

Inference of natural selection on quantitative traits



Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Nico Riedel

aus Mettmann

Köln 2016

Berichterstatter: Prof. Dr. Johannes Berg

Prof. Dr. Michael Lässig

Tag der mündlichen Prüfung: 28.06.2016

Abstract

The concept of evolution, which was introduced by Charles Darwin in 1859, and also its mathematical description by the theory of population genetics are well-established. Population genetics describes the development of a population under the influence of mutations, creating new genetic variants, and natural selection, increasing the frequency of favorable phenotypes. Yet, the experimental verification of selective forces acting on species has proven difficult. With new experimental techniques that have been established in the field of quantitative genetics, like the sequencing of DNA or measurements of gene expression levels, it has become possible to find signs of natural selection on the level of the genome.

In this thesis, I develop a statistical test based on population genetics theory that can infer lineage-specific differences in selection between multiple lines of a species. The test employs data from quantitative trait experiments and uses a log-likelihood scoring to quantify the evidence for different selective scenarios. I show that the use of multiple lines increases both the power and the scope of selection inference. Extensive numerical simulations demonstrate that the test can distinguish selection from neutral evolution as well as different scenarios of lineage-specific evolution. The principle of maximum entropy is used to derive a modified version of the selection test that accounts for the multiple testing problem arising when many traits are tested for selection at the same time. The developed test is applied to two published plant datasets and a published dataset of gene expression levels in three yeast lines. In all cases, I find signs of selection not seen with a two-line test. For the yeast dataset I find pervasive adaptation linked to stress resistance both on the level of individual genes as well as for larger gene modules consisting of several genes, like protein complexes and pathways. This adaptation signal is also reflected on the protein levels.

Kurzzusammenfassung

Sowohl das Konzept der Evolution, welches 1859 von Charles Darwin eingeführt wurde, als auch die mathematische Beschreibung durch die Populationsgenetik sind seit langem etabliert. Die Populationsgenetik beschreibt die Entwicklung einer Population unter dem Einfluss von Mutationen, welche neue genetische Varianten erzeugen, und der natürlichen Selektion, welche die Häufigkeit der günstigen Phenotypen erhöht. Jedoch hat es sich als schwierig erwiesen, diese selektiven Kräfte experimentell nachzuweisen. Neue experimentelle Techniken die sich im Feld der quantitativen Genetik etabliert haben, wie das Sequenzieren der DNA oder der Messung von Genexpressionsleveln, haben es ermöglicht, Spuren der natürlichen Selektion auf dem Level des Genoms nachzuweisen.

In dieser Dissertation entwickle ich einen statistischen Test welcher auf der Theorie der Populationsgenetik basiert und mit welchem man linienspezifische Unterschiede in der Selektion zwischen verschiedenen Linien einer Spezies nachweisen kann. Dieser Test verwendet Resultate von Experimenten über quantitative Merkmale und verwendet einen Log-Likelihood-Quotienten um die Evidenz für verschiedene selektive Szenarien zu quantifizieren. Ich weise nach, dass die Verwendung von mehreren Linien sowohl die Leistungsfähigkeit als auch den Anwendungsbereich des Selektionstests vergrößert. Umfangreiche numerische Simulationen zeigen, dass der Test zwischen Selektion und neutraler Evolution als auch zwischen verschiedenen linienspezifischen Evolutionsszenarien unterscheiden kann. Das Prinzip der maximalen Entropie wird verwendet um eine modifizierte Version des Selektionstests herzuleiten, welche das multiple Testproblem berücksichtigt, welches auftritt, wenn viele Merkmale gleichzeitig auf Selektion getestet werden. Der entwickelte Test wird auf zwei publizierte Pflanzendatensätze sowie einen publizierten Datensatz zu Genexpressionsleveln in drei Hefelinien angewandt. In allen Fällen finde ich Hinweise für Selektion, welche nicht durch einen Zwei-Linien-Test entdeckt werden. Für den Hefedatensatz finde ich weit verbreitete selektive Anpassung die mit Stressresistenz verbunden ist, sowohl auf dem Level einzelner Gene als auch für größere Genmodule, die aus mehreren Genen bestehen, wie zum Beispiel Proteinkomplexen. Diese Adaption kann auch für die Proteinlevel nachgewiesen werden.

Contents

1	Introduction	9
2	Genetics and evolution	11
2.1	The Genome	12
2.1.1	DNA	12
2.1.2	Genes and proteins	12
2.2	Quantitative traits and QTL mapping	13
2.2.1	Quantitative traits	13
2.2.2	QTL mapping	13
2.3	Population genetics and evolution	16
2.4	Connection to statistical physics	18
3	Multiple-line selection model	21
3.1	Idea of the selection test	21
3.2	Derivation of the model	22
3.3	Inference of selection and log-likelihood scoring of evolutionary scenarios	28
3.4	Advantages of multiple-line testing	31
3.4.1	Increase in the number of detected loci	31
3.4.2	Two vs. three lines at a constant number of crosses	35
3.5	Statistical power of the selection test	38
3.6	Robustness of the model assumptions	40
3.6.1	Epistasis and multiple segregating loci	40
3.6.2	Evolutionary timescales	45
3.7	Comparison to other selection tests	46
4	Multiple Testing and Maximum Entropy	49
4.1	Classical multiple testing corrections	50
4.1.1	Holm-Bonferroni Correction	50
4.1.2	Benjamini–Hochberg procedure	50
4.2	Conditioning on the trait difference	51

4.2.1	Pedagogical example	52
4.2.2	Derivation of the ascertained neutral scenario	55
5	Short evolutionary times	63
5.1	Two lines	63
5.2	Three lines	66
5.3	Statistical power of the short-time test	69
6	Selection on Plant Quantitative Traits	71
6.1	Maize photoperiodic response traits	71
6.2	<i>Mimulus</i> floral traits	74
7	Gene expression evolution in the yeast <i>S. pombe</i>	79
7.1	Yeast dataset	80
7.1.1	Available data	80
7.1.2	State configurations	81
7.2	Selection on individual genes	84
7.2.1	State configurations per gene	84
7.2.2	Selection analysis	86
7.2.3	Multiple testing correction	89
7.2.4	Adjusting for the high QTL false discovery rate	89
7.2.5	Lines with highest trait divergence	91
7.2.6	Pleiotropy	93
7.3	Selection on gene modules	96
7.3.1	State configurations per gene module	96
7.3.2	Selection analysis	97
7.3.3	Stress-specific gene modules under selection	99
7.3.4	Biological functions of gene modules under selection	101
7.4	Protein level changes	107
7.5	Possible use of protein QTL data	109
7.6	Comparison to other selection tests	110
7.6.1	Orr test	110
7.6.2	Test for genome-wide level of selection	111
8	Conclusions	113
	Appendix	114

Chapter 1

Introduction

In the past decades the field of quantitative genetics has experienced a tremendous progress. Quantitative genetics deals with traits of an organism that vary continuously (called quantitative traits), like the height of a plant or the expression level of a gene, and their underlying genetic basis. The development of the method of QTL analysis allowed to map observed changes in the phenotype, like changes in morphology, to the genome. This method helped to uncover the genetic basis of quantitative traits (Mackay 2004; Mackay et al. 2009). In numerous organisms like crop, cattle, or yeast genetic loci underlying many traits have been determined (Marullo et al. 2007; Goddard and Hayes 2009). In recent years, new, advanced experimental techniques have allowed the mapping of all gene expression or protein levels of an organism simultaneously, taking quantitative trait analysis from a level of individual macroscopic phenotypes to the level of changes in genes underlying these phenotypes (Hoheisel 2006; Bantscheff et al. 2007). Furthermore, crossing multiple lines has allowed to study a greater genetic diversity underlying many quantitative traits. Yet, many challenges still persist. Identifying the genes underlying quantitative trait loci has proven difficult and was only possible with further experimental effort for individual cases (Fanara et al. 2002; De Luca et al. 2003) and interactions between different genetic loci have also complicated the picture of the genetic architecture underlying quantitative traits.

On the theoretical side, the field of population genetics has been long established, building the theoretical foundation for the stochastic evolution of populations under the influence of natural selection, mutations and random fluctuations in the composition of the population (called genetic drift) (Hartl et al. 1997). The theory of population genetics uses stochastic differential equations to describe the time evolution of a population which allows to gain insights into the dynamics of the evolutionary process. Established results describe the probability for new, beneficial mutations to spread in a population as well as the time it takes to fixation. There is also an close connection between the fields of population genetics and statistical physics (de Vladar and Bar-

ton 2011b). The time evolution of allele frequencies in a population can be described using stochastic differential equations and the steady state of a population, balancing the forces of selection (increasing the fitness) and random genetic drift (decreasing the fitness), can be determined using methods of statistical mechanics.

In this thesis, I combine aspects of both fields to study the evolutionary history of quantitative traits. Based on established results from population genetics theory I construct a model for the evolution of quantitative traits. This model quantifies the strength of evidence for selection acting on a particular trait. I define a log-likelihood score that weights different selective scenarios against each other. The model is defined for an arbitrary number of lines and I show that using multiple lines increases both the power and the scope of selection inference. First, a test based on three or more lines detects selection with strongly increased statistical significance. Second, a multiple-line test allows to distinguish different lineage-specific selection scenarios, unlike in the case of two lines. Extensive numerical simulations show the ability of the test to distinguish neutral evolution from selection as well as different scenarios of lineage-specific selection. I show explicitly how the sensitivity of the test depends on the number of lines. Different violations of the model assumptions, like epistasis or shorter evolutionary timescales, are investigated, showing the overall robustness of the model. In addition to the case of long evolutionary times considered in the original model, I also give a solution for short evolutionary times. The multiple testing problem that arises when many traits are tested for selection is considered. Using the principle of maximum entropy, I derive a modified version of the developed selection test that accounts for a possible ascertainment bias.

I apply the multiple-line test to QTL data on floral character traits in plant species of the *Mimulus* genus and on photoperiodic traits in different maize lines, where signatures of lineage-specific selection are found that are not seen in a two-line test. Finally, I use a dataset of expression QTL (eQTL) for three lines of the yeast species *Schizosaccharomyces pombe* to study the adaptation of oxidative stress response both for individual genes as well as for gene modules, like protein complexes or pathways. I consistently find high levels of selection on both genes and gene modules. The analysis of gene modules exclusively under selection in the stress condition uncovers the adaptation of stress response on the level of individual genes and protein complexes. An analysis of the protein levels connected to the expression levels shows that the selection on the transcriptional level is also reflected in translational changes.

Chapter 2

Genetics and evolution

The field of genetics has undergone a rapid development during the last decades. Building upon the success of the discovery of DNA as the carrier of the genetic information necessary for the functioning and reproduction of an organism, the understanding of the structure and functionality of the genome has progressed. Many new experimental techniques changed genetics into a quantitative field (Hoheisel 2006; Bantscheff et al. 2007). Nowadays it is possible to measure the genetic sequence of numerous organisms, including humans (Metzker 2010). The field of quantitative genetics contains a rich variety of interesting mathematical problems and challenges connected to statistical physics. For example, the evolution of species is a stochastic system. Many processes inside organisms like the activity and regulation of genes and its time dynamics is a many-body problem with strong interaction between the individual elements of the system. Numerous inference problems appear in this context, where the result of a process is observed (like in the evolution of species), while the underlying dynamics are unknown. The contribution of statistical physics is to build appropriate mathematical descriptions of the underlying processes which for example allow to infer model parameters from experimental observations.

In this chapter I introduce the basic genetic concepts necessary to understand the background and motivation for the selection test developed in this thesis. I start with a short introduction to DNA and the genome. The notion of quantitative traits and quantitative trait loci (QTL) is explained, which are key concepts used for the selection test. I end the introductory chapter with the topics of evolution and population genetics, which puts the evolution of a population in a quantitative context described by stochastic differential equations.

2.1 The Genome

2.1.1 DNA

The DNA (deoxyribonucleic acid) stores all the inheritable information of an organism. It is a linear molecule consisting of the four nucleotides cytosine (C), guanine (G), adenine (A), or thymine (T). It is structured as a double helix of two complementary strands of nucleotides with pairwise bonds between the nucleotides C and G as well as A and T. During reproduction the whole DNA of an organism is duplicated. The genome denotes the sum of all DNA present in an organism, which is often distributed on different chromosomes, that are separate DNA molecules. The sequence information contained in the genome is called the genotype of an organism and can be measured through DNA sequencing (Metzker 2010).

2.1.2 Genes and proteins

Genes are the central elements of the genome. Genes serve as the blueprint for proteins, which are biomolecules that perform a vast number of functions in an organism. The coding region of a gene is marked by a characteristic starting sequence, also called promoter, and is read off by a protein called polymerase. In the step called transcription, the polymerase synthesizes a single stranded RNA molecule with the complimentary nucleotide sequence. This messenger RNA (mRNA) is in turn read off by the ribosome, where three nucleotides are read off at a time, forming a codon. Each codon is assigned an amino acid (the smallest components of a protein), where different codons can map to the same amino acid. This map from nucleotide codons to amino acids is called the genetic code and is almost universal among all life forms (Crick et al. 1961). All codons are read off sequentially to produce a sequence of amino acids, which form the protein sequence of the underlying gene. This step is called translation. In summary, the conversion of genetic information starts with the DNA sequence of a gene and is mediated via the mRNA to form a protein as a final product. This process is also known as the central dogma of molecular biology (Crick et al. 1970). A gene can exist in different variants which are called alleles.

Not all genes are needed in an organism at all times or in all tissues. Therefore, a complex gene regulatory system exists that controls the gene expression (gene expression denotes the amount of mRNA produced from a gene in the step of transcription). An example for important elements of gene regulation are transcription factor binding sites, which are typically located within the promoter sequence close to the starting point of a gene. These sites can bind proteins, that are called transcription factors, which can modulate the binding probability of the polymerase which in turn modulates the gene expression levels. Many genes are coding for proteins that can alter the

expression of other genes. Thus, all genes are part of a complex regulatory network with many interactions between genes.

2.2 Quantitative traits and QTL mapping

2.2.1 Quantitative traits

While the genome provides all genetic information - the genotype - of an organism, the phenotype of an organism is the set of all observable characteristics or traits. These can be morphological traits like size or color but also more microscopic traits like the expression level of a gene or a protein level. The phenotype of an organism is a result of two main factors, the genotype and environmental factors. While the genotype provides the basic information for all features of an organism, environmental factors like nutrition, temperature, or the surrounding ecosystem can influence the phenotype as well. Understanding the genotype-phenotype map, which captures the relationship between the genotypes of an organism and its observed phenotypes, is one of the central problems of genetics (Doerge 2002; Mackay et al. 2009). While the complete genotype-phenotype map is still out of reach, the genetic basis of individual phenotypes, like e.g. yield traits in maize (Xing and Zhang 2010), has already been uncovered.

Some traits have a very simple genetic basis, with only a single genetic locus affecting the trait and with only a few possible characteristics, see Figure 2.1 (A). These traits are called Mendelian traits, as the Mendelian inheritance patterns can directly be observed. An example for this is the flower color of pea plants, as originally discovered by Mendel (Mendel 1866).

In contrast to this, most traits have a complex genetic basis with many genetic regions influencing the trait (Mackay et al. 2009). The combined influence of these genetic loci leads to an effectively continuous scale of trait characteristics, see Figure 2.1 (B). These traits are called quantitative traits. Examples for quantitative traits are flower size (Chen 2009), leaf number (Coles et al. 2010), or bristle number in flies (Mackay and Lyman 2005). Any genetic region that has an influence on a quantitative trait is called a quantitative trait locus (QTL). QTL can be genes linked to the phenotype or regulatory elements that alter the expression of these genes. Often, a QTL might be caused by allelic variants such as single nucleotide polymorphisms (SNPs), where only changes in individual nucleotides occur, in coding as well as non-coding regions (Stam and Laurie 1996; Harbison et al. 2004; Jordan et al. 2006; Zheng et al. 2010).

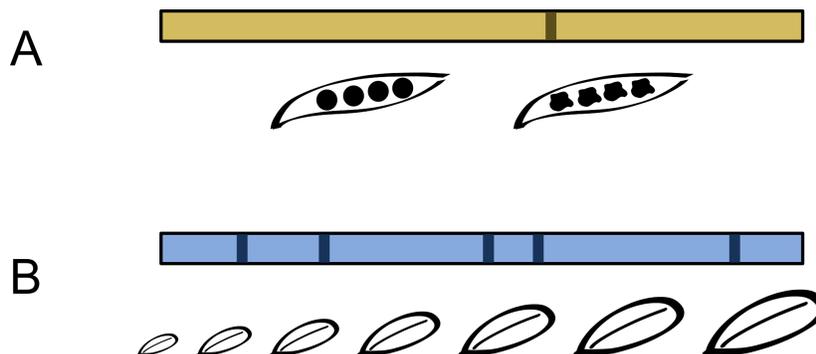


Figure 2.1: **Mendelian traits and quantitative traits.** (A) Mendelian traits are only governed by a single genetic locus (marked by the dark bar) in the genome and can only take on few distinct and clearly defined characteristics. (B) Quantitative traits are affected by many genetic loci. Taken together the effects of these loci generate an approximately continuous spectrum of possible trait values.

2.2.2 QTL mapping

Understanding the complex genetic basis of quantitative traits is a key step towards the genotype-phenotype map and is important in many different contexts like the understanding of complex diseases in humans (Plomin et al. 2009). QTL analysis allows to identify the QTL underlying a quantitative trait by linking phenotypic variation to genetic markers. This is done by crossing individuals of different strains (also called lines) of a species. These might be different breeding lines for cattle or crops (Goddard and Hayes 2009; Xing and Zhang 2010) or different strains of wild yeast isolates that were collected in different environments (Marullo et al. 2007; Gerke et al. 2009). In the crossed individuals the two genomes of lines 1 and 2 recombine (which happens in the second offspring generation or F_2 generation for diploid organisms that have a duplicate chromosome set) and the alleles of the lines get reshuffled. In these crosses the genome is a mixture of DNA stretches originating from line 1 and stretches originating from line 2 (see Figure 2.2 (B)).

For the crosses the trait values T of the quantitative trait (e.g. plant height measured in cm) as well as the genotype in form of genetic markers are measured. The goal of QTL mapping is to find genetic markers which are correlated with the trait value, i.e. to find markers where the trait value is on average higher when the allele of that marker is inherited from line 1, compared to the crosses with the marker allele from line 2. The genetic markers need to differ between the lines and to be evenly distributed across the genome with a sufficient density. These markers can be for example single nucleotide polymorphisms (SNPs) (Wicks et al. 2001) or microsatellites (repetitive DNA elements) (Somers et al. 2004). With QTL mapping the QTL cannot

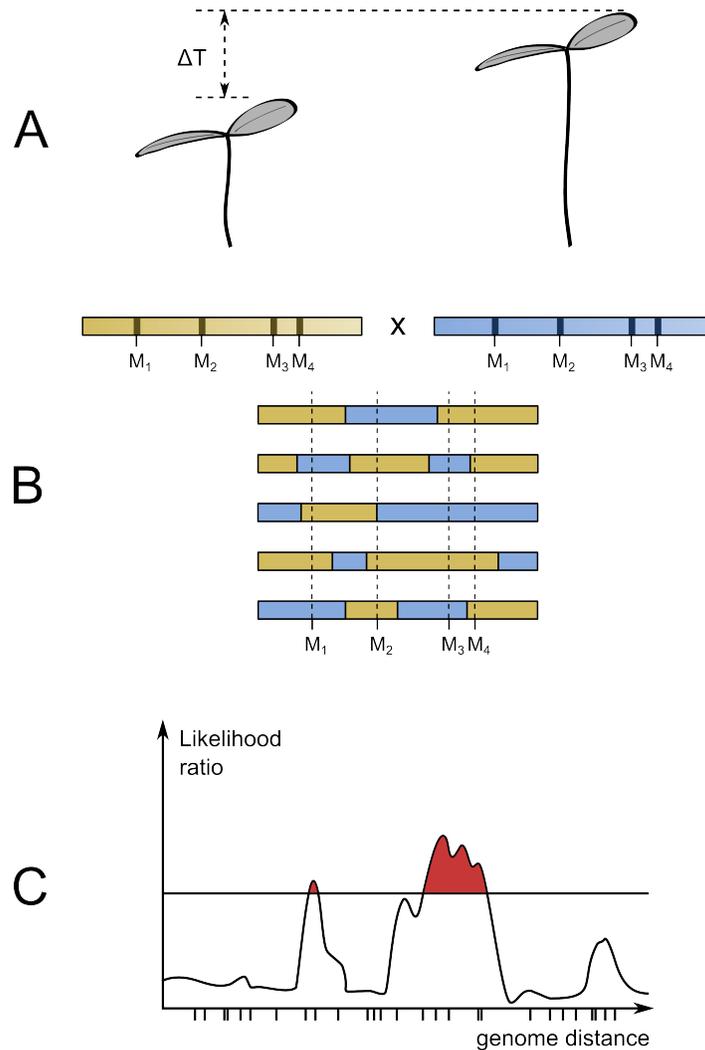


Figure 2.2: **The principle of QTL mapping.** (A) The goal of a QTL mapping experiment is to determine the genetic basis underlying trait differences ΔT between different lines of a species. (B) To achieve this goal, individuals of the lines are crossed. The genome of the offsprings (the F_2 generation for diploid organisms) is a mixture of the genomes of the parental lines created by recombination events between the genomes. A certain number of markers ($M_1 \dots M_4$) spread over the entire genome is measured in the offsprings to determine from which line they originate. (C) QTL mapping algorithms are used to determine which markers are correlated with the trait. If the trait value of the crosses is on average higher when a marker originates from line 1, it is likely that a QTL close to the marker is affecting the trait. This likelihood is quantified by the QTL mapping algorithm. Markers for which the likelihood is above a significance threshold (red areas) are in linkage with a QTL. Since the QTL cannot be measured directly but only by linkage to nearby markers, the QTL mapping does not reveal the exact gene or mutation causing the QTL, but allows to infer the number and effects of QTL affecting a trait.

be identified directly. Instead, the markers are in genetic linkage with nearby QTL due to the rare frequency of recombination events, as the whole genetic region around a marker gets inherited from the same line (Lynch et al. 1998). Thus, QTL mapping only identifies markers that are close to a QTL, but it does not unravel the nature of the QTL itself. As another restriction of QTL mapping only the genetic diversity present between the lines can be uncovered. Since all QTL that have the same allele in both lines also have that allele in all crosses, no difference of the trait effect can be observed between the lines for these QTL.

QTL mapping is a computationally nontrivial task. Many effects like interactions between QTL (called epistasis) (Wang et al. 1999), multiple testing problems (arising due to the high number of possible trait-marker pairs) (Doerge and Churchill 1996), or a limited number of recombination events lead to complications in the analysis. There are many different QTL mapping algorithms employing different techniques like the interval mapping (Lander and Botstein 1989), composite interval mapping (Zeng 1994), or multiple trait mapping (Jiang and Zeng 1995; Kao et al. 1999). Up to today, quantitative traits of many organisms have been mapped, ranging from bristle numbers or wing shape in *Drosophila* to yield traits in rice (Dilda and Mackay 2002; Mackay 2004; Mezey et al. 2005; Mackay and Lyman 2005; Bernier et al. 2007; Flint and Mackay 2009; Xing and Zhang 2010). But only in few cases and with additional experimental effort it was possible to fine-map individual QTL, identifying the gene or even the nucleotide change responsible for the QTL (Pasyukova et al. 2000; Fanara et al. 2002; De Luca et al. 2003; Moehring and Mackay 2004; Harbison et al. 2004; Jordan et al. 2006). In recent years, QTL analysis has also been expanded onto multiple-line crosses, where crosses between all possible pairs of the lines or backcrosses to a common line are performed. It has been shown that utilizing information from several lines drastically increases the power and accuracy of QTL identification (Rebai and Goffinet 2000; Steinhoff et al. 2011) and the genetic variability that can be accessed (Blanc et al. 2006).

2.3 Population genetics and evolution

Evolution describes the change of the genetic composition of a population over time. These changes can occur on the species level, with the creation of new species, or on the molecular levels, with e.g. changes in gene expression levels. Random genetic mutations lead to a genetic diversity in a population by creating new genetic variants. Natural selection acts on the level of different genetic variants of a population. Some of the genetic variants might have favorable phenotypes that have a greater reproductive success than other phenotypes. Since the individuals with favorable phenotypes accumulate a larger number of offsprings over the generations, this often leads to the

prevalence of the favorable phenotype.

The fitness of an organism is the measure of its reproductive success and is defined as the average number of offsprings of an individual (Wrightian fitness; for subtleties between slightly different definitions of fitness see also (Orr 2009)). If the fitness of all variants in a population is the same, the population is evolving neutrally. In this case, new mutants can become prevalent in the population without any effect on the reproductive success, which is called random genetic drift.

Population genetics is the mathematical description of the evolutionary process at the population level (Hartl et al. 1997). Central to population genetics is the evolution of a population of N individuals under the effect of mutations, selection and reproduction. The evolution of a population is a stochastic process that can be described by a stochastic differential equation. While selection is a directed process, increasing the fraction of the fittest individual in the population, reproductive fluctuations lead to undirected changes in the composition of the population. These stochastic fluctuations are larger for smaller population sizes and tend to zero for very large populations, leading to an effectively deterministic evolution of allele frequencies in the population. Mutations introduce new genotypes in the population that can have different fitness values than the ancestral genotype.

In the simplest case, a population consisting of two genotypes a and b with (Malthusian) fitness F_a and F_b , respectively, evolves under the action of selection and reproductive fluctuations as (Lässig 2007)

$$\frac{d}{dt}N_{a/b}(t) = F_{a/b}N_{a/b}(t) + \chi_{a/b}(t), \quad (2.1)$$

where the population size $N_{a/b}(t)$ of individuals with genotype a or b , respectively, evolves according to a simple exponential growth law with a growth rate $F_{a/b}$. $\chi_{a/b}(t)$ is a noise term that describes the reproductive fluctuations in the population. It is defined as a Gaussian random variable with mean $\langle \chi_{a/b}(t) \rangle = 0$ and variance $\langle \chi_a(t)\chi_b(t) \rangle = N_a(t)\delta(t-t')\delta_{a,b}$ that describes uncorrelated white noise. The evolution of the population fraction $x(t) = N_a/(N_a + N_b)$ can be written in term of a Fokker-Planck equation that captures the time evolution of the probability distribution of the $x(t)$:

$$\frac{\partial}{\partial t}P(x, t) = \frac{1}{2N} \frac{\partial^2}{\partial x^2}x(1-x)P(x, t) - \Delta F_{a,b}(t) \frac{\partial}{\partial x}x(1-x)P(x, t), \quad (2.2)$$

where N is the total population size and $\Delta F_{a,b}(t) = F_a(t) - F_b(t)$ (Kimura 1962; Lässig 2007). Without mutations introducing new genotypes, the population eventually reaches the fix points $x = 1$ or $x = 0$. At $x = 1$ the whole population is monomorphic and only consists of individuals of genotype a , called fixation of the genotype, while at $x = 0$ the genotype a is lost.

In this context there are many interesting questions arising: What is the probability of fixation of a newly introduced mutation? How does this fixation probability depend on the fitness of the mutant? What is the distribution of fitness values of new mutants? What is the typical time to fixation? How do different mutations that arise at the same time interact with each other?

The fitness of new mutants depends on the fitness landscape the population is living in. A fitness landscape is a theoretical construct that maps every genotype onto a fitness value, where a population tends towards genotypes with the highest fitness ('fitness peaks') (Orr 2005). As the genotype typically consists of many factors the fitness landscape is often very high-dimensional. A single mutation can have vastly different effects on fitness depending on the type of the fitness landscape. A fitness landscape can be smooth, with the fitness contributions of different parts of the genotype adding up linearly. Otherwise there can be interactions between different alleles, called epistasis, where for example a certain combination of alleles that individually have a positive effect on fitness, leads to a decrease in fitness. This can make the fitness landscape more 'rugged', create multiple peaks, and make it more difficult for a population to reach the global fitness optimum (Whitlock et al. 1995).

When new mutations are allowed in the model, the fix points where the population becomes monomorphic are not the end of the dynamics. Instead, new mutations arise from time to time and either get fixed or are lost. For this, the weak mutation limit is typically assumed where the time for a new mutation to arise is much longer than the time to fixation. Otherwise multiple competing mutations can arise, leading to a more complicated dynamics which is known under the name of clonal interference (Gerrish and Lenski 1998). For a new mutation to become fixed in the population, it has to overcome the random reproductive fluctuations. When only a single individual with the new mutation is present at the beginning, it can easily go extinct even when the mutant has a higher fitness than the rest of the population. It can be shown that the substitution rate $u_{a \rightarrow b}$ that describes the fixation of genotype b (after the genotype arises via a mutation) in a population with genotype a is (Kimura 1962)

$$u_{a \rightarrow b} = N\mu_{a \rightarrow b} \frac{1 - \exp(-2\Delta F_{ab})}{1 - \exp(-2N\Delta F_{ab})}, \quad (2.3)$$

where $\mu_{a \rightarrow b}$ is the mutation rate for creating genotype b from genotype a . This result will be used later for the population genetics model underlying my selection test.

2.4 Connection to statistical physics

Even though statistical physics does not directly relate to the topics of genetics and evolution many concepts of statistical physics can be transferred to these fields of research

and can help to deepen the mathematical understanding in many cases. As mentioned before, population genetics deals with the stochastic evolution of a population, well described by stochastic differential equations. Many-body problems arise at numerous levels of genetics and evolution, as in the context of gene regulatory networks, with many genes that alter the expression of other genes, or at the level of species, with different species competing for limited resources in an environment. Evolution takes place on timescales that often span millions or billions of years, but only the species and genomes present today can be observed. This leads to many inference problems: given the genomes of today's species, what is the evolutionary history of the lines and how are the species related? One of the great success stories was the inference of the famous tree of life, which describes the evolutionary relations of all species, from the comparison of their genetic sequences (Woese and Fox 1977; Delsuc et al. 2005).

In the case of population genetics the connection to statistical physics goes even deeper. The system of an evolving population of many individuals can readily be mapped onto a thermodynamic system. Analogously to macroscopic observables that arise from the underlying microscopic dynamics in a thermodynamic system, the time evolution of averaged quantities (like allele frequencies or trait values) are considered for the large number of individuals of an evolving population (de Vladar and Barton 2011a; de Vladar and Barton 2011b). In population genetics the population size determines the magnitude of fluctuations in the population composition and is playing the role of inverse temperature. In addition, the fitness plays the role of the (negative) energy of the system, as the population tends to the state of maximum fitness and the steady state distribution of allele frequencies in certain evolutionary models has been found to follow a Boltzmann distribution (Sella and Hirsh 2005). A free fitness can be defined which increases during the evolutionary dynamics and that reaches its maximum in the steady state of the system (Iwasa 1988; Sella and Hirsh 2005; Barton and de Vladar 2009; Barton and Coe 2009). As for the free energy in thermodynamics the free fitness balances two concepts: the maximization of fitness (corresponding to the minimization of energy) and the maximization of entropy. Also the non-equilibrium aspects of adaptation to changing environments with a constant fitness flux in the steady state (Mustonen and Lässig 2009; Mustonen and Lässig 2010) as well as the connection to disordered systems (Mustonen and Lässig 2008) have been explored. Overall, this shows the close connection between population genetics and thermodynamic systems.

In this thesis, I combine the two fields of quantitative trait analysis and population genetics. I develop a novel statistical framework to test different evolutionary hypotheses for multiple QTL lines. This test is based on a population genetics model describing the evolution of traits under different selective scenarios. This test allows to infer the evolutionary history of traits and can determine lineage-specific differences in the selection pressure between different lines.

Chapter 3

Multiple-line selection model

As explained in the previous chapter, QTL analysis can be used to obtain information on the number and effect sizes of QTL affecting a trait. The aim of this thesis is to quantify the evolutionary forces acting on a trait using information from QTL analysis. In this chapter, I develop a selection test that can infer lineage-specific selection on quantitative traits using multiple QTL lines. The model used for the selection test is derived using established population genetics theory. A log-likelihood score is introduced that tests different selective and neutral scenarios against each other. I provide arguments for the superiority of a multiple-line selection test as compared to a two-line test and perform numerical simulations to show its feasibility. Finally, I analyze the robustness of the selection test when several of the model assumptions are violated.

3.1 Idea of the selection test

When a trait diverges between two lines there are two basic evolutionary mechanisms that could have affected the change in trait values: First, natural selection that acts as a directed force, increasing or decreasing the trait value in one (or potentially all) of the lines. Second, when an allele is established in a population only due to random fluctuations in the composition of the population, this is called genetic drift (Lande 1976). For example two different plant lines grow in two geographically separated regions where individuals of one of the lines have on average a larger size. This can reflect either an adaptation of one (or both) of the lines to the different habitats, or random genetic changes that accumulated after the reproductive isolation following the geographic separation of the lines. The idea to distinguish these two cases, which was developed by Orr (Orr 1998), is the following: Under selection consistent changes of the QTL in one line would be expected, i.e. the QTL at different positions in the genome would all affect the trait value in the same way (e.g. 9 out of 10 QTL would increase the trait value in line 1). Under genetic drift a more even distribution of alleles

that increase or decrease the trait value would be expected. Of course, an imbalance of alleles can also be created in this case, yet one would typically expect less extreme imbalances as in the selective case.

Here, I use population genetics theory to develop a model that quantifies the evidence for natural selection versus genetic drift acting on a quantitative trait. I construct a population genetics model of QTL evolving in n haploid populations in the weak-mutation regime with full recombination. Trait and fitness are linear functions of the states of the loci. The effects of inter-locus epistasis, simultaneous polymorphism and lack of recombination will be examined in section 3.6.

3.2 Derivation of the model

I consider a quantitative trait T affected by L QTL labeled $l = 1 \dots L$. I assume QTL analysis has been performed for a number of n lines yielding estimates for the number, position and average effects of the QTL on the trait that are used as parameters in the model (see Figure 3.1). Due to the poor resolution of QTL mapping which is limited by the frequency of recombination events in the crosses (Lynch et al. 1998) the particular genetic feature affecting a QTL is unknown. Usually, many genes lie within the range of a QTL marker (Lynch et al. 1998) and only further experiments can clarify the molecular basis of a QTL (Mackay et al. 2009). Yet, we know that each locus is characterized by a genotype, and the genotype at each locus affects the trait in a particular way. As an example, consider a trait affected by a transcription factor. A transcription factor is a molecule that binds to a regulatory region close to a gene to alter its gene expression. In that case, the regulatory region of a gene may affect the trait. This regulatory region consists of a specific sequence of nucleotides that affect the binding strength of a particular transcription factor. A change in that sequence can alter the binding strength of the transcription factor and thus alter the expression of a gene which finally affects the value of a quantitative trait. I approximate the relationship between trait and locus by a two-state variable q with states "on" (functional binding site in the regulatory region) or "off" (non-functional site). For simplification and since genetic details of a QTL are not known I describe each locus l by a number of macroscopic effect states q_l (states for short). That way, I only allow for few possible states at each locus and each genotype is assigned to one of this states. In general, different genotypes at a locus correspond to the same state (there are many different sequences with a functional binding site, and even more without). In the following, I denote the number of genotypes corresponding to a state by ω_q .

In general, each of the n lines could have a different state at a given locus. As there are limitations in QTL mapping, the information on the effect state of a locus is indirect. In most cases it is not known what genetic feature close to a marker

determines the state of the locus (as there are typically many genes linked to one marker). Instead, for each allele at a locus, QTL analysis gives the effect a particular allele has on the trait averaged over many crosses. And due to the uncertainty on the estimated effects, different alleles with a very similar effect on the trait cannot be distinguished. QTL studies involving crosses of 4 different lines (Blanc et al. 2006; Coles et al. 2010) showed that most QTL fall into a two state scheme, where a QTL either increases or decreases the trait value in a line. For this reason, I restrict myself to a two-state model with states $q_l = \pm 1$ per locus, effectively focusing on the genetic feature that has the largest effect on the trait. Yet, in a QTL study with 25 different maize lines, many different alleles could be found at one locus (Buckler et al. 2009), in which case an extension to a multiple state model would be necessary.

I assume a linear trait model without trait epistasis (inter-locus epistasis) where the state at each locus contributes additively to the trait

$$T(\{q_l\}) = \sum_{l=1}^L a_l q_l, \quad (3.1)$$

where $\{q_l\}$ denotes the set of states of all loci in a single line. The additive QTL effects a_l are obtained from experiment as the average trait contribution of a locus averaged over many crosses between different lines, as is the state q_l of a particular allele (see Figure 3.1). Without loss of generality I assume $a_l \geq 0$ and $q_l = \pm 1$, such that $q_l = +1$ (termed the + state) results in a higher trait value than $q_l = -1$ (the - state).

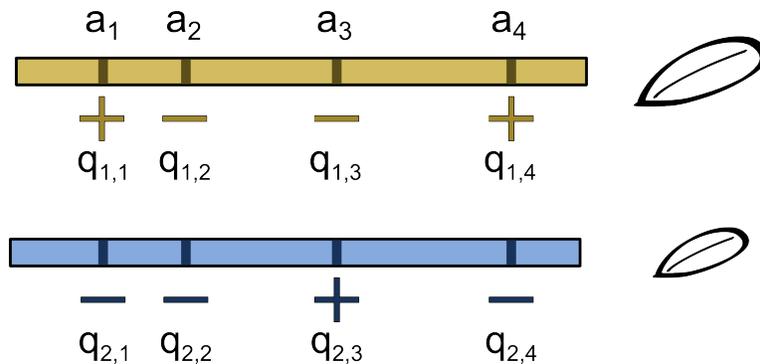


Figure 3.1: **QTL data available from QTL experiments.** From a QTL experiment one obtains the following data necessary for the selection test: the additive effect a_l that is fixed for locus l in all lines $i = 1, \dots, n$ and the state $q_{i,l}$ which can take the values ± 1 and which can be different for each line (assuming that there are only two possible states). The trait value is the sum of contributions from all loci, see eq. (3.1).

Furthermore, I assume a linear Malthusian fitness (log-fitness) landscape

$$F(\{q_l\}) = sT(\{q_l\}) = \sum_{l=1}^L sa_l q_l, \quad (3.2)$$

with the selection coefficient s . That is, if the trait is under selection (with selection strength s), the fitness increases linearly with the trait value. Fitness is the measure for the reproductive success of an individual and in population genetics fitness can be related to the probability of the establishment of an allele. Under the assumption of a linear fitness landscape, the effect of the state of a locus on both trait and fitness is independent of the states of other loci. This assumption will be examined and relaxed in section 3.6. The model contains no environmental component (quantitative traits often behave differently across different environments such as e.g. varying temperatures; such effects are not considered here) and (for a diploid population) no dominance (a dominant allele masks the effect of a recessive allele).

I use established results from population genetics theory to derive the probability of observing the possible states $q = \pm$ at a given locus. In population genetics the time evolution of a population of effective population size N is considered under the influence of selection, described by the fitness F , stochastic fluctuations of the reproductive process (genetic drift) and mutations (with a mutation rate μ) (Lässig 2007). While mutations introduce new alleles into a population, selection and genetic drift finally lead to their fixation. In the case of low mutation rates (also called weak-mutation regime, which is the relevant regime for eukaryotes and most prokaryotes (Stewart and Plotkin 2013)) the time until a new mutation arises is much longer than the time needed for the fixation of a new allele. Thus, the population is monomorphic (i.e. has the same alleles in all individuals) most of the time. The other case, where several mutations segregate in one population and compete against each other is called clonal interference (Gerrish and Lenski 1998). In summary, mutations introduce new alleles at a certain rate and these mutations are either fixed in the population with a certain probability or lost. Kimura (Kimura 1962) derived the rate of fixation of a new mutation in the weak-mutation regime. This can be done by solving the Fokker-Planck equation (2.2) in the stationary case, when $\partial P(x, t)/\partial t = 0$. This yields the probability of fixation of a new mutation with fitness advantage ΔF as $\frac{1 - \exp(-2\Delta F)}{1 - \exp(-2N\Delta F)}$. Multiplying this fixation probability with the probability of creating an individual with this mutation in the population, which is μN , yields the substitution rate

$$u_{q \rightarrow -q} = \mu N \frac{1 - \exp(-2\Delta F)}{1 - \exp(-2N\Delta F)}, \quad (3.3)$$

which depends on the fitness difference ΔF between the two possible states q and $-q$, the mutation rate μ , and the effective population size N . When considering multiple

loci, these loci do not segregate independently, but they are linked by the genetic sequence they share (called genetic linkage). There are two possible mechanisms that can break this linkage: recombination and a low mutation rate. Recombination events lead to a mixing of the genomes of different individuals. If a recombination event happens between two loci, one locus is inherited from one individual while the second locus is inherited from the other individual, which allows the loci to mix between different genomes. A low mutation rate leads to a monomorphic population where the mutations at different loci fix one by one and do not interfere. Thus, assuming the weak-mutation regime is sufficient to guarantee an independent segregation of the loci.

At low mutation rates, most loci are monomorphic at a given point in time, but may differ between lines (due to mutations that fix in a given population before the next mutation occurs). The state statistics $P(q)$ of a locus describes the probability that this locus in a given line is in state q . In the limit of long evolutionary times between lines, this statistics no longer changes with time, so the probability $P(q)$ is stationary (equilibrium). Under neutral evolution the genetic sequence constituting the locus is allowed to evolve freely and the equilibrium probability $P(q)$ depends only on the number of sequence variants ω_q of the locus corresponding to state q . One obtains

$$P(q|\Omega) = \frac{\omega_q}{\omega_+ + \omega_-} = \frac{\exp(\Omega q)}{\exp(\Omega) + \exp(-\Omega)}, \quad (3.4)$$

where I have introduced the multiplicity factor $\Omega = (1/2) \log(\omega_+/\omega_-)$ of a locus which allows to capture the imbalance between + and - state in a single parameter. Loci with a high multiplicity factor have more sequence variants for the + state and thus a higher probability to be in that state. In the example with the transcription factor binding site, the number of sequences with a functioning binding site ω_+ is much lower than the number of sequences without such a site ω_- , leading to $P(q = +1) \ll 1$ in the absence of selection (see also Figure 3.2). The multiplicity factor of a locus quantifies the asymmetry between the + and - state in the absence of selection, and correspondingly the relative number of mutations at a locus increasing or decreasing the trait.

Under selection, however, the fitness differences between the states can also create an additional bias towards one of the states. From eq. (3.3) one obtains that the ratio between the transition rates $u_{+\rightarrow-}/u_{-\rightarrow+}$ simplifies to $\mu_{+\rightarrow-}/\mu_{-\rightarrow+} \exp(2N\Delta F) \propto \exp(2N\Delta F)$ for $N\Delta F \gg 1 \gg \Delta F$ (strong selection) and the probability to be in state q is proportional to $\exp(2NF)$ (Iwasa 1988; Berg et al. 2004; Sella and Hirsh 2005; Lässig 2007). The bias in mutation rates between states is exactly captured by the multiplicity factor Ω . Thus the probability of states for one locus becomes

$$P(q|Ns, \Omega, a) = \frac{\exp(2Nsaq + \Omega q)}{\exp(2Nsa + \Omega) + \exp(-2Nsa - \Omega)}, \quad (3.5)$$

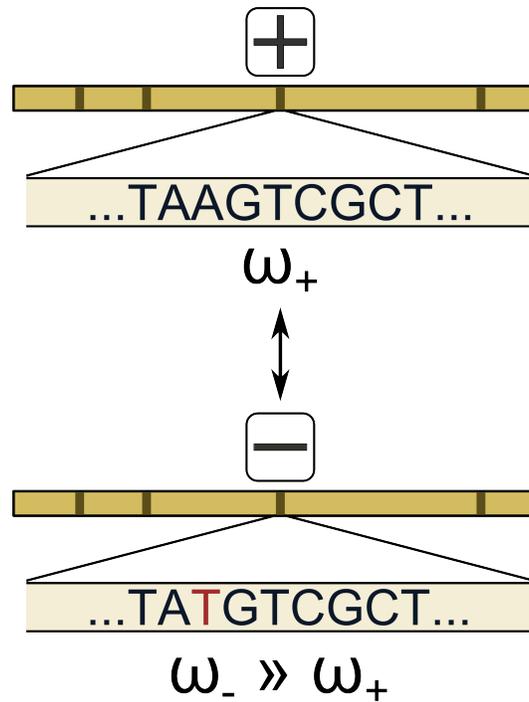


Figure 3.2: **Multiplicity factor of a QTL.** A QTL is typically governed by a longer genetic sequence (e.g. a transcription factor binding site) and many sequence variants contribute to the two effective states $q = \pm$. The number of sequence variants contributing to the $+$ and $-$ state, ω_+ and ω_- , can be very asymmetrically distributed. For example the $+$ state can correspond to a functional transcription factor binding site (top), while the $-$ state corresponds to a non-functional binding site (bottom). Since the number of sequence variants which correspond to a non-functional binding site is much larger, there is an asymmetry between the number of sequence variants corresponding to the two states, $\omega_- \gg \omega_+$. This asymmetry is contained in the multiplicity factor $\Omega = (1/2) \log(\omega_+/\omega_-)$.

where I used the linear fitness function from eq. (3.2). I define a selection coefficient $\sigma = Nsa$ for each locus proportional to the additive effect a , which is the relevant quantity determining the strength of the selection. Since the effective population size N is usually unknown and N and s always appear together, only their product will be estimated later.

If there are n lines, each line $i = 1 \dots n$ can be in a different state q at that locus. Thus, one locus consists of the set of states $(q_1, q_2 \dots, q_n)$. Assuming the lines evolve independently of each other, the joint probability distribution in the limit of long evolutionary time factorizes over lines, so the state statistics for a given locus is

$$P_{\text{undiv}}(q_1, \dots, q_n | Ns_1, \dots, Ns_n, \Omega, a) = \frac{1}{Z} e^{\sum_{i=1}^n (2Ns_i a + \Omega) q_i}, \quad (3.6)$$

where $Z = \sum_{q_1, \dots, q_n = \pm 1} e^{\sum_{i=1}^n (2Ns_i a + \Omega) q_i}$. Here, each line can be under a different selection pressure s_i but the multiplicity factor is fixed, which reflects the same genetic basis of that locus in all lines.

Here, one needs to consider one subtlety arising from QTL analysis based on crosses between individuals from different lines: In the crosses only the effects of loci differing in their state q in at least two lines can be determined. A locus that has the same allele in all lines will also have that allele in all crosses between the lines, rendering its trait contribution invisible. This means that, independent of the number of lines, there are always two state configurations $q_1 = q_2 = \dots = q_n = \pm 1$ that cannot be observed. From eq. (3.6) the result for the diverged loci directly follows as

$$P(q_1, \dots, q_n | Ns_1, \dots, Ns_n, \Omega, a) = \frac{1}{Z'} e^{\sum_{i=1}^n (2Ns_i a + \Omega) q_i}, \quad (3.7)$$

where I allow only diverged configurations (q_1, \dots, q_n) and the normalization constant has changed to $Z' = \sum'_{q_1, \dots, q_n = \pm 1} e^{\sum_{i=1}^n (2Ns_i a + \Omega) q_i}$ where the sum \sum' excludes the two unobserved configurations $q_1 = q_2 = \dots = q_n = \pm 1$.

Under the linear fitness model (3.2), states at different loci are statistically independent, so the state statistics for several loci is the product of (3.7) over loci

$$P(\{q_{i,l}\} | \{Ns_i\}, \{\Omega_l\}, \{a_l\}) = \prod_{l=1}^{L_{\text{div}}} P(q_{1,l}, \dots, q_{n,l} | Ns_1, \dots, Ns_n, \Omega_l, a_l), \quad (3.8)$$

where the number of loci with different states in at least two lines is denoted by L_{div} . With this state statistics I can assign a probability to each of the possible configurations $\{q_{i,l}\}$ given the selection strength of the lines and the multiplicity factors of the loci.

3.3 Inference of selection and log-likelihood scoring of evolutionary scenarios

The state statistics (3.8) can be used to infer the parameters of the model (selection strengths Ns_i for different lines and the multiplicity factors Ω_l at different loci) from experimental data on the states $\{q_{i,l}\}$ across lines and loci and on the additive effects $\{a_l\}$. Denoting the position of the maximum of a function $f(x)$ over x by $x^* = \underset{x}{\operatorname{argmax}} f(x)$, the maximum-likelihood estimates of the free parameters $\{Ns_i, \Omega_l\}$ are obtained by maximizing (3.8) with respect to the free parameters

$$\{Ns_i^*, \Omega_l^*\} = \underset{\{Ns_i, \Omega_l\}}{\operatorname{argmax}} P(\{q_{i,l}\} | \{Ns_i\}, \{\Omega_l\}, \{a_l\}) . \quad (3.9)$$

There are some limitations to the inference of multiplicity factors and selection strengths. In order to obtain a proper maximum likelihood estimate a minimal amount of data is required. For the estimation of the selection strength Ns_i the number of loci L is crucial, since the amount of data per selection parameter increases with L . For the estimation of the multiplicity factors the number of lines n is the important factor.

The number of lines in QTL studies is typically very low, with most studies in the range of 2 – 4 lines (Rebai and Goffinet 1993; Brem and Kruglyak 2005; Blanc et al. 2006; Coles et al. 2010; Steinhoff et al. 2011). For a low number of lines some restrictions on the estimation of the multiplicity factors exist. For two lines there are in principle four different state configurations per locus (see also Table 3.1) $(q_1, q_2) = (+, +), (+, -), (-, +)$ and $(-, -)$. Two of the configurations are not diverged between the lines and thus cannot be observed in experiment. The only observable loci are in states $(+, -)$ or $(-, +)$, for which the multiplicity factor Ω cancels out (see also Table 3.1). Hence the state statistics (3.7) does not depend on the multiplicity factors for two lines, making their inference impossible. This means on the one hand that one cannot access information about the multiplicity factors from two lines. On the other hand, there are less free parameters to be estimated in the maximum likelihood function (3.9) as the Ω_l drop out, leaving only the Ns_i to be estimated.

For three lines one either has configurations with two + states (e.g. $(+, +, -)$) that have a term $+\Omega$ in the state statistics (3.7) (see also Table 3.2) or configurations with two – states (e.g. $(-, -, +)$) that have a contribution of $-\Omega$. For the maximum likelihood estimation of Ω via eq. (3.9) both cases lead to extreme values of Ω^* : the estimate diverges to $+\infty$ for a surplus of + states and to $-\infty$ for a surplus of – states. This is caused by the insufficient amount of data per locus that leads to overfitting, categorizing the configurations into two classes: $\Omega^* = +\infty$ only allows configurations with two + states and assigns zero probability for configurations with two – states and vice versa for $\Omega^* = -\infty$. This overfitting is reduced with an increasing number of

			relative
q_1	q_2	probabilities	
-	-	$e^{-2N(s_1+s_2)a-2\Omega}$	
+	-	$e^{2N(s_1-s_2)a}$	
-	+	$e^{-2N(s_1-s_2)a}$	
+	+	$e^{2N(s_1+s_2)a+2\Omega}$	

Table 3.1: **Relative probabilities for the state configurations in two lines.** The relative probabilities of eq. (3.7) (excluding the normalization Z') are given for two lines. Only loci with states $(+, -)$ and $(-, +)$ in the two lines are detected in the crosses. For these states, only the difference of selection strength $s_1 - s_2$ of the two lines and the additive a enter the state statistics (3.7). The multiplicity factor Ω cancels out for the diverged state configurations.

				relative
q_1	q_2	q_3	probabilities	
-	-	-	$e^{2N(-s_1-s_2-s_3)a-3\Omega}$	
-	-	+	$e^{2N(-s_1-s_2+s_3)a-\Omega}$	
-	+	-	$e^{2N(-s_1+s_2-s_3)a-\Omega}$	
+	-	-	$e^{2N(+s_1-s_2-s_3)a-\Omega}$	
+	+	-	$e^{2N(+s_1+s_2-s_3)a+\Omega}$	
+	-	+	$e^{2N(+s_1-s_2+s_3)a+\Omega}$	
-	+	+	$e^{2N(-s_1+s_2+s_3)a+\Omega}$	
+	+	+	$e^{2N(+s_1+s_2+s_3)a+3\Omega}$	

Table 3.2: **Relative probabilities for the two unobservable and the six observable state configurations of a locus for three lines.** In contrast to the case of two lines, configurations that show no difference between two lines (e.g. $(q_1, q_2) = (+, +)$) can now be observed (as in $(q_1, q_2, q_3) = (+, +, -)$). Still, there are two configurations $(+, +, +)$ and $(-, -, -)$ that remain unobservable. For three lines, the multiplicity factor Ω enters the relative probabilities of all possible state configurations

lines (for 4 lines one has the three cases $\Omega^* = +\infty$ for e.g. $(+, +, +, -)$, $\Omega^* = 0$ for e.g. $(+, +, -, -)$ and $\Omega^* = -\infty$ for e.g. $(-, -, -, +)$) but the amount of knowledge that can be gained about the multiplicity factors from a small number of lines remains limited. Yet, it would be a mistake not to include the multiplicity factors into the analysis since a bias towards a certain state in many loci under neutral evolution might be misinterpreted as selection. Still, even with this very limited knowledge about the multiplicity factors selection can be inferred. Given that one observes for example only loci with two $+$ states (such that all Ω_l are positive) one can ask if an accumulation of the $-$ state is seen in one of the lines. Such a case would hint towards a difference in selection strength between the lines, as opposed to a case where the $-$ states are evenly distributed across the lines, as expected under neutral evolution.

Another restriction caused by the multiplicity factors is that selection strengths can only be determined relative to each other. The likelihood (3.9) depends on the states $q_{a,l}$ via $\sum_{a,l}(Ns_a a_l + \Omega_l)q_{a,l}$. A constant shift of the selection strength by an amount of s_0 in all lines, $s'_i = s_i + s_0$, can always be compensated by a shift of the multiplicity factors by $\Omega'_l = \Omega_l + 2Ns_0 a_l$ leaving the likelihood (3.9) unchanged. This makes the estimation of selection strength only possible up to an additive constant, which only fixes the fitness differences between the lines. Given a uniform selection strength $s_i = \bar{s}$ in all lines and multiplicity factors $\Omega_l = 0$, this situation cannot be distinguished from the case of neutral evolution with $s_i = 0$ but with multiplicity factors $\Omega_l = 2N\bar{s}a_l$. This is also true for the case of two lines where the multiplicity factors cancel out but only differences of the selection coefficients appear in the state statistics (see Table 3.1), so that no uniform mode of selection is detectable. Thus, for the rest of the thesis I only consider cases of lineage-specific selection and determine selection strengths relative to each other. Using further information on multiplicity factors (for instance from mutation accumulation experiments (Rice and Townsend 2012a)), or further assumptions (for instance that multiplicity factors are uncorrelated with the effect sizes, or are on average non-negative) one can also obtain information on absolute selection strengths from eq. (3.8).

Often the number of loci gained from QTL experiments is limited. When only few loci are known for a trait, the inference of all parameters may be unreliable due to overfitting. In this case it is convenient to restrict the parameter space of the s_i and test specific hypotheses against each other. For example, one can have a completely neutral scenario, where all selection coefficients are set to zero with $(s_1, s_2, \dots, s_n) = (0, 0, \dots, 0)$, a case of lineage-specific selection with only one line being under selection with $(s_1, s_2, \dots, s_n) = (s_1, 0, \dots, 0)$, or several lines under different selection pressures with $(s_1, s_2, \dots, s_n) = (s_1, s, \dots, s)$.

I define a log-likelihood score that quantifies the evidence for two such evolutionary scenarios compared against each other. For two scenarios P and Q the log-likelihood

score is defined as

$$S_{Q,P} = \sum_{l=1}^{L_{\text{div}}} \ln \left(\frac{Q(q_{1,l}, q_{2,l}, \dots, q_{n,l} | Ns_1^*, \dots, Ns_n^*, \Omega_l^*, a_l)}{P(q_{1,l}, q_{2,l}, \dots, q_{n,l} | Ns_1^{*'}, \dots, Ns_n^{*'}, \Omega_l^{*'}, a_l)} \right), \quad (3.10)$$

where for both scenarios the maximum likelihood estimate (3.9) for the corresponding model parameters is computed separately. The score is positive if the configuration $(q_{1,l}, \dots, q_{n,l})$ is more in accord with scenario Q and negative if it is more in accord with scenario P .

Both these scenarios are described by statistics of the form (3.8) but differ in their parameter values. The remaining selection parameters are estimated together with the multiplicity factors according to (3.9), giving the maximum likelihood estimate $\{Ns_i^*, \Omega_l^*\}$ for scenario Q and $\{Ns_i^{*'}, \Omega_l^{*'}\}$ for scenario P .

When two scenarios with different numbers of free parameters are tested against each other, the log-likelihood score is generally biased towards the scenario with more parameters. A simple way to correct this bias is the Bayesian information criterion (BIC) (Schwarz 1978). The BIC is a model selection criterion and penalizes a surplus in model parameters. When an increase in the number of parameters is not leading to a clearly increased fitting accuracy the model with more parameters will be rejected. Under the BIC correction, the log-likelihood score (3.10) is decreased by an offset

$$S_{Q,P}^{\text{BIC}} = S_{Q,P} - k/2 \ln L_{\text{div}}, \quad (3.11)$$

where k is the excess number of parameters of scenario Q compared to scenario P (assuming that Q has more free parameters).

3.4 Advantages of multiple-line testing

3.4.1 Increase in the number of detected loci

In this section I will highlight the advantages of multiple-line crosses for the inference of selection. There is a simple reason why a QTL selection test benefits from the use of multiple lines. Since only loci with different states in at least two lines can be observed in QTL analysis, a certain fraction of loci affecting the trait remains hidden since their effect cannot be measured. Increasing the number of lines increases the probability that a locus is diverged in at least one line, making it accessible to QTL analysis (see Figure 3.3). For two lines, there are 2 out of 4 possible states per locus that are diverged, $(q_1, q_2) = (+, -)$ and $(-, +)$ while loci with the states $(+, +)$ and $(-, -)$ cannot be observed. Increasing the number of lines to three, there are already 6 out of 8 states that are diverged, $(+, +, -)$, $(+, -, +)$, $(-, +, +)$, $(+, -, -)$, $(-, +, -)$,

and $(-, -, +)$. In the case that both states appear with equal probability this would mean a reduction of the fraction of unobserved loci from 50% in two lines to 25% in three lines. Increasing the number of lines further would lead to a further reduction of the proportion of unobserved loci. In general, the probability for a locus to remain unobserved under the effect of selection strength s_i and a multiplicity factor Ω is given by

$$\gamma(n|s_i, \Omega) = \prod_{i=1}^n P(+1|Ns_i, \Omega, a) + \prod_{i=1}^n P(-1|Ns_i, \Omega, a), \quad (3.12)$$

where I used eq. (3.5) to calculate the probability for the two unobserved states with $q_1 = q_2 = \dots = q_L = \pm 1$.

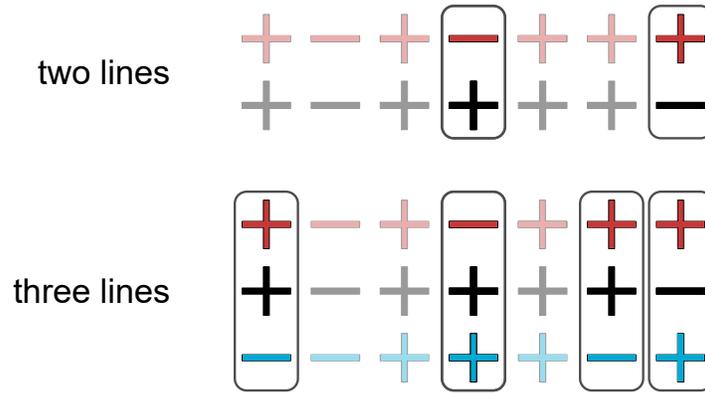


Figure 3.3: **Increasing the number of lines increases the number of diverged loci.** Top: A quantitative trait might be influenced by many loci. But many of them might not be observable (shown transparently) since they are not diverged between the two lines used for QTL analysis. Bottom: Increasing the number of lines typically increases the number of diverged loci, making QTL analysis more powerful.

To test how the number of lines affects the estimation of model parameters as in eq. (3.9) and the scoring of scenarios as in eq. (3.10), I compare selective and neutral hypotheses against each other using artificial data. For this I generate artificial QTL data under different scenarios, which I label for easy reference. In the first, neutral scenario P_0 , the selection strength on all lines is zero $(s_1, s_2, \dots, s_n) = (0, 0, \dots, 0)$. In the second scenario Q_1 , only line 1 is under selection $(s_1, s_2, \dots, s_n) = (s, 0, \dots, 0)$.

The selection strength for scenario Q_1 is chosen as $Ns = 10$, such that one obtains an average selection coefficient $\sigma = Nsa = 1$ (corresponding to a probability of 0.88 for a locus to be in the + state) with mean additive effect $a = 0.1$. The multiplicity factors Ω_l are set for both scenarios to ± 0.2 for half of the loci each (a value of $\Omega = 0.2$ corresponds to a 50% higher chance to have a + state than a - state under neutral

evolution). The additive effects $\{a_l\}$ are drawn from a gamma distribution, as in (Orr 1998; Zeng 1992), with shape parameter $\alpha = 2$ and rate parameter $\beta = 20$. After choosing the effects $\{a_l\}$, their values are fixed and are taken to be known explicitly (in practice obtained through experiments using QTL crosses).

In each run, a set of states $\{q_{1,l}, \dots, q_{n,l}\}$ is drawn for $L = 20$ loci from the probability distribution (3.6) using scenario Q_1 . For the subset of loci with different states in at least two lines I use eq. (3.9) to calculate the maximum likelihood values for the model parameters of both scenario P_0 and Q_1 , respectively. The log-likelihood score (3.10) is obtained by inserting both state statistics for Q_1 and P_0 with their respective estimated model parameters. To gauge the statistical significance of a given value of this score, I also estimate the probability of reaching the same score or higher under the neutral scenario P_0 . For this, I repeatedly draw configurations from the state statistics (3.6) under the neutral scenario P_0 , using the same input parameters for the Ω_l and the a_l . For each neutral configuration I calculate the score (3.10) as mentioned before. I define a p -value as the fraction of scores under neutral scenario P_0 that are equal to or larger than the score obtained under selective scenario Q_1 . This p -value quantifies how likely it is to obtain the observed score under the neutral scenario (type I error rate). The lower the p -value the less likely it is to obtain the score under scenario P_0 , strengthening the evidence for scenario Q_1 . To gauge how frequently a positive score occurs in favour of scenario Q_1 with selection on any of the lines 1, 2, or 3, the configurations drawn from the null model P_0 are sorted according to the size of their trait values T_1, T_2, T_3 (such that the trait with the highest trait value gets assigned the selection coefficient s of scenario Q_1). The results of the simulations for different numbers of lines can be found in Figure 3.4.

As expected, the mean log-likelihood score for the selective scenario Q_1 increases with the number of lines, which is the consequence of the increase in the number of detectable loci L_{div} . The increase in score is largest when going from two to three lines. Adding further lines only leads to a diminishing increase. The score finally saturates when all existing loci are detectable. Together with the increase in score a decrease in the average p -value can be observed, since more loci allow for a better distinction between the selective and the neutral scenario.

A theoretical expression for the dependency of the mean score on the number of lines can easily be derived. Assuming that the score contribution is on average the same for each locus, leading to a linear increase of the score with the number of loci, I obtain the relationship $S(n)/S_2 = \frac{1-\gamma(n|s_i, \Omega)}{1-\gamma(2|s_i, \Omega)}$, where $1 - \gamma(n|s_i, \Omega)$ is the number of diverged loci in n lines with $\gamma(n|s_i, \Omega)$ defined in eq. (3.12). Since I average over many loci to obtain S , the multiplicity factor Ω appearing here has to be understood as an average multiplicity factor. One can rewrite this relationship to give the score for n

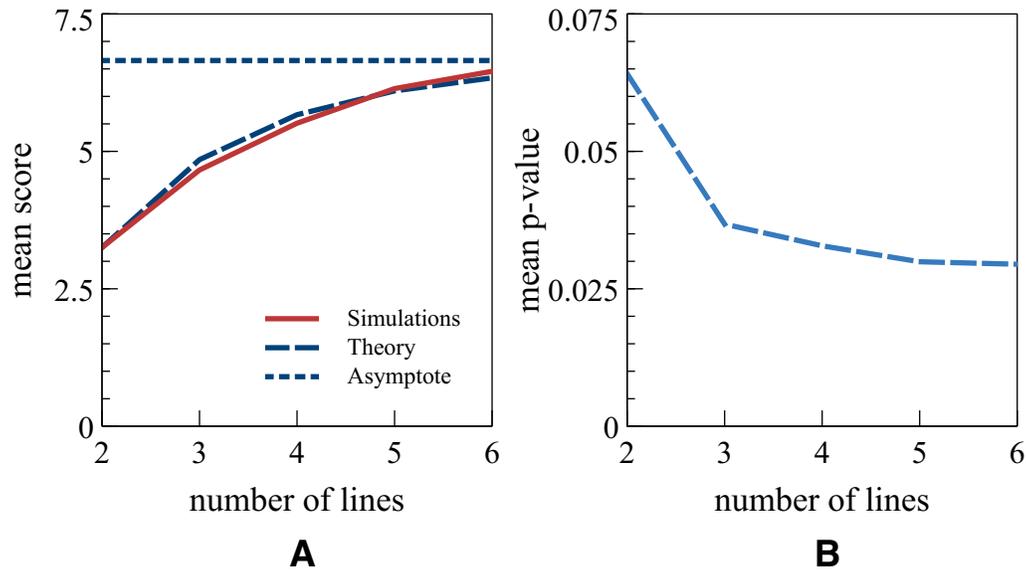


Figure 3.4: **Log-likelihood score and its statistical significance for different numbers of lines.** (A) The average log-likelihood score (3.10) is shown for many realizations of lineage-specific scenario Q_1 compared to neutral scenario P_0 for different numbers of lines at a fixed total number of $L = 20$ loci. With an increasing number of QTL lines more loci are diverged between the lines and thus are detectable. This leads to an increase of the score with the number of lines. The increase in score is largest from two to three lines. For a large number of lines an asymptotic value of the score is reached, as all loci become detectable. The theoretical prediction of the score increase (3.13) fits the simulation results very well. (B) As the score increases with an increasing number of lines the expected p -value for scenario Q_1 decreases, reflecting a better distinction between scenarios Q_1 and P_0 .

lines as a function of the score obtained for two lines S_2

$$S(n) = S_2 \frac{1 - \gamma(n|s_i, \Omega)}{1 - \gamma(2|s_i, \Omega)}. \quad (3.13)$$

This function increases monotonically with n and reaches an asymptotic value of $S_2 \frac{1}{1 - \gamma(2|s_i, \Omega)}$ when all loci become detectable for a large number of lines (see also Figure 3.4).

Nevertheless, the number of detectable QTL, and hence the statistical signal of selection can remain small, even when the number of lines is large. This is the case if selection strength is so high in all n lines that all or nearly all QTL have the same state in all lines. This can be seen from the expression for the fraction of unobserved loci $\gamma(n|s_i, \Omega)$, which tends to 1 as all s_i go to $\pm\infty$. For a selection coefficient $\sigma_i = Ns_i a > 2$ in all lines, less than 5% of the loci are observable for three lines (assuming $\Omega = 0$). In practice, this particular problem can be remedied by including one line with small selection pressure on the trait.

So far, the increase in statistical power in multiple-line tests is due to an increase in the number of diverged loci L_{div} with the number of lines. In order to address other, qualitative effects arising when the number of lines is increased, L_{div} is kept fixed for the remainder of this thesis.

3.4.2 Two vs. three lines at a constant number of crosses

Another interesting point to be addressed when comparing two- and multiple-line crosses is the efficient use of experimental crosses. Performing a large number of crosses would be preferable as this increases the statistical power of QTL mapping and thus the number of detected QTL. Yet, the necessary experimental effort limits the number of crosses that can be performed, which can go up to a few hundreds (Lynch et al. 1998). Performing a QTL analysis for multiple lines increases the experimental effort even further, as more pairwise crosses need to be performed between the different lines (or, depending on the mating design, with a common base line). For this reason I compare two- and three-line tests while keeping the total number of crosses constant. Given a fixed number of crosses that can be performed, should those crosses be concentrated on two lines, or should crosses between all possible pairs of three lines be performed (with fewer crosses between each pair of lines)?

Here, I use a simulation of a very simple scenario to test this hypothesis. A QTL mapping of a simulated QTL experiment is used to assess the power to detect selection in both cases. A quantitative trait is simulated for three lines by drawing a number of $L_{\text{div}} = 10$ diverged loci from the state statistics (3.7). I again use scenario Q_1 , where line 1 is under selection while the other lines evolve neutrally, as described above.

For the QTL mapping I simulate 100 single nucleotide polymorphism (SNP) markers, with 10 of the markers (equally spaced, i.e. markers 1, 11, 21, ...) being perfect QTL, meaning that they are perfectly linked to the QTL and their states always reflect the state of the linked QTL. Crosses between the different lines are simulated. In one case, a number of M_{tot} crosses is performed between lines 1 and 2 only, while in the other case $M_{tot}/3$ crosses each are performed for the pairwise combinations of lines 1 and 2, lines 1 and 3 as well as lines 2 and 3. The recombination probability between two adjacent markers is set to 0.25 such that the different QTL markers are passed on approximately independently. The trait value of each cross is determined from the state configurations at the QTL markers and the additive effects a_l using the linear trait model (3.1). For the QTL mapping I use the random forest QTL mapping algorithm as described in (Michaelson et al. 2010; Clément-Ziza et al. 2014) which can be used to perform two- and multiple-line QTL mapping. The QTL mapping results in a q -value for each of the markers, quantifying the probability that this marker is a true QTL. I only keep the markers with $q < 0.05$, assuming that these are the QTL affecting the trait. For each QTL the additive effects a'_l (note that I do not assume knowledge about the true underlying effects a_l) can be estimated by sorting the crosses into the two groups with states \pm at that particular marker and calculating the mean trait value for each of the groups. The difference of these means gives an estimate for a'_l . For three lines the estimates for a'_l coming from the different pairwise crosses are averaged over. The states q'_l for the QTL are set to $+$ if the nucleotide in that line is associated with a higher trait value and set to $-$ if it is associated with a lower trait value. For the resulting estimated state configurations $\{q'_l\}$ and additive effects $\{a'_l\}$ the score (3.10) is calculated, comparing scenarios Q_1 and P_0 .

As can be seen in Figure 3.5 (A) the mean score for the three-line crosses surpasses the score for the two-line case at around 200 total crosses. Similarly, at a high number of crosses the p -value is lowest for three lines (Figure 3.5 (B)). Thus, the three-line design is more effective at detecting selection given a sufficient number of crosses. However, for a small number of crosses the two-line test is more effective. The existence of two regimes can be understood as follows: On the one hand, there are typically more diverged QTL in three lines, as shown in the previous section. For a very high number of crosses the power of QTL mapping is high enough to detect all loci, leading to more detected loci in three lines (see Figure 3.5). On the other hand, for a diverged locus not all crosses are informative in three lines. If there is for example a locus with the state configuration $(+, -, -)$, only crosses between lines 1 and 2 as well as lines 1 and 3 are informative, while crosses between line 2 and 3 give no information about that QTL since it is not diverged between the lines. Thus for three lines more QTL are diverged but for each diverged QTL there are only $2/3$ of the crosses that are informative, making it more difficult to detect the QTL.

While I only consider a very simple model for QTL mapping here, where the exact

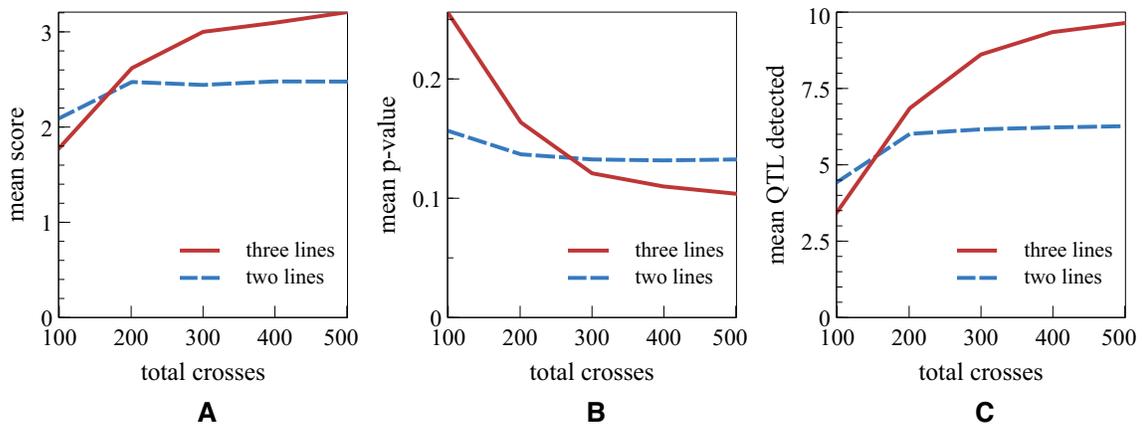


Figure 3.5: **Comparing the selection test on QTL mapping data for two and three lines at a constant number of crosses.** (A) The log-likelihood score (3.10) averaged over 1000 runs is plotted with pairwise crosses either between three lines or between two lines against the total number of crosses. The mean score for scenarios Q_1 tested against P_0 as a function of the total number of crosses increases faster for three lines with a crossover at around 200 crosses. (B) Similarly, the mean p -value decreases faster for three lines, making detection of selection easier for a high number of crosses. (C) For three lines more QTL are diverged compared to a two-line design. As a result, more QTL can be detected in the limit of a large number of crosses, where all diverged loci are also detected.

results depend on the details for the simulation, this still points out that it can be beneficial for the detection of selection to distribute a total number of available crosses on two lines. The crossover region at around 200 to 300 crosses is reasonable when compared to number of crosses taken in real QTL experiments (Lynch et al. 1998).

3.5 Statistical power of the selection test

In this section I want to determine under which conditions the detection of selection is possible. For that, several different evolutionary scenarios are tested against each other at a varying degree of selection strength. I use the scenarios P_0 and Q_1 as defined in section 3.4.1. Additionally, I define a second lineage-specific scenario Q_2 with $(s_1 = +s/2, s_2 = +s/2, s_3 = -s/2)$, where selection acts in different directions in the different lines. Finally, I define the only possible selective scenario for two lines Q_{two} with $(s_1 = +s/2, s_2 = -s/2)$ (which is identical to any other scenario with a difference in selection coefficients of $\Delta s = s$ since a constant mode \bar{s} with $(s_1 = +s/2 + \bar{s}, s_2 = -s/2 + \bar{s})$ vanishes), and the neutral two-line scenario P_{two} with $(s_1 = 0, s_2 = 0)$. Simulations are performed as described in section 3.4.1 but this time varying the selection strength in the selective scenarios. For three lines I test selective against neutral scenarios (Q_1 vs. P_0 , Q_2 vs. P_0) and different lineage-specific selection scenarios against each other (Q_2 vs. Q_1). Also for two lines selection is tested against neutral evolution (Q_{two} vs. P_{two}).

As can be seen in Figure 3.6 the score distributions of the scenarios Q_1 and P_0 separate as the selection strength s increases, making distinction between selection and neutral evolution more reliable. Also the fraction of p -values below a threshold of 0.05 in favor of the selective scenario increases steeply with increasing selection strength (see Figure 3.7). This shows that the test can clearly distinguish between selection and neutral evolution as well as different lineage selective scenarios. The test works in a suitable parameter regime, allowing to detect selection for only few available loci ($L \gtrsim 4$ loci for $Nsa = 1$) and reasonable selection strength ($Nsa = 1$ corresponds to a probability of 0.88 for a locus to be in the + state for $\Omega = 0$).

3.6 Robustness of the model assumptions

The selection test is based on a simple population genetics model and contains several simplifying assumptions. In this section, I explore how the test behaves in cases where the model assumptions are violated such that the model is no accurate description of the dynamics anymore. In particular, the model assumes a linear trait model, a linear fitness landscape, a low mutation rate leading to a mostly monomorphic populations and long evolutionary times, such that the state statistics is in equilibrium. Therefore,

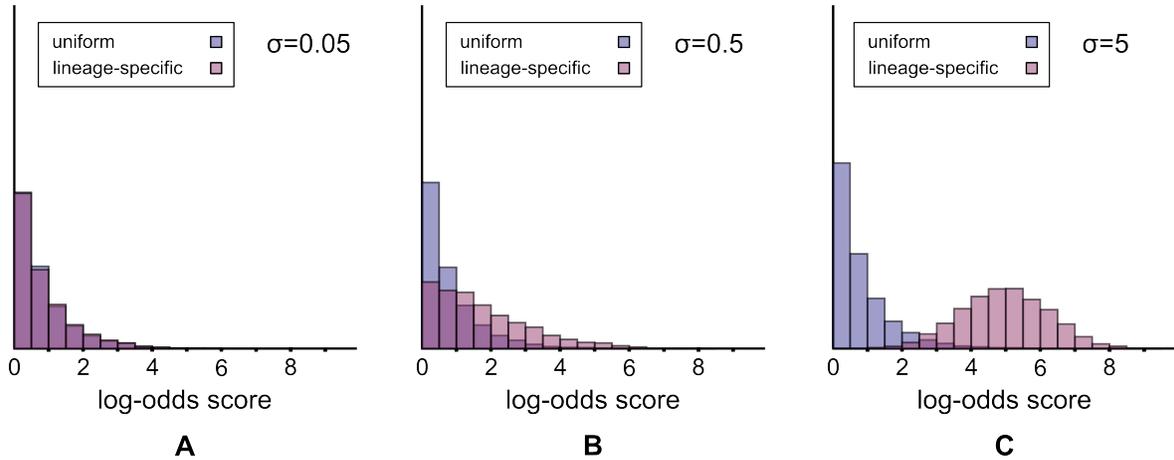


Figure 3.6: **Distribution of the log-likelihood score for selective scenario Q_1 and neutral scenario P_0 at different selection strengths.** (A)-(C) While for small selection strength selective and neutral scenario are not distinguishable, the separation of the score distributions increases with increasing selection strength, making the detection of selection more reliable.

I perform finite-population simulations that capture these features not considered in my model. The simulations explore a regime with and without multiple segregating loci, two kinds of epistasis, fitness and phenotype epistasis, and a range of evolutionary times.

3.6.1 Epistasis and multiple segregating loci

Wright-Fisher model For the simulations I use the Wright-Fisher model (Wright 1931) that explicitly describes the dynamics of a population of N individuals under the action of reproduction, selection, and mutations. The model considers non-overlapping generations, such that in one time-step the old generation is entirely replaced by the new generation. The population size is kept fixed during the simulation. To model reproduction, each individual in the new generation chooses one individual of the old generation at random from which it inherits the genome. Selection enters the step of reproduction such that the probability that an individual of the old generation is chosen as parent is dependent to its fitness. Mutations enter the model by switching the state of each locus in each individual with probability μ before the new generation is formed. I assume no recombination between the individuals, such that all loci are perfectly linked for each individual and only mutations can change the composition of loci in an individual.

An important quantity which has to be considered in population genetics is the

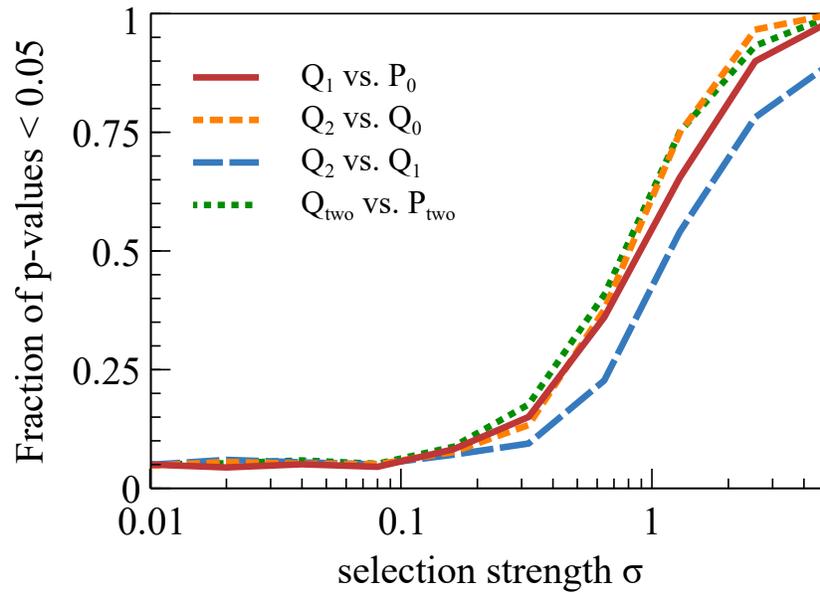


Figure 3.7: **The power of the selection test rises steeply with selection strength.** The fractions of p -values in favor of the selective scenarios under a threshold of 0.05 are given. The scenarios tested against each other are Q_1 vs. P_0 , Q_2 vs. P_0 , Q_2 vs. Q_1 , and Q_{two} vs. P_{two} . In all cases the fraction of statistically significant and correctly identified cases rises steeply with the selection strength. At very high selection strength all cases are correctly identified. This shows that the selection test can successfully identify selection and can distinguish between different selective scenarios.

number of polymorphic loci (which are loci for which two or more alleles segregate in the population) that are typically present in a population at one point in time. If the mutation rate μ is small (as assumed in my model), the population is monomorphic most of the time and if a new mutation occurs, the time required for that mutation to get lost or to spread to fixation is much shorter than the time interval between two successive mutations. If the mutation rate is large, however, there will be several mutations present in the population at a given time. These mutations will compete against each other (if they are beneficial), since without recombination only one of them can go to fixation. This phenomenon is called clonal interference (Gerrish and Lenski 1998). A consequence of clonal interference is that the adaptation speed is slowed down compared to a situation with recombination, since not all beneficial mutations can be fixed (Gerrish and Lenski 1998). A quantity to measure the degree of clonal interference in the system is $2\mu LN \ln N$, where L is the number of loci and N is the population size (Wilke 2004). $2\mu LN \ln N > 1$ is the regime where clonal interference plays a role (where on average at least one beneficial mutation arises per generation). Since I assume a low mutation rate in my model, such that there is no clonal interference, I will perform simulations with the Wright-Fisher model both with and without clonal interference to see the effect on the test performance.

Phenotype epistasis First, I consider phenotype epistasis. This means that the trait value is not only determined by the additive contributions of each separate locus, but also interactions between loci. To model phenotype epistasis, I add a pairwise interaction term to the linear relationship between QTL states and the trait value (3.1), yielding

$$T(\{q_l\}) = \sum_{l=1}^L a_l q_l + \sum_{l,m=1}^L J_{lm} q_l q_m. \quad (3.14)$$

J is a $L \times L$ symmetric matrix describing the interactions between loci. The interaction coefficients $\{J_{lm}\}$ are drawn from the same gamma distribution as the effects $\{a_l\}$ (and are assigned random signs). However, the average value of the $\{J_{lm}\}$ is varied relatively to that of the $\{a_l\}$ by multiplying them with a factor J_0/L . Then, for $J_0 = 1$ the cumulative contribution to the trait from the epistatic interaction $\sum_{l,m} |J_{l,m}|$ is on average as large as the contributions from the linear term $\sum_l a_l$. The regime of large J_0 corresponds to significant epistasis: in this regime the trait value T can change significantly with the change of state of a single locus. I assume that, as it is generally the case, the epistatic interactions $\{J_{lm}\}$ are not known.

I perform numerical simulations with the Wright-Fisher model using the epistatic trait model (3.14). Starting from a random initial configuration $\{q_l\}$ for $L = 15$ loci, a Wright-Fisher model is simulated with three independent populations of 100 individuals each evolving over M generations. I simulate the population under selective

scenario Q_1 . At the end of each run, the configuration of loci with the largest fraction in the population is used to calculate the score (3.10), comparing scenarios Q_1 and P_0 . I perform simulations both at a high mutation rate that leads to multiple segregating loci (mutation rate $\mu = 0.0002$ over $M = 3000$ generations, resulting in $2\mu LN \ln N \sim 27.6 \gg 1$ which is clearly in the regime of clonal interference) and in a second regime with low mutation rates ($\mu = 2.5 \times 10^{-5}$ over $M = 25000$ generations, with $2\mu LN \ln N \sim 0.35$; M is increased compared to the first case to compensate for the lower number of mutations in order to keep the total number of mutations approximately constant), where there is typically at most a single segregating locus.

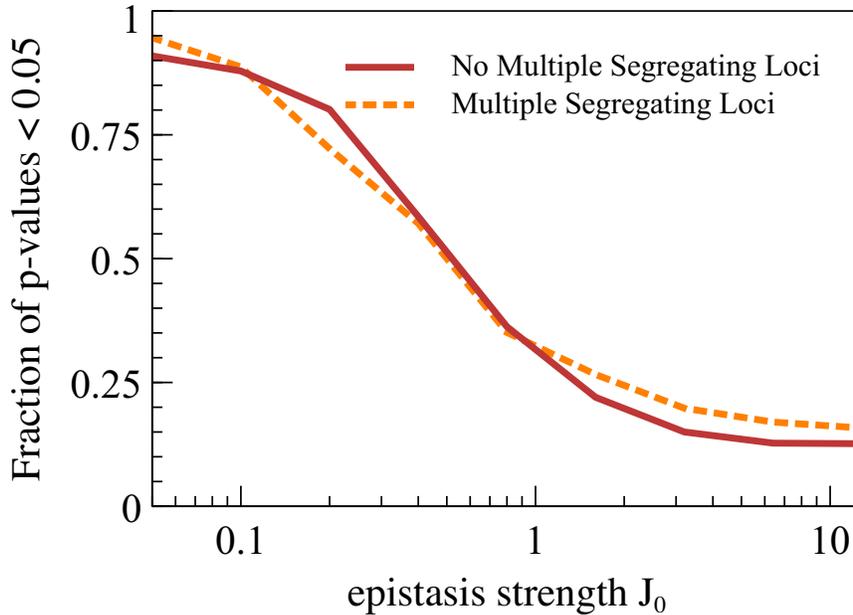


Figure 3.8: **Statistical significance of the selection test in the case of phenotype epistasis.** The expected p -value of the likelihood score (3.10) comparing selective scenario Q_1 and neutral scenario P_0 on data generated under scenario Q_1 assuming the epistatic trait model (3.14). With increasing coupling strength J_0 of the epistatic interactions the linear trait model assumed by the selection test becomes more and more inaccurate. This leads to a decrease in the sensitivity of the selection test. For very high coupling strength the test is not able to distinguish selective and neutral scenario anymore, but it retains some of its power for only moderate selection strength. There is only little difference between the cases with and without multiple segregating loci.

As J_0 is increased, the effect of each locus on the trait becomes coupled to the states of other loci and the linear trait model (3.1) becomes increasingly inaccurate. This results in a decrease in the sensitivity of the test, as can be seen in Figure (3.8).

While at very high coupling strength between the loci the test is not able to distinguish between selection and neutral evolution anymore, it still retains some power for low and moderate couplings. There is barely any difference to be seen between the cases with and without multiple segregating loci, suggesting that the test is robust with respect to a violation of the assumption of the weak mutation regime. If the epistatic interactions (3.8) would be known, these could be implemented into the test as well, restoring the sensitivity of the test.

Fitness epistasis Next, I test a second possible form of epistasis, fitness epistasis. In the case of fitness epistasis, the fitness function does not follow the linear form (3.2) but shows some kind of non-linear behavior with respect to the trait value. Here, I choose the special case of a quadratic fitness function

$$F(\{q_l\}) = -s_e(T(\{q_l\}) - T_0)^2, \quad (3.15)$$

where $T(\{q_l\})$ is the original, linear fitness function (3.2). The fitness function now has a parabolic shape with parameter T_0 giving the trait value with the maximal fitness and the parameter s_e determining how quickly fitness decreases away from the maximum. To compare fitness epistasis to the case of a linear fitness landscape, the fitness parameters T_0 and s_e are chosen such that mean and variance of the distribution of trait values T equals those without epistasis at a given value of the selection strength Ns . Each line can have a different set of fitness parameters $T_{0,i}$ and $s_{e,i}$, corresponding to different selection strength s_i for the linear model. In this way, the scenarios with and without fitness epistasis can be compared directly even though the shapes of the fitness functions are completely different. Again, I perform numerical simulations in regimes with and without multiple segregating loci, analogously to the case of phenotype epistasis. Figure 3.9 shows that there is no big difference between the cases with and without epistasis. The feature of the plot showing that the quadratic fitness function actually leads to an improvement of the sensitivity of the test is merely an artifact arising due to the very different shapes of the two fitness functions - fixing mean and variance for the trait values is not enough to perfectly compare the two fitness functions. As before there is no difference between the cases with and without multiple segregating loci. This shows that the test is not particularly sensitive to the choice of the fitness function (linear or quadratic fitness function).

Pleiotropy Beyond epistasis, the results of a QTL-based test for selection are potentially limited by pleiotropic effects: a subset of QTL of one trait may affect a second, unknown trait. If this unknown trait is under selection, but not the first, a QTL-based test may erroneously lead to the conclusion that the first trait is under selection (because some of its loci show a signal of selection induced by the second, unknown trait).

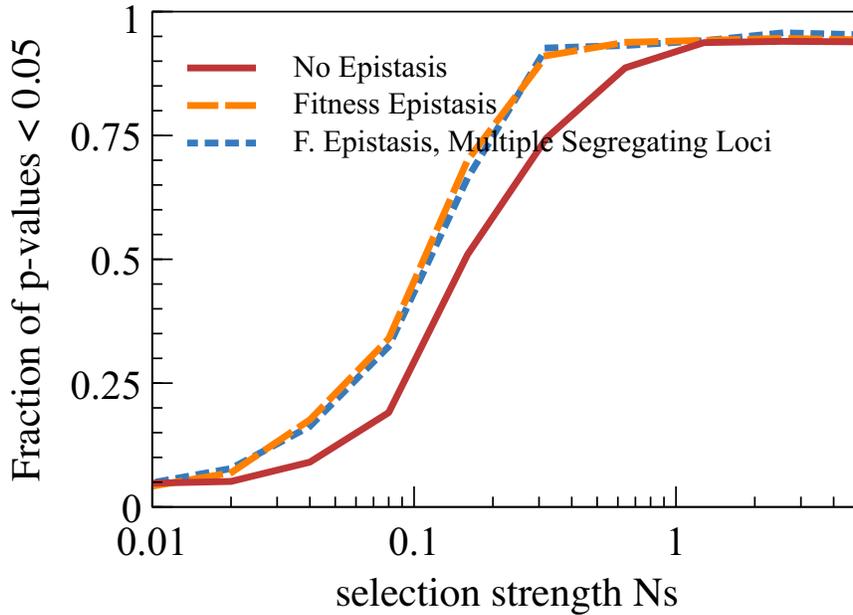


Figure 3.9: **Statistical significance of the selection test in the case of fitness epistasis.** The selection test (3.10) is applied to data generated under a model with a quadratic fitness function. The expected p -value was calculated as in the case of phenotype epistasis before, this time applying the selection test to data generated both under a model with a quadratic fitness function (3.15) and the linear version (3.2). The simulation results show that there is no qualitative difference between the cases with and without fitness epistasis. The sensitivity of the test is even larger in the case with fitness epistasis. This is, however, only an artifact of the very different fitness functions that cannot be compared perfectly (even though mean and variance of the distribution of trait values T is the same in both cases). Again there is no difference between the cases with and without multiple segregating loci. Simulation parameters: fitness parameters for each line $T_{0,i}$ and $s_{e,i}$ for the case with fitness epistasis are always chosen to match mean and variance of the trait value distributions for the three lines of the linear fitness case at a selection strength of $(Ns, 0, 0)$ for scenario Q_1 . Other simulation parameters as in Figure 3.14.

Hence, the evidence for selection from QTL statistics pertains to the trait the loci were identified for, or some unknown trait with substantial overlap of QTL loci with the trait under study. Conversely, the trait under study may be under selection (favouring + states, say), but some of its loci affect another trait also under selection, favouring – states. If the second trait is unknown, the test would infer a selection strength on the first trait that is too low. With a small number of lines or loci, the signal of selection may even be lost altogether.

Relaxed purifying selection Another limitation of the selection test concerns the interpretation of selection parameters. It is not possible to distinguish positive selection from relaxed purifying selection. Positive selection means a change in selection strength in a line that e.g. makes the increase of the trait value from a neutral baseline favorable. Relaxed purifying selection however denotes the situation, where all lines were under selection favouring a certain trait value (Fraser 2011). When the purifying selection pressure disappears in one of the lines, the trait value will tend back to its neutral baseline. Both cases lead to a pattern in the state configurations where the optimal trait value is very different between the lines and no distinction is possible. Instead, the test is sensitive to any kind of change in the selection strength acting on the lines. This problem is already known for different QTL-based selection tests and discussed in the work of (Fraser 2011).

3.6.2 Evolutionary timescales

Next, I assess the assumption of long evolutionary times since the divergence of the lines. The state statistics (3.8) was derived in the steady state and is reached a long time after the divergence of the different lines. This means that all loci have changed state often since the last common ancestor, such that the state statistics does not depend anymore on the initial state of the locus and the phylogenetic relationship between the lines. The equilibration time depends, besides the mutation rate, on the strength of selection and the size of the mutational targets. In a regime of short evolutionary times, most loci have not changed their state (and are thus not detected in crosses) and most diverged loci have undergone a single change of state over the phylogeny. With a sufficient number of lines, the two scenarios can be distinguished easily on the basis of the QTL states in all lines; in the limit of short times the states are compatible with a single mutation event in the phylogeny (for each diverged locus). Yet, this assumption might rarely be fulfilled perfectly in experiment. Therefore, I probe the statistical power of the equilibrium test at different evolutionary times. A version of the selection test working in the regime of short evolutionary times is developed in chapter 5.

I perform numerical simulations analogous to the ones described in section 3.5. But instead of drawing configurations $\{q_{1,l}, \dots, q_{n,l}\}$ from the equilibrium distribution (3.8),

I simulate transitions between states for a number of t time steps at each locus with substitution rates $\mu \frac{4Ns_i a}{1 - e^{-4Ns_i a}}$ and $\mu \frac{-4Ns_i a}{1 - e^{4Ns_i a}}$ for the transition from $-$ to $+$ and vice versa (Kimura 1962), see also chapter 5. The phylogenetic relationship between the lines is given by the phylogenetic tree in Figure 5.1. To simulate the transition between short and long evolutionary times, the average number of substitutions per locus μt is varied. At low values for μt only few loci are diverged, corresponding to the short time limit. At high values for μt each locus has switched states often, corresponding to long evolutionary times. To keep the number of diverged loci L_{div} constant, the number of total loci is varied with the mutation rate (L_{div} is smaller than the total number of mutable QTL loci L when the expected number of substitutions per locus is smaller than 1) and only events with a fixed number of L_{div} diverged loci are kept. For each configuration the log-likelihood score (3.10) is calculated and a p -value is obtained by repeated simulations of a neutral model on the observed diverged loci (keeping the additive effects a_i and the multiplicity factors Ω_i fixed).

The change of the sensitivity of the selection test (3.10) that is built on the assumption of long evolutionary times at varying levels of substitutions per locus is shown in Figure 3.10. The statistical power decreases only slightly when going from long to short evolutionary times and the test retains some of its statistical power even as μt becomes small, making it feasible to use even when the assumption of long evolutionary times is not completely satisfied. The statistics of states in this limit of short evolutionary times is derived in chapter 5.

3.7 Comparison to other selection tests

I give a brief summary of the connection to other selection tests on quantitative traits. The test developed here follows the test of Orr (Orr 1998) in that I use a two-state model at each locus and infer selection from the statistics of $+$ states (increasing the trait value) and $-$ states (decreasing the trait value). Also I condition the allele statistics on the phenotypic difference to deal with a potential bias introduced by testing multiple traits (see next chapter). Unlike Orr, I use population genetic models to compare the empirical allele statistics with the statistics observed under different evolutionary scenarios. The approach of Rice and Townsend (Rice and Townsend 2012a) is similar in spirit, but uses information from mutation accumulation experiments, which go beyond standard QTL analysis. Numerical simulations performed to assess the statistical power of the test are similar to Rice and Townsend (Rice and Townsend 2012b) (which however focus on the connection to the variance of the distribution of additive effects) and for the multiple testing simulations in the next chapter a scenario analogous to Anderson and Slatkin (Anderson and Slatkin 2003) is used.

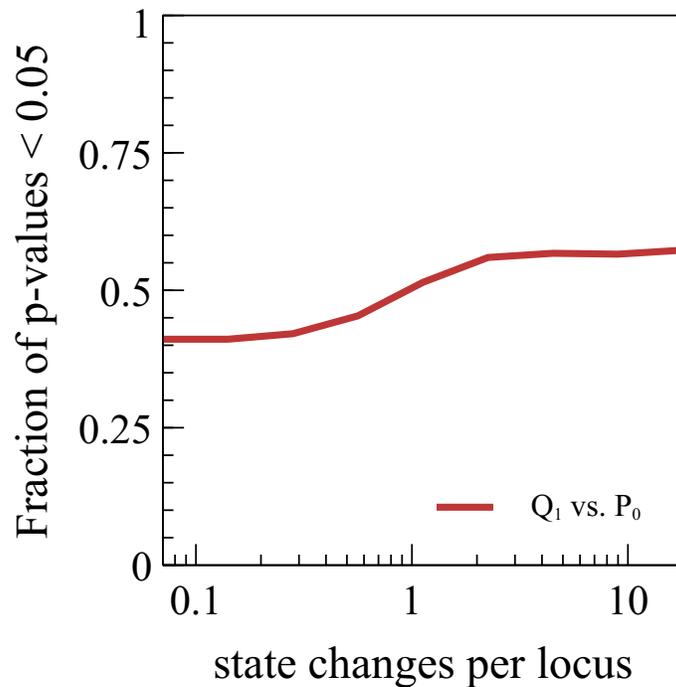


Figure 3.10: **The power of the selection test based on equilibrium statistics (3.8) over different evolutionary times.** The significance of the three-line equilibrium selection test decreases only slightly with decreasing number of state changes per locus since the last common ancestor (corresponding to shorter evolutionary timescales). Both at intermediate times and even for short evolutionary times the equilibrium test retains most of its power. Parameters: evolutionary scenario Q_1 tested against P_0 , average selection coefficient $\sigma = Nsa = 1$, number of diverged loci $L_{\text{div}} = 10$, phylogenetic tree as given in Figure 5.1 with $t_1 = t_2 = 50$ time steps, $\mu = 0.0002 - 0.06$.

Chapter 4

Multiple Testing and Maximum Entropy

In this chapter, I address the multiple testing problem that arises when analysing multiple traits from a QTL analysis and discuss several possible methods dealing with this problem: the Holm-Bonferroni method, the Benjamini-Hochberg procedure, and a method based on conditioning on the trait difference. I show that the state statistics conditioned on the trait difference can be obtained using the maximum entropy principle. I derive the conditioned state statistics and test the ability of the methods to reduce the false positive rate in a multiple testing scenario.

When the selection test is applied to multiple traits of the same QTL study it can happen that some of the traits get a positive test result even in the absence of selection. This, however, should not come as a surprise since, if the number of tested traits is high enough, one expects that some traits obtain high test scores by pure chance. This is known as the problem of multiple testing. There are several established methods that account for the multiple testing bias, like the Holm-Bonferroni method or the Benjamini-Hochberg procedure, which are briefly described in the next section.

Similarly, there might be a bias in the way the traits were selected for analysis in the QTL study. From a possibly very large pool of traits only few traits were selected for analysis because they showed a clear divergence between the lines. This induces an ascertainment bias on the level of the pool of possible traits, which is harder to quantify. It might not even be clear how many traits the trait pool comprises. There are cases like expression QTL (eQTL) studies, where all gene expression levels in an organism are considered and the total number of traits is clearly defined (Fraser et al. 2010). In this case a standard multiple testing correction is applicable. In other cases, where e.g. different leaf or flower characteristics of a plant are measured, the number of traits that can possibly be measured is unclear. One possible way to correct for this ascertainment bias was proposed by Orr (1998): The state statistics (3.8) will be conditioned on the

observed trait difference. That way, information coming from a difference in trait values is neglected from the analysis, circumventing a possible ascertainment bias. Instead, one answers the question if one still can distinguish neutral evolution and selection, given a certain trait difference between the lines. While this general idea was proposed by Orr (1998), I will show in the following that the state statistics conditioned on the observed trait difference can be obtained by the method of maximum entropy.

4.1 Classical multiple testing corrections

In this section I discuss some of the standard multiple testing corrections that can be applied if the total number of traits is known.

4.1.1 Holm-Bonferroni Correction

A suitable multiple testing correction is the Holm–Bonferroni correction (Holm 1979), which has the advantage that no independence of the different hypotheses needs to be assumed. This is particularly important in QTL analysis, since different traits can be affected by the same genetic loci. The Holm–Bonferroni correction controls the family-wise error rate (FWER), i.e. the false positive rate not only for a single trait but for a whole set of traits. The FWER is the probability that at least one false positive is observed in the set of traits. If there are m traits for which scenario Q is tested against the null hypothesis of scenario P , one calculates the log-likelihood score $S_{Q,P}$ (3.10) and the corresponding p -values p_j for all m traits. The traits are then ranked according to their p -values with the highest p -values first. Next, one searches for the first trait j^* for which $p_j > \alpha/(m+1-j)$, where α is the significance threshold for the family-wise error rate. Scenario P can then be rejected for the traits $1, \dots, j^* - 1$, but not for traits j^*, \dots, m .

The Holm-Bonferroni correction is a rather conservative and strict approach, since it focuses on the family-wise error rate. Thus, it barely allows any false positives in the whole set of tested traits and possibly neglects many traits that are under selection (meaning a high type II error rate). However, it might be desirable to find more true positive cases while accepting a higher false positive rate (type I error rate), especially for a very large number of traits as in the case of eQTL studies.

4.1.2 Benjamini–Hochberg procedure

A less strict test that controls the false discovery rate (FDR) instead of the FWER at the expense of a higher false positive rate is the Benjamini–Hochberg procedure (Benjamini and Hochberg 1995). For this procedure one again tests scenario Q against the

null hypothesis of scenario P for m traits and calculates $S_{Q,P}$ as well as the corresponding p -values p_j . The traits are again ranked according to their p -values, but now one searches for the largest trait index j^* , for which $p_j \leq \frac{j}{m}\alpha$, where α is the significance threshold for the false discovery rate. Now the null hypothesis of scenario P can be rejected for $j = 1 \dots j^*$. As this correction is not focusing on keeping the number of false positives in the whole set of traits small, it is especially suited for datasets with large number of traits, where a small fraction of false positives in the entire dataset is acceptable.

4.2 Conditioning on the trait difference

As emphasized by Orr (Orr 1998), a large trait difference between two lines alone is not sufficient evidence for lineage-specific selection. The proposal of Orr for this case is to use, in place of (3.8), a state statistics conditioned on the empirical trait difference between two lines by restricting the states to those giving rise to the observed trait difference $T_1 - T_2$. In doing so, the part of the evidence for selection that comes from the trait difference between two lines is discarded. Instead, one focuses on differences in the state statistics between neutral and selective scenario given that they produce the same trait difference. Orr defines the trait difference as

$$R = T_1 - T_2 = \sum_{l=1}^L a_l(q_{1,l} - q_{2,l}) \quad (4.1)$$

for the case of two lines (Orr 1998). Here, I generalize this notion to the case of n lines and denote the maximal trait difference across two lines

$$R_{\max} = \sum_{l=1}^L a_l(q_{1,l} - q_{2,l}), \quad (4.2)$$

where the lines are ordered such that line 1 has the largest trait value $T_1 = \sum_{l=1}^L a_l q_{1,l}$ and line 2 has the smallest trait value T_2 .

The goal is to calculate the state statistics (3.8) conditioned on R_{\max} , $P(\{q_{i,l}\} | R_{\max})$. This statistics can then be used in the log-likelihood score (3.10) in place of the neutral null model. The calculation to obtain $P(\{q_{i,l}\} | R_{\max})$ is based on the principle of maximum entropy.

The principle of maximum entropy was first stated by E. T. Jaynes (Jaynes 1957). This general principle applies to situations with incomplete knowledge on the probability distribution $p(x)$ of some variable x . This distribution must be consistent with any prior information on x one might have (for instance the mean value of x), but

otherwise it should be as unbiased as possible. The principle of maximum entropy states that the distribution which best describes the incomplete state of knowledge is the distribution which maximizes the information entropy

$$H(X) = - \sum_x p(x) \ln p(x) \quad (4.3)$$

with respect to $p(x)$ subject to the constraints resulting from prior information. The principle of maximum entropy is a central part of statistical physics, where the maximum entropy principle gives rise to the well known Boltzmann distribution $p(x) \propto e^{-\beta E(x)}$ which is derived for the configuration x of an ensemble of particles subject to a constraint on the mean value of the energy $E(x)$. Other applications of the principle of maximum entropy are in image reconstruction (Narayan and Nityananda 1986), language modelling (Berger et al. 1996), and neural networks (Mora and Bialek 2011). In the context of quantitative traits, the principle of maximum entropy and the associated calculus of exponential distributions has been used to estimate unobserved allele frequencies and to infer selection from trait observables (Prügel-Bennett and Shapiro 1994; Ruttray 1995; Prügel-Bennett and Shapiro 1997; Berg et al. 2004; Mustonen and Lässig 2005; Lässig 2007; Mustonen et al. 2008; Barton and de Vladar 2009; de Vladar and Barton 2011a; Nourmohammad et al. 2013a; Nourmohammad et al. 2013b).

4.2.1 Pedagogical example

Here, I give a simple concrete example to demonstrate the link between ascertainment bias and the maximum entropy principle. For this, I derive a probability distribution subject to a constraint on the mean value of the variable using the maximum entropy principle.

Consider a uniform distribution $p_0(x)$ in the interval $[0, 1]$. Now sets of ten numbers x_i , $i = 1, \dots, 10$ are drawn independently from the distribution $p_0(x)$ and their sum $m = \sum_{i=1}^{10} x_i$ is considered (see Figure 4.1 left). The value of m will fluctuate from set to set with an average of 5 and each contributing value x_i will come from the uniform distribution $p_0(x)$. Now, I filter the results and retain only those sets whose sum is close to $\bar{m} \neq 5$. While the sum of values m is fixed to one particular value, the individual x_i contributing to the sum are still variable and are originally drawn from the uniform distribution $p_0(x)$. Yet, this external condition creates a bias in the original distribution $p_0(x)$, and for $m > 5$ one finds that larger values x appear with a higher probability compared to the uniform distribution (see Figure 4.1). So clearly, the effective distribution $p_0(x)$ from which the x_i are drawn under conditioning on \bar{m} is different. This is the ascertainment bias induced by conditioning the sum of each set. Now the question is, how does the restriction of the sum m influence the distribution $p_0(x)$ of the single x_i ? This is answered by the principle of maximum

entropy, which allows to determine the exact form of this biased distribution $p(x)$. The relative information entropy between the original (uniform) distribution $p_0(x) = 1$ for $x \in [0, 1]$ and biased distribution $p(x)$ has to be maximized:

$$H(p) = - \int_0^1 dx p(x) \log \frac{p(x)}{p_0(x)}, \quad (4.4)$$

subject to the constraints

$$\int_0^1 dx p(x) = 1, \quad \int_0^1 dx xp(x) = \frac{\bar{m}}{N}, \quad (4.5)$$

where $N = 10$ is the size of each set. Here, the first condition ensures the normalization of the distribution and the second condition fixes the average contribution of a single value x to \bar{m}/N . The two constraints can be incorporated using two Lagrange multipliers. The method of Lagrange multipliers is a method to maximize a function subject to one or more constraints. In order to maximize a function $f(x)$ subject to a constraint $g(x) = 0$, one maximizes the function $\mathcal{L}(x, \lambda) = f(x) + \lambda g(x)$, where λ is a constant that needs to be determined to fulfill the constraint. This is easily generalized to multiple constraints, where one additional Lagrange multiplier is added for each constraint. The maximum of the function is then found by taking the partial derivatives with respect to all variables and Lagrange parameters and setting them to zero.

In this case the method of Lagrange multipliers for the relative information entropy together with the constraints (4.5) yields

$$\begin{aligned} \hat{H}(p) = & - \int_0^1 dx p(x) \log p(x) + \lambda_1 \left(\int_0^1 dx p(x) - 1 \right) \\ & + \lambda_2 \left(\int_0^1 dx xp(x) - \bar{m}/N \right), \end{aligned} \quad (4.6)$$

which needs to be maximized with respect to p , λ_1 , and λ_2 . Here, I inserted $p_0(x) = 1$ for $x \in [0, 1]$. The derivatives with respect to λ_1 and λ_2 automatically give the two constraints (4.5). The functional derivative with respect to p yields

$$\frac{\partial \hat{H}}{\partial p} = -(\log p(x) + 1) + \lambda_1 + \lambda_2 x = 0. \quad (4.7)$$

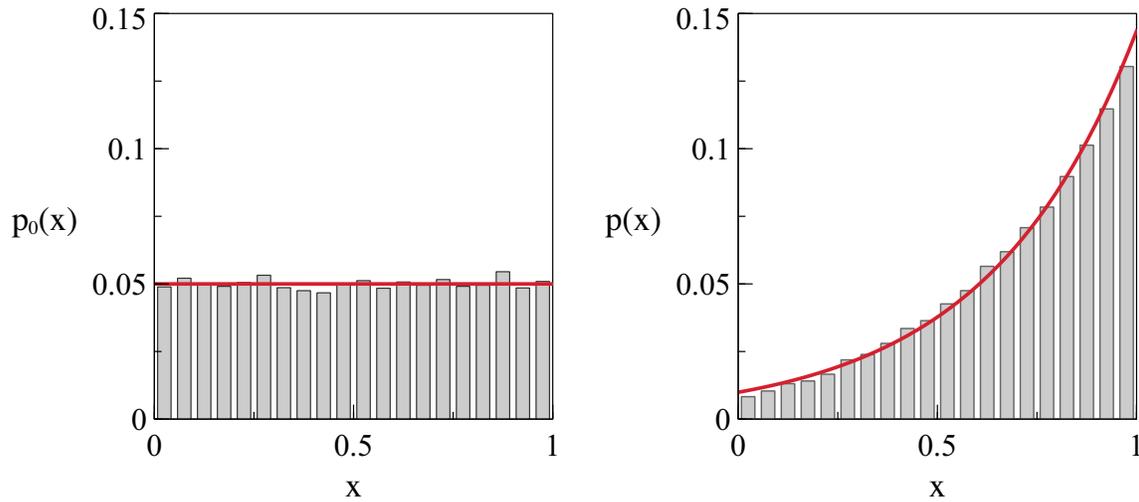


Figure 4.1: **A biased distribution can be inferred with the maximum entropy method.** Left: Numbers drawn from a uniform distribution in the interval $[0, 1]$. Right: Only sets of ten numbers drawn from $p(x)$ are considered that have a sum close to $m = 8$. This leads to a bias in the distribution $p(x)$ which now has a mean of 0.8. The exact form of the distribution can be calculated analytically using the method of maximum entropy. This analytical result takes on an exponential form (red line) and agrees perfectly with the histogram obtained by drawing sets of ten numbers randomly from the uniform and only keeping 1000 sets with a sum in the interval $7.95 < m < 8.05$.

Solving with respect to $p(x)$ yields

$$p(x) = e^{\lambda_2 x + \lambda_1 - 1}. \quad (4.8)$$

Ascertainment bias thus makes x exponentially rather than uniformly distributed. The values for λ_1 and λ_2 have to be determined by inserting $p(x)$ from (4.8) into the constraints (4.5) and solving for λ_1 and λ_2 . For the case of $\bar{m} = 8$ and $N = 10$ one obtains $\lambda_1 \approx -1.62$ and $\lambda_2 \approx 2.67$. The result for $p(x)$ shown in Figure 4.1 agrees perfectly with the histogram of numbers in sets with a constrained sum (note that a perfect agreement is only expected for a large number of N ; as seen from Figure 4.1, $N = 10$ already leads to a reasonably good approximation).

Suppose one did not know whether the original distribution $p(x)$ from which the data were drawn was uniform or not and one had access only to data subject to the known constraint. If the distribution of the empirical data deviates from or agrees with the maximum entropy distribution $p(x)$, then this deviation or agreement could be used to quantify the likelihood that the original data came from the uniform distribution (vs. an alternative hypothesis).

I now adopt this approach for the log-likelihood testing with the log-likelihood score (3.10). For this, the neutral (uniform) scenario is replaced by the maximum entropy distribution of the neutral scenario conditioned on the trait difference (tested against an alternative, selective hypothesis). That way, I test if the data rather originate from a completely neutral state statistics, which is however biased due to a higher trait difference, or if they originate from a proper selective scenario.

4.2.2 Derivation of the ascertained neutral scenario

I now apply the principle of maximum entropy analogously to the state statistics (3.7) to derive the neutral scenario $P_h(q_1, \dots, q_n | \Omega, a)$ conditioned on the trait difference R_{\max} from the original neutral scenario P_0 . This introduces an additional parameter h which is used to adjust the trait difference R_{\max} in the neutral scenario. The distribution $P_h(q_1, \dots, q_n | \Omega, a)$ is obtained by maximizing the information entropy

$$\begin{aligned} H(P) = & - \sum_{q_i = \pm 1} P_h(q_1, \dots, q_n | \Omega, a) \log \left(\frac{P_h(q_1, \dots, q_n | \Omega, a)}{P_0(q_1, \dots, q_n | \Omega)} \right) \\ & + \lambda_0 \left(\sum_{q_i = \pm 1} P_h(q_1, \dots, q_n | \Omega, a) - 1 \right) \\ & + h \left(a \sum_{q_i = \pm 1} (q_1 - q_2) P_h(q_1, \dots, q_n | \Omega, a) - \frac{R_{\max}}{L_{\text{div}}} \right), \end{aligned} \quad (4.9)$$

with respect to $P_h(q_1, \dots, q_n | \Omega, a)$. Here, the first term denotes the unconstrained information entropy and the other two terms again comprise the Lagrange multipliers for the constraints. The second term ensures the normalization of the distribution P_h with a normalization parameter λ_0 and the third parameter fixes the trait difference to R_{\max} with the additional parameter h . Again the sums \sum' over all possible states $q_i = \pm 1, i = 1, \dots, n$ exclude the undiverged states $q_1 = \dots = q_n = \pm 1$.

Taking the functional derivative with respect to P gives, similar to the result in the previous section

$$\frac{\partial \hat{H}}{\partial P} = - \left(\log \frac{P_h}{P_0} + 1 \right) + \lambda_0 + ha(q_1 - q_2) = 0. \quad (4.10)$$

Solving this equation with respect to P_h gives

$$\begin{aligned} P_h(q_1, \dots, q_n | \Omega, a) &= \exp(ha(q_1 - q_2)) P_0(q_1, \dots, q_n | \Omega) / \tilde{Z} \\ &= \frac{\exp(ha(q_1 - q_2) + \Omega \sum_{i=1}^n q_i)}{\sum'_{\{q'_1, q'_2, \dots, q'_n = \pm 1\}} \exp(ha(q'_1 - q'_2) + \Omega \sum_{i=1}^n q'_i)}, \end{aligned} \quad (4.11)$$

where I have the new normalization constant $\tilde{Z} = \lambda_0 - 1$. Note that equation (4.11) is only the result for a single locus. As defined before in (3.8) one obtains the full state statistics by taking the product over all loci $P_h(\{q_{i,l}\} | \{\Omega_l\}, \{a_l\}) = \prod_{l=1}^{L_{\text{div}}} P_h(q_{1,l}, \dots, q_{n,l} | \Omega_l, a_l)$, where the parameter h is the same for all loci.

As can be seen in equation (4.11), the old neutral state statistics P_0 gets multiplied by a factor of $\exp(ha(q_1 - q_2))$ with an additional parameter h that determines the trait difference $R_{\max} = \sum_l a_l (q_{1,l} - q_{2,l})$. A larger value of h leads to a higher probability for a + state in line 1 and a higher probability for a - state in line 2, thus leading to an increasing trait difference R_{\max} . This increase of the trait value in line 1 and decrease in line 2 always happens symmetrically around a mean trait value (which can itself be biased towards a higher or lower trait value by the combined effect of the multiplicity factors Ω_l which are the same for all lines) (see Figure 4.2 a). This can be distinguished from other selective scenarios, where for example only one line is under selection and has a trait value diverged from the value of the other lines (see Figure 4.2 b).

The maximum-entropy statistics P_h conditioned on R_{\max} will be used to describe the state statistics under neutral evolution and with ascertainment bias. One can now compare a selective scenario against the neutral scenario under conditioning on R_{\max} using the log-likelihood score

$$S_{Q, P_h} = \sum_{l=1}^L \ln \left(\frac{Q(q_{1,l}, q_{2,l}, \dots, q_{n,l} | N s'_1, \dots, N s'_n, \Omega'_l, a_l)}{P_{h^*}(q_{1,l}, q_{2,l}, \dots, q_{n,l} | \Omega_l^*, a_l)} \right). \quad (4.12)$$

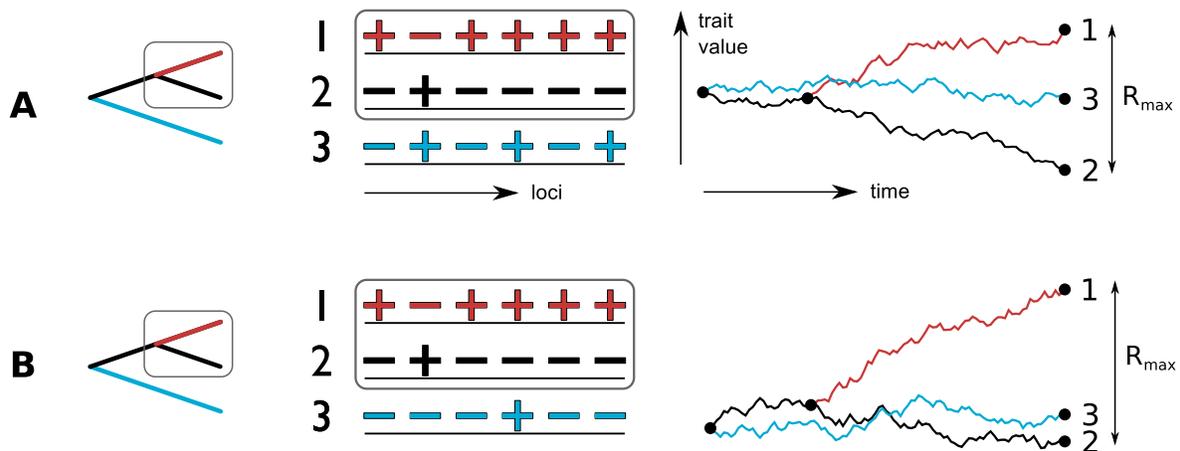


Figure 4.2: **Maximum entropy solution for neutral evolution under ascertainment bias can be distinguished from selection.** (A) The maximum entropy solution for neutral evolution conditioned on the trait difference R_{\max} (4.11) predicts that the trait contribution of the line with the highest trait value (line 1) and the line with the lowest trait value (line 2) symmetrically diverge from the mean trait value (which can itself be biased due to the effect of the multiplicity parameters Ω_l which are however the same for all lines). (B) In many selective scenarios, however, the situation is different. For example, only one line (line 1) can be under selection for a high trait value. This leads to one line with a trait value very different from the trait values of the other lines. Therefore, the selective scenario becomes distinguishable from the neutral scenario under ascertainment bias. Note that the distinction between neutral and selective scenarios is not possible when only two lines (line 1 and 2) are available. In this case, both scenarios would simply produce the same trait difference R_{\max} and one would have no possibility to distinguish between them.

This log-likelihood score now also depends on the ascertainment parameter h . The value of the ascertainment parameter h^* are estimated together with the values of the multiplicity parameters for scenario P_h by maximizing the log-likelihood score with respect to all parameters.

If one considers the maximum entropy solution (4.11) for the case of two lines, one sees that the probabilities for the two observable states $(+, -)$ and $(-, +)$ $\exp(ha(q_1 - q_2))/C$ match the probabilities of the selective scenario (see Table 3.2) $\exp(2N\Delta sa(q_1 - q_2))/C$ (the multiplicity parameters cancel out in the case of two lines). Here, the ascertainment parameter h and the selection parameter $2N\Delta s$ can be identified with each other, yielding the same functional form. Maximizing the log-likelihood score (4.12) with respect to Δs for the selective scenario Q and with respect to h for the neutral scenario then lead to exactly the same state statistics, making distinction between the two cases impossible. As a result, the log-likelihood score comparing evolution under selection at equilibrium with the neutral statistics conditioned on the observed trait value is exactly zero. Hence, for two lines at equilibrium it is not possible to distinguish neutral evolution with ascertainment bias from the effect of selection.¹

This is, however, completely different for more than two lines. In this case the state statistics in the selective scenario in equilibrium (3.8) differs from the neutral scenario and the log-likelihood score (4.12) generally gives non-zero results, making distinction between neutral evolution and selection possible. (However, there is again a particular selection scenario, $s_1 = +s$, $s_2 = -s$, $s_3 = 0$, \dots , $s_n = 0$, that is not distinguishable from neutral evolution conditioned on R_{max} .)

To test the statistical power of the different approaches to the multiple testing problem, I perform numerical simulations using a simple multiple testing scenario where a trait is picked from a larger set of traits. This multiple testing scenario follows the lines of Anderson and Slatkin (Anderson and Slatkin 2003). For this I draw state configurations $\{q_{l,i}\}$ for $m = 5$ traits at random from the neutral scenario P_0 for three lines. Then the traits are ordered according to the maximal phenotypic difference R_{max} occurring between any two of the three lines. To create an ascertainment bias I only pick the trait with the largest phenotypic difference for further analysis. For this trait I test for selection using selective scenario Q_1 against the neutral scenario. This is done in three different ways: First, by using the completely neutral scenario P_0 without any correction for ascertainment bias. Second, using the conditioned neutral scenario P_h

¹A key difference of my log-likelihood score to Orr's test is that Orr not only uses the empirically observed additive effects $\{a_l\}$ available from crossing experiments, but also additive effects drawn from a plausible distribution $P(a)$. Orr's test can appear to yield significant results when calculating the trait difference R using the additive effects empirically determined from crosses, but using a different set of additive effects drawn from some distribution $P(a)$ for p -value computations. Consistent with this, Rice and Townsend found that the outcome of Orr's test strongly depends on the assumptions made on that distribution and that the test can produce nonsensical results (Rice and Townsend 2012b) in particular cases.

and third, by using scenario P_0 and applying the Holm-Bonferroni correction. In each case the score (4.12) is computed, testing selective scenario Q_1 against the corresponding neutral scenario, as well as the corresponding p -value. This procedure is repeated many times over to get an estimate for the false discovery rate (FDR) in all cases. Here, a significance threshold of $p < 0.05$ ($p < 0.05/5 = 0.01$ in the case of the Holm-Bonferroni correction) is applied. The false discovery rates (which can be found in Table 3.1) are clearly elevated without any correction while the conditioning on R_{\max} and even more pronounced the Holm-Bonferroni correction lead to a reduced FDR. This result for the conditioning on R_{\max} is in accord with Anderson and Slatkin (Anderson and Slatkin 2003), who found that the Orr test, which uses a similar correction scheme, also led to conservative test statistics.

method	false discovery rate
no correction	0.20
conditioning on R_{\max}	0.11
Holm-Bonferroni	0.044

Table 4.1: **False discovery rates for the different methods accounting for the multiple testing bias.** While applying no correction leads to a clearly elevated false discovery rate, this can be significantly reduced by applying the neutral scenario conditioned on the phenotypic difference. The Holm-Bonferroni procedure yields the smallest false discovery rate, but here the total number of traits has to be known.

Next, I simulate one trait under the selective scenario Q_1 and the other traits under the neutral scenario P_0 . Then, I determine in which fraction of the cases the trait under selection is correctly identified with a p -value $p < 0.05$ ($p < 0.05/5 = 0.01$ in the case of the Holm-Bonferroni correction) for all three methods (true positive rate), given that the trait under selection turns out as the trait with the highest trait difference. As can be seen in Figure 4.3, the true positive rate in the case without any correction is highest, but also suffers from the highest false positive rate (which can be read off from the value of the curve at the very left of the plot, where the selection strength $\sigma \rightarrow 0$ corresponds to a neutral scenario). Both the conditioning on the phenotypic difference and the Holm-Bonferroni correction have a lower true positive rate, but at a much lower false positive rate. For high selection strength σ the Holm-Bonferroni method has a higher statistical power to detect selection and should thus be preferred (the fraction of significant p -values is higher for the conditioning on R_{\max} for small values of σ only because it operates at a higher effective false positive rate as the Holm-Bonferroni method). However, the Holm-Bonferroni method can only be applied if the total number of traits is known, which is not always possible.

While the maximum trait difference R_{\max} is a plausible observable on the basis of

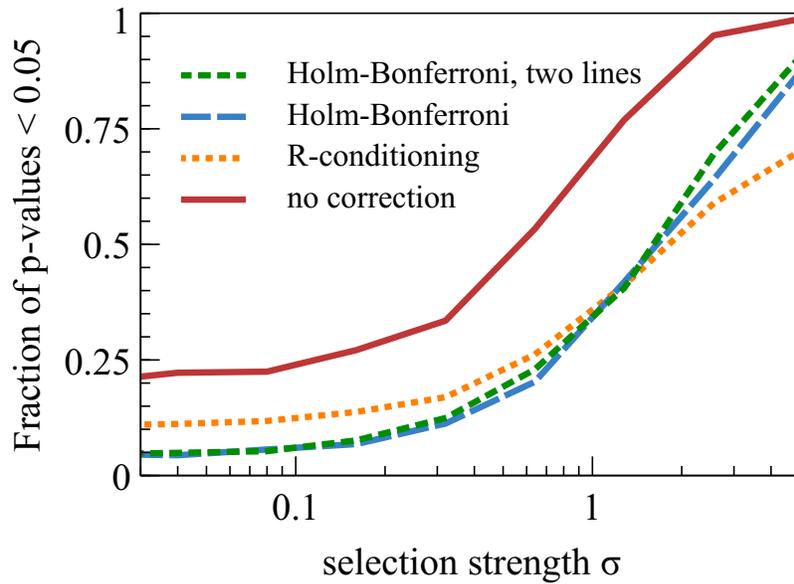


Figure 4.3: **True positive rate of different methods dealing with the multiple testing problem.** While the true positive rate is highest without any multiple testing correction, it also has the highest false positive rate of 0.20 (which can be read off as the value on the very left of the plot, where the selection strength σ tends to zero, corresponding to the neutral case). The Holm-Bonferroni correction works both for two and three lines and has a higher true positive rate at high selection strength than the conditioning on the phenotypic difference. The higher true positive rate of the conditioning on R_{\max} for low values of σ is only achieved along with a higher false positive rate. The conditioning on R_{\max} for two lines is not possible and would only result in a flat line at the false positive threshold since selection and the conditioned neutral distribution cannot be distinguished.

which traits can be selected from a larger pool, it is by no means the only one. For instance, with three lines, traits could in principle be selected based on the difference between the trait in line 1 and the trait mean in line 2 and 3, $R_\Delta = T_1 - \frac{T_2+T_3}{2}$. One would use this observable when looking specifically for traits with lineage-specific selection acting on line 1. For $s_2 = s_3 \equiv s_0$ the fitness (3.2) can be written

$$F(T_1, T_2, T_3) = \bar{s}(T_1 + T_2 + T_3) + \hat{s} \left(T_1 - \frac{T_2 + T_3}{2} \right) \quad (4.13)$$

with $\bar{s} = (s_1 + 2s_0)/3$ and $\hat{s} = s_1 - \bar{s} = 2(s_1 - s_0)/3$. The maximum entropy distribution conditioned on R_Δ is proportional to $\exp(hR_\Delta)$ and thus again differs from the equilibrium distribution $\propto \exp(\beta F(T_1, T_2, T_3))$, except in the special case $\bar{s} = 0$.

Chapter 5

Short evolutionary times

In this chapter, I consider the limit of short evolutionary times as opposed to the equilibrium model in chapter 3. But the assumption of long evolutionary times, such that the state statistics is in equilibrium, might not always be valid. Because of this and because it is usually difficult to consider arbitrary evolutionary times (where the evolutionary times and mutation rates have to be known), this limit of short evolutionary times is worthwhile considering. The limit of short evolutionary times is characterized by $n\mu t \ll 1$, which denotes the average number of mutations per locus (where n is the number of lines, μ the mutation rate per locus and generation, and t the number of generations since the divergence of the lines). In this regime, most loci are not diverged between the lines and the loci that are diverged have undergone a single mutation. Nevertheless, the total number of diverged loci $n\mu tL$ should be at least of order one.

In the short-time limit the state statistics also depends on the phylogenetic tree of the lines, which is assumed to be known. Since the mutation affecting one locus can happen at different branches of the phylogenetic tree, its probability of occurrence depends also on the length of the branch, which is proportional to the evolutionary time. Also not all state configurations (q_1, q_2, q_3) can be reached from the ancestral state c , which is the state of the locus in the common ancestor of the lines, by a single mutation. As the size of the phylogenetic tree and especially the number of possible trees grow very quickly with the number of lines, it is infeasible to calculate the state statistics for general number of lines n . Here, I only consider the cases $n = 2$ and $n = 3$. In the following, I assume the phylogenetic tree for three lines as defined in Figure 5.1.

5.1 Two lines

First, I consider the case of two lines. Here, both possible diverged final states for a locus $(q_1, q_2) = (+, -)$ and $(-, +)$ can be reached from either ancestral state $c = \pm$.

The transitions between states of a locus (resulting in the fixation of the new state at that locus in the population) with additive effect a in a population of size N under the influence of a selection strength s occur with transition rates

$$u_+(Ns, a) = \mu_+ \frac{2Nsa}{1 - e^{-2Nsa}} \quad (5.1)$$

for the transition from ancestral state $c = -$ to state $+$ and

$$u_-(Ns, a) = \mu_- \frac{-2Nsa}{1 - e^{2Nsa}} \quad (5.2)$$

for the transition from ancestral state $c = +$ to state $-$. This result follows from eq. (2.3) when ΔF is assumed to be small. Here, μ_+ and μ_- denote the mutation rates for the transitions $- \rightarrow +$ and $+ \rightarrow -$, respectively. In the absence of selection a difference in the mutation rates can still lead to a bias in the probability for a transition to the $+$ or $-$ state, much like the multiplicity factor Ω in the case of long evolutionary times. Since the mutation rates μ_+ and μ_- are not defined for individual nucleotides but as effective mutation rates between the states of a locus that can consist of many nucleotides, a bias between these rates can be very likely, as already explained for the case of long evolutionary times. Therefore, I again introduce a parameter Ω that, similar to the multiplicity factor in the case of long evolutionary times, quantifies the bias towards the one or the other transition under neutral evolution. Since only the relative difference $\frac{\mu_+}{\mu_-}$ between the two mutation rates counts in the remainder of the calculation, one parameter is enough to encode the information about the bias in mutation rates. Without loss of generality, I set μ_- to 1 and replace μ_+ with e^Ω . This yields the relative transition rates

$$u_c(Ns, a, \Omega) = e^{\Omega(1-c)/2} s_c(Ns, a) \quad (5.3)$$

with

$$s_c(Ns, a) = \frac{-2Nsa}{1 - e^{2Nsa}} \quad (5.4)$$

for an ancestral state $c = \pm 1$.

The probability for a transition to occur in line i for a given ancestral state c follows now as

$$P(i|Ns_1, Ns_2, a, c) = \frac{u_c(Ns_i, a, \Omega)}{u_c(Ns_1, a, \Omega) + u_c(Ns_2, a, \Omega)} = \frac{s_c(Ns_i, a)}{s_c(Ns_1, a) + s_c(Ns_2, a)}, \quad (5.5)$$

since the mutational bias Ω included in $u_c(a, Ns, \Omega)$ cancels out. The evolutionary times of the lines since the ancestor do not enter the equations for two lines, since in this case only a trivial phylogenetic tree with two equally long branches is possible and the time parameter t cancels out due to the normalization.

Often it might not be possible to infer the ancestral states c for the loci. In this case one can average over both possible ancestral states $c = \pm$ with an appropriate prior. Here, I make the assumption that the state statistics of the ancestral line was in equilibrium before the divergence of the two lines. Without any selection acting on the ancestral line the bias in mutation rates alone would influence the state statistics. For two mutation rates μ_+ and μ_- the probability for an ancestral state c would be $P(c) = \frac{\mu_c}{\mu_+ + \mu_-} = \frac{1}{1 + \frac{\mu_- c}{\mu_+}}$. Additionally, the ancestral line can be subject to selection with selection strength s_{anc} , which gives rise to a factor of $e^{Ns_{\text{anc}}a}$. Switching to the notation with the Ω parameter, I obtain

$$P_{\text{anc}}(c, Ns_{\text{anc}}, \Omega) = \frac{1}{1 + e^{-c\Omega}} e^{Ns_{\text{anc}}a}. \quad (5.6)$$

With this result for the prior probability, one can average the probability to obtain one of the possible states $(q_1, q_2) = (+, -)$ and $(-, +)$ over the ancestral states. One obtains

$$P(q_1, q_2 | Ns_1, Ns_2, a, \Omega) = \frac{P_{\text{anc}}(q_2, Ns_{\text{anc}}, \Omega) u_{q_2}(Ns_1, a, \Omega) + P_{\text{anc}}(q_1, Ns_{\text{anc}}, \Omega) u_{q_1}(Ns_2, a, \Omega)}{\sum_{c=\pm 1} \sum_{i=1}^2 P_{\text{anc}}(c, Ns_{\text{anc}}, \Omega) u_c(Ns_i, a, \Omega)}. \quad (5.7)$$

Contrary to equation (5.5) both ancestral states $c = \pm$ enter the equation, such that Ω does not cancel out directly. Yet, the terms $P_{\text{anc}}(c, Ns_{\text{anc}}, \Omega)$ and $u_c(Ns, a, \Omega)$ appear only in combination in eq. (5.7). As it turns out, the dependency on Ω is the same for both $c = +$ and $c = -$. One gets $P_{\text{anc}}(\pm, Ns_{\text{anc}}, \Omega) u_{\pm}(Ns, a, \Omega) \propto \frac{e^{\Omega}}{1 + e^{\Omega}}$, such that the dependency on Ω cancels out also in the case of unknown ancestral states. The reason why Ω cancels out in this case as well is easily understood: if there is a bias towards the $+$ state, the probability to start with $c = +$, $P_{\text{anc}}(+, Ns_{\text{anc}}, \Omega)$, is high, but the transition probability $u_+(Ns, a, \Omega)$ to the $-$ state is low. If the bias is towards the $-$ state, the probability to start with $c = +$ is low, but the transition probability to the $-$ state is high, resulting in the same probability for both cases.

Thus, I can write the two-line result as

$$P(q_1, q_2 | Ns_1, Ns_2, a) = \frac{P_{\text{anc}}^{\text{two}}(q_2, Ns_{\text{anc}}) s_{q_2}(Ns_1, a) + P_{\text{anc}}^{\text{two}}(q_1, Ns_{\text{anc}}) s_{q_1}(Ns_2, a)}{\sum_{c=\pm 1} \sum_{i=1}^2 P_{\text{anc}}^{\text{two}}(c, Ns_{\text{anc}}) s_c(Ns_i, a)}, \quad (5.8)$$

which is now independent of Ω . Here, I define $P_{\text{anc}}^{\text{two}}(c, Ns_{\text{anc}}) = e^{Ns_{\text{anc}}a}$.

The distribution (5.8) conditioned on the phenotypic difference can now be obtained analogously to the equilibrium case in chapter 4. The result $P(q_1, q_2 | Ns_1, Ns_2, a, R_{\text{max}})$

has again introduced a new term $\exp(ha(q_1 - q_2))$ leading to the trait difference R_{\max} by adjusting the parameter h . But now, for the short-time dynamics, this result for the neutral scenario under conditioning on R_{\max} has not the same form as the selective scenario anymore, unlike the case of long evolutionary times (see section 4.2.2). Thus, it is in principle possible to distinguish neutral evolution and selection in this case. Yet, as the numerical simulations in section 5.3 show, the test has barely any statistical power in this case.

5.2 Three lines

In the case of three lines, the phylogenetic tree of the lines becomes important as well. Two lines will be more closely related while a third line is the outgroup. In the following I define line 3 as the outgroup, resulting in a phylogenetic tree as given in Figure 5.1. There are two time intervals t_1 and t_2 denoting the times between the speciation events and an additional selection strength s_{12} for the ancestor of the two more closely related lines 1 and 2 (see Figure 5.1).

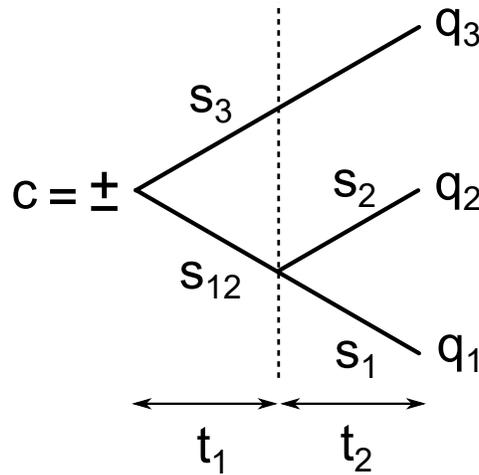


Figure 5.1: **Phylogenetic tree for three lines.** In the short-time limit the phylogenetic tree is important for the calculation of the state statistics. The branch lengths t_1 and t_2 , the ancestral state c , and the selection strengths s_1 , s_2 , s_3 , and s_{12} determine the relative mutation probabilities in the different branches.

For four of the six possible diverged configurations a unique ancestor can be assigned: Denoting line 3 as the outgroup, configurations $(q_1, q_2, q_3) = (+, -, -)$ and $(-, +, -)$ diverged from the ancestral state $c = -$, configurations $(-, +, +)$ and $(+, -, +)$

from ancestor $c = +$. Configurations $(+, +, -)$ and $(-, -, +)$ can either be reached by a mutation in the ancestor of lines 1 and 2 or a mutation in line 3 (see also Figure 5.2).

Under neutral evolution the relative probability for a state configuration (q_1, q_2, q_3) depends only on the length of the branch in the phylogenetic tree (which is proportional to the passed evolutionary time) in which the mutation occurs. The multiplicity factors Ω_l of the loci would in principle also play a role, but they again cancel out, as described in the previous section for the case of two lines. The relative probability for the transition from $c = +$ to $(-, -, +)$ for example is proportional to t_1 since the mutation can only happen in the ancestor of lines 1 and 2, while the configuration $(+, -, +)$ can only be reached by a mutation after the divergence. Thus the relative probability is proportional to t_2 in that case.

With selection included the relative transition probabilities also depend on the selection strength in the different branches. In the most general case, there are the selection strength s_1, s_2, s_3, s_{12} (see Figure 5.1) and s_{anc} (for the ancestral line).

As already for the two-line result in eq. (5.8), the relative transition rates for the different state configurations are now composed of the branch length of the phylogenetic tree in which the mutation happens, the probability $P_{\text{anc}}(c, N s_{\text{anc}})$ for ancestral state c and the transition rates $s_c(Ns, a)$ for that change of state. Here, I again assume that the ancestral states are unknown and need to be averaged over. A straightforward extension to the case of known ancestral states is easily possible. For the state configurations $(-, -, +)$ and $(+, +, -)$ that can be reached from either ancestral state the relative probabilities for both transitions from $c = +$ and $c = -$ are added up, weighted with their probability of occurrence $P(+, N s_{\text{anc}})$ and $P(-, N s_{\text{anc}})$, respectively. In total I obtain

$$\begin{aligned}
(-, -, +) & \quad (t_1 + t_2)P(-, N s_{\text{anc}})s_-(N s_3, a) + t_1P(+, N s_{\text{anc}})s_+(N s_{12}, a) \\
(-, +, -) & \quad t_2P(-, N s_{\text{anc}})s_-(N s_2, a) \\
(+, -, -) & \quad t_2P(-, N s_{\text{anc}})s_-(N s_1, a) \\
(+, +, -) & \quad t_1P(-, N s_{\text{anc}})s_-(N s_{12}, a) + (t_1 + t_2)P(+, N s_{\text{anc}})s_+(N s_3, a) \\
(+, -, +) & \quad t_2P(+, N s_{\text{anc}})s_+(N s_2, a) \\
(-, +, +) & \quad t_2P(+, N s_{\text{anc}})s_+(N s_1, a), \tag{5.9}
\end{aligned}$$

for the relative probabilities of the state configurations. The absolute probabilities are obtained by normalizing by the sum of all relative probabilities. As in the case of two lines, only combinations of $P(c, N s_{\text{anc}})s_c(Ns, a)$ appear, such that the multiplicity factors Ω_l again cancel out. One can write the short-time probability distribution

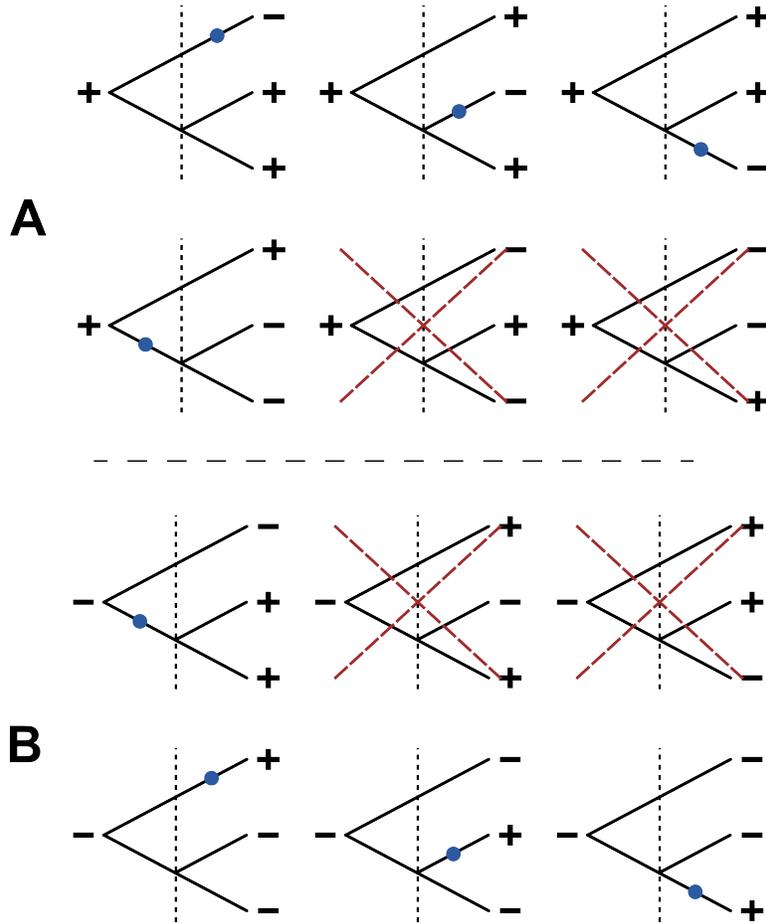


Figure 5.2: **Allowed transitions for three lines.** The phylogenetic tree for three lines is constraining the three-line states that can be reached by a single mutation from ancestral state c . (A) Four of the six possible diverged states can be reached from ancestral state $c = +$, while states $(-, +, -)$ and $(+, -, -)$ cannot be reached by a single mutation. (B) While the two states $(+, +, -)$ and $(-, -, +)$ (where the outgroup line is in a different state) can also be reached from the ancestral state $c = -$, two other states $(-, +, +)$ and $(-, +, -)$ that could be reached from $c = +$ are now inaccessible. Thus, four of the six configurations can be assigned to a unique ancestral state.

$Q_s(q_1, q_2, q_3|a)$ in a more compact form as

$$Q_s(q_1, q_2, q_3|a) = \frac{1}{Z} \left(t_2 P(k, N_{s_{\text{anc}}}) s_k(N_{s_{q=-k}}, a) + \delta_{q_1, q_2} t_1 \cdot [P(k, N_{s_{\text{anc}}}) s_k(N_{s_3}, a) + P(-k, N_{s_{\text{anc}}}) s_{-k}(N_{s_{12}}, a)] \right). \quad (5.10)$$

Here, I define the shorthand $k = q_1 + q_2 + q_3 = \pm 1$, and $N_{s_{q=-k}}$ denotes the selection strength of the line with the minority state (e.g. N_{s_3} for the configuration $(-, -, +)$) and $Z = \sum'_{q_1, q_2, q_3 = \pm 1} Q_s(q_1, q_2, q_3|a)$. Again, the two states with $q_1 = q_2 = q_3 = \pm 1$ are excluded from this sum.

The neutral distribution $P_s(q_1, q_2, q_3|a)$ is straightforwardly derived from eq. (5.10) by setting all selection parameters to zero. This leads to $P(c, 0) = 1$ and $s_c(a, 0) = 1$. Since the multiplicity factors cancel out, the only dependency that is left is the one on the branch length of the phylogenetic tree. One obtains

$$P_s(q_1, q_2, q_3) = \frac{t_2 + \delta_{q_1, q_2} 2t_1}{6t_2 + 4t_1}. \quad (5.11)$$

Again, the distribution (5.10) conditioned on the maximal phenotypic difference R_{max} can be obtained as before, yielding an additional term $\exp(ha(q_1 - q_2))$ to fix R_{max} .

5.3 Statistical power of the short-time test

Now eq. (5.10) and (5.11) can be used to compare different selective scenarios for short times using the log-likelihood score (3.10). For this, the parameters of the scenarios are estimated by maximum likelihood, where one has the new selection parameter s_{anc} of the ancestral line but no more multiplicity parameters Ω_l . The evolutionary times t_1 and t_2 are assumed to be known and will not be estimated by maximum likelihood.

To explore the statistical power of the selection test in the short-time regime, I perform numerical simulations as in section 3.5 but with the short-time scenario (5.10) applied on simulated data created under a short-time dynamics. Similarly to the equilibrium case I define the selective scenario Q_1 with selection parameters $(s_1, s_2, s_3, s_{12}, s_{\text{anc}}) = (s, 0, 0, 0, 0)$, the neutral scenario P_0 with selection parameters $(0, 0, 0, 0, 0)$ as well as the two-line scenarios Q_{two} with $(s_1, s_2, s_{\text{anc}}) = (s, 0, 0)$ and P_{two} with $(0, 0, 0)$. The evolutionary timescales are chosen as $t_1 = t_2$ (since only the relative values of t_1 and t_2 play a role). For better comparability to the equilibrium case I consider sets of $L_{\text{div}} = 10$ diverged loci drawn from the short-time state statistics (5.10). The additive effects $\{a_l\}$ are again drawn from a gamma distribution with parameters $\alpha = 2$ and $\beta = 20$ (mean effect $a = 0.1$ per locus).

Figure 5.3 shows that both the three-line and the two-line version of the short-time test are able to distinguish selection and neutral evolution. The statistical power of the three-line test is somewhat lower in the short-time case, yet the test works in a reasonable parameter range. For two lines it is effectively impossible to detect selection using the test incorporating the conditioning on the phenotypic difference R_{\max} . So even though neutral and selective state statistics are now different, unlike in the case of long evolutionary times, the difference turns out to be too small to utilize it for the selection test.

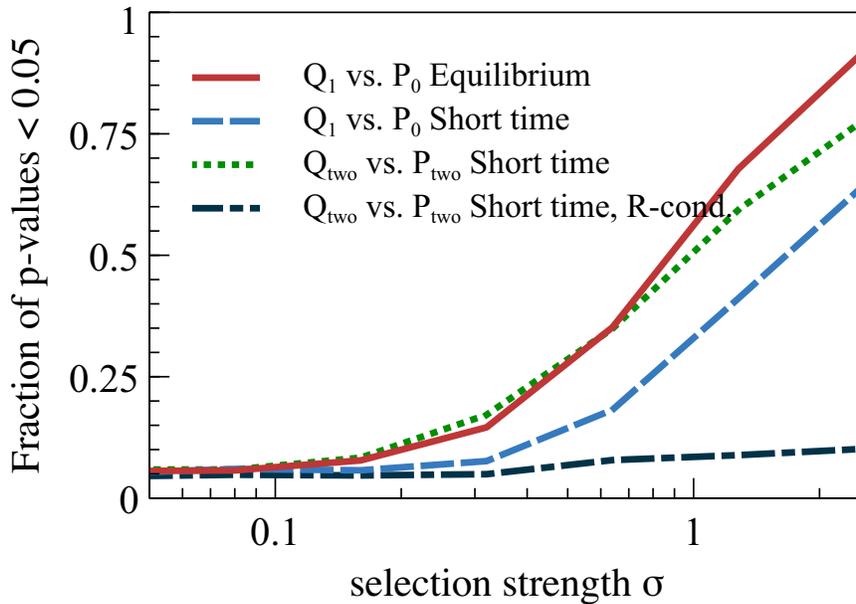


Figure 5.3: **Statistical power of the short-time selection test.** The statistical power of the short-time selection test applied to simulated data under the short-time dynamics is compared to the equilibrium selection test applied to simulated data under the long-time dynamics. Again, the fraction of significant scores ($P \leq 0.05$) for configurations drawn under selective scenario Q_1 are plotted against the selection strength σ . The curve for the comparison of scenarios Q_1 and P_0 in equilibrium for three lines is taken from Figure 3.7. The comparison of three-line scenarios Q_1 and P_0 under the short-time dynamics gives a somewhat lower but similar significance. This shows that also for short evolutionary times and without the knowledge of the ancestral states of the loci (which is not assumed here) the inference of selection is possible. Also in the case of two lines the comparison of Q_{two} and P_{two} the statistical power rises quickly with selection strength σ . For the case of two lines with conditioning on the phenotypic difference the selective scenario now gives a different result than the neutral scenario, unlike in the case of long evolutionary times. Yet, the statistical power is only marginal in this case, making the conditioned two-line test infeasible.

Chapter 6

Selection on Plant Quantitative Traits

In this chapter I will apply the multiple-line selection test on several datasets. To probe the feasibility of the test under realistic circumstances I will use data from two plant quantitative trait studies, one comparing four different maize lines and one comparing three different lines of the genus *Mimulus*. In both studies the environmental differences behind the divergence of the traits under investigation are known and different selection pressures among the lines may reasonably be expected. In the next chapter I will apply the selection test to a far bigger study of expression QTL (eQTL) in three lines of *Schizosaccharomyces pombe*, where thousands of expression levels serve as traits, to infer which expression levels are under selection.

6.1 Maize photoperiodic response traits

First, I investigate a QTL study comparing four different maize lines (*Zea mays* L. subsp. *mays*) (Coles et al. 2010). Two of the maize lines (CML254, Ki14) originate from tropical regions while the other two lines (B73, B97) originate from temperate regions. While the tropical lines encounter a short-day environment with relatively stable day length over their growth period, the temperate lines are used to a long-day environment. In the study, both the tropical and temperate lines were grown in both environments. While the temperate lines are not affected by day length and have an equally long growth period in both environments, the tropical lines react to the long-day environment with an increased length of the growth period and a higher number of leaves compared to the short-day environment. In their study, Coles et al. (2010) investigated the genetic basis of photoperiodic response of several traits, which is defined as the difference in trait value between short- and long-day environment, using QTL analysis. In particular, they measured the 'growing degree days to anthesis'

(GDDTA) and the 'growing degree days to silking' (GDDTS). 'Growing degree days' is a measure for the growth time weighted by the temperature for each day, which accounts for a higher growth rate at higher temperatures. Anthesis corresponds to the time point of completed flower development and silking corresponds to the time point where the silk fibres emerge in maize. Additionally, the traits plant height (PH), ear height (EH; the ear holds the maize kernels), and total number of leaves (TLN) were measured.

For maize it has been shown that the architecture of quantitative traits such as flowering time and leaf size accurately follows a model with additive trait effects and only weak epistatic effects (Buckler et al. 2009; Tian et al. 2011). In Coles et al. (2010), the additive effect of alleles from different QTL and the corresponding experimental errors are given. For each locus it is specified which lines harbour an allele with the same effect on the trait (within experimental error). To obtain the states $\{q_{i,l}\}$ for each QTL in the different lines I determine all QTL which show alleles that have one of two experimentally distinguishable effects on the trait. In those cases, I can unambiguously assign the + state to the lines with the higher trait value and the - state to the allele with the lower trait value.

While for the majority of loci the two-state assumption holds true, about one third of the QTL did not fall into the two-state pattern as they either have more than two significantly different alleles or have an unclear assignment of alleles (for example line 1 has allele a , line 2 has the distinct allele b , and the allele of line 3 is consistent with either allele a or b within the measurement uncertainty of the trait values). The additive effects of the two-state loci are determined as the mean of the absolute values of the trait effects of all lines (including both the + and the - states). The resulting states $\{q_{i,l}\}$ and additive effects $\{a_l\}$ can be found in Table 6.1. Here, the two traits EH and TLN are not included since they both had less than four two-state QTL, which is not enough to test for selection.

To test for selection I define the four line selective scenario Q_4 with ($N_{SB73} = Ns$, $N_{SB97} = Ns$, $N_{SCML254} = +Ns$, $N_{SKi14} = +Ns$), that has a single selection parameter. In this scenario selection increases the trait value in the tropical lines and decreases the trait value in the temperate lines. To account for a possible ascertainment bias, I compare selective scenario Q_4 both against the neutral model P_h conditioned on the trait difference R_{\max} and the neutral model P_0 together with a Holm-Bonferroni correction. For each scenario, I calculate the multiplicity parameters Ω_l and (for the selective scenario Q_4) selection strength s according to eq. (3.9). For scenario P_h I condition the null model on the pair of lines with the highest trait difference for each trait. As the total number of traits is not known, I assume here a number of $m = 6$ traits for the Holm-Bonferroni correction, which corresponds to the number of traits tested in Coles (2010). However, one has to be careful with this point, since the possible pool of traits might be larger. Wherever necessary I use the BIC correction (3.11) to

additive effect a_l	B73	B97	CML254	Ki14
GDDTA [GDD]:				
4.73	–	–	+	+
3.85	–	–	+	–
4.43	–	–	+	–
11.13	–	–	+	+
GDDTS [GDD]:				
6.33	–	–	+	+
6.20	–	–	+	–
4.68	+	–	+	+
5.68	–	–	+	+
plant height [cm]:				
1.10	–	–	+	+
1.25	+	+	–	–
1.73	+	–	+	+
2.10	–	+	+	–

Table 6.1: Additive QTL effects for three quantitative traits in the four maize lines B73, B97, CML254, and Ki14 estimated from (Coles et al. 2010): growing degree day to anthesis (GDDTA), growing degree day to silking (GDDTS), and plant height.

correct the log-likelihood scores (3.10) when the number of free parameters is different between the tested scenarios (as is the case when testing Q_4 against P_0). Note that this leaves the p -values unaffected.

First, I consider the GDDTA trait, which measures the time to full flower development. For tropical lines, which are not adapted to long day length, the flowering time is reduced for specimens grown in temperate latitudes compared to tropical environments (Coles et al. 2010). For the temperate lines no difference in flowering time is observed between the different environments. For this trait, 4 out of 7 loci show a clear two-state pattern. I first compare scenario Q_4 against the neutral scenario P_h conditioned on trait difference R_{23} between lines 2 and 3. In this case, the straightforward maximum likelihood estimate of the parameter h fails, since all states in the high line are + states and all states in the low line are – states, leading to a diverging $h \rightarrow \infty$. I use a lower-bound estimate for h by determining the value h for which the probability to see this extreme configuration equals p_e . $p_e = 0.1$ is chosen to obtain a conservative estimate for h . For consistency, Ns is determined in the same way. In this case the log-likelihood score (3.10) $S_{Q_4, P_h} = 2.77$ is positive and thus in favour of the selective scenario. I test the significance of this score by repeated simulations under scenario P_h at fixed additive effects $\{a_l\}$. For each configuration drawn from P_h I sort the lines according to their trait values T . In this way one can account for the possibility that

under neutral evolution fluctuations create patterns of lineage-specific selection in any of the lines (rather than only in what is called line 1 here). This results in a p -value of $p = 0.07$. Similarly, I test Q_4 against P_0 , which results in a score $S = 5.1$ and a p -value of $p = 0.048$ consistent with the result for P_h . Here, the more restrictive p -value threshold of $p < 0.05/m = 0.01$ for the Holm-Bonferroni correction has to be applied (here $m = 5$ has to be used, as this will be the smallest p -value of all traits considered).

Next, I test the GDDTS trait, where 4 of 6 loci show a two-state pattern. The test of scenario Q_4 against P_h results in a score $S_{Q_4, P_h} = 2.77$, and a p -value of 0.048 again in favour of the selective scenario. Again P_h is conditioned on R_{23} and the lower bound for h is used as described above. The test of Q_4 against P_0 gives a score $S_{Q_4, P_0} = 5.1$ with a corresponding p -value $p = 0.03$, which is consistent with the test under conditioning. Last, the plant height trait is tested, where 4 out of 6 loci are two-state loci. Here, both the conditioned test ($S_{Q_4, P_h} = -0.61$, $p = 0.42$) and the unconditioned test ($S_{Q_4, P_0} = -0.9$, $p = 0.978$) are in favour of the neutral scenario. The results are summarized in Table 6.3.

For comparison I also apply Orr's sign test (Orr 1998) (not the equal effects version) to this dataset. Since the Orr test is a two-line test, I use for each trait the two lines with the largest trait difference for testing, where one would expect the strongest signal for selection. Following Orr, the additive effects $\{a_l\}$ are taken from a gamma distribution whose parameters for each trait are estimated by maximum likelihood. Then the probability to find at least the observed number of + states in the high line given the observed trait difference R or greater is calculated according to eq. (4) in Orr (1998). Since I can use all loci that are diverged between two lines (ignoring possible inconsistencies or additional alleles in the other lines and thus also taking into account the loci that violate the two-state assumption when considering all lines), one can actually use more loci than used for the four line test. For both the GDDTA and the GDDTS trait the lines B73 and CML254 show the largest trait difference. For the GDDTA trait 6 out of 6 diverged loci have the + state in the tropical line CML254 (see again Table 6.1). Here, the Orr test gives a p -value of $p = 0.13$. For the GDDTS trait, 5 out of 5 diverged loci go in the + direction with a p -value of $p = 0.2$. The plant height trait has only 2 out of 3 diverged loci in the + state for the line pairs B73 and CML254 as well as B97 and CML254 with $p = 0.99$. So no clear signal of selection can be found using the Orr test.

6.2 *Mimulus* floral traits

As a second example I consider a QTL study comparing three different plant species of the genus *Mimulus* (Chen 2009). For the species *M. guttatus*, *M. platycalyx*, and *M. micranthus*, which I label as lines 1, 2 and 3, respectively, different floral characters

were measured. For each of the loci influencing the traits it turns out that there were two lines with a very similar effect on the trait value (within experimental error) and one with a significantly different effect. I assigned the same + (−) state to the two lines with the similar allelic effect if they had the higher (lower) trait value than the third line. The allelic effects were assigned analogously to the case of the maize traits in the previous section. Here, I restrict myself to the two traits with the highest number of available QTL, the corolla width and the corolla length trait. The resulting state configurations $\{q_{i,l}\}$ and additive effects a_l can be found in Table 6.2.

additive effect a_l	<i>M. gutt.</i>	<i>M. platy.</i>	<i>M. micr.</i>
corolla width [mm]:			
0.41	−	+	−
0.74	+	−	−
0.39	+	−	+
0.59	+	−	−
0.28	+	+	−
0.64	+	−	−
0.37	+	−	+
corolla length [mm]:			
0.67	+	−	−
0.41	+	+	−
0.21	+	−	+
0.60	+	−	−
0.27	−	+	+
0.51	+	+	−

Table 6.2: Additive QTL effects for two flower traits of the *Mimulus* species *M. guttatus*, *M. platycalyx* and *M. micranthus* estimated from (Chen 2009).

As with the maize traits before, I test the different scenarios on the *Mimulus* traits, where I use scenario Q_1 (as defined in section 3.4.1) as the selective scenario. This time the number of traits in the study was $m = 5$. I start with the corolla width trait where 7 QTL are available. In this case I condition scenario P_h on the maximal trait difference R_{12} between lines 1 and 2. The comparison of scenarios Q_1 and P_h yields a log-likelihood score of $S_{Q_1, P_h} = 0.38$ with a corresponding p -value $p = 0.13$ slightly favouring the selective scenario. The unconditioned test yields a score of $S_{Q_1, P_0} = 1.9$ and a p -value of $p = 0.05$, in good agreement with the results of the conditioned test.

A preference for a selective model is in agreement with the different reproductive modes of these species (Chen 2009): line 1 reproduces predominantly by outcrossing (so that large floral characters are needed to attract pollinators), whereas line 2 and line 3 are mostly self-pollinating (but still maintain a certain degree of outcrossing). In

the latter species, large petals are less indispensable for reproduction, but nevertheless require resources to develop and maintain.

The second trait is the corolla length trait with 6 known QTL and the maximal trait difference is R_{13} between lines 1 and 3. Here, the test gives $S_{Q_1, P_h} = -0.43$ and $p = 0.54$, so the neutral hypothesis cannot be rejected. A similar result is obtained for the Holm-Bonferroni procedure ($S_{Q_1, P_0} = 1.0$, $p = 0.14$). The results are summarized in Table 6.3.

For comparison, I again apply the Orr test on the two lines with the highest trait differences as done for the maize traits. For the corolla width trait 5 out of 6 loci in lines 1 and 2 have the + state. Here, the Orr test yields a p -value of $p = 0.42$. Comparing lines 1 and 3 (with also 5 out of 6 + loci but with different additive effects) gives $p = 0.29$. For the corolla length trait there are 4 out of 5 diverged loci with the + state between lines 1 and 3 and 3 out of 4 diverged loci between lines 1 and 2. The Orr test yield p -values $p = 0.48$ and $p = 0.72$, respectively. Thus, no sign for selection can be found for these traits using the Orr test.

In both case studies, the statistical significance of the evidence for a particular evolutionary scenario is limited by the number of identified trait loci. With a higher number of crosses in the original studies, identifying more trait loci, a stronger statistical signal would be possible.

maize study	Ns_4	S	p
GDDTA			
Q_4 vs. P_h	40	2.8	0.07
Q_4 vs. P_0	40	5.1	0.045
GDDTS			
Q_4 vs. P_h	27	2.8	0.048
Q_4 vs. P_0	27	5.1	0.030
plant height			
Q_4 vs. P_h	0.77	-0.61	0.42
Q_4 vs. P_0	0.77	-0.90	0.978
<i>Mimulus</i> study			
<i>Mimulus</i> study	Ns_1	S	p
corolla width			
Q_1 vs. P_h	2.2	0.38	0.13
Q_1 vs. P_0	2.2	1.9	0.05
corolla length			
Q_1 vs. P_h	2.2	-0.43	0.54
Q_1 vs. P_0	2.2	1.0	0.14

Table 6.3: Summary of results for the QTL data of the maize (Coles et al. 2010) and *Mimulus* (Chen 2009) studies. Different evolutionary scenarios are tested against each other using both conditioning on the trait difference (P_h) as well as the Holm–Bonferroni correction (P_0). Ns_1 and Ns_4 denote the inferred selection strengths of the selective scenarios Q_1 and Q_4 , S is the log-likelihood score obtained and p the corresponding p -value. In *Mimulus*, corolla width shows some evidence of selection and in maize the photoperiodic response traits GDDTA and GDDTS.

Chapter 7

Gene expression evolution in the yeast *S. pombe*

The role of gene expression levels as major drivers of adaptation and speciation has been hypothesised a long time ago (King and Wilson 1975). Yet, providing evidence for adaptation on specific genes or pathways has proven to be a non-trivial task. In this chapter, I will study adaptation for stress resistance on gene expression levels of individual genes as well as protein complexes in yeast. I apply the selection test to an eQTL study of three different lines of the species *Schizosaccharomyces pombe*. In particular, I focus on the differences in gene expression between stress and non-stress condition. The eQTL dataset provides a large amount of eQTL per gene, allowing to test for selection both on individual genes and on larger gene modules with clearly defined biological function. I find signatures of pervasive selection on genes and gene modules that are connected to biological mechanisms involved in stress response.

The yeast species *S. pombe* is a model organism for eukaryotic molecular biology that has been extensively studied. It is only distantly related to the most commonly used yeast model organism *Saccharomyces cerevisiae* with a divergence time of ~ 400 million years and with many genes of *S. pombe* more similar to their mammalian homologues than to *S. cerevisiae* (Sipiczki 2000). Nowadays, a diverse collection of wild isolates from all over the world is available with samples of *S. pombe* mostly gathered from fruits, alcoholic beverages, and fermentation processes (Brown et al. 2011). Yet, most of the research on *S. pombe* was conducted on the standard laboratory strain 968 (Leupold 1950) that does not capture the full ecological diversity of the species. An analysis of the evolution of the different isolates of the organism can provide insight about the forces of evolution acting on *S. pombe* and about the genetic targets that mediate the adaptation to different environments.

7.1 Yeast dataset

7.1.1 Available data

I use data from a QTL experiment on crosses between all possible pairs of three different strains of *S. pombe* (Clément-Ziza et al. unpublished dataset). The following three strains were used in the QTL experiment: the standard laboratory strain 968 and the two wild isolates Y0036, which was isolated from South African beverage, and DBVPG2812, which was isolated from Italian grape must (Jeffares et al. 2015; Brown et al. 2011). In the following I label these lines as line 1, 2, and 3, respectively. In the QTL study (Clément-Ziza et al. unpublished dataset) the traits considered were the expression levels of the of the different *S. pombe* strains, with expression QTL (eQTL) having an impact on the level of expression of the genes in the different strains. 46 crosses were performed between line 1 and 2, 44 crosses between lines 1 and 3, and 43 crosses between lines 2 and 3.

In their study, the gene expression levels of all crosses were measured using RNA-seq under two different conditions, first under oxidative stress through the addition of H_2O_2 and second without any stress applied. This allows to study the changes in gene expression levels under stress and to identify genes involved in the stress response that have specifically adapted in one or more of the lines.

The following data are available from the experiment:

- **Expression data:** For each of the crosses as well as for individuals from the pure lines the expression data of 5260 genes were measured using RNA-seq. RNA-seq is a next-generation sequencing technique that measures the abundance of RNA molecules in a sample, which represent the expression levels of the genes.
- **Genotype:** A total of 1693 genetic markers (SNPs) were analyzed. For each cross one can determine from which line a marker is inherited.
- **eQTL:** The QTL mapping looks for statistical correlations between markers and expression levels. If a high expression level is always observed when the genotype of a given marker is inherited from the first line, the probability is high that a QTL close to this marker affects that expression level (see also section 2.2.2). A list of eQTL that are significant for a certain q -value threshold is given. The q -value denotes the false discovery rate (FDR) for a particular gene-marker pair. The number of available QTL at different false discovery rates can be found in Table 7.1.

Clement-Ziza et al. used the expression data under both conditions for the simultaneous QTL mapping as described in Ackermann et al. (2013) that implements the

FDR	static QTL	cond. stress QTL	cond. non-stress QTL
0.05	2527	963	239
0.1	3932	1389	341
0.15	5511	1754	456
0.2	7173	2034	569
0.25	9079	2351	670

Table 7.1: Number of significant QTL at different levels of the QTL false discovery rate (FDR) for the three different QTL categories of the *S. pombe* dataset: static (QTL detected in both conditions), 'conditional stress' (detected in stress condition only) and 'conditional non-stress'. The static QTL category contains most QTL, while the category of 'conditional non-stress' QTL is the smallest category. The number of detected QTL drastically increases with an increasing FDR.

random forest technique (Michaelson et al. 2009; Michaelson et al. 2010). This technique has the advantage that it strongly increases the statistical power to detect QTL that are present under both conditions.

This extraordinarily powerful QTL mapping of (Clément-Ziza et al. unpublished dataset) yields a much higher number of QTL per gene and allowing to test individual expression levels for selection, which was not possible before (Fraser et al. 2011).

The simultaneous QTL mapping in both conditions also allows to distinguish different kinds of QTL concerning the two different conditions (Ackermann et al. 2013): First, static QTL, which can be detected in both stress and non-stress condition with the same effect strength. Second, conditional QTL that are detected in only one of the conditions and are thus either specific to the stress- or the non-stress condition. This classification of QTL allows to analyze specific subsets of QTL, e.g. only those QTL that are found under the stress condition and thus allows to study the genes that are directly involved in (or at least affected by) the stress response of the organism. The highest number of QTL can be found for the static category (Table 7.1), while less QTL are found for the 'conditional stress' category. Even smaller are the numbers for the 'conditional non-stress' QTL. Using this dataset of yeast eQTL, I aim to understand how frequently selection acts on gene expression levels and how strong the selection pressures are. Furthermore, I will use predefined gene modules (sets of interacting genes as defined in (Ryan et al. 2012), among which are many protein complexes) to test if selection acts on biologically relevant units that combine many genes of interest.

7.1.2 State configurations

To apply the selection test the state configurations (q_1, q_2, q_3) have to be determined. Only QTL that fulfill the two-state assumption and the assumption of additive trait

effects can be used for the selection model. If for a locus more than two significantly different alleles per locus exist (i.e. a locus has a different effect in each of the lines) or if the pairwise trait differences between the lines are not compatible with a single, additive trait effect, I neglect that locus in the following. Yet, one has to be careful with this step as this can be a potential source of bias, since a part of the available data is neglected for the selection analysis. In the following, I also explore the number of QTL following an additive three state model and what fraction of the loci show epistasis.

FDR	static QTL	cond. stress QTL	cond. non-stress QTL
0.05	1075 (43%)	552 (57%)	107 (45%)
0.1	1747 (44%)	796 (57%)	147 (43%)
0.15	2534 (46%)	976 (56%)	192 (42%)
0.2	3340 (47%)	1091 (54%)	231 (41%)
0.25	4341 (48%)	1269 (54%)	292 (44%)

Table 7.2: Number of *S. pombe* QTL with valid state configuration that follow the two-state assumption and the assumption of additivity. The QTL are again given for different significance thresholds for the QTL and sorted according to the four QTL categories as previously done in Table 7.1. The percentage given in brackets denotes the fraction of QTL that are assigned a valid state configuration for each category.

For all available QTL I calculate the state configurations, as described in detail in appendix A1. The number of QTL with valid state configuration can be found in Table 7.2. As can be seen by comparison of Tables 7.2 and 7.1 on average 49% of the QTL follow the two-state and the additivity assumption at a FDR of 0.25. The fraction of valid state configurations is particularly high for the QTL that are conditional for the stress condition. There is a slight dependency of the fraction of valid state configurations on the FDR: the fraction increases with the FDR for the static QTL and decreases for the 'conditional stress' QTL.

The remaining 51% of QTL violate either the two-state assumption or the assumption of additivity. To further test which of the assumptions is not fulfilled, I recalculate the state configurations while allowing for three different states per locus. In this case, the consistency check between the additive effects of the crosses simplifies. Yet, to fulfill the additivity assumption, one consistency relation is still necessary. Taking a_{13} as the largest additive effect one needs to have $a_{13} = a_{12} + a_{23}$ (within the uncertainty of the a 's), see also Figure 7.1. This is because three arbitrary fixed trait values T_1 , T_2 , and T_3 only allow for two independent values for the trait differences a_{12} , a_{13} , and a_{23} while the third trait difference is always determined by the other two.

Reestimating the state configurations with the possibility of three-state loci shows that only 1.6% of the QTL follow the additive three-state assumption, which is much

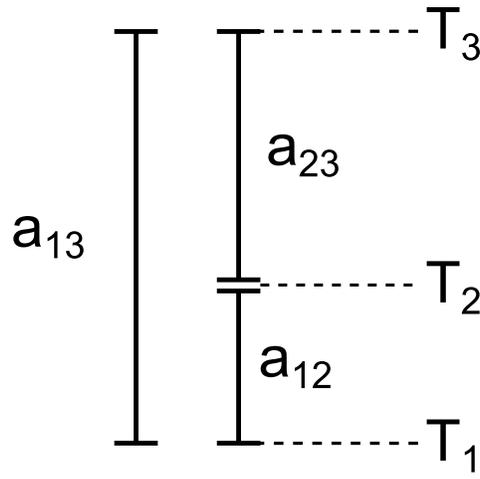


Figure 7.1: In a model that allows for three different trait values T_1 , T_2 , and T_3 , the trait differences only have two degrees of freedom, as they have to obey the condition $a_{13} = a_{12} + a_{23}$ (assuming that a_{13} is the largest trait difference). A two state model with only two different trait values is reached by setting either $a_{12} = 0$ or $a_{23} = 0$.

less than the 49% of two-state QTL. But this also implies that the uncertainties σ_a of the additive effects are so large that possible three-state loci where two states have a similar trait value are considered two-state loci. Thus, the two-state assumption seems to be reasonable considering this uncertainty. The remaining 49% of QTL cannot be explained by any additive model. The most probable explanation is that these QTL are affected by epistasis. An influence of the QTL mapping on this result cannot be ruled out, but the QTL mapping only influences the choice of QTL, while the non-additive behaviour is observed using only the measured trait values and genotypes.

Yet, this result cannot be explained with the simplest case of epistasis where there is one other locus epistatically interacting with the locus that shows the non-additive behavior. The non-additive behavior is observed as an average over many QTL crosses, where the genetic background is different in each crossed individual. Thus, for a single pair of interacting loci this averaging procedure should at least lead to a strong decrease in the strength of the detected interaction. Rather, this is an interaction with the genetic background (Rebai et al. 1997; Blanc et al. 2006; Lari pe et al. 2012), where even averaged over many possible background configurations a non-vanishing epistatic interaction remains. As already described before in Blanc et al. (2006) pairwise crosses of three lines can be used to quantify the strength and pervasiveness of the epistatic interaction with the genetic background. This is not possible for two lines as the consistency check between the additive effects observed for the different pairwise crosses is not available. For this dataset epistatic interactions with the genetic background are

very pervasive, as it concerns nearly half of the loci. Similarly high levels of epistasis has been found before for the yeast *S. cerevisiae*, where a QTL analysis of five wild isolate lines showed epistatic interactions with the genetic background in 69% of all QTL (Cubillos et al. 2011). In general this supports the hypothesis that epistasis is common in the genetic architecture of quantitative traits (Mackay 2014).

7.2 Selection on individual genes

First, I consider selection acting on individual genes. The expression level of a gene is the smallest unit for which selection can be investigated here. I identify genes with a sufficient number of state configurations to test for selection. I perform the selection test on the set of static QTL as well as a combination of static and conditional QTL. I apply corrections for the multiple testing bias, the high FDR of the QTL, and pleiotropy. In the next section larger sets of genes, that are e.g. part of a pathway or protein complex, are considered. In that case the focus is on higher level biological functions, with e.g. the output of a pathway being the feature under selection.

7.2.1 State configurations per gene

To determine the number of state configurations per gene I sort the QTL according to the targeted genes. In Table 7.3 the distribution of state configurations per gene is given for a FDR of 0.25. As can be seen the number of genes decreases rapidly for higher numbers of state configurations. Only for the static QTL a higher number of genes with 4 – 6 state configurations can be found.

state configurations	0	1	2	3	4	5	6
static	2392	1820	728	239	63	12	6
cond. stress	4211	851	177	20	1	0	0
cond. non-stress	4959	289	9	3	0	0	0

Table 7.3: Number of state configurations per gene at a FDR of 0.25. The number of genes with more state configurations is rapidly decreasing. Only for the static QTL category there is a higher number of genes with more than four QTL, which are necessary to apply the selection test.

As mentioned in section 3.5 a minimum of four state configurations is needed to perform the selection test on three lines. The number of genes with at least four state configurations is given for each QTL category in Table 7.4. Only for the category of static QTL a reasonable number of genes with enough QTL can be found. For all other categories there are no testable genes available. For the static QTL the number of testable genes also strongly depends on the chosen FDR of the QTL. For low FDR's

barely any genes with sufficient state configurations are available and only when a higher FDR cutoff is chosen a higher number of testable genes is reached.

There are only few genes that have a sufficient number of state configurations to apply the selection test. To deal with the problem, I use the following strategies: a) Instead of only allowing QTL that have a $FDR \leq 0.05$ and thus have a very high evidence, I extend my selection analysis to QTL with a higher FDR of ≤ 0.25 . This drastically increases the number of QTL and testable genes but also increases the probability that some of the QTL in the analysis are false positives that can skew the results of the selection test. If, however, the fraction of false positive QTL is low, one might hope that this has only a minor effect on the results of the selection test. In section 7.2.4 I describe and apply a strategy to correct for the possible bias introduced by the false positive QTL. b) No gene has sufficient QTL from the categories of 'conditional stress' or 'conditional non-stress' QTL alone. To make use of the conditional QTL, I combine the QTL for each of the conditions with the static QTL. By forming the two categories 'static + conditional stress' and 'static + conditional non-stress', I consider the complete set of QTL that can be found under the stress or the non-stress condition (including both the QTL that are specific to that condition as well as the unspecific QTL that can be found in any condition), respectively. This allows an inclusion of the stress/non-stress specific QTL into the selection analysis. A separate analysis of the stress/non-stress specific QTL would be preferable, as it would yield results specific to the stress response, which is however not possible here. As can be seen in Table 7.5, combining the two QTL categories leads to a clear increase in the number of genes with sufficient state configurations. c) Additionally, I combine sets of genes into biologically relevant gene modules to further increase the number of available state configurations per tested unit. That way, not only selection on single gene expression levels, but on larger biological units can be tested. A definition of the gene modules and the selection analysis for these can be found in section 7.3.

7.2.2 Selection analysis

For the available genes with at least four state configurations in the categories static, 'static + conditional stress', and 'static + conditional non-stress' I apply the selection test. For this, I use the scenario Q_1 with only one free selection parameter $(s_1, s_2, s_3) = (s, 0, 0)$ and scenario Q_2 with two free selection parameters $(s_1, s_2, 0)$. Q_2 is the most general three line selection model, as only the relative differences in selection parameters can be inferred. For scenario Q_1 I determine for each gene, which line is the most diverged and assign the selection parameter s to this line. For this, I calculate the differences between the trait values $|\Delta T_{12}| = |T_1 - T_2|$, $|\Delta T_{13}|$, and $|\Delta T_{23}|$. The line which is included in the two highest of these trait differences is the most diverged line (e.g. line 2 if $|\Delta T_{12}|$ and $|\Delta T_{23}|$ are the highest trait differences). Both scenarios are

FDR	static	cond. stress	cond. non-stress
0.05	1	0	0
0.1	3	0	0
0.15	13	0	0
0.2	39	0	0
0.25	81	1	0

Table 7.4: Number of *S. pombe* genes with at least four state configurations. Only for the category of static QTL enough genes with at least four state configurations are available. All other QTL categories have no genes with enough state configurations to apply the selection test. For the static category there is additionally a strong dependency on the false discovery rate cutoff for the QTL. At low FDR barely any gene has enough QTL; only when QTL with a higher FDR are taken into account more testable genes are available. Even then the testable genes are only a small subset of the 5260 measured genes.

FDR	static + cond. stress	static + cond. non-stress
0.05	1	1
0.1	7	3
0.15	20	17
0.2	52	50
0.25	124	93

Table 7.5: Number of *S. pombe* genes with at least four QTL for the two combined categories static and 'conditional stress' QTL as well as static and 'conditional non-stress' QTL. In comparison to the individual datasets in Table 7.4 a clear increase in the number of testable genes can be observed, especially in the case of static and 'conditional stress' QTL combined (which are the two largest categories in terms of number of QTL). This strategy of combining QTL categories also allows to include the conditional QTL into the selection test.

tested against the neutral scenario P_0 . The scores and p -values for each scenario are determined for all genes. The calculation of the p -values for this dataset is described in appendix A.2.

As can be seen in Figure 7.2 (red histogram) a large fraction of the genes is found to be under selection both using scenario Q_1 and scenario Q_2 . About 30% of all tested genes have a significant p -value for scenario Q_1 and similarly 33% of the genes for scenario Q_2 using the static QTL (see also Table 7.6). When the set of 'conditional stress' QTL is added to the static QTL the fraction of significant genes slightly increases to 35% for scenario Q_1 and to 39% for scenario Q_2 . There are more genes found to be under selection using scenario Q_2 . Since scenario Q_2 has one additional free selection parameter compared to scenario Q_1 , it has usually less statistical power to detect selection than scenario Q_1 if scenario Q_1 provides an accurate description of the data. Since scenario Q_2 yields the stronger signal for selection, scenario Q_1 with only one line diverged from the other two might not be the most precise description for the present data. Thus, the most general selection scenario Q_2 should be preferred in this case.

This clear result suggests that selection on gene expression levels is widespread for the tested lines of the organism *S. pombe*. The study of selection on gene expression levels in *S. cerevisiae* by Fraser et al. (2010) also found signs for pervasive selection, while the fraction of selection (a lower bound of 10% of genes under selection was found) was lower than the fraction found for *S. pombe* in this dataset. A pervasive role of directional selection on gene expression levels was also found in other species like primates (Romero et al. 2012).

What distinguishes this study from most other selection studies on QTL (Fraser et al. 2010; Bullard et al. 2010; Fraser et al. 2011) is that it allows to identify individual genes under selection in contrast to larger sets of genes as e.g. pathways (which will also be considered in the following) allowing to probe selection on an even smaller genetic level. This is possible due to the large number of eQTL identified in (Clément-Ziza et al. unpublished dataset) together with the strategy to allow for a high FDR for the eQTL that drastically increases the number of eQTL.

category	genes tested	$p < 0.05$ Q_1	$p < 0.05$ Q_2
static	81	24 (30%)	27 (33%)
static + cond. stress	124	44 (35%)	48 (39%)
static + cond. non-stress	93	27 (29%)	34 (37%)

Table 7.6: **Number of *S. pombe* genes with significant p -values.** Number of genes that are tested for selection and number of genes that are found to be under selection ($p < 0.05$) under the scenarios Q_1 and Q_2 , respectively. Numbers are given for the static QTL set as well as the combined sets of 'static + conditional stress' QTL and 'static + conditional non-stress' QTL. In all cases a high fraction of genes under selection was found, with the highest fraction (39%) found in the case of the 'static + conditional stress' genes and scenario Q_2 .

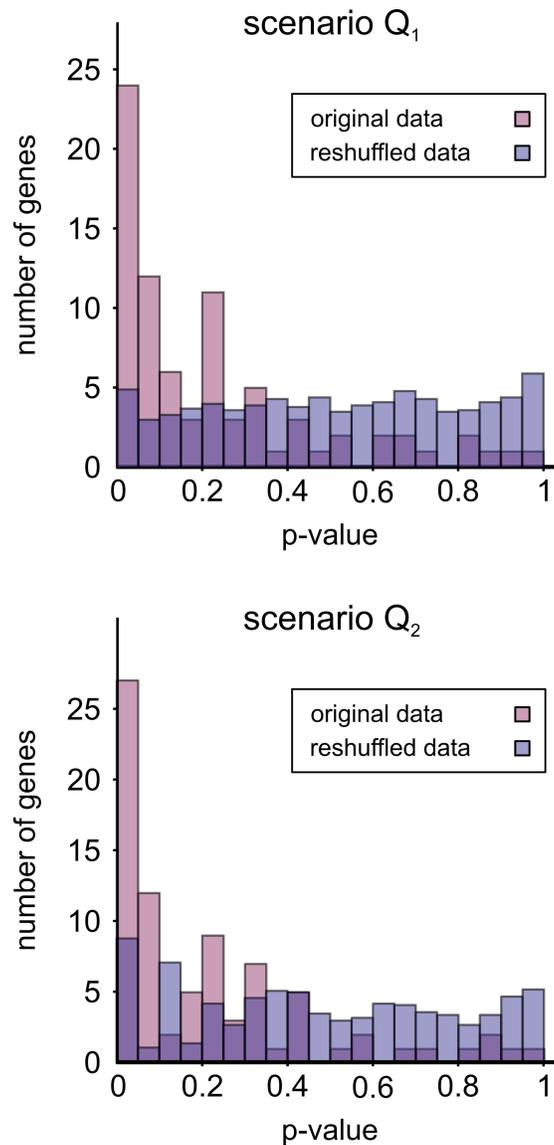


Figure 7.2: **A large fraction of *S. pombe* genes is found to be under selection.** The distribution of p -values calculated for all genes with at least four state configurations using all static QTL and a FDR cutoff of 0.25 for the QTL. The log-likelihood score (3.10) was calculated testing selective scenarios Q_1 (top) and Q_2 (bottom) against the neutral scenario P_0 and a p -value was calculated for each of the scores. For both scenarios Q_1 (30%) and Q_2 (33%) a high fraction of genes has significant scores in favour of the selective scenario. To account for the multiple testing bias, a null distribution of p -values is calculated by randomly reshuffling all state configurations of the tested genes and calculating scores and p -values for these reshuffled data. This is repeated 10 times to obtain an average null distribution for the p -values. Comparing actual and reshuffled p -value distribution one sees that, even though some fraction of the significant genes will be false positives, there are many more significant genes found than expected by chance, confirming the high fraction of genes under selective pressure.

7.2.3 Multiple testing correction

Since many genes are tested for selection, one expects a fraction of genes with a significant p -value even under completely neutral evolution. To deal with this multiple testing problem and to quantify the number of false positive genes, I randomly reshuffle the state configurations of all tested genes (that have at least four state configurations) and recalculate their score and p -values. This is repeated 10 times to obtain an average of the null distribution of p -values, keeping the set of genes, number of QTL and additive effects fixed. Figure 7.2 (blue histogram) shows that the null distribution dramatically differs from the p -value distribution for the real data. While the null distribution is nearly uniform with some deviations that do not come from insufficient averaging but from the limited number of tested genes, the distribution obtained for the data is strongly biased towards small p -values with a clear peak at $p < 0.05$. This shows a very strong signal of selection in the dataset. When comparing the number of genes with significant p -values for the two distributions, 5 out of the 24 genes that are significant under scenario Q_1 are expected by chance and around 8 out of the 27 significant genes under scenario Q_2 , suggesting that the majority of significant genes are true positives. After adjusting for the number of false positive genes, one still expects 19 out of 81 genes (23%) to be under selection for both scenario Q_1 and scenario Q_2 . So a high fraction of all tested genes is found to be under selection.

To determine more precisely which of the individual genes is under selection, I use the Benjamini-Hochberg procedure as described in section 4.1. With this procedure an individual false discovery rate (q -value) is calculated for each gene that accounts for the multiple testing problem arising due to the large number of tested traits. The distribution of q -values for both scenarios can be seen in Figure 7.3. The q -value allows to quantify the probability for each significant gene to be a false positive and allows to identify the genes where the prediction of selection is most robust. Here, I again choose a significance threshold of 0.05. 6 of the genes have a q -value below 0.05 for scenario Q_1 and 11 genes have $q < 0.05$ for scenario Q_2 , showing that the selection test is able to identify genes under selection with certainty even in a large dataset with thousands of genes and more than hundred genes tested for selection.

7.2.4 Adjusting for the high QTL false discovery rate

To incorporate more QTL in the selection analysis, I increased the FDR threshold for the genes. This strongly increased the number of available QTL and led to a number of genes with sufficient state configurations for the selection test. Yet, it also increases the number of false positive QTL that are incorporated in the analysis. False positive QTL contributing a state configuration to a gene might change the outcome of the selection test, either leading to a significant score where no selection is present or hiding selection with a state configuration that does not fit the pattern of the other state configurations

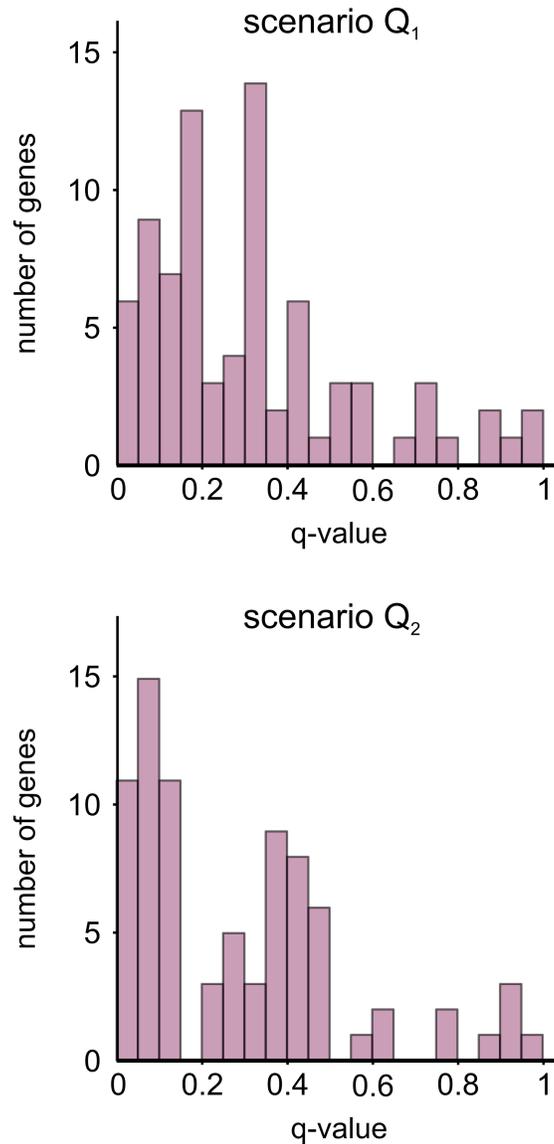


Figure 7.3: **False discovery rate (q -value) for all tested *S. pombe* genes.** For both scenario Q_1 and Q_2 there are genes with a q -value below 0.05, showing strong evidence of selection. With 11 genes with $q < 0.05$, scenario Q_2 seems better suited to describe the selection in the data than scenario Q_1 , which identified only 6 genes with $q < 0.05$.

for the gene. To correct for the false discovery rate of the individual genes, I introduce a modified log-likelihood score, where the contribution of each score is weighted by the expected true discovery rate (which is 1 minus the FDR). I use a modified version of the log-likelihood score (3.10)

$$S_{Q,P} = \sum_{l=1}^{L_{\text{div}}} (1 - \text{FDR}_l) \ln \left(\frac{Q(q_{1,l}, q_{2,l}, \dots, q_{n,l} | N s_1^*, \dots, N s_n^*, \Omega_l^*, a_l)}{P(q_{1,l}, q_{2,l}, \dots, q_{n,l} | N s_1^{*'}, \dots, N s_n^{*'}, \Omega_l^{*'}, a_l)} \right), \quad (7.1)$$

where FDR_l is the false discovery rate for locus l . With this FDR-adjusted log-likelihood score QTL with a higher FDR contribute less to the score, accounting for the larger uncertainty for these QTL. I recalculate all scores and p -values using the adjusted score (7.1). The results in Figure 7.4 show a very similar picture for the p -value distribution compared to the unadjusted distribution in Figure 7.2. The exact numbers of significant genes of the FDR-adjusted score are found in Table 7.7. The very similar number of genes under selection (there is at most a difference of 3 significant genes more or less for any category) show that the higher FDR threshold used for the QTL barely affects the results of the selection test. In the following, I only use the results obtained using the FDR-adjusted log-likelihood score (7.1).

category	genes tested	$p < 0.05$ Q_1	$p < 0.05$ Q_2
static	81	23	29
static + cond. stress	124	47	45
static + cond. non-stress	93	27	34

Table 7.7: **Number of *S. pombe* genes with significant p -values using the FDR-adjusted log-likelihood score (7.1).** Comparison with Table 7.6 shows that the number of genes under selection are nearly identical, indicating only a small influence of the QTL with elevated FDR rates on the results of the selection test. In two cases (static QTL, scenario Q_2 and 'static + conditional stress' QTL, scenario Q_1) there are even more genes found to be under selection using the adjusted score. This is not unexpected as a gene that is actually under selection might get assigned a low log-likelihood score if a false positive QTL with a state configuration that is not fitting the selective pattern is added.

7.2.5 Lines with highest trait divergence

With a large fraction of genes under selection the next question is, which of the lines show the strongest sign for adaptation, i.e. which lines have the highest level of trait divergence. For this I calculate the line with the highest trait difference compared to the other two lines for each tested gene (which is the line that gets the selection coefficient in scenario Q_1). The number of genes with diverged lines 1, 2, and 3 for

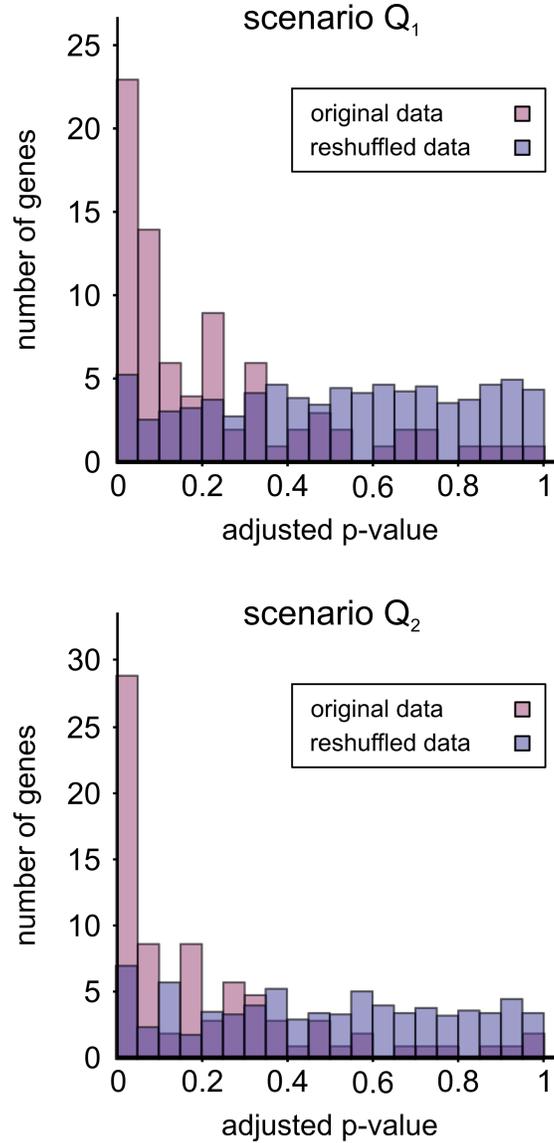


Figure 7.4: p -values for the *S. pombe* genes tested under the FDR-adjusted log-likelihood score. The distribution of p -values under the FDR-adjusted log-likelihood score (7.1) for the static QTL has nearly the same shape as the unadjusted distribution (Figure 7.2). The number of genes with p -values below 0.05 are similarly high for both scenario Q_1 and Q_2 . This indicates that the elevated FDRs in some of the QTL have no large effect of the result of the selection test.

both all tested genes as well as for the significant genes only can be found in Table 7.8. It can be clearly seen that line 2 is most frequently the most diverged line, while this is rarely the case for line 1. A possible explanation for the absence of trait divergence in line 1, which is the lab strain, is that the constant environment experienced under laboratory conditions leads to reduced selective pressure compared to the other lines such that no strong changes in expression levels are observed. No bias between up- or downregulated genes could be found (data not shown). This suggests that line 2 has changed its gene expression levels more than the other lines and possibly being under the highest selection pressure. Another explanation could be the frameshift in the gene *swc5* of line 2, the South African line, which alters the chromosomal histone deposition (Clément-Ziza et al. 2014). This effective knockout of *swc5* affects the expression of many of the genes, with the majority of genes downregulated. The signal of selection seen in the data might be, at least partially, affected by this knockout phenotype. But as shown in the next section, the results of the selection test are nearly unaffected when all QTL connected to *swc5* are removed from the analysis (see also Table 7.8).

category	all			significant		
	1	2	3	1	2	3
static	9	38	34	0	17	6
static + cond. stress	10	68	46	0	36	11
static + cond. non-stress	11	39	43	0	18	9
static, no <i>swc5</i> -QTL	12	37	32	0	17	6

Table 7.8: ***S. pombe* lines with the largest trait divergence for the genes.** The number of genes in which lines 1, 2, or 3 have the largest trait difference (and get assigned the selection coefficient in scenario Q_1) are listed for the different QTL categories. Numbers are given for both the set of all tested genes and the set of genes with a significant score (with $p < 0.05$). In all cases, but especially for the group of significant genes, line 2 is the most diverged line in the majority of cases while this is never the case for line 1, pointing towards strong lineage-specific selection in these lines. Excluding the QTL connected to the *swc5* knockout in line 2 that is known to affect a large number of gene expression levels leads to very similar results, demonstrating no major effect of this gene on the high number of diverged genes in line 2 found to be under selection.

7.2.6 Pleiotropy

Another point that has to be considered is pleiotropy. In this context pleiotropy describes the situation where a QTL influences several gene expression levels (Wagner and Zhang 2011), such that an evolutionary force on one of the genes might also affect

the second gene via the shared QTL (see also section 3.6.1). Since there are only 1693 markers used for the eQTL analysis but more than 10000 detected eQTL, many of the markers have to affect multiple genes, implying a high level of pleiotropy. To test the level of pleiotropy I calculate for each marker how many of the 81 genes tested for selection (using the static QTL) have a QTL affected by that marker. The analysis is complicated by the long marker intervals associated with a single QTL. Since the resolution of QTL analysis is limited by the frequency of recombination events in the crosses, one QTL can be associated with many markers. Also, the marker intervals of different QTL can overlap even though the underlying genetic variant that is causal for the QTL is not the same. This leads to an overestimate of the prevalence of pleiotropy. A high number of genes affected by one marker might not mean a high level of pleiotropy since different loci within the regions of the marker intervals might cause the QTL effect.

In Figure 7.5 the number of genes affected by the individual markers can be seen. The most prominent features of the distribution are the marker hotspots (that affect many genes) in the region approximately between marker no. 300 and 600, around marker no. 920, and around marker no. 1580. The large region between marker no. 300 and 600 corresponds to the known region of reduced recombination in line 1 (see section 7.1.2) which leads to very long marker intervals associated to each detected QTL that add up in the histogram but which do not imply a high number of QTL in that region. The marker hotspot around marker no. 1580 corresponds to the known knockout of the *swc5* gene in line 2 (see section 7.2.2) which was shown to affect many gene expression levels (Clément-Ziza et al. 2014). The underlying nature of hotspot marker 920 is not known.

The knockout of the *swc5* gene was shown to alter the expression levels of many genes, with most of the genes being downregulated (Clément-Ziza et al. 2014). To determine the effects of this knockout on the results of the selection test, I repeat the selection analysis on the static QTL excluding all QTL associated with the hotspot marker at position 1580. The results showed barely any differences with 23 genes under selection scenario Q_1 (previously: 23 genes) and 30 genes for scenario Q_2 (previously 29 genes). Also the results for the largest trait divergence in the lines are consistent (see Table 7.8). The exclusion of the other hotspots at marker no. 920 and marker no. 515 gives very similar results (data not shown) with only a small reduction in the number of genes detected for selection even though the number of available QTL is reduced. This shows the robustness of the selection result with respect to the individual hotspot markers, which, even though they affect many gene expression levels, have no large scale effect on the selective signatures within the dataset.

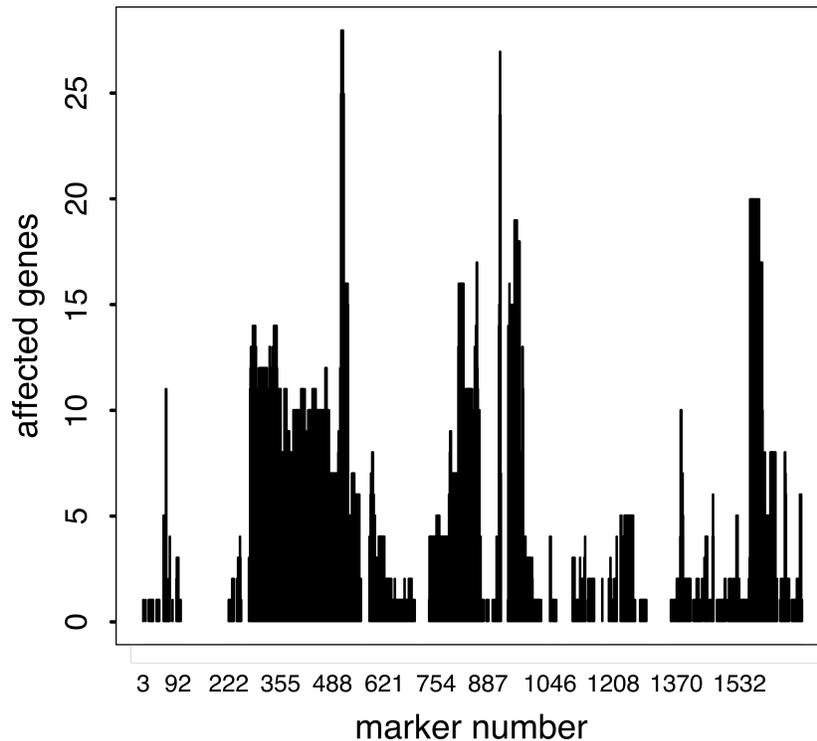


Figure 7.5: **Number of *S. pombe* genes associated with each marker.** For each marker the number of genes with a QTL connected to that marker is given (of all 81 genes tested for selection using the static QTL at a QTL-FDR cutoff of 0.25). Because the resolution of QTL mapping is limited by the distance between recombination events in the crosses, some QTL are associated with a large marker interval. Due to this effect the distribution overestimates the amount of pleiotropy. Two genes sharing a QTL marker might not display a case of pleiotropy if the underlying causal QTL are located at different places of the QTL marker intervals. One example for this is the region between markers no. 300 and 600, which corresponds to the known region of reduced recombination in line 1 (see section 7.1.2) leading to very long marker intervals connected to single QTL that add up in the histogram. Other QTL hotspots are around marker no. 1580, which is connected to the known knockout of the *swc5* that affects many gene expression levels (Clément-Ziza et al. 2014), and around marker no. 920.

7.3 Selection on gene modules

Up to now, I only looked at the selection pressure on individual genes, but selection might also act on larger units that perform a higher-level biological function for the organism. For this, I use the gene modules that combine several genes defined in Ryan et al. (2012). As for the genes, I look for gene modules under selection using both the static as well as the 'static + conditional stress' and 'static + conditional non-stress' QTL. I determine the gene modules exclusively found to be under selection using the conditional stress QTL and determine the biological function of those gene modules.

In their paper Ryan et al. (2012) created a genome-scale epistasis map of about 60% of the nonessential genome, creating genetic interaction profiles of the genes. A similarity score measures the similarity of the genetic interaction profiles between the genes. A hierarchical clustering algorithm is used to group genes with similar interaction profiles in modules. They found 297 functional modules containing a total of 992 genes, with the size of the gene modules ranging from 2 genes (for around 2/3 of the modules) up to 26 genes (Ryan et al. 2012). For many of the gene modules a clear biological function could be assigned and some could be identified with known protein complexes as for example the large and small ribosomal subunit.

7.3.1 State configurations per gene module

For the selection analysis, I combine all QTL of the genes contributing to a module. This allows me to test biologically relevant gene modules that might not have been accessible before, as not enough QTL were available for the individual genes. Since the genetic interaction profiles of the genes within one module are similar, I can directly combine the state configurations of the genes in a gene module. An increase in the expression level of one gene in a module should have a similar effect as the increase in another gene (e.g. the number of a protein complex increases if the expression of all its components increases).

The state configurations of all genes in a module were combined and gene modules with at least four state configurations were used for further analysis. Table 7.9 lists the number of gene modules with at least four state configurations for different FDR thresholds for the QTL. On average, the individual genes contribute approximately 1.5 state configurations to the gene modules, with most of the genes only contributing a single state configuration and only 10 of the 992 genes contributing 4 or more state configurations. This shows that the set of QTL used for the selection test of the gene modules is very different from the set used for the genes such that genes that could not be tested individually are now accessible for the selection test. Since the gene modules combine many genes, already at low FDR rates gene modules with sufficient state configurations are available. Even the set of 'conditional stress' QTL alone yields

several gene modules with more than four state configurations. This allows to test this QTL set separately, without the addition of the static QTL, which was not possible using individual genes. This does not hold true for the set of 'conditional non-stress' QTL where not a single gene module has sufficient state configurations. At the highest FDR of 0.25, 65 out of 297 gene modules can be tested for the static QTL, implying a much better coverage than for the individual genes, where only 81 out of 5260 genes could be tested (Table 7.4). Yet, the set of genes in the gene modules is itself only a limited subset of all genes. Comparing the different sets of QTL, the combined set of 'static + conditional stress' QTL has the highest number of testable gene modules, similarly to the case of the genes.

FDR	static	cond. stress	cond. non-stress	static + cond. stress	static + cond. non-stress
0.05	7	2	0	13	7
0.1	15	4	0	25	18
0.15	29	5	0	44	32
0.2	48	8	0	72	53
0.25	65	11	0	90	70

Table 7.9: **Number of gene modules for *S. pombe* with at least four state configurations.** Since the gene modules combine several genes, gene modules with at least four state configurations are already available at the lowest FDR of 0.05, in contrast to the case of the genes. Even for the set of stress QTL alone some genes with sufficient state configurations can be found, allowing to analyze this set of QTL separately, which was not possible for the genes. This is not true for the set of non-stress QTL, where no gene module has four or more state configurations. At the highest FDR, 65-90 out of 297 gene modules can be tested for selection. As in the case of the genes before, the combined dataset of 'static + conditional stress' QTL clearly has the highest number of testable gene modules.

7.3.2 Selection analysis

Next, I test the gene modules for selection. For this, I calculate score and p -value for each gene module separately, as already described for the genes in the previous section. For this, I again use the FDR-adjusted log-likelihood score (7.1). As can be seen in Figure 7.6, a clear selection signal is present also for the gene modules with a much larger fraction of gene modules with significant p -values than expected when comparing to the null distribution. Yet the fraction of gene modules under selection is smaller than the fraction of genes under selection (see also Table 7.10). An exception is the set of 'conditional stress' QTL, which could not be tested for the case of the genes. Here, 55% (6 out of 11) of the tested gene modules had a significant score,

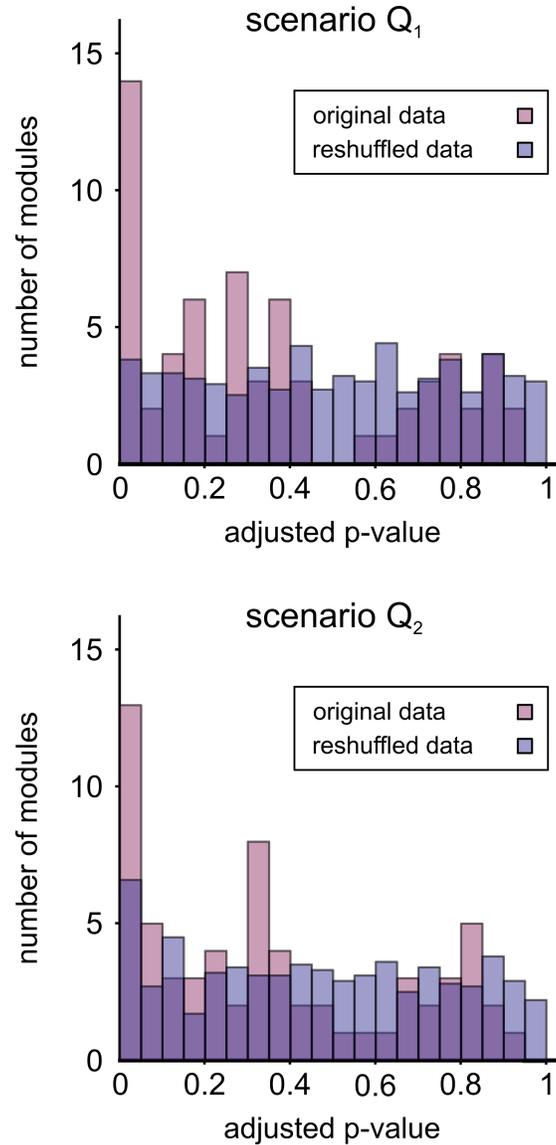


Figure 7.6: p -value histogram for the gene modules using the static QTL and the FDR-adjusted log-likelihood score (7.1). Also for the gene modules the selection test shows a clear sign of selection for a part of the gene modules (red distribution). Even though the fraction of gene modules under selection (with $p < 0.05$) is smaller than for the case of the genes (see also Table 7.10), it is still elevated compared to the null distribution of gene modules with reshuffled state configurations (blue distribution).

which is much higher than the highest fraction of significant genes (39% for 'static + conditional stress' QTL, scenario Q_2). This gives a clear hint that for the QTL specific to the stress response selection pressure is much more pervasive.

For comparison, a similar but simpler approach tests for selection on pathways in the yeast *S. cerevisiae* and *S. bayanus* (Bullard et al. 2010). This test only looks at *cis*-regulatory changes in gene expression and also finds signs of selection on pathways, but with a lower prevalence.

category	modules tested	$p < 0.05 Q_1$	$p < 0.05 Q_2$
static	65	14 (22%)	13 (20%)
static + cond. stress	90	24 (27%)	25 (28%)
static + cond. non-stress	70	17 (24%)	15 (21%)
cond. stress only	11	6 (55%)	6 (55%)

Table 7.10: **Number of gene modules for *S. pombe* found to be under selection using the FDR-adjusted log-likelihood score (7.1).** While the percentage of gene modules under selection is smaller than in the case of the genes, it is still clearly elevated compared to expectations under a neutral null model. Especially striking is the fraction of gene modules under selection for the 'conditional stress' QTL, which can now be used separately for the selection test. Here, 55% of the 11 tested gene modules are under selection, far higher than the highest percentage (39%) reached for the genes using the 'static + conditional stress' QTL with scenario Q_2 . This suggests that the QTL responsible for stress response are experiencing especially strong selection pressures.

Also for the most diverged lines the picture is very similar to the results for the genes, see Table 7.11. In many cases lines 2 and 3 are the most diverged lines while this is still rarely the case for line 1. Especially for the 'static + conditional stress' QTL line 2 is again by far the most diverged line with nearly 3 out of 4 significant gene modules. Even more striking, for the set of 'conditional stress' QTL alone, all six significant genes are diverged in line 2. This shows that line 2 has a very different phenotype than the other two lines when considering the set of 'conditional stress' QTL. This might hint towards a difference in the response to stress of line 2 that was formed through adaptation. Again there is no clear bias towards the up- or downregulation of the gene modules (data not shown).

7.3.3 Stress-specific gene modules under selection

In this and the next section I want to answer two questions. First, are there differences in selection pressure between the QTL categories? And second, what are the biological functions of the genes and gene modules that are found to be under selection? For this, I start with the analysis of the gene modules, since they are more tightly connected to

category	all			significant		
	1	2	3	1	2	3
static	17	25	23	2	7	5
static + cond. stress	15	47	28	0	18	7
static + cond. non-stress	16	28	26	1	8	6
cond. stress only	2	8	1	0	6	0

Table 7.11: *S. pombe* lines with the largest trait divergence for the tested gene modules. Similarly to the results for the genes, line 1 is rarely the most diverged line, especially when considering the significant gene modules, while especially line 2 but also line 3 are often the most diverged ones. In the case of the 'static + conditional stress' QTL and even more pronounced for the stress QTL alone line 2 is the most diverged line for the majority of gene modules. This could hint to a role of selection in the phenotypic difference of the lines for the gene modules connected to stress response.

known biological functions, giving a more precise picture about the biological relevance of the selection test results.

To reveal the differences between the QTL categories, I determine which gene modules under selection are found exclusively in one of the QTL categories and which are shared between them. Especially, I want to focus on the gene modules that are exclusively found in the 'static + conditional stress' QTL set and the set containing only the stress QTL. This might help to understand which gene modules are under selection specifically for the stress response of the organisms. As can be seen in the Venn diagram in Figure 7.7, the majority of gene modules found to be under selection for the 'static + conditional stress' QTL set are exclusive for this QTL set and cannot be found by using either the static or 'static + conditional non-stress' dataset. Nearly all of the remaining significant gene modules are shared between all QTL categories. This clearly suggests that a subset of the gene modules might have a strong connection to stress response. Also it shows that the combination of 'static + conditional stress' QTL is useful to reveal the selective pressures connected to the stress condition. Even though the 'conditional stress' QTL contribute only a minority of the QTL to the combined QTL set (see again Table 7.2), half of the gene modules found to be under selection are exclusive for this dataset, highlighting the importance of the contribution of the 'conditional stress' QTL.

Some significant gene modules exclusive to the 'static + conditional stress' QTL set are expected just because this combined QTL set has the highest number of available QTL, such that more testable gene modules are expected. To test if there are more gene modules exclusive to the 'static + conditional stress' QTL set than expected by the increase in testable gene modules, I compare these numbers to the numbers of testable gene modules that are exclusive to the individual conditions, see Figure 7.7 (middle).

As can be seen, a much smaller fraction of testable gene modules are exclusive to the 'static + cond. stress' QTL set and an exact Fisher test confirms that more gene modules under selection are exclusive to this condition than expected ($p = 0.013$). This again indicates that a higher selection pressure seems to act on gene modules connected to the stress response.

The results for the 'conditional stress' QTL set alone produce very similar results as the 'static + conditional stress' QTL set (see Figure 7.7 right). Four of the six gene modules that are found to be under selection for the 'conditional stress' QTL set are already found to be under selection exclusively in the 'static + conditional stress' QTL set. One gene module under selection is shared with all other QTL sets and is thus found to be under selection either using the static QTL only as well as the 'conditional stress' QTL only, which are non-overlapping and thus provide different QTL for selection analysis. So this gene module is found to be under selection using two completely different QTL sets, providing an even stronger sign for selection than a positive test in any of the QTL sets alone. The last gene module under selection is exclusive for the 'conditional stress' QTL. This means that the static QTL contributing to this gene module changed the result of the selection test from significant to non-significant. Overall, the trend is even stronger for the 'conditional stress' QTL set alone, where a clear majority of significant gene modules are specific to the stress condition, confirming the role of the significant gene modules in stress response.

7.3.4 Biological functions of gene modules under selection

Next, I want to investigate the biological functions of the gene modules under selection. For this, I will use all gene modules that are found to be significant using the 'static + conditional stress' QTL set. With this QTL set a clear surplus of gene modules under selection could be detected, indicating a role of these gene modules in stress response. Additionally, this combined QTL set covers a much larger fraction of tested gene modules, giving a broader overview over the selective pressures for the majority of gene modules. The 'conditional stress' QTL set has a much clearer connection to stress response, with all QTL found in this set exclusively appearing under the stress condition. Yet, only a small fraction of all gene modules could be tested for this subset due to a low number of QTL, lacking the broadness of the combined QTL set. Also most of the significant gene modules found for the stress QTL set are found for the combined QTL set as well and do not add new insights.

The gene modules found in Ryan et al. (2012) cover a larger fraction of all *S. pombe* genes. 997 of 5260 genes are included in the gene modules and cover a broad field of biological functions as for example transcription, translation, DNA metabolism, mitosis, or RNA processing. 64 out of the known 297 gene modules could be assigned a clear biological function which allows to gain a better insight into the biological

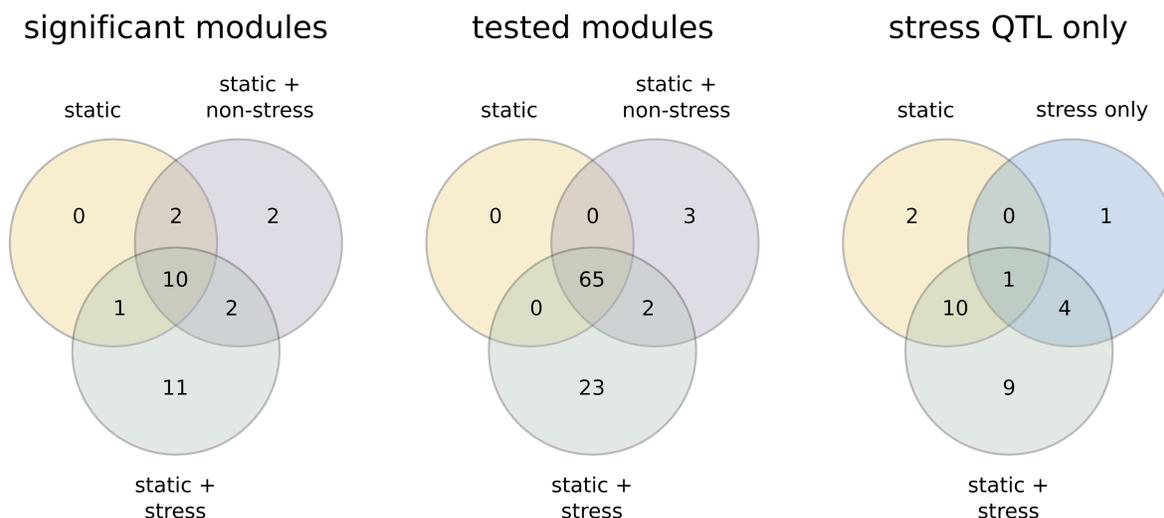


Figure 7.7: Venn diagram for significant *S. pombe* gene modules as well as all tested gene modules for the different QTL categories. Left: The number of significant gene modules that are exclusive to or shared by the three QTL sets static, 'static + conditional stress', and 'static + conditional non-stress'. Most remarkably, 11 of the 24 gene modules are specific for the 'static + conditional stress' QTL set. This speaks for a stronger selection on the set of QTL that can only be observed in the stress condition and are thus involved in the stress response of the organism. Middle: Since the 'static + conditional stress' QTL set contains the highest number of QTL, some significant gene modules exclusive to this set are expected. To test if there is an excess of significant gene modules in this QTL set, I compare this to the numbers of tested gene modules exclusive to the 'static + conditional stress' QTL sets. As can be seen, a significantly smaller fraction of tested gene modules is exclusive to the 'static + conditional stress' QTL sets (Fisher's exact test, $p = 0.013$). This indicates that selection pressure is much more pervasive for the gene modules that are specific for the stress response of the organism. Right: The result for the number of significant gene modules found using the 'conditional stress' QTL only is even more pronounced. While only one of the six significant gene modules is shared with the static QTL and the 'static + conditional non-stress' QTL (not shown here), four significant gene modules are in common with the 'static + conditional stress' QTL set while one gene module has been found to be under selection that has not been significant for the 'static + conditional stress' QTL set.

functions under selection. Because the function of the remaining gene modules remains unclear, so far not the full range of biological functions connected to the gene modules can be utilized. Gene modules directly connected to stress response are not available, such that one cannot expect to find these. Rather, one might expect gene modules that are reasonably connected to mechanisms included in stress response.

The gene modules that are found to be under selection exclusively in the 'static + conditional stress' QTL set and their respective biological function (gene module numbers as defined in Ryan et al. (2012) are given) can be found in Table 7.12. For each gene module also the most common gene ontology (GO) annotations, which name the biological processes and cellular components connected to the genes, of the associated genes obtained from the PomBase database (Wood et al. 2011) are given in Table 7.12. In the GO annotations of the gene modules several GO terms can be found repeatedly. First, there are several gene modules connected to the regulation of transcription and translation (gene modules no. 64, 114, 136, 181). It is well known that cells respond to stress by a broad and coordinated change in gene expression, which allows the organism to quickly adapt protein levels to changing environmental demands (Holcik and Sonenberg 2005). As an example, the histone H3-K9 demethylation of gene module no. 181 is correlated with transcriptional repression (Rosenfeld et al. 2009). Besides the importance of the regulation of transcription also the regulation of translation is important for the rapid response to stress (Holcik and Sonenberg 2005). The large ribosomal subunit (gene module no. 136) is one of the central protein complexes involved in translation. It was shown that organisms respond to a stress condition by a general reduction of transcription which allows to save up to 50% of the cellular energy and increase the production rate of proteins involved in the stress responses (Holcik and Sonenberg 2005). This makes the regulation of translation a very important mechanism for stress response.

Additionally, gene modules connected to different kinds of stress responses are found in the analysis. Most prominently, the two gene modules no. 86 and 114 are involved in the electron transport chain. They comprise several protein complexes in the mitochondrial membrane and are responsible for the cellular respiration, converting biochemical energy from nutrients to ATP (Joseph-Horne et al. 2001). As reactive oxygen species are formed within this metabolic pathway, and in particular during situations with environmental stress, this pathway is one of the major sources of oxidative stress inherent to the organism (Herrero et al. 2008). Furthermore, gene module no. 86 is also involved in the arginine biosynthetic process which is the direct precursor of nitric oxid, a major responder to oxidative stress (Astuti et al. 2016). Another example for a gene module involved into the response to oxidative stress is the gene module no. 253, which is the only gene module found to be under selection using the stress QTL alone. Although this gene module could not be assigned a clear biological function, it consists of some key genes of the large ribosomal subunit (*rpp201*, *rpl4302* and *rpl2301*) as

well as the heat shock protein *hsp104*, which is known to be involved in the response to heat stress but also oxidative stress in the yeast *S. cerevisiae* (Collinson and Dawes 1992; Godon et al. 1998).

As an additional factor, several modules are connected to different kinds of stress that are not directly connected to the stress applied to the *S. pombe* lines, but which might share some common mechanisms for stress response. Gene module no. 64 is connected to the response to glucose starvation and gene module no. 170 to the response to salt stress. Finally, a connection to the regulation of the mitotic cell cycle can be found in gene module no. 170; it is known that the cell cycle is typically arrested as a response to stress (Boonstra 2003).

Overall, the analysis of the significant gene modules exclusive to the 'static + conditional stress' QTL show a tight connection to biological functions related to stress response.

Interestingly, the pathways found to be under selection in Bullard et al. (2010) for the species *S. cerevisiae* and *S. bayanus* are connected to similar biological functions with keywords like ribosome biogenesis, respiration, translation, and response to stress appearing as well. This might hint to common pathways as typical targets for selection in different environments.

module	relative trait value	involved genes	GO annotation
64		<i>ssp2, amk2</i>	regulation of transcription, response to glucose starvation
69		<i>elp1, elp2, elp3, elp6, ctu1, ctu2, dph3</i>	tRNA modification, tRNA metabolic process
86		<i>tsf1, arg3, arg6, arg11, arg12, lys2, lys3, lys4, lys12</i> , 2 unassigned	arginine biosynthetic process, urea cycle, NAD binding, mitochondrial matrix
114		<i>atp2, atp15, sep1</i>	mitochondrial proton transport, ATP biosynthetic process, regulation of transcription
122		<i>rsn1, rpl3001</i>	vesicle-mediated transport, cytoplasmic translation
136		<i>rpl2102, rpl1702, rpl2801, rpl1601, rpl702, rpl1902, rpl3702, rpl3801</i>	large ribosomal subunit, translation
170		<i>pck1, cch1, rga8, mug123, pmp1, yam8</i> , 1 unassigned	signal transduction, response to salt stress,

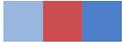
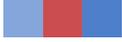
175		<i>dsc1, dsc2, dsc3,</i> <i>dsc4, dsc5</i>	protein ubiquitination
181		<i>pbr1, phf1, phf2,</i> <i>mit1, mlo2, 1 unassigned</i>	histone H3-K9 demethylation protein ubiquitination, regulation of transcription
182		<i>gar2, tma23, rok1</i>	ribosome biogenesis
224		<i>nxt1, mex67, 1 unassigned</i>	ribosome biogenesis, nucleocytoplasmic transport
253		<i>gal1, cxr1, abc3, rpp201,</i> <i>fcf2, gdh1, tho4, mpc1,</i> <i>rpl4302, rpl2301, hsp104,</i> 10 unassigned	cytosolic large ribosomal subunit, regulation of transcription, cellular amino acid metabolic process, cellular response to misfolded protein

Table 7.12: **Biological functions of the significant *S. pombe* gene modules exclusive to the 'static + conditional stress' QTL set.** For each gene module the most common gene ontology (GO) annotations are given, which name the biological processes and cellular components connected to the genes. Many gene modules are connected to biological functions involved in stress responses, as the regulation of transcription and translation, the electron transport chain, several stress response mechanisms (oxidative stress, salt stress, glucose starvation), and regulation of the cell cycle. All gene modules were exclusive to the 'static + conditional stress' dataset, except for gene module no. 253, which was only found to be under selection in the 'conditional stress' QTL set alone. The relative trait values of the modules in the different lines are given in form of a heat map in the second column, where the three colors correspond to lines 1 - 3 (from left to right). Red corresponds to the highest possible trait value (with all loci contributing a + state for the line), while blue corresponds to the lowest possible trait value (with all loci contributing a - state). White corresponds to an equal number of + and - states.

Additional stress specific protein complexes Since not all protein complexes that are known to be connected to the response to oxidative stress are included in the gene module dataset, I additionally test two well known protein complexes related to stress response for selection, the proteasome and the respiratory chain complexes. The proteasome degrades misfolded or damaged proteins (e.g. due to oxidative stress) and is known to be involved in the response to oxidative stress (Aiken et al. 2011). It consists of the 20S and the 19S subcomplexes. The 20S subcomplex again has two subunits, the 20S alpha subunit and the 20S beta subunit. I test these protein complexes for selection by combining the QTL of all genes contributing to these protein complexes. I both test the complexes individually for selection, as well as the combination of QTL

of all complexes that constitute the proteasome. The list of all genes contributing to the protein complexes were obtained from the PomBase database (Wood et al. 2011). The 20S alpha subunit consists of 5 genes. Three of the genes contribute one state configuration and one gene contributes two state configurations. The selection test gives no significant result for either the Q_1 scenario ($S = 1.27$, $p = 0.31$) or the Q_2 scenario ($S = 1.47$, $p = 0.31$). The 20S beta subunit consists of 7 genes with 6 of them contributing a single state configuration. Here, the selection test gives a significant score for both scenarios ($S = 4.51$, $p = 0.017$ for scenario Q_1 and $S = 6.59$, $p = 0.0056$ for scenario Q_2). Most intriguingly, all of the state configurations consistently have a + state for line 2 and a - state for line 3. So 6 different genes of the same protein complex show a consistent change in expression, displaying a remarkable case of parallel evolution. The 19S subunit consists of 20 genes with 11 resulting state configurations. Also here the selection test gives a positive result for the selective scenario ($S = 3.65$, $p = 0.025$ for scenario Q_1 and $S = 3.65$, $p = 0.035$ for scenario Q_2). Finally, as a consistency check I also apply the selection test on the combined set of state configurations from all proteasome subunits. This can be done since the increase in any gene expression contributing to the complex is expected to act to increase the number of produced proteasome complexes. This results in an even stronger selection result ($S = 8.75$, $p = 1.6 \times 10^{-4}$ for scenario Q_1 and $S = 9.36$, $p = 1.2 \times 10^{-4}$ for scenario Q_2), showing the consistency of the state configurations between the subunits.

protein complex	S	p
proteasome 20S alpha	1.47	0.31
proteasome 20S beta	6.59	0.0056
proteasome 19S	3.65	0.035
proteasome combined	9.36	1.2×10^{-4}
respiratory chain complex I	4.39	0.0042
respiratory chain complex IV	1.29	0.33
respiratory chain combined	5.24	0.013

Table 7.13: Summary of the selection test results for the proteasome and the respiratory chain complexes. In both cases selection is found both on individual subunits/complexes as well as on the combined complex. Score and p -value are given for scenario Q_2 .

The second group of complexes are the respiratory chain complexes. There are four of these complexes that are located in the mitochondrial membrane and are part of the respiratory chain, which is one of the major sources of oxidative stress as already discussed for the stress-exclusive gene modules under selection. Complex I is not found in *S. pombe* but replaced by two mitochondrial NADH dehydrogenases (Friedrich et al. 1995; Joseph-Horne et al. 2001). While one of the 2 genes coding for the NADH dehydrogenases has no state configuration, the other gene has 4 state configurations

resulting in a selective signal ($S = 3.7$, $p = 0.044$ for scenario Q_1 and $S = 4.39$, $p = 0.0042$ for scenario Q_2). Complex II has 4 genes but no state configuration. Complex III has 7 genes but only 3 state configurations, which is not enough to perform the selection test, since at least 4 state configurations are necessary. Complex IV has 11 associated genes and 9 state configurations, but the selective test is not significant here ($S = 1.29$, $p = 0.27$ for scenario Q_1 and $S = 1.29$, $p = 0.33$ for scenario Q_2). When the configurations of all complexes are combined, which allows to incorporate the state configurations of complex III and which tests for the consistency of the direction of change between the complexes, I again retrieve a selective signal (Q_1 : $S = 4.52$, $p = 0.0097$; Q_2 : $S = 5.24$, $p = 0.013$). The evidence for selection is weaker than for the proteasome, but a selective pressure still seems to be present. Taken together, this shows a pervasive selective pressure on biological pathways and complexes related to stress response. The summary of the results can be found in Table 7.13.

The functional analysis for the genes yields a less clear picture. A GO analysis using topGO (Alexa et al. 2006) yielded no significant GO terms for any set of significant genes (static QTL, 'static + conditional stress' QTL, 'static + conditional non-stress' QTL, 'static + conditional stress' exclusive QTL ; all measured 5260 *S. pombe* genes used as background gene set), showing no clear preference for any biological functions. Some of the genes under selection are again connected to the GO terms found for the significant gene modules that are exclusive for the 'static + conditional stress' QTL (regulation of transcription: *cdc10*, *mst1*; translation: *sdo1*, *tif11*; respiratory chain: *idh1*; cell cycle regulation: *cut9*, *apc14*, *cdc10*) but the evidence for the contribution of certain biological functions is clearly lower.

7.4 Protein level changes

A central question remaining to be answered is the question if the selection observed on the gene expression levels is also reflected in the protein levels. As the biological function of the genes is mediated by their protein products, a change of protein levels is expected as a result of selective pressures. In this section, I compare expression and protein levels of the lines obtained from (Clément-Ziza et al. unpublished dataset). Protein levels are available for 2180 of the 5260 *S. pombe* genes and are given in log-scale. The protein levels are normalized and centered around zero. The method used to obtain the protein levels is a combination of mass spectrometry and targeted sequencing (Schmidt et al. 2008) which leads to the measurement of several peptide groups corresponding to a single gene. I use these data to test the hypothesis that for genes under selection a change in expression level is also reflected in the protein level. For this, I divide the genes tested for selection (using the 'static + conditional stress' dataset) into two groups: the genes found to be under selection ($p < 0.05$ for

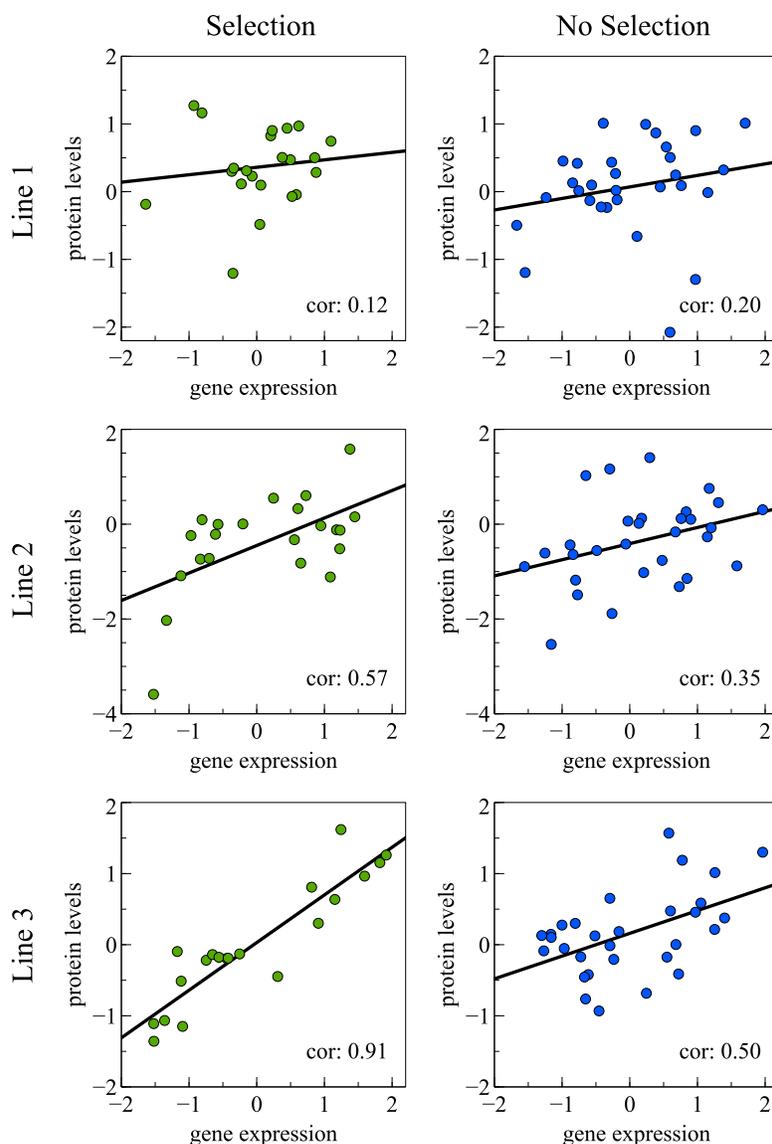


Figure 7.8: **Protein levels in *S. pombe* are stronger correlated with gene expression levels under selection.** Gene expression levels are plotted against protein levels for all three *S. pombe* lines, distinguished between genes under selection and genes not under selection ('static + conditional stress' dataset, selected genes with $p < 0.05$ for scenario Q_2). All gene expression levels as well as all protein levels are normalized and centered around zero (of which only a subset is plotted here). For the genes under selection a stronger correlation could be found between gene expression and protein levels for lines 2 and 3, meaning that a change in expression level has a stronger effect on the protein levels. This is expected if the selective forces not only act to change expression levels but also protein levels as the product of a gene. No clear signal could be found for line 1 which is the the laboratory line. This might point towards the absence of strong selective forces in this line, which could be connected to the constant, well-defined environment experienced by the laboratory line.

scenario Q_2) and those not under selection. If selection is acting on the protein levels as well, a high or low expression level would also lead to a high or low protein level. When no selection is present, a less tight connection between gene expression levels and protein levels might be assumed, as protein levels are often buffered and follow gene expression levels only to a lower extent (Holcik and Sonenberg 2005). For both groups and for each line I plot the gene expression levels against the mean protein levels averaged over all peptides that are assigned to a gene (if the protein level is available at all), see Figure 7.8. It shows that the protein levels for genes under selection are more strongly correlated with the expression levels for line 2 and especially for line 3, which are the lines that showed the strongest level of trait divergence (see Table 7.8). A linear regression performed on the data also shows that the slope of the fitted line is steeper for the genes under selection, such that an increase in gene expression leads to a stronger increase in protein level. From this one can conclude that the selection acting on the gene expression levels is also reflected in the protein levels. Genes that are under selection for a higher or lower gene expression level are also showing higher or lower protein levels. These results are robust with respect to the choice of the threshold for selection; choosing a threshold as low as $p < 0.01$ leads to very similar correlation coefficients ($r = 0.90$ for line 3) while an increase of the threshold leads to a slow decrease in the correlation.

Line 1 shows no clear correlation between gene expression levels and protein levels. This is in line with the result already seen for the selection test in section 7.2.5, where line 1 was almost never the diverged line. This might reflect that line 1, the laboratory strain, is cultured under constant conditions such that selective forces as well as the need for a tight control of protein levels might play a less important role.

The verification of the results of the selection test on the protein levels shows that the selective forces leave traces both on the transcriptional as well as the translational level. It might be possible that the protein level is the main target of selection.

7.5 Possible use of protein QTL data

The selection test could be extended to data on protein QTL (pQTL). pQTL would allow to test selection acting not only on the transcriptional level but also on the translational level. Genes that are affected by both eQTL and pQTL would allow to examine the consistency of state configurations obtained from both datasets. For genes found to be under selection using eQTL only, pQTL state configurations could be added to the analysis to test if this increases the selection score (in which case the pQTL state configurations would be in line with the eQTL configurations). A problem arising in this case would be that additive effects for the two kinds of QTL are not directly comparable, as the additive effects for the eQTL are given in units of mRNA

produced, while the additive effects for the pQTL are given in units of the resulting protein level. A possible solution would be to calculate the effects of each eQTL on the protein levels, such that all additive effects can be given on the level of protein changes.

A different ansatz would be to focus on the pQTL data only and to divide them into two groups: pQTL with known eQTL (reflecting transcriptional changes) and without known eQTL (translational changes). If for a pQTL an eQTL is known at the same marker position, it is likely that the original change happened on the transcriptional level and is passed on to the translational level. If no known eQTL exists this might either reflect a change in protein level that happens after the transcription, or it might be that the eQTL is present but was not detected due to e.g. insufficient power of the eQTL study or a too small additive effect. Using this classification, the selection score could be split up as $S_{\text{tot}} = S_{\text{transcription}} + S_{\text{translation}}$ with a score contribution $S_{\text{transcription}}$ from changes on the transcriptional level (pQTL with known eQTL) and a contribution $S_{\text{translation}}$ from the translational level (pQTL without known eQTL). This would allow to quantify the dominant modes of selection, i.e. it could be tested if there is a majority of changes on the transcriptional or the translational level.

7.6 Comparison to other selection tests

In this section, I will compare the results of my selection test for the yeast eQTL dataset to other available selection tests. For comparison I will use the Orr test (Orr 1998), and the test of Fraser, Moses, and Schadt (Fraser et al. 2010).

7.6.1 Orr test

The test of Orr (Orr 1998) employs a similar strategy as our selection test (for more details see also section 3.7) but is only restricted to two lines and does not implement a population genetics model comparing neutral and selective scenario. Instead it tests against a simple neutral null model with equal probabilities for the + and – states. As for the plant QTL datasets in section 6.1 before, I apply the Orr test to the two lines with the largest trait divergence for each gene. I only consider the QTL with a valid state configuration for three lines, neglecting QTL that do not follow the two-state assumption or show epistatic behaviour. This excludes some QTL that would have a valid state configuration for two lines but not for three lines and leads to less genes with enough QTL for a selection test. Yet, this leads to a better comparability to the results of the three-line test, as QTL with possible epistatic interactions that might create a bias in the two-line data are neglected.

In contrast to the selection test presented in this thesis the Orr test does not find a sign of selection, being unable to reject the neutral null hypothesis at a p -value threshold

of 0.05 even for a single gene, as can be seen in Figure 7.9. Yet, the p -value distribution is biased to lower p -values, suggesting that selection could possibly be detected if more QTL would be known for the individual genes. When the previously neglected QTL that have an invalid three-line state configuration but a valid two-line configuration are incorporated as well, leading to more QTL per gene and more testable genes, very similar results are obtained with again no gene found to be under selection (data not shown). This result shows the advantage of a multiple-line test for the detection of selection.

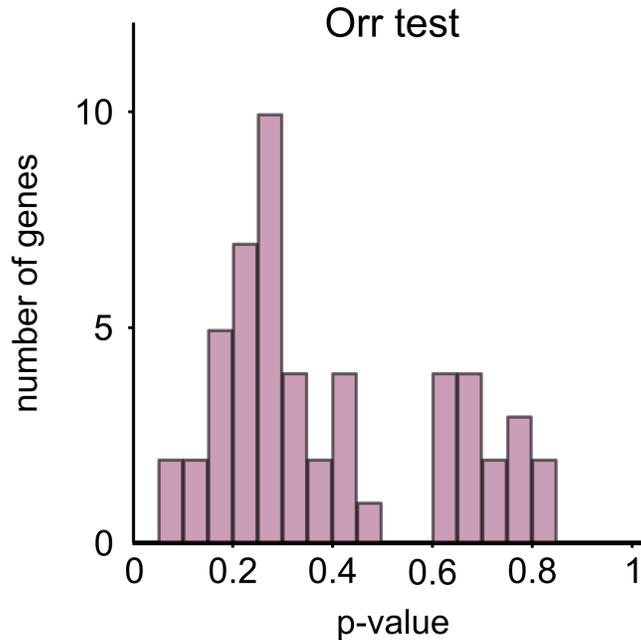


Figure 7.9: **The Orr test cannot reject neutral evolution for the *S. pombe* eQTL dataset.** The Orr test (Orr 1998) is applied to the two *S. pombe* lines with the highest trait divergence for each gene, using only the QTL that also have a valid three-line state configuration. The Orr test cannot reject the null hypothesis of neutral evolution for a single gene. So this two-line test cannot find the pervasive selection pressure on the gene expression levels found with the multiple-line selection test.

7.6.2 Test for genome-wide level of selection

The test of Fraser, Moses, and Schadt (Fraser et al. 2010) uses pairs of *cis*- and *trans*-eQTL affecting a single gene expression level to test for the global level of selection. If more reinforcing *cis*- and *trans*-QTL pairs (with state configurations affecting the

expression level in the same direction) than opposing QTL pairs (with state configurations acting in different directions) were found across all genes, this points to an overall excess of selection acting on the gene expression levels. As only pairs of eQTL per gene were considered, the test cannot identify individual genes under selection but determine the global level of selection. With the large number of eQTL found in the *S. pombe* dataset used in this chapter, the identification of individual genes under selection becomes possible.

To apply the test on the *S. pombe* dataset, I determine for each QTL with valid state configuration, if it is a *cis*- or *trans*-QTL (with a *cis*-QTL within the range of 50 kb around the considered gene as in (Fraser et al. 2010); I do not require the *trans*-QTL to be on a different chromosome than the gene since *S. pombe* only has 3 chromosomes as opposed to the 16 chromosomes of *S. cerevisiae* used in (Fraser et al. 2010)). As the majority of eQTL found were *trans*-QTL (Clément-Ziza et al. 2014), many genes have multiple *trans*-QTL. For each gene with multiple *trans*-QTL, I only pick the QTL with the highest statistical significance. In total, I obtain 72 genes with a pair of *cis*- and *trans*-QTL. For the further analysis I only consider the state configurations of lines 2 and 3, as these showed the highest level of trait divergence (see section 7.2.2) and since line 1 had problems recombining with the other lines (see section 7.1.2). The number of reinforcing and opposing QTL pairs can be found in the contingency table in Table 7.14. Under neutral evolution the same number of QTL pairs would be expected in all categories (which would be 18 pairs per category in this case). This shows a clear excess of reinforcing pairs (chi-square test: $p = 0.002$) with 14 more reinforcing pairs as expected by chance (corresponding to 19% of all 72 pairs) and with 28 more reinforcing than opposing pairs (39% of all 72 pairs). This is in perfect agreement with the levels of selection found by the multiple-line selection test (see section 7.2.2), supporting the hypothesis of widespread adaptation on gene expression levels in *S. pombe*.

	<i>cis</i> $q_2 > q_3$	<i>cis</i> $q_2 < q_3$
<i>trans</i> $q_2 > q_3$	22	10
<i>trans</i> $q_2 < q_3$	12	28

Table 7.14: Number of reinforcing and opposing *cis*- and *trans*-QTL pairs for the *S. pombe* lines 2 and 3. Reinforcing QTL pairs affect the trait in the same direction (states $q_2 > q_3$ or $q_2 < q_3$ in both cases; upper left and lower right number) while opposing QTL pairs affect the trait in different directions (states $q_2 > q_3$ in one line and $q_2 < q_3$ in the other line; upper right and lower left number) A clear excess of reinforcing QTL pairs can be seen compared to a scenario of neutral evolution, where 18 pairs would be expected in each of the four categories. 14 more reinforcing pairs are seen as expected by chance, corresponding to 19% of all 72 pairs.

Chapter 8

Conclusions

In this thesis, I combine information from QTL experiments with population genetics theory to infer the evolutionary history of quantitative traits. I define a model of trait evolution that is based on population genetics theory and that quantifies the evidence for different evolutionary scenarios of a trait. A log-likelihood score weights different selective scenarios against each other and numerical simulations show that the selection test can distinguish between neutral evolution and selection as well as between different lineage-specific selection scenarios. In addition to the long evolutionary times that are assumed in the original model, I also derive the state statistics for two and three lines in the limit of short evolutionary times.

A major advance compared to previous selection tests is the usage of multiple QTL lines. I show that multiple lines clearly increase the statistical power to detect selection by increasing the genetic diversity that is uncovered. An explicit expression for the increase in statistical power to detect selection as a function of the number of lines is given. Additionally, a scenario is explored where a fixed number of QTL crosses is either performed on two lines or distributed on crosses between all possible pairs of three lines. I find that for a sufficient number of crosses it can be beneficial for the detection of selection to distribute the crosses on three lines, as more QTL can be detected.

Applying the selection test to very large numbers of lines poses interesting challenges in connection with the number of alleles per locus and the rapid growth of the number of possible evolutionary scenarios. At the same time, the need for experimental crosses between three or more different lines is a major bottleneck of the multiple-line test. Due to the additional experimental work involved, there are currently few datasets on QTL and their additive effects in more lines than two. However recent studies employing crosses of 25 maize lines and detecting around 30-40 QTL per trait (Buckler et al. 2009; Tian et al. 2011) as well as the three-line yeast eQTL dataset analyzed in this thesis, detecting QTL acting on all gene expression levels of the organism, give a

promising outlook to the future.

To address the problem of ascertainment bias that arises when multiple traits are tested for selection, I use, besides the established Holm-Bonferroni correction and the Benjamini-Hochberg procedure, the conditioning on the phenotypic difference as proposed by Orr (Orr 1998). I show that the state statistics conditioned on the phenotypic difference can be obtained using the maximum entropy principle. It turns out that for two lines selection cannot be distinguished from the conditioned state statistics under neutral evolution. Using three or more lines, however, a distinction between neutral evolution and selection is easily possible. I derive the conditioned state statistics and explore the capabilities of all methods to reduce the false positive rate in a multiple testing scenario.

The developed selection test can be a major tool helping to uncover the genetic basis of adaptation of quantitative traits. One of the most interesting applications might be studies of expression QTL or protein QTL. Using such datasets, the genome-wide level of selection on transcription and translation can be uncovered and selective pressures specific to certain conditions, like changing stress levels or light periods, can be detected. This might help closing the gap between changes observed in macroscopic phenotypes, like yield traits in crops, and the genetic changes on the level of individual genes or pathways, leading to a more detailed picture of the selective forces and their action on the level of the genome.

Here, the developed selection test is applied on several multiple-line QTL datasets. For both tested plant QTL studies, one on photoperiodic response traits in Maize (Coles et al. 2010) and one on *Mimulus* floral traits (Chen 2009), I find a signature of lineage-specific selection not seen in two-line tests. A QTL study on gene expression levels in yeast (Clément-Ziza et al. unpublished dataset) is used to uncover the genetic basis of adaptation to stress response. A high level of selection on both genes as well as gene modules that consist of several genes and that have clearly defined biological functions, like protein complexes or pathways, could be detected. Unlike in most other studies on QTL adaptation, individual genes under selection could be identified. The biology behind the gene modules found to be under selection exclusively in the stress condition is explored and interesting connections to known stress response mechanisms are uncovered. A strong connection between expression levels and protein levels is found for genes under selection, with the transcriptional changes also confirmed on the level of translation, hinting towards protein levels as the actual target of selection.

A possible further application of my selection test is the inference of adaptation on combined datasets of expression QTL and protein QTL. With this, the action of selection on the tightly coupled levels of transcription and translation could be studied, separating the contributions of selection acting on changes of expression levels as well as protein levels and quantifying how the action of selection on gene expression levels is reflected in changes of protein levels.

Appendix A

A.1 Calculation of the state configurations of the eQTL

The step necessary before the selection test can be applied is to calculate the three-line state configurations (q_1, q_2, q_3) for each of the QTL. Here, I have to consider that the two-state assumption and the assumption of additive trait effects needs to be fulfilled. If for a locus more than two significantly different alleles exist (i.e. a locus has a different effect in each of the lines) or if the pairwise trait differences between the lines are not compatible with a single, additive trait effect, I neglect that locus in the following. This step is necessary in order to be able to apply the framework of my selection test. Yet, one has to be careful with this step as this can be a potential source of bias. In the following, I also explore how many QTL follow an additive three-state model and what fraction of the loci show epistasis.

To obtain the state configurations q_i , only information from crosses between all possible pairs of the lines is available. To test if a given QTL fulfills the additive two-state assumption, consistency between the pairwise crosses needs to be checked. Starting with a significant QTL-gene pair, I calculate the additive effects for each of the pairwise crosses. For a given QTL I first look up which of the crosses between line 1 and 2 have the genotype of line 1 and which the genotype of line 2. Then I group the gene expression levels according to their genotype at the QTL marker position and calculate the mean gene expression levels and the standard deviation for each group. The difference between the mean gene expression levels of alleles 1 and 2 gives the estimate for the additive effect a_{12} together with its uncertainty $\sigma_{a_{12}}$ of the line pair 1 and 2. The same is repeated for the pairwise crosses of lines 1 and 3 as well as lines 2 and 3, resulting in additive effects a_{13} and a_{23} and uncertainties $\sigma_{a_{13}}$ and $\sigma_{a_{23}}$ for each of the pairwise crosses.

Next, I have to check the consistency of these additive effects. I first take the pair of lines with the largest difference in gene expression and assign a + state to the line with the higher gene expression level and the - state to the line with the lower gene expression level. In order to obtain a valid configuration the remaining two crosses need

to be consistent with this. For both remaining crosses I check if a) the additive effects are consistent to the largest additive effect within error (e.g. $|a_{12} - a_{13}| < \sqrt{\sigma_{a_{12}}^2 + \sigma_{a_{13}}^2}$ if a_{12} is the largest additive effect) and b) if a is consistent with zero (e.g. $|a_{13}| < \sigma_{a_{13}}$). It should be noted that these two points are non-exclusive (e.g. one of the crosses could have such a high uncertainty in its additive effect that it is both consistent with the additive effect of the other cross and consistent with zero). One gets a valid state configuration (q_1, q_2, q_3) for the locus if a) is true for one of the remaining crosses and b) is true for the other remaining cross.

For example if the cross of line 1 and line 2 has the highest expression difference and expression is higher in line 1, I assign $q_1 = +$ and $q_2 = -$ such that only the state of q_3 needs to be deduced. If the cross of lines 1 and 3 gives a trait difference that is consistent with zero (fulfilling condition b), one can deduce that $q_3 = q_1 = +$. If furthermore the cross of lines 2 and 3 gives an expression difference a_{23} consistent with a_{12} (fulfilling condition a), this is consistent with $q_3 = -q_2 = +$. This yields the final state configuration $(q_1, q_2, q_3) = (+, -, +)$.

The additive effect for a locus is calculated as the mean of the two additive effects of the pairwise crosses with significantly different states (e.g. a_{12} and a_{23} in the case of a state configuration $(+, -, +)$).

In some cases it is not possible to calculate one of the additive effects, e.g. if all markers at the considered position on the chromosome originate from the same line, such that no difference between the mean expression levels of the two lines can be calculated. In these cases I only use the two remaining pairwise crosses to determine the state configuration of the locus. I apply the same procedure when there is only one cross with a different marker, as it is the case for one large region in the genome where there is no recombination between lines 1 and 3 and all crosses but one carry the alleles of line 1 in this region. This region coincides with a chromosomal inversion in line 1 (Brown et al. 2011; Clément-Ziza et al. 2014).

A.2 Calculation of p -values for the eQTL scores

To obtain a p -value for each of the scores, a null distribution of scores is calculated separately for each gene by repeatedly simulating instances of the null model for the given number of QTL and given additive effects. For this, two possibilities exist. Either the state configurations are directly drawn from scenario P_0 , or the state configurations are obtained by randomly reshuffling the state configurations of the QTL within the dataset (i.e. all state configurations of all the genes are reshuffled, allowing exchange of configurations between the genes). The second option has the advantage that it leaves the overall distribution of state configurations unchanged, including possible biases in the distribution. When looking at the relative frequency of the different state

configurations in the dataset, no bias towards any configuration is observed. With only deviations of about 1% around the expected mean of 16.7% for each of the six possible state configurations no significant deviation from a uniform distribution for the state configurations is present ((+, +, -): 15.3%, (+, -, +): 16.8%, (-, +, +): 16.3%, (+, -, -): 17.2%, (-, +, -): 17.6%, (-, -, +): 16.7%, chi-squared test: $p = 0.95$). Thus, the two options to obtain the state configurations for the null model are approximately equal. For simplicity, I implement the first option, calculating the null distribution of scores from state configurations drawn from P_0 . The p -value for each gene is then calculated as the probability that the observed score is obtained from the null distribution, as already described in section 3.4.1.

Bibliography

- M. Ackermann, W. Sikora-Wohlfeld, and A. Beyer. Impact of natural genetic variation on gene expression dynamics. *PLoS Genet*, 9(6):e1003514, 2013.
- C. T. Aiken, R. M. Kaake, X. Wang, and L. Huang. Oxidative stress-mediated regulation of proteasome complexes. *Molecular & Cellular Proteomics*, 10(5):R110–006924, 2011.
- A. Alexa, J. Rahnenführer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607, 2006.
- E. C. Anderson and M. Slatkin. Orr’s quantitative trait loci sign test under conditions of trait ascertainment. *Genetics*, 165(1):445–446, 2003.
- R. I. Astuti, D. Watanabe, and H. Takagi. Nitric oxide signaling and its role in oxidative stress response in *Schizosaccharomyces pombe*. *Nitric Oxide*, 52:29–40, 2016.
- M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, and B. Kuster. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry*, 389(4):1017–1031, 2007.
- N. H. Barton and J. Coe. On the application of statistical physics to evolutionary biology. *Journal of theoretical biology*, 259(2):317–324, 2009.
- N. H. Barton and H. P. de Vladar. Statistical mechanics and the evolution of polygenic quantitative traits. *Genetics*, 181(3):997–1011, 2009.
- Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.
- J. Berg, S. Willmann, and M. Lässig. Adaptive evolution of transcription factor binding sites. *BMC Evol. Biol.*, 4(1):42, 2004.

- A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71, 1996.
- J. Bernier, A. Kumar, V. Ramaiah, D. Spaner, and G. Atlin. A large-effect QTL for grain yield under reproductive-stage drought stress in upland rice. *Crop Science*, 47(2):507–516, 2007.
- G. Blanc, A. Charcosset, B. Mangin, A. Gallais, and L. Moreau. Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize. *Theor. Appl. Genet.*, 113(2):206–224, 2006.
- J. Boonstra. *Regulation of G1 Phase Progression*. Springer Science & Business Media, 2003.
- R. B. Brem and L. Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. USA*, 102(5):1572–1577, 2005.
- W. R. Brown, G. Liti, C. Rosa, S. James, I. Roberts, V. Robert, N. Jolly, W. Tang, P. Baumann, C. Green, et al. A geographically diverse collection of *Schizosaccharomyces pombe* isolates shows limited phenotypic variation but extensive karyotypic diversity. *G3: Genes, Genomes, Genetics*, 1(7):615–626, 2011.
- E. S. Buckler, J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown, C. Browne, E. Ersoz, S. Flint-Garcia, A. Garcia, J. C. Glaubitz, M. M. Goodman, C. Harjes, K. Guill, D. E. Kroon, S. Larsson, N. K. Lepak, H. Li, S. E. Mitchell, G. Pressoir, J. A. Peiffer, M. O. Rosas, T. R. Rocheford, M. C. Romay, S. Romero, S. Salvo, H. S. Villeda, H. Sofia da Silva, Q. Sun, F. Tian, N. Upadyayula, D. Ware, H. Yates, J. Yu, Z. Zhang, S. Kresovich, and M. D. McMullen. The genetic architecture of maize flowering time. *Science*, 325(5941):714–718, 2009.
- J. H. Bullard, Y. Mostovoy, S. Dudoit, and R. B. Brem. Polygenic and directional regulatory evolution across pathways in *Saccharomyces*. *Proc. Natl. Acad. Sci. USA*, 107(11):5058–5063, 2010.
- C. Chen. *Lineage specific inference about QTL evolution among three Mimulus species of contrasting relationship and inbreeding*. PhD thesis, University of British Columbia, 2009.
- M. Clément-Ziza, F. X. Marsellach, S. Codlin, M. A. Papadakis, S. Reinhardt, M. Rodríguez-López, S. Martin, S. Marguerat, A. Schmidt, E. Lee, C. T. Workman, J. Bähler, and A. Beyer. Natural genetic variation impacts expression levels of coding, non-coding, and antisense transcripts in fission yeast. *Molecular systems biology*, 10(11):764, 2014.

- M. Clément-Ziza, J. Großbach, and A. Beyer. unpublished dataset.
- N. D. Coles, M. D. McMullen, P. J. Balint-Kurti, R. C. Pratt, and J. B. Holland. Genetic control of photoperiod sensitivity in maize revealed by joint multiple population analysis. *Genetics*, 184(3):799–812, 2010.
- L. P. Collinson and I. W. Dawes. Inducibility of the response of yeast cells to peroxide stress. *Microbiology*, 138(2):329–335, 1992.
- F. Crick, L. Barnett, S. Brenner, and R. J. Watts-Tobin. *General nature of the genetic code for proteins*. Macmillan Journals Limited, 1961.
- F. Crick et al. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.
- F. A. Cubillos, E. Billi, E. Zörgö, L. Parts, P. Fargier, S. Omholt, A. Blomberg, J. Warringer, E. J. Louis, and G. Liti. Assessing the complex architecture of polygenic traits in diverged yeast populations. *Molecular ecology*, 20(7):1401–1413, 2011.
- M. De Luca, N. V. Roshina, G. L. Geiger-Thornsberry, R. F. Lyman, E. G. Pasyukova, and T. F. C. Mackay. Dopa decarboxylase (Ddc) affects variation in *Drosophila* longevity. *Nat. Genet.*, 34(4):429–433, 2003.
- H. P. de Vladar and N. H. Barton. The statistical mechanics of a polygenic character under stabilizing selection, mutation and drift. *Journal of The Royal Society Interface*, 8(58):720–739, 2011a.
- H. P. de Vladar and N. H. Barton. The contribution of statistical physics to evolutionary biology. *Trends Ecol. Evol.*, 26:424–432, 2011b.
- F. Delsuc, H. Brinkmann, and H. Philippe. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6(5):361–375, 2005.
- C. L. Dilda and T. F. C. Mackay. The genetic architecture of *Drosophila* sensory bristle number. *Genetics*, 162(4):1655–1674, 2002.
- R. W. Doerge. Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics*, 3(1):43–52, 2002.
- R. W. Doerge and G. A. Churchill. Permutation tests for multiple loci affecting a quantitative character. *Genetics*, 142(1):285–294, 1996.
- J. J. Fanara, K. O. Robinson, S. M. Rollmann, R. R. H. Anholt, and T. F. C. Mackay. Vanaso is a candidate quantitative trait gene for *Drosophila* olfactory behavior. *Genetics*, 162(3):1321–1328, 2002.

- J. Flint and T. F. Mackay. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res.*, 19(5):723–733, 2009.
- H. B. Fraser. Genome-wide approaches to the study of adaptive gene expression evolution. *Bioessays*, 33(6):469–477, 2011.
- H. B. Fraser, A. M. Moses, and E. E. Schadt. Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proc. Natl. Acad. Sci. USA*, 107(7):2977–2982, 2010.
- H. B. Fraser, T. Babak, J. Tsang, Y. Zhou, B. Zhang, M. Mehrabian, and E. E. Schadt. Systematic detection of polygenic cis-regulatory evolution. *PLoS Genet*, 7(3):e1002023, 2011.
- T. Friedrich, K. Steinmüller, and H. Weiss. The proton-pumping respiratory complex I of bacteria and mitochondria and its homologue in chloroplasts. *FEBS letters*, 367(2):107–111, 1995.
- J. Gerke, K. Lorenz, and B. Cohen. Genetic interactions between transcription factors cause natural variation in yeast. *Science*, 323(5913):498–501, 2009.
- P. J. Gerrish and R. E. Lenski. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102:127–144, 1998.
- M. E. Goddard and B. J. Hayes. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics*, 10(6):381–391, 2009.
- C. Godon, G. Lagniel, J. Lee, J.-M. Buhler, S. Kieffer, M. Perrot, H. Boucherie, M. B. Toledano, and J. Labarre. The H₂O₂ stimulon in *Saccharomyces cerevisiae*. *Journal of Biological Chemistry*, 273(35):22480–22489, 1998.
- S. T. Harbison, A. H. Yamamoto, J. J. Fanara, K. K. Norga, and T. F. C. Mackay. Quantitative trait loci affecting starvation resistance in *Drosophila melanogaster*. *Genetics*, 166(4):1807–1823, 2004.
- D. L. Hartl, A. G. Clark, and A. G. Clark. *Principles of population genetics*, volume 116. Sinauer associates Sunderland, 1997.
- E. Herrero, J. Ros, G. Bellí, and E. Cabiscol. Redox control and oxidative stress in yeast cells. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1780(11):1217–1235, 2008.

- J. D. Hoheisel. Microarray technology: beyond transcript profiling and genotype analysis. *Nature reviews genetics*, 7(3):200–210, 2006.
- M. Holcik and N. Sonenberg. Translational control in stress and apoptosis. *Nature reviews Molecular cell biology*, 6(4):318–327, 2005.
- S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.
- Y. Iwasa. Free fitness that always increases in evolution. *J. Theor. Biol.*, 135(3):265 – 281, 1988.
- E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, 1957.
- D. C. Jeffares, C. Rallis, A. Rieux, D. Speed, M. Převorovský, T. Mourier, F. X. Marsellach, Z. Iqbal, W. Lau, T. M. Cheng, et al. The genomic and phenotypic diversity of *Schizosaccharomyces pombe*. *Nature genetics*, 47(3):235–241, 2015.
- C. Jiang and Z.-B. Zeng. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, 140(3):1111–1127, 1995.
- K. W. Jordan, T. J. Morgan, and T. F. C. Mackay. Quantitative trait loci for locomotor behavior in *Drosophila melanogaster*. *Genetics*, 174(1):271–284, 2006.
- T. Joseph-Horne, D. W. Hollomon, and P. M. Wood. Fungal respiration: a fusion of standard and alternative components. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1504(2):179–195, 2001.
- C.-H. Kao, Z.-B. Zeng, and R. D. Teasdale. Multiple interval mapping for quantitative trait loci. *Genetics*, 152(3):1203–1216, 1999.
- M. Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47(6):713–719, 1962.
- M. King and A. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116, 1975.
- R. Lande. Natural selection and random genetic drift in phenotypic evolution. *Evolution*, pages 314–334, 1976.
- E. S. Lander and D. Botstein. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1):185–199, 1989.

- A. Larièpe, B. Mangin, S. Jasson, V. Combes, F. Dumas, P. Jamin, C. Lariagon, D. Jolivot, D. Madur, J. Fievet, et al. The genetic basis of heterosis: multiparental quantitative trait loci mapping reveals contrasted levels of apparent overdominance among traits of agronomical interest in maize (*Zea mays* L.). *Genetics*, 190(2): 795–811, 2012.
- M. Lässig. From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinformatics*, 8(Suppl 6):S7, 2007.
- U. Leupold. Die Vererbung von Homothallie und Heterothallie bei *Schizosaccharomyces pombe*. *CR Trav. Lab. Carlsberg Ser. Physiol.*, 24:381–480, 1950.
- M. Lynch, B. Walsh, et al. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA, 1998.
- T. F. Mackay. The genetic architecture of quantitative traits: lessons from *Drosophila*. *Current opinion in genetics & development*, 14(3):253–257, 2004.
- T. F. Mackay. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Reviews Genetics*, 15(1):22–33, 2014.
- T. F. Mackay and R. F. Lyman. *Drosophila* bristles and the nature of quantitative genetic variation. *Philos. T. Roy. Soc. B*, 360(1459):1513–1527, 2005.
- T. F. C. Mackay, E. A. Stone, and J. F. Ayroles. The genetics of quantitative traits: challenges and prospects. *Nat. Rev. Genet.*, 10:565–577, 2009.
- P. Marullo, M. Aigle, M. Bely, I. Masneuf-Pomarede, P. Durrens, D. Dubourdieu, and G. Yvert. Single QTL mapping and nucleotide-level resolution of a physiologic trait in wine *Saccharomyces cerevisiae* strains. *FEMS yeast research*, 7(6):941–952, 2007.
- G. Mendel. Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brunn 4: 3*, 44, 1866.
- M. L. Metzker. Sequencing technologies - the next generation. *Nature reviews genetics*, 11(1):31–46, 2010.
- J. G. Mezey, D. Houle, and S. V. Nuzhdin. Naturally segregating quantitative trait loci affecting wing shape of *Drosophila melanogaster*. *Genetics*, 169(4):2101–2113, 2005.
- J. J. Michaelson, S. Loguercio, and A. Beyer. Detection and interpretation of expression quantitative trait loci (eQTL). *Methods*, 48(3):265–276, 2009.

- J. J. Michaelson, R. Alberts, K. Schughart, and A. Beyer. Data-driven assessment of eQTL mapping methods. *BMC genomics*, 11(1):502, 2010.
- A. J. Moehring and T. F. C. Mackay. The quantitative genetic basis of male mating behavior in *Drosophila melanogaster*. *Genetics*, 167(3):1249–1263, 2004.
- T. Mora and W. Bialek. Are biological systems poised at criticality? *Journal of Statistical Physics*, 144(2):268–302, 2011.
- V. Mustonen and M. Lässig. Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. *Proc. Natl. Acad. Sci. USA*, 102(44):15936–15941, 2005.
- V. Mustonen and M. Lässig. Molecular evolution under fitness fluctuations. *Physical review letters*, 100(10):108101, 2008.
- V. Mustonen and M. Lässig. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends in Genetics*, 25(3):111–119, 2009.
- V. Mustonen and M. Lässig. Fitness flux and ubiquity of adaptive evolution. *Proceedings of the National Academy of Sciences*, 107(9):4248–4253, 2010.
- V. Mustonen, J. Kinney, C. G. Callan, and M. Lässig. Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites. *Proc. Natl. Acad. Sci. USA*, 105(34):12376–12381, 2008.
- R. Narayan and R. Nityananda. Maximum entropy image restoration in astronomy. *Annu. Rev. Astron. Astr.*, 24(1):127–170, 1986.
- A. Nourmohammad, T. Held, and M. Lässig. Universality and predictability in molecular quantitative genetics. *Curr. Opin. Genet. Dev.*, 23(6):684 – 693, 2013a.
- A. Nourmohammad, S. Schiffels, and M. Lässig. Evolution of molecular phenotypes under stabilizing selection. *J. Stat. Mech. - Theory E.*, 2013, 2013b.
- H. A. Orr. Testing natural selection vs. genetic drift in phenotypic evolution using quantitative trait locus data. *Genetics*, 149(4):2099–2104, 1998.
- H. A. Orr. The genetic theory of adaptation: a brief history. *Nature Reviews Genetics*, 6(2):119–127, 2005.
- H. A. Orr. Fitness and its role in evolutionary genetics. *Nature Reviews Genetics*, 10(8):531–539, 2009.

- E. G. Pasyukova, C. Vieira, and T. F. C. Mackay. Deficiency mapping of quantitative trait loci affecting longevity in *Drosophila melanogaster*. *Genetics*, 156(3):1129–1146, 2000.
- R. Plomin, C. M. Haworth, and O. S. Davis. Common disorders are quantitative traits. *Nature Reviews Genetics*, 10(12):872–878, 2009.
- A. Prügel-Bennett and J. L. Shapiro. Analysis of genetic algorithms using statistical mechanics. *Phys. Rev. Lett.*, 72:1305–1309, 1994.
- A. Prügel-Bennett and J. L. Shapiro. The dynamics of a genetic algorithm for simple random Ising systems. *Physica D*, 104(1):75 – 114, 1997.
- A. Rebai and B. Goffinet. Power of tests for QTL detection using replicated progenies derived from a diallel cross. *Theor. Appl. Genet.*, 86(8):1014–1022, 1993.
- A. Rebai and B. Goffinet. More about quantitative trait locus mapping with diallel designs. *Genet. Res.*, 75:243–247, 2000.
- A. Rebai, P. Blanchard, D. Perret, and P. Vincourt. Mapping quantitative trait loci controlling silking date in a diallel cross among four lines of maize. *Theoretical and applied genetics*, 95(3):451–459, 1997.
- D. P. Rice and J. P. Townsend. A test for selection employing quantitative trait locus and mutation accumulation data. *Genetics*, 190(4):1533–1545, 2012a.
- D. P. Rice and J. P. Townsend. Resampling QTL effects in the QTL sign test leads to incongruous sensitivity to variance in effect size. *G3*, 2(8):905–911, 2012b.
- I. G. Romero, I. Ruvinsky, and Y. Gilad. Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics*, 13(7):505–516, 2012.
- J. A. Rosenfeld, Z. Wang, D. E. Schones, K. Zhao, R. DeSalle, and M. Q. Zhang. Determination of enriched histone modifications in non-genic portions of the human genome. *BMC genomics*, 10(1):143, 2009.
- M. Ruttray. The dynamics of a genetic algorithm under stabilizing selection. *Complex Syst.*, 9(3):213–234, 1995.
- C. J. Ryan, A. Roguev, K. Patrick, J. Xu, H. Jahari, Z. Tong, P. Beltrao, M. Shales, H. Qu, S. R. Collins, et al. Hierarchical modularity and the evolution of genetic interactomes across species. *Molecular cell*, 46(5):691–704, 2012.

- A. Schmidt, N. Gehlenborg, B. Bodenmiller, L. N. Mueller, D. Campbell, M. Mueller, R. Aebersold, and B. Domon. An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. *Molecular & Cellular Proteomics*, 7(11):2138–2150, 2008.
- G. Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6(2):461–464, 1978.
- G. Sella and A. E. Hirsh. The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci. USA*, 102(27):9541–9546, 2005.
- M. Sipiczki. Where does fission yeast sit on the tree of life. *Genome Biol*, 1(2):1011–1, 2000.
- D. J. Somers, P. Isaac, and K. Edwards. A high-density microsatellite consensus map for bread wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics*, 109(6):1105–1114, 2004.
- L. F. Stam and C. C. Laurie. Molecular dissection of a major gene effect on a quantitative trait: the level of alcohol dehydrogenase expression in *Drosophila melanogaster*. *Genetics*, 144(4):1559–1564, 1996.
- J. Steinhoff, W. Liu, H. P. Maurer, T. Würschum, H. L. C. Friedrich, N. Ranc, and J. C. Reif. Multiple-line cross quantitative trait locus mapping in European elite maize. *Crop Sci.*, 51:2505–2516, 2011.
- A. J. Stewart and J. B. Plotkin. The evolution of complex gene regulation by low-specificity binding sites. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1768):20131313, 2013.
- F. Tian, P. J. Bradbury, P. J. Brown, H. Hung, Q. Sun, S. Flint-Garcia, T. R. Rocheford, M. D. McMullen, J. B. Holland, and E. S. Buckler. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.*, 43(2):159–162, 2011.
- G. P. Wagner and J. Zhang. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nature Reviews Genetics*, 12(3):204–213, 2011.
- D. Wang, J. Zhu, Z. Li, and A. Paterson. Mapping QTLs with epistatic effects and QTL \times environment interactions by mixed linear model approaches. *Theoretical and Applied Genetics*, 99(7-8):1255–1264, 1999.
- M. C. Whitlock, P. C. Phillips, F. B.-G. Moore, and S. J. Tonsor. Multiple fitness peaks and epistasis. *Annual Review of Ecology and Systematics*, pages 601–629, 1995.

- S. R. Wicks, R. T. Yeh, W. R. Gish, R. H. Waterston, and R. H. Plasterk. Rapid gene mapping in *Caenorhabditis elegans* using a high density polymorphism map. *Nature genetics*, 28(2):160–164, 2001.
- C. O. Wilke. The speed of adaptation in large asexual populations. *Genetics*, 167(4):2045–2053, 2004.
- C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.
- V. Wood, M. A. Harris, M. D. McDowall, K. Rutherford, B. W. Vaughan, D. M. Staines, M. Aslett, A. Lock, J. Bähler, P. J. Kersey, et al. Pombase: a comprehensive online resource for fission yeast. *Nucleic acids research*, page gkr853, 2011.
- S. Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931.
- Y. Xing and Q. Zhang. Genetic and molecular bases of rice yield. *Annual review of plant biology*, 61:421–442, 2010.
- Z. B. Zeng. Correcting the bias of Wright’s estimates of the number of genes affecting a quantitative character: a further improved method. *Genetics*, 131(4):987–1001, 1992.
- Z.-B. Zeng. Precision mapping of quantitative trait loci. *Genetics*, 136(4):1457–1468, 1994.
- W. Zheng, H. Zhao, E. Mancera, L. M. Steinmetz, and M. Snyder. Genetic analysis of variation in transcription factor binding in yeast. *Nature*, 464(7292):1187–1191, 2010.

Acknowledgements

First of all, I would like to thank Johannes Berg for the outstanding supervision and the fruitful discussions over the years. He helped me advancing my project continuously. A special thanks goes to my collaborators Andreas Beyer, Mathieu Clément-Ziza, and Jan Großbach from the CECAD who provided the yeast eQTL dataset and who had the patience to answer all my biological questions concerning the dataset. I also like to thank Michael Lässig for the helpful advice and discussions concerning the published paper. I would like to acknowledge my colleagues Nina, Nandita, Stephan, Mara, Simone, Daniel, Thorsten, Fernanda, and especially my officemates Chau, Pras, and Simon that contributed to a nice working atmosphere and an entertaining lunch time. Finally, I would like to express my gratitude to Andrea, who always supported me and who was always able to make me smile throughout my journey to the PhD.

Teilpublikationen

Riedel, N., Khatri, B. S., Lässig, M., & Berg, J. (2015). Multiple-line inference of selection on quantitative traits. *Genetics*, 201(1), 305-322.

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit einschließlich Tabellen, Karten und Abbildungen, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie abgesehen von unten angegebenen Teilpublikationen noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Johannes Berg betreut worden.

Nico Riedel