DETECTING SIGNATURES OF NATURAL SELECTION IN GENETIC DATA

BY AATISH BHATIA

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Physics and Astronomy

Written under the direction of

Dr. Gyan Bhanot

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2013

ABSTRACT OF THE DISSERTATION

Detecting Signatures of Natural Selection in Genetic Data

by Aatish Bhatia

Dissertation Director: Dr. Gyan Bhanot

I report on three studies where I identify signatures of natural selection in humans, and dissect the genetic architecture of complex phenotypic traits in yeast. In chapter 2, I discuss the results of a quantitative trait mapping study, where we showed that yeast growth can be characterized by multiple biologically-relevant growth parameters obtained by fitting yeast growth OD data to a sigmoid function. We identified quantitative trait loci (QTL) and gene-gene interactions driving variation in these yeast growth parameters. We analyzed the environment dependence of these QTLs and gene-gene interactions, and identified a common gene, *FLO8*, which interacts with other genes in an environment specific fashion to affect distinct growth phenotypes. In chapter 3, I describe our published study where we applied quantitative trait locus mapping to wildtype yeast strains, and identified linked clusters of genetic variants that contributed to variation in the sporulation efficiency of these strains. In chapter 4, I describe our work on identifying signatures of natural selection in the human lineage, specifically in the Maasai people in East Africa. Our work suggests that the Maasai have undergone recent diet induced positive natural selection that may confer protection against hyperlipidemia and cardiac diseases.

Acknowledgements

At the time of writing this document, the work described in chapter 2 is being prepared for publication. The data was collected by Chenchen Zhu, Manu M. Tekkedil and Lars Steinmetz at the European Molecular Biology Laboratory, Heidelberg, Germany, and Julien Gagneur at the Department of Chemistry and Biochemistry, Ludwig-Maximilians-Universität München, Munich, Germany. The work described in this chapter is a collaboration with these authors and with Anupama Yadav and Himanshu Sinha at the Tata Institute of Fundamental Research, Mumbai, India. I conducted all the data analysis in this chapter. I co-wrote the paper with Anupama Yadav, Himanshu Sinha and Gyan Bhanot, with additional input from Julien Gagneur and Lars Steinmetz.

The analysis described in chapter 3 is published in PLOS One [1]. I am joint first author in this publication with Parul Tomar. In this study, I was responsible for the LOD score analysis, subsequent p-value computations and clustering of SNPs based on linkage disequilibrium. Liyang Diao performed the population structure corrections, Gyan Bhanot performed the binomial analysis, and Shweta Ramdas and Parul Tomar performed SNP data annotation. Parul Tomar collected the experimental data. I co-wrote the paper with Parul Tomar, Shweta Ramdas, Himanshu Sinha and Gyan Bhanot.

The analysis described in chapter 4 is published in PLOS One [2]. I am joint first author in this publication with Kshitij Wagh. I co-wrote the paper with Kshitij Wagh and Gyan Bhanot. I was responsible for the Structure computation, F_{ST} computation, clustering of SNPs based on linkage disequilibrium, SNP annotation, and I assisted with the iHS and XP-EHH computations. The lab-work required for sequencing was performed by myself, Kshitij Wagh, Vijay Ravikumar, and Michael Boemo, under the guidance of Ming Yao and Shridar Ganesan at the Cancer Institute of New Jersey. This

work benefited from numerous conversations with Asad Naqvi and Sergio Lukic at the Institute of Advanced Study.

I would like to thank the Rutgers University Physics Department, my Ph.D. thesis committee, and especially Ronald Ransome, for their patience, encouragement and support in allowing me to switch fields into quantitative genetics and population genetics during my graduate studies. I would also like to acknowledge the financial support provided by the American Physical Society and the Indo-U.S. Science and Technology Forum, and the generous support and accommodation provided by the Tata Institute of Fundamental Research. I am also very grateful to Anupama Yadav, Parul Tomar and Himanshu Sinha at the Tata Institute of Fundamental Research, and to Marton Toth and Monica Driscoll in the Department of Molecular Biology and Biochemistry, Rutgers University for their patience, time and support in valued experimental collaborations. I would like to thank Ming Yao and Shridar Ganesan at the Cancer Institute of New Jersey for their guidance, and for allowing someone who had never held a pipette to work in their lab. I would like to thank everyone who aided, abetted, supported and encouraged my forays into physics education and science outreach: Michael Gentile, Deepak Iyer, Darakhshan Mir, Jean Patrick Antoine, Saurabh Jha, Mohan Kalelkar, Suzanne Brahmia, David Maiullo, Ronald Ransome, Michael Manhart, Simon Knapen, Manjul Apratim, Stephanie Wortel, Meghan Groome, John Boccio and Carl Grossman. I have learnt a lot from every one of you. I'm also thankful to Chef Rachel Weston and her friendly crew at the Better World Cafe, whose excellent meals got me through graduate school, and to Kshitij Wagh for his constant source of ideas and support. And finally, I owe a debt of gratitude to Gyan Bhanot for his encouragement, guidance and support.

Dedication

To my parents, who put my education above all else, and to my paternal grandfather, whose last words to me in person were, "*Prove something wrong.*"

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The fundamental forces that shape the genome

Natural selection is the primary mechanism through which organisms become better adapted to their environment. The vast majority of random mutations have either a deleterious or a neutral effect on the reproductive success of an organism. Occasionally, a mutation, or a series of mutations arise, which enhance the ability of an organism to pass on copies of its genes to future generations. It is these rare events that make organisms better adapted to their environment.

Selection pressures are different for different organisms, depending on their life cycles, the ecological niches they occupy, their method of reproduction, and the timescales over which biological forces alter their genomes. Neutral forces such as mutation, recombination and genetic drift (discussed below) alter the genome over long timescales. On the other hand, selection can act rapidly. These forces create a complex interplay of timescales over which must be carefully analyzed to detect signatures of natural selection.

The overall question we begin to address in this thesis is: how can we identify genetic signatures of adaptation in genetic data?

In this thesis, I restrict my attention to eukaryotes, specifically to yeast (*S. cerevisiae*) and humans. In natural populations such as human groups, it is difficult to isolate the genetic loci underlying complex traits [3], [4], [5]. Our understanding of eukaryote evolution has therefore greatly benefited by quantitative genetic studies conducted in experimental populations of yeast [6]. The abundance of phenotype data and high-resolution genotype data has led to genomic scans and experimental methods that

can precisely dissect the genomic architecture of complex phenotypes [3], [7]. Furthermore, we can identify the gene and interactions governing complex traits with a high degree of accuracy. In chapters 2 and 3, I describe my work identifying genetic loci, as well as gene-gene interactions and gene-environment interactions, that contribute towards complex quantitative phenotypic traits in yeast. In Chapter 4, I describe my work identifying signals of natural selection in humans, specifically in the members of the Maasai population in East Africa.

A defining aspect of life is that organisms can pass on their genomes, with modification, to the next generation. Darwin used the term "descent with modification" to describe this phenomenon. Today we understand the molecular basis by which DNA is translated into proteins. However, the mapping from DNA variants to phenotypic changes is highly complex and there are notably few instances where the presence or absence of a single mutation determines a phenotype. Complex traits, such as height or disease risk in humans, or growth and aging in yeast, are regulated by the interaction of multiple genes, each of which may have a small effect on the overall phenotype [8], [9], [10], [11]. The phenotypic effect of a gene can also depend on the genotypes at many other genes (the genetic background). And lastly, the effect of a gene, and its interactions with other genes, are highly dependent on the environmental context.

Experimental techniques in molecular biology such as knock-down gene studies and genome-wide associations studies allow us to dissect the genetic basis of such complex phenotypes. In particular, a host of statistical tools under the category of *quantitative trait mapping* uses the genotype and phenotype data of a large number of individuals to identify causal genetic variants that drive a change in phenotype [12], [13].

In chapter 2, I report on the results of such a mapping study, where we show that yeast growth can be characterized by three different growth parameters, which we expect to be regulated by a number of distinct genetic loci. Using sequence and growth data for different carbon sources in crosses between highly divergent strains of yeast, we identified the genes and gene-gene interactions that drive variation in these growth parameters. By mapping these traits in a variety of environmental conditions, we analyzed the environment dependence of these genes and gene-gene interactions,

and identified specific genes that interact in an environment-specific manner to alter growth phenotypes.

In chapter 3, I describe our published study where we applied quantitative trait mapping to wild-type yeast strains (i.e. those isolated from the wild as opposed to clinical or lab-adapted isolates), and identified linked clusters of genetic loci which contribute to variation in sporulation efficiency among these strains [1]. Such an analysis, which attempts to map quantitative traits using a limited set of wildtype strains, is quite challenging. One specific challenge is the need to correct for population structure. In this study, we addressed this using some recently developed statistical tools to correct for the bias from the relatedness (population substructure) of the yeast samples.

In chapter 4, I discuss our published work [2] on identifying diet induced selection in humans. Compared to yeast genetics, the study of natural selection in humans requires us to address a completely different set of issues, which are specific to multicellular, non-clonal, sexually reproducing organisms. The widespread availability of genome-wide sequencing and large-scale polymorphism data on diverse human populations allows us to investigate selection at a finer resolution than previously possible [14], [15], [16], [17], [18], [19]. To understand the challenges involved in detecting natural selection in humans, and to understand the statistical methods to identify these signals, we first need to understand the evolutionary selection pressures that shape the genome.

In the rest of the introduction, we give a brief description of the four fundamental 'forces' that shape the genomes of organisms, and then describe how to use this understanding to detect selection. Using the understanding of these major forces, we will develop the rationale for a fitness landscape, where multiple genes interact and drive phenotypes in an environment dependent fashion. Finally, we describe how we can use this overall picture to identify quantitative trait loci by dissecting the genetic basis of complex phenotypes, in single celled eukaryotes in chapters 2 and 3 and in multicellular eukaryotes in chapter 4.

1.1.1 The first force: Mutations: increasing genetic diversity

In eukaryotes, mutations are local genomic changes that arise from a variety of effects, such as copying/editing errors in DNA replication, effects of line elements, errors in DNA damage repair, etc. The most common type of mutations are *point mutations*, which change a single letter of DNA. A single DNA position (nucleotide) that has two or more forms circulating in a population is known as a *single nucleotide polymorphism* (SNP). Mutations can also present themselves as abnormal numbers of repeats of a DNA sequence, known as copy number variation (CNV). These manifest as insertions or deletions, and occur mainly due to errors in DNA replication, particularly while copying highly repetitive DNA sequences. Insertions can also be caused by self-replicating transposable elements.

Mutations can be classified on the basis of their heritability. Somatic mutations are those that occur within an individual lifetime, but which cannot be passed onto the next generation. Many cancers are the result of such somatic mutations. On the other hand, germ-line mutations are mutations that can be passed onto the next generation. In multicellular, sexually reproducing species, these are mutations in sex cells - sperm or egg cells. In single celled organisms, somatic mutations are generally the same as germ-line mutations (exceptions are cases where there is asymmetric segregation of DNA errors in cell division).

It is often said that mutations occur at random. This statement is true in the following limited sense: mutations occur without any regard to their potential contribution to the fitness of an organism. However, due to the biochemical properties of DNA, certain mutations are more likely to occur. For example, the most common point mutations are C $\leftrightarrow$ T or A $\leftrightarrow$ G transitions [20]. This is because the amino acids C/T (one-ring pyrimidines) and A/G (two-ring purines) are structurally similar. Furthermore, the probability of mutation is not uniform on the genome, and genomes have mutational *hot spots* where the mutation rate can be an order of magnitude higher than in mutational *cold spots* [20].

Of the point mutations that fall within genes, *synonymous mutations* are those that

do not change the protein sequence. This is possible because the genetic code is degenerate, with many of the amino acids encoded by several triplets of bases. On the other hand, *non-synonymous mutations* are those that do alter the amino acid, and hence affect the overall protein sequence. These altered proteins can then have functional consequences for the organism. For example, a non-synonymous mutation in the genes *BRCA1* or *BRCA2* significantly increases an individual's chances of developing breast cancer [21]. Indeed, non-synonymous mutations are the most common way in which an altered genotype brings about a change in phenotype. We will identify and discuss many such protein-altering mutations in each of the chapters that follow.

Assume for simplicity that mutations are the only force shaping the genome. If a segment of DNA has a local mutation rate μ per base per generation, then the frequency p of this variant will change only when a mutation occurs. Hence,

$$p' = (1 - \mu)p$$

where p and p' are the frequencies of a particular variant (allele) at this locus in successive generations.

The change in frequency in a single generation is

$$\Delta p = -\mu p$$

As μ is typically very small, we can justify taking the continuum limit:

$$\frac{dp}{dt} = -\mu p$$

The solution of this equation is a familiar exponential decay, i.e. $p = p_0 e^{-\mu t}$, where t is the number of generations that have elapsed. Hence, the timescale over which mutation brings about significant changes is allele frequencies is given by $t_\mu \approx \frac{1}{\mu}$.

For humans, the mutation rate is approximately $\mu = 2.5 \times 10^{-8}$ per nucleotide per generation [20]. An average human gene is on the order of 10 kilobases in size (10 kb = 10,000 nucleotides) [22]. The occurrence of mutations within a gene can be modeled as a Poisson process. The probability of at least one mutation occurring in such a gene in a single generation is given by $1 - e^{-\mu d}$ where $d = 10,000$ is the gene size (note

that $1 - e^{-\mu d} \approx \mu d$ for small μd). Thus, for an average gene in humans, the mutation rate is $\mu_{gene} \approx 2.5 \times 10^{-4}$. The timescale corresponding to this ($t_\mu = \frac{1}{\mu_{gene}}$) is about 4,000 generations, or somewhere between 80,000 to 100,000 years. This is the timescale over which neutral mutations build diversity in a 10 kb gene in the absence of other effects. However, due to genetic drift (discussed below), the timescale over which such mutations will attain a reasonable frequency in the population is even longer.

1.1.2 The second force: Recombination: breaking down genetic correlations

In sexually reproducing organisms, the process known as genetic recombination shuffles the genomes of the parents in their offspring. Recombination allows beneficial mutations occurring in different lineages to rapidly combined in an offspring. Furthermore, recombinations allows for genes to occur in new combinations, which can lead to a fitness advantage. Indeed, from the perspective of evolution, the primary benefit of sex is to dramatically increase the pace of adaptation, as sex provides a mechanism to 'mix and match' beneficial mutations that have evolved in separate lineages. In contrast, asexual populations can never combine beneficial mutations occurring in different individuals. Instead, every beneficial mutation in asexual reproduction must arise independently in the same lineage, which significantly slows down the pace of adaptation. In chapter 2, we identify situations where recombination between genetically divergent strains of yeast leads to a phenotypic change in growth because mutations from distinct parental lineages interact in the offspring. Studying growth parameters in these hybrid strains allows us to identify functional loci associated with the growth phenotype.

Somatic cells in diploid organisms such as humans have two copies of each chromosome - a paternally derived copy and a maternally derived copy. On the other hand, sex cells (sperm or egg) are haploid, and contain only a single set of chromosomes. When an individual produces sex cells in early embryogenesis, their paternally and maternally inherited chromosomes are pairwise aligned and shuffled at several points, so that each sex cell is a different reshuffling of their chromosomes. This is the process of recombination, through which each individual passes on to their progeny a haploid mosaic

of their own paternally and maternally inherited DNA. Each of our chromosomes is a shuffled copy of our grandparents chromosomes, one from the maternal grandparents and the other from the paternal grandparents.

While mutations are a constant source of new genomic changes, recombination can spread these changes across lineages, by shuffling the genomes of unrelated individuals. These forces increase the pace of adaptation in a population, and also lead to an increase in genetic diversity. One of the key challenges of detecting signs of natural selection is to identify the genetic fingerprints of selection in the past, which have been obscured by generations of subsequent mutation and chromosomal shuffling. In chapter 4 we discuss the methods that we used to detect such signatures in humans.

To understand how recombination shuffles the genome, consider a pair of bi-allelic loci, X and Y on the same chromosome. At each locus, we have two alleles, a wild type allele and a mutant allele. We denote these as 0 and 1 respectively. Let $2N$ be the sample size (twice the number of individuals for a diploid species). The degree to which the alleles at these two loci are correlated in the population is a measure of the amount of recombination that has occurred. To capture this effect, we define a quantity D as follows:

$$\begin{aligned} D &= \langle XY \rangle - \langle X \rangle \langle Y \rangle \\ &= \frac{1}{2N} \sum_i x_i y_i - \frac{1}{2N} \sum_i x_i \frac{1}{2N} \sum_i y_i \\ &= P_{XY} - P_X P_Y \end{aligned}$$

Here, P_X and P_Y denote the frequency of the mutant genotype at X and Y, and P_{XY} denotes the frequency of the double mutant individuals at X and Y. In the population genetics literature, the quantity D is known as the linkage disequilibrium coefficient, and the degree to which it deviates from zero is a measure of the relatedness of the two loci. D captures the extent to which the observed two-locus frequencies are

dependent. It is simply related to the Pearson correlation coefficient r as follows:

$$\begin{aligned} r &= \frac{\langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle}{\sigma_X \sigma_Y} \\ &= \frac{\langle XY \rangle - \langle X \rangle \langle Y \rangle}{\sigma_X \sigma_Y} \\ &= \frac{D}{\sqrt{P_X(1-P_X)P_Y(1-P_Y)}} \end{aligned}$$

In asexual species, the population consists of clones of the previous generation. Hence, apart from the effects of mutations, the correlation between different loci on the same chromosome is very high. The extreme case of perfect correlation is known as perfect linkage disequilibrium.

In contrast, in a sexually reproducing population, recombination will work to decouple loci over time. If the probability of a recombination event occurring between the two loci in a single generation is r, then the probability that the two loci will remain linked is $1 - r$. In the remaining fraction r of the time, recombination shuffles the two loci. Then, assuming random mating, the probability of arriving at the mutant genotype at both loci is rP_XP_Y. Thus, the expected frequency of individuals with the double mutant genotype in the next generation is the sum of these two effects. The first effect reduces the probability and the second increases it. Thus,

$$\begin{aligned} P'_{XY} &= (1-r)P_{XY} + rP_XP_Y \\ P'_{XY} - P_XP_Y &= (1-r)(P_{XY} - P_XP_Y) \\ D' &= (1-r)D \end{aligned}$$

Here the prime indicates the next generation. We have assumed that the allele frequencies remain constant in the next generation. This is typically the case for large population sizes, assuming random mating and a sufficiently small mutation rate and selection pressure.

Over t generations, for small r,

$$D_t = (1-r)^t D_0 \approx e^{-rt} D_0$$

Hence linkage disequilibrium decays exponentially in time. In particular, we can

define a timescale $t_r \approx \frac{1}{r}$ over which recombination breaks down linkage between alleles. As we will see, recombination erodes the genetic signature of selection over these timescales. The accumulation of new mutations and the shuffling of chromosomes due to recombination are the major forces that obscure signatures of positive selection which occurred in the past.

1.1.3 The third force: Genetic Drift: finite population size reduces genetic diversity.

Genetic drift is the stochastic variation in allele frequencies that arises due to a finite population size. Given a population size of N individuals, mating is modeled as the random union of 2N alleles (the factor of two accounts for diploid populations). This introduces a stochastic variation in allele frequencies from generation to generation.

If in a given generation the allele frequency at a given locus is p, in the next generation, the probability of sampling exactly i of the same allele out of the $2N$ alleles available, is given by the binomial distribution:

$$P(i) = \binom{2N}{i} p^i (1-p)^{2N-i}$$

i is a binomially distributed variable, and therefore has mean $\langle i \rangle = 2Np$ and variance $\sigma_i^2 = 2Np(1-p)$. Therefore, $\langle p' \rangle = \frac{\langle i \rangle}{2N} = p$ and its variance is $\sigma_{p'}^2 = \sigma_i^2/(2N)^2 = p(1-p)/2N$.

The heterozygosity of a population is defined as the probability that two randomly chosen alleles are different. Genetic drift causes the heterozygosity to decay over time. We can demonstrate this by calculating the expected value of heterozygosity after a

single generation of mating.

$$\begin{aligned} H &= \langle 2p'(1-p') \rangle \\ &= 2\left(\langle p' \rangle - \langle p'^2 \rangle \right) \\ &= 2\left(\langle p' \rangle - \sigma_{p'}^2 - \langle p' \rangle^2 \right) \\ &= 2p\,(1-p)\left(1 - \frac{1}{2N}\right) \\ &= H_0\left(1 - \frac{1}{2N}\right) \end{aligned}$$

Over t generations,

$$\begin{aligned} H_t &= H_0\left(1 - \frac{1}{2N}\right)^t \\ &\approx H_0 \exp\left(-\frac{t}{2N}\right) \end{aligned}$$

where the approximation holds in the limit of large population size. Note that when $N \to \infty$, $H_t = H_0$, i.e. in an infinite population, in the absence of mutation or selection, the heterozygosity remains unchanged over generations. The equation above also shows that the heterozygosity decays to zero over a timescale $t_{drift} \approx 2N$ generations. Hence genetic drift has the effect of reducing the genetic diversity in a population over time.

In realistic situations, the population size fluctuates over time. This generalizes the above equation to:

$$\begin{aligned} H_t &= H_0\left(1 - \frac{1}{2N_0}\right)\left(1 - \frac{1}{2N_1}\right)\left(1 - \frac{1}{2N_2}\right)\cdots \\ &= H_0 \prod_{i=0}^{t-1}\left(1 - \frac{1}{2N_i}\right) \\ &\approx H_0\left(1 - \sum_{i=0}^{t-1}\frac{1}{2N_i}\right)\left(\text{ to lowest order in } \frac{1}{N_i}\right) \\ &= H_0\left(1 - \frac{t}{2N_{eff}}\right) \\ &\approx H_0 \exp\left(-\frac{t}{2N_{eff}}\right)\left(\text{ to lowest order in } \frac{1}{N_{eff}}\right) \end{aligned}$$

Thus, the heterozygosity at a locus in a population with a variable population size falls off exponentially over a timescale $t_{drift} \approx 2N_{eff}$.

The quantity N_{eff} is called the *effective population size*, and it determines the overall rate at which genetic diversity decays. It is defined as the harmonic mean of the population size over time.

$$\frac{t}{N_{eff}} = \sum_{i=0}^{t-1} \frac{1}{2N_i}$$

Since N_{eff} is a harmonic mean, it will be biased towards generations with the lowest population sizes. This is known as a genetic bottleneck effect, where species that have had low population numbers in the past continue to exhibit low genetic diversity.

Although the human population is very large today ($\approx$ 7 billion individuals), for much of human history, the effective population size of breeding humans was on the order of 10,000 individuals [23]. This corresponds to a timescale $t_{drift} \approx 20,000$ generations, or about half a million years. Drift is a very slow force in humans.

1.1.4 The balance between drift and mutation

If drift serves to reduces the heterozygosity in a population, why is it that populations are not devoid of all genetic diversity at neutral loci? The answer is that mutations oppose the effects of genetic drift.

We have seen that due to finite sampling effects, genetic drift reduces the heterozygosity in a single generation by the following amount:

$$H_1 = H_0 \left(1 - \frac{1}{2N}\right)$$
$$\implies \Delta H_{drift} = -\frac{1}{2N} H_0$$

Now, instead, consider a population of infinite size with a mutation rate μ. The heterozygosity in a given generation can be related to that in the previous generation as follows:

$$H_1 = H_0 + (1 - H_0)\left(1 - (1 - \mu)^2\right)$$

In words, the equation above says that there are two ways to be heterozygous in the next generation. Either you select two alleles that are non-identical in the first generation, and this happens with probability H_0. Or, you select two alleles that are identical in the first generation (with probability $1 - H_0$), and at least one of them

undergoes a mutation. The factor of $1-(1-\mu)^2$ is the probability that a mutation occurred in at least one of the two alleles, since $(1-\mu)^2$ is the probability that no mutation occurred in both alleles.

$$\begin{aligned} H_1 &= H_0 + (1-H_0)\left(1-(1-\mu)^2\right) \\ &\approx H_0 + 2\mu(1-H_0) \\ \implies \Delta H_\mu &= 2\mu(1-H_0) \end{aligned}$$

Combining these equations, we see that the overall change in heterozygosity in a single generation has contributions from two terms: a drift term, and a mutation term. Thus,

$$\begin{aligned} \Delta H &= \Delta H_{drift} + \Delta H_\mu \\ &= -\frac{1}{2N} H_0 + 2\mu(1-H_0) \end{aligned}$$

Mutations increase the genetic diversity, and drift decreases it. These two opposing forces attain an equilibrium when,

$$\begin{aligned} &\Delta H = 0 \\ \implies &2\mu(1-H_{eq}) - \frac{1}{2N} H_{eq} = 0 \\ \implies &H_{eq} = \frac{4N\mu}{1+4N\mu} \end{aligned}$$

When $4N\mu \approx 1$, we see interesting dynamics between drift and mutation. If $2\mu >> \frac{1}{2N}$, mutation dominates over drift, and we reach the limiting case where all $2N$ alleles are different from each other. On the other hand, if $2\mu << \frac{1}{2N}$, then the mutation rate is insufficient to offset the depletion of genetic diversity due to drift, and one of the alleles will attain a frequency of 1 (fixation).

1.1.5 The fourth force: Selection: results in adaptations

We have discussed the three neutral forces that alter gene frequencies and affect genetic diversity. We saw that mutation and drift act in opposite directions, and that there is a regime in which they can balance each other. However, this assumes that new mutations have a neutral effect and their fate is decided by chance.

An interesting situation arises when a new mutation confers a selective advantage to an individual. In the context of population genetics, a selective advantage amounts to a higher reproductive success. Hence, for a mutation to be under positive selection, it must provide a reproductive advantage to its bearer, i.e. it must increase the probability of its occurring in the next generation. If this effect is sufficiently strong, then the mutation will spread through the population at an exponential rate. The timescale for this spread is smaller than the timescale over which recombination breaks down linkage, or the timescale over which mutation builds diversity and breaks down linkage. The result is to create a region of the genome where diversity is markedly reduced, i.e. to create genomic islands where many individuals in the population share the selected mutation and the region around it (until it is broken up by recombination). This phenomenon of local reduction in genomic diversity around a selected mutation is known as a *selective sweep*. These selective sweeps leave behind a footprint of reduced genomic diversity, and these can be detected using numerous statistical tests which we will discuss and employ when identifying signatures of natural selection in the Maasai in chapter 4.

We can understand selection quantitatively as follows. Consider the fate of two alleles at a single locus, that differ in the extent to which they affect the reproductive success of the individuals that carry them. Suppose that, in a certain generation, the alleles have frequency p and $q = 1 - p$. Assume that the fitness of each allele provides a multiplicative factor to its frequency in the next generation. Then, we have that

$$p' = p\frac{w_1}{w}$$
$$q' = q\frac{w_2}{w}$$

where the primes indicate the next generation. Here $\frac{w_1}{w}$ and $\frac{w_2}{w}$ are the normalized selection coefficients (or fitness coefficients), and w is a normalization constant:

$$p' + q' = 1$$
$$\implies w = w_1 p + w_2(1 - p)$$

Thus w is also the mean fitness of the population with respect to this allele.

The change in frequency in a single generation is given by

$$\begin{aligned}\Delta p &= p' - p \\ &= p\left(\frac{w_1}{w} - 1\right)\end{aligned}$$

with some simplification, this can be written as

$$\Delta p = \frac{(w_1 - w_2)}{w} pq \tag{1.1}$$

This can be written in the more suggestive form first expressed by Sewall Wright in 1932 [24] [25].

$$\Delta p = \frac{pq}{w}\frac{dw}{dp}$$

This equation encapsulates the idea that natural selection drives allele frequencies to maximize the mean fitness. The change in frequency of an allele that is under selection depends on two key factors: the genetic variance $2pq$ and on the slope of the fitness function $\frac{dw}{dp}$. The latter term led Wright to the notion of a fitness landscape. In the case of two interacting loci, the fitness landscape is a two dimensional surface defined over the space of possible genotype frequencies. Wright visualized fitness as a very high-dimensional surface, defined over the range of frequencies of the vast number of genetic loci that interact to produce a complex phenotype. In this model, populations are driven to adaptive peaks in the landscape. If selection were the only operative force, populations would remain stuck in local peaks of this fitness landscape. However, mutation and drift can drive a population to explore the space away from these peaks, into fitness valleys and towards other fitness peaks.

There are few lessons we can take away from this idea of a fitness landscape. First, that population size governs the dynamics of populations in the landscape. Small population sizes are more strongly driven by drift and furthermore, selection has a weaker effect in such populations. Therefore smaller populations are less likely to be trapped in a local adaptive peak and can drift further in fitness space. This hypothesis, known as the shifting balance theory, was proposed by Wright to explain how populations approach ever higher fitness peaks. Secondly, the entire fitness landscape is strongly

dependent on the environment in which the population exists, which changes with time. Therefore, the landscape is far from a static entity - peaks and valleys shift as the environment varies. Classic examples of selection in response to environmental conditions are the genetic adaptations for high altitude existence in residents of the Tibetan Plateau [26], and lactase persistence in cattle herders [18], [27], [28], [29].

Finally, the fitness landscape emphasizes the important notion that genotypes are not independent. The existence of fitness peaks implies that different genes do not have a simple additive effect, but instead interact with each other to create the fitness landscape. Such gene-gene interactions are known as epistatic interactions, and we investigate their effects further when we study the gene interaction networks in yeast in chapter 2.

1.1.6 Effect of selection on allele frequencies

Let us consider the trajectory of a mutation that is undergoing positive selection. Taking the continuum limit, we can write equation 1.1 as

$$\frac{dp}{dt} = fp(1-p) \tag{1.2}$$

where f is the difference in relative fitness of the two alleles. This equation is the widely recognized logistic differential equation, whose solution is a sigmoidal function

$$p(t) = \frac{p_0}{p_0 + (1-p_0)e^{-ft}}$$

This teaches us that the rate of growth of an allele under selection is initially slow, then enters a period of exponential growth as the allele spreads to a sizable fraction of the population. The growth rate peters out as the allele frequency is close to fixation ($p = 1$). The timescale corresponding to these selective sweeps is inversely proportional to the fitness advantage of the allele $t_{sel} \approx \frac{1}{f}$. A selective advantage of 1 in a 1000 can cause a sizable change in the frequency of an allele in just 1,000 generations. This is a much shorter timescale compared to $t_{recombination}$, t_{drift} or $t_{mutation}$. Therefore, the hallmark of a selective sweep is a region of the genome over which genetic diversity is rapidly depleted, until the forces of recombination and mutation restore the diversity around the selected locus.

In real populations, a finite population size causes genetic drift to interact with selection. The detailed calculation of this effect was done by Kimura in 1962 [30]. In a seminal paper, he showed that the probability of a selected mutation being fixed in the population (i.e. attaining frequency of 1), assuming an initial frequency of $\frac{1}{2N}$, is given by

$$P(\text{fixation}) = \frac{1 - e^{-s}}{1 - e^{-2Ns}}$$

Hence, selection can interact with drift when the fitness advantage of an allele $s \approx \frac{1}{2N}$. For $s >> \frac{1}{2N}$, drift plays no role in selection.

1.1.7 Timescales of selection

The genetic signature of selection is that a single genetic variant (allele) rises rapidly in frequency in the population. Neighboring alleles that are linked to the selected mutation will also rise in frequency, and this is known as genetic hitchhiking. This results in a local reduction in genomic diversity, known as a selective sweep. The signature of a selective sweep is therefore lowered diversity combined with an abundance of rare alleles that have risen to a high frequency. All methods that identify selection over long timescales are designed to detect this signature. Over longer timescales, mutation and recombination will re-introduce diversity into this genetic locus. On average in humans, a chromosomal segment 100 kilobases in length will have had more than one recombination event in 30,000 years, and this breaks down the pattern of linkage. Therefore methods that are based on linkage disequilibrium will only work over a timescale up to about 30,000 years, after which recombination has effectively broken down the linkage. Over much longer timescales, (i.e. on the order of half a million years) mutations will re-appear with reasonable frequency in this region, restoring the diversity. Eventually, these neutral forces will overwrite the signature of selection. Therefore, we only have access to selective sweeps that have occurred in the not-too distant past.

The following is a summary of signatures used to detect selection, ranked according to the timescale over which they are effective. For a review of these methods, see [15].

1. High proportion of protein altering mutations (millions of years)

2. Reduced genetic diversity, abundance of high-frequency derived alleles ($< 250{,}000$ years)

3. Differences between population groups ($< 50{,}000$ years)

4. Methods based on linkage disequilibrium ($< 30{,}000$ years)

In chapter 4 we use numerous statistics designed to identify population differences and linkage disequilibrium based signals to identify genes undergoing recent, positive, natural selection.

1.1.8 Challenges of Genome-Wide Association Studies

One of the key issues in computational scans of selection is that it is difficult to isolate the effect of a single mutation in a population. Our ability to separate the phenotypic effects of two distinct genetic variants relies on the extent to which the mutations occur separately in the population, and on the frequency with which the variants occur in the population [31]. This issue is particularly important in clonal populations where recombination is rare or absent, because this leads to high correlation (linkage disequilibrium) between loci. Furthermore, if the individuals in a population differ in degree of relatedness, this can lead to spurious associations between markers due to the population stratification [32]. Genome-wide association studies (GWAS) of naturally occurring populations typically need very large sample sizes to overcome these challenges [31].

An ideal genetic mapping study would involve a large sample of equally related individuals (to avoid spurious associations due to population structure), low linkage disequilibrium (so that distinct genotypes are assorted randomly), and genotype frequencies of 50% for a bi-allelic locus (allowing for equal statistical power to detect the affect of both the alleles). Furthermore, we should have a high-resolution map of genetic markers to maximize our ability to isolate causal variants. Such a situation can be engineered in yeast, by studying a large number of segregants that are created by crossing two genetically divergent parental strains.

1.1.9 Dissecting the genetic architecture of complex traits using quantitative trait mapping

Mancera *et al.* [7] created a high resolution genetic map of all segregants that arose from 56 meiotic crosses between two genetically divergent strains of *Saccharomyces cerevisiae*. The segregants are the products of meiosis between the parental genomes (each cross resulting in four segregants). Since recombination happens at random locations in the genome, the large number of recombination events between the parents ensures that distant loci are effectively uncorrelated in the segregants. Furthermore, by genotyping all meiotic offspring at nucleotides where the parental strains differed, Mancera *et al.* [7] ensured that the genotype frequencies in the offspring was approximately 50% (the exception being asymmetrical crossover events, which occurred rarely). One can then measure any yeast phenotype of these strains, across a variety of environmental conditions, and use the genotype and phenotype data to conduct a high resolution genetic mapping study.

The central idea behind quantitative trait mapping is to identify associations between the phenotype and the genotype of an organism at a given marker. A quantitative trait locus, or QTL, is a marker where individuals with different genotypes have significantly different phenotypic means. For a complex trait, multiple QTLs may contribute towards phenotypic variance in the population, each QTL explaining a fraction of the variance.

In the quantitative trait mapping projects of chapters 2 and 3, we fit the phenotypic data at a locus to two hypotheses, one of which is that there is a genotype/phenotype association, and the other is a null hypothesis. The hypothesis of genotype/phenotype association being tested is that phenotype data fits to different means for each genotype at the locus (i.e. the marker is a QTL). The null hypothesis is that the phenotype data fits a single overall mean (no QTL present). The parameters of these models are inferred using maximum likelihood. The strength of the evidence for the presence of a QTL at a locus is evaluated using the LOD score, which is the log (base 10) of the ratio of the likelihood of the QTL hypothesis to that of the null hypothesis. A LOD score

of 3 implies that the hypothesis that a QTL is present is 1000 fold more likely than the null hypothesis that there is no QTL. The details and implementation of LOD scores to identify QTLs and environment specific QTL-QTL interactions are discussed in chapters 2 and 3.

Chapter 2

The *FLO8* locus regulates yeast growth plasticity through environment specific epistatic interactions

2.1 Introduction

A single genotype can exhibit different phenotypes in different environments, a property which is known as phenotypic plasticity [33]. The extent of phenotypic plasticity is quantifiable as the rate of change of the phenotype with respect to an environmental variable (reviewed in [34], [35]). A gene-environment interaction or GEI is characterized by measurable differences in plasticity as the alleles at a locus are varied. Such loci, called GEI QTLs, contribute to the variation in phenotypes seen across environments in populations. It is possible to identify such GEI QTLs by identifying the genotypes which cause a differential change in phenotype under a change in the environment [36], [37].

There are two proposed models about how organisms mediate plasticity [38], [39]: The first model, called the "allelic sensitivity model", asserts that there are loci with a direct effect on the phenotype and differential effects of its alleles in different environments contribute to variation in plasticity or GEI. The second, called the "regulatory gene model", posits that regulatory genes affect the expression of other genes, which directly affect the phenotype, in an environment dependent way. The two models are overlapping to the extent that regulatory genes may affect the expression of allelic sensitivity genes resulting in variation in plasticity. There is no clear distinction between the kind of genes which can fall under either model. There is also the possibility that the genotype/phenotype association is a complex scenario, with plasticity being governed by epistatic interactions between the effects of the QTL and genes that it affects.

In a unicellular, non-motile organism such as *Saccharomyces cerevisiae*, carbon

sources act both as an energy source as well as signaling molecules [40]. Their availability affects not only growth, but also other processes, such as stress response and metabolism [41]. In its evolutionary history, yeast must have encountered and adapted to a variety of carbon sources, both fermentable high growth sources such as glucose, fructose, and maltose, as well as non-fermentable, slow growth sources such as glycerol, and ethanol (reviewed in [42], [6]). Mapping studies have shown that different QTL contribute to variation in growth in the presence of different types of carbon sources [43], [44]. However, how these QTLs vary for fermentable and non fermentable carbon sources is not well understood.

Yeast growth can be studied via measurements of several phenotypes, including colony size [45], biomass [46] and growth kinetics [47]. However, it is known that different carbon sources can have independent effects on these measures of growth [48], [43]. The mechanism by which different growth phenotypes are affected by changing carbon sources is not clearly understood, nor is it known which genes regulate growth plasticity.

One can now ask: How plastic are the three growth parameters across diverse environments? Do the same QTLs mediate plasticity across functionally dissimilar environments? Do these QTLs vary for different growth parameters? Do epistatic interactions amongst these QTLs contribute to variation in growth plasticity? In this study we attempted to understand these questions by studying the genetic factors driving growth across a varied set of carbon sources.

Two genetically and phenotypically divergent yeast strains, S96 (a laboratory strain) and YJM789 (a clinical isolate) [49], [50] were selected for mapping. From a previously studied set of high-resolution genotyped meiotic segregants [7], we grew 157 non-flocculating segregants separately in the presence of fructose, maltose and glycerol. Three growth parameters: lag time, doubling time and biomass accumulated were calculated by fitting the growth data to a sigmoid function (described below). Using the phenotypic values obtained from these curve fits, we identified (mapped) the QTLs driving the variation in each parameter in the presence of each of the three carbon sources, and then mapped gene $\times$ environment interactions (GEI QTLs). This allowed

us to identify the QTLs that confer variation within a single environment, as well as the GEI QTLs that confer variation in plasticity across environments. Finally, we studied epistatic interactions between the resulting single environment and GEI QTLs.

Our analysis showed that genetic regulators of growth plasticity are both environment specific and growth parameter specific. We also found that common QTLs and QTL-QTL interactions contribute to variation in lag time and doubling time, and further, that a largely dissimilar set of QTLs and interactions governs the variation in biomass. A key finding of our study was that different genetic regulators govern growth rate and biomass. In addition, we identified an overlap in QTLs and QTL-QTL interactions, including the presence of a common interaction with the *FLO8* locus in fructose and glycerol. This demonstrated that growth in these dissimilar media have some common regulators. Furthermore, we found that this *FLO8* locus interacts with multiple QTLs to regulate distinct growth parameters, indicating that a regulatory gene may affect different target structural genes to manifest phenotypic plasticity.

2.2 Materials and Methods

2.2.1 Strains and Growth Conditions

We measured optical density growth profiles for the two parental yeast strains S288C, YJM879, and for 157 meiotic segregants, obtained from the collection of Mancera *et al.* [7]. The strains were grown separately in the presence of Glycerol, Fructose, Maltose, and in YPD (yeast-extract peptone dextrose). In each environmental condition, we measured two experimental replicates for each segregant, and 12 replicates for each parental strain. Optical density data was measured for each strain in intervals of 15 minutes over a period of 50 hours.

2.2.2 Curve Fitting

We used a custom Python script to fit the data to growth curves. Each set of optical density measurements was fit to a sigmoidal curve with the functional form:

$$OD(t) = \frac{m}{1 + \exp(-2c(t - t_{\frac{1}{2}}))} \tag{2.1}$$

The optical density (OD), which is proportional to the number of yeast cells, is a function of three parameters - c, m and $t_{\frac{1}{2}}$. The parameter m is the maximum growth attained by a strain, in units of optical density. The parameter $t_{\frac{1}{2}}$ is the time, measured in minutes, at which the number of yeast cells are one half of their maximum amount. When $t \to t_{\frac{1}{2}}$, the yeast growth curve is approximately described by an exponential, i.e. $OD(t) \approx \frac{m}{2} e^{c(t - t_{\frac{1}{2}})}$. Hence $t_{\frac{1}{2}}$ can be thought of as the time after which a yeast strain undergoes exponential growth. The doubling time $\delta = \frac{\ln(2)}{c}$, measured in minutes, is the time for a strain to double in number while in exponential growth.

One concern in using a sigmoidal model to curve-fit growth curves is that the growth rate c and the time to exponential growth $t_{\frac{1}{2}}$ may be dependent on the initial number of cells. Hence, any experimental variation in initial OD may affect the measurement of these growth parameters. Here, we demonstrate that this is not a significant issue, and that variation in the initial number of yeast cells has a small effect on c and $t_{\frac{1}{2}}$.

As $OD(t) \propto N(t)$, the growth rate is given by:

$$\frac{1}{N(t)} \frac{dN(t)}{dt} = c(1 - \tanh(c(t - t_{\frac{1}{2}})))$$

This simplifies considerably at $t = t_{\frac{1}{2}}$,

$$\left. \frac{1}{N(t)} \frac{dN(t)}{dt} \right|_{t = t_{\frac{1}{2}}} = c$$

implying that, near $t_{\frac{1}{2}}$, the growth is well approximated by exponential growth.

At the initial timepoint $t = 0$,

$$N_0 = OD(0) = \frac{m}{1 + e^{2ct_{\frac{1}{2}}}}$$

We can relate the uncertainty in N_0 to the uncertainty in c and $t_{\frac{1}{2}}$ using the following relation:

$$\sigma_{N_0}^2 = \left(\frac{\partial N_0}{\partial c} \right)^2 \sigma_c^2 + \left(\frac{\partial N_0}{\partial t_{\frac{1}{2}}} \right)^2 \sigma_{t_{\frac{1}{2}}}^2 + 2 \frac{\partial N_0}{\partial c} \frac{\partial N_0}{\partial t_{\frac{1}{2}}} \sigma_c \sigma_{t_{\frac{1}{2}}} \operatorname{Corr}(c, t_{\frac{1}{2}})$$

which reduces to

$$\sigma_{N_0}^2 = \left(\frac{1}{2}mct_{\frac{1}{2}}\operatorname{sech}^2(ct_{\frac{1}{2}})\right)^2\left(\left(\frac{\sigma_{t_{\frac{1}{2}}}}{t_{\frac{1}{2}}}\right)^2+\left(\frac{\sigma_c}{c}\right)^2+2\frac{\sigma_{t_{\frac{1}{2}}}}{t_{\frac{1}{2}}}\frac{\sigma_c}{c}\operatorname{Corr}(c,t_{\frac{1}{2}})\right)$$

And, using the relations $N_0e^{ct_{\frac{1}{2}}} = \frac{m}{2}\operatorname{sech}(ct_{\frac{1}{2}})$ and $\operatorname{sech}(ct_{\frac{1}{2}})e^{ct_{\frac{1}{2}}} = 1+\tanh(ct_{\frac{1}{2}})$, we can simplify this to:

$$\frac{1}{\left[ct_{\frac{1}{2}}(1+\tanh(ct_{\frac{1}{2}}))\right]^2}\left(\frac{\sigma_{N_0}}{N_0}\right)^2 = \left(\frac{\sigma_{t_{\frac{1}{2}}}}{t_{\frac{1}{2}}}\right)^2+\left(\frac{\sigma_c}{c}\right)^2+2\frac{\sigma_{t_{\frac{1}{2}}}}{t_{\frac{1}{2}}}\frac{\sigma_c}{c}\operatorname{Corr}(c,t_{\frac{1}{2}})$$

This is the desired relation between uncertainty in N_0 to uncertainty in c and uncertainty in $t_{\frac{1}{2}}$. If we assume that uncertainty in c and uncertainty in $t_{\frac{1}{2}}$ are on the same order of magnitude (as seen in the data), i.e.

$$\frac{\sigma_c}{c} = (1+\epsilon)\frac{\sigma_{t_{\frac{1}{2}}}}{t_{\frac{1}{2}}}$$

Then, to first order in ϵ,

$$\frac{1}{\sqrt{2(1+\epsilon)(1+\operatorname{Corr}(c,t_{\frac{1}{2}}))}}\frac{1}{ct_{\frac{1}{2}}(1+\tanh(ct_{\frac{1}{2}}))}\frac{\sigma_{N_0}}{N_0} = \frac{\sigma_{t_{\frac{1}{2}}}}{t_{\frac{1}{2}}}$$

As an example, for S96 in Maltose, $c \approx \ln(2)/167$ min^{-1}, $t_{\frac{1}{2}} \approx 760$ min, so $\sqrt{2}ct_{\frac{1}{2}}(1+\tanh(ct_{\frac{1}{2}})) \approx 8.9$. Inserting these values,

$$\frac{1}{\sqrt{(1+\epsilon)(1+\operatorname{Corr}(c,t_{\frac{1}{2}}))}}\frac{1}{8.9}\frac{\sigma_{N_0}}{N_0} = \frac{\sigma_{t_{\frac{1}{2}}}}{t_{\frac{1}{2}}}$$

Hence, an uncertainty of $\frac{\sigma_{N_0}}{N_0}$ in initial OD leads to at most $\frac{1}{8.9} \approx 0.11$ as much uncertainty in $t_{\frac{1}{2}}$ (for strain S96), and this factor is even lower if lag time and growth rate are correlated (as seen in the data). A similar argument holds true for other yeast strains analyzed. This suggests that, under the assumption that the data is well approximated by a sigmoid curve, the growth parameters that are inferred from the curve fit are not strongly affected by variation in the initial number of yeast cells.

2.2.3 Mapping single QTLs

For each strain, we measured 9 phenotypic traits: 3 growth parameters (doubling time, lag time, biomass), each of which was measured in 3 environmental conditions (in the

presence of glycerol, maltose, and fructose). Genotype data for the parental strains and segregants was obtained from Mancera *et al.* [7], and filtered to include only single nucleotide markers, which resulted in 48,934 markers.

We used the R/qtl [12], [13] package to construct a genetic map and identify QTLs separately for each of the 9 conditions. QTLs were identified using the LOD score, which is the $\log_{10}$ of the ratio of the likelihood of the experimental hypothesis to the likelihood of the null hypothesis. A LOD score of 3 implies that the hypothesis in question is 1000 times more likely than the null hypothesis at a given locus.

In the case of a non-interacting QTL, we compare the likelihood of the data given the following two hypotheses:

$$H_1 : y_i = \mu + \beta g_i + \epsilon_i$$

$$H_0 : y_i = \mu + \epsilon_i$$

Here, y_i is the phenotype of strain i, g_i is a genotype variable (0 or 1) and ϵ_i is a noise variable, with zero mean and fixed variance, representing stochastic variation in the measurements.

H_1 is the hypothesis that the two genotypes have different means (i.e. a QTL is present), and H_0 is the null model that both genotypes have the same mean (no QTL present). The parameter β captures the effect of the QTL.

We define a likelihood function for each hypothesis:

$$\mathcal{L}(H_1) = \prod_i \phi(y_i|\mu + \beta g_i, \sigma^2)$$

$$\mathcal{L}(H_0) = \prod_i \phi(y_i|\mu, \sigma^2)$$

where ϕ is the density function for the normal distribution, and the parameters μ, β and σ^2 are obtained for each hypothesis by maximizing the likelihood.

The LOD score of interest for a single environment QTL is then $LOD = \log(\frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)})$.

$$\mathcal{L}(H_0) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \prod_i \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

$$\log \mathcal{L}(H_0) = \frac{n}{2} \log\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2$$

Similarly, we have that

$$\log \mathcal{L}(H_1) = \frac{n}{2} \log \left(\frac{1}{2\pi\sigma^2} \right) - \frac{1}{2\sigma^2} \left(\sum_{i \in G_1} (x_i - \mu_1)^2 + \sum_{i \in G_2} (x_i - \mu_2)^2 \right)$$

where $\mu_2 = \mu_1 + \beta$

We can simplify the maximum likelihood estimations by noting that:

$$\sum_{i \in G_j} (x_i - \mu_j)^2 = \sum_{i \in G_j} (x_i - \bar{x}_j)^2 + n_j (\bar{x}_j - \mu_j)^2$$

where $\bar{x}_j = \frac{1}{n_j} \sum_{i \in G_j} x_i$ is the mean of samples with the j^{th} genotype.

Maximizing the likelihood with respect to μ_j, we see that the maximum likelihood estimate of μ_j is just the sample mean:

$$\hat{\mu}_j = \bar{x}_j$$

Similarly, maximizing the likelihood with respect to σ^2 shows that the maximum likelihood estimate of σ^2 is the weighted average of the variances of the genotype variances.

$$\hat{\sigma}^2 = \sum_j \frac{\sum_{i \in G_j} (x_i - \mu_j)^2}{n} = \sum_j \frac{n_j}{n} {\sigma_j}^2$$

The LOD score is then $\log \mathcal{L}(H_1) - \log \mathcal{L}(H_0)$, which can be simplified to:

$$\begin{aligned} LOD &= \frac{n}{2} \log \left(\frac{\sum_i (x_i - \bar{x})^2}{\sum_{i \in G_1} (x_i - \bar{x}_1)^2 + \sum_{i \in G_2} (x_i - \bar{x}_2)^2} \right) \\ &= \frac{n}{2} \log \left(\frac{n\sigma^2}{n_1 \sigma_1^2 + n_2 \sigma_2^2} \right) \end{aligned} \tag{2.2}$$

where σ^2 is the sample variance, and σ_1^2 and σ_2^2 are the variances of samples with genotypes 1 and 2.

The size of the effect of the QTL (β) is then given by the difference in sample means for the two genotypes, i.e. $\hat{\beta} = \hat{\mu}_2 - \hat{\mu}_1 = \bar{x}_2 - \bar{x}_1$.

We used the *scanone* function in R/qtl to compute this LOD score using the Haley-Knott regression algorithm [12], [13]. This is an interval mapping method and has the advantage over marker regression that it can impute data at missing markers and inspect positions between markers. We compute p-values in R/qtl with a permutation

test (1,000 permutations). The null distribution was obtained by measuring the highest genome-wide LOD score obtained from each permutation of the permutation test [12], [13].

2.2.4 Mapping QTL-environment interactions

A QTL-environment interaction occurs when the effect of a QTL is environment dependent. We identify such QTLs by pooling data from two environmental conditions and including the effect of the environment as a covariate. This requires us to compare the following two hypotheses:

$$H_I : y_i = \mu + \beta_g g_i + \beta_x x_i + \gamma g_i x_i + \epsilon_i$$

$$H_A : y_i = \mu + \beta_g g_i + \beta_x x_i + \epsilon_i$$

The new variable x_i is an environmental covariate that is 0 or 1 depending on the environment of the strain. As before, the parameters μ, β_g, β_x, γ are all obtained by maximizing the likelihood.

In H_A, the effect of the environment is modeled as an additive covariate, i.e. the phenotype is the sum of a constant QTL effect (β_g) and a constant environment dependent effect (β_x). In H_I, the effect of the environment is modelled as an interactive covariate. The term γ captures the effect of the QTL-Environment interaction.

To identify a QTL-environment interaction, the LOD score of interest is $LOD(H_I) - LOD(H_A)$. These scores were calculated by again using the *scanone* function in R/qtl (using the Haley-Knott regression algorithm), including the environmental variable as an additive and interactive covariate. We compute p-values as before (100 permutations) [12], [13].

2.2.5 Mapping QTL-QTL interactions

A QTL-QTL interaction occurs when the effect of a QTL at a single locus depends on the genotype at some other locus. We identified the presence of QTL-QTL interactions

by comparing the following hypotheses:

$$H_I : y_i = \mu + \beta_1 g_{1i} + \beta_2 g_{2i} + \gamma g_{1i} g_{2i} + \epsilon_i$$

$$H_A : y_i = \mu + \beta_1 g_{1i} + \beta_2 g_{2i} + \epsilon_i$$

Here g_{1i} and g_{2i} are binary variables that specify the genotypes at two loci. As before, μ, β_1, β_2 and γ are inferred from the data using maximum likelihood. The parameters β_1 and β_2 quantify the individual effect of each QTL, and γ quantifies the effect of the QTL-QTL interaction.

The LOD score of interest in identifying QTL-QTL interactions is $LOD(H_I) - LOD(H_A)$. The log likelihood of each hypothesis can be written as

$$\log \mathcal{L} = \frac{n}{2} \log \left(\frac{1}{2\pi\sigma^2} \right) - \frac{1}{2\sigma^2} \sum_{j \in \{00,01,10,11\}} \left(\sum_{i \in G_j} (x_i - \bar{x}_j)^2 + n_j (\bar{x}_j - \mu_j)^2 \right) \quad (2.3)$$

where

$$\mu_{00} = \mu$$

$$\mu_{10} = \mu + \beta_1$$

$$\mu_{01} = \mu + \beta_2$$

$$\mu_{11} = \mu + \beta_1 + \beta_2 + \gamma$$

for the interactive hypothesis H_I.

Maximizing Equation 2.3 with respect to μ, β_1, β_2 and γ shows that

$$\hat{\mu} = \bar{x}_{00}$$

$$\hat{\beta}_1 = \bar{x}_{10} - \hat{\mu}$$

$$\hat{\beta}_2 = \bar{x}_{01} - \hat{\mu}$$

$$\hat{\gamma} = \bar{x}_{11} - \hat{\mu} - \hat{\beta}_1 - \hat{\beta}_2$$

For the additive hypothesis H_A, there is one fewer parameter:

$$\mu_{00} = \mu$$

$$\mu_{10} = \mu + \beta_1$$

$$\mu_{01} = \mu + \beta_2$$

$$\mu_{11} = \mu + \beta_1 + \beta_2$$

Maximizing Equation 2.3 with respect to μ, β_1 and β_2 gives a set of three equations:

$$\begin{aligned} n_{00}(\bar{x}_{00}-\mu)+n_{10}(\bar{x}_{10}-\mu-\beta_1)+n_{01}(\bar{x}_{01}-\mu-\beta_2)+n_{11}(\bar{x}_{11}-\mu-\beta_1-\beta_2)&=0 \\ n_{10}(\bar{x}_{10}-\mu-\beta_1)+n_{11}(\bar{x}_{11}-\mu-\beta_1-\beta_2)&=0 \\ n_{01}(\bar{x}_{01}-\mu-\beta_2)+n_{11}(\bar{x}_{11}-\mu-\beta_1-\beta_2)&=0 \end{aligned}$$

Which can be solved to provide maximum likelihood estimates for μ, β_1, and β_2.

$$\begin{aligned} \alpha &= n_{01}n_{10}(n_{00}+n_{11})+n_{00}n_{11}(n_{01}+n_{10}) \\ \hat{\mu} &= (n_{00}(n_{10}n_{11}+n_{01}(n_{10}+n_{11}))\bar{x}_{00}+n_{01}n_{10}n_{11}(\bar{x}_{01}+\bar{x}_{10}-\bar{x}_{11}))/\alpha \\ \hat{\beta}_1 &= ((n_{00}+n_{10})n_{01}n_{11}(\bar{x}_{11}-\bar{x}_{01})+n_{00}n_{10}(n_{01}+n_{11})(\bar{x}_{10}-\bar{x}_{00}))/\alpha \\ \hat{\beta}_2 &= ((n_{00}+n_{01})n_{10}n_{11}(\bar{x}_{11}-\bar{x}_{10})+n_{00}n_{01}(n_{10}+n_{11})(\bar{x}_{01}-\bar{x}_{00}))/\alpha \end{aligned} \tag{2.4}$$

The LOD score is then $\log \mathcal{L}(H_I) - \log \mathcal{L}(H_A)$

$$\begin{aligned} LOD = \frac{n}{2}\log\Bigg(&\sum_{i\in G_{00}}(x_i-\hat{\mu})^2+\sum_{i\in G_{10}}(x_i-\hat{\mu}-\hat{\beta}_1)^2+\sum_{i\in G_{01}}(x_i-\hat{\mu}-\hat{\beta}_2)^2 \\ &+\sum_{i\in G_{11}}(x_i-\hat{\mu}-\hat{\beta}_1-\hat{\beta}_2)^2\Bigg) \\ -\frac{n}{2}\log\Bigg(&\sum_{i\in G_{00}}(x_i-\bar{x}_{00})^2+\sum_{i\in G_{10}}(x_i-\bar{x}_{10})^2+\sum_{i\in G_{01}}(x_i-\bar{x}_{01})^2+\sum_{i\in G_{11}}(x_i-\bar{x}_{11})^2\Bigg) \end{aligned} \tag{2.5}$$

where the parameters $\hat{\mu}$, $\hat{\beta}_1$, $\hat{\beta}_2$ are obtained using equation 2.4.

We used a custom-written python script to compute this LOD score for pairwise comparisons among a set of markers. Our script did not impute missing genotypes. We compute p-values in python with a permutation test (1,000 permutations) where the null distribution consisted of the highest LOD score obtained among all pairwise comparisons for each permutation of the phenotype.

2.2.6 Mapping QTL-QTL-environment interactions

So far we have investigated two locus interactions, where the phenotype data is mapped to a relation of the form

$$y = \mu + \sum_j \beta_g g_j + \sum_{j,k} \gamma_{jk} g_j g_k + \epsilon$$

In the equation above, we have dropped the i subscript on y, g, and ϵ for the sake of clarity. This model accounts for single locus effects as well as pairwise interactions. However, we can also investigate three-point interactions such as QTL $\times$ QTL $\times$ QTL interactions or QTL $\times$ QTL $\times$ environment interactions by investigating the effect of an additional term of the form $\delta g_1 g_2 g_3$ where g_3 could be an additional genotype locus, or it could be an environmental covariate variable x. The term δ captures the effect size of the QTL $\times$ QTL $\times$ environment or QTL $\times$ QTL $\times$ QTL interaction.

Concretely, we compare the following hypotheses:

$$H_{3I} : y_i = \mu + \beta_1 g_{1i} + \beta_2 g_{2i} + \beta_3 g_{3i} + \gamma_{12} g_{1i} g_{2i} + \gamma_{13} g_{1i} g_{3i} + \gamma_{23} g_{2i} g_{3i} + \delta g_{1i} g_{2i} g_{3i} + \epsilon_i$$

$$H_{2I} : y_i = \mu + \beta_1 g_{1i} + \beta_2 g_{2i} + \beta_3 g_{3i} + \gamma_{12} g_{1i} g_{2i} + \gamma_{13} g_{1i} g_{3i} + \gamma_{23} g_{2i} g_{3i} + \epsilon_i$$

As before, the log likelihood of either hypothesis can be written as:

$$\log \mathcal{L} = \frac{n}{2} \log \left(\frac{1}{2\pi\sigma^2} \right) - \frac{1}{2\sigma^2} \sum_{j \in \{\mathcal{Z}_2 \times \mathcal{Z}_2 \times \mathcal{Z}_2\}} \left(\sum_{i \in G_j} (x_i - \bar{x}_j)^2 + n_j (\bar{x}_j - \mu_j)^2 \right)$$

Maximizing $\log \mathcal{L}(H_{3I})$ gives us the maximum likelihood estimate of the parameters of H_{3I}:

$$\begin{aligned}
\hat{\mu} &= \bar{x}_{000} \\
\hat{\beta}_1 &= \bar{x}_{100} - \hat{\mu} \\
\hat{\beta}_2 &= \bar{x}_{010} - \hat{\mu} \\
\hat{\beta}_3 &= \bar{x}_{001} - \hat{\mu} \\
\hat{\gamma}_{12} &= \bar{x}_{110} - \hat{\beta}_1 - \hat{\beta}_2 - \hat{\mu} \\
\hat{\gamma}_{23} &= \bar{x}_{011} - \hat{\beta}_2 - \hat{\beta}_3 - \hat{\mu} \\
\hat{\gamma}_{13} &= \bar{x}_{101} - \hat{\beta}_1 - \hat{\beta}_3 - \hat{\mu} \\
\hat{\delta} &= \bar{x}_{111} - \hat{\gamma}_{12} - \hat{\gamma}_{23} - \hat{\gamma}_{13} - \hat{\beta}_1 - \hat{\beta}_2 - \hat{\beta}_3 - \hat{\mu}
\end{aligned}$$

To maximize the likelihood of H_{2I}, we differentiate $\log \mathcal{L}(H_{2I})$ with respect to μ, β_1, β_2, β_3, γ_{12}, γ_{23}, γ_{13}, resulting in a set of 7 simultaneous linear equations:

$$\begin{aligned}
&2n_{000}(\bar{x}_{000}-\mu)+2n_{100}(\bar{x}_{100}-\beta_1-\mu)+2n_{010}(\bar{x}_{010}-\beta_2-\mu)+2n_{001}(\bar{x}_{001}-\beta_3-\mu)\\
&\quad+2n_{110}(\bar{x}_{110}-\beta_1-\beta_2-\gamma_{12}-\mu)+2n_{101}(\bar{x}_{101}-\beta_1-\beta_3-\gamma_{13}-\mu)\\
&\quad+2n_{011}(\bar{x}_{011}-\beta_2-\beta_3-\gamma_{23}-\mu)+2n_{111}(\bar{x}_{111}-\beta_1-\beta_2-\beta_3-\gamma_{12}-\gamma_{13}-\gamma_{23}-\mu)=0
\end{aligned}$$

$$\begin{aligned}
&2n_{100}(\bar{x}_{100}-\beta_1-\mu)+2n_{110}(\bar{x}_{110}-\beta_1-\beta_2-\gamma_{12}-\mu)\\
&\quad+2n_{101}(\bar{x}_{101}-\beta_1-\beta_3-\gamma\gamma_{13}-\mu)+2n_{111}(\bar{x}_{111}-\beta_1-\beta_2-\beta_3-\gamma_{12}-\gamma_{13}-\gamma_{23}-\mu)=0
\end{aligned}$$

$$\begin{aligned}
&2n_{010}(\bar{x}_{010}-\beta_2-\mu)+2n_{110}(\bar{x}_{110}-\beta_1-\beta_2-\gamma_{12}-\mu)\\
&\quad+2n_{011}(\bar{x}_{011}-\beta_2-\beta_3-\gamma_{23}-\mu)+2n_{111}(\bar{x}_{111}-\beta_1-\beta_2-\beta_3-\gamma_{12}-\gamma_{13}-\gamma_{23}-\mu)=0
\end{aligned}$$

$$\begin{aligned}
&2n_{001}(\bar{x}_{001}-\beta_3-\mu)+2n_{101}(\bar{x}_{101}-\beta_1-\beta_3-\gamma_{13}-\mu)\\
&\quad+2n_{011}(\bar{x}_{011}-\beta_2-\beta_3-\gamma_{23}-\mu)+2n_{111}(\bar{x}_{111}-\beta_1-\beta_2-\beta_3-\gamma_{12}-\gamma_{13}-\gamma_{23}-\mu)=0
\end{aligned}$$

$$2n_{110}(\bar{x}_{110}-\beta_1-\beta_2-\gamma_{12}-\mu)+2n_{111}(\bar{x}_{111}-\beta_1-\beta_2-\beta_3-\gamma_{12}-\gamma_{13}-\gamma_{23}-\mu)=0$$

$$2n_{011}(\bar{x}_{011}-\beta_2-\beta_3-\gamma_{23}-\mu)+2n_{111}(\bar{x}_{111}-\beta_1-\beta_2-\beta_3-\gamma_{12}-\gamma_{13}-\gamma_{23}-\mu)=0$$

$$2n_{101}(\bar{x}_{101}-\beta_1-\beta_3-\gamma_{13}-\mu)+2n_{111}(\bar{x}_{111}-\beta_1-\beta_2-\beta_3-\gamma_{12}-\gamma_{13}-\gamma_{23}-\mu)=0$$

The solution to these simultaneous equations gives us the maximum likelihood estimates $\hat{\mu}$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\gamma}_{12}$, $\hat{\gamma}_{23}$, $\hat{\gamma}_{13}$

$$\begin{aligned}
\alpha = (&n_{000}n_{001}n_{010}n_{011}n_{100}n_{101}n_{110}+n_{001}n_{010}n_{011}n_{100}n_{101}n_{110}n_{111}\\
&+n_{000}(n_{001}n_{010}n_{011}n_{100}n_{101}+n_{010}n_{011}n_{100}n_{101}n_{110}\\
&+n_{001}(n_{010}n_{011}n_{100}+n_{011}n_{100}n_{101}+n_{010}(n_{011}+n_{100})n_{101})n_{110})n_{111})
\end{aligned}$$

$$\begin{aligned}
\hat{\mu}\cdot\alpha = n_{000}(&n_{001}n_{010}n_{011}n_{100}n_{101}n_{110}+n_{010}n_{011}n_{100}n_{101}n_{110}n_{111}\\
&+n_{001}(n_{010}n_{011}n_{100}n_{101}\\
&+(n_{010}n_{011}n_{100}+n_{011}n_{100}n_{101}+n_{010}(n_{011}+n_{100})n_{101})n_{110})n_{111})\bar{x}_{000}\\
&+n_{001}n_{010}n_{011}n_{100}n_{101}n_{110}n_{111}(\bar{x}_{001}+\bar{x}_{010}-\bar{x}_{011}+\bar{x}_{100}-\bar{x}_{101}-\bar{x}_{110}+\bar{x}_{111})
\end{aligned}$$

$$\begin{aligned}
\hat{\beta}_1\cdot\alpha = n_{000}(&n_{010}n_{011}n_{100}n_{101}n_{110}n_{111}(-\bar{x}_{000}+\bar{x}_{100})\\
&+n_{001}(n_{011}n_{100}n_{101}n_{110}n_{111}(-\bar{x}_{000}+\bar{x}_{100})\\
&+n_{010}(n_{100}n_{101}n_{110}n_{111}(-\bar{x}_{000}+\bar{x}_{100})\\
&+n_{011}(-n_{100}(n_{101}n_{110}+(n_{101}+n_{110})n_{111})(\bar{x}_{000}-\bar{x}_{100})\\
&+n_{101}n_{110}n_{111}(-\bar{x}_{001}-\bar{x}_{010}+\bar{x}_{011}+\bar{x}_{101}+\bar{x}_{110}-\bar{x}_{111})))))\\
&+n_{001}n_{010}n_{011}n_{100}n_{101}n_{110}n_{111}(-\bar{x}_{001}-\bar{x}_{010}+\bar{x}_{011}+\bar{x}_{101}+\bar{x}_{110}-\bar{x}_{111})
\end{aligned}$$

$$\hat{\beta}_2 \cdot \alpha = n_{000}(n_{010}n_{011}n_{100}n_{101}n_{110}n_{111}(-\bar{x}_{000} + \bar{x}_{010}) + n_{001}(-n_{010}(n_{011}n_{100}n_{101}n_{110} + (n_{011}n_{100}n_{101} + n_{100}n_{101}n_{110} + n_{011}(n_{100} + n_{101})n_{110})n_{111})(\bar{x}_{000} - \bar{x}_{010}) + n_{011}n_{100}n_{101}n_{110}n_{111}(-\bar{x}_{001} + \bar{x}_{011} - \bar{x}_{100} + \bar{x}_{101} + \bar{x}_{110} - \bar{x}_{111}))) + n_{001}n_{010}n_{011}n_{100}n_{101}n_{110}n_{111}(-\bar{x}_{001} + \bar{x}_{011} - \bar{x}_{100} + \bar{x}_{101} + \bar{x}_{110} - \bar{x}_{111})$$

$$\hat{\beta}_3 \cdot \alpha = n_{000}(-n_{001}(n_{010}n_{011}n_{100}n_{101}n_{110} + n_{011}n_{100}n_{101}n_{110}n_{111} + n_{010}(n_{011}n_{100}n_{101} + n_{100}n_{101}n_{110} + n_{011}(n_{100} + n_{101})n_{110})n_{111})(\bar{x}_{000} - \bar{x}_{001}) + n_{010}n_{011}n_{100}n_{101}n_{110}n_{111}(-\bar{x}_{010} + \bar{x}_{011} - \bar{x}_{100} + \bar{x}_{101} + \bar{x}_{110} - \bar{x}_{111})) + n_{001}n_{010}n_{011}n_{100}n_{101}n_{110}n_{111}(-\bar{x}_{010} + \bar{x}_{011} - \bar{x}_{100} + \bar{x}_{101} + \bar{x}_{110} - \bar{x}_{111})$$

$$\begin{aligned}\hat{\gamma}_{12} \cdot \alpha &= n_{001}n_{010}n_{011}n_{100}n_{101}n_{110}n_{111}(\bar{x}_{001} - \bar{x}_{011} - \bar{x}_{101} + \bar{x}_{111}) \\ &+ n_{000}(n_{010}n_{011}n_{100}n_{101}n_{110}n_{111}(\bar{x}_{000} - \bar{x}_{010} - \bar{x}_{100} + \bar{x}_{110}) \\ &+ n_{001}(n_{011}n_{100}n_{101}n_{110}n_{111}(\bar{x}_{001} - \bar{x}_{011} - \bar{x}_{101} + \bar{x}_{111}) \\ &+ n_{010}(n_{100}n_{101}n_{110}n_{111}(\bar{x}_{000} - \bar{x}_{010} - \bar{x}_{100} + \bar{x}_{110}) \\ &+ n_{011}(n_{101}n_{110}n_{111}(\bar{x}_{001} - \bar{x}_{011} - \bar{x}_{101} + \bar{x}_{111}) + n_{100}(n_{110}n_{111}(\bar{x}_{000} - \bar{x}_{010} - \bar{x}_{100} + \bar{x}_{110}) \\ &+ n_{101}(n_{110}(\bar{x}_{000} - \bar{x}_{010} - \bar{x}_{100} + \bar{x}_{110}) + n_{111}(\bar{x}_{001} - \bar{x}_{011} - \bar{x}_{101} + \bar{x}_{111})))))))\end{aligned}$$

$$\hat{\gamma}_{23} \cdot \alpha = n_{001}n_{010}n_{011}n_{100}n_{101}n_{110}n_{111}(\bar{x}_{100} - \bar{x}_{101} - \bar{x}_{110} + \bar{x}_{111}) + n_{000}(n_{010}n_{011}n_{100}n_{101}n_{110}n_{111}(\bar{x}_{100} - \bar{x}_{101} - \bar{x}_{110} + \bar{x}_{111}) + n_{001}(n_{011}n_{100}n_{101}n_{110}n_{111}(\bar{x}_{100} - \bar{x}_{101} - \bar{x}_{110} + \bar{x}_{111}) + n_{010}(n_{011}(n_{100}n_{101}n_{110} + n_{101}n_{110}n_{111} + n_{100}(n_{101} + n_{110})n_{111})(\bar{x}_{000} - \bar{x}_{001} - \bar{x}_{010} + \bar{x}_{011}) + n_{100}n_{101}n_{110}n_{111}(\bar{x}_{100} - \bar{x}_{101} - \bar{x}_{110} + \bar{x}_{111}))))$$

$$\begin{aligned}\hat{\gamma}_{13} \cdot \alpha &= n_{001}n_{010}n_{011}n_{100}n_{101}n_{110}n_{111}(\bar{x}_{010} - \bar{x}_{011} - \bar{x}_{110} + \bar{x}_{111}) \\ &+ n_{000}(n_{010}n_{011}n_{100}n_{101}n_{110}n_{111}(\bar{x}_{010} - \bar{x}_{011} - \bar{x}_{110} + \bar{x}_{111}) \\ &+ n_{001}(n_{011}n_{100}n_{101}n_{110}n_{111}(\bar{x}_{000} - \bar{x}_{001} - \bar{x}_{100} + \bar{x}_{101}) \\ &+ n_{010}(n_{100}n_{101}n_{110}n_{111}(\bar{x}_{000} - \bar{x}_{001} - \bar{x}_{100} + \bar{x}_{101}) \\ &+ n_{011}(n_{101}n_{110}n_{111}(\bar{x}_{010} - \bar{x}_{011} - \bar{x}_{110} + \bar{x}_{111}) \\ &+ n_{100}(n_{101}(n_{110} + n_{111})(\bar{x}_{000} - \bar{x}_{001} - \bar{x}_{100} + \bar{x}_{101}) + n_{110}n_{111}(\bar{x}_{010} - \bar{x}_{011} - \bar{x}_{110} + \bar{x}_{111}))))))\end{aligned}$$

The LOD score is then $\log \mathcal{L}(H_{3I}) - \log \mathcal{L}(H_{2I})$

$$LOD = \frac{n}{2} \log \Bigg(\sum_{i \in G_{000}} (x_i - \hat{\mu})^2 + \sum_{i \in G_{100}} (x_i - \hat{\mu} - \hat{\beta}_1)^2 + \sum_{i \in G_{010}} (x_i - \hat{\mu} - \hat{\beta}_2)^2 + \sum_{i \in G_{001}} (x_i - \hat{\mu} - \hat{\beta}_3)^2 + \sum_{i \in G_{110}} (x_i - \hat{\mu} - \hat{\beta}_1 - \hat{\beta}_2 - \hat{\gamma}_{12})^2 + \sum_{i \in G_{011}} (x_i - \hat{\mu} - \hat{\beta}_2 - \hat{\beta}_3 - \hat{\gamma}_{23})^2 + \sum_{i \in G_{101}} (x_i - \hat{\mu} - \hat{\beta}_1 - \hat{\beta}_3 - \hat{\gamma}_{13})^2 + \sum_{i \in G_{111}} (x_i - \hat{\mu} - \hat{\beta}_1 - \hat{\beta}_2 - \hat{\beta}_2 - \hat{\gamma}_{12} - \hat{\gamma}_{23} - \hat{\gamma}_{13})^2 \Bigg) - \frac{n}{2} \log \Bigg(\sum_{i \in G_{000}} (x_i - \bar{x}_{000})^2 + \sum_{i \in G_{100}} (x_i - \bar{x}_{100})^2 + \sum_{i \in G_{010}} (x_i - \bar{x}_{010})^2 + \sum_{i \in G_{001}} (x_i - \bar{x}_{001})^2 + \sum_{i \in G_{110}} (x_i - \bar{x}_{110})^2 + \sum_{i \in G_{011}} (x_i - \bar{x}_{011})^2 + \sum_{i \in G_{101}} (x_i - \bar{x}_{101})^2 + \sum_{i \in G_{111}} (x_i - \bar{x}_{111})^2 \Bigg) \tag{2.6}$$

This LOD score can then be used to investigate the presence of QTL × QTL × QTL interactions or QTL × QTL × environment interactions.

2.3 Results

2.3.1 Growth rate is not a strong predictor of overall growth

Three growth parameters, doubling time, lag time, and biomass accumulated, were measured for parental haploid strains S96 (denoted by 'S') and YJM789 (denoted by 'Y'), and 157 of their haploid meiotic segregants, separately in the presence of fructose, glycerol, and maltose as the sole carbon source (see methods below). Figure 2.1 schematically illustrates the effect of separately varying each growth parameter.

We observed that doubling time and lag time consistently showed a high correlation of $r^2 \approx 0.7$ to 0.8 (Table 2.1). In contrast, doubling time and lag time both showed a wide range of correlations with biomass ($r^2 \approx 0.4$ to 0.8), indicating that the growth rate (doubling time or lag time) does not predict overall growth (accumulated biomass). In agreement with these results, we find substantial overlap between doubling time and lag time QTLs, whereas these QTL do not have a strong effect on biomass (see below).

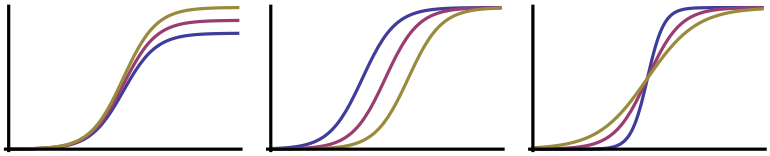


Figure 2.1: **Yeast growth curves are fit to sigmoidal curves that are a function of three growth parameters.** The figure shows the effect of varying each growth parameter independently: maximum OD, corresponding to overall biomass, (left), time to exponential growth, corresponding to time spend in the lag phase of growth (center), and doubling time (right).

Our results indicate that growth is a multi-variable phenotype, and that growth rate (lag time and doubling time) and biomass capture different aspects of growth.

Table 2.1: **Correlations between Doubling Time, Lag Time, and Biomass in a fixed environmental condition**

Environmental Condition	Doubling Time vs Lag Time	Doubling Time vs Biomass	Lag Time vs Biomass
Glycerol	0.73	0.41	0.40
Maltose	0.82	0.54	0.56
Fructose	0.77	0.79	0.57

This table lists correlation coefficients calculated between different growth phenotypes, calculated over both parental strains S and Y and 157 segregants. The calculation was repeated in the three media conditions. In all conditions, the three growth parameters are not strongly correlated, suggesting that growth is a multi-dimensional phenotype and each phenotype should be considered separately.

2.3.2 Yeast grown in glycerol and fructose show similar growth patterns

We observed that fructose, a readily fermentable sugar and glycerol, a non-fermentable sugar, showed the highest correlation for all growth parameters. In contrast, growth parameters in fructose and maltose were not correlated, even though they are both readily fermentable sugars (Table 2.2). The higher correlation between growth phenotypes in glycerol and fructose can be explained by the common QTLs and QTL-QTL interactions that we identified in these conditions (see below).

The difference between the growth parameters of the S and Y parents was largest in

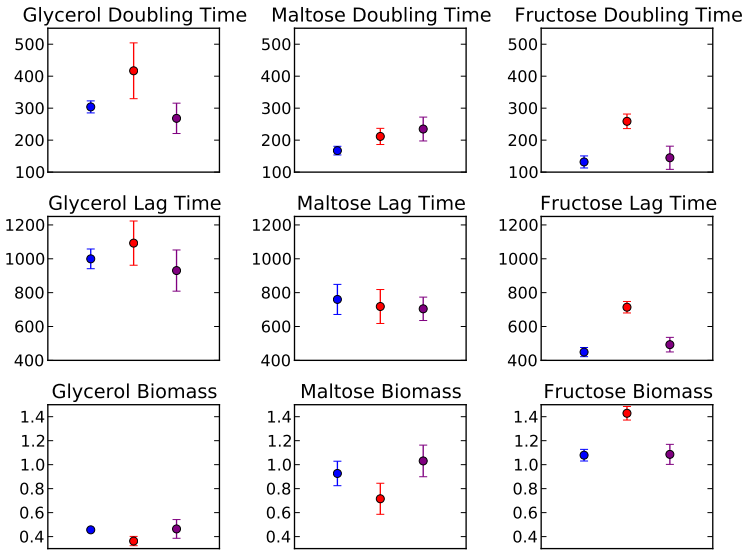


Figure 2.2: **Average growth phenotypes for parental strains S96 (blue), YJM789 (red) and 157 segregants (purple).** Error bars represent $\pm 1SE$. Varying the sugar source results in marked differences in yeast growth parameters.

fructose, where Y showed a longer doubling time, longer lag time and a higher maximum OD than S, indicating that Y has a preference for growth efficiency (quantified by peak biomass attained) over growth rate in the presence of fructose (Figures 2.2 and 2.3). In both glycerol and fructose, growth was slower in Y than in S (as quantified by doubling time and lag time). No difference was observed between the parental strains in Maltose. The S strain showed similar growth patterns in maltose and fructose, whereas the Y strain grew slower but with a higher efficiency in fructose compared to maltose.

2.3.3 A QTL in *FLO8* increases growth efficiency and decreases growth rate for the Y allele in glycerol and fructose

We identified a strong QTL in fructose and glycerol, with the LOD score peaking consistently at chr 5, position 377,186 bp, which is located in the gene *FLO8* (Figure 2.4). This QTL has a strong effect on doubling time, lag time, and biomass in fructose (LOD = 6.93, 15.88, 6.68 respectively), and on lag time in glycerol (LOD = 6.60) (Table 2.3, Figures 2.5 and 2.6). The 1.5 LOD support interval for this QTL extends over a 15

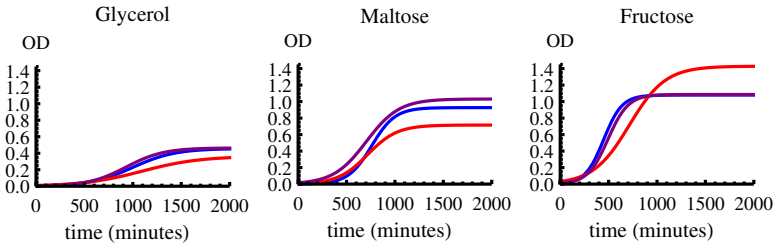


Figure 2.3: **Average growth curves for parental strains S96 (blue), YJM789 (red) and 157 segregants (purple).** These curves were generated using the average values of doubling time, lag time, and biomass in each condition. The segregants are more similar to S96 than to YJM789 in their growth profile.

Table 2.2: **Correlations of growth phenotypes across Glycerol, Maltose, and Fructose rich environments**

Growth Parameter	Glycerol vs Maltose	Maltose vs Fructose	Glycerol vs Fructose
Doubling Time	0.05	-0.09	0.28
Lag Time	0.10	-0.01	0.66
Biomass	0.13	0.02	0.20

This table lists correlation coefficients for a given growth phenotype between different environmental conditions, calculated over both parents and 157 segregants. Low correlations indicate high dissimilarity in growth phenotypes between Glycerol and Maltose, and between Maltose and Fructose. All growth phenotypes show a low to medium correlation between Glycerol and Fructose.

kb region from 364,321 bp to 379,328 bp, containing genes *SSA4, RTT105, NUP157, MAM1, GLE2, FLO8, KAP123* and the peak LOD score is consistently identified at the same SNP in the *FLO8* gene across four different phenotypes (Table 2.3). Hence, we refer to this locus as the *FLO8* locus in the rest of this manuscript.

FLO8 is a transcriptional activator that binds to the promoter of *FLO11*, a gene that is required for filamentous growth (pseudohyphal growth in diploids and invasive growth in haploids) [51], [52]. Activation of *FLO8* contributes towards filamentous growth [51], [53] and is required for flocculation [54]. The S96 strain has a nonsense mutation in this gene that prevents haploid invasive growth, and functional *FLO8* transformants of S96 show a partial haploid invasion phenotype [55]. Our results show that, apart from its known role in filamentous growth, *FLO8* may have a carbon source

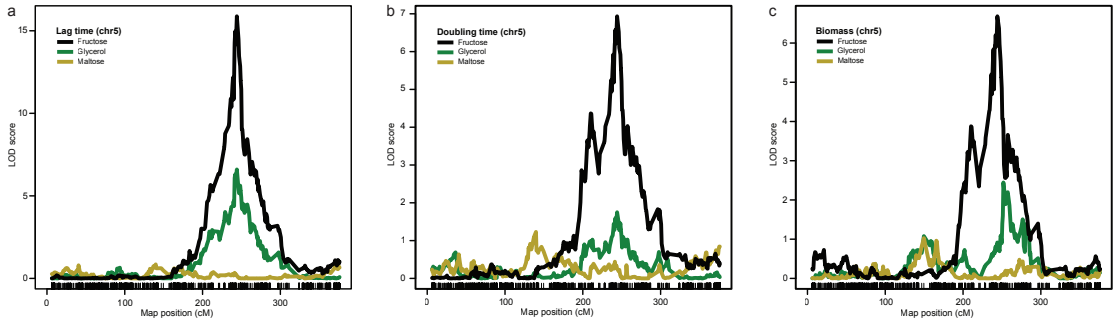


Figure 2.4: **Chromosome 5 LOD Curves indicate the presence of the *FLO8* QTL in Fructose (lag time, doubling time, biomass) and Glycerol (lag time).**

Table 2.3: **Single environment QTLs**

	Glycerol	Maltose	Fructose
Doubling Time			chr5@377,186 (6.93, 21.3%) $p < 0.001$
Lag Time	chr5@377,186 (6.60, 19.4%) $p < 0.001$	chr12@30,768 (3.69, 11.5%) $p < 0.04$	chr5@377,186 (15.88, 42.3%) $p < 0.001$
Biomass		chr2@83,942 (3.89, 12.2%) $p < 0.05$	chr5@377,186 (6.68, 20.6%) $p < 0.001$

This table lists QTLs that were identified as significant ($p < 0.05$) based on 1,000 permutation tests. Each entry lists the chromosome position (in bp), LOD score, and permutation test p-value.

specific effect on growth kinetics.

The *FLO8* QTL contributed to 42.3% of the phenotypic variance in lag time in fructose, and 19.4 % of the variance in lag time in glycerol. However, it showed no significant effect on any growth phenotype in maltose (Table 2.3 and Figure 2.6). The conclusion from these observations is that this QTL contributes to the high correlation in lag time between fructose and glycerol noted previously (Table 2.2).

The *FLO8* QTL was identified as having a large effect ($LOD > 6.9$) on all growth parameters in fructose (Table 2.3 and Figure 2.5). This observation is consistent with

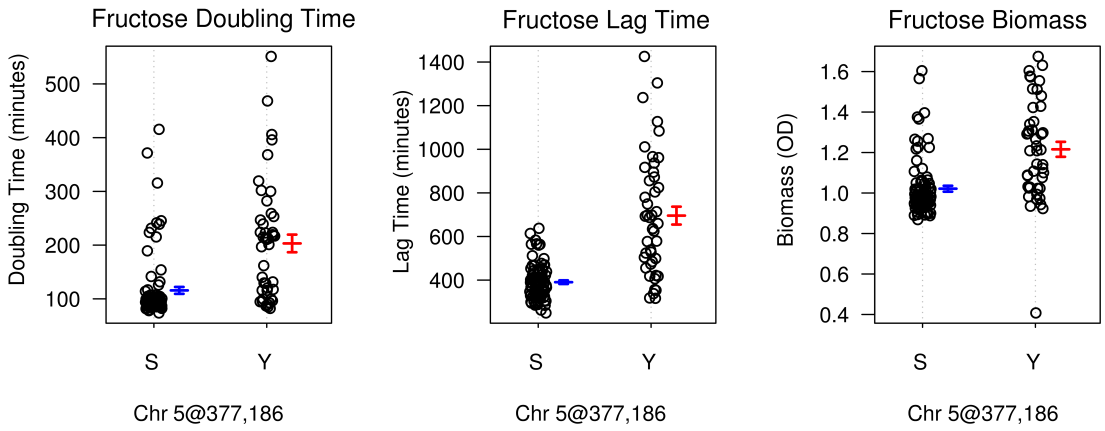


Figure 2.5: **Effect of *FLO8* QTL on all growth parameters for strains grown in the fructose media condition.** S and Y correspond to S96 and YJM789 allele at the marker. The Y allele at this marker is associated with a preference for growth efficiency over growth rate, as evidenced by longer doubling time and lag time, and higher biomass.

the fact that there were high correlations among all growth parameters for strains grown in fructose (Table 2.1). Furthermore, the direction of the effect of this QTL is consistent with the interpretation that the Y strain has adapted to prioritize growth efficiency over growth rate in fructose (Figure 2.5).

In maltose, we identified a QTL on chr 2 that had an intermediate effect on biomass (LOD = 3.89), with the LOD score peaking at position 83,492 bp (1.5 LOD support interval = 77,256 bp to 90,683 bp). This QTL contributed to 12.2% of the variance in biomass in maltose and the interval included the genes *NUP170*, *ATG8*, *ILS1*, *SSA3*, *AAR2* and *RPS8A*. The gene *SSA3*, an ATPase belonging to the *Hsp70* family, is known to interact with the maltose-responsive transcriptional activator *MAL63* [56], and is therefore a candidate that may contribute to biomass variation.

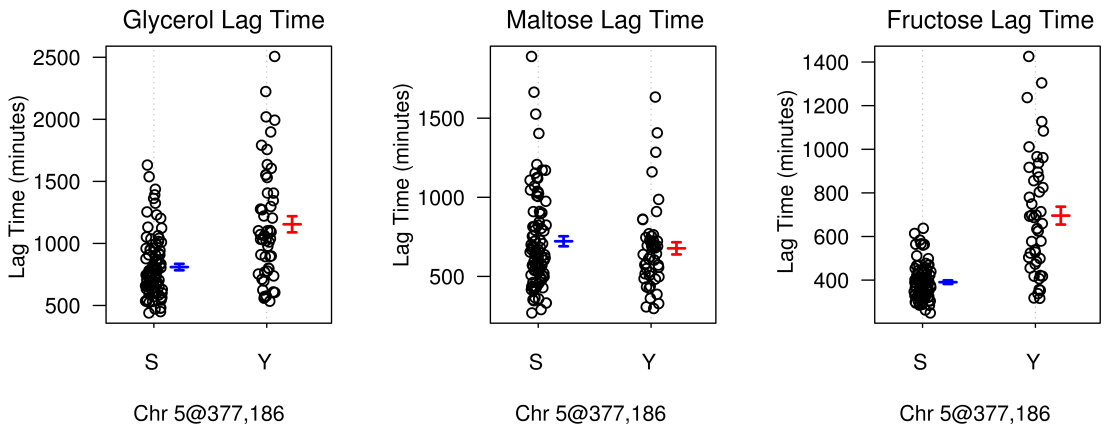


Figure 2.6: **Effect of *FLO8* QTL on lag time for all three media conditions.** S and Y correspond to S96 and YJM789 allele at the marker. This QTL has a significant effect in glycerol and fructose media conditions.

2.3.4 Many QTLs reversed their phenotypic effect with a change in carbon source (antagonistic gene-environment interaction)

We mapped QTLs that interact with the environment (GEI QTLs) for all growth parameters across all three pairs of environments. These GEI QTLs fall into three categories: scale effect QTLs (whose effect is in the same direction in the two environments), antagonistic effect QTLs (whose effect is in the opposite direction in the two environments), and environment specific QTLs (whose effect is only present in a single environment) (Figure 2.7).

We observed a large number of GEI QTL that were not identified in single QTL mapping (Table 2.4). Many of these QTLs, while they had a small (i.e. not significant) effect in each individual environment, yet they interacted antagonistically between environments, leading to their identification as GEI QTLs. Hence, our results demonstrated that a more comprehensive set of QTLs contributing to phenotypic plasticity can be identified by mapping gene-environment interactions, as QTLs that have small individual effects can have significant pleiotropic effects when environmental conditions are varied.

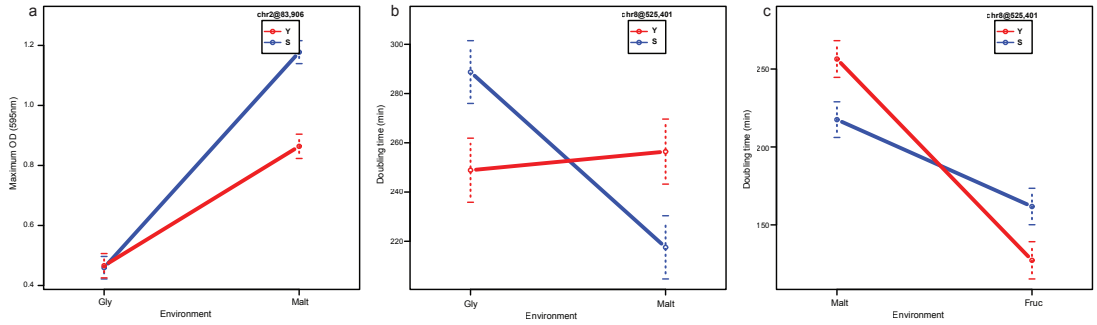


Figure 2.7: **Phenotype effect plots of GEI QTLs demonstrate clear gene-environment interactions.** (a) A chr2 QTL showing a Maltose specific effect. (b and c) A chr 8 QTL showing an antagonistic interaction (crossover) where the S allele has a larger doubling time in Fructose and Glycerol, and Y allele has a larger doubling time in Maltose.

2.3.5 Growth is modulated by common QTLs of similar effect in glycerol and fructose

We observed that eleven GEI QTLs were responsible for the differences in growth rate in glycerol or fructose and maltose (Table 2.4). In contrast, only two GEI QTLs were identified when comparing growth rate in glycerol to that in fructose. These results demonstrate that even among small effect QTLs, similar loci affect growth rates in glycerol and fructose. This corroborates well with the higher correlation among growth phenotypes in glycerol and fructose, and the fact that the *FLO8* QTL was significant even in single environment mapping in these conditions.

2.3.6 *FLO8* regulates growth through environment specific and growth parameter specific epistatic interactions with loci on chr 7 and chr 13

In each medium, we examined pairs of single environment QTL candidates of any effect size ($p < 1.0$, permutation test) to test significance for QTL-QTL interactions ($p < 0.05$, permutation test). This analysis showed that QTLs of weak independent

Table 2.4: **Gene-Environment Interaction QTLs**

	Glycerol vs Maltose	Glycerol vs Fructose	Maltose vs Fructose
Doubling Time	chr2@90,357 (1.93) $p = 0.01$		chr2@91,797 (3.51) $p < 0.01$
			chr3@105,155 (2.50) $p < 0.01$
	chr4@22,874 (1.69) $p = 0.04$		
	chr5@364,321 (1.78) $p = 0.04$		chr5@364,321 (4.49) $p < 0.01$
	chr8@525,401 (2.04) $p = 0.01$		chr8@525,401 (2.17) $p = 0.01$
	chr10@394,687 (2.12) $p = 0.01$	chr10@404,421 (2.68) $p = 0.01$	
			chr11@219,509 (2.29) $p = 0.01$
			chr12@388,274 (2.11) $p = 0.01$
	chr13@407,190 (1.79) $p = 0.03$		
		chr14@411,669 (2.06) $p = 0.03$	
	chr16@398,558 (2.76) $p = 0.01$		chr16@347,462 (1.85) $p = 0.03$
Lag Time	chr2@90,194 (3.22) $p = 0.01$		chr2@90,194 (3.02) $p < 0.01$
			chr3@74,327 (2.24) $p = 0.01$
	chr4@22,874 (2.67) $p = 0.01$		
	chr5@364,321 (5.25) $p < 0.01$		chr5@364,321 (7.01) $p < 0.01$
			chr8@521,191 (2.28) $p < 0.01$
			chr12@386,022 (2.49) $p < 0.01$
	chr13@163,283 (2.58) $p = 0.02$		chr13@159,055 (1.98) $p = 0.04$
Biomass	chr2@83,906 (3.74) $p = 0.03$		chr2@83,906 (3.79) $p = 0.03$
		chr3@97,696 (2.72) $p = 0.01$	
		chr5@378,582 (2.24) $p = 0.04$	

This table lists QTLs showing a gene-environment interaction that was identified as significant ($p < 0.05$) using 100 permutation tests. The presence of a GEI QTL indicates the presence of a genotype environment interaction. Each entry lists the chromosome position (in bp), LOD score, and permutation test p-value.

effect (i.e. they did not pass genome-wide significance threshold as a single-environment QTL) were involved in statistically significant QTL-QTL interactions (Table 2.5).

In glycerol and fructose, a common interaction was identified between the *FLO8* locus and a locus on chr13, affecting lag time in glycerol (interaction LOD score = 2.6, Figure 2.8), and fructose (interaction LOD score = 3.1, Figure 2.8) (Table 2.5). The 1.5 LOD support interval for the chr13 QTL extended over a 31 kb region from 146,198 to 177,237 bp. In both media conditions, the interaction demonstrated that the strains with the S allele at chr13 and the Y allele at chr5 had the longest lag time (Figure 2.8).

In glycerol, the *FLO8* locus interacted strongly with a locus on chr7 to alter biomass, peaking between 26,299 to 33,274 bp (interaction LOD score = 3.26, Table 2.5). This interaction had a dominant effect on biomass, leading to a 1.5 fold increase in maximum optical density for strains with the Y allele at chr5 and S allele at chr7 (Figure 2.9). However, as the chr7 QTL had a weak effect in a single environment (LOD = 1.6), we could not measure the 1.5 LOD support interval (a 1.0 LOD support interval extended

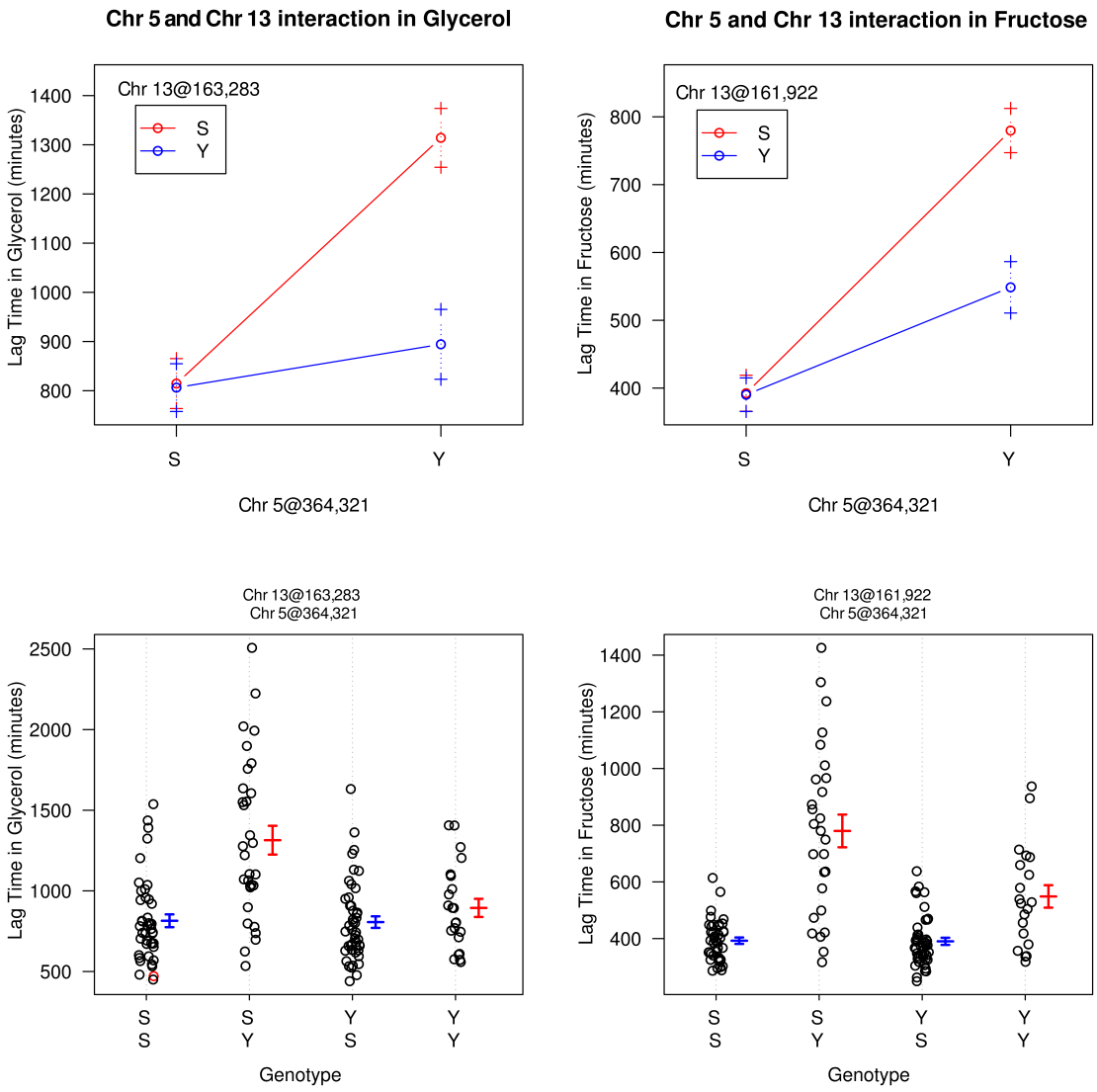


Figure 2.8: **An interaction between the *FLO8* QTL and a locus on chr13 regulates lag time in the presence of Glycerol and in Fructose.** S and Y correspond to S96 and YJM789 allele at each marker. Strains with the S allele at chr13 and the Y allele at chr5 have a longer lag time.

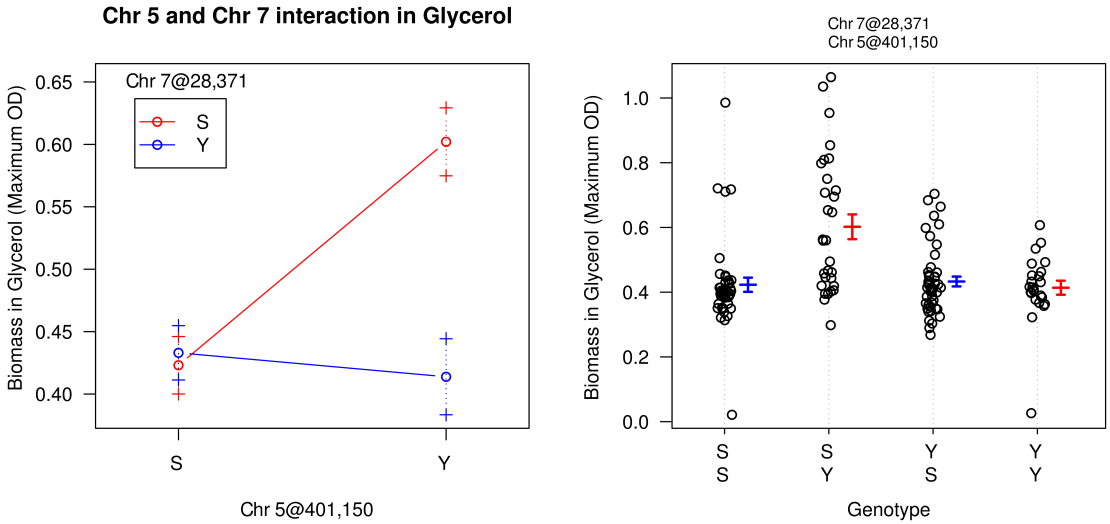


Figure 2.9: **An interaction between the *FLO8* QTL and a locus on chr7 regulates biomass in the presence of Glycerol.** S and Y correspond to S96 and YJM789 allele at each marker. Strains with the S allele at chr7 and the Y allele at chr5 have a higher biomass.

from 9.4 to 67.38 kb). Furthermore, the interaction between the *FLO8* locus and the chr13 locus arose only when mapping lag time in glycerol and fructose (Table 2.5). Similarly, we identified QTL-QTL interactions that are only associated with doubling time (Table 2.5).

Thus, epistatic QTL interactions described above were both environment specific and parameter specific, highlighting the ability of yeast populations to independently vary separate aspects of growth in response to their environment.

2.4 Discussion

Non-motile organisms need to exhibit high phenotypic plasticity in order to adapt to changing environments. Growth is a key phenotype which should exhibit such plasticity in these organisms. For *Saccharomyces cerevisiae*, the availability and utilizability of carbon sources affects every aspect of growth (Reviewed in [40]), and many genes are known to respond to a change in the type and level of carbon sources [57], [58]. Attempts have been made to identify loci contributing to variation in yeast growth

Table 2.5: **QTL-QTL interactions**

	Glycerol	Maltose	Fructose
Doubling Time	chr8@26,177 - chr8@525,401 (2.565) $p = 0.035$		chr11@219,509 - chr12@96,033 (3.018) $p = 0.02$
Lag Time	chr5@364,321 - chr13@163,283 (2.601) $p = 0.012$ chr5@377,186 - chr13@163,283 (2.379) $p = 0.029$		chr5@364,321 - chr13@161,922 (3.101) $p = 0.009$ chr5@377,186 - chr13@161,922 (2.519) $p = 0.04$
Biomass	chr5@401,150 - chr7@28,371 (3.258) $p = 0.007$		

This table lists QTL-QTL interactions that were identified as significant ($p < 0.05$) based on 1,000 permutation tests. Each entry lists the chromosome positions of the interacting SNPs (in bp), LOD score, and permutation test p-value.

under different environmental conditions [43], [44]. However, the mechanisms by which genetic interactions affecting different aspects of yeast growth are modulated by nature of carbon sources is not well understood.

In the present study, we mapped variation in phenotypic plasticity across three carbon sources for three growth parameters. Using two genetically divergent yeast strains, S96 (a lab strain) and YJM789 (a clinical isolate), and their meiotic recombinants, we found a variety of loci associated with carbon source dependent phenotypic plasticity for all three growth parameters. We also found that this variation was attributable to different sets of GEI QTLs for growth rate (lag time and doubling time) and for biomass. We identified epistatic interactions between some of these GEI QTLs which contributed to their environment specificity.

Our study identified a strong QTL at chr5, position 377,186 bp, located in the *FLO8* gene which had different effects on doubling time, lag time, and biomass in fructose and glycerol. It is known that the *FLO8* gene is functional in YJM789 but not in S96 [55]. Our study found that the Y allele of this locus contributes to a slower growth rate but higher biomass accumulation in fructose and glycerol, but not in maltose. Furthermore, we find that *FLO8* regulates growth kinetics through distinct, environment-specific interactions with other QTLs, that separately affect growth rate and biomass.

2.4.1 Gene-environment interactions demonstrated scale effects, environment specific effects, and crossover effects

By mapping gene-environment interactions across pairwise comparisons of growth media, we identified GEI QTLs in the following three categories: a) scale effect interactions occur when single QTLs contributed to variation in growth in both environments, with different effect sizes. An example is the *FLO8* locus across glycerol and fructose which affects both doubling time and lag time; b) environment-specific interaction QTLs. These contribute to growth in only one of the two media (to the limit of our mapping resolution). The maltose specific chr2 locus affecting biomass showed such an effect; c) crossover effect QTLs. In this case, one parental allele increases the phenotype in one condition, and the other parental allele increases the phenotype in another condition. Six of the twelve GEI QTLs fall into this category. Such crossover interactions can occur if one allele is sensitive to an environmental variable and the other allele shows a resilient phenotype across environment. Alternatively, the parental strains may have adapted to different environmental conditions at these loci.

Many genome-wide significant GEI QTLs were not significant QTLs in single environment mapping. Instead, these GEI QTLs had a weak effect in a specific environment, but had a significant effect when the carbon source was varied, in the form of a crossover effect of the phenotype. This result has two broad implications. First, it stresses the added statistical power of GEI mapping to identify loci involved in crossover interactions. Secondly, it shows that the differences that we identified between growth in fructose/glycerol and growth in maltose are regulated by multiple crossover interactions.

2.4.2 Growth in functionally dissimilar carbon sources, fructose and glycerol, is regulated by common QTLs and epistatic interactions

Fructose, a monosaccharide, and maltose, a disaccharide, are both readily fermentable sugars and support faster growth rate and larger biomass than in glycerol, a non-fermentable carbon source. This is readily seen for both the lab strain S96 and the

clinical isolate YJM789 (Figure 2.2).

The high phenotypic correlations observed for all growth parameters when comparing fructose-rich and glycerol-rich environments can be partly explained by the overlap in QTLs and QTL-QTLs interactions in these conditions. The *FLO8* locus contributed to variation in lag time in glycerol, and to all three growth parameters in fructose (explaining 42.3% of the variation in lag time). On the other hand, two distinct QTLs (on chr2 and chr12) were identified for growth in maltose (Table 4.5).

The *FLO8* locus was not identified on mapping in glucose, also a readily fermentable sugar (results not presented). Furthermore, we identified a common interaction between *FLO8* and a chr13 QTL for lag time in glycerol and fructose, whereas no significant QTL-QTL interactions were detected in maltose.

Our results support the interpretation that *FLO8* regulates growth kinetics in a similar manner in a non-fermentable carbon source, glycerol, and in a fermentable carbon source, fructose. This mechanism of growth regulation is absent in other fermentable carbon sources like glucose and maltose. *FLO8* is an invasive growth specific transcription factor, and although expressed throughout growth kinetics, it activates invasive growth only during nutrient limitation (i.e. in the absence of a fermentable carbon source). In our study, the *FLO8* locus was found to affect growth parameters in nutrient rich conditions (lag phase and exponential phase in fructose) through interactions conserved across nutrient limited conditions (glycerol).

A large number of gene-environment interactions were identified in our study when comparing fructose or glycerol to maltose, but not when comparing fructose to glycerol. This suggests that, apart from the *FLO8* locus, many other alleles differentially affect growth for the two fermentable carbon sources (Table 2.4). Many of these GEI QTLs showed crossover interactions, indicating that the S and Y alleles are antagonistically adapted to the two fermentable carbon sources.

2.4.3 Yeast regulates growth rate and biomass through different sets of QTLs and epistatic interactions

Yeast growth is usually measured as a single gross phenotype. Previous studies have shown that different phases of growth are differentially affected by various environmental conditions [48], [43].

In our study, we noted that while there was a high correlation between lag time and doubling time, these parameters showed a low correlation with total biomass accumulated. This implies two things: a) the lag time (i.e. time for yeast to adapt to a nutrient condition and enter exponential phase) is highly predictive of the doubling time; b) the growth rate is not a good predictor of the overall biomass accumulated. Hence our study emphasizes that growth is characterized by growth rate and growth yield, and variations in these traits have a partially overlapping genetic basis. We identified a QTL that regulates all three phases of growth (the *FLO8* locus in fructose), as well as QTLs having a parameter specific affect. Gene-environment mapping identified partially overlapping sets of QTLs that regulate growth parameter specific interactions between glycerol/fructose and maltose (Table 2.4).

The *FLO8* locus interacts with a locus on chr13 to affect lag time in fructose and glycerol (Figure 2.8), and with a locus on chr7 to affect biomass in glycerol (Figure 2.9). This demonstrates that a common gene (*FLO8*) can have different genetic interactors that differentially regulate lag time and doubling time. Our study indicates that growth kinetic parameters in *S. cerevisiae* are differentially regulated in an environment specific manner through interactions with a common regulator, *FLO8*. Such modular functional mechanisms may have provided yeast with the flexibility to alter each growth phase independently to optimize its fitness in the varied environments encountered in its evolutionary history.

2.4.4 Phenotypic plasticity arises through epistatic interactions between environment-specific QTLs (allelic sensitivity hypothesis) and environment-specific regulatory interactions (gene regulatory hypothesis)

Two models have been proposed to explain genetic control of phenotypic plasticity. The allelic sensitivity model proposes that the plasticity of a population is contributed by genes which directly alter the phenotype in changing environments. In contrast, the gene regulatory model suggests that regulatory genes render plasticity by activating or repressing structural genes in an environment specific manner [38]. The two hypotheses are not mutually exclusive and it is possible that both types of genes contribute to plasticity of a complex phenotype like growth. Our study supports such a complex scenario, because we find that the *FLO8* locus, a transcriptional regulator, affects growth rate and biomass directly (Figure 2.5, Table 2.3) as well as through epistatic interactions with other loci (Figures 2.8 and 2.8, Table 2.5).

In glycerol, the *FLO8* locus interacts strongly with a QTL on chr7 to affect biomass. This interaction increases the overall biomass accumulated by 50% (measured in OD) for the segregants with the Y/S combination of alleles at chr5/chr7. One of the genes in the chr7 QTL is *RTG2*, a transcription factor that senses mitochondrial dysfunction [59]. Glycerol is a respiratory medium, requiring the TCA cycle and the glyoxylate cycle activity for growth and *RTG2* is known to affect expression of the enzymes involved in these cycles [59]. Crucially, *RTG2* and *FLO8* both increase *FLO11* activity, resulting in an invasive growth under glucose limited conditions (Reviewed in [52]). Hence, it is possible that *RGT2* and *FLO8* interact either via *FLO11* or independently to affect biomass in glycerol. It has been shown that polymorphisms in transcription factors directly affecting the phenotype can have environment specific interactions [60]. Our study shows that epistatic interaction between loci not directly associated with growth can have carbon source specific effects on various growth parameters.

An example of a regulatory locus showing environmental specificity is the GEI QTL

on chr2 (Table 2.4). This locus was identified when maltose was compared to fructose/glycerol, but not when fructose was compared to glycerol. The 1.5 LOD support interval for the chr2 QTL extends over a 13 kb region containing the genes (NUP170, ATG8, ILS1, SSA3, AAR2 and RPS8A). One of these genes, *SSA3* has been shown to form a subcomplex with the maltose responsive transcription factor *MAL63*, in the presence of maltose [56]. This is consistent with our finding that this locus was identified in GEI mapping (in all growth parameters) specifically for comparisons involving maltose.

Summary: Yeast growth is a highly composite phenotype, and its various measures (biomass, growth rate and lag time) show plasticity in different carbon sources related to different QTLs. Much of this plasticity is a result of carbon source specific gene-gene interactions. Furthermore, these interactions are parameter specific suggesting that yeast has the ability to modulate different aspects of growth independently to maximize its fitness across varied environments. The candidate genes located in the QTLs identified in this study are both regulatory and structural genes which interact to contribute to variation in phenotypic plasticity.

2.5 Future Directions

Thus far, we have applied quantitative trait mapping to the phenotypic means of populations. The hypothesis being investigated is: does the genotype at a given set of loci affect the phenotypic mean of the samples? The standard assumption in mapping quantitative trait loci (QTL) is that the phenotypic variance remains unchanged [13]. However, there can be valid biological reasons for phenotypic variance to be affected by changes in the genotype. Indeed, it is quite likely that fluctuations (standard deviation) in phenotypic output may be proportional to the phenotypic abundance (mean). Alternatively, there may be gene-gene interactions present in a strain that regulate the phenotypic output. These regulation mechanisms may differ for strains and hence be disrupted in the hybrid offspring. In the above study, we observe environment-dependent differences in the phenotypic variance of certain growth traits. Specifically, the yeast strain YJM789 shows a larger phenotypic variance than S96 (haploid form of

S288c) for doubling time and lag time when grown with glycerol as the carbon source (Figure 2.2). This leads to the following biological question: do the genetic drivers of phenotypic variance differ from the drivers of phenotypic mean in these growth parameters?

We can address this question by mapping QTL that drive variation in the phenotypic variance of samples. Concretely, we can investigate the following three hypotheses, given by their corresponding likelihood functions.

$$\mathcal{L}(H_M) = \prod_i \phi(y_i|\mu + \beta g_i, \sigma^2)$$
$$\mathcal{L}(H_V) = \prod_i \phi(y_i|\mu, \sigma^2 + \alpha g_i)$$
$$\mathcal{L}(H_0) = \prod_i \phi(y_i|\mu, \sigma^2)$$

where ϕ is the density function for the normal distribution, and the parameters μ, σ^2, β and α are obtained for each hypothesis by maximizing the likelihood. H_M is the hypothesis that the genotype affects the phenotypic mean, H_V is the hypothesis that the genotype affects the phenotypic variance, and H_0 is the null hypothesis that genotype does not affect the phenotypic mean nor the phenotypic variance. In the equations above, if we assume the gaussian form for ϕ,

$$\phi(x_i|\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

the log likelihoods of the three hypotheses are then given by

$$\log \mathcal{L}(H_0) = \frac{n}{2}\log\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2$$
$$\log \mathcal{L}(H_M) = \frac{n}{2}\log\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}\left(\sum_{i\in G_1}(x_i - \mu_1)^2 + \sum_{i\in G_2}(x_i - \mu_2)^2\right)$$
$$\log \mathcal{L}(H_V) = \frac{n_1}{2}\log\left(\frac{1}{2\pi\sigma_1^2}\right) + \frac{n_2}{2}\log\left(\frac{1}{2\pi\sigma_2^2}\right) - \frac{1}{2\sigma_1^2}\left(\sum_{i\in G_1}(x_i - \mu)^2\right) - \frac{1}{2\sigma_2^2}\left(\sum_{i\in G_2}(x_i - \mu)^2\right)$$

For the null hypothesis, maximizing $\log \mathcal{L}(H_0)$ shows that the mean and variance estimators are the phenotypic mean and the phenotypic variance, respectively.

$$\hat{\mu} = \bar{x}$$
$$\hat{\sigma}^2 = \sigma^2$$

For the H_M hypothesis, maximizing $\log \mathcal{L}(H_M)$ gives us the parameter estimates:

$$\hat{\mu}_j = \bar{x}_j$$
$$\hat{\sigma}^2 = \sum_j \frac{n_j}{n} {\sigma_j}^2$$

For the H_V hypothesis, maximizing $\log \mathcal{L}(H_V)$ gives us that the parameter estimates:

$$\hat{\mu} = \bar{x}$$
$$\hat{\sigma_j}^2 = {\sigma_j}^2$$

Inserting these values back into the log likelihoods gives us the respective LOD scores:

$$LOD_{mean} = \log\left(\frac{\mathcal{L}(H_M)}{\mathcal{L}(H_0)}\right) = \frac{n}{2}\log\left(\frac{n\sigma^2}{n_1\sigma_1^2 + n_2\sigma_2^2}\right)$$
$$LOD_{variance} = \log\left(\frac{\mathcal{L}(H_V)}{\mathcal{L}(H_0)}\right) = \frac{n_1}{2}\log\left(\frac{\sigma^2}{\sigma_1^2}\right) + \frac{n_2}{2}\log\left(\frac{\sigma^2}{\sigma_2^2}\right)$$

where $\sigma^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$ is the overall variance, and σ_1^2 and σ_2^2 are the variance for samples with genotypes 1 and 2, respectively.

These log of odds scores $LOD_{variance}$ and LOD_{mean} can then be used to identify loci that contribute to variation specifically in the phenotypic variance or in the phenotypic mean. Furthermore, $LOD_{mean} - LOD_{variance} = \log\left(\frac{\mathcal{L}(H_M)}{\mathcal{L}(H_V)}\right)$ can identify which hypothesis is a better fit for a given genetic locus. We can use these LOD scores to address the question of whether the genetic drivers of the phenotypic variance differ from the drivers of the phenotypic mean. We can also extend these methods to an unexplored direction of QTL mapping where, analogously to the discussion in the Methods section of this chapter, we can identify gene-gene interactions and gene-environment interactions that may regulate phenotypic variance.

Chapter 3

Sporulation Genes Associated with Sporulation Efficiency in Natural Isolates of Yeast

3.1 Introduction

Sporulation is a response to nutrient deprivation in which yeast exits mitotic cell cycle and enters into meiosis, leading to spore formation [61]. About 400 genes have been shown to modulate sporulation [62], [63] and more than 1,000 genes are known to change expression during sporulation [64], [65]. Sporulation efficiency, defined as the fraction of cells that sporulate in a culture, varies among strains and has been identified as a quantitative trait that is modulated by at least 10 genes [66], [67], [68]. However, many of these studies have been performed using laboratory strains [66], [67], which face distinct selective pressures as compared to wild type strains.

The lack of information about traits in natural populations has limited our understanding of the potential effects of evolution, selection pressure, life history and environment on trait variation and its mechanism of action. Sporulation is triggered as a response to nutrient deprivation. As natural isolates face strong selection pressure to adapt to nutrient changes in their environment, it is reasonable that mechanisms causing variation in sporulation efficiency in natural isolates may be very different from those operating in laboratory strains.

Several previous studies have shown variation of sporulation efficiency among natural isolates of yeast, such as clinical, oak and wine strains [68], [69], [70], [71]. To understand this variation among a larger set of natural isolates and to identify some of the genetic factors contributing to this phenotype, we measured the sporulation efficiency of strains in the SGRP collection [72]. While a previous study has shown large variation in sporulation efficiency in SGRP strains [71], our goal was to examine

whether the genes that have been implicated in sporulation to date [61], [63] also contribute to sporulation efficiency variation in these SGRP strains. This would help us understand how sporulation efficiency variation is modulated in natural isolates from diverse environmental niches.

To identify loci associated with sporulation efficiency in the SGRP collection, we used two methods of association mapping (described below) on a set of 397 sporulation and sporulation-associated genes. After correcting for population structure in the SGRP strains, we identified two significant clusters of SNPs in strong linkage disequilibrium that were strongly associated with high sporulation efficiency. The SNPs were found in the genes *HOS4*, *MCK1*, *SET3*, *SPO74* and other candidate genes.

3.2 Materials and Methods

3.2.1 Yeast Strains and Culture Conditions

Yeast strains were obtained from the Saccharomyces Genome Resequencing Project (SGRP) [72]. All strains were grown under standard media and growth conditions. To measure sporulation efficiency, strains were first grown in YPD (yeast extract, peptone and dextrose) from a starting optical density (OD) at 600nm of 0.2 to a final OD of 1.0. Their cell cycle was then synchronized by growing them in YPA (yeast extract, peptone and acetate) from a starting OD of 0.2 to final OD of 1.0 at 30°C, shaking at 250rpm [73]. Approximately 1×10^7 cells from this synchronized culture were then incubated in liquid sporulation medium (1% potassium acetate supplemented with amino acid mixture) at 30°C for the duration of experiment.

3.2.2 Estimation of Sporulation Efficiency

For each strain, three biological replicates were used and approximately 1,000 cells were counted per replicate per strain. Sporulation efficiency was measured as the ratio of tetrads and dyads produced by a strain, to the number of cells (expressed as a percentage). For each strain, sporulation efficiency was measured every two days until saturation was reached for three consecutive readings.

3.2.3 Sequence Data

The sequence and SNP data for all strains was obtained from the SGRP project (`http://www.sanger.ac.uk/research/projects/genomeinformatics/sgrp.html`; downloaded in February 2012). Sequence alignments using the *Saccharomyces cerevisiae* genome as reference were performed for each gene being analyzed, starting from 500 base pairs upstream of the gene. Alignment was performed using the SGRP tool 'alicat.pl' (available for download at the SGRP database). Based in this alignment, variant loci were identified and were analyzed for association with the phenotype.

3.2.4 LOD Score Analysis

The data consisted of 42,003 SNPs with phenotype data for 32 strains. These SNPs were filtered to include only bi-allelic SNPs with no missing data and with minor allele frequency $\geq 2/32$, leaving 10,481 SNPs. For each SNP, we calculated the LOD score [13], which is the log (base 10) of the ratio of the likelihood of the data given the hypothesis that there is a QTL to the likelihood of the data given the hypothesis that there is no QTL at a locus. A LOD score of 3.0 implies that the likelihood that there is a QTL (i.e. the data are drawn from a distribution where the two genotypes have different phenotypic means) is 1,000 times greater than the likelihood that there is no QTL (i.e. the data are drawn from a distribution where the two genotypes have the same phenotypic mean). Let q_1 and q_2 be the fraction of strains having allele 1 and 2, respectively, and x be the total number of strains. Let v_1 and v_2 be the phenotype variances of strains with alleles 1 and 2, and v be the overall phenotype variance. Then, for each SNP, the LOD score is given by

$$LOD = \frac{x}{2}\log(\frac{v}{q_1 v_1 + q_2 v_2})$$

Permutation tests of up to 10^6 permutations were performed to assign an empirical p-value to each SNP. This test approximates the probability of observing a LOD score greater than or equal to a certain value, under the null hypothesis that there is no QTL at this SNP. To correct for multiple hypothesis testing (Bonferroni correction), we first grouped the 10,481 SNPs in the filtered data into clusters containing SNPs

that were in perfect linkage disequilibrium. This identified $n = 1709$ distinct clusters. The permutation test p-value was multiplied by n to obtain the Bonferroni corrected p-value. This left us with 2 clusters of SNPs with p-value < 0.03.

3.2.5 Binomial Analysis

As a check on the LOD analysis, we also performed a binomial test on the data. The data consisted of 42,003 variant loci in genes potentially associated with sporulation and a measured sporulation efficiency value for 32 strains. After retaining only bi-allelic SNPs with no missing data and restricting to loci with minor allele frequency (MAF) $> 5/32$ (0.16), 4,664 SNPs remained. The strains were stratified into 3 sets, broadly based on the sporulation efficiency classification used by Cubillos *et al.* [71], ranging from 0/1, 2 and 3. Set S1 contained 15 poor sporulation efficiency strains, with sporulation efficiency from 0% to 24%; set S2 contained 8 intermediate efficiency strains with sporulation efficiency from 25% to 74%, and S3 contained 9 high sporulation efficiency strains, with sporulation efficiency from 75% to 100%. Thus, the *a priori* probabilities for a strain chosen at random to belong to set S1, S2, and S3 were 0.47, 0.25 and 0.28 respectively.

For each allele, a binomial test was applied to determine whether an allele at a SNP was significantly associated with set S1 (low sporulation efficiency) or with set S3 (high sporulation efficiency).

Let n be the number of samples with the major alleles and k the number of major alleles in class S1. Also, let p to be the *a priori* probability for an allele to occur in class S1 (0.47). If there is no association between the major allele and low sporulation efficiency, the probability P of obtaining k or more major alleles in class S1 is given by:

$$P = \sum_{m=k}^{n} \binom{n}{m} p^m (1-p)^{n-m}$$

This is the p-value, or the probability of obtaining an association as extreme as the one seen in the data by chance, when the null hypothesis is true, i.e. when there is no association between the allele and sporulation efficiency. For the 4,664 SNPs

that remained after filtering, the p-value was computed as described above to test for the association of both minor and major alleles with high or low sporulation efficiency (4 comparisons per SNP). We used a significance threshold of $p < 0.05$. For our final results, we retained only those SNPs identified as statistically significant by the LOD score analysis and by the binomial test, as being associated with the sporulation phenotype (Table 3.2, Table S4 in [1]). Table S2 in [1] lists the LOD score, binomial test p-values, genotypes and mean phenotypes for the 69 SNPs that were identified with a LOD score > 2.5.

3.3 Results

3.3.1 Sporulation Efficiency Variation in SGRP collection strains

The sporulation efficiency of the 36 sequenced, genetically diverse and highly polymorphic *S. cerevisiae* strains in the SGRP collection showed extensive variation, ranging from zero for strains that did not sporulate: 322134S, 378604X, 273614N, YIIc17-E5, poor (1-25%) for DBVPG6044, K11, DBVPG1106, Y9, intermediate (25-49%) for DBVPG1788, YJM975, YJM978, high (50-74%) for Y12, Y55, BC187, DBVPG6040, L-1528, and very high (75-100%) for L-1374, UWOPS05-227.2, SK1, YPS606, YPS128 (Table 3.1 and Table S3 in [1]). Approximately one third (11 out of 32) of the strains failed to sporulate and their sporulation efficiency was set to zero in the association analysis. The inability of these isolates to sporulate may simply reflect the fact that the lab condition for temperature, media, aeration, etc. [74] used may not be appropriate for sporulation in these natural strains. Alternately, these strains may have inherently low sporulation efficiency and may have developed alternate mechanisms to cope with nutrient deprivation, e.g. pseudo-hyphae as in case of YJM981 and 322134S.

In addition to a wide spectrum of sporulation efficiencies, these strains also showed a specific pattern in the kinetics of sporulation, with the high sporulation efficiency strains showing fast sporulation kinetics and the low sporulation efficiency strains showing slow sporulation kinetics. For example, the strain YPS128 had maximum sporulation efficiency of 99.5% and reached saturation within 48 h. On the other hand DBVPG1788

Strains	Mean sporulation efficiency (%)[a]	Sporulation efficiency (from Cubillos *et al.* [11])[b]
273614N	NS	+++
322134S	NS	NA
378604X	NS	NA
BC187	61.3±0.9	++
DBVPG1106	22.4±1.2	+++
DBVPG1373	NA	+
DBVPG1788	40.4±0.9	NA
DBVPG1853	NS	+
DBVPG6040	67.4±1.2	+
DBVPG6044	6.0±0.9	++
DBVPG6765	NS	+++
K11	19.9±2.4	–
L-1374	76.6±1.4	+++
L-1528	70.2±1.0	+++
NCYC110	NS	+++
NCYC361	NA	–
S288c	NS	NA
SK1	92.4±1.8	+++
UWOPS03–461.4	86.8±1.2	+++
UWOPS05–217.3	88.5±1.2	+++
UWOPS05–227.2	85.2±1.8	+++
UWOPS83–787.3	98.6±0.4	+++
UWOPS87–2421	89.9±1.0	+++
W303	NA	NA
Y12	54.0±1.5	+
Y55	73.7±1.7	+++
Y9	22.1±2.8	–
YIIc17_E5	NS	++
YJM975	48.2±2.2	+++
YJM978	40.2±1.0	+++
YJM981	NS	+++
YPS128	99.0±0.6	+++
YPS606	97.9±0.5	+++
YS2	NS	–
YS4	NS	–
YS9	NA	NA

(a) Mean (with standard deviation) sporulation efficiency of each strain at saturation, *i.e.* when sporulation efficiency did not vary for three consecutive time points. (b) Sporulation efficiency as reported by Cubillos *et al.* [11]. The scale indicates: (+++) high, (++) medium, (+) low sporulation efficiency, (−) none, (NA) not applicable (either the strain was haploid or did not grow in YPA), (NS) did not sporulate and zero sporulation efficiency.
doi:10.1371/journal.pone.0069765.t001

Table 3.1: **Sporulation efficiency measurement of SGRP strains. Table reproduced from [1]**

had a maximum sporulation efficiency of 41.0% and took 8 days to reach this efficiency. Keeping the strain for a longer time in the sporulation media condition did not increase their sporulation efficiency (Figure 3.1, Table S3 in [1]). A comparison of sporulation efficiency estimated at 23°C in [71] with our estimates at 30°C showed notable differences (see Table 3.1). In the two studies, 16 strains had consistent efficiencies in both studies, indicating a significant effect of temperature dependence in sporulation efficiency. The results of our analysis are therefore relevant at 30°C. Future studies will be necessary to understand the effect of temperature on sporulation efficiency.

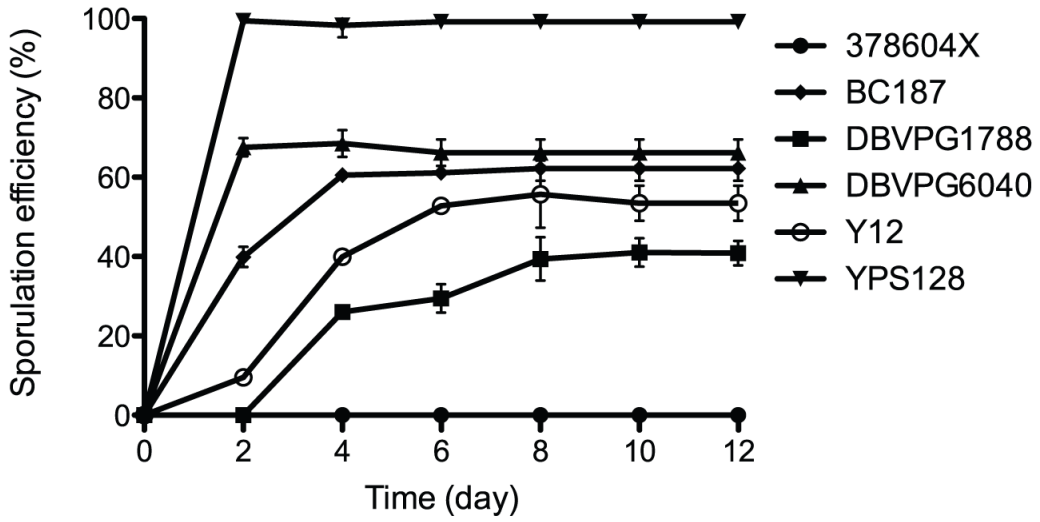


Figure 3.1: **Kinetics of sporulation efficiency measurements of representative *S. cerevisiae* SGRP strains.** Sporulation efficiency of each strain was measured till saturation, i.e. when sporulation efficiency did not vary for three consecutive time points. The data is plotted as mean and standard deviation of 3 independent biological replicates. Figure reproduced from [1].

3.3.2 SNP Variation in Sporulation Genes

Given our limited sample size, we searched for genotype/phenotype associations only among sporulation and sporulation associated genes. A survey of the literature identified a comprehensive list of 397 genes [63], [64], [65], [75] which included genes required

for metabolic adaptation, early, middle and late sporulation genes (meiosis, spore formation and general stress response genes), mitochondrial and autophagy genes and also genes which were induced upon sporulation but had unknown function (Table S1 in [1]). We looked for variation in these genes across all strains by identifying variant alleles from the SGRP alignment of all 32 strains. In total, we found 42,003 SNPs across these genes. The presence of variation allowed us to look for genetic determinants of variation of sporulation efficiency in these strains.

3.3.3 Association Mapping of Sporulation Efficiency

We used two methods to identify SNPs in genes that were associated with an increase or decrease in sporulation efficiency. The first method used the LOD score to identify SNPs in which the genotype was strongly associated with the sporulation efficiency phenotype. A high LOD score was evidence for the presence of a quantitative trait locus, where the two genotypes at a locus had significantly different phenotype averages. The second method binned the strains into three classes of high, intermediate and low sporulation efficiency and then applied a binomial test (see Methods) to identify SNPs in association with high and low sporulation. Both methods identified 31 SNPs in 24 different genes (Bonferroni corrected p-value < 0.03, permutation test) associated with sporulation efficiency variation (Table 3.2, Table S4 in [1]).

3.3.4 Population Structure Correction

Recently, the SGRP collection has been proposed for use in yeast GWAS studies [76], [77]. However, several issues have been raised about using this collection, including high type I errors (false positives) in determining causative loci [76], as underlying population structure can lead to spurious associations [77]. Using STRUCTURE [78] to determine population structure, and data for 201 phenotypes (not including sporulation efficiency), Diao and Chen [77], used extensive simulations and several GWAS methods on a genome wide set of tag SNPs to show that the mixed linear model EMMAX-KLA (a model with local ancestry and the kinship matrix as covariates) was the most effective at reducing type I errors and correcting for population structure in these strains. EMMAX-KLA

SNPs in Linked Cluster	LOD score	Bonferroni Corrected p-value (n = 1,709)	Sporulation Efficiency of Minor Allele	Sporulation Efficiency of Major Allele
HOS4:1038, HOS4:1206, MLS1:-21, SPR6:-434, TEP1:219	4.47	0.004	92.27	28.95
CCR4:2016, CDC10:-126, CIS1:174, DOA1:1152, EMI5:-10, GIP1:213, HOS4:1384, HPR1:1137, HPR1:1239, HPR1:1293, MAF1:761, MCK1:1112, MPC54:678, PEP12:201, PEP12:294, RAS2:924, RME1:63, SEF1:1254, SET3:1783, SHC1:213, SPO74:16, SPO75:1842, SPR6:519, SPR6:426, SSN8:-484, VID28:1410	3.50	0.026	92.67	31.38

doi:10.1371/journal.pone.0069765.t002

Table 3.2: **Clusters of SNPs with genome-wide significant LOD scores (Bonferroni corrected p-value < 0.03). Table reproduced from [1]**

was applied to our phenotype data to identify the tag SNPs that were significantly associated with the sporulation phenotype after correcting for population structure ($p < 0.05$). We verified that the SNPs that we identified as statistically significant using the LOD score and binomial test were in perfect linkage disequilibrium ($r^2 = 1$) with the tag SNPs identified as statistically significant using EMMAX-KLA, demonstrating that the association with sporulation efficiency remained after correcting for population structure (Table S4 in [1]).

3.3.5 Candidate SNPs and Genes Associated with Sporulation Efficiency

The SNPs that were identified as statistically significant by our two association analyses fell into two linkage blocks, one with a LOD score of 4.47 (Bonferroni corrected $p < 0.004$, permutation test) and another with a LOD score of 3.5 (Bonferroni corrected $p < 0.026$, permutation test). The first linkage block contained 5 SNPs whereas the

second linkage block contained 26 SNPs. These blocks of linked SNPs contained SNPs that were associated with sporulation efficiency in the SGRP strains. The SNPs in these clusters showed perfect linkage ($r^2 = 1$) i.e. they segregated in an identical manner across the yeast strains and they were not all contiguous in the genome. This suggests a residual population structure effect due to a small sample size. As a result, we could not computationally determine which of the SNPs in our clusters were causally associated with the phenotype and which were non-causal and in linkage with other causal variants. It will be necessary to analyse additional strains or perform additional experiments on the SGRP strains to answer this question.

Gene annotations were performed for these potentially functional SNPs to classify them as regulatory, synonymous or non-synonymous. We found that 20 genes had SNPs in their coding region and 5 genes (*CDC10*, *EMI5*, *MLS1*, *SPR6* and *SSN8*) had SNPs in their un-translated region. One gene, *SPR6*, had SNPs in both coding and regulatory regions (Table S4 in [1]). Interestingly, deletions of *EMI5*, *MLS1* and *SSN8* have been reported to decrease sporulation efficiency [79] and *CDC10* deletion abrogates sporulation [62].

Four of the 26 coding SNPs were non-synonymous. These may affect sporulation efficiency altering binding ability, or the extent of functionality, or the flux through the pathway which may alter protein levels. Two of the 4 non-synonymous substitutions were in *SET3* (A1783T), a repressor of sporulation specific genes [80] and *HOS4* (A1384G), a component of *Set3* complex and a suppressor of early and middle sporulation specific genes [81]. A possible reduction in protein function due to these mutations in the repressors, *Set3* and *Hos4*, could lead to an increase in sporulation efficiency in strains with these SNPs. The other two non-synonymous substitutions were *MCK1* (C1112A) and *SPO74* (C16A), deletions of which lead to decrease [62] and absence [82] of sporulation respectively. Among these four non-synonymous substitutions, the only one non-conservative substitution in *Mck1* (T371K) lies within its putative kinase domain, a positive regulator of meiosis and spore formation [83].

Two of the genes, *HOS4* and *SPR6* (a gene of unknown function expressed during sporulation and interacting with sporulation genes [84]), were present in both significant

clusters (Table 3.2), suggesting their role as potential candidates for variation in sporulation efficiency across SGRP strains. However, an experimental validation is required to confirm their actual role, either by performing reciprocal hemizygosity analysis [3] or by constructing allele replacement strains.

3.4 Discussion

A limited understanding of traits in natural populations is one of the biggest challenges in genetic association studies. This lack of information about phenotypes in the wild has limited our knowledge about the role played by evolution, life history, environment and selection pressure in driving these processes. In this study, we have tried to understand the genetic basis of variation in sporulation efficiency in natural isolates of yeast using the SGRP collection. Since sporulation is triggered as a response to nutrient deprivation, we expected that the genetic factors contributing to variation in sporulation efficiency might be different for lab strains compared to natural isolates. To identify such differences, we measured sporulation efficiency of *S. cerevisiae* strains in the SGRP collection and found a large variation in sporulation efficiencies ranging from 0% to 100%, which we then used to identify the genetic basis of variation in sporulation efficiency of these wild yeast strains.

Our study suggests that both regulatory and coding variants may be responsible for variation in sporulation efficiency. Four out of twenty six (15%) of the SNPs identified to be associated with sporulation efficiency were non-synonymous, and occurred in the genes *HOS4*, *MCK1*, *SET3* and *SPO74*. We list these genes as candidate drivers of variation in sporulation efficiency in the SGRP collection. Previous studies have identified roles for sporulation genes (*IME1*, *RME1*) and sporulation-associated genes (*FKH2*, *PMS1*, *RAS2*, *RSF1*, *SWS2*), as well as non-sporulation pathway genes (*MKT1*, *TAO3*) in maintaining this variation [66], [67], [68]. Our results showed that in the SGRP collection, a different set of genetic factors contribute to variation in sporulation efficiency.

S. cerevisiae is a powerful system for quantitative trait genetics and has advanced

our understanding of the genotype-phenotype relationship of these traits. With decreasing cost of sequencing and high-throughput phenotyping, yeast has become a model for GWAS studies [76], [77]. Our results provide another example of how GWAS studies in the SGRP collection can identify known and new candidates for sporulation efficiency variation in natural strains of yeast. Thus, it provides insight into how the selection pressure due to changes in the environmental conditions of natural isolates (such as nutrient availability) can drive evolution of a phenotype (such as variation in sporulation efficiency).

Chapter 4

Detecting signs of recent, positive selection in the Maasai people of East Africa

4.1 Introduction

The Maasai are a pastoralist, Nilotic people living primarily in southern Kenya and northern Tanzania. An economy traditionally based on herds of cattle, sheep, and goats led to a diet rich in lactose, fat, and cholesterol consisting largely of milk, meat, and blood. Although their cholesterol intake is high ($600 - 2000$ mg/day), and 66% of their calories come from fat, their total serum cholesterol levels average 135 mg/100 ml [85], [86], [87], [88]. In comparison, a study consisting of cohorts from seven countries (Croatia, Finland, Greece, Italy, Japan, Netherlands, USA) found that the average dietary cholesterol intakes are $141-612$ mg/day and serum cholesterol levels range from $160 - 266$ mg/100 ml [89]. Although African children generally have lower cholesterol levels ($115-137$ mg/100 ml for 7-8 year olds) than other populations [90], the fact that adult Maasai have very low cholesterol levels, in spite of a high cholesterol diet, is quite remarkable. The Maasai also have low rates of cholelithiasis (especially cholesterol gallstones), low blood pressure, and low incidence of atherosclerotic coronary artery disease [85], [86], [87], [91]. Various hypotheses to understand this puzzle have been proposed, such as: physical fitness and freedom from emotional stress [91], [92], a hypo-cholesterolaemic factor in milk [93] and saponins derived from herbs [94]. However, the hypo-cholesterolaemic factor was never found, and the model of [91], [92] could not explain the low frequencies of heart disease in older Maasai men who lead sedentary lives after age 24, when their warrior (Murran/Moran) period ends [95], [96].

Additional clues emerged from a controlled experiment [86] on 23 healthy Maasai adults (11 experimental, 12 control) between the ages of 20 and 24 years. All study

subjects were fed a basic high calorie, cholesterol-free diet for 8 weeks, including trace amounts (1 micro-curie) of radioactively labeled Cholesterol-4-^{14}C. The eleven subjects in the treatment group were fed 2 g of crystalline cholesterol per day in addition to the basic diet. Blood and fecal samples were collected at the start of the study, weekly for 8 weeks and at the end of 9, 16 and 24 weeks. Using the radioactive tracer to quantitate/normalize the measurements, the data were analyzed to characterize metabolic patterns, namely, the amounts of dietary cholesterol absorbed, synthesized and excreted. The study found that, in spite of the additional 2 g/day ingestion of cholesterol in the experimental group, there were no significant differences in serum cholesterol, phospholipids, triglyceride levels and lipoprotein patterns between the experimental and control groups. Both groups had identical turnover rates for cholesterol, with no evidence for cholesterol storage in the experimental group. In a similar study in American subjects, Mattson *et al.* [97] found that total serum cholesterol increased linearly with dietary cholesterol with 11.8 mg/100 ml increase for every 100 mg/1000 kcal increase in dietary cholesterol over the range $100-317$ mg/1000 kcal. Were this relation to hold in the Maasai, an increase of 66 mg/100 ml total cholesterol levels would be expected in the above experiment, contrary to the observed cholesterol homeostasis. The observed cholesterol homeostasis could not be attributed to a hypo-cholesterolaemic factor, or to saponins, which were absent from the Maasai study diet. The authors concluded that the Maasai have some basically different genetic traits that result in their having superior biologic mechanisms for protection from hypercholesterolemia [87].

It is widely accepted that there is a strong genetic component in the risk of hypercholesterolemia, atherosclerosis and heart disease [98], [99], [100], [101]. Typically, genome-wide association studies (GWAS) focus on markers for increased risk of disease [102], [103], [104], [105], [106] and to a lesser extent on protective polymorphisms. Such protective polymorphisms are known to arise as adaptations and can be identified in selection studies. For example, many studies have identified polymorphisms conferring lactase persistence in Northern Europeans, which arose with the advent of cattle breeding [28]. Just as in Europe, pastoralism arose in East Africa around 4,000-10,000

years ago [107] leading to selection for lactase persistence [18]. In the Maasai, pastoralism led to a lactose rich, high fat, high cholesterol diet of milk, meat and blood [88]. It is quite reasonable that, in a time span similar to that which conferred lactase persistence in Europeans, selection pressure in the Maasai from such a diet might result in genetic adaptations against diseases such as hypercholesterolemia and atherosclerosis.

Motivated by this possibility, we performed a genome wide scan for selection in 156 founder individuals from the Maasai of Kinyawa, Kenya (MKK) using the HapMap 3 SNP (single nucleotide polymorphism) data [19] to identify genomic regions under recent selection. We also used SNP data from 110 HapMap 3 founder individuals from the Luhya population from Webuye, Kenya (LWK) as a reference group. Three complementary metrics to detect selection were applied: the Fixation Index (F_{ST}) [108], the Cross Population Extended Haplotype Homozygosity (XP-EHH) [17], and the Integrated Haplotype Score (iHS) [14], [16]. Note that the phased data used for iHS and XP-EHH was from HapMap3 Release 2, which has fewer individuals (143 and 90 for MKK and LWK respectively) whereas the data for F_{ST} was from HapMap Release 3, which had more individuals (156 and 110 respectively). Our analysis consistently identified strong, recent selection in genes involved in lipid metabolism and lactase persistence in the Maasai (MKK) samples. Several of the regions under selection in MKK contained specific polymorphisms known to protect against hyperlipidemia in other populations. Sanger sequencing of DNA from six MKK samples showed that the GC-14010 polymorphism in the Minichromosome Maintenance Complex Component (*MCM6*) gene, known to confer adult lactase persistence in East Africans [18], is segregating in the Maasai at a frequency of 58%. These results suggest that the regions identified contain polymorphisms that confer lactase persistence and protection from hypercholesterolemia in the Maasai. The wider consequence of our study is that consistent dietary pressure can induce strong selection in complex pathways in a short time ($150 - 400$ generations).

4.2 Numerical methods to detect recent positive selection

4.2.1 The HapMap database

We analyzed single nucleotide polymorphism (SNP) data and phased haplotype data collected and made publicly available by the International HapMap Project [19]. HapMap 3 release 3. SNP **genotype** data was downloaded from `http://snp.cshl.org/` for founder individuals from the Maasai in Kinyawa, Kenya (MKK) (n = 156), the Luhya in Webuye, Kenya (LWK) (n = 110), African-Americans in Southwest USA (ASW) (n = 53), the Yoruba in Ibadan, Nigeria (YRI) (n = 147), and Utah residents of Northern and Western European ancestry (CEU) (n = 112). Using PLINK (`http://pngu.mgh.harvard.edu/~purcell/plink/`) [109], we filtered the data to retain only the SNPs that were common to all populations. HapMap 3 release 2 autosomal **haplotype** data for the MKK (n = 143) and LWK (n = 90) was also downloaded from `http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/2009-02_phaseIII/HapMap3_r2/`. The data was phased using IMPUTE++ [110]. SNPs were pre-filtered for Hardy Weinberg equilibrium and for low frequency of Mendel errors (`http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/2010-05_phaseIII/00README.txt`). Genetic maps were downloaded from `http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/2008-03_rel22_B36/rates/` to obtain the genetic map position of the SNPs in centiMorgans.

4.2.2 Addressing Population Structure using STRUCTURE

Using PLINK [109], the genotype data for MKK, LWK, YRI, ASW and CEU was further filtered to exclude SNPs with minor allele frequency $< 1\%$ or SNPs where more than 1% of the genotype data was missing. Restricting the samples to founders resulted in 1,325,342 common SNPs for 578 individuals. To analyze the population structure of these populations, we further restricted the genotype data to a random subset of 1% of these SNPs (12,999 SNPs) and ran the no admixture model in STRUCTURE [111] version 2.3. We found that k = 6 ancestral populations fit the data best (Table 4.1).

Thinning the dataset was necessary to reduce the likelihood of SNPs in linkage disequilibrium, which is a requirement for the no admixture model. Adding more SNPs

did not result in significant gains in statistical power. We used 10,000 steps in the "burnin" period, and 20,000 steps as the number of MCMC iterations. We ran the simulation over several values for k = number of inferred (ancestral) populations, and obtained the log likelihoods for the fits as shown in Table 4.1.

4.2.3 F_{ST} computation

Using PLINK, we retained 1,175,055 autosomal SNPs in Hardy Weinberg equilibrium ($p > 0.05$) and with minor allele frequency $> 5\%$ in either population (LWK and MKK). We then computed F_{ST} using the method of [108]. Two tests were used to assess statistical significance, a Bonferroni corrected permutation test (p-value p_B), and an empirical p-value that compared the F_{ST} of a SNP to the F_{ST} distribution of intergenic SNPs. Gene positions were from the human genome build 37 (GRCh37/hg19) available at `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/knownGene.txt.gz`. To avoid linkage with genes and promoter regions, we define intergenic regions as those that are at least 50 kb away from the start or stop site of a gene. For the remaining genic or near-gene SNPs, we calculated an empirical p-value (p_E) given by the fraction of intergenic SNPs with greater F_{ST}. This procedure identified 1,232 SNPs with $p_B < 8.6 \times 10^{-6}$ and $p_E < 0.001$ that are the top candidates for selection using F_{ST}. These SNPs were then clustered into regions of high linkage (Table 4.3) using the method described below (details of the F_{ST} calculation are below).

4.2.4 iHS computation

Autosomal haplotype data for 991,737 SNPs in MKK with minor allele frequency $> 10\%$ were used to calculate raw iHS scores as in Voight *et al.* [16]. These raw iHS scores were binned on the basis of derived allele-frequency, and the scores in each bin were standard normalized to zero mean and unit variance. Genomic sliding windows of 50 SNPs were ranked by the percentage of SNPs with $|iHS| > 2$. The SNPs with $|iHS| > 2$ that occurred in the top 0.02% of non-overlapping windows were selected as top candidates for selection by iHS. These were then clustered into regions of high linkage (Table 4.5) using the method described below (details of the iHS calculation are below).

4.2.5 XP-EHH computation

Autosomal haplotype data for 1,373,755 SNPs in MKK and LWK was mapped to genomic locations in the human genome, build 37 (GRCh37). XP-EHH scores were calculated using the code at `http://hgdp.uchicago.edu/Software/xpehh.tar`. The XP-EHH scores were fit to a normal distribution, which identified the threshold for genome-wide significance to be XP-EHH ≥ 4.796 (Bonferroni corrected $p < 0.05$, two-tailed test). The SNPs that exceeded this threshold were chosen as top candidates for selection by XP-EHH (Table S5 in [2]). These SNPs were clustered into regions of high linkage (Table 4.5, Table S6 in [2]) using the method described below (further details of the XP-EHH calculation are included below).

4.2.6 LD clustering of SNPs

The SNPs identified as candidates for selection by each of the above methods were clustered using genotype r^2 as an estimator of linkage disequilibrium. We used the criterion that for a SNP to be included in a cluster, it must have genotype $r^2 \geq 0.25$ with at least one other SNP in the cluster (the justification for this choice of cutoff is given below).

More concretely, for the SNPs identified by the methods above, we used PLINK to extract a file of raw genotype data from the HapMap genotype data file for MKK. These files contained a matrix of genotype values, whose columns were labeled by SNPs and rows labeled by individuals. We imported this genotype matrix into the statistical package R, to calculate a SNP $\times$ SNP Pearson correlation matrix. This correlation matrix was then used to construct a SNP $\times$ SNP adjacency matrix whose entries are 1 if $r^2 \geq 0.25$ and 0 if $r^2 < 0.25$. The problem of finding linked clusters of SNPs then translates to identifying the connected components of the graph described by this adjacency matrix. This computation was performed in Python using the NetworkX package (`http://networkx.lanl.gov/`) [112].

4.2.7 Sequencing loci in *LCT/MCM6* and *RAB3GAP1*

Forward and reverse primers for Sanger sequencing were chosen using *Primer3* (`http://frodo.wi.mit.edu/primer3/`), and checked for absence of homologies to other parts of the human genome using *BLAT* [113]. The details of the primers, the loci sequenced and the samples used are in Appendix S5 of [2].

4.2.8 Further details on the fixation index F_{ST}

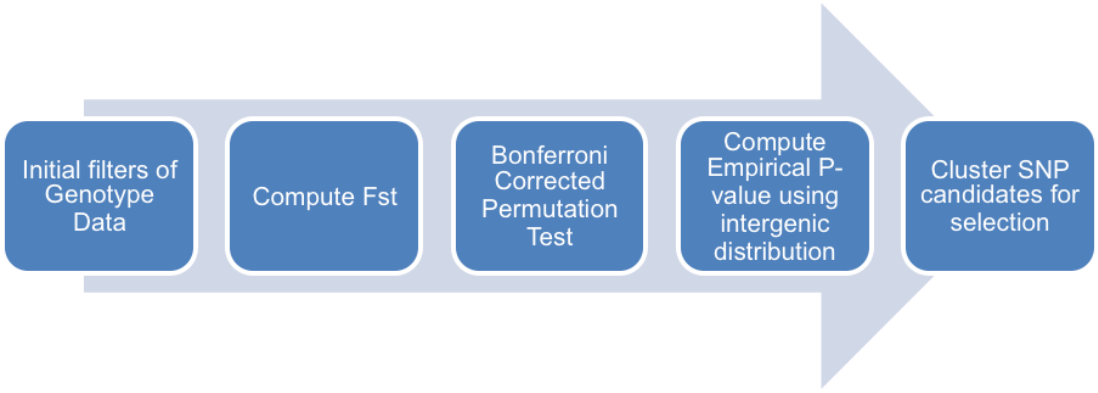


Figure 4.1: **Workflow illustrating the stages in the F_{ST} analysis of the genotype data.**

The overall workflow used in the F_{ST} analysis is shown in Figure 4.1. Genotype data from HapMap 3 release 3 [19] was downloaded and pruned to 1,175,055 autosomal SNPs with minor allele frequency > 0.05 in MKK (n = 143 founders) and LWK (n = 100 founders). To reduce the chance of incorporating SNPs with genotyping errors, we imposed a Hardy-Weinberg equilibrium p-value cutoff < 0.05 in either population (as calculated by PLINK [109]), and excluded SNPs with genotype missing in $> 5\%$ of samples. F_{ST} was computed using the method of Reynolds, Weir and Cockerham [108].

F_{ST} computation details

The fixation index F_{ST} is the fraction of total variance in the genotype frequencies of a population that is due to the variance between the populations. Concretely, if we define $\sigma^2_{population}$ to be the component of variance between populations and $\sigma^2_{individuals}$

to be the component of variance between individuals within a population, and σ^2_{gamete} to be the component of variance between gametes within an individuals

$$F_{ST} = \frac{\sigma^2_{population}}{\sigma^2_{population} + \sigma^2_{individual} + \sigma^2_{gamete}}$$

We can now simplify each term:

$$\begin{aligned}
\sigma^2_{population} &= \frac{1}{n_1 + n_2}\left(\sum_{pop1}(p_1 - p)^2 + \sum_{pop2}(p_2 - p)^2\right) \\
&= \frac{n_1 p_1^2 + n_2 p_2^2}{n_1 + n_2} - p^2 \\
&= \langle p_j^2 \rangle - p^2
\end{aligned}$$

where we have defined:

$$\langle p_j^2 \rangle \equiv \frac{n_1 p_1^2 + n_2 p_2^2}{n_1 + n_2}$$

and p is the average frequency:

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Similarly,

$$\begin{aligned}
\sigma^2_{individual} &= \frac{1}{n_1 + n_2}\left(\sum_{individuals \in pop1}(p_i - p_1)^2 + \sum_{individuals \in pop2}(p_i - p_2)^2\right) \\
&= \frac{n_1}{n_1 + n_2}\frac{p_1(1 - p_1)}{2} + \frac{n_2}{n_1 + n_2}\frac{p_2(1 - p_2)}{2} \\
&= \frac{1}{2}\left(p - \langle p_j^2 \rangle\right)
\end{aligned}$$

where we have used that $\sigma_j^2 = \frac{p_j(1-p_j)}{2}$ for a single population. This can be derived assuming Hardy-Weinberg equilibrium of genotype frequencies as follows:

$$\begin{aligned}
\sigma_j^2 &= \frac{1}{n}\sum_i (p_i - p_j)^2 = \langle p_i^2 \rangle - p_j^2 \\
&= (p_j^2 \times 1 + 2p_j q_j \times \left(\frac{1}{2}\right)^2 + q_j^2 \times 0) - p_j^2 \\
&= \frac{p_j(1 - p_j)}{2}
\end{aligned}$$

Finally,

$$\begin{aligned}\sigma^2_{gamete} &= \frac{1}{n_1+n_2}\left(\sum_{genotypes\in pop1}\sigma_g^2 + \sum_{genotypes\in pop1}\sigma_g^2\right)\\ &= \frac{1}{n_1+n_2}\left(n_1\frac{p_1(1-p_1)}{2} + n_2\frac{p_2(1-p_2)}{2}\right)\\ &= \frac{1}{2}\left(p - \langle p_j^2\rangle\right)\end{aligned}$$

$$\begin{aligned}\sum_{genotypes\in popj}\sigma_g^2 &= \frac{1}{2n_j}2n_j\left(p_j^2\sigma_{00}^2 + 2p_jq_j\sigma_{het}^2 + q_j^2\sigma_{11}^2\right)\\ &= p_j^2\times 0 + 2p_jq_j\times\frac{1}{4} + q_j^2\times 0\\ &= \frac{p_j(1-p_j)}{2}\end{aligned}$$

In summary,

$$\langle p_j^2\rangle \equiv \frac{n_1p_1^2 + n_2p_2^2}{n_1+n_2}$$

$$F_{ST} = \frac{\sigma^2_{population}}{\sigma^2_{population}+\sigma^2_{individual}+\sigma^2_{gamete}} = \frac{\langle p_j^2\rangle - p^2}{p(1-p)}$$

In 1984, Weir and Cockerham derived an unbiased estimator for F_{ST} that accounts for the bias associated with sampling a population [108]. Their result is summarized below, and we used this estimator of F_{ST} in the results that follow. If n_1 and n_2 are the number of MKK (Maasai) and LWK (Luhya) individuals measured at a locus l, and p_1 and p_2 are the derived allele frequencies at this locus in the two populations, define a_l and b_l as:

$$a_l = (p_1-p_2)^2 - \frac{(n_1+n_2)(2n_1p_1(1-p_1)+2n_2p_2(1-p_2)}{4n_1n_2(n_1+n_2-1)}$$

$$b_l = \frac{2n_1p_1(1-p_1)+2n_2p_2(1-p_2)}{n_1+n_2-1}$$

Then,

$$F_{ST} = \frac{a_l}{a_l+b_l}$$

Bonferroni corrected permutation p-value p_B for F_{ST}

At every SNP we compute a p-value for F_{ST} using a permutation test. The null hypothesis is that all rearrangements of the alleles among the two populations are equally

probable. The Bonferroni corrected permutation p-value p_B is then n times the probability that the value of F_{ST} in the null-distribution exceeds the observed value of F_{ST}, where n = number of hypotheses (SNPs) tested.

For each SNP, there are $2(n_1 + n_2)$ alleles in the combined population. We define a partition of the data by assigning $2n_1$ alleles to MKK and the rest to LWK. The permutation p-value p is the sum, over all such partitions, of the probability that the F_{ST} value obtained in a partition is greater than or equal to the F_{ST} value x obtained in the actual data. Thus,

$$p(x) = \sum_{part} Prob(part)\ \theta(F_{ST}(part) - x) \tag{4.1}$$

where the θ is a step function that ensures that only partitions with $F_{ST} > x$ contribute to the sum.

Let $n_1(--)$, $n_2(--)$ and $n_1(-+)$, $n_2(-+)$ be the number of mutant homozygous and heterozygous individuals in the MKK and LWK cohorts. Then the total number of mutant alleles N in the combined population is given by:

$$N = 2(n_1(--) + n_2(--)) + n_1(-+) + n_2(-+)$$

Since we know the genotypes of the samples, N is known from the data. In the $2n_1$ alleles assigned to MKK, let there be n mutant alleles. Then,

$$p_1(part) = \frac{n}{2n_1}$$
$$p_2(part) = \frac{N - n}{2n_2}$$

Using these values of $p_1(part)$ and $p_2(part)$ (which are specific to the partition), one can compute $F_{ST}(part)$ using the formulae in the previous section. The values that n take are limited to the open interval: $[n_{min}, n_{max}]$, with $n_{min} = max(0, N - 2n_2)$ and $n_{max} = min(2n_1, N)$. Hence, we can rewrite equation 4.1 as:

$$p(x) = \sum_{n=[n_{min},n_{max}]} \frac{\binom{N}{n}\binom{L-N}{L_1-n}}{\binom{L}{L_1}}\ \theta(F_{ST}(n) - x) \tag{4.2}$$

where, $L = 2n_1 + 2n_2$, $L_1 = 2n_1$ and $L_2 = 2n_2$. The factor in the numerator is the number of ways of assigning n mutant alleles and $N - n$ non-mutant alleles to the $2n_1$

loci in the MKK samples. The normalization factor in the denominator accounts for all possible ways of choosing $2n_1$ alleles from $2(n_1 + n_2)$ alleles. For each of the 1,175,055 SNPs tested, we computed the sum in Equation 4.2 to obtain the permutation p-value $p(x)$ for each measured value of $F_{ST} = x$. From this, we obtained a Bonferroni corrected p-value: $p_B(x) = 1175055 \times p(x)$.

Empirical p-value p_E using the F_{ST} distribution of intergenic SNPs

As a further filter, the F_{ST} values of SNPs in non-intergenic regions were compared to the F_{ST} distribution of intergenic SNPs. SNP annotations were obtained for version hg19 of the human genome from `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/knownGene.txt.gz`. Intergenic SNPs were defined as those located more than 50 Kb away from the start and stop sites of all known genes. This 50Kb buffer was used to exclude promoter regions (possibly conserved due to purifying selection) and to minimize the effect of LD with genic SNPs.

The F_{ST} distribution of 351,254 intergenic SNPs was used to compute an empirical p-value p_E for the remaining non-intergenic SNPs as the fraction of intergenic SNPs with higher F_{ST}. 1,232 SNPs within genes or within 50 kb of genes with $p_B < 8.610^{-6}$ and $p_E < 0.001$ were retained for clustering. These are shown in Supplementary Table 3a in [2] and summarized in Tables 4.3 and 4.4.

Clustering significant SNPs using Linkage Disequilibrium

The SNPs thus identified as selection candidates do not all represent independent selection events. During a selective sweep, many neighboring linked SNPs can hitchhike along with the selected allele, and thus show correspondingly high scores for selection. In order to identify such linked regions in which high F_{ST} SNPs occur, the 1,232 SNPs identified above were clustered into contiguous genomic regions using genotypic r^2 in MKK as a measure of linkage disequilibrium. For each population, a SNP was assigned to a cluster if it had a genotype $r^2 \geq 0.25$ with at least one other SNP in the cluster.

The value $r^2 \geq 0.25$ has been shown to correspond to a genetic distance of $0.01 - 0.02$ cM across a varied set of population growth models [114]. Assuming a genomic

average recombination rate of 1 cM/Mb, this is equivalent to a physical distance on the order of 10 kb in each direction. The probability for two or more SNPs from a randomly chosen set of 1,232 SNPs to occur within 10 kb is close to one percent, hence we conclude that $r^2 \geq 0.25$ is a reasonably stringent cutoff for linkage.

Using XP-EHH to identify the population in which a sweep has occurred

Assuming that only one of the two populations has undergone a selective sweep at a given locus, we identified the population in which the sweep is more likely to have occurred by comparing the local haplotype diversity across populations. Concretely, for each cluster identified by F_{ST}, we label it as selection candidate in MKK if the maximum normalized XP-EHH score of a SNP in the cluster is > 3. A positive value for XP-EHH indicates that the MKK carry the longer-range haplotypes.

This procedure identified 26 clusters (containing 318 SNPs) as selection candidates in MKK (Supplementary Table 1a in [2]). 9 of these clusters include SNPs that exceed the genome-wide significance threshold for XP-EHH (XP-EHH > 4.8, Bonferroni corrected $p < 0.05$, two-tailed). In Table 4.7, we list the intersection of clusters that are identified as genome-wide significant by at least two out of the three methods used (F_{ST}, iHS, XP-EHH). Supplementary Table 2 in [2] shows the concordance of our results with those of the HapMap consortium [19].

The remaining SNPs were either singletons (did not occur in clusters) or were in clusters that could not be confidently assigned to the MKK. In Table 4.4, we list the non-synonymous SNPs with most significant genome-wide F_{ST}. These are our top candidates for possible functional polymorphisms.

4.2.9 Further details on the integrated haplotype score: iHS

Selection events not only sweep functional loci to high frequency, they also reduce haplotype diversity in the region around the selected locus because of hitchhiking. The Extended Haplotypic Homozygosity (EHH) statistic [14] exploits this principle and provides a criterion for detecting SNPs under selection within a population, without reference to another population. EHH measures the diversity of extended haplotypes

containing a chosen core haplotype, as a function of the distance from the core. Recently, much work has focused on using EHH to define various measures (such as iHS [16] or iES [115]) which can identify recent selection. Here, we use iHS to detect recent selective sweeps in the Maasai.

EHH is the probability, as a function of distance from the core SNP, that two randomly selected haplotypes that share the core SNP will be identical. It is defined as follows:

$$EHH(x) = \frac{\sum_{i=1}^{h(x)} \binom{n_i(x)}{2}}{\binom{n}{2}}$$

where $n_i(x)$ is the number of samples of a particular haplotype i (up to a distance x), $h(x)$ is the total number of distinct unique haplotypes in this distance, and n is the total number of samples. Thus, $n = \sum_{i=1}^{h(x)} n_i(x)$.

iHS is defined as the log of the ratio of the integrated EHH score for haplotypes containing the ancestral allele to the integrated EHH score for haplotypes containing the derived allele [16].

$$\text{unstandardized } iHS = \log\left(\frac{\int EHH_{ancestral}(x)dx}{\int EHH_{derived}(x)dx}\right)$$

Since iHS is a local measure and does not use a reference population, haplotypes for both alleles share the same genomic environment (local mutation rate, recombination rate) and are subject to identical population demography. Hence iHS does not suffer from ambiguities associated with genomic and population structure specific variations such as differences in recombination rates, demographic history, population bottlenecks, etc. As a result, significantly large values of iHS are more likely to be due to selection.

High values of iHS occur when haplotype diversity is reduced because of selection induced hitchhiking, which leads to more extended haplotypes for the selected allele and a consequent slower fall-off of EHH on either side of the selected locus. A high iHS scoring SNP typically has one allele associated with longer haplotypes and lower neighborhood diversity compared to the other allele (see Figure 1a in [16]). In a selective sweep, hitchhiking causes both functional SNPs as well as any SNPs in their neighborhood to have amplified iHS scores. In fact, simulations show that a high density of high-scoring SNPs is a better indicator of a selective sweep than high iHS score of a

single SNP [16].

Computational details and p-value significance

Autosomal haplotype data phased with IMPUTE++ [110] was downloaded on 10.24.2010 from: `http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/2009-02_phaseIII/HapMap3_r2/`. SNPs were pre-filtered for Hardy Weinberg equilibrium and had low frequency of Mendel errors (see `http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/2010-05_phaseIII/00README.txt`). We further pruned the data by removing SNPs with MAF < 0.1. After applying these filters, we analyzed 991,737 SNPs. The ancestral allele information was downloaded from the NCBI ftp server (`ftp://ftp.ncbi.nih.gov/snp/database/shared_data/`) on 7.5.2011. Genetic maps were downloaded from HapMap website (`http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/2008-03_rel22_B36/rates/` as accessed on 9/22/2010). These maps did not contain the genetic distances for all the SNPs used, and hence we interpolated the genetic distances. For the final results genetic distances were converted to GRCh37.

To identify potential genomic regions under selection in the MKK data, we followed the protocol described in [16]. We calculated the integrals of EHH for each SNP using the genetic distances over a domain of integration such that at least one allele of the SNP has an EHH > 0.05. We then binned these raw scores according to derived allele frequency.

Since the total number of haplotypes is 286, the frequencies of alleles are integral multiples of 1/286 (=0.0035). Hence, we used each frequency to comprise a single bin, and obtained 229 bins spanning frequencies 0.1 to 0.9 (SNPs with minor allele frequency < 0.1 were filtered out). We computed the mean and standard deviation of iHS values for each frequency bin and standard normalized the raw iHS scores to have zero mean and unit variance. We then considered sliding windows of 50 consecutive SNPs and noted the fraction of SNPs with normalized-iHS (hereafter iHS) ≥ 2. This fraction of high-scoring SNPs in a window is the statistic used for detecting selective SNPs [16]. Using sliding windows is advantageous over using fixed, gene-centric windows as done

in [16], since it does not rely on the choice of start position. In [16], the top 1% of non-overlapping windows were candidates for selective sweep. As we used sliding windows of size 50 SNPs, we corrected for this by looking at 1/50th of the top 1% of such windows with the constraint that they do not overlap. Overall we analyzed 990,659 windows and chose top 196 non-overlapping windows. Some of these chosen windows were adjacent to each other, and are likely to represent sweeps extending more than 50 SNPs. To merge such windows, we listed all SNPs with $|iHS| \geq 2$ (high-scoring SNPs) in these top windows (listed in Supplementary Table 2b), and clustered them using a genotype r^2 cutoff of 0.5. This clustering has the advantage that it does not impose an ad-hoc window size, but is based on local patterns of LD. We kept clusters with size greater than the least number of high-scoring SNPs in the top windows (20). These clusters are candidate regions for selective sweeps in the Maasai, and are given in Supplementary Table 2b in [2]. We then used the UCSC genome browser to identify genes and GWAS SNPs that lie in these regions.

4.2.10 Further details on cross-population extended haplotype homozygosity: XP-EHH

A selective sweep results in the rapid rise in the frequency of beneficial alleles accompanied by a reduction in haplotype diversity in the neighborhood of functional mutations due to a hitch-hiking effect (see [25] for a discussion). The key idea behind methods to identify selective sweeps is to use metrics that probe such reduced haplotype diversity. The statistic EHH (Extended Haplotype Homozygosity) [14] is one such metric. It measures the reduction in haplotype diversity by computing the probability that two extended haplotypes around a given locus are the same, given that they have the same allele at the locus. While selection decreases haplotype diversity, recombination increases it. Since recombination rates vary widely across the genome within and between populations, the EHH statistic can be interpreted as a measure of selection only after suitable normalization. The iHS statistic [16] compares the integrated EHH profiles between two alleles at a given SNP in the same population (iHS is discussed in more detail in Supplementary Appendix 3). On the other hand, the XP-EHH (Cross

Population Extended Haplotype Homozygosity) statistic (defined below) compares the integrated EHH profiles between two populations at the same SNP [17].

The iHS statistic is expected to be more reliable when one cannot find a good reference population (i.e. when the demographic history of potential reference populations is unknown or very different from the target population), but has low power when the selected allele is close to fixation. On the other hand, XP-EHH is expected to be more reliable if a reference population with a similar demographic history is available, and if the allele under selection is close to fixation in one of the populations. For XP-EHH, we used the Luhya (LWK) samples as the reference population to compare to the Maasai (MKK) samples. The motivation to choose the LWK samples was that they were closest to MKK with respect to overall population structure (discussed in Results).

Computing XP-EHH requires the computation of EHH in each population. For a bi-allelic SNP with alleles a and A, the EHH is defined as follows:

$$EHH(x) = \frac{\sum_{i=1}^{h_x} \binom{n_i}{2}}{\binom{n_a}{2} + \binom{n_A}{2}}$$

Here n_a and n_A are the number of haplotypes with alleles a and A respectively, n_i is the count of the i^{th} haplotype in a population and h_x represents the number of distinct haplotypes in a genomic region up to a distance x from the locus. The unstandardized $XP-EHH$ statistic is then defined as:

$$XP-EHH(x) \text{ (before standardization) } = \log \frac{\int_D EHH_{pop1}(x)dx}{\int_D EHH_{pop2}(x)dx}$$

In Eq. (2), pop1 and pop2 represent the two populations (pop1 = MKK and pop2 = LWK in our case). The integration domain D (cutoff over the x integration) was chosen so that the EHH values for both populations have fallen to sufficiently small values. We chose the cutoff as the distance at which EHH for both the populations combined was $0.03 - 0.04$. The unstandardized XP-EHH scores from Eq. (2) were standard normalized and p-value cutoffs were obtained (after correcting for multiple hypothesis testing) from a Gaussian fit to the resulting data. Since XP-EHH (unlike iHS) is not sensitive to allele frequencies, there is no need to stratify the data into frequency bins before determining significance.

Autosomal haplotype data phased with IMPUTE++ was downloaded on 10.24.2010 from: `http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/2009-02_phaseIII/HapMap3_r2/`. SNPs were pre-filtered for Hardy Weinberg equilibrium and had low frequency of Mendel errors (see `http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/2010-05_phaseIII/00README.txt`). Common SNPs between the MKK and the LWK were retained, and un- standardized XP-EHH scores were computed using the program by Joe Pickrell at `http://hgdp.uchicago.edu/Software/`. SNPs with unique positions in the dbSNP build 131 (GRCh37) were retained, leaving 1,373,756 SNPs. As expected, the distribution of XP- EHH was close to Gaussian (Fig. 1). We used IGOR Pro (`http://www.wavemetrics.com/products/igorpro/igorpro.htm`) to fit the data to a Gaussian, using the Levenberg- Marquardt method for curve-fitting. Using this fit, we obtained the cutoff of XP-EHH > 4.7958 at 95% genome-wide significance levels (two-tailed Bonferroni corrected p = 0.05, n = 1,373,756). SNPs passing this threshold are candidates for selection in MKK and are listed in Supplementary Table 3a in [2]. High scoring XP-EHH SNPs seemed to naturally form clusters when mapped to chromosomal regions. To identify regions associated with selective sweeps, high scoring SNPs were clustered using the same scheme as was used for F_{ST} and iHS. Clusters of SNPs were defined as sets of SNPs that had genotype $r^2 \geq 0.25$ for at least two SNPs in the cluster. These clusters of SNPs are listed in Supplementary Table 3b in [2].

4.3 Results

4.3.1 Population Structure

Two of the methods used to detect selection (F_{ST} and XP-EHH) require a genetically similar reference population. A comparison of F_{ST} among HapMap populations shows that the MKK and African-Americans from South-west USA (ASW) have the lowest average F_{ST} (0.0145), followed by MKK and the Luhya in Webuye, Kenya (LWK) (0.017), while F_{ST} between MKK and Yoruba from Nigeria (YRI) is significantly higher (0.027) (Table S6 in [19]). However, a plot of the first two principal components from a PCA analysis of the African populations and Utah residents with Northern and Western

European ancestry from the CEPH collection (CEU) (Figure S2c in [19]) shows that the MKK are genetically closer to LWK.

To understand the degree of admixture in the populations ASW, CEU, LWK, MKK and YRI, we used STRUCTURE [111] on a randomly sampled subset of 12,999 SNPs from the HapMap 3 dataset. Without using any population identification information, STRUCTURE found that the data fits best to 6 ancestral populations (Figure 4.2). In agreement with [19], [116], the STRUCTURE results show that whereas the CEU and YRI are genetically homogenous, the LWK, ASW and MKK are admixed, with a 20% CEU admixture in ASW. The LWK and ASW also have a large admixture with YRI (66% and 76% respectively), while MKK have a smaller admixture with YRI (10%). In addition, the STRUCTURE results indicate that MKK have a 15% admixture with two populations that are not sampled in the HapMap study. We also see a small admixture between MKK and LWK, which is expected, given their geographical proximity. These results are largely consistent with linguistic phylogeny; whereas the Maasai speak a Nilo-Saharan language, the Luhya and the Yoruba speak Niger-Congo languages, also spoken by African ancestors of African Americans [116].

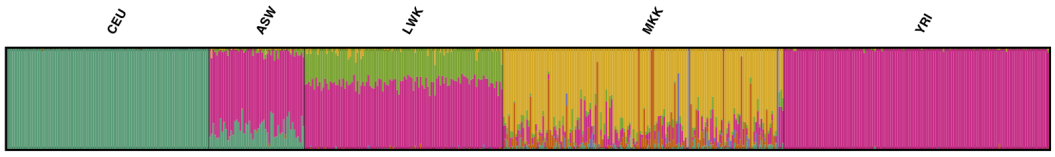


Figure 4.2: **Population structure components for individuals from CEU, ASW, LWK, MKK and YRI.** Results from STRUCTURE version 2.3 on genotype data for 12,999 randomly selected SNPs in 578 founder (unrelated) individuals from the CEU, ASW, LWK, MKK and YRI HapMap populations. The no-admixture model showed that the data was best fit by 6 inferred ancestral populations. Each column represents an individual, and the colors indicate the fractions of their genotype attributable to ancestry from each of the 6 inferred populations.

This analysis shows that k=6 populations is overwhelmingly the most likely fit to the data under the no admixture model. We then used *distruct* [117] to create the images shown in Figures 4.2 and 4.3. The 20% European admixture in ASW is clearly visible.

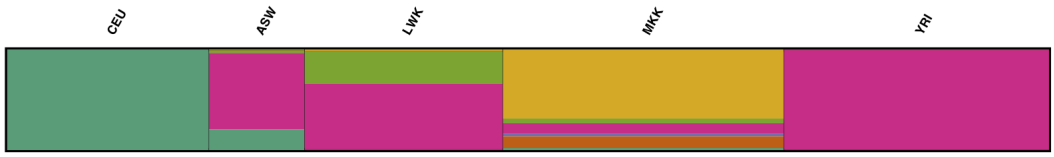


Figure 4.3: **Population structure components for populations CEU, ASW, LWK, MKK and YRI.** Results from STRUCTURE version 2.3 on genotype data for 12,999 randomly selected SNPs in 578 founder (unrelated) individuals from the CEU, ASW, LWK, MKK and YRI HapMap populations. The no-admixture model showed that the data was best fit by 6 inferred ancestral populations. Each column represents a sample population, and the colors indicate the fractions of their genotype attributable to ancestry from each of the 6 inferred populations.

Table 4.1: **Log Likelihood values of STRUCTURE analysis**

k = Number of Inferred Populations	$\ln P(X\|K)$ = Estimated Log Probability of Fit	Relative Log Probability of Fit
1	-7124377.3	-382966.4
2	-6792102.1	-50691.2
3	-6756124.2	-14713.3
4	-6752204.2	-10793.3
5	-6769565.8	-28154.9
6	-6741410.9	0
7	-7247376.2	-505965.3
8	-7242335.3	-500924.4

The log likelihood values of the fits from Structure or various numbers of inferred populations.

It is also clear that the Maasai are highly admixed, with a clear African admixture from YRI and LWK as well as a 15% admixture with two other populations not represented in HapMap (Table 4.2).

To further quantify the genetic similarity of MKK, LWK, ASW and YRI to the six ancestral populations, we assigned a six component vector to each of these populations, whose coordinates were the fraction of the ancestral components represented in them. A comparison of the cosine similarity of these vectors showed that the largest overlap was between MKK and LWK (0.18), followed by MKK and ASW (0.16). Based on their closer proximity to MKK in the PCA plot, as well as closer cosine similarity, we chose the LWK as the appropriate reference population for the F_{ST} and XP-EHH analysis.

Table 4.2: **Population structure components identified by STRUCTURE**

HapMap Population ID	Teal	Brown	Purple	Pink	Green	Yellow	Number of founders
CEU	0.998	0	0	0	0	0	112
ASW	0.202	0.003	0.002	0.759	0.026	0.009	53
LWK	0.001	0.004	0.002	0.655	0.32	0.018	110
MKK	0.02	0.118	0.028	0.101	0.046	0.687	156
YRI	0.001	0.002	0.001	0.992	0.003	0.002	147

The percentage contributions to CEU, ASW, LWK, MKK and YRI from the 6 ancestral population groups inferred by STRUCTURE. The ancestral groups are labeled as different colors, indicated on Figure 4.2.

4.3.2 Selection based on F_{ST}

We calculated F_{ST} between MKK (n = 156) and LWK (n = 110) as in [108] for 1,175,055 SNPs common to both populations that passed filters for minor allele frequency, genotyping rate, and consistency with Hardy-Weinberg equilibrium. Statistical significance was assessed using a Bonferroni corrected permutation test p-value p_B (described in Methods). Within the SNPs that passed this filter, we identified those deviating significantly from neutral evolution using an empirical p-value (p_E) based on the F_{ST} distribution of intergenic SNPs. This identified 1,232 SNPs with $p_B < 8.6 \times 10^6$ and $p_E < 0.001$ (Table S1 in [2]) which were either genic or within 50 kb of genes.

In a recent selective sweep, many neighboring SNPs may remain linked due to genetic hitchhiking. To identify such regions, we grouped the genome-wide significant SNPs identified by F_{ST} into clusters based on linkage disequilibrium using the criterion that each SNP has genotype $r^2 \geq 0.25$ with at least one other SNP in the cluster (described in Methods). Each cluster so identified is a candidate for a selective sweep in one of the two populations. To identify the population in which the sweep is most likely to have occurred, we compared the local haplotype diversity in each population using the XP-EHH score [17]. For each cluster identified by F_{ST}, we label it as a selection candidate in MKK if the maximum XP-EHH score of a SNP in the cluster is > 3. A positive value for XP-EHH indicates that the MKK carry the longer-range haplotypes. This procedure identified 26 clusters (containing 318 SNPs) as candidate regions for selective sweeps in MKK (Table S2 in [2]). Nine of these clusters include SNPs that exceed the genome-wide significance threshold for XP-EHH (XP-EHH > 4.79580, Bonferroni corrected $p < 0.05$,

two-tailed). The most significant genomic regions and non-synonymous SNP candidates under selection in MKK by F_{ST} are listed in Table 4.3 and Table 4.4 respectively. Note that the isolated SNPs identified in Table 4.4 have high F_{ST} with respect to at least two of the three possible reference African populations (ASW, LWK and YRI). This suggests that the results shown there do not strongly depend on the choice of the reference population.

Chr	Start location	Stop location	Genes in region	Number of High Fst SNPs (empirical p-value <0.001)	Max Fst within cluster	Max XP-EHH score within cluster
2	135036696	136726567	RAB3GAP1, ZRANB3, DARS, R3HDM1, TMEM163, YSK4, LCT, UBXN4, MCM6, MGAT5, CCNT2	123	0.382	12.202
2	78305622	78500655	-	33	0.311	3.805
12	56402204	56754137	PAN2, OBFC2B, SLC39A5, APOF, STAT2, CS, RNF41, IKZF4, SMARCC2	28	0.283	3.024
3	191929784	191990575	FGF12	13	0.272	5.222
5	115126388	115223035	ATG12, AP3S1	7	0.266	3.870
2	163048404	163152351	IFIH1, FAP	19	0.261	3.108
7	99053816	99436198	ZNF498, CYP3A4, CPSF4, CYP3A7, CYP3A43	17	0.260	3.290
1	12296232	12319994	VPS13D	4	0.253	3.060
22	49978502	50077531	-	4	0.244	3.732
5	32128179	32159329	GOLPH3	5	0.242	3.062
5	14747247	14750823	ANKH	4	0.237	6.800
14	36033703	36201722	RALGAPA1	4	0.221	3.517
2	136917330	136921703	-	2	0.218	8.549
1	198692364	198745866	PTPRC	2	0.212	3.138
2	137580234	137595545	-	4	0.209	4.871
12	111414527	111502280	CUX2	5	0.209	3.393
17	75423198	75431978	SEPT9	3	0.200	5.024
18	66714832	66724690	CCDC102B	4	0.200	5.704
1	74807337	74842787	TNNI3K	3	0.193	3.993
3	185752767	185805993	ETV5	3	0.192	4.569

1,232 SNPs with significant Fst scores (p_B<8.6E−6, p_E<0.001) were clustered into contiguous genomic regions of linkage disequilibrium. A cluster was defined as a collection of SNPs in a genomic region where each SNP had genotype R^2≥0.25 with at least one other SNP in the cluster. Clusters containing a SNP with maximum XP-EHH score >3 were identified as being MKK associated. The 22 top clusters are ranked by the highest Fst value for a SNP pair in a cluster. The complete set of clusters identified by Fst is in Table S2.
doi:10.1371/journal.pone.0044751.t001

Table 4.3: **Top 20 genomic regions identified as selection candidates in MKK using the F_{ST} statistic and clustering. Table reproduced from [2]**

4.3.3 Selection based on iHS

Recent selective sweeps amplify beneficial mutations and reduce haplotype diversity due to the hitchhiking effect. The Extended Haplotype Homozygosity [14] (EHH) statistic identifies such events without using a reference population. $EHH(x)$ measures the probability that two randomly selected haplotypes sharing the same allele at a SNP are identical up to genomic distance x. At each SNP, we computed the unstandardized

Rsid of SNP	Chr	Position	Gene	Bonferroni corrected Permutation p-value (pB)	Empirical p-value (pE) using distribution of non-coding SNPs	Fst MKK vs LWK	Fst MKK vs YRI	Fst MKK vs ASW
rs2241883	2	88424066	FABP1	1.72E−12	3.13E−05	0.250	0.172	0.152
rs961360	2	136393658	R3HDM1	3.13E−08	3.13E−04	0.199	0.288	0.447
rs6997753	8	142487937	FLJ43860	4.87E−08	3.59E−04	0.194	0.138	0.006
rs531503	7	100377082	ZAN	3.83E−07	5.47E−04	0.182	0.014	0.073
rs17014118	4	89319296	HERC6	4.42E−07	6.06E−04	0.180	0.178	0.045
rs2271586	11	3659993	ART5	4.76E−07	6.06E−04	0.180	0.034	0.004
rs10930046	2	163137983	IFIH1	1.24E−06	6.86E−04	0.176	0.279	0.128
rs1051334	12	71523134	TSPAN8	1.36E−06	6.86E−04	0.176	0.173	0.104
rs10475299	5	5461233	KIAA0947	1.46E−06	6.86E−04	0.175	0.160	0.198
rs1918496	12	56722060	PAN2	3.06E−06	8.17E−04	0.171	0.296	0.074
rs13389745	2	65298657	CEP68	3.84E−06	8.17E−04	0.172	0.115	0.052
rs846266	7	42088222	GLI3	2.54E−06	9.42E−04	0.169	0.150	0.059
rs3813227	2	73651967	ALMS1	6.02E−06	9.82E−04	0.167	0.173	0.034

The most significant non-synonymous SNPs identified as candidates for selection by Fst. The complete list of 1,232 SNPs identified as selection candidates by Fst (p_B <8.6E−6 and p_E <0.001) is in Table S1.
doi:10.1371/journal.pone.0044751.t002

Table 4.4: **The most significant non-synonymous SNPs under selection in MKK using F_{ST}, with LWK as the reference population. Table reproduced from [2]**

Integrated Haplotype Score [16] (iHS), defined as the logarithm of the ratio of the integrated EHH scores for the ancestral allele and the derived allele. Stratifying the data into bins by the derived allele frequency of the SNPs, the scores within each bin were then normalized to have zero mean and unit standard deviation. The iHS statistic is less sensitive to demographic history (e.g. population bottlenecks) and to local differences in recombination rates, because such factors have similar effects on ancestral and derived alleles, and tend to cancel in the ratio [16]. If either allele is under selection, the reduced haplotype diversity around it will tend to increase the absolute value of iHS.

Following the protocols in [16], raw iHS scores for 991,737 SNPs in MKK (n = 143 individuals) that passed filters (minor allele frequency cutoff, consistency with Hardy-Weinberg equilibrium) were binned by derived allele frequency and standard normalized within each bin (described in Methods). Genomic regions were scored by the fraction of high scoring iHS SNPs ($|iHS|> 2$) using a sliding window of 50 SNPs. The top 0.02% of non-overlapping SNP windows identified 196 regions likely to be under selection (Table S3 in [2]). These were further grouped on the basis of linkage disequilibrium using the same criterion as for F_{ST} (genotype $r^2 \geq 0.25$). The most significant regions identified

as candidates for selection in MKK are in Table 4.5 (the complete list is in Table S4 in [2]).

Chr	Cluster start position (GRCh37)	Cluster end position (GRCh37)	Genes	Max |iHS| in cluster	# of SNPs in cluster with |iHS| >2
2	134221398	137892309	LCT, MGAT5, NCKAP5, DARS, ZRANB3, R3HDM1, TMEM163, RAB3GAP1, THSD7B, CCNT2, YSK4, UBXN4, MCM6	6.339	545
13	30496779	30565298	–	5.234	26
7	20373632	20468718	ITGB8	5.012	45
2	176089888	176422005	–	4.626	69
11	110532348	110663647	ARHGAP20	4.480	36
9	83127968	83382243	–	4.471	59
5	14657062	14753764	FAM105B, ANKH	4.429	23
18	66652846	66765215	CCDC102B	4.402	33
11	34025053	34189564	CAPRIN1, NAT10, ABTB2	4.375	22
2	179421694	179606538	TTN	4.289	28
14	105792959	105907642	PACS2, MTA1	4.228	20
5	108990708	109217428	MAN2A1	4.219	50
9	107973277	108067684	SLC44A1	4.192	34
9	3869844	3919130	GLIS3	4.185	23
7	99053816	99314986	ZNF789, CPSF4, ATP5J2, FAM200A, ZNF655, ZNF498, CYP3A7, ZKSCAN5, CYP3A5	4.120	24
9	13812037	13867306	–	4.066	23
11	75470813	75678647	UVRAG, DGAT2	4.059	48
2	12294875	12366781	–	4.041	24
14	97426813	97505011	–	4.025	24
8	145839058	146082167	COMMD5, LOC100287170, LOC100129596, ARHGAP39, RPL8, ZNF7, ZNF251, ZNF34, LOC100287297, ZNF517	3.955	22

Using a sliding window of 50 SNPs wide, genomic regions were scored for the fraction of SNPs with |iHS|>2. The top 0.02% of non-overlapping windows were identified and merged into genomic clusters based on genotype R^2 using the same criterion as in Table 1. Clusters are ranked by the maximum |iHS| value in the cluster. Complete lists of genome-wide significant SNPs and regions identified by iHS are in Tables S2a and S2b respectively.
doi:10.1371/journal.pone.0044751.t003

Table 4.5: **The most significant genomic regions under selection in MKK using iHS. Table reproduced from [2]**

4.3.4 Selection based on XP-EHH

The third method used to identify selective sweeps in MKK was the Cross Population Extended Haplotype Homozygosity statistic (XP-EHH) [17]. This statistic compares the EHH profiles for bi-allelic SNPs between two populations. It is defined as the log of the ratio of the integrals of the EHH profiles for a given allele between the two populations (described in Methods). The comparison between populations normalizes the effects of large-scale variations in recombination rates on haplotype diversity, and has a higher statistical power to detect sweeps that are close to fixation [17].

Using the LWK cohort (n = 90) as the reference population for MKK (n = 143), XP-EHH was calculated for 1,373,755 SNPs that passed various filters (further details

provided below). Following [17], we assigned p-values using a Gaussian fit after standard normalizing the XP-EHH distribution. SNPs with Bonferroni corrected p-value < 0.05 (two-tailed) were chosen as potentially significant candidates for selection. These are listed in Table S5 in [2]. We also clustered these candidate SNPs (using the genotype $r^2 \geq 0.25$ criterion as before) to identify putative regions under selection in MKK (Table S6 in [2]). The most significant regions thus identified are listed in Table 4.6.

Chr	Start Position	End Position	Genes	Number of SNPs	Max XP-EHH
2	135058615	137017060	R3HDM1, MGAT5, RAB3GAP1, LCT, DARS, ZRANB3, MCM6, TMEM163, ACMSD, CCNT2, YSK4, UBXN4, CXCR4	572	12.182
5	14681797	14751400	FAM105B, ANKH	25	6.800
18	66712510	66731187	CCDC102B	12	5.587
5	115885282	115922669	SEMA6A	21	5.482
18	66768031	66777543	–	5	5.324
20	4513311	4522535	–	10	5.313
13	104870241	104880533	–	7	5.183
4	64594290	64639661	–	16	5.149
2	134507165	134561145	–	12	5.062
16	75360734	75364940	CFDP1	2	5.040
17	75427551	75428021	SEPT9	2	5.024
3	191943578	191989642	FGF12	10	5.019
11	117610387	117620420	DSCAML1	8	4.989

SNPs with positive genome-wide significant XP-EHH scores (XP-EHH ≥4.796, two-tailed Bonferroni corrected p≤0.05) were grouped into contiguous genomic clusters using genotype R^2 using the same criterion as in Table 1. Overlapping clusters were merged. Column E lists the number of significant SNPs in each each cluster. Complete lists of genome-wide significant SNPs and clusters identified by XP-EHH are in Tables S3a and S3b.
doi:10.1371/journal.pone.0044751.t004

Table 4.6: **The most significant genomic regions under selection in MKK using XP-EHH, with LWK as the reference population. Table reproduced from [2]**

4.3.5 Overlap of high scoring regions

The metrics we use probe for different signatures of selection, and hence, genomic regions which are identified by more than one metric are more likely to be true positives. Using a concordance between at least two of the metrics, we identified seven genomic regions as strong candidates for selection (Table 4.7). There was also overlap between the regions identified by our methods and those identified by the International HapMap Consortium for MKK (they used a statistic they call CMS or Composite of Multiple Signals) [19]. These regions of concordance are listed in Table S7 in [2] and summarized in Table 4.7. Figure 4.4 shows the results for all three metrics for chromosome 2. The significant selection in a region in Chr2q21 of size $1.0 - 1.7$ Mb is clearly visible in

Figure 4.4a. Figure 4.4b shows details of this region which contains a large number of polymorphisms with significant high scores by all three metrics (discussed below).

Chr	Genomic Extent	Significant by (Method)	Genes in Region	Number of SNPs identified by each Method
2	135058615–136726567	Fst, iHS, XP-EHH	MGAT5, TMEM163, ACMSD, CCNT2, YSK4, RAB3GAP1, ZRANB3, R3HDM1, UBXN4, LCT, MCM6, DARS	Fst: 123, iHS: 545, XP-EHH: 572
3	191943578–191989642	Fst, XP-EHH	FGF12	Fst:13, XP-EHH: 10
5	14747247–14750823	Fst, iHS, XP-EHH	ANKH	Fst: 4, iHS: 23, XP-EHH: 25
5	115885574–115885672	Fst,XP-EHH	SEMA6A	Fst: 2, XP-EHH: 21
7	99053816–99314986	Fst, iHS	ZNF789, CPSF4, ATP5J2, FAM200A, ZNF655, ZNF498, CYP3A7, ZKSCAN5, CYP3A5	Fst: 17, iHS: 24
17	75427551–75428021	Fst, XP-EHH	SEPT9	Fst: 3, XP-EHH: 2
18	66714832–66724690	Fst, iHS, XP-EHH	CCDC102B	Fst: 4, iHS: 33, XP-EHH: 12

Genomic regions identified as genome-wide significant by at least two of the three methods - Fst, iHS and XP-EHH.
doi:10.1371/journal.pone.0044751.t005

Table 4.7: **Concordant genomic regions identified by at least two of three metrics as candidates for selection in MKK. Table reproduced from [2]**

We found that the non-synonymous SNP with the highest genome-wide significant F_{ST} was rs2241883 in the gene Fatty Acid binding Protein 1, Liver (*FABP1*, alternative name *LFABP*) (Table 4.4 and Figure 4.4a). The SNP rs2241883 is a T→C non-synonymous transition which encodes a Threonine to Alanine (T94A) change in the protein *LFABP*, which is expressed in liver. The C allele was associated with total tri-glyceride and low density lipoprotein (LDL) cholesterol levels in Germans [118], and with Apolipoprotein B (ApoB) levels induced by a high fat diet in French-Canadians [119]. The MKK have high F_{ST} at this SNP, relative to all the other three African populations in Hapmap (Table 4.4). The allele frequency of the C allele is also highest (0.44) in MKK compared to all other HapMap3 populations (in which the frequency ranges from $0.09 - 0.32$). These results suggest that the rs2241883 polymorphism is under selection in the Maasai.

4.3.6 Maasai are under Selection in a 1.7 Mb Region on Chr2q21 for Lactase Persistence

The largest cluster under selection in Maasai, identified by all the metrics, was a 1.7 Mb region on Chr2q21 (Figures 4.4a, 4.4b, Tables 4.3, 4.4, 4.5, 4.6). The region includes the Lactase (LCT) gene, which encodes the Lactase protein, as well as the gene *MCM6*, which contains intronic regulatory regions for LCT [18], [120], [121], [122]. Specific

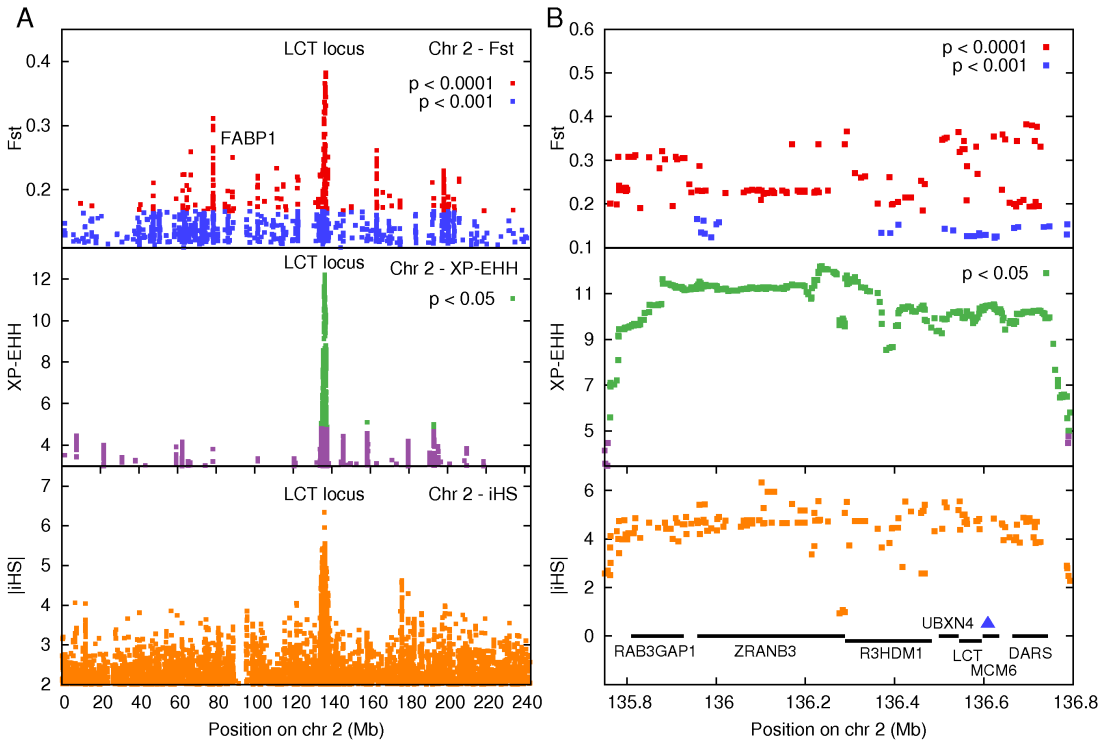


Figure 4.4: **(a) Genome-wide significant scores identifying candidate regions under selection on Chromosome 2.** Chromosome wide plot of SNPs with significant scores using F_{ST} (empirical p-value < 0.001 and Bonferroni corrected permutation test $p_B < 8.610^6$), iHS (normalized $|iHS|> 2$), and XP-EHH (XP-EHH ≥ 4.796, two-tailed Bonferroni corrected $p \leq 0.05$). The SNPs thus identified were clustered on the basis of linkage disequilibrium to identify contiguous genomic regions that are candidates for selections (Table 4.3, 4.4, 4.5, 4.6, 4.7). The locus containing the genes *LCT* and *MCM6* (135-137 Mb) was identified by all three metrics as the top candidate for selection. The non-synonymous T→C polymorphism at rs2241883 in the *FABP1* gene had most significant genome-wide F_{ST} ($F_{ST} = 0.25$, $p_E = 3.1310^5$). The MKK samples have a high frequency (0.45) of the protective C allele, known to be associated with low cholesterol levels in Europeans. **(b) Inset of the *LCT* locus on Chromosome 2.** An inset of the F_{ST}, iHS and XP-EHH scores for SNPs in the 1 Mb locus (from 135.8-136.8 Mb) on Chr 2 containing the genes *LCT* and *MCM6*. The uniformly high values for all three metrics in this region suggest that this locus has undergone strong selection pressure. The blue marker indicates the position of the lactase associated SNP in *MCM6* that we sequenced, which was polymorphic in MKK with frequency $p_C = 0.58 \pm 0.14$ (68% CI) for the protective C allele. Figure reproduced from [2]

polymorphisms in these regions are known to confer lactase persistence in Europeans and Africans [18], [120]. Our results are in agreement with other studies that have also shown that this region is under recent, positive selection in the Maasai [18], [17], [14], [16], [123], [29].

To identify specific polymorphisms for adult lactase persistence in the Maasai, we sequenced DNA from six founder MKK samples (HapMap IDs: NA21367, NA21379, NA21454, NA21519, NA21522, NA21650) at five loci in MCM6 (G/C-14010, rs41525747, rs4988235, rs41380347 and rs182549), which are known to be associated with lactase persistence in Africans and Europeans [18]. We found that the GC-14010 polymorphism in the *MCM6* gene is segregating in these samples ($n_{GG} = 1$, $n_{GC} = 3$, $n_{CC} = 2$). We estimated the frequency of the beneficial (C) allele in the MKK samples to be $p_C = 0.58 \pm 0.14$ (68% CI from finite size sampling - details in Appendix S5 in [2]). This is in agreement with Tishkoff *et al.* [18], who showed that this allele is significantly associated with lactase persistence, has significantly reduced haplotype diversity indicative of a selective sweep, and is segregating at high frequency in Maasai individuals from Kenya.

4.3.7 The Selected Locus on Chr2q21 Contains Polymorphisms Associated with Cholesterol Levels

The selected locus on Chr2q21 contains polymorphisms that have been associated with cholesterol levels in various GWAS studies [124], [125], [126]. The SNP rs7570971 in *RAB3GAP1*, not found in the HapMap data for the MKK, is associated with total cholesterol levels in a GWAS of $> 100,000$ individuals of European descent [124]. However, the six MKK samples we sequenced were homozygous at this locus in the Maasai for the allele associated with an increase in total cholesterol levels in the samples with European descent.

A study in a Finnish cohort identified polymorphisms in LCT associated with total cholesterol and Low Density Lipoprotein C (LDL-C) levels [125]. The authors found that the lactase persistence genotype in Finns, as defined by the genotype for SNP rs4988235, was associated with lower cholesterol values. Several SNPs in and around

the gene *LCT* were associated with total cholesterol and LDL-C levels, with stronger associations in males than females. This study also found that the G allele at the synonymous SNP rs2304371 in the *LCT* gene was associated with highest LDL-C levels in males. The same SNP was identified by our methods as a selection candidate (Tables S1, S2, S3 in [2]). However, once again, the major allele in the MKK (frequency 87%) was the one associated with higher LDL-C levels.

4.3.8 The CYP3A Locus is a Candidate for Selection in Maasai

On Chromosome 7, a 261 kb wide region spanning the entire Cytochrome P450 Subfamily 3A (*CYP3A*) locus was identified as a candidate for selection by F_{ST} and iHS (Tables 4.3, 4.5). All *CYP* genes in this locus contain SNPs with genome-wide significant F_{ST} or iHS scores, including: *CYP3A4* (a potent oxidizer of steroids and drugs), *CYP3A5* (involved in oxidation of fatty acids and steroids in the liver), *CYP3A7* (the main *CYP* enzyme expressed in fetal livers) and *CYP3A43* (involved in testosterone metabolism). The *CYP* proteins play an important role in drug metabolism and in the synthesis of steroids from cholesterol [127].

4.4 Discussion

In spite of diet that is rich in fat and cholesterol, the Maasai have low blood cholesterol levels and low incidence of heart disease and atherosclerosis. Cholesterol challenge studies in the 1970s [86] demonstrated that the Maasai are able to maintain cholesterol homeostasis in response to elevated levels of dietary cholesterol, and suggested that the mechanism of cholesterol homeostasis may have a genetic basis. In the present study, we used HapMap 3 data to investigate this possibility. Using 90-110 unrelated LWK individuals as a reference population, three complementary metrics (F_{ST}, iHS and XP-EHH) were used to identify SNPs and chromosomal regions under selection in 143-156 unrelated MKK (Maasai) individuals in HapMap 3. The genomic regions and genes identified as selection candidates in MKK are shown in Tables 4.4, 4.5, and 4.6 for the F_{ST}, iHS, and XP-EHH metrics respectively. We identified seven genomic regions as

strong candidates for selection using concordance between at least two of the metrics (Table 4.7). We now discuss some of the most interesting SNPs and regions identified for the role they may play in lactase persistence and lipid pathway selection in the Maasai.

Using F_{ST}, the most significant non-synonymous SNP was the polymorphism rs2241883 located at 88.42 Mb on Chromosome 2 (Figure 4.4a, Table 4.4). This is a Threonine to Alanine substitution (T94A) in exon 3 of the FABP1 (or LFABP) gene, a fatty acid binding protein expressed in liver. This locus was not detected by iHS or XP-EHH, suggesting either an increased local recombination rate or a more ancient selective sweep. The T94A polymorphism was strongly associated with lower levels of plasma triglycerides and LDL-cholesterol levels in a study of 826 individuals from Northern Germany [118]. A study of plasma concentrations of ApoB in 623 French Canadian men found that carriers of the A94 allele were protected against high ApoB levels when consuming a high fat and saturated fat diet, possibly because of diminished function of the protein *LFABP* due to a disruption in ligand binding [118]. *LFABP* knockout mice fed a high cholesterol, high saturated fat diet were protected against diet-induced obesity and lower levels of hepatic triglycerides compared to control mice, despite the absence of discernible differences in energy levels, food intake, or mal-absorption of fat induced obesity [128], [129]. The study concluded that "*LFABP* may function as a metabolic sensor in regulating lipid homeostasis" [128]. The protective C allele of this SNP is segregating in the Maasai at allele-frequency 0.44, suggesting that the effect of the T94A mutation on the LFABP pathway may be partly responsible for the homeostatic regulation of blood cholesterol in Maasai [85], [86], [87].

We found evidence for a strong recent selective sweep in a 1.7 Mb region on Chr2q21 (Figure 4.4, Table 4.3, 4.4, 4.5, 4.6, 4.7). This region is known to harbor polymorphisms conferring lactase persistence in Kenyans, and has been shown to be under strong recent selection. Tishkoff *et al.* [18] performed a phenotype-genotype association study for lactase persistence on 470 Tanzanians, Kenyans and Sudanese who were genotyped at 123 SNPs, in a 3 Mb region surrounding the *LCT* and *MCM6* genes. The SNP known as G/C-14010 was found to have the most significant association with the lactase

persistence phenotype in Kenyan Nilo-Saharan and Tanzanian Afro-Asiatic populations, as well as in a meta-analysis of all the populations combined. Tishkoff *et al.* observed the C-14010 allele to occur at 32% frequency in Kenyan populations. As this SNP is in the upstream regulatory region of the gene *LCT*, the authors also studied the effect of this polymorphism on expression using luciferase assays in intestinal cells. They found that the C-14010 allele leads to a significantly higher expression. Furthermore, an iHS analysis of the haplotype background on which the SNP occurs indicated that the SNP is under selection in Kenyans and Tanzanians. We found that in the MKK samples from HapMap the C-14010 allele is segregating at high frequency (0.58). Thus, our results confirm the findings of Tishkoff *et al.*, that C-14010 contributes towards selection for lactase persistence in the MKK samples from HapMap.

In addition to lactase persistence, the GWAS studies of [124] and [125] indicate that, in Europeans, the locus on Chr2q21 is associated with cholesterol levels. As this locus is also identified by our analysis, it may be associated with cholesterol levels in the Maasai. However, the allelic variants of the GWAS SNPs of [124], [125] that have high frequency in MKK are associated with an increase in cholesterol levels in Europeans. This might reflect the possibility that Europeans and Maasai have different sets of functional polymorphisms at this locus responsible for lower cholesterol levels: indeed it is known that the Maasai have an African polymorphism associated with lactase persistence, different from the one found in Europeans. It could also be that in the Maasai, the SNPs identified in our study are not themselves functional, but linked to functional variants that are not genotyped. Given the extended linkage disequilibrium (LD) in this region due to a selective sweep in both Europeans and the Maasai, this last possibility is especially important. The differing effects of the SNPs identified in the Maasai, as compared with the Europeans, could arise from the effects of differing modifier alleles at different loci in this region. These possibilities emphasize the difficulties associated with identifying true functional polymorphisms because of potential population specificity of SNP based studies. However, given the GWAS findings, and the strong signal of selection in MKK seen in our analysis, the LCT locus is a candidate region for identifying genotypic variants associated with cholesterol regulation in the

Maasai.

We also identified a 261 kb locus on Chr 7 (the *CYP3A* locus) to be under selection using F_{ST} and iHS (Tables 4.3 and 4.5). This locus has been identified in re-sequencing studies and genome-wide scans to be under positive selection in Africans and non-Africans [16], [130], [131] and is also under positive selection for salt sensitivity in equatorial populations [130], [132]. This locus contains the *CYP3A* (cytochrome P450, subfamily 3A) family of genes which are involved in cholesterol metabolism and steroid biosynthesis [127]. This family contains *CYP3A5*, a gene involved in fatty acid oxidation in liver, as well as *CYP3A7*, a gene encoding a *CYP* enzyme expressed in fetal livers. Variants in *CYP3A5* have been shown to reduce the efficacy of certain statins, drugs used to lower cholesterol biosynthesis [133]. Thus, the selection pressure at this locus, as identified by our analysis, coupled with its role in cholesterol metabolism, suggests that the *CYP3A* locus is an important candidate for cholesterol homeostasis in the Maasai.

Several other clusters identified to be under selection in MKK contain genes related to cholesterol metabolism, cholesterol biosynthesis and atherosclerosis. On Chr12q13, we identified a region spanning many genes with one of the highest F_{ST} signals (Table 4.3). This locus contains the Apolipoprotein F (*APOF*) gene, involved in cholesterol transport and esterification [127], whose over-expression in mice reduces high density lipoprotein (HDL) cholesterol levels [134]. A cluster identified by iHS on chromosome 11q13.5 contains the gene Diacylglycerol O-acyltransferase 2 (*DGAT2*) (Table 4.5). This gene is involved in biosynthesis of triglycerols [135], [136] and has been implicated in hyperlipidemia [137] and fatty liver disease [138]. Another cluster on Chr7p21.1 identified by iHS, contains the Integrin Beta 8 (*ITGB8*) gene (Table 4.5) implicated as a quantitative trait locus (QTL) for fibrinogen plasma levels in a study involving 3600 Native Americans [139]. Fibrinogen levels are associated with risks for several cardiovascular diseases [140], and play a role in the pathogenesis of atherosclerosis [139]. XP-EHH identified a genome-wide significant region on chromosome 16q22.2-22.3, containing the gene Craniofacial Development Protein 1 (*CFDP1*) (Table 4.6). A GWAS showed that this region is associated with low levels of HDL cholesterol in 400

French-Canadians [141].

Our results identified several genes and loci involved in cholesterol metabolism as selection candidates in the Maasai. Thus, our findings suggest that the Maasai are adapted for a high-cholesterol and high-fat diet. The traditional diet of the Maasai is rich in saturated fats and cholesterol, and low in carbohydrates. Similar ketogenic diets are often used to treat epileptic seizures in children [142], [143]. Early complications of these diets include hypertriglyceridemia, hypercholesterolemia, and low levels of HDL, and late complications include osteopenia, renal stones, and cardiomyopathy [143], [144]. This suggests that a diet rich in fat and cholesterol from childhood can exert a strong diet-induced selection pressure on survival and reproductive success.

4.5 Future Directions

The Extended Haplotype Homozygosity (EHH) is a quadratic measure of diversity, and is the probability that two randomly selected haplotypes are identical up to a distance x from the core SNP. It is quantified as follows:

$$EHH(x) = \frac{\sum \binom{n_i}{2}}{\binom{\sum n_i}{2}} \approx \frac{\sum n_i^2}{(\sum n_i)^2} = \sum p_i^2$$

where n_i are the number of haplotypes of type i that are identical up to a distance x. The total number of haplotypes $= \sum_i n_i$.

The integrated haplotype score (iHS) is then defined (up to a negative sign) as the log ratio of the integral of this statistic for the two alleles (derived allele and ancestral allele) at a bi-allelic SNP.

$$iHS(x) = \log \frac{\int EHH_D(x)dx}{\int EHH_A(x)dx}$$

The central idea is that this quadratic measure of diversity, when integrated, is able to quantify the difference in the pattern of diversity surrounding a locus under selection from one that is evolving under neutral forces (see Figure 4.5). This raises the question: how do other measures of diversity, such as entropy based measures, perform in detecting selection?

We began to address this question as follows. Consider the Tsallis entropy [145] - one of the many generalizations of the Shannon entropy - which is defined as:

$$S_q(x) = \frac{1 - \sum_i p_i^q(x)}{q - 1}$$

For q = 1, this reduces to the Shannon entropy:

$$\lim_{q \to 1} S_q(x) = S_1(x) = -\sum_i p_i(x) \log p_i(x)$$

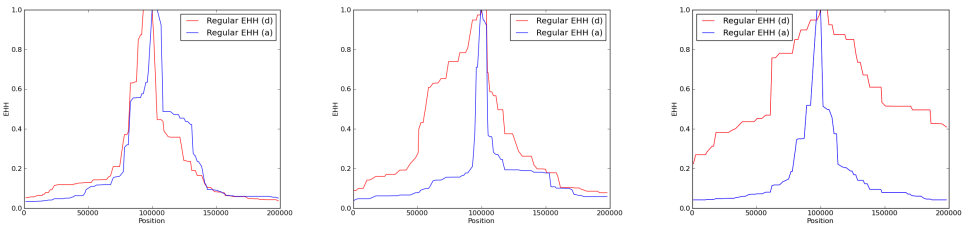
For q = 2,

$$S_2(x) = 1 - EHH(x)$$

This suggests a possible generalization of EHH:

$$S_q(x) = \frac{1 - \sum_i p_i^q(x)}{q - 1}$$

We then measure the statistical power of these metrics to detect selective sweeps in simulated SNP data. We simulate haplotype data at a number of SNPs using *mbs* [146], a modification of Hudson's coalescent simulation software *ms*, that includes the effect of selection. Using *mbs*, we can vary the population size N, as well as neutral parameters such as the effective mutation and recombination rates ($4N\mu$ and $4Nr$ respectively, where μ and r are the mutation and recombination rates per site per generation). We can also include the effect of selection on a single SNP, parameterized by the selection coefficient $4Ns$.



(a) (b) (c)

Figure 4.5: **Plot of Extended Haplotype Homozygosity (EHH) surrounding the derived allele and the ancestral allele for varied selection coefficients ($4Ns$) measured on haplotype data simulated using *mbs* [146].** (a) $4Ns = 0$ (neutral evolution) (b) $4Ns = 100$ (moderate selection) (c) $4Ns = 500$ (strong selection)

One can then take the simulated haplotype data generated by *mbs* and model the ascertainment bias involved in sampling real haplotype data (low frequency variants are less likely to be detected than high frequency variants).

For the N haplotypes, we created a binary distance matrix $C_{N\times N}(x)$ whose entries are 1 for identical haplotypes and 0 for dissimilar haplotypes. The haplotype frequencies p_i are then obtained from the eigenvalues $\lambda_i(x)$ of $C_{N\times N}(x)$.

$$p_i = \frac{\lambda_i}{\sum \lambda_i(x)}$$

From which we can calculate $\sum_i p_i^q$. Alternatively, we can calculate $\sum_i p_i^q$ in the following equivalent way:

$$\sum_i p_i^q = \frac{Tr(C^q)}{Tr(C)^q}$$

The Tsallis Entropy is then:

$$S_q(x) = \frac{1 - \sum_i p_i^q(x)}{q-1}$$

From which we can calculate an integrated score for the two alleles at a bi-allelic SNP.

$$IHS_q = \log \frac{\int S_{q_D}(x)dx}{\int S_{q_A}(x)dx}$$

$$\overline{IHS}_q = \log \frac{\int 1 - S_{q_D}(x)dx}{\int 1 - S_{q_A}(x)dx}$$

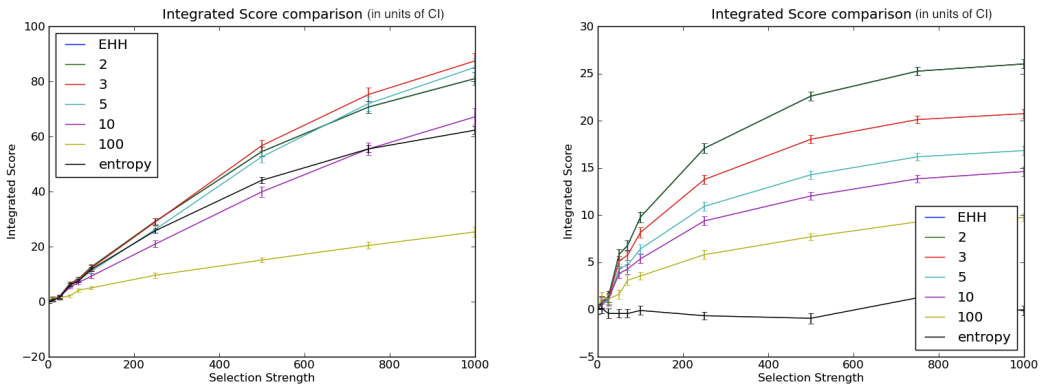
We studied the ability of IHS_q and $\overline{IHS}_q$ to detect selection sweeps for different values of q. Higher powers of q indicate a weighting that is more strongly biased by the frequency of the most frequent haplotype. The following figure shows the values of

$$\frac{IHS_q \mid_{\text{selection}} - IHS_q \mid_{\text{neutral}}}{CI(IHS_q \mid_{\text{neutral}})}$$

and similarly for $\overline{IHS}_q$, where CI represents the width of a 95% confidence interval.

Figure 4.6 indicates that IHS_q is not strongly dependent on the value of q, for $q < 10$. Furthermore, IHS_q shows a linear dependence on selection strength, and a larger difference beween scores of selected loci and neutral loci, as compared to the conventional integrated haplotype score $= \overline{IHS}_2$.

We further assessed the dependence of IHS_q on the recombination rate $\rho = 4Nr$. Ideally, these integrated scores should not depend strongly on the recombination rate



(a) (b)

Figure 4.6: **Comparison of performance of two integrated selection metrics, IHS_q and $\overline{IHS}_q$, to detect sweeps of varied selection strengths.** The plots shows the value of the difference between the average selected score and the average neutral score, divided by the 95% the confidence interval of the neutral score. 100 simulation replicates were run to generate these figures. Error bars indicate $\pm 1SE$. (a) $IHS_q = \log \frac{\int S_{q_D}(x)dx}{\int S_{q_A}(x)dx}$ (b) $\overline{IHS}_q = \log \frac{\int 1-S_{q_D}(x)dx}{\int 1-S_{q_A}(x)dx}$

because the effect of recombination is common to haplotypes with the ancestral allele and those with the derived allele, and should therefore cancel out in the ratio. This agrees with the simulation results (Figure 4.7).

Hence, the Tsallis Entropy may provide a framework to extend the integrated Haplotype Score (iHS) of Voight *et al.* [16]. In particular, a future project in this direction could explore some of the following:

- Include a model of ascertainment bias to test the performance of selection metrics on ascertained data

- Optimize selection metrics to reduce the dependence on neutral forces such as the recombination rate $\rho = 4Nr$, the mutation rate $\theta = 4N\mu$, and the frequency of the derived allele.

- Investigate the functions $f(q)$ that optimize the ability of the integrated score obtained from $S = \int f(q)S_q dq$ or $S = \sum_q f_q S_q$ to detect selective sweeps, and reduce the overall false positive rate over realistic values of ρ, θ and selection strength.

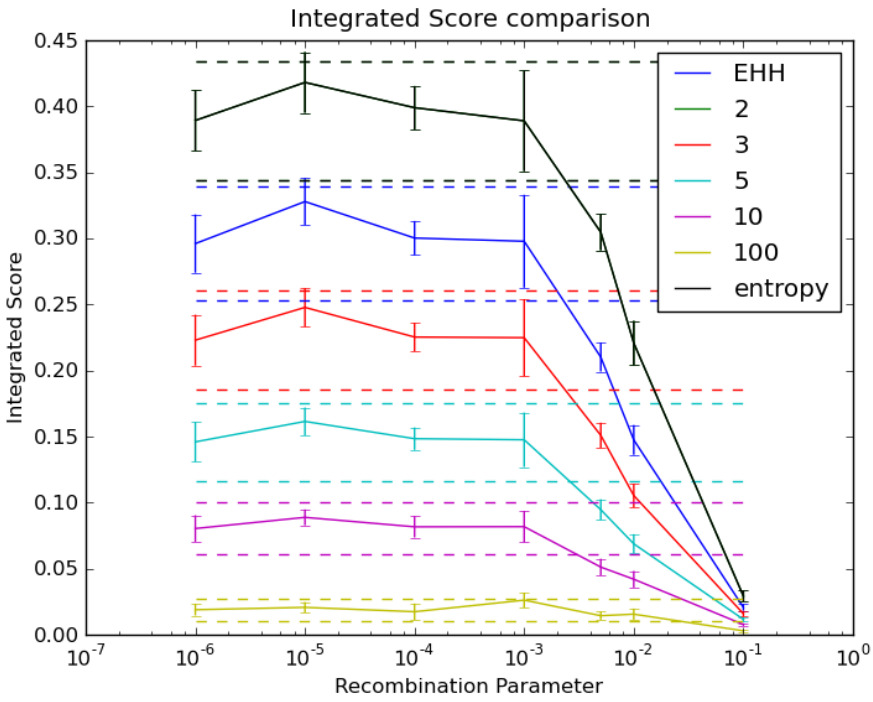


Figure 4.7: **Dependence of selection metric IHS_q on recombination rate.** For a wide range of realistic recombination parameters, the selection metrics are not significantly affected by the variation in recombination rate. Error bars indicate $\pm 1SE$.

- Determine whether a non-binary distance matrix $C_{N\times N}(x)$ (such as a correlation matrix) can improve the performance of IHS_q.

References

[1] Tomar P, Bhatia A, Ramdas S, Diao L, Bhanot G, et al. (2013) Sporulation genes associated with sporulation efficiency in natural isolates of yeast. PloS one 8: e69765.

[2] Wagh K, Bhatia A, Alexe G, Reddy A, Ravikumar V, et al. (2012) Lactase persistence and lipid pathway selection in the maasai. PloS one 7: e44751.

[3] Steinmetz LM, Sinha H, Richards DR, Spiegelman JI, Oefner PJ, et al. (2002) Dissecting the architecture of a quantitative trait locus in yeast. Nature 416: 326–330.

[4] Flint J, Mott R (2001) Finding the molecular basis of quatitative traits: successes and pitfalls. Nature Reviews Genetics 2: 437–445.

[5] Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nature genetics 11: 241.

[6] Feldmann H (2012) Yeast: Molecular and Cell Biology. John Wiley & Sons.

[7] Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454: 479–485.

[8] Gibson G (2012) Rare and common variants: twenty arguments. Nature Reviews Genetics 13: 135–145.

[9] Reich DE, Lander ES (2001) On the allelic spectrum of human disease. TRENDS in Genetics 17: 502–510.

[10] Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease–common variant or not? Human molecular genetics 11: 2417–2423.

[11] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.

[12] Broman KW, Wu H, Sen Ś, Churchill GA (2003) *R/qtl*: *QTL* mapping in experimental crosses. Bioinformatics 19: 889–890.

[13] Broman KW, Sen S (2009) A guide to *QTL* mapping with *R-qtl*. Springer.

[14] Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–7.

[15] Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, et al. (2006) Positive natural selection in the human lineage. Science (New York, NY) 312: 1614–20.

[16] Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS biology 4: e72.

[17] Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–8.

[18] Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nature genetics 39: 31–40.

[19] Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–8.

[20] Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297–304.

[21] King MC, Marks JH, Mandell JB, et al. (2003) Breast and ovarian cancer risks due to inherited mutations in brca1 and brca2. Science 302: 643–646.

[22] Strachan T, Read A (2004) Human Molecular Genetics 3. Garland Science. URL `http://books.google.com/books?id=g4hC63UrPbUC`.

[23] Hawks J, Hunley K, Lee SH, Wolpoff M (2000) Population bottlenecks and pleistocene human evolution. Molecular Biology and Evolution 17: 2–22.

[24] Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: Proceedings of the sixth international congress on genetics. volume 1, pp. 356–366.

[25] Gillespie JH (2010) Population genetics: a concise guide. JHU Press.

[26] Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, et al. (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329: 75–78.

[27] Ingram CJE, Raga TO, Tarekegn A, Browning SL, Elamin MF, et al. (2009) Multiple rare variants as a cause of a common phenotype: several different lactase persistence associated alleles in a single ethnic group. Journal of molecular evolution 69: 579–88.

[28] Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG (2009) The origins of lactase persistence in Europe. PLoS computational biology 5: e1000491.

[29] Coelho M, Luiselli D, Bertorelle G, Lopes AI, Seixas S, et al. (2005) Microsatellite variation and evolution of human lactase persistence. Human genetics 117: 329–39.

[30] Kimura M (1962) On the probability of fixation of mutant genes in a population. Genetics 47: 713.

[31] McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics 9: 356–369.

[32] Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics 11: 459–463.

[33] Pigliucci M (2001) Phenotypic plasticity: beyond nature and nurture. Johns Hopkins University Press.

[34] Scheiner SM (1993) Genetics and evolution of phenotypic plasticity. Annual review of ecology and systematics : 35–68.

[35] Pigliucci M (2005) Evolution of phenotypic plasticity: where are we going now? Trends in Ecology & Evolution 20: 481–486.

[36] Wu R (1998) The detection of plasticity genes in heterogeneous environments. Evolution : 967–977.

[37] Shook DR, Johnson TE (1999) Quantitative trait loci affecting survival and fertility-related traits in *Caenorhabditis elegans* show genotype-environment interactions, pleiotropy and epistasis. Genetics 153: 1233–1243.

[38] Via S, Gomulkiewicz R, De Jong G, Scheiner SM, Schlichting CD, et al. (1995) Adaptive phenotypic plasticity: consensus and controversy. Trends in Ecology & Evolution 10: 212–217.

[39] Schlichting CD, Pigliucci M (1995) Gene regulation, quantitative genetics and the evolution of reaction norms. Evolutionary Ecology 9: 154–168.

[40] Broach JR (2012) Nutritional control of growth and development in yeast. Genetics 192: 73–105.

[41] Zaman S, Lippman SI, Zhao X, Broach JR (2008) How saccharomyces responds to nutrients. Annual review of genetics 42: 27–81.

[42] Gancedo JM (1998) Yeast carbon catabolite repression. Microbiology and molecular biology reviews 62: 334–361.

[43] Cubillos FA, Billi E, Zörgö E, Parts L, Fargier P, et al. (2011) Assessing the complex architecture of polygenic traits in diverged yeast populations. Molecular ecology 20: 1401–1413.

[44] Bloom JS, Ehrenreich IM, Loo WT, Lite TLV, Kruglyak L (2013) Finding the sources of missing heritability in a yeast cross. Nature 494: 234–237.

[45] Tong AHY, Evangelista M, Parsons AB, Xu H, Bader GD, et al. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science 294: 2364–2368.

[46] Perlstein EO, Ruderfer DM, Ramachandran G, Haggarty SJ, Kruglyak L, et al. (2006) Revealing complex traits with small molecules and naturally recombinant yeast strains. Chemistry & biology 13: 319–327.

[47] Warringer J, Ericson E, Fernandez L, Nerman O, Blomberg A (2003) High-resolution yeast phenomics resolves different physiological features in the saline response. Proceedings of the national academy of sciences 100: 15724–15729.

[48] Warringer J, Anevski D, Liu B, Blomberg A (2008) Chemogenetic fingerprinting by analysis of cellular growth dynamics. BMC chemical biology 8: 3.

[49] Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, et al. (2002) Systematic screen for human disease genes in yeast. Nature genetics 31: 400–404.

[50] Wei W, McCusker JH, Hyman RW, Jones T, Ning Y, et al. (2007) Genome sequencing and comparative analysis of saccharomyces cerevisiae strain yjm789. Proceedings of the National Academy of Sciences 104: 12825–12830.

[51] Rupp S, Summers E, Lo HJ, Madhani H, Fink G (1999) *MAP* kinase and *cAMP* filamentation signaling pathways converge on the unusually large promoter of the yeast *FLO11* gene. The EMBO journal 18: 1257–1269.

[52] Cullen PJ, Sprague GF (2012) The regulation of filamentous growth in yeast. Genetics 190: 23–49.

[53] Chen RE, Thorner J (2010) Systematic epistasis analysis of the contributions of protein kinase a-and mitogen-activated protein kinase-dependent signaling to nutrient limitation-evoked responses in the yeast saccharomyces cerevisiae. Genetics 185: 855–870.

[54] Kobayashi O, Suda H, Ohtani T, Sone H (1996) Molecular cloning and analysis of the dominant flocculation gene *FLO8* from *Saccharomyces cerevisiae*. Molecular and General Genetics MGG 251: 707–715.

[55] Liu H, Styles CA, Fink GR (1996) Saccharomyces cerevisiae *S288C* has a mutation in *FL08*, a gene required for filamentous growth. Genetics 144: 967–978.

[56] Ran F, Bali M, Michels CA (2008) *Hsp90/Hsp70* chaperone machine regulation of the saccharomyces mal-activator as determined in vivo using noninducible and constitutive mutant alleles. Genetics 179: 331–343.

[57] Daran-Lapujade P, Jansen ML, Daran JM, van Gulik W, de Winde JH, et al. (2004) Role of transcriptional regulation in controlling fluxes in central carbon metabolism of saccharomyces cerevisiae a chemostat culture study. Journal of Biological Chemistry 279: 9125–9138.

[58] Wu L, Ashraf MHN, Facci M, Wang R, Paterson PG, et al. (2004) Dietary approach to attenuate oxidative stress, hypertension, and inflammation in the cardiovascular system. Proceedings of the National Academy of Sciences of the United States of America 101: 7094–7099.

[59] Liao X, Butow RA (1993) *RTG1* and *RTG2*: Two yeast genes required for a novel path of communication from mitochondria to the nucleus. Cell 72: 61–71.

[60] Gerke J, Lorenz K, Ramnarine S, Cohen B (2010) Gene–environment interactions at nucleotide resolution. PLoS genetics 6: e1001144.

[61] Neiman AM (2011) Sporulation in the budding yeast saccharomyces cerevisiae. Genetics 189: 737–765.

[62] Enyenihi AH, Saunders WS (2003) Large-scale functional genomic analysis of sporulation and meiosis in saccharomyces cerevisiae. Genetics 163: 47–54.

[63] Neiman AM (2005) Ascospore formation in the yeast saccharomyces cerevisiae. Microbiology and molecular biology reviews 69: 565–584.

[64] Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, et al. (1998) The transcriptional program of sporulation in budding yeast. Science 282: 699–705.

[65] Primig M, Williams RM, Winzeler EA, Tevzadze GG, Conway AR, et al. (2000) The core meiotic transcriptome in budding yeasts. Nature genetics 26: 415–423.

[66] Ben-Ari G, Zenvirth D, Sherman A, David L, Klutstein M, et al. (2006) Four linked genes participate in controlling sporulation efficiency in budding yeast. PLoS genetics 2: e195.

[67] Deutschbauer AM, Davis RW (2005) Quantitative trait loci mapped to single-nucleotide resolution in yeast. Nature genetics 37: 1333–1340.

[68] Gerke J, Lorenz K, Cohen B (2009) Genetic interactions between transcription factors cause natural variation in yeast. Science 323: 498–501.

[69] Gerke JP, Chen CT, Cohen BA (2006) Natural isolates of saccharomyces cerevisiae display complex genetic variation in sporulation efficiency. Genetics 174: 985–997.

[70] Mortimer RK (2000) Evolution and variation of the yeast (saccharomyces) genome. Genome Research 10: 403–409.

[71] Cubillos FA, Louis EJ, Liti G (2009) Generation of a large set of genetically tractable haploid and diploid saccharomyces strains. FEMS yeast research 9: 1217–1225.

[72] Liti G, Carter DM, Moses AM, Warringer J, Parts L, et al. (2009) Population genomics of domestic and wild yeasts. Nature 458: 337–341.

[73] Sherman F, Fink GR, Hicks JB (1986) Laboratory course manual for methods in yeast genetics. Cold Spring Harbor Laboratory.

[74] Codon A, Gasent-Ramirez J, Benitez T (1995) Factors which affect the frequency of sporulation and tetrad formation in saccharomyces cerevisiae baker's yeasts. Applied and environmental microbiology 61: 630–638.

[75] Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, et al. (2012) Saccharomyces genome database: the genomics resource of budding yeast. Nucleic acids research 40: D700–D705.

[76] Connelly CF, Akey JM (2012) On the prospects of whole-genome association mapping in saccharomyces cerevisiae. Genetics 191: 1345–1353.

[77] Diao L, Chen KC (2012) Local ancestry corrects for population structure in saccharomyces cerevisiae genome-wide association studies. Genetics 192: 1503–1511.

[78] Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

[79] Deutschbauer AM, Williams RM, Chu AM, Davis RW (2002) Parallel phenotypic analysis of sporulation and postgermination growth in saccharomyces cerevisiae. Proceedings of the National Academy of Sciences 99: 15530–15535.

[80] Aasland R, Gibson TJ, Stewart AF (1995) The phd finger: implications for chromatin-mediated transcriptional regulation. Trends in biochemical sciences 20: 56–59.

[81] Pijnappel WP, Schaft D, Roguev A, Shevchenko A, Tekotte H, et al. (2001) The s. cerevisiae set3 complex includes two histone deacetylases, hos2 and hst1, and is a meiotic-specific repressor of the sporulation gene program. Genes & development 15: 2991–3004.

[82] Nickas ME, Schwartz C, Neiman AM (2003) Ady4p and spo74p are components of the meiotic spindle pole body that promote growth of the prospore membrane in saccharomyces cerevisiae. Eukaryotic Cell 2: 431–445.

[83] Neigeborn L, Mitchell AP (1991) The yeast mck1 gene encodes a protein kinase homolog that activates early meiotic gene expression. Genes & development 5: 533–548.

[84] Kallal L, Bhattacharyya M, Grove S, Iannacone R, Pugh T, et al. (1990) Functional analysis of the sporulation-specific spr6 gene of saccharomyces cerevisiae. Current genetics 18: 293–301.

[85] Biss K, Ho KJ, Mikkelson B, Lewis L, Taylor CB (1971) Some unique biologic characteristics of the Masai of East Africa. The New England journal of medicine 284: 694–9.

[86] Ho KJ, Biss K, Mikkelson B, Lewis LA, Taylor CB (1971) The Masai of East Africa: some unique biological characteristics. Archives of pathology 91: 387–410.

[87] Taylor CB, Ho KJ (1971) Studies on the Masai. The American journal of clinical nutrition 24: 1291–3.

[88] Å rhem K (1989) The Cultural Connotations of Milk, Meat, and Blood in the Pastoral Maasai Diet. Anthropos 84: 1–23.

[89] Kromhout D, Menotti A, Bloemberg B, Aravanis C, Blackburn H, et al. (1995) Dietary saturated and transfatty acids and cholesterol and 25-year mortality from coronary heart disease: the seven countries study. Preventive medicine 24: 308–315.

[90] Brotons C, Ribera A, Perich RM, Abrodos D, Magaña P, et al. (1998) Worldwide distribution of blood lipids and lipoproteins in childhood and adolescence: a review study. Atherosclerosis 139: 1–9.

[91] Mann GV, Shaffer RD, Anderson RS, Sandstead HH (1964) Cardiovascular disease in the Masai. Journal of atherosclerosis research 4: 289–312.

[92] Mann GV, Shaffer RD, Rich A (1965) Physical fitness and immunity to heart-disease in Masai. Lancet 2: 1308–10.

[93] Gibney MJ, Burstyn PG (1980) Milk, serum cholesterol, and the Maasai. A hypothesis. Atherosclerosis 35: 339–43.

[94] Johns T, Mahunnah RL, Sanaya P, Chapman L, Ticktin T (1999) Saponins and phenolic content in plant dietary additives of a traditional subsistence community, the Batemi of Ngorongoro District, Tanzania. Journal of ethnopharmacology 66: 1–10.

[95] Coast E (2001) Maasai demography. Ph.D. thesis, University of London, University College London. URL `http://eprints.lse.ac.uk/264/1/Maasai_Demography_PhD.pdf`.

[96] Hollis AC (1910) A Note on the Masai System of Relationship and Other Matters Connected Therewith. The Journal of the Royal Anthropological Institute of Great Britain and Ireland 40: 473–482.

[97] Mattson F, Erickson B, Kligman A (1972) Effect of dietary cholesterol on serum cholesterol in man. The American journal of clinical nutrition 25: 589–594.

[98] Rader DJ, Cohen J, Hobbs HH (2003) Monogenic hypercholesterolemia: new insights in pathogenesis and treatment. The Journal of clinical investigation 111: 1795–803.

[99] Lusis AJ (2000) Atherosclerosis. Nature 407: 233–41.

[100] Lusis AJ, Fogelman AM, Fonarow GC (2004) Genetic basis of atherosclerosis: part II: clinical implications. Circulation 110: 2066–71.

[101] Lusis AJ, Fogelman AM, Fonarow GC (2004) Genetic basis of atherosclerosis: part I: new genes and pathways. Circulation 110: 1868–73.

[102] Wang WYS, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nature reviews Genetics 6: 109–18.

[103] Hardy J, Singleton A (2009) Genomewide association studies and human disease. The New England journal of medicine 360: 1759–68.

[104] Ku CS, Loy EY, Pawitan Y, Chia KS (2010) The pursuit of genome-wide association studies: where are we now? Journal of human genetics 55: 195–206.

[105] Johnson AD, O'Donnell CJ (2009) An open access database of genome-wide association results. BMC medical genetics 10: 6.

[106] Manolio TA (2010) Genomewide association studies and assessment of the risk of disease. The New England journal of medicine 363: 166–76.

[107] Hanotte O, Bradley DG, Ochieng JW, Verjee Y, Hill EW, et al. (2002) African pastoralism: genetic imprints of origins and migrations. Science (New York, NY) 296: 336–9.

[108] Weir BS, Cockerham CC (1984) Estimating f-statistics for the analysis of population structure. evolution : 1358–1370.

[109] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81: 559–575.

[110] Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics 5: e1000529.

[111] Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. Molecular Ecology Resources 9: 1322–1332.

[112] Hagberg A, Swart P, S Chult D (2008) Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Laboratory (LANL).

[113] Kent WJ (2002) Blatthe blast-like alignment tool. Genome research 12: 656–664.

[114] Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. The American Journal of Human Genetics 69: 1–14.

[115] Tang K, Thornton KR, Stoneking M (2007) A new approach for using genome scans to detect recent positive selection in the human genome. PLoS biology 5: e171.

[116] Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, et al. (2009) The genetic structure and history of Africans and African Americans. Science (New York, NY) 324: 1035–44.

[117] Rosenberg NA (2004) Distruct: a program for the graphical display of population structure. Molecular Ecology Notes 4: 137–138.

[118] Fisher E, Weikert C, Klapper M, Lindner I, Möhlig M, et al. (2007) L-FABP T94A is associated with fasting triglycerides and LDL-cholesterol in women. Molecular genetics and metabolism 91: 278–84.

[119] Robitaille J, Brouillette C, Lemieux S, Pérusse L, Gaudet D, et al. (2004) Plasma concentrations of apolipoprotein B are modulated by a gene–diet interaction effect between the LFABP T94A polymorphism and dietary fat intake in French-Canadian men. Molecular genetics and metabolism 82: 296–303.

[120] Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, et al. (2002) Identification of a variant associated with adult-type hypolactasia. Nature genetics 30: 233–7.

[121] Ingram CJ, Elamin MF, Mulcare CA, Weale ME, Tarekegn A, et al. (2007) A novel polymorphism associated with lactose tolerance in africa: multiple causes for lactase persistence? Human genetics 120: 779–788.

[122] Enattah NS, Jensen TG, Nielsen M, Lewinski R, Kuokkanen M, et al. (2008) Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. The American Journal of Human Genetics 82: 57–72.

[123] Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, et al. (2004) Genetic signatures of strong recent positive selection at the lactase gene. American journal of human genetics 74: 1111–20.

[124] Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466: 707–13.

[125] Silander K, Alanne M, Kristiansson K, Saarela O, Ripatti S, et al. (2008) Gender differences in genetic risk profiles for cardiovascular disease. PloS one 3: e3615.

[126] Ma L, Yang J, Runesha HB, Tanaka T, Ferrucci L, et al. (2010) Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the Framingham heart study data. BMC medical genetics 11: 55.

[127] Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research 33: D501–4.

[128] Newberry EP, Xie Y, Kennedy SM, Luo J, Davidson NO (2006) Protection against Western diet-induced obesity and hepatic steatosis in liver fatty acid-binding protein knockout mice. Hepatology (Baltimore, Md) 44: 1191–205.

[129] Newberry EP, Kennedy SM, Xie Y, Luo J, Davidson NO (2009) Diet-induced alterations in intestinal and extrahepatic lipid metabolism in liver fatty acid binding protein knockout mice. Molecular and cellular biochemistry 326: 79–86.

[130] Thompson EE, Kuttab-Boulos H, Witonsky D, Yang L, Roe BA, et al. (2004) CYP3A variation and the evolution of salt-sensitivity variants. American journal of human genetics 75: 1059–69.

[131] Chen X, Wang H, Zhou G, Zhang X, Dong X, et al. (2009) Molecular population genetics of human CYP3A locus: signatures of positive selection and implications for evolutionary environmental medicine. Environmental health perspectives 117: 1541–8.

[132] Kuehl P, Zhang J, Lin Y, Lamba J, Assem M, et al. (2001) Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. Nature genetics 27: 383–91.

[133] Kivistö KT, Niemi M, Schaeffeler E, Pitkälä K, Tilvis R, et al. (2004) Lipid-lowering response to statins is affected by CYP3A5 polymorphism. Pharmacogenetics 14: 523–5.

[134] Lagor WR, Brown RJ, Toh SA, Millar JS, Fuki IV, et al. (2009) Overexpression of apolipoprotein F reduces HDL cholesterol levels in vivo. Arteriosclerosis, thrombosis, and vascular biology 29: 40–6.

[135] Lardizabal KD, Mai JT, Wagner NW, Wyrick A, Voelker T, et al. (2001) DGAT2 is a new diacylglycerol acyltransferase gene family: purification, cloning, and expression in insect cells of two polypeptides from Mortierella ramanniana with diacylglycerol acyltransferase activity. The Journal of biological chemistry 276: 38862–9.

[136] Cases S, Stone SJ, Zhou P, Yen E, Tow B, et al. (2001) Cloning of DGAT2, a second mammalian diacylglycerol acyltransferase, and related family members. The Journal of biological chemistry 276: 38870–6.

[137] Yu XX, Murray SF, Pandey SK, Booten SL, Bao D, et al. (2005) Antisense oligonucleotide reduction of DGAT2 expression improves hepatic steatosis and hyperlipidemia in obese mice. Hepatology (Baltimore, Md) 42: 362–71.

[138] Kantartzis K, Machicao F, Machann J, Schick F, Fritsche A, et al. (2009) The DGAT2 gene is a candidate for the dissociation between fatty liver and insulin resistance in humans. Clinical science (London, England : 1979) 116: 531–7.

[139] Best LG, North KE, Li X, Palmieri V, Umans JG, et al. (2008) Linkage study of fibrinogen levels: the Strong Heart Family Study. BMC medical genetics 9: 77.

[140] Danesh J, Lewington S, Thompson SG, Lowe GDO, Collins R, et al. (2005) Plasma fibrinogen level and the risk of major cardiovascular diseases and non-vascular mortality: an individual participant meta-analysis. JAMA : the journal of the American Medical Association 294: 1799–809.

[141] Dastani Z, Pajukanta P, Marcil M, Rudzicz N, Ruel I, et al. (2010) Fine mapping and association studies of a high-density lipoprotein cholesterol linkage region on chromosome 16 in French-Canadian subjects. European journal of human genetics : EJHG 18: 342–7.

[142] Neal EG, Chaffe H, Schwartz RH, Lawson MS, Edwards N, et al. (2008) The ketogenic diet for the treatment of childhood epilepsy: a randomised controlled trial. Lancet neurology 7: 500–6.

[143] Shorvon SD, Perucca E, Fish D, Dodson WE, editors (2004) The Treatment of Epilepsy. Wiley-Blackwell, 2 edition, 913 pp. doi:10.1002/9780470752463. URL `http://doi.wiley.com/10.1002/9780470752463`.

[144] Kang HC, Chung DE, Kim DW, Kim HD (2004) Early- and late-onset complications of the ketogenic diet for intractable epilepsy. Epilepsia 45: 1116–23.

[145] Tsallis C (1988) Possible generalization of boltzmann-gibbs statistics. Journal of statistical physics 52: 479–487.

[146] Teshima K, Innan H (2009) mbs: modifying hudson's ms software to generate samples of dna sequences with a biallelic site under selection. BMC bioinformatics 10: 166.