lines. In Figure 4.5 we plot the coverage distribution for Mut2/3 at each of the loci where Mut4 has an in-del mutation, including the double hit loci involving Mut2/3. We found that Mut2 displayed the lowest coverage for most loci in the plot. We also calculated the probability of nine double hits occurring among three Mut lines but not involving Mut2 and Mut4. This probability, $P = 0.026$, is not significant with a cutoff value of 0.01. Therefore, given low Mut2 sequencing coverage, it is reasonable to expect nine double hits among three lines without a specific combination of two lines.

## 4.6 A broad sequence context for mutagenesis and statistical test to examine association of a mutation in a given poly(dA:dT) tract with nearby poly(dA:dT) tracts

Mutational hotspots occur in repetitive DNA such as HP tracts and dinucleotide repeats [93, 20, 21, 29, 89, 85]. While such mutational biases have been identified at a local sequence level (within ~80 bp), larger genomic contexts were not thought to contribute or may be difficult to find (e.g. [33, 81, 13].

To test if DNA sequences surrounding in-del mutations in the Mut4 generation 160 line were enriched for poly(dA:dT) tracts, non-overlapping 50 to 4000 bp windows, centered around size-matched 5-14 nt poly(dA:dT) tracts, were analyzed. Windows either contained (64 sites) or lacked (39290, 32891, 24959, 14743, 8773, 4780, 2654, 1972 sites for 50, 100, 200, 500, 1000, 2000, 3000, 4000 bp windows, respectively) an in-del mutation in a poly(dA:dT) tract. This was done because mutations in poly(dA:dT) tracts represented the majority (~ 70%) of mutations detected in the Mut4 line and would thus provide the best opportunity to find broad sequence contexts. A Negative Binomial (NB) model was fitted, where the number of poly(dA:dT) tracks in a fixed window size was counted, excluding the center site, to account for the reasonable small mean and over-dispersion that cannot be predicted by a simple Poisson model. A goodness of fit test was then performed for the two distributions where nearby bins were combined so that they have an expected value of at least five for a fixed window size. This was done to make sure the negative binomial distribution is appropriate. We then tested if there was a difference between nearby poly(dA:dT) tract occurrences for windows with and without an poly(dA:dT)

Figure 4.5: Coverage plot in Mut2/3 for every mutation site in Mut4. The coverage at the corresponding position in the Mut2 and 3 lines is shown for each Mut4 in-del site. Blue indicates coverage for Mut2, red, coverage for Mut3. The Y-axis indicates coverage; the X-axis lists all 54 in-del mutations occurring within 9-14 nt HP tracts. The yellow colored positions on the X-axis represent the hotspot coordinates between Mut2 and Mut3; the pink colored positions represent the hotspot coordinates between Mut3 and Mut4.



tract mutation (the difference between the mean parameters equals to 0), using likelihood-based methods. This performed well for testing equality of mean counts modeled by a negative binomial distribution, even when the over-dispersion parameter of one group was twice that of the other group ([1]; Figure 4.6).

We also carried out an association test of nearby AT content in fixed window sizes with and without an in-del mutations (Figure 4.7). For each fixed window size, we computed the percentage of AT content excluding the center poly(dA:dT) tract. For each of the two distributions in a fixed window size, we first fitted a normal distribution, and then used the Bootstrap Kolmogorov-Smirnov test, which executes a bootstrap version of the univariate Kolmogorov-Smirnov test to correct coverage when distributions compared are not entirely continuous [83]. This was performed to assess the fitness for each of the distributions (all $P$-values were $> 0.1$ and were not significant). Furthermore, we used an F-test to make sure that the two distributions in the fixed window size have equal variance (all $P$-values were $> 0.1$, not significant). Lastly, we used a Z-test to compare the mean between the two distributions.

Based on the above findings we tested, using conservative statistical methods that accounts for the need to compare small (detected mutations) and large (potential sites in the genome) data sets, whether windows surrounding poly(dA:dT) mutations in Mut4 contained specific sequence patterns, focusing on nearby poly(dA:dT) because a previous visual scan of mutated sequences suggested an enrichment of such tracts [106] excluding the mutated site. As shown in Figure 4.7, the occurrence of a mutation in a poly(dA:dT) tract was highly associated with the number of nearby 5 to 14 nt poly(dA:dT) tracts for

Figure 4.6: Sliding window analysis for in-del mutations in poly(dA:dT) tracts in the Mut4 line. The number of 5-14 nt poly(dA:dT) tracts was counted under different window sizes(50-4000 bp).This was determined for windows centered on poly(dA:dT) tracts with (red) and without (blue) an in-del mutation. The center sites were excluded from the counting analysis. The X-axis displays the number of poly(dA:dT) tracts contained within each window. The Y-axis shows the frequency for which each poly(dA:dT) tract was observed. The fitted size (S) and mean ($\mu$) for each of the two distributions in a fixed window size is listed. The *P*-value of the likelihood-based method used to compare the means of two Negative Binomial distributions ([1]) is shown for each window

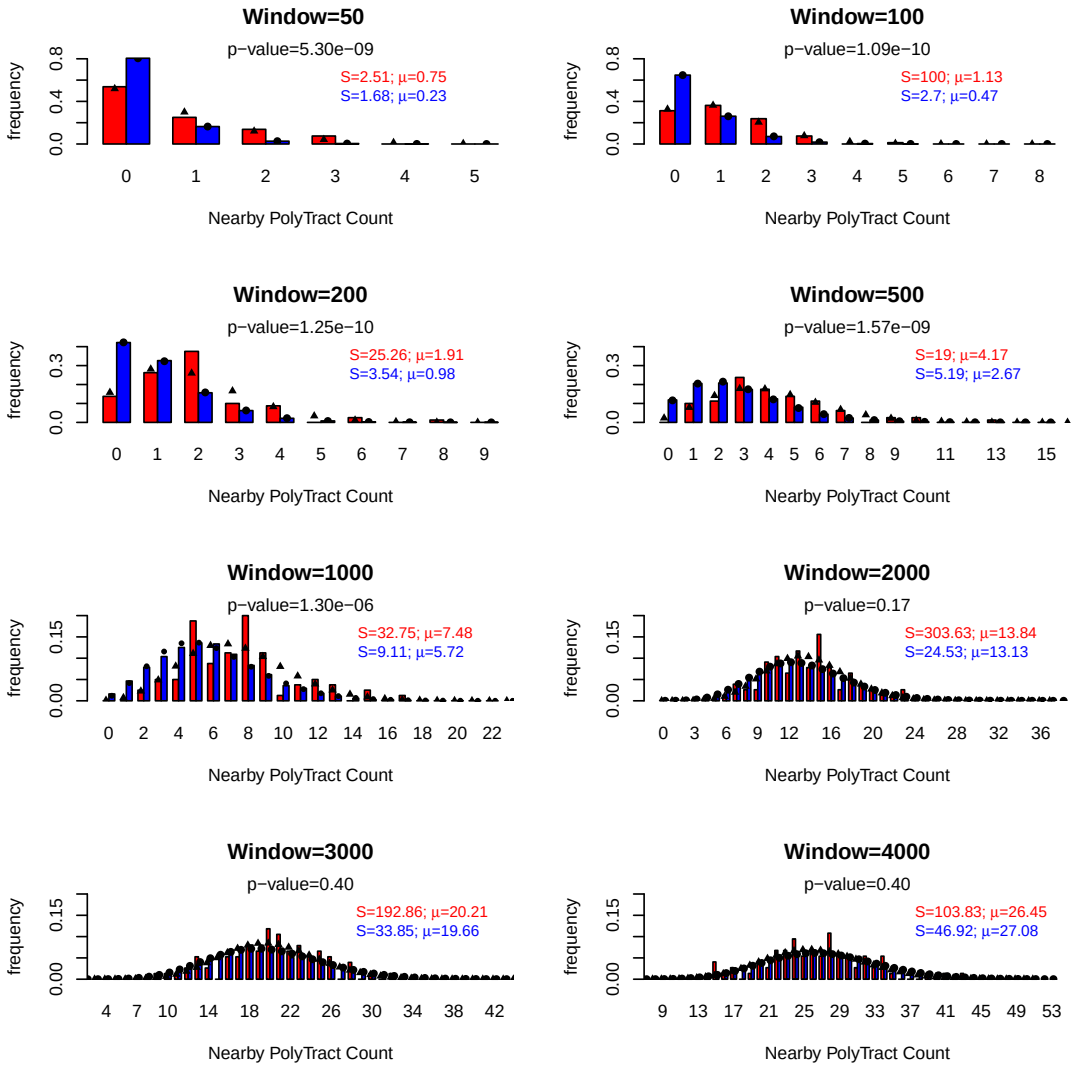
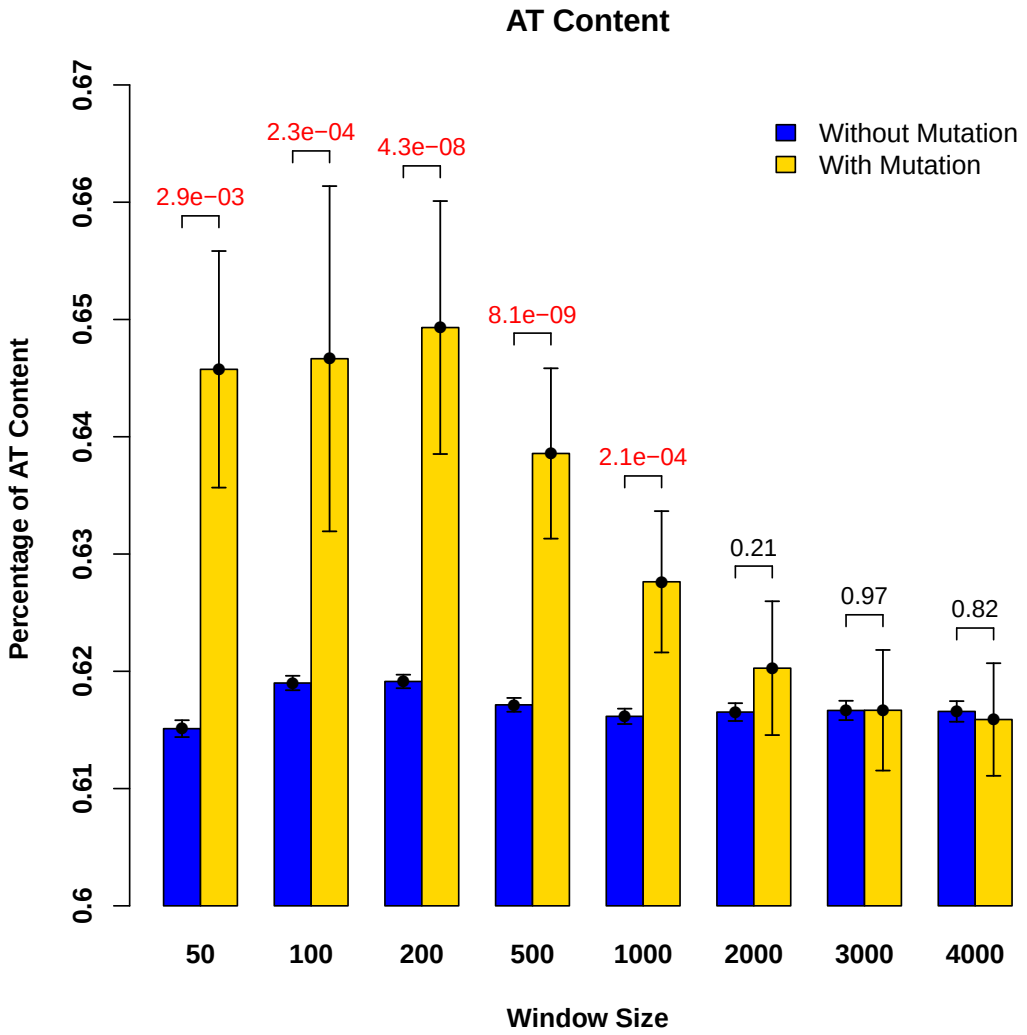


Figure 4.7: Association testing of nearby AT content in a fixed window size in the Mut4 line. AT content was determined for each window (50-4000 bp) under conditions where the centered poly(dA:dT) tract, with or without an in-del mutation, was excluded. The X-axis shows the fixed window size and the Y-axis displays the mean AT content observed among all windows for the fixed window size. For a given fixed window size, we included the 5% error for each of the two distributions and also grouped the two distributions. The *P*-value for a Z-test used to compare the means of the two distributions is shown for each window. Red represents significance ($P < 0.05$) and black represents a lack of significance ($P > 0.05$)



windows of 50 to 1000 bp. An even stronger correlation was seen for the same window sizes when data for the Mut2 and Mut3 poly(dA:dT) tracts were included (data not shown). A statistical association was not seen for 2000 to 4000 bp windows. If the 50 bp surrounding the mutated poly(dA:dT) tract is excluded, the genomic context of the mutated poly(dA:dT) tracts still contains significantly higher poly(dA:dT) tracts than unmutated HP tracts for window sizes of 100 ($P$ =$1.7 \times 10^{-3}$), 200 ($P = 2.0 \times 10^{-5}$), 500 ($P = 4.8 \times 10^{-5}$) and 1000 ($P = 9.2 \times 10^{-4}$) bp. Together, these analysis show that a larger genomic context plays a role in the formation of mutations at HP tracts. No significant association was found for the Mut4 line when a window analysis (window sizes 50, 100, 200) was performed to examine a correlation between single base change mutations and nearby poly(dA:dT) tracts ($P \geq 0.1$ for all windows).

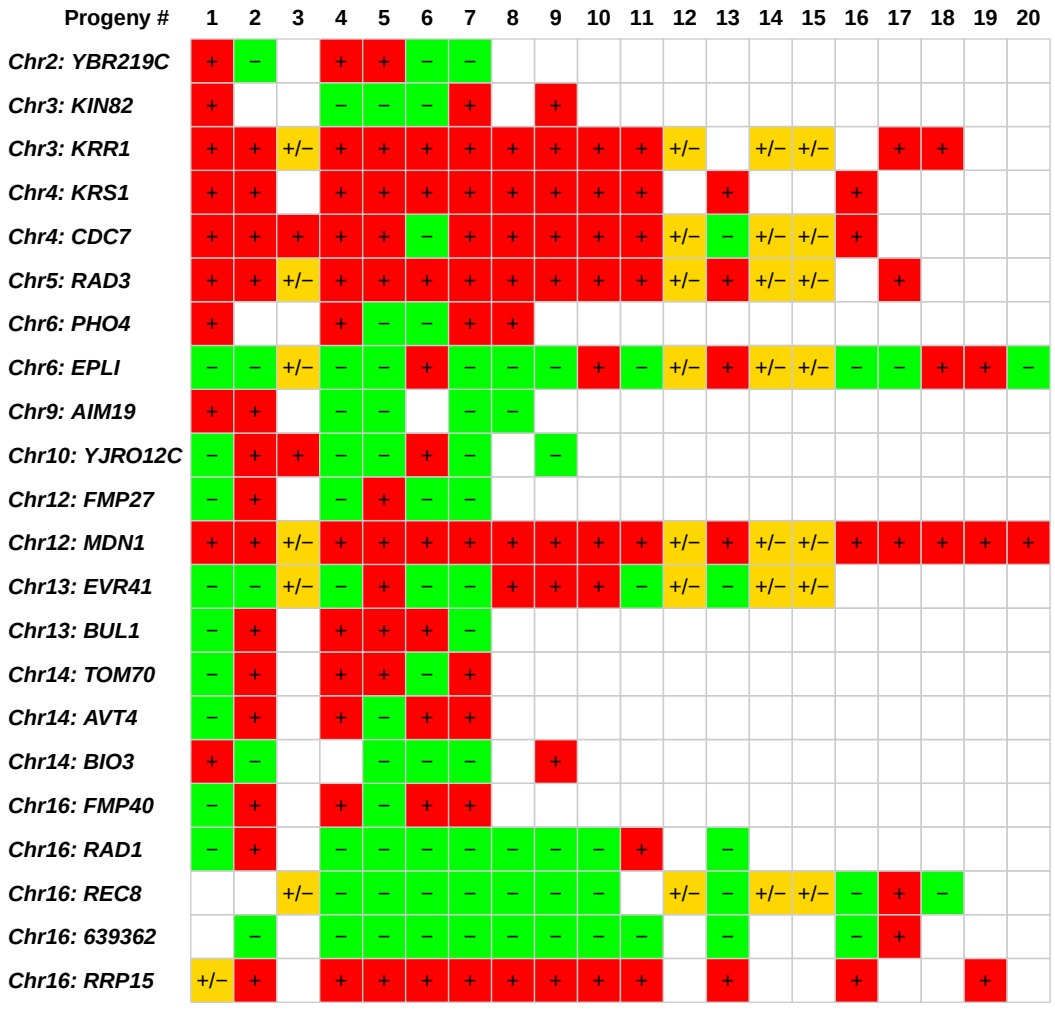
## 4.7 Spore genotyping and detection of chromosome aneuploidy

Chromosomal DNA was extracted from haploid spore clones obtained from the Mut4 diploid line at generation 160 and then genotyped for 20 heterozygous mutations identified in the diploid by Maria Rogacheva in Eric Alani's group in Cornell University (Figure 4.8). For 15 of the 20 mutations, at least one wild-type and one mutant allele were detected in spore progeny. For the other five mutations (mapping to frameshift mutations in the essential *KRR1, KRS1, RAD3, MDN1 and RRP15* open reading frames) only the wild-type allele was detected. Deviations from a 1:1 ratio were detected at some loci (e.g. *EPL1, CDC7, REC8*, Chromosome XVI:509432, Chromosome XVI:639362).

In some cases the skew can be explained by linkage to a recessive lethal gene (e.g. *CDC7* is located 5 cM from *KRS1*). In other cases (e.g. the 300 KB region in Chromosome XVI), the best explanation is that a crossover defect contributes to the linkage defect. Curiously, for three loci on a 300 KB stretch of chromosome XVI we detected a skew from 1:1 ratio that appeared more extreme for markers closer to the *RRP15* locus. One explanation for this pattern is that the 300 KB region is linked to the *rrp15* lethal mutation on chromosome XVI; we have no evidence for another recessive lethal mutation in this region. However, such tight linkage is unusual for baker's yeast, which shows on average a physical distance of 3 KB for each cM of map distance (http://www.yeastgenome.org/).

One way to explain the segregation patterns described above is that there is a defect in meiotic crossing over that increases linkage between distant markers. Such a defect would disrupt chromosome segregation in Meiosis I because crossovers are critical for ensuring a proper division. As shown in Figure 4.8, we have evidence of an aneuploidy phenotype in the Mut4 line. 20 Mut4 spore clones were genotyped for the presence of wild-type and mutant alleles as described in [106]. Briefly, chromosomal DNA encompassing mutations located on 11 chromosomes was amplified by PCR and then analyzed by Sanger sequencing. Five of twenty genotyped spore clones display evidence of chromosome aneuploidy as shown by the presence of both alleles of specific mutations. Interestingly, in all five cases both alleles can be found for mutations that cause a recessive lethal mutation. This suggests that the meiotic chromosome aneuploidy phenotype could contribute to low spore viability; however, we believe that this likely to be minimal because such aneuploidy can rescue the inviability of spores that contain recessive lethal mutations. Such large-scale aneuploidy has been seen in Drosophila male meiosis in mutants that show severe defects in

Figure 4.8: Mut4 spore clones genotyped for the presence of wild-type and mutant alleles located on eleven chromosomes. Chromosomal DNA from 20 spore clones obtained by dissecting tetrads from the Mut4 line were PCR amplified using primers specific to the above loci. The resulting DNA was then subjected to Sanger sequencing. Spore clones were genotype using the following designations: +, presence of wild-type allele; -, presence of mutant allele; +/-, presence of both wild-type and mutant alleles. Note that spore progeny 3, 12, 14, and 15 failed to mate with either *MATa* or *MATalpha* mating testers



meiotic chromosome segregation and assemble chromosome-free spindles [11]. Because there are no obvious candidates or chromosome rearrangements that could explain this phenotype, we hypothesize that the meiotic aneuploidy defect resulted from mutational load.

## 4.8 Discussion

Individual HP tracts are known to be sensitive to in-del mutations, which are caused primarily by DNA slippage during DNA replication [93, 20, 21, 29, 89, 85]. Such slippage events are not thought to be influenced by local sequence context, including adjacent HP tracts [33]. In this study we show that the likelihood of a mutation in a given poly(dA:dT) tract is increased by the presence of poly(dA:dT) tracts in a 1 KB region centered on the given tract. Due to the size of the hotspot region, ~ 1 KB, it would have been very difficult to identify such a broad DNA sequence context by creating specific reporter constructs or searching for the association of mutations with unique DNA sequence motifs. It is important to note that our work is distinct from bioinformatic studies of Denver *et al*. [20], who observed that the *C. elegans* genome contains distinct clusters of HP tracts in autosomal arms. They hypothesized that such sites could be hotspots for recombination but also suggested that certain types of nearly tandem repeat clusters could serve as hotspots for slippage-mediated deletions.

As outlined in Chatper 1, a number of molecular, population genetic, and bioinformatic studies have shown that mutation rate varies across the eukaryotic genome. For example, Hawk et al. [34] showed in baker's yeast that the mutation rate of a microsatellite reporter placed at different chromosomal po-

sitions could vary by 16-fold; however, they were unable to identify a specific motif/chromosomal signature associated with shared mutations. Why might clusters of poly(dA:dT) tracts create mutational hotspots? One possibility is that clusters of these tracts form a secondary structure such as bent or flexible DNA that would predispose DNA polymerase to slippage [39, 87]. If such structures exist, they are likely to be unstable, because we were unable to detect in acrylamide gels a change in the expected mobility of ~ 400 bp DNA fragments containing the DNA sequence in which in-dels were detected (data not shown). Alternatively, poly(dA:dT) tracts have been shown to be stiff, resist bending, and could affect mutagenesis by excluding nucleosomes [82, 100]. A third possibility is that DNA polymerase stalling at HP tracts facilitates polymerase switching, perhaps to a DNA polymerase that replicates adjoining HP tract with lower fidelity [63]. Work by Kim *et al*. [47] support such an idea. They found that indel mutations in HP tracts under high-transcription conditions were partially dependent on the function of polymerase zeta, an error-prone translesion DNA polymerase. A fourth possibility is that a cluster of HP tracts confers an increased mutation rate through increased transcription because it is known that poly(dA:dT) tracts serve as ubiquitous promoters [90, 42]. It will be important in the future to develop model systems to distinguish between these possible mechanisms.

Moreover, We used conservative statistical methods to determine if broader sequence contexts were associated with mutagenesis by examining the Mut4 sequencing data and that of two other lines, Mut2 and 3 [106], to look for specific sites mutated in two of three Mut generation 160 lines. We found using the most conservative approach that the probability of identifying nine independent mutations at multiple sites by chance was low ($P = 7.15 \times 10^{-3}$). We were

unable to identify any associations for the nine mutations with respect to origins of DNA replication (ORC and Mcm2 binding sites; [105], centromere position, and Ty-element density (http://www.yeastgenome.org). It is possible that the small size of our data set precludes the identification of a specific pattern, or that complex non-overlapping parameters create mutational hotspots at these sites.

In summary, through mutation accumulation experiments performed in MMR defective yeast strains, we identified novel mutational hotspots in the genome and found a new pattern in which the likelihood of a in-del mutation in a given poly(dA:dT) tract is increased by the presence of nearby poly(dA:dT) tracts. The identification of mutational hotspots supports the idea that natural selection occurs in a landscape where certain sequences and regions of the genome are mutated at higher frequency. Such information provides important clues on targets for evolvability in cell types that are mutators due to defects in specific repair processes [36, 18, 91, 61].

Our work also supports the idea that the primary role of MMR is to remove in-del mutations in HP tracts [93, 106]. Such in-del mutations occur during DNA replication primarily as the result of slippage by DNA polymerases [89, 93, 29, 39]. In wild-type yeast DNA slippage events are rarely detected in HP tracts due to the detection and removal of slippage intermediates by MMR [78]. Based on the observation that 25% of yeast ORFs have HP tracts 8 nt or longer [93] and 56% of ORFs have HP tracts 5 nt or longer (S288c reference genome), Tran *et al*. [93] hypothesized that the high rate of mutation in HP tracts could explain "the high rates of recessive lethal mutations that accumulate in diploid Mmr- (*pms1* and *msh2*) yeast." They also suggested that "the lack of MMR in cancer tissue could lead to inactivation of genes with long homonu-

cleotide runs that are important for cancer progression and for secondary effects of cancer." Mutations in four MMR genes confer predisposition to hereditary, non-polyposis colorectal cancer (HNPCC) [48, 64]. Our genome-wide analysis of mutations observed in MMR defective lines, coupled with the detection of recessive lethal mutations seen as frameshift mutations in HP tracts (Table 4.7), confirms the Tran et al. [93] hypothesis and supports the idea that inactivation of genes with HP tracts is critical for cancer progression in MMR deficient tumors. Genes with long HP runs are mutated in MMR deficient tumors (reviewed in [84]) and thus are likely to contribute to the cancer specificity observed in MMR mutants.

## BIBLIOGRAPHY


[1] I. B. Aban, G. R. Cutter, and N. Mavinga. Inferences and power analysis concerning two negative binomial distributions with an application to mri lesion counts data. *Comput Stat Data Anal*, 53(3):820–833–, 2008.

[2] Frdric Achard, Hugh D Eva, Hans-Jrgen Stibig, Philippe Mayaux, Javier Gallego, Timothy Richards, and Jean-Paul Malingreau. Determination of deforestation rates of the world's humid tropical forests. *Science*, 297(5583):999–1002, Aug 2002.

[3] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4):248–249, Apr 2010.

[4] M. Alcaide, D. Serrano, J. J. Negro, J. L. Tella, T. Laaksonen, C. Muller, A. Gal, and E. Korpimaki. Population fragmentation leads to isolation by distance but not genetic impoverishment in the philopatric lesser kestrel: a comparison with the widespread and sympatric eurasian kestrel. *Heredity*, 102(2):190–8–, 2009.

[5] P. F. Arndt, T. Hwa, and D. A. Petrov. Substantial regional variation in substitution rates in the human genome: importance of gc content, gene density, and telomere-specific effects. *J Mol Evol*, 60(6):748–63, 2005. eng.

[6] N. Arora, A. Nater, C. P. van Schaik, E. P. Willems, M. A. van Noordwijk, B. Goossens, N. Morf, M. Bastian, C. Knott, H. Morrogh-Bernard, N. Kuze, T. Kanamori, J. Pamungkas, D. Perwitasari-Farajallah, E. Verschoor, K. Warren, and M. Krutzen. Effects of pleistocene glaciations and rivers on the population structure of bornean orangutans (pongo pygmaeus). *Proc Natl Acad Sci U S A*, 107(50):21376–81, 2010. eng.

[7] C. F. Baer, M. M. Miyamoto, and D. R. Denver. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat Rev Genet*, 8(8):619–31–, 2007.

[8] Jeffrey A Bailey, Robert Baertsch, W. James Kent, David Haussler, and Evan E Eichler. Hotspots of mammalian chromosomal evolution. *Genome Biol*, 5(4):R23, 2004.

[9] H. J. Bandelt, V. Macaulay, and M. Richards. Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtdna. *Mol Phylogenet Evol*, 16(1):8–28, 2000. eng.

[10] A. R. Boyko, S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez, K. E. Lohmueller, M. D. Adams, S. Schmidt, J. J. Sninsky, S. R. Sunyaev, T. J. White, R. Nielsen, A. G. Clark, and C. D. Bustamante. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*, 4(5):e1000083, 2008. eng.

[11] E. Bucciarelli, M. G. Giansanti, S. Bonaccorsi, and M. Gatti. Spindle assembly and cytokinesis in the absence of chromosomes during drosophila male meiosis. *J Cell Biol*, 160(7):993–9–, 2003.

[12] C. D. Bustamante, A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz, S. Glanowski, D. M. Tanenbaum, T. J. White, J. J. Sninsky, R. D. Hernandez, D. Civello, M. D. Adams, M. Cargill, and A. G. Clark. Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062):1153–7, 2005. eng.

[13] K. A. Canella and M. M. Seidman. Mutation spectra in supf: approaches to elucidating sequence context effects. *Mutat Res*, 450(1-2):61–73–, 2000.

[14] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, Oct 2004.

[15] A. Coventry, L. M. Bull-Otterson, X. Liu, A. G. Clark, T. J. Maxwell, J. Crosby, J. E. Hixson, T. J. Rea, D. M. Muzny, L. R. Lewis, D. A. Wheeler, A. Sabo, C. Lusk, K. G. Weiss, H. Akbar, A. Cree, A. C. Hawes, I. Newsham, R. T. Varghese, D. Villasana, S. Gross, V. Joshi, J. Santibanez, M. Morgan, K. Chang, W. H. Iv, A. R. Templeton, E. Boerwinkle, R. Gibbs, and C. F. Sing. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun*, 1:131, 2010. eng.

[16] Kimura M. Crow, J.F. *An Introduction to Population Genetics Theory*. Harper and Row, 1970.

[17] A. Datta and S. Jinks-Robertson. Association of increased spontaneous mutation rates with high levels of transcription in yeast. *Science*, 268(5217):1616–9, 1995. eng.

[18] A. Demogines, A. Wong, C. Aquadro, and E. Alani. Incompatibilities involving yeast mismatch repair genes: a role for genetic modifiers and implications for disease penetrance and variation in genomic mutation rates. *PLoS Genet*, 4(6):e1000103–, 2008.

[19] D. R Denver, P. C Dolan, L. J Wilhelm, W Sung, J. I Lucas-Lledo, D. K Howe, S. C Lewis, K Okamoto, W. K Thomas, M Lynch, and C. F Baer. A genome-wide view of caenorhabditis elegans base-substitution mutation processes. *Proceedings of the National Academy of Sciences of the United States of America*, 106(38):16310–16314, 2009.

[20] D. R. Denver, K. Morris, A. Kewalramani, K. E. Harris, A. Chow, S. Estes, M. Lynch, and W. K. Thomas. Abundance, distribution, and mutation rates of homopolymeric nucleotide runs in the genome of caenorhabditis elegans. *J Mol Evol*, 58(5):584–95–, 2004.

[21] Dee R Denver, Seth Feinberg, Suzanne Estes, W Kelley Thomas, and Michael Lynch. Mutation rates, spectra and hotspots in mismatch repair-deficient caenorhabditis elegans. *Genetics*, 170(1):107–13, 2005.

[22] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput dna sequencing. *Nucleic Acids Res*, 36(16):e105, 2008. eng.

[23] T. Dos Vultos, O. Mestre, J. Rauzier, M. Golec, N. Rastogi, V. Rasolofo, T. Tonjum, C. Sola, I. Matic, and B. Gicquel. Evolution and diversity of clonal bacteria: the paradigm of mycobacterium tuberculosis. *PLoS One*, 3(2):e1538, 2008. eng.

[24] A. J. Drummond and A. Rambaut. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, 7:214, 2007. eng.

[25] R. M. Durbin. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.

[26] E. E. Eichler and D. Sankoff. Structural dynamics of eukaryotic chromosome evolution. *Science*, 301(5634):793–7–, 2003.

[27] W. J. Ewens. *Mathematical Population Genetics*. Springer, New York, 1979.

[28] Guri Giaever, Angela M Chu, Li Ni, Carla Connelly, Linda Riles, Steeve V—[Eacute]—Ronneau, Sally Dow, Ankuta Lucau-Danila, Keith An-

derson, Bruno Andr—[Eacute], Adam P Arkin, Anna Astromoff, Mohamed El Bakkoury, Rhonda Bangham, Rocio Benito, Sophie Brachat, Stefano Campanaro, Matt Curtiss, Karen Davis, Adam Deutschbauer, Karl-Dieter Entian, Patrick Flaherty, Francoise Foury, David J Garfinkel, Mark Gerstein, Deanna Gotte, Ulrich G—[Uuml]—Ldener, Johannes H Hegemann, Svenja Hempel, Zelek Herman, Daniel F Jaramillo, Diane E Kelly, Steven L Kelly, Peter K—[Ouml]—Tter, Darlene Labonte, David C Lamb, Ning Lan, Hong Liang, Hong Liao, Lucy Liu, Chuanyun Luo, Marc Lussier, Rong Mao, Patrice Menard, Siew Loon Ooi, Jose L Revuelta, Christopher J Roberts, Matthias Rose, Petra Ross-Macdonald, Bart Scherens, Greg Schimmack, Brenda Shafer, Daniel D Shoemaker, Sharon Sookhai-Mahadeo, Reginald K Storms, Jeffrey N Strathern, Giorgio Valle, Marleen Voet, Guido Volckaert, Ching-yun Wang, Teresa R Ward, Julie Wilhelmy, Elizabeth A Winzeler, Yonghong Yang, Grace Yen, Elaine Youngman, Kexin Yu, Howard Bussey, Jef D Boeke, Michael Snyder, Peter Philippsen, Ronald W Davis, and Mark Johnston. Functional profiling of the saccharomyces cerevisiae genome. *Nature*, 418(6896):387, 2002.

[29] H. Gragg, B. D. Harfe, and S. Jinks-Robertson. Base composition of mononucleotide runs affects dna polymerase slippage and removal of frameshift intermediates by mismatch repair in saccharomyces cerevisiae. *Mol Cell Biol*, 22(24):8756–62, 2002. eng.

[30] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5):696–704, 2003. eng.

[31] Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, John L Rinn, Eric S Lander, and Aviv Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nat Biotech*, 28(5):503–510, May 2010.

[32] R. C. Hardison, K. M. Roskin, S. Yang, M. Diekhans, W. J. Kent, R. Weber, L. Elnitski, J. Li, M. O'Connor, D. Kolbe, S. Schwartz, T. S. Furey, S. Whelan, N. Goldman, A. Smit, W. Miller, F. Chiaromonte, and D. Haussler. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res*, 13(1):13–26, 2003. eng.

[33] B. D. Harfe and S. Jinks-Robertson. Sequence composition and context ef-

fects on the generation and repair of frameshift intermediates in mononucleotide runs in saccharomyces cerevisiae. *Genetics*, 156(2):571–8, 2000. eng.

[34] J. D. Hawk, L. Stefanovic, J. C. Boyer, T. D. Petes, and R. A. Farber. Variation in efficiency of dna mismatch repair at different sites in the yeast genome. *Proc Natl Acad Sci U S A*, 102(24):8639–43, 2005. eng.

[35] Lawrence R. Heaney. A synopsis of climatic and vegetational change in southeast asia. *Climatic Change*, 19(1):53–61, 1991.

[36] J. A. Heck, J. L. Argueso, Z. Gemici, R. G. Reeves, A. Bernard, C. F. Aquadro, and E. Alani. Negative epistasis between natural variants of the saccharomyces cerevisiae mlh1 and pms1 genes results in a defect in mismatch repair. *Proc Natl Acad Sci U S A*, 103(9):3256–61–, 2006.

[37] J. A. Heck, D. Gresham, D. Botstein, and E. Alani. Accumulation of recessive lethal mutations in saccharomyces cerevisiae mlh1 mismatch repair mutants is not associated with gross chromosomal rearrangements. *Genetics*, 174(1):519–23, 2006. eng.

[38] Dale J Hedges, Dale Hedges, Dan Burges, Eric Powell, Cherylyn Almonte, Jia Huang, Stuart Young, Benjamin Boese, Mike Schmidt, Margaret A Pericak-Vance, Eden Martin, Xinmin Zhang, Timothy T Harkins, and Stephan Zchner. Exome sequencing of a multigenerational human pedigree. *PLoS One*, 4(12):e8232, 2009.

[39] S. E. Hile and K. A. Eckert. Dna polymerase kappa produces interrupted mutations and displays polar pausing within mononucleotide microsatellite sequences. *Nucleic Acids Res*, 36(2):688–96–, 2008.

[40] S. M. Huse, J. A. Huber, H. G. Morrison, M. L. Sogin, and D. M. Welch. Accuracy and quality of massively parallel dna pyrosequencing. *Genome Biol*, 8(7):R143–, 2007.

[41] D. G. Hwang and P. Green. Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A*, 101(39):13994–4001, 2004. eng.

[42] Ravi R. Iyer, Anna Pluciennik, Vickers Burdett, and Paul L. Modrich. Dna mismatch repair: functions and mechanisms. *Chemical Reviews*, 106(2):302–323, 2005.

[43] Ravi R Iyer, Anna Pluciennik, Vickers Burdett, and Paul L Modrich. Dna mismatch repair: functions and mechanisms. *Chem Rev*, 106(2):302–323, Feb 2006.

[44] F. A Kadyrov, S. F Holmes, M. E Arana, O. A Lukianova, M O'donnell, T. A Kunkel, and P Modrich. Saccharomyces cerevisiae mutl is a mismatch repair endonuclease. *Journal of Biological Chemistry*, 282(51):37181–37190, 2007.

[45] Farid A Kadyrov, Leonid Dzantiev, Nicoleta Constantin, and Paul Modrich. Endonucleolytic function of mutl in human mismatch repair. *Cell*, 126(2):297–308, 2006.

[46] P. D Keightley, U Trivedi, M Thomson, F Oliver, S Kumar, and M. L Blaxter. Analysis of the genome sequences of three drosophila melanogaster spontaneous mutation accumulation lines. *Genome Research*, 19(7):1195–1201, 2009.

[47] N. Kim, A. L. Abdulovic, R. Gealy, M. J. Lippert, and S. Jinks-Robertson. Transcription-associated mutagenesis in yeast is directly proportional to the level of gene expression and influenced by the direction of dna replication. *DNA Repair (Amst)*, 6(9):1285–96–, 2007.

[48] T. A. Kunkel and D. A. Erie. Dna mismatch repair. *Annu Rev Biochem*, 74:681–710, 2005. eng.

[49] William Lee, Zhaoshi Jiang, Jinfeng Liu, Peter M Haverty, Yinghui Guan, Jeremy Stinson, Peng Yue, Yan Zhang, Krishna P Pant, Deepali Bhatt, Connie Ha, Stephanie Johnson, Michael I Kennemer, Sankar Mohan, Igor Nazarenko, Colin Watanabe, Andrew B Sparks, David S Shames, Robert Gentleman, Frederic J de Sauvage, Howard Stern, Ajay Pandita, Dennis G Ballinger, Radoje Drmanac, Zora Modrusan, Somasekar Seshagiri, and Zemin Zhang. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature*, 465(7297):473–477, May 2010.

[50] H. Li and N. Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform*, 11(5):473–83, 2010. eng.

[51] Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform*, 11(5):473–483, Sep 2010.

[52] Heng Li, Jue Ruan, and Richard Durbin. Mapping short dna sequenc-

ing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–1858, Nov 2008.

[53] Ruiqiang Li, Wei Fan, Geng Tian, Hongmei Zhu, Lin He, Jing Cai, Quanfei Huang, Qingle Cai, Bo Li, Yinqi Bai, Zhihe Zhang, Yaping Zhang, Wen Wang, Jun Li, Fuwen Wei, Heng Li, Min Jian, Jianwen Li, Zhaolei Zhang, Rasmus Nielsen, Dawei Li, Wanjun Gu, Zhentao Yang, Zhaoling Xuan, Oliver A. Ryder, Frederick Chi-Ching Leung, Yan Zhou, Jianjun Cao, Xiao Sun, Yonggui Fu, Xiaodong Fang, Xiaosen Guo, Bo Wang, Rong Hou, Fujun Shen, Bo Mu, Peixiang Ni, Runmao Lin, Wubin Qian, Guodong Wang, Chang Yu, Wenhui Nie, Jinhuan Wang, Zhigang Wu, Huiqing Liang, Jiumeng Min, Qi Wu, Shifeng Cheng, Jue Ruan, Mingwei Wang, Zhongbin Shi, Ming Wen, Binghang Liu, Xiaoli Ren, Huisong Zheng, Dong Dong, Kathleen Cook, Gao Shan, Hao Zhang, Carolin Kosiol, Xueying Xie, Zuhong Lu, Hancheng Zheng, Yingrui Li, Cynthia C. Steiner, Tommy Tsan-Yuk Lam, Siyuan Lin, Qinghui Zhang, Guoqing Li, Jing Tian, Timing Gong, Hongde Liu, Dejin Zhang, Lin Fang, Chen Ye, Juanbin Zhang, Wenbo Hu, Anlong Xu, Yuanyuan Ren, Guojie Zhang, Michael W. Bruford, Qibin Li, Lijia Ma, Yiran Guo, Na An, Yujie Hu, Yang Zheng, Yongyong Shi, Zhiqiang Li, Qing Liu, Yanling Chen, Jing Zhao, Ning Qu, Shancen Zhao, Feng Tian, Xiaoling Wang, Haiyin Wang, Lizhi Xu, Xiao Liu, Tomas Vinar, Yajun Wang, Tak-Wah Lam, Siu-Ming Yiu, Shiping Liu, Hemin Zhang, Desheng Li, Yan Huang, Xia Wang, Guohua Yang, Zhi Jiang, Junyi Wang, Nan Qin, Li Li, Jingxiang Li, Lars Bolund, Karsten Kristiansen, Gane Ka-Shu Wong, Maynard Olson, Xiuqing Zhang, Songgang Li, Huanming Yang, Jian Wang, and Jun Wang. The sequence and de novo assembly of the giant panda genome. *Nature*, 463(7279):311–317, January 2010.

[54] Ruiqiang Li, Yingrui Li, Xiaodong Fang, Huanming Yang, Jian Wang, Karsten Kristiansen, and Jun Wang. Snp detection for massively parallel whole-genome resequencing. *Genome Res*, 19(6):1124–1132, Jun 2009.

[55] Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, Aug 2009.

[56] W. H. Li and L. A. Sadler. Low nucleotide diversity in man. *Genetics*, 129(2):513–23–, 1991.

[57] Y. Li, N. Vinckenbosch, G. Tian, E. Huerta-Sanchez, T. Jiang, H. Jiang, A. Albrechtsen, G. Andersen, H. Cao, T. Korneliussen, N. Grarup, Y. Guo, I. Hellman, X. Jin, Q. Li, J. Liu, X. Liu, T. Sparso, M. Tang, H. Wu, R. Wu,

C. Yu, H. Zheng, A. Astrup, L. Bolund, J. Holmkvist, T. Jorgensen, K. Kristiansen, O. Schmitz, T. W. Schwartz, X. Zhang, R. Li, H. Yang, J. Wang, T. Hansen, O. Pedersen, and R. Nielsen. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet*, 42(11):969–72, 2010. eng.

[58] Ryan Lister, Brian D Gregory, and Joseph R Ecker. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Curr Opin Plant Biol*, 12(2):107–118, Apr 2009.

[59] Gianni Liti, David M Carter, Alan M Moses, Jonas Warringer, Leopold Parts, Stephen A James, Robert P Davey, Ian N Roberts, Austin Burt, Vassiliki Koufopanou, Isheng J Tsai, Casey M Bergman, Douda Bensasson, Michael J T O—[Rsquo]—Kelly, Alexander van Oudenaarden, David B H Barton, Elizabeth Bailes, Alex N Nguyen, Matthew Jones, Michael A Quail, Ian Goodhead, Sarah Sims, Frances Smith, Anders Blomberg, Richard Durbin, and Edward J Louis. Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337, 2009.

[60] D. P. Locke, L. W. Hillier, W. C. Warren, K. C. Worley, L. V. Nazareth, D. M. Muzny, S. P. Yang, Z. Wang, A. T. Chinwalla, P. Minx, M. Mitreva, L. Cook, K. D. Delehaunty, C. Fronick, H. Schmidt, L. A. Fulton, R. S. Fulton, J. O. Nelson, V. Magrini, C. Pohl, T. A. Graves, C. Markovic, A. Cree, H. H. Dinh, J. Hume, C. L. Kovar, G. R. Fowler, G. Lunter, S. Meader, A. Heger, C. P. Ponting, T. Marques-Bonet, C. Alkan, L. Chen, Z. Cheng, J. M. Kidd, E. E. Eichler, S. White, S. Searle, A. J. Vilella, Y. Chen, P. Flicek, J. Ma, B. Raney, B. Suh, R. Burhans, J. Herrero, D. Haussler, R. Faria, O. Fernando, F. Darre, D. Farre, E. Gazave, M. Oliva, A. Navarro, R. Roberto, O. Capozzi, N. Archidiacono, G. Della Valle, S. Purgato, M. Rocchi, M. K. Konkel, J. A. Walker, B. Ullmer, M. A. Batzer, A. F. Smit, R. Hubley, C. Casola, D. R. Schrider, M. W. Hahn, V. Quesada, X. S. Puente, G. R. Ordonez, C. Lopez-Otin, T. Vinar, B. Brejova, A. Ratan, R. S. Harris, W. Miller, C. Kosiol, H. A. Lawson, V. Taliwal, A. L. Martins, A. Siepel, A. Roychoudhury, X. Ma, J. Degenhardt, C. D. Bustamante, R. N. Gutenkunst, T. Mailund, J. Y. Dutheil, A. Hobolth, M. H. Schierup, O. A. Ryder, Y. Yoshinaga, P. J. de Jong, G. M. Weinstock, J. Rogers, E. R. Mardis, and R. A. Gibbs. Comparative and demographic analysis of orang-utan genomes. *Nature*, 469(7331):529–33, 2011. eng.

[61] L. A. Loeb. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nat Rev Cancer*, 11(6):450–7–, 2011.

[62] K. E. Lohmueller, A. R. Indap, S. Schmidt, A. R. Boyko, R. D. Hernandez,

M. J. Hubisz, J. J. Sninsky, T. J. White, S. R. Sunyaev, R. Nielsen, A. G. Clark, and C. D. Bustamante. Proportionally more deleterious genetic variation in european than in african populations. *Nature*, 451(7181):994–7, 2008. eng.

[63] Susan T Lovett. Polymerase switching in dna replication. *Mol Cell*, 27(4):523–526, Aug 2007.

[64] H. T. Lynch, P. M. Lynch, S. J. Lanspa, C. L. Snyder, J. F. Lynch, and C. R. Boland. Review of the lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clin Genet*, 76(1):1–18, 2009. eng.

[65] M. Lynch, W. Sung, K. Morris, N. Coffey, C. R. Landry, E. B. Dopman, W. J. Dickinson, K. Okamoto, S. Kulkarni, D. L. Hartl, and W. K. Thomas. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A*, 105(27):9272–7, 2008. eng.

[66] G. T. Marsischky, N. Filosi, M. F. Kane, and R. Kolodner. Redundancy of saccharomyces cerevisiae msh3 and msh6 in msh2-dependent mismatch repair. *Genes Dev*, 10(4):407–20, 1996. eng.

[67] G. Matassi, P. M. Sharp, and C. Gautier. Chromosomal location effects on gene sequence evolution in mammals. *Curr Biol*, 9(15):786–91, 1999. eng.

[68] S. D. McCulloch and T. A. Kunkel. The fidelity of dna synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res*, 18(1):148–61, 2008. eng.

[69] E. Meijaard and S. Wich. Putting orang-utan population trends into perspective. *Curr Biol*, 17(14):R540–, 2007.

[70] Michael L Metzker. Emerging technologies in dna sequencing. *Genome Res*, 15(12):1767–1776, Dec 2005.

[71] Matthew Meyerson, Stacey Gabriel, and Gad Getz. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*, 11(10):685–696, Oct 2010.

[72] P. Modrich and R. Lahue. Mismatch repair in replication fidelity, genetic recombination, and cancer biology. *Annu Rev Biochem*, 65:101–33, 1996. eng.

[73] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, 320(5881):1344–1349, June 2008.

[74] A. Nater, P. Nietlisbach, N. Arora, C. P. van Schaik, M. A. van Noordwijk, E. P. Willems, I. Singleton, S. A. Wich, B. Goossens, K. S. Warren, E. J. Verschoor, D. Perwitasari-Farajallah, J. Pamungkas, and M. Krutzen. Sex-biased dispersal and volcanic activities shaped phylogeographic patterns of extant orangutans (genus: Pongo). *Mol Biol Evol*, 2011. Eng.

[75] Sarah B Ng, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, Jay Shendure, and Michael J Bamshad. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet*, 42(1):30–35, January 2010.

[76] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443, 2011.

[77] K. T. Nishant, N. D. Singh, and E. Alani. Genomic mutation rates: what high-throughput methods can tell us. *Bioessays*, 31(9):912–20, 2009. eng.

[78] K. T. Nishant, W. Wei, E. Mancera, J. L. Argueso, A. Schlattl, N. Delhomme, X. Ma, C. D. Bustamante, J. O. Korbel, Z. Gu, L. M. Steinmetz, and E. Alani. The baker's yeast diploid genome is remarkably stable in vegetative growth and meiosis. *PLoS Genet*, 6(9), 2010. eng.

[79] Chandra Shekhar Pareek, Rafal Smoczynski, and Andrzej Tretyn. Sequencing technologies and genome sequencing. *J Appl Genet*, Jun 2011.

[80] D. Posada. jmodeltest: phylogenetic model averaging. *Mol Biol Evol*, 25(7):1253–6, 2008. eng.

[81] I. B. Rogozin, B. A. Malyarchuk, Y. I. Pavlov, and L. Milanesi. From context-dependence of mutations to molecular mechanisms of mutagenesis. *Pac Symp Biocomput*, pages 409–20–, 2005.

[82] E. Segal and J. Widom. Poly(da:dt) tracts: major determinants of nucleosome organization. *Curr Opin Struct Biol*, 19(1):65–71–, 2009.

[83] JS Sekhon. Multivariate and propensity score matching software with automated balance optimization: The matching package for r. *Journal of Statistical Software*, 42(7):1–52, 6 2011.

[84] S. N. Shah, S. E. Hile, and K. A. Eckert. Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. *Cancer Res*, 70(2):431–5–, 2010.

[85] E. A. Sia, R. J. Kokoska, M. Dominska, P. Greenwell, and T. D. Petes. Microsatellite instability in yeast: dependence on repeat unit size and dna mismatch repair genes. *Mol Cell Biol*, 17(5):2851–8, 1997. eng.

[86] Wich S. A. Husson S. Atmoko S. U. Leighton M. Rosen N. Traylor-Holzer K. Lacy R. Byers O. Singleton, I. Orangutan population and habitat viability assessment: Final report. Technical report, IUCN/SSC Conservation Breeding Specialist Group, Apple Valley, MN, USA, 2004.

[87] M. Snyder, A. R. Buchman, and R. W. Davis. Bent dna at a yeast autonomously replicating sequence. *Nature*, 324(6092):87–9–, 1986.

[88] J. A. Stamatoyannopoulos, I. Adzhubei, R. E. Thurman, G. V. Kryukov, S. M. Mirkin, and S. R. Sunyaev. Human mutation rate associated with dna replication timing. *Nat Genet*, 41(4):393–5, 2009. eng.

[89] G. Streisinger, Y. Okada, J. Emrich, J. Newton, A. Tsugita, E. Terzaghi, and M. Inouye. Frameshift mutations and the genetic code. this paper is dedicated to professor theodosius dobzhansky on the occasion of his 66th birthday. *Cold Spring Harb Symp Quant Biol*, 31:77–84–, 1966.

[90] K. Struhl. Naturally occurring poly(da-dt) sequences are upstream promoter elements for constitutive transcription in yeast. *Proc Natl Acad Sci U S A*, 82(24):8419–23–, 1985.

[91] F. Taddei, M. Radman, J. Maynard-Smith, B. Toupance, P. H. Gouyon, and B. Godelle. Role of mutator alleles in adaptive evolution. *Nature*, 387(6634):700–2–, 1997.

[92] L. Teytelman, M. B. Eisen, and J. Rine. Silent but not static: accelerated base-pair substitution in silenced chromatin of budding yeasts. *PLoS Genet*, 4(11):e1000247, 2008. eng.

[93] H. T. Tran, J. D. Keen, M. Kricker, M. A. Resnick, and D. A. Gordenin. Hy-

permutability of homonucleotide runs in mismatch repair and dna polymerase proofreading yeast mutants. *Mol Cell Biol*, 17(5):2859–65, 1997. eng.

[94] P. T. Tran, J. A. Simon, and R. M. Liskay. Interactions of exo1p with components of mutlalpha in saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, 98(17):9760–5, 2001. eng.

[95] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech*, 28(5):511–515, May 2010.

[96] M. A. van Noordwijk and C. P. van Schaik. Development of ecological competence in sumatran orangutans. *Am J Phys Anthropol*, 127(1):79–94–, 2005.

[97] C. P. van Schaik, M. Ancrenaz, G. Borgen, B. Galdikas, C. D. Knott, I. Singleton, A. Suzuki, S. S. Utami, and M. Merrill. Orangutan cultures and the evolution of material culture. *Science*, 299(5603):102–5–, 2003.

[98] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry,

S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guig, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001.

[99] K. S. Warren, E. J. Verschoor, S. Langenhuijzen, Heriyanto, R. A. Swan, L. Vigilant, and J. L. Heeney. Speciation and intrasubspecific variation of bornean orangutans, pongo pygmaeus pygmaeus. *Mol Biol Evol*, 18(4):472–80, 2001. eng.

[100] S. Washietl, R. Machne, and N. Goldman. Evolutionary footprints of nucleosome positions in yeast. *Trends Genet*, 24(12):583–7, 2008. eng.

[101] G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, 7(2):256–76–, 1975.

[102] W. Wei, J. H. McCusker, R. W. Hyman, T. Jones, Y. Ning, Z. Cao, Z. Gu, D. Bruno, M. Miranda, M. Nguyen, J. Wilhelmy, C. Komp, R. Tamse, X. Wang, P. Jia, P. Luedi, P. J. Oefner, L. David, F. S. Dietrich, Y. Li, R. W. Davis, and L. M. Steinmetz. Genome sequencing and comparative analysis of saccharomyces cerevisiae strain yjm789. *Proc Natl Acad Sci U S A*, 104(31):12825–30, 2007. eng.

[103] K. H. Wolfe, P. M. Sharp, and W. H. Li. Mutation rates differ among regions of the mammalian genome. *Nature*, 337(6204):283–5, 1989. eng.

[104] B. E. Wright. A biochemical mechanism for nonrandom mutations and evolution. *J Bacteriol*, 182(11):2993–3001–, 2000.

[105] W. Xu, J. G. Aparicio, O. M. Aparicio, and S. Tavare. Genome-wide mapping of orc and mcm2p binding sites on tiling arrays and identification of essential ars consensus sequences in s. cerevisiae. *BMC Genomics*, 7:276, 2006. eng.

[106] S. Zanders, X. Ma, A. Roychoudhury, R. D. Hernandez, A. Demogines, B. Barker, Z. Gu, C. D. Bustamante, and E. Alani. Detection of heterozygous mutations in the genome of mismatch repair defective diploid yeast using a bayesian approach. *Genetics*, 186(2):493–503, 2010. eng.

[107] D. R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res*, 18(5):821–9, 2008. eng.

[108] Jun Zhang, Rod Chiodini, Ahmed Badr, and Genfa Zhang. The impact of next-generation sequencing on genomics. *J Genet Genomics*, 38(3):95–109, Mar 2011.