Discovering genetic interactions based on natural genetic variation

Dissertation

by

Marit Ackermann

Submitted to

Fakultät Statistik,

Technische Universität Dortmund

in Fulfillment of

the Requirements for the Degree of

Doktorin der Naturwissenschaften

Dresden, July 2012

Referees:
Prof. Dr. Katja Ickstadt
Prof. Dr. Jörg Rahnenführer
Supervisor and external referee:
Dr. Andreas Beyer

Date of Oral Examination: September 24, 2012

Acknowledgements

This work would not have been possible without the help and support of many people. To all of them, I am much obliged.

I am deeply grateful to my supervisor Andreas Beyer who gave me the opportunity to conduct research in his group, who steadily and cautiously guided me throughout my PhD and whose door was always open for discussions, which often went beyond the mere scientific problems. I also thank him for enabling me to see so many beautiful places all over the world, where he sent me to present our work.

But even more I wish to thank him for bringing together such a diverse group of very skilled international researchers with different scientific and cultural backgrounds. This made my daily work an adventure on its own, inspired us to deep discussions and many cheerful moments. In particular, I would like to thank Mathieu for many insightful conversations, especially on genetics problems, for his career advice as a very skilled and experienced researcher, but first and foremost for bringing French flair and lifestyle to Dresden and for being the best Parisian friend I could ever imagine. Special thanks goes to Weronika for our fruitful collaboration on a "two weeks" project, which I am really looking forward to continue. I thank Jake for all our discussions on statistical topics, for his inspiring work and thoughts about life and science.

And of course I would like to thank Claudia and Thomas for solving all our problems with administrative issues and overheated servers in almost no time; Ralf for always finding a free CPU, for not even being angry about an emergency call on a Saturday morning and for providing the best pumpkins in the world; and my dear Mandy for becoming so much more for me than just a perfect administrative assistant.

My special thanks goes to Professor Katja Ickstadt and Professor Holger Schwender for their advice on the methodological orientation of this thesis, their support on all statistical issues as well as for many helpful discussions. And although I regrettably never attended any of his lectures, I owe a lot to Professor Götz Trenkler, who convinced me to study Statistics in Dortmund by writing me a personal letter.

My work heavily relies on very good experimental data. Therefore, I would like to thank Richard Mott and his group at the Wellcome Trust Centre for Human Genetics for their help on the HS mouse data and for inviting me to Oxford. I am also thankful to Gerald de Haan and his group at the University of Groningen for providing the BXD gene expression data and for many fruitful conversations about the genetic regulation of hematopoiesis.

But first and foremost, I owe a debt of gratitude to my family for their love, faith and constant support. Ich danke meiner Familie über alles dafür, dass sie immer für mich da ist.

Abstract

Complex traits can be attributed to the effect of two or more genes and their interaction with each other as well as the environment. Unraveling the genetic cause of these traits, especially with regard to disease etiology, is a major goal of current research in statistical genetics. Much effort has been invested in the development of methods detecting genetic loci that are linked to variation of disease traits or intermediate molecular phenotypes such as gene expression levels.

A very important aspect to be considered in the modeling of genotype-phenotype associations is that genes often interact with each other in a non-additive fashion, a phenomenon called epistasis. A special case of an epistatic interaction is an allele incompatibility, which is characterized by the inviability of all individuals carrying a certain combination of alleles at two distinct loci in the genome. The relevance and distribution of allele incompatibilities has not been investigated on a genome-wide scale in mammals.

In this thesis, I propose a method for inferring allele incompatibilities that is exclusively based on DNA sequence information. We make use of genome-wide SNP data of parent-child trios and inspect 3×3 contingency tables for detecting pairs of alleles from different genomic positions that are under-represented in the population. Our method detected substantially more imbalanced allele pairs than what we got in simulations assuming no interactions. We could validate a significant number of the interactions with external data and we found that interacting loci are enriched for genes involved in developmental processes.

Genes do not only interact with one another, their regulatory activity also depends on the environment or cellular context. The impact of genetic variation on gene expression will therefore also depend on cell types or on the cellular state. This aspect has long been neglected in the inference of genetic loci that are linked to gene expression variation (expression quantitative trait loci, eQTL). There is thus a need to develop methods for analyzing the variation of eQTL between different cell types and to assess the impact of genetic variation on expression dynamics rather than just static expression levels.

In the second part of this thesis, I show that defining and detecting eQTL regulating expression dynamics is non-trivial. I propose to distinguish "static", "conditional" and "dynamic" eQTL and suggest new strategies for mapping these eQTL classes. By using murine mRNA expression data from four stages of hematopoiesis, we demonstrate that eQTL from the above three classes yield associations with different modes of expression regulation. Intriguingly, dynamic and conditional eQTL complement one another although they are based on integration of the same expression data. We reveal substantial effects of individual genetic variation on cell state specific expression regulation.

Contents

List of Figures

List of Tables

List of Acronyms and Symbols

Acronyms

A	adenine
ANOVA	analysis of variance
BXD	panel of inbred strains derived from C57BL/6J (B) and DBA/2J (D)
C	cytosine
DNA	deoxyribonucleic acid
eQTL	expression quantitative trait locus
FDR	false discovery rate
G	guanine
GO	Gene Ontology
HS	heterogeneous stock
HSC	hematopoietic stem cell
HWE	Hardy-Weinberg equilibrium
ImAP	imbalanced allele pair frequencies
LD	linkage disequilibrium
MAF	minor allele frequency
mRNA	messenger ribonucleic acid
QTL	quantitative trait locus

RF	Random Forests
RIL	recombinant inbred line
SF	selection frequency
SNP	single nucleotide polymorphism
T	thymine
TF	transcription factor
U	uracil

Symbols

c	number of cell types
χ^2_{jk}	ImAP test statistic for markers j and k
$\chi^2_{\mathbf{obs}}$	ImAP test statistic for the observed data
$\chi^2_{\mathbf{perm}}$	ImAP test statistic for a given combination of pseudo-genotypes
$\mathbf{df}_{jk}$	degrees of freedom of χ^2_{jk}
G_{jk}	frequency of a genotype combination on markers j and k
g_l	genotype; takes values AA, Aa, aa for l, $l = 1, 2, 3$.
N	number of genes
n	number of samples
$\mathcal{O}_j$	set of trios for which there is genotype information on marker j
P	number of permutations
R^2	Pearson's correlation coefficient
w_{ijk}^{probe}	quality weight of probe j in sample i of cell type k
$\mathbf{X}$	predictor matrix for eQTL mapping
X_{ij}	genotype indicator of individual i on marker j
$\mathbf{x}_m$	genotype vector for marker m across all cell types

$\mathbf{y}_{k_1k_2}^{\text{diff}}$	vector of mRNA expression differences between cell types k_1 and k_2
y_{ijk}	mRNA expression level of gene j in sample i of cell type k
$\mathbf{y}_j$	mRNA expression vector of gene j across all cell types
$\bar{y}_{ij}$	average mRNA expression of gene j in sample i
y_{ijk}^{probe}	mRNA expression level of probe j in sample i of cell type k
$\bar{y}_{ij}^{\text{probe}}$	weighted average probe level of gene j in sample i
$\mathbf{z}$	cell type factor variable with as many levels as there are types

Chapter 1

Introduction

An organism's traits are determined in large part by the interplay between its genetic makeup and the physical and social environment that it inhabits. Very often, the genetic component of a trait is itself the result of an interplay of many small effect changes of the genetic material. In that case it is called a complex trait. This comprises the organism's physical appearance and fitness as well as its interaction with the community and the diseases it is affected by. Heritability, i.e. the genetic contribution to the phenotypic variance, has been estimated to be extremely large for some complex traits, e.g. more than 80% for schizophrenia and more than 90% for autism (Maher, 2008). Hence, there is a need to identify the genetic determinants of complex traits in order to fully understand their etiology and, for the case of diseases, for developing suitable means to cure them.

Gene association studies have revealed that very often the collective effect of these genetic factors is more than just the sum of their single effects. Broken down to the level of two genes, this phenomenon is called epistasis: the simultaneous perturbation of two genes leads to a phenotype that is not expected based on the phenotypes of the individual genes.

Classical genetics used to investigate gene-gene interactions by artificially introducing perturbations at two defined loci in the genome and then observed their effect on a phenotype of interest. This procedure is very laborious and thus restricts the analysis to a small number of candidate loci. The genomic era with its high-throughput measurements of molecular entities has paved the way to systematically study genetic crosstalk and its influence on a variety of phenotypes on a large scale. Microarrays and, more recently, genome-wide sequencing technologies allow to capture the natural genetic variation within a defined population on a genome-wide scale and thus enable us to directly infer the influence of multiple genetic alterations on complex traits without the need for any experimental intervention. Model organisms such as yeast or mouse have proven to be particularly useful here, as they can relatively easily be kept in a controlled environment at low costs, allowing to attribute any phenotypic differences to variation in the genetic blueprint. The laboratory mouse has attracted special attention in human hereditary disease research, because of its close similarities to human physiology and disease etiology.

Together with the emergence of new experimental methodologies, a range of statistical methods for analyzing the relationship between the genetic makeup and a phenotype has been developed. Presumably the most prominent among them are linkage and association tests. They are based on the concept that the genomic position of the mutation causing the disease, called the disease locus, is physically close to a genetic locus whose variation in the population we were able to measure (the marker locus). If this is the case, then both loci will be inherited together, they are said to be linked. A sample of affected individuals will carry the same genetic information on the disease locus and therefore also on the marker locus. Hence, a position in the genome that is identical among individuals suffering from the disease is likely to be close to the disease causing mutation.

Under the assumption of a given genetic model, linkage analysis tests for the cosegregation of a certain manifestation of the genetic information at the marker locus together with the disease phenotype. Statistical inference is drawn from a likelihood ratio model comparing the likelihood of the data under the assumption of linkage with the likelihood under independence (Thomas, 2004, LOD score). Model-free linkage approaches are mostly based on inferring the over-representation of genetic material at a given locus among affected sib pairs (Elston, 1998).

Association studies directly infer the statistical association of the disease phenotype with a candidate gene region either in a population of unrelated individuals or a set of families. Hence, these studies test whether certain combinations of a trait characteristic and the genetic information at a given locus in the genome occur together more often than expected by chance. This directly leads to the application of χ^2 tests or, for family based designs, the transmission disequilibrium test (TDT), a version of the classical McNemar's test (Elston, 1998; Spielman et al., 1993). Genome-wide association studies (GWAS) mainly fit a battery of linear or logistic regression models, where each candidate causal locus is separately used as a predictor for disease state or phenotype.

Despite their popularity, these methods have long been limited in the sense that they infer the effect of each gene on a phenotype separately, without taking into account epistatic effects, although it is known from mutational studies that these are ubiquitous (Phillips, 2008). One of the main reasons for that was the explosion in the number of tests that have to be conducted in order to infer all pairwise interactions on a genome-wide scale and the lack of large-sized studies for reaching sufficient power.

In recent years, several approaches try to fill this gap. Linear and logistic regression models can be easily extended to contain interaction effects (Cordell, 2002). In order to model the effect of many predictors simultaneously, penalized likelihood approaches as well as Bayesian hierarchical models have come into play (Yi, 2011). In addition, several machine learning approaches such as multifactor dimensionality reduction (Ritchie et al., 2001), Random Forests

(Kim et al., 2009; Schwarz et al., 2010) or logic regression (Schwender and Ickstadt, 2008; Schwender, 2011; Ickstadt et al., 2008) have been applied to detect or at least take into account epistasis in genome-wide association studies. Since under certain conditions the interaction term in a logistic regression model corresponds to the correlation of the two genetic loci, epistasis can also be inferred by directly calculating an association between the loci within a set of individuals exhibiting the same phenotype (e.g. all cases). This results in a χ^2 test for independence (Cordell, 2009). Similar approaches have been proposed for family-based studies, which require the incorporation of kinship relations into the analysis as these will otherwise confound the results (Devlin et al., 2001; Rabinowitz and Laird, 2000). Beyond that, family structure can be regarded as an additional piece of information, which can be exploited to increase the power of a method to infer epistatic interactions (Schaid, 1999).

A very extreme form of epistasis is the incompatibility between two particular genetic variants at two distinct loci. Such incompatibility leads to the inviability of the organism. Therefore, only individuals not exhibiting the variant combination can be observed. Hence, classical case-control designs will not be able to detect them. These so-called Dobzhansky-Müller incompatibilities (Orr, 1996) have been investigated on a small scale predominantly by the plants genetics community, but a systematic search on the genome-level in mammalian species is lacking. However, such an analysis would provide insights into the extent and structure of epistatis throughout the genome as well as its relevance for the general health of the individual. The availability of more and more fully genotyped mouse or even human families might enable us to fill this gap. With the lessons learned from inferring epistatic interactions in family-based studies using statistical genetics approaches, we aimed at developing a method that detects such incompatibility of two genes, based solely on genetic data of populations with known kinship relations. Chapter 3 will present an extension and modification of a χ^2-like test to discover pairs of loci across the genome that together cause severe phenotypes while not being harmful on their own.

Alterations of the genetic makeup do not directly act on complex phenotypes like disease symptoms, their effect is rather mediated by changes of molecular processes taking place on the cell level. Therefore, these molecular traits, e.g. gene expression levels, protein or metabolite abundances, are often studied as intermediate phenotypes to gain insights into the genetic regulation of higher level traits. Moreover, understanding the development of traits on the molecular level is prerequisite for deriving strategies for disease cure, since this is where drugs can most directly intervene.

A genetic locus that acts upon one particular molecular trait, namely the expression level of a gene, is called an expression quantitative trait locus (eQTL). This regulatory relationship can be regarded as a special kind of genetic interaction, where the information

at one genetic locus influences the amount of usable information (i.e. information that can be translated into a functional protein) of another gene. From a methodological point of view, the expression levels of this gene can thus be modeled as a function of the genetic information at the eQTL locus. Thereby, eQTL studies exploit genome-wide measurements of natural genetic variation together with high-throughput gene expression measurements for virtually every gene in the genome.

Methods for eQTL mapping can be divided into univariate and multivariate (or single-locus versus multi-locus) mapping methods. Traditional genetics approaches, such as Haley-Knott regression (Haley and Knott, 1992) or composite interval mapping (Zeng, 1994), test each genetic locus one by one for its association with the expression level of a given gene ignoring the effect of other loci. These strategies have therefore only limited potential to detect factors with joint effects on gene expression, not to mention epistatic interactions. However, genes or their products do not act independently from the rest of the genome. Instead, they are embedded in regulatory pathways and complexes implying that the expression level of a gene is influenced by a set of genetic factors (its network neighborhood). Therefore, multi-locus methods regard eQTL mapping as a feature selection problem: the expression of a gene is predicted (explained) using a set of genetic marker loci (Broman and Speed, 2002). Each genetic locus is scored with respect to how informative it is for the prediction task while taking into account the effect of multiple genetic loci. These methods often rely on penalized regression algorithms such as LARS (Efron et al., 2004), partial least squares regression (Chun and Keleş, 2009) or machine learning techniques, e.g. Random Forests (Breiman, 2001), which is based on an ensemble of regression trees.

Our group has previously shown through investigation of both simulated and experimental data that multi-marker mapping methods clearly outperform single-marker methods (Michaelson et al., 2010). Moreover, we participated in the "DREAM5 SYSGEN A - In-silico network challenge" (`http://wiki.c2b2.columbia.edu/dream/index.php/D5c3`, Prill et al., 2010), which was set up to provide synthetic gene expression and genotype data that mimic the structure of real gene-regulatory networks, facilitating the assessment of the accuracy and sensitivity of eQTL mapping approaches that are currently used by the community. In the course of our participation, we showed that the combination of several multivariate eQTL mapping methods into committees outperforms the individual methods (Ackermann et al., 2012). We tested different committees of the following methods: Random Forests, the Lasso (Tibshirani, 1996) and the Elastic Net (Zou and Hastie, 2005). We also showed that our proposed approaches lead to a much higher average precision than the other DREAM challenge contributions, at the cost of slightly lower average sensitivity.

While the impact of genetic interactions on gene expression levels is already being acknowledged by the community, less attention is paid to the interaction between genetic causes of

gene expression variation and a cell's state or environment. It has been shown that eQTL are to a large extent tissue-specific (Dimas et al., 2009; Fu et al., 2012), i.e. different sets of genetic loci influence the expression of a given gene in different contexts. Two consequences arise from this: (i) not every gene-regulatory relationship can be deduced from a data sample that was collected in a defined tissue or cell state; (ii) eQTL studies conducted in one cell type cannot directly be transferred to another cell type, e.g. to explain the etiology of a disease related to this cell type (Powell et al., 2011). These issues often are not taken into consideration when conducting eQTL studies. Most of these studies still infer eQTL in only a single cell type or cell line.

Furthermore, under certain circumstances cells can undergo severe morphological or functional changes, which will be induced by extensive changes in their gene expression landscape. This will, for example, be the case during the differentiation of a stem cell, the response of a cell to a changed environment or after treatment with a drug. It can be assumed that the accompanying changes of gene expression levels are under tight genetic control. However, there are only very few studies that directly infer eQTL associated with gene expression changes (Smith and Kruglyak, 2008), and maybe no such study has been conducted in mammals. Beyond that, a comprehensive consideration of methodological issues arising when inferring eQTL across tissues or during dynamic changes of the cell is lacking.

In Chapter 4 of this thesis, I propose to consider gene-regulatory relationships as dynamic processes and classify eQTL into three categories: those that are related to a gene's expression levels regardless of the cell state, those that are specific for particular cellular states or environmental conditions and finally eQTL that impact on the dynamics of expression changes. New statistical methods will be presented and compared that allow to detect these different kinds of eQTL.

This thesis is structured in the following way: In the next chapter I introduce the biological concepts being essential for this work. The main Chapters 3 and 4 then adopt two very different perspectives on deciphering molecular mechanisms underlying complex traits. Chapter 3 describes a statistical test allowing to detect allele incompatibilities on a genome-wide scale. The test is based on genetic data of a set of individuals with known family relations. In Chapter 4 I present a systematic classification of genetic variation impacting gene expression during dynamic processes. Different classes of eQTL are contrasted and strategies for detecting them are proposed and compared with each other. The applicability and biological relevance of both methodological developments presented in Chapters 3 and 4 is demonstrated on publicly available mouse data sets. Finally, all results are summarized and discussed in Chapter 5 and an outlook of how to extend and combine the different approaches of inferring the influence of natural genetic variation on complex traits is presented.

Chapter 2

Biological Background

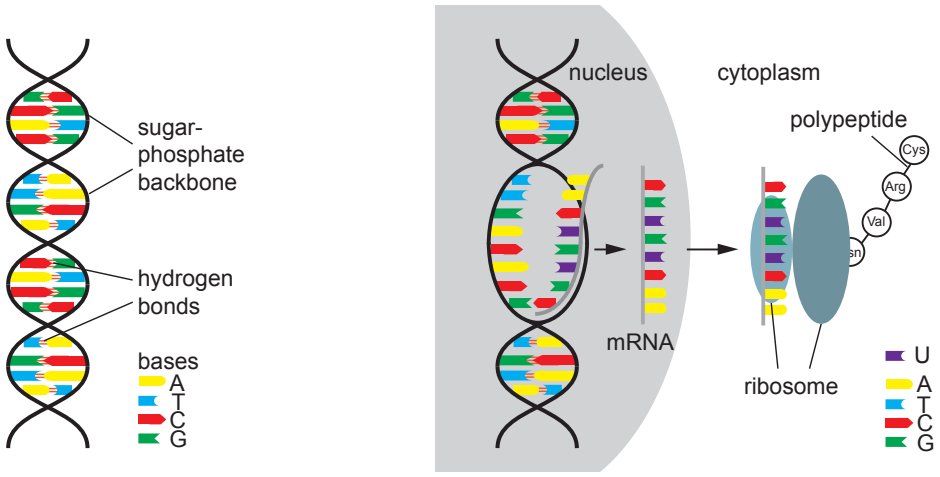
This chapter briefly introduces basic genetic concepts and terms that are used throughout this thesis. If not stated otherwise, all explanations apply to mammalian species, in particular mouse and human, which are very similar in terms of molecular genetics. A more comprehensive introduction to the concepts of molecular cell biology that are explained in this chapter can be found in Alberts (2004) and Alberts (2002). Additional references are given whenever necessary.

2.1 From DNA to protein and beyond

The *genome* of an eukaryotic organism is the ensemble of the heritable information, which is necessary to control its structure, development and virtually all its activities. It is stored in the nucleus of almost every cell it is comprised of and is made up of *deoxyribonucleic acid (DNA)*. If the organism possesses mitochondria, they contain their own genome. DNA consists of two long chains (*strands*) of nucleotides, which are connected by hydrogen bounds to form a twisted double-helical structure (Figure 2.1A). Each nucleotide consists of a deoxyribose sugar, a phosphate group and one of four different nitrogen-containing bases - *adenine (A)*, *cytosine (C)*, *guanine (G)* or *thymine (T)*. The alternating sugar-phosphate complexes build the backbone of the two strands of DNA, while the bases point to the inside of the double-helix. Thereby, each two bases form a pair, in which A always pairs with T and G with C. This defined pairing results in a complementarity of the two strands, which is important for the transcription and replication of the DNA.

DNA is organized into chromosomes, which correspond to DNA molecules. In each cell there might be one or several copies of each chromosome. Most cells of nearly all mammals are *diploid*, i.e. they carry two sets of chromosomes in their nuclei, one inherited from the mother, the other one from the father. Some parts of the DNA form functional units called *genes*, which encode the information needed to build proteins. Proteins are the major components of each cell and responsible for almost all of its activities.

The transformation of the genetic information into a protein is a two-step procedure often called the *Central Dogma of Molecular Biology* (Figure 2.1B). In the first step, the DNA



A DNA structure **B** From DNA to protein

Figure 2.1: From DNA to protein. **A** DNA is a double-helix of polymers made up of nucleotides. Sugars and phosphate groups build the backbone, while pairs of nitrogen-containing bases connected by hydrogen bonds face each other in the center of the helix. **B** One strand of the DNA is transcribed to mRNA, which is transported to the cytoplasm. In the ribosome, mRNA is translated into polypeptides, the building blocks of proteins.

sequence is copied by polymerases into *messenger ribonucleic acid (mRNA)* in a process called *transcription*. The chemical structure of mRNA is very similar to that of DNA, except that its nucleotides consist of ribose sugars instead of deoxyribose and thymine (T) is replaced by *uracil (U)*. Transcription takes place in the nucleus and is followed by some post-processing steps, which stabilize and rearrange the mRNA before proteins can be constructed from it. Subsequently, the mRNA is transported to the cytoplasm of the cell where it is translated into a protein in the ribosome. This process is called *translation*. Each triplet of mRNA bases codes for one of twenty amino acids (some triplets code for the same amino acid), the building blocks of peptides, which in turn are assembled to proteins. These often fold into 3-dimensional structures and undergo post-translational modifications transforming them into functional proteins.

Proteins are the major players in the establishment of cellular structure and traits. Any observable trait resulting from the expression of one or many genes is called a *phenotype*. *Mutations*, i.e. modifications of the genetic information encoded in the DNA, can eventually lead to pathological phenotypes of the cell (such as cell death or abnormal growth) and consequently induce diseases.

2.2 Natural genetic variation

Although by far the largest part of the genome is identical between individuals of the same species, there exist genetic variations, such as deletions, insertions or substitutions of single bases or larger parts of DNA, underlying the phenotypical diversity among individuals. At each genetic *locus*, i.e. a specific position in the genome, each of the variants of the DNA sequence occurring in the population represents an *allele*. The allele that is observed less frequently in the population is called the *minor allele* as opposed to the *major allele*. The frequency of the minor allele is denoted by *minor allele frequency (MAF)*. A combination of alleles at neighboring loci that is transmitted together to the next generation is called a *haplotype*. If a variation in only one of the two copies of a locus is sufficient to affect a given phenotype, its effect is called *dominant*. If on the contrary, both copies need to carry a variation in order to change the phenotype, it is referred to as a *recessive* allele. The combination of the two alleles at one locus is called the *genotype*.

The length of a natural genetic variation may vary from one base up to the length of one or several genes. If a variation of one base is present in at least 1% of the population, it is called a *single nucleotide polymorphism (SNP)*. In most cases, there are two possible allele variants at a SNP locus. This implies three possible genotypes: *homozygous major* (both copies of a chromosome carry the more frequent SNP variant), *heterozygous* (the copies carry different SNP variants), *homozygous minor* (both copies of a chromosome carry the less frequent SNP variant). If a SNP, a point variation in the genomic sequence, results in a change of the protein code, it is called a *non-synonymous* SNP. However, SNPs might also affect a phenotype if they are, for example, located in a non-coding regulatory region of gene expression.

2.3 Genetic inheritance

Most of the traits that describe an individual, including physical appearance, social skills or diseases, are influenced by the organism's genetic background. This genetic makeup is inherited from the parents as a unique mixture of their own genetic information, causing the variation among traits that makes each individual unique.

Each mammalian cell normally contains two copies of each chromosome. Since the two copies contain very similar information but are not 100% identical due to naturally occurring genetic variation (Section 2.2), they are called maternal and paternal *homologs*. To avoid that successive generations accumulate more and more copies of each chromosome, only one copy is passed on by each parent to its offspring. Thus, special cells, called gametes and containing only a single (*haploid*) set of chromosomes, have to be created from diploid

somatic cells in a process called *meiosis* before the actual reproduction can take place. During meiosis, homologous chromosomes are replicated, paired with their corresponding maternal or paternal homolog and eventually separated in two sequential cell divisions (Figure 2.2). This results in four haploid cells, each containing a random assortment of one maternal or paternal homolog of each chromosome. If two individuals, a female and a male, reproduce, their gametes will fuse and create a diploid cell from which a new organism will develop.

The most important step during meiosis is the pairing of homologous chromosomes. While the pairs of homologs are physically close to each other, fragments of homologous chromosomes can be exchanged, an event referred to as *recombination* or *crossover* (Figure 2.2). Recombination is a major source of genetic variation among individuals, since it rearranges maternally and paternally derived genetic information. Together with the random assignment of maternal and paternal homologs to gametes, recombination allows the reshuffling of alleles, which in turn provides the basis for genetic diversity and evolution.

The frequency of crossover, called the *recombination rate*, varies across the genome. Moreover, some allele combinations, maybe even across chromosomes, might occur less frequently than others due to selection pressure, non-random mating and other causes. Hence, there are alleles which are inherited together more (or less) often than expected from their marginal allele frequencies. Two alleles for which such a non-random association is observed, are said to be in *linkage disequilibrium (LD)* (Balding et al., 2007, p. 909ff.). If there are no such events as selection or genetic drift, i.e. under the assumption of an ideal, infinitely large population, allele and genotype frequencies should remain constant over generations, a state which is called *Hardy-Weinberg equilibrium (HWE)* (Balding et al., 2007, p. 1243).

Since there is only a limited number of recombinations per generation (on average two or three per meiosis for a human chromosome), a large part of the genetic information on a chromosome is inherited together. These chromosomal regions can be represented by genetic *markers*, single SNPs at a defined position in the genome that are representative for the ancestry of the locus.

2.4 Genetic interactions

Nowadays, the generic term *genetic interaction* is used in different contexts and with a wide range of interpretations. In its most stringent definition, it refers to the term *epistasis*, which was first established by William Bateson (Bateson, 1909). He described the phenomenon that the effect of an allele A at locus A is masked by the effect of an allele B at locus B. A prominent example of such a case (described here in a simplified way) is the hair color of mice which is determined by an "albino" locus A and a "color" locus C as shown in Table 2.1. Locus C determines whether the hair of the mouse is grey or black, the C allele

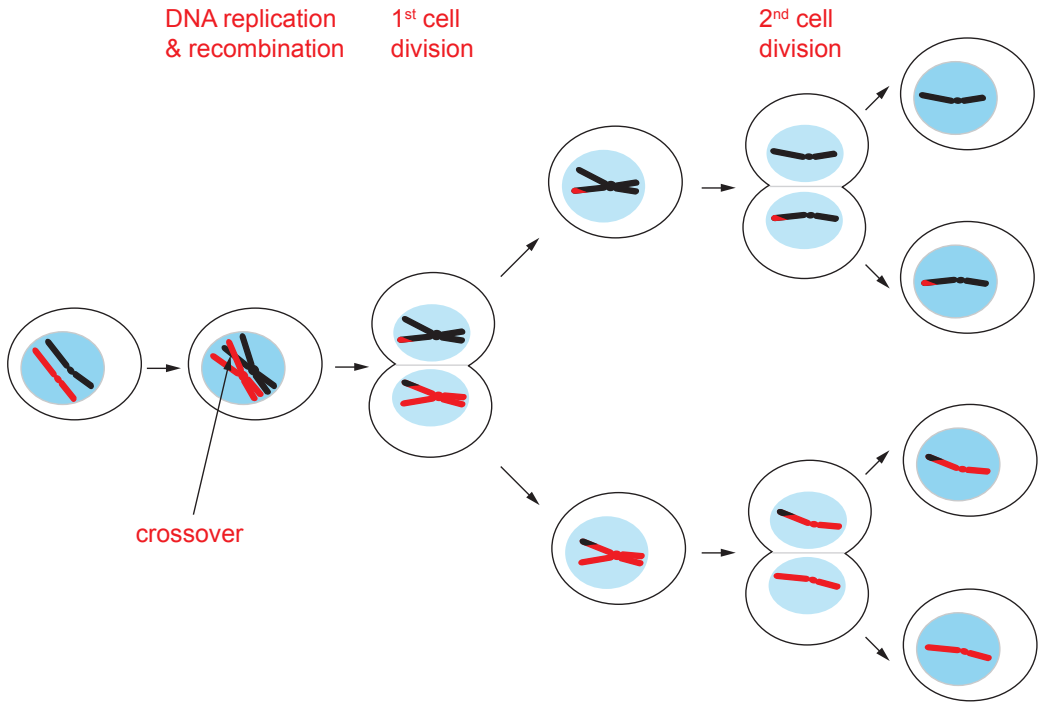


Figure 2.2: Meisosis. For the sake of simplicity only one pair of homologous chromosomes is shown. The DNA of each chromosome in the nucleus is replicated, whereby pairs of homologous chromosome copies are in physical proximity. This allows crossover between the maternal and paternal copy of the same chromosome and subsequently leads to an exchange of genetic material between the chromosomes (recombination). Afterwards, the paternal and maternal homolog are separated in a first cell division, and finally both copies are partitioned off in a second cell division. Hence, meiosis generates four haploid cells each containing only one copy of each chromosome (where maternally and paternally derived chromosomes are mixed randomly).

being dominant for grey. However, the color effect is masked by the recessive a allele at the "albino" locus, i.e. the hair is colorless if the mouse has the aa genotype at locus A, regardless of the allele at locus C.

Epistasis is always defined with respect to a certain trait, i.e. two loci with an epistatic effect on one phenotype might be independent with respect to another phenotype. Many genetic interaction studies have considered the fitness or the growth rate of an organism as the phenotype being effected by epistasis (Beltrao et al., 2010; Costanzo et al., 2010; Schuldiner et al., 2005; Tong et al., 2001). In the most extreme case of such a setting, a

Table 2.1: Epistatic control of hair color in mice. Hair color of the mouse depends on two loci A and C. Mice carrying at least one dominant *C* allele have grey hair while mice carrying two *c* alleles at locus C have black hair. The effect of the color locus is masked by the effect of the recessive *a* allele at locus A causing an albino phenotype with white hair color.

	Genotype at locus C		
Genotype at locus A	**CC**	**Cc**	**cc**
AA	grey	grey	black
Aa	grey	grey	black
aa	white	white	white

specific combination of alleles at the epistatic loci could be lethal for the organism. We call this scenario an *allele incompatibility*. Since individuals inheriting an incompatible allele combination are not viable, these combinations will not be observed in the population.

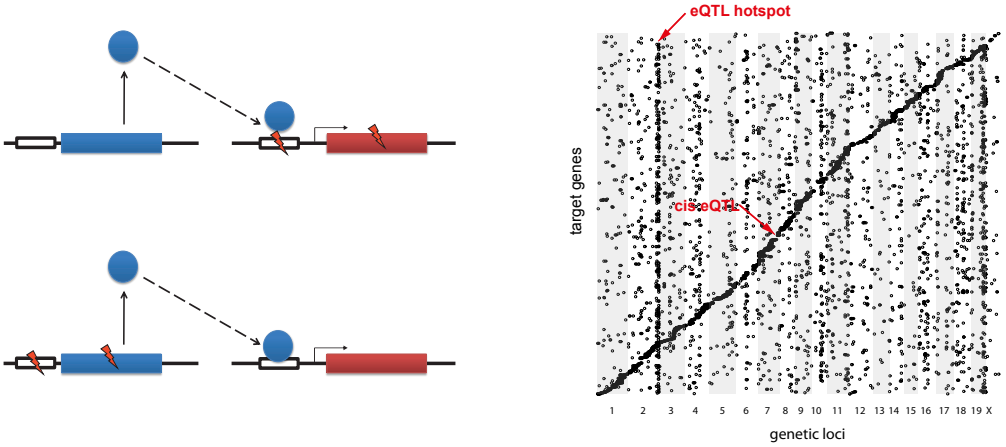
The above biologically motivated definition of epistasis deviates from the statistical definition of epistasis. It was brought up first by Fisher (1918). It describes a non-additive effect of two predictors (in this case the genotype at two genetic loci) on a quantitative phenotype in a linear model (Cordell, 2002). In this sense, epistasis is now used interchangeably with the term genetic interaction in statistical genetics. It is important to note that additivity is always defined with respect to a determined scale of the predictors, i.e. predictors that are interacting on one scale might become additive when being transformed to another scale.

The term genetic interaction is sometimes also used to describe biochemical interactions between gene products. However, we refer to these as *protein interactions*.

In this thesis, gene-regulatory relationships are also included into the broader understanding of genetic interactions. Gene-regulatory relationships describe the impact of the alleles at a genetic locus on the transcription level of a gene at either the same or a distant locus and are explained in detail in the following section.

2.5 Genetic control of gene expression

As protein synthesis from DNA is a multi-stage process, the amount of protein that is produced can be controlled in each of the steps described in Section 2.1. For example, there is regulation on the amount of transcribed mRNA, its post-processing and degradation, its transport to the cytoplasm, the amount of translated mRNA and on the activity of the protein itself. The first step of protein synthesis, the transcription, is under complex control of several factors, proteins interacting with the DNA or the histones (special proteins that are needed to package DNA into a stable structure). The initiation of transcription is mainly



A *cis* and *trans* eQTL **B** eQTL map

Figure 2.3: Cis and trans acting eQTL. A The transcription of a target gene (red) is regulated through the binding of a TF (blue circle) in its promoter region (small white rectangle). The TF gene is encoded at a distant locus (blue rectangle). The eQTL influencing the expression level of the transcribed gene can be located in the coding region of the gene or its promoter (upper panel), in which case it is called a *cis* eQTL. *trans* eQTL (lower panel) are located in the locus of the TF gene, which might be very distant from the target gene locus. **B** The regulatory relationships between eQTL and their targets can be visualized in an eQTL map. Each point represents an eQTL - target gene pair, where markers are displayed on the x-axis and genes on the y-axis. Points on the diagonal show *cis* regulation, while off-diagonal points belong to *trans* eQTL. Vertical bands show eQTL loci that control a large number of target genes. They are called hotspots.

controlled by *transcription factors (TFs)*, proteins that bind individually or in complexes to the regulatory region of a gene, which is called the target gene. There, they either promote or repress the recruitment of RNA polymerase to the DNA.

Since transcriptional regulation is key to the control of protein levels in cells and thus the manifestation of phenotypes, it is crucial to understand its molecular mechanisms. The combined analysis of natural genetic variation and gene expression data, known as *genetical genomics* (Jansen, 2001), has proven its value for the elucidation of how genotype affects gene expression levels. Natural genetic variations can impact the regulation of transcription, for example by changing the sequence of the promoter region of the gene whose expression is regulated, or by modifying the 3-dimensional structure of the TF, thereby modifying its binding affinity and in turn its regulatory activity. Of course, genetic variation might also affect the coding region of the regulated gene itself and thus hamper the translation into the protein or change the protein sequence. Because it causes a quantitative variation of a trait

(the gene expression), a genetic locus that contains such a variation is called an *expression quantitative trait locus (eQTL)*. (Any variation that causes changes in a quantitative trait different from gene expression is called a *quantitative trait locus (QTL)*.) An eQTL can either act in *cis* or *trans* (Figure 2.3A), depending on whether the variation occurs in one locus with the gene whose expression is regulated or further away at a distant locus. The expression levels of a gene can be controlled by more than one eQTL just as an eQTL can influence the expression of a number of target genes. A locus that regulates a large number of targets is called an *eQTL hotspot*. The relationship between genetic loci and target genes can be visualized in an eQTL map, an example of which is shown in Figure 2.3B.

2.6 Hematopoiesis

In mammals, the blood is composed of a large variety of cells with very different functions in immune system and oxygen supply. Blood cells are roughly classified into white and red cells as well as platelets. Red blood cells, also called *erythrocytes*, are the most abundant cell type in the blood and transport oxygen to and carbonic dioxide from every cell of the organism. White blood cells (*leukocytes*) can be divided into three main groups: *granulocytes*, *monocytes* and *lymphocytes*. Granulocytes are further distinguished into *neutrophils*, *basophils* and *eosinophils* (Figure 2.4). All these different kinds of white blood cells play different roles in the immune response of the organism to different kinds of stimuli and pathogens.

Since blood cells have to fulfill a wide range of tasks and are thus needed in different quantities under specific conditions, there is a need for tight control of blood cell frequencies. Moreover, mature blood cells cannot divide and have a limited life time. Therefore, the stock of blood cells has to be continuously replenished throughout the life of the organism. This is achieved by the complex process of *hematopoiesis*, the development of blood cells from one common cell type, the *hematopoietic stem cell (HSC)*. HSCs reside in the bone marrow and have two unique properties: the ability of self-renewal and the potential to differentiate into any kind of blood cell. This differentiation takes place in several stages of progressive lineage restriction, which are depicted in Figure 2.4. First, the HSC is committed to either the *myeloid* or *lymphoid* cell line. Myeloid progenitor cells can still differentiate into erythrocytes as well as all white blood cells except lymphocytes. Lymphoid precursors give rise to the different kinds of lymphocytes, e.g. B and T cells. These differentiation processes are tightly genetically controlled by the expression of specific (sets of) TFs that activate and/or repress specific lineage commitments. In adult mammals, differentiation takes place in the bone marrow and also depends on signals transmitted through direct contact of the HSC with the bone marrow. Mature blood cells are released into the blood stream, which transports them to the tissue where they are required.

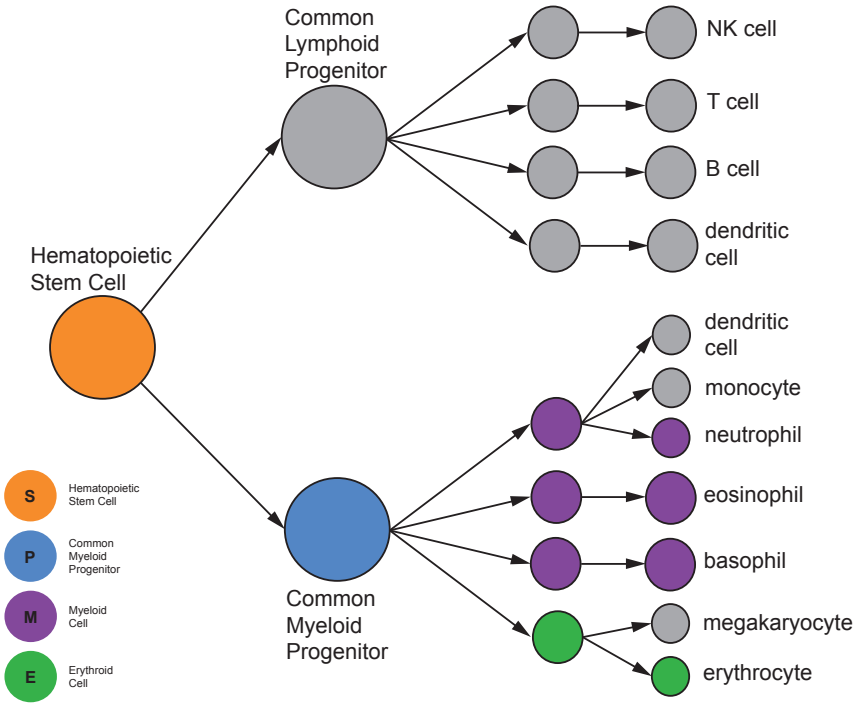
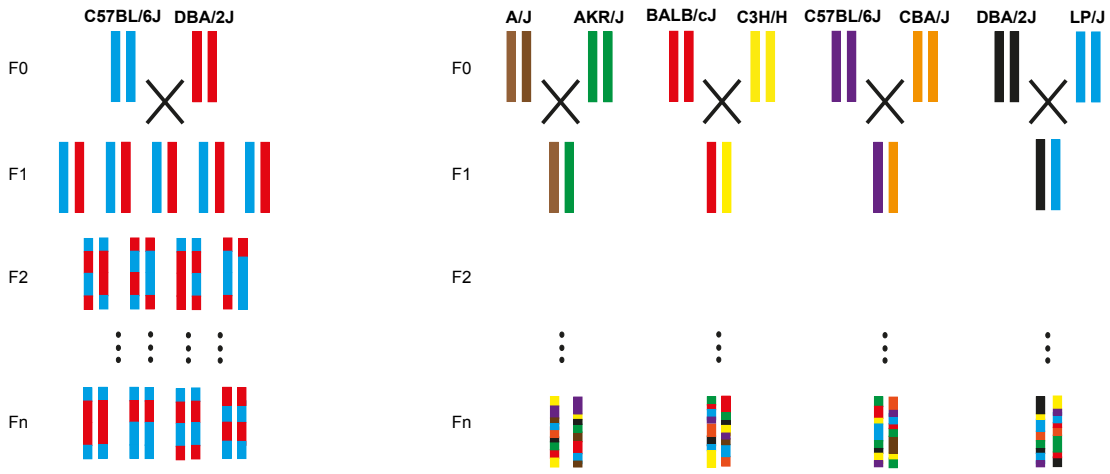


Figure 2.4: Schematic overview of hematopoiesis. A multipotent HSC differentiates into either a lymphoid or a myeloid progenitor cell. Progenitors then undergo several phases of lineage restriction until they are terminally differentiated into mature blood cells. The colors indicate hematopoietic cell types that are considered in Chapter 4 of this thesis.

2.7 Mouse resources for genetic studies

The laboratory mouse provides a powerful resource to study the impact of genetic variation on qualitative as well as quantitative traits. Each mouse strain carries a different mosaic of alleles, each of which represents a perturbation of the genetic system with a putative influence on a given trait. Through the events of recombination and segregation (Section 2.3), the allele distributions are randomized among the offspring of each mouse cross. Consequently, a pool of mouse lines can be used to track genetic loci containing variants that affect a phenotype (Rockman, 2008). For this reason, many different types of mouse genetic resources have been established since the 1950s. Most of them apply different breeding schemes on two or a panel of inbred strains, resulting in mouse stocks with different genetic architectures, spatial distribution of variation and allele frequencies (Roberts et al., 2007).

Recombinant inbred lines (RILs) are derived from inbreeding two parental strains over many generations such that the final progeny, all individuals belonging to the same line, carry an identical mix of the parental genomes on both of their chromosomes, i.e. they



A BXD strains **B** HS stock

Figure 2.5: Mouse breeding schemes. A The BXD inbred lines are derived from crossing two parental strains (F0 generation), C57BL/6J (B) and DBA/2J (D). The F1 offspring carry one chromosome from each parental strain. Due to recombination events during meiosis, each of the chromosomes of the progeny of each pair of F1 animals, the F2 generation, is a different mix of the B and D chromosomes. After many generations of inbreeding pairs of F2 mice, a panel of mice is obtained, each carrying two identical (homozygous) chromosomes, which are a mosaic of the original parental B and D chromosomes. **B** The HS outbred mice descend from eight inbred strains which are crossed amongst each other and subsequently mated randomly for many generations. The chromosomes of members of the final mouse stock are heterozygous assortments of the founder chromosomes.

have a homozygous set of chromosomes (Figure 2.5A). A popular and widely used panel of inbred lines are the BXD mouse lines, for which there exists dense genotype information on 88 extant and extinct lines (`www.genenetwork.org`). This panel was derived from the two parental strains C57BL/6J (B) and DBA/2J (D), which are known to differ a lot in their genetic variation and many clinically relevant phenotypes, among them many hematopoietic traits.

However, RILs often exhibit two major problems: (i) The cross of only two parents restricts the genetic variety among the lines. This results in a limited resolution of the loci that are found to be linked to a phenotype. Therefore, such loci often contain more than one - possibly a large number of - gene(s), so that no conclusion about the causal polymorphism of a given trait can be drawn. (ii) Many inbred lines quickly suffer from infertility or severe fitness defects that might eventually lead to the extinction of the line. This further restricts the palette of genetic variation that will eventually be used to detect QTL.

An alternative mouse resource that was created in order to circumvent the above mentioned

problems is the *heterogeneous stock (HS)*. Since it is derived from multiple initial parental strains, it contains a higher level of genetic diversity and thus provides a better resolution for (e)QTL mapping. Yet, these and other more diverse mouse resources like the collaborative cross still suffer from the problem of loosing inviable lines during the first generations of inbreeding (Chesler et al., 2008). In Chapter 3 of this thesis, we use data obtained from an HS derived from eight founder strains: A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6J, CBA/J, DBA/2J and LP/J (Hitzemann et al., 1994; Shifman et al., 2006). The breeding protocol differs from RILs in the way that progeny are mated randomly after an initial eight-way cross phase. Hence, the final population of outbred mice carry heterozygous chromosomes, each of which is a fine-grained composition of the founder chromosomes (Figure 2.5B). While the HS allows a more precise mapping of QTL, the higher variation and heterozygosity also pose challenges in the statistical analysis of the data. In particular, when genetic loci are used as predictors for modeling a quantitative trait, they need to be coded as either one three-level factor or two two-level dummy variables (as opposed to one binary predictor for RILs). Moreover, the higher resolution of the stock increases the sheer number of predictors, and thus exacerbates the "small n, large p" problems faced in many QTL studies. Finally, while the genotype of each RIL can be reproduced (and thus tested) as often as needed, the genotype of each mouse in the HS is unique and cannot be replicated.

Chapter 3

Systematic detection of epistatic interactions based on allele pair frequencies

The work presented in this chapter led to the following publication:

Ackermann, **Marit** and A. Beyer (2012). Systematic detection of epistatic interactions based on allele pair frequencies. *PLoS Genetics 8* (2), e1002463.

3.1 Introduction to allele incompatibilities

The simultaneous perturbation of two epistatically interacting genes leads to a phenotype that is not expected based on the phenotypes of the individual genes. Understanding these phenomena is indispensable for explaining multi-factorial traits and diseases (Cordell, 2002). In addition, epistatic interactions provide important insights into the functional organization of molecular pathways (Kelley and Ideker, 2005; Beyer et al., 2007). Much effort has therefore been put into the development of methods to discover epistatic interactions, mostly in linkage and association studies (Cordell, 2002; Hoh and Ott, 2003; Marchini et al., 2005; Phillips, 2008; Cordell, 2009; An et al., 2009; Liu et al., 2011; Wang et al., 2010).

Epistasis is always defined with respect to a specific phenotype and describes a non-additive interaction effect of two genes on that phenotype (Section 2.4). Most gene interaction studies explicitly measure a phenotype such as growth rate or viability (Beltrao et al., 2010; Costanzo et al., 2010; Schuldiner et al., 2005; Tong et al., 2001). However, one can also study implicit phenotypes by searching for the over- or under-representation of certain allele pairs in a given population. Such allele pairs are examples of Dobzhansky-Müller incompatibilities: they establish a fitness bias in favor of individuals inheriting the over-represented allele combination (Orr, 1996). In their most extreme form such incompatibilities are embryonic lethal. Genes harboring these alleles are clearly in epistasis, as none of the alleles alone has a fitness effect. Only the presence of specific allele pairs in one individual exposes the phenotype. In this context, an implicit phenotype is a trait that is not explicitly measured in the sample but whose regulators can still be inferred from the genotype data.

Whereas several such incompatibilities are known in plants (see Bomblies and Weigel (2007) and references therein), only very few allele incompatibilities have been reported in mammals (Montagutelli et al., 1996; Payseur and Place, 2007). A small number of recent studies have explored this idea for the genome-level identification of epistatic interactions: if a large number of individuals is genotyped at a large number of genomic positions, it becomes possible to test all allele pairs for over- and under-representation in that population (Williams et al., 2001; Payseur and Place, 2007; Lawrence et al., 2009). For example, (Williams et al., 2001) provide a map of distant LD in mouse RILs giving some indication about the distribution of imbalanced allele pair frequencies in the genome. However, even though some methodological progress has been made (Payseur and Place, 2007), previous studies could hardly identify a significant number of interactions. The main obstacle is the humongous number of statistical hypotheses tested when comparing all markers in a genome against all markers. When correcting for multiple hypothesis testing one is usually left with very few or even no significant allele pairs.

Here, we propose to address this problem by exploiting the additional information gained from studying family trios. We show that by analyzing a sufficiently large number of individuals with known family structure it becomes possible to detect substantially more interactions than what is expected if all markers were independent. Our method, called "Imbalanced Allele Pair frequencies" (ImAP), relies on sequence data only, making it applicable to the many already available SNP studies without the need for additional phenotype measurements. ImAP is based on inspecting 3×3 contingency tables that track the frequencies of all possible two-locus allele combinations in heterozygous individuals (assuming a diploid genome). The test that we propose is similar to a χ^2 test in that it compares the observed frequencies in this table to expected frequencies assuming independence. However, our version corrects the expected frequencies for confounding factors such as family structure or allelic drift (Griffiths, 2000). ImAP is described in detail in Section 3.2.

In Section 3.3 we apply ImAP to genotype data from a population of $2,002$ heterozygous mice with known family structure and identify 168 LD block pairs with imbalanced alleles. Using simulations we can show that this number is significantly larger than expected under the null hypothesis even after correcting for multiple hypothesis testing. The significance of the top scoring interactions between the LD blocks could be independently confirmed using a large collection of RILs. The number of significant allele pair imbalances that we detected is surprisingly large and was not expected based on the published evidence. Section 3.4 discusses important outcomes and consequences of our analysis. The R implementation of ImAP can be found in Appendix D.

3.2 The ImAP procedure

An overview of ImAP is given in Figure 3.1. Panel A shows the core step of ImAP, a χ^2-type test comparing the observed frequency of the joint occurrence of a certain diallelic genotype in one locus (with alleles A and a) together with a certain genotype in a second locus (with alleles B and b) with the frequency expected based on the genotypes of the parents under the null hypothesis (i.e. assuming no epistasis). The two loci are required to be distant enough from each other in order not to get false positive results due to local linkage. This results in a score χ^2_{obs} quantifying the deviation of allele pair frequencies from their expected values that is already corrected for inherent population structure. Subsequently, the significance of the score is assessed with a permutation approach using pseudo-controls that are derived from the genotypes that parents could have transmitted to their offspring (Figure 3.1B).

We apply this framework in two steps: First, we only analyze genomic blocks with high local LD using representative markers. In a second step we drill down to individual marker pairs. To further verify our results, we established a simulation procedure that mimics the mating structure of the pedigree under the assumption of independence.

3.2.1 The ImAP test statistic

The calculation of the test statistic can be divided into five steps which are depicted in Panel A of Figure 3.1. Steps one to three apply on single markers, while the last two steps consider the interaction between two marker loci.

1. Let $\mathcal{O}_j$ be the set of all parent child trios for which we have complete genotype information. This set might differ between markers due to missing values. Hence, for each marker only those trios are taken into account for which there are no missing values in the genotypes of both the parents and the offspring.

2. For each child in $\mathcal{O}_j$, calculate the probability to inherit each genotype based on the genotypes of the parents. This calculation is based on Mendelian laws.

   Let $X_{ij}(g_l) \in \{0, 1\}$ be the genotype indicator of a diploid child $i \in \mathcal{O}_j$. $g_l,\ l = 1, 2, 3$, can take one of the three values (AA), (Aa), (aa), where A is the major allele and a the minor allele on marker j. $\hat{X}_{ij}(g_l)$ is the corresponding expected genotype probability.

   The expected genotype of individual i on marker j is derived from the genotypes of the parents under the assumption of equal chances of inheriting each of the two possible alleles from each of the parents. The resulting probabilities for all possible parental genotype combinations are shown in Table 3.1.

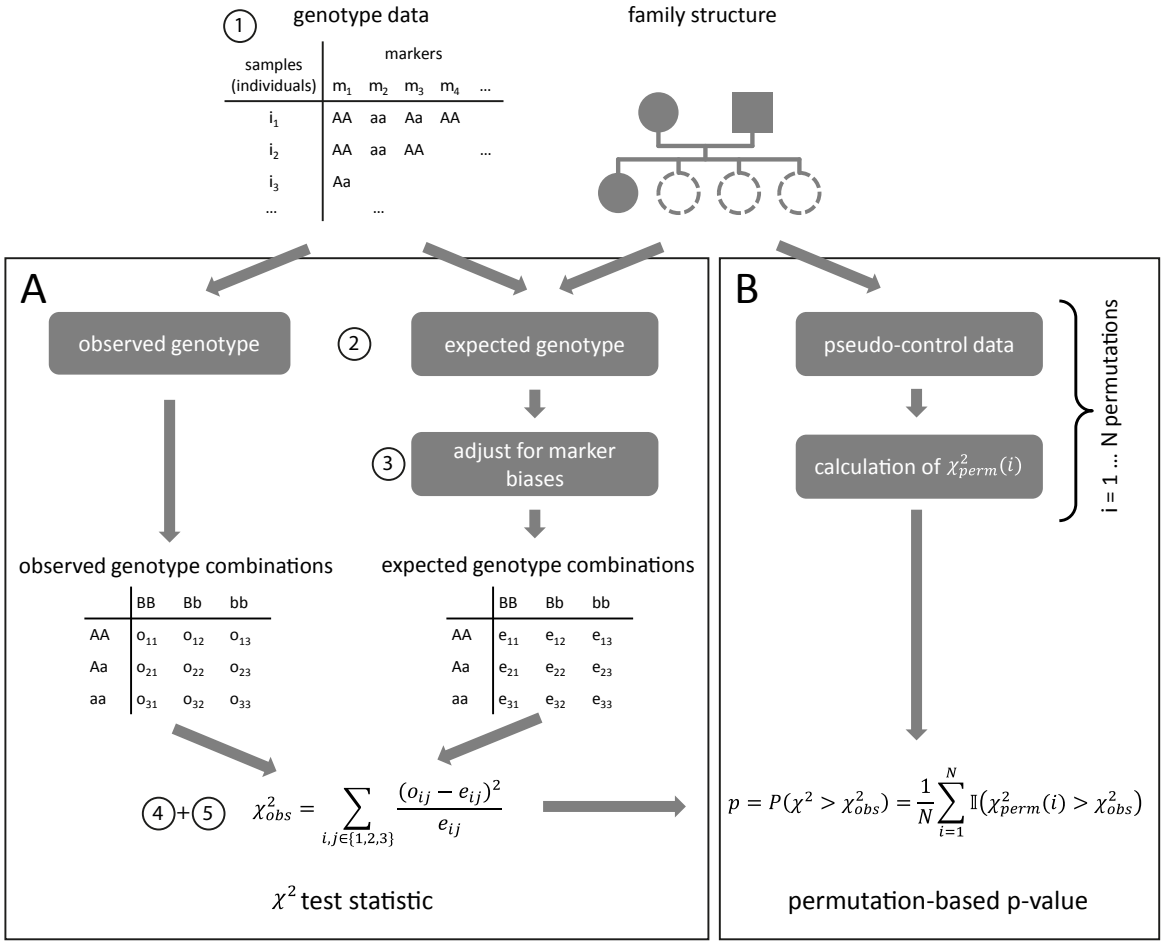


Figure 3.1: Schematic overview of the ImAP procedure. Panel A shows the calculation of the test statistic (numbers indicate the steps described in Section 3.2.1), panel B depicts the calculation of the p-values. Family information is used for both parts.

3. Correct the expected genotypes for possible confounding factors such as segregation distortion. There might be a preference in the inheritance of a certain genotype on one marker in the population which is independent of interaction effects, e.g. if this genotype leads to increased fitness. In order to correct the expected frequencies for allele selection that is independent of other loci we multiply each individual's expected genotype by the ratio of the sample-wide observed and expected frequencies for the corresponding marker (based on all samples):

Table 3.1: Expected genotypes. Probabilities of each of the possible genotypes in the offspring for each combination of parental alleles on a given marker.

		Offspring		
Parent 1	**Parent 2**	**AA**	**Aa**	**aa**
AA	**AA**	1	0	0
AA	**Aa**	0.5	0.5	0
AA	**aa**	0	1	0
Aa	**Aa**	0.25	0.5	0.25
Aa	**aa**	0	0.5	0.5
aa	**aa**	0	0	1

$$\hat{X}'_{ij}(g_l) = \hat{X}_{ij}(g_l) \cdot \frac{\sum\limits_{o \in \mathcal{O}_j} X_{oj}(g_m)}{\sum\limits_{o \in \mathcal{O}_j} \hat{X}_{oj}(g_l)}. \tag{3.1}$$

Normalize the corrected expectation so that the probabilities for each marker sum up to one:

$$\hat{X}^{\text{adj}}_{ij}(g_l) = \frac{\hat{X}'_{ij}(g_l)}{\sum\limits_{k \in \{1,2,3\}} \hat{X}'_{ij}(g_k)}. \tag{3.2}$$

This guarantees an adjustment of expected allele frequencies in cases where the observed frequency of a marker in the population deviates from the theoretically expected values.

4. Next, the observed and expected number of times each combination of genotypes appears on two distant markers can be inferred.

   Let $G_{jk}(g_{l_j}, g_{l_k})$ be the observed frequency of the genotype combination (g_{l_j}, g_{l_k}) on markers j and k, $\hat{G}_{jk}(g_{l_j}, g_{l_k})$ the corresponding expected frequency. They are obtained by summing over all individuals $i \in \mathcal{O}_{jk} = (\mathcal{O}_j \bigcap \mathcal{O}_k)$:

$$G_{jk}(g_{l_j}, g_{l_k}) = \sum_{i \in \mathcal{O}_{jk}} \left(X_{ij}(g_{l_j}) = 1 \wedge X_{ik}(g_{l_k}) = 1 \right), \tag{3.3}$$

$$\hat{G}_{jk}(g_{l_j}, g_{l_k}) = \sum_{i \in \mathcal{O}_{jk}} \hat{X}_{ij}(g_{l_j}) \cdot \hat{X}_{ik}(g_{l_k}). \tag{3.4}$$

This step results in the 3×3 tables in the boxes "observed genotype combination" and "expected genotype combination" in Figure 3.1. Using the product of the marginal probabilities of the single marker genotypes for calculating the probability of the genotype combination in Equation (3.4) mimics the assumption of no epistatic effects under the null hypothesis. Thus, G_{jk} and $\hat{G}_{jk}$ can be used to derive a χ^2 like statistic comparing observed and expected genotype combinations on two markers.

5. The test statistic is obtained by first calculating the squared difference of observed and expected frequencies for each genotype combination (g_{l_j}, g_{l_k}) of two markers j and k divided by the corresponding expected frequency. The final score for a marker pair is the sum of these values over all nine possible genotype combinations:

$$\chi^2_{jk} = \sum_{l_j, l_k \in \{1,2,3\}} \frac{(G_{jk}(g_{l_j}, g_{l_k}) - \hat{G}_{jk}(g_{l_j}, g_{l_k}))^2}{\hat{G}_{jk}(g_{l_j}, g_{l_k})}. \tag{3.5}$$

3.2.2 Permutation p-values

The significance of the imbalances observed for each marker pair is assessed with a permutation approach based on pseudo-controls. This approach has already been adopted in related problems (Li, Qing et al., 2009) and is outlined in Figure 3.1B.

For each parent-child trio we infer the four genotypes that the child could have inherited from its parents at each marker j. They consist of (a subset of) the three possible genotypes (AA), (Aa) and (aa), with relative frequencies as given in Table 3.1. These genotypes are then randomly combined over markers to form pseudo-offspring genomes in which each of the possible 16 genotype pair combinations could in principle appear for each marker pair jk. The R package `trio` (Schwender et al., 2012) was used to infer pseudo-controls for our application example in Section 3.3.

The pseudo-genotypes allow us to assess the significance of the ImAP test statistic of each marker pair by calculating an empirical marker pair specific null distribution based on $N = 10,000$ permutations (random combinations of pseudo-genotypes). The permutation p-value is calculated as the fraction of pseudo-control test statistics χ^2_{perm} exceeding the

observed score:

$$p = P(\chi^2_{\text{perm}} > \chi^2_{\text{obs}}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(\chi^2_{\text{perm}}(i) > \chi^2_{\text{obs}}\right). \tag{3.6}$$

We correct p-values for the multiple hypothesis testing problem by applying the Benjamini-Hochberg approach for the false discovery rate (FDR, Benjamini and Hochberg, 1995).

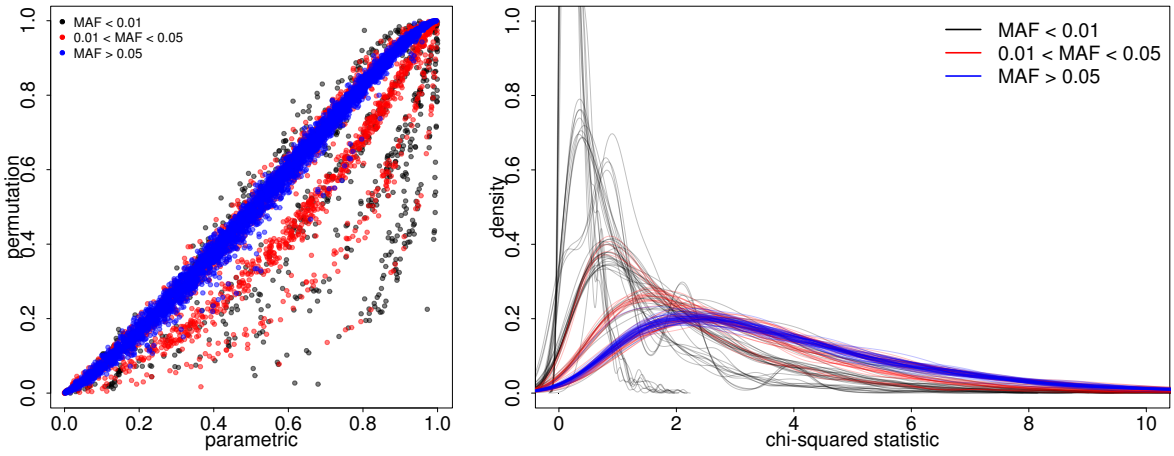
A natural approach for obtaining p-values would be the use of the χ^2 distribution of the ImAP test statistic. The degrees of freedom of this distribution depend on the marker pair and are given as $\text{df}_{jk} = (|j|-1)\cdot(|k|-1)$, where $|j|$ and $|k|$ are the actual number of genotypes present in the population for a marker pair jk. Analytical p-values for each marker pair could simply be derived from these distributions. However, we found that the distribution of these parametric p-values differed conditionally on the MAF of the markers (Figure 3.2A). The χ^2 distribution based p-values tend to be too conservative when the MAF is small. The underlying cause is a shift in the distribution of the test statistics depending on the MAF (Figure 3.2B). This phenomenon was greatly reduced when we changed to the permutation based p-value calculation (Figure 3.2A).

3.2.3 Fine mapping of interesting loci

In order to speed up the calculations but still retain an acceptable resolution of loci with potentially interacting genes, we pursued the following strategy.

In a first run of ImAP we split the data into blocks of high LD. This is done again with the package `trio`, which provides an algorithm by Gabriel et al. (2002) (as described in Wall and Pritchard, 2003) that uses confidence intervals on Lewontin's D' to estimate LD block borders in parent-offspring data (see Section A.1 for a definition of D'). Afterwards, one representative marker is chosen randomly among all markers with a minimum number of missing values in each LD block and ImAP is applied to all possible combinations of these representatives on different chromosomes. The restriction to markers on different chromosomes is applied to rule out false positive results due to local linkage disequilibrium.

Subsequently, we identify all block pairs which were assigned an FDR below 0.5 and repeat the analysis using all markers from those blocks. In this way we restrict testing of individual marker pairs to genomic regions that are suggestive for interactions. Finally, we select the highest scoring marker pairs from each locus pair as the 'interacting pairs'. This two-step approach allows an accurate mapping of epistatic interactions over the whole genome while simultaneously restricting the number of tests and the computing time to a more reasonable level.



A Permutation vs analytical p-values

B χ^2 distribution depending on MAF

Figure 3.2: Dependence of analytical p-values on MAF. A Permutation p-values vs analytical p-values based on the χ^2 distribution. The color code shows different MAF of the markers. The smaller the MAF, the more the analytical p-values are conservative. **B** Exemplary distributions of the test statistics depending on the MAF of the markers. The scores follow a χ^2 distribution with increasing degrees of freedom for larger MAF.

3.2.4 Pedigree simulation

The pseudo-control data was used to compute p-values. In order to also correct for multiple hypothesis testing and for testing of any other possible biases in our data we simulated the mating process in the mouse population assuming independence of the markers but adhering to the original pedigree structure.

The simulation starts with the first generation of mice for which we have genotype information (F0 generation). Using fastPHASE (Scheet and Stephens, 2006) we infer the haplotypes of these individuals. fastPHASE is based on the notion that haplotypes cluster into locally restricted groups which can be described using a Hidden Markov model. As opposed to other methods, fastPHASE assumes that due to recombination events the group membership changes continuously across the chromosome and not only at the block borders.

Obtaining the haplotypes of the F0 generation allows us to initialize the mating process. For each mother and father of an F1 individual we start with randomly choosing whether they pass on the maternal or the paternal allele of the first marker on a chromosome to the offspring. Then, using either general or sex-specific recombination rates (Supplementary Material in Shifman et al., 2006), we sample whether the second marker is inherited from the same chromosome or whether a recombination took place during meiosis. This procedure is continued until a complete chromosome is assembled that is passed on to the offspring. The

whole process is repeated until all generations are simulated. Subsequently, we randomly add 0.01% genotyping errors (making sure we do not introduce any Mendelian errors) as well as the same missing values as in the original data.

Since the simulation only accounts for local linkage but not for any other influences on allele frequencies, these data should not contain any true gene-gene interactions. The proportion of false positive findings should be comparable to the original data due to the same error rates and missing values.

3.3 Application to mouse genotype data

3.3.1 The heterozygous mouse stock genotype data

We applied ImAP to search for potential epistatic interactions using outbred HS mice as described in Section 2.7. We are using the genotype data of $2,002$ individuals that were genotyped at $10,168$ markers. Importantly, the pedigree of these $2,002$ individuals is almost completely known. The HS consists of 84 families, some of which are large, while others are only nuclear families. These families were derived from 40 mating pairs of mice from the original stock after more than 50 generations of random mating. Genotypes were obtained with the Illumina BeadArray platform achieving call rates of 99.86%, the genotyping accuracy was greater than 99.9% (Shifman et al., 2006).

After removing individuals with more than 5% missing data, we were left with $2,000$ individuals. In addition, we excluded markers with more than 5% missing values and/or a MAF less than 0.1. Since we observed a rather poor quality of the genotypes on the X chromosome with relatively few markers passing the quality criteria, we discarded data from this chromosome altogether. The filtering resulted in $8,091$ markers used for the subsequent analysis.

We did not have to discard any SNPs due to lack of HWE as is generally done in genome-wide association studies. Instead, ImAP corrects for the disequilibrium (Section 3.2.1). In the first run of our analysis, 230 out of $1,159$ markers had correction factors greater than 1.1 or smaller than 0.9. There are several explanations for the deviation from HWE, for example natural selection, genetic drift or segregation distortion (Griffiths, 2000; McLean et al., 1994). Even though it might not be possible to distinguish the source of disequilibrium, our correction can be applied anyway.

3.3.2 Interactions between LD block representatives

When applying ImAP to the HS mouse data, we limited our analysis to markers residing on different chromosomes in order to exclude local LD (Payseur and Place, 2007). An

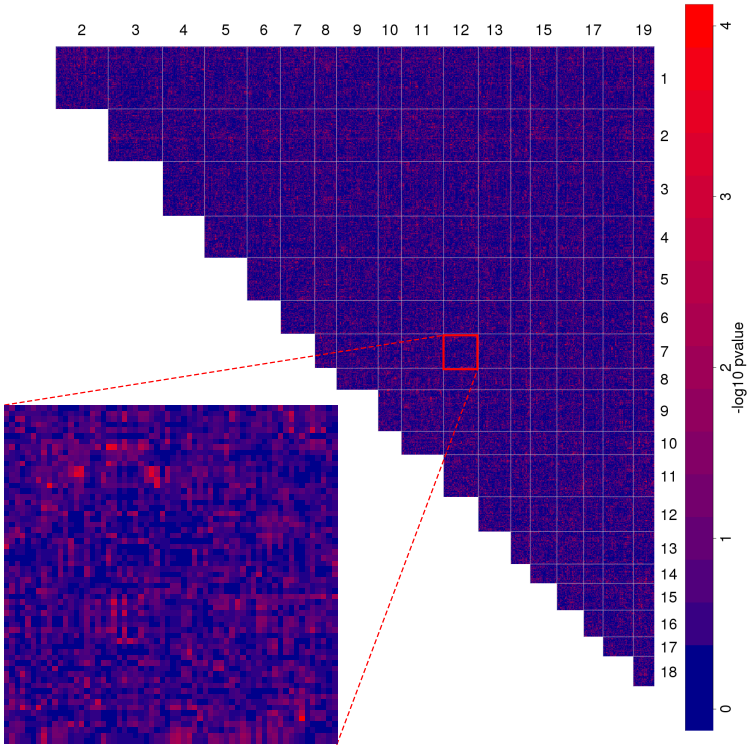


Figure 3.3: Genome-wide map of allele incompatibilities. The heatmap shows the negative $\log_{10}$ p-values of each LD block combination on different chromosomes. Light red spots show putatively interacting loci. Inset shows an enlargement of chromosome 7 versus chromosome 12.

alternative approach would have been to determine local LD first and subsequently apply ImAP to regions outside local LD. As described in Section 3.2.3, we first applied ImAP to a reduced set of $1,159$ markers, one per LD block.

Figure 3.3 shows the spatial distribution of the interactions at the level of LD blocks in a genome-wide map. As expected, most block pairs do not interact. At a p-value cutoff of 0.0001 we identify 168 interactions between 272 distinct loci (i.e. LD blocks). This p-value corresponds to an FDR of 0.5. Although we did not achieve very low FDR values, they were still markedly lower than in five simulated data sets. In two of the simulations the minimum FDR was above 0.5.

Most of the loci interact with only one other locus, only 10 loci participate in more than 2 interactions (Figure 3.5). Not surprisingly, there are more significant interactions between large chromosomes with many measured markers than between small chromosomes (Figure 3.4). However, we also found remarkable differences in the relative number of interactions per chromosome. Especially chromosomes 2, 12 and 19 incorporate more loci carrying

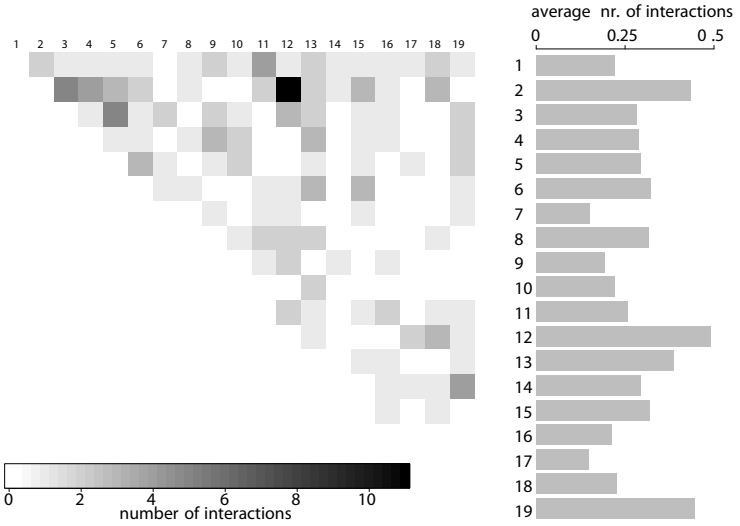


Figure 3.4: Number of interactions per autosome pair. Results are based on the 168 significant LD block pairs involving 272 loci. The barplot on the right shows the average number of interactions per LD block for each chromosome. Chromosomes 2, 12, and 19 show the highest participation in interactions while the fewest interactions per LD block are on chromosome 17.

allelic incompatibilities than other chromosomes. To see whether the number of interactors per chromosome is different from what would be expected by chance, we simulated the 168 interacting marker pairs 100,000 times and compared the distribution of the number of interactors per chromosome to the observed values. At a nominal 5% significance level, three chromosomes (2, 7, and 12) differ from their expected values. At this significance level, we expect less than one chromosome to differ significantly by chance. Hence, there is significant variation of the number of interacting LD blocks between chromosomes.

In order to rule out the possibility of false positive findings due to increased numbers of missing values or small MAF on some markers, we compared the distributions of missing values and MAF between block representatives from significant block pairs to those of non-significant pairs (Figure 3.6). There are no significant differences between the proportion of missing values (Wilcoxon rank sum test, p-value 0.67). The MAF tends to be even higher in the significant blocks compared to the other blocks. Thus, our results are not biased by missing genotypes or differences in MAF.

The histograms in Figure 3.7 compare the distribution of the p-values that we obtained by applying ImAP to the original block representative data with those resulting from five simulations. While the histograms of the simulated data sets resemble those of uniformly distributed p-values under the null hypothesis of no interacting loci, the original data show a clear peak in the low p-value range. The simulated pedigrees contain significantly

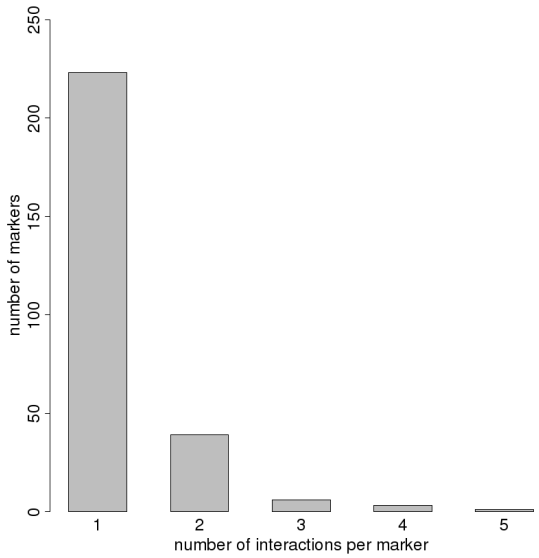


Figure 3.5: Number of interactions per LD block. Number of interactions for each of the 272 loci involved in the 168 LD block interactions with $p \leq 0.0001$. 6, 3 and 1 loci have 3, 4 and 5 interactors, respectively.

less interactions with low p-values than the real data (one-sided Kolmogorov-Smirnov test p-values $< 10^{-23}$). The p-value distribution of the observed genotypes is also significantly different from a uniform distribution (one-sided Kolmogorov-Smirnov test, $p < 10^{-69}$). This is not the case for all but one of the simulations (p-values 0.991, 0.587, 0.994, $< 10^{-12}$, 0.995). Both results confirm that there are more imbalances in allele pair frequencies than expected by chance.

This difference between the real and simulated data can now be quantified to make suggestions about the number of true allelic incompatibilities in the HS mouse population. For example, at $p \leq 0.0001$ (corresponding to an FDR < 0.5) we find between 26 and 58 more significant block pairs in the original data compared to the simulations.

As can be seen in the inset of Figure 3.3, each chromosome pair exhibits only few such interacting pairs that are often surrounded by less significant markers due to local linkage. To further increase the resolution in these interesting regions, we performed fine mapping of all marker pairs in the significant block pairs.

3.3.3 Fine mapping of interactions

For the second step of the analysis we chose all LD blocks that were involved in at least one significant interaction. There might be one or more interacting markers within each LD block and the above analysis does not reveal which markers within a region are involved

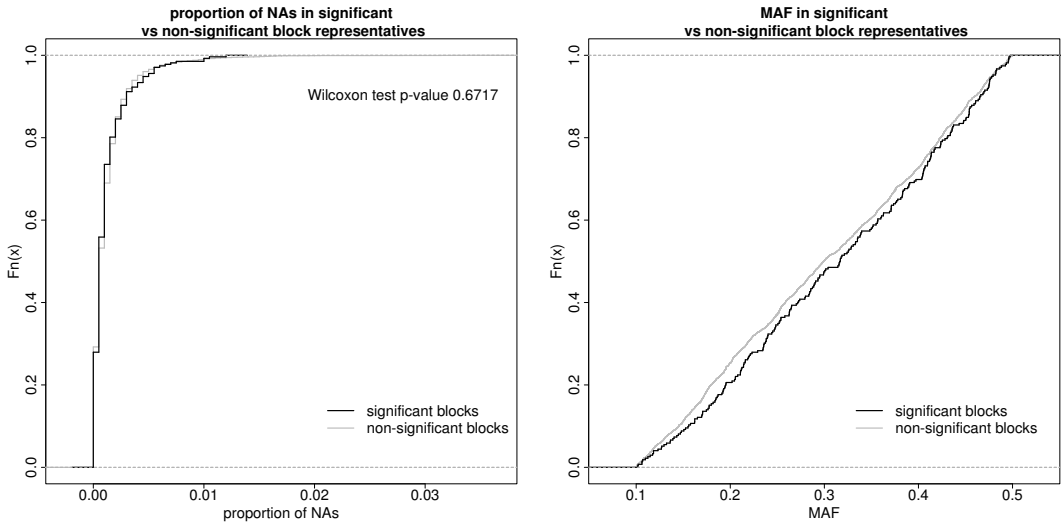
in the interactions. We repeated the calculation of the test statistics, null distribution and p-values with all markers in those blocks to find the SNP pairs with the highest signal in each significant block pair. This resulted in $1,464$ marker pairs with a p-value < 0.0005 (Tables S3 and S4 in Ackermann and Beyer, 2012), since each block pair could contain more than one significant marker pair. Note that the interpretation of the newly calculated p-values has to be done with care since a large number of the tested marker pairs is already assumed to be interacting (they were chosen from interacting LD blocks) and because markers inside LD blocks are highly correlated (i.e. not independent). Therefore, it is difficult to correct for multiple hypothesis testing. However, we can still use the p-values to rank the interactions, i.e. to identify the most likely interacting marker inside each LD block.

3.3.4 Overlap with published mouse RIL data

Only few allele incompatibilities in mouse have been reported so far (Montagutelli et al., 1996; Payseur and Place, 2007). We are not aware of any analysis that quantitatively examines the number of such interactions that can be expected in the whole genome. An overview of the distribution of allele imbalances in RIL is given in Williams et al., 2001. The authors inferred the correlation between locus pairs as a measure for distant LD. The strains used in this study are partly identical to the progenitors of the HS. Thus, it is reasonable to assume at least partial overlap of incompatible locus pairs between our study and the RIL data.

We therefore investigated the distant LD of markers that were genotyped in the RIL as well as in the HS mice. We downloaded the genotype data for 322 inbred mouse strains (`www.genenetwork.org`) and recalculated Pearson's correlation coefficient (R^2) as well as the MAF of the common markers. This allowed us to apply the same quality constraints (MAF > 0.1) to the RIL data as to the HS genotypes. Moreover, only marker pairs on different chromosomes were considered. After the filtering, 584 markers constituting 777 informative pairs were used for the analysis.

Figure 3.8 compares the overall distribution of distant LD in the RIL data with the distribution of markers showing high ImAP scores. There is a significant difference between the background distribution of R^2 of common marker pairs on different chromosomes and the R^2 of the top ImAP pairs (one-sided Kolmogorov-Smirnov test, p-value 0.0004). Marker pairs with a significant ImAP score tend to be more in distant LD than other marker pairs. More specifically, 292 out of the 777 marker pairs have an absolute correlation above 0.2. Thus, a significant number of interactions obtained from the HS can be independently confirmed in a different set of mouse populations.



A Missing values **B** MAF

Figure 3.6: Relationship between ImAP scores and missing values or MAF. The Figures show the cumulative distribution functions of the proportion of missing values (**A**) and MAF (**B**) of representative markers of significant and non-significant LD block pairs.

3.3.5 Functional enrichment of ImAP interactions

As a second check of the relevance of the ImAP outcome, we investigated if the genes mapping to loci that participate in high ranking interactions are enriched for relevant functional categories. The Gene Ontology (GO) (Ashburner et al., 2000) provides a systematic ordering of genes into functional classes in a tree-like structure. We use the GO "Biological Process" ontology to annotate genes related to marker loci with their biological function.

ImAP detects interactions between markers, not genes. Thus, in order to perform an enrichment analysis we have to assign gene functions to markers. A conservative solution to this problem is to assign to a marker j the functions of all genes encoded between the flanking markers $j-1$ and $j+1$. If there actually exists a functional enrichment among genes causing allele incompatibilities this enrichment will be 'diluted' due to this procedure. However, since we do not know the causal genes *a priory* there is no other rigorous way of performing such GO enrichment. This strategy also prevents a bias in GO enrichment due to local gene clusters with similar annotation.

We further restricted the enrichment analysis to interacting marker pairs whose 3×3 genotype table contained exactly one cell with a zero entry. This corresponds to locus pairs where one allele pair combination was not observed at all in the sample and can thus be assumed to be lethal. We reasoned that genes involved in such an interaction have functions

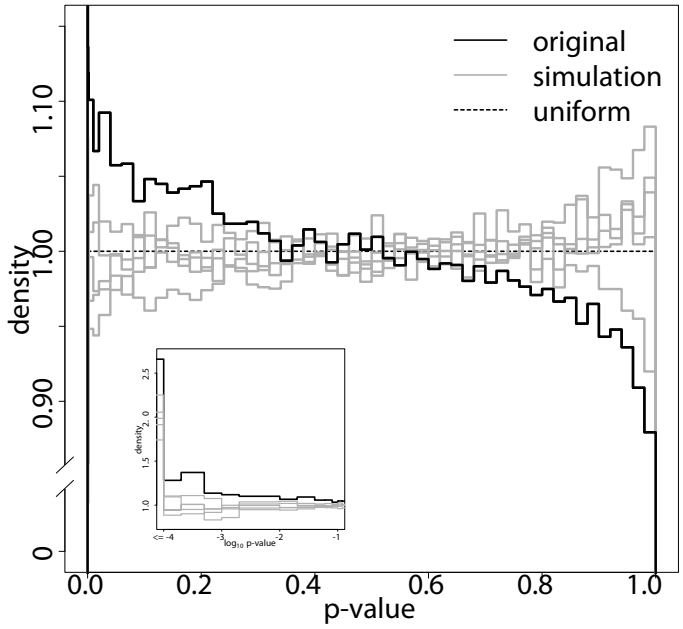


Figure 3.7: ImAP p-value distribution. Distribution of the p-values of the original data (black) and five simulations under the null hypothesis of no allelic incompatibilities (grey). The y-axis is concentrated on the interesting area of high density. The inset shows a zoom on the small p-values in $\log_{10}$ scale.

related to organism development. The mapping of genes and their associated GO terms to these markers resulted in $1,314$ markers having at least one GO term assigned to them. Seventy-three of these markers are involved in one of the significant interactions.

The enrichment test was conducted using the topGO algorithm (Alexa et al., 2006). An advantage of topGO is that it corrects for multiple hypothesis testing, particularly taking into account the nested structure of the GO tree. Since the multiple hypothesis testing correction is inherent in the algorithm, the authors suggest to use the unadjusted p-values as a ranking criterion. We call all terms significant with a p-value < 0.01 based on the "weighting" algorithm of topGO.

The top ranking GO biological process terms for the original data as well as for an exemplary simulation are shown in Tables B.1 and B.2 in the appendix. We found more significant and more relevant GO terms in the original data compared to the simulation. As expected, many of the significant GO terms are related to developmental processes such as germ cell layer development and development of brain, lung and epithelium. A lot of interesting terms had p-values just above the threshold of 0.01 (e.g. stem cell maintenance

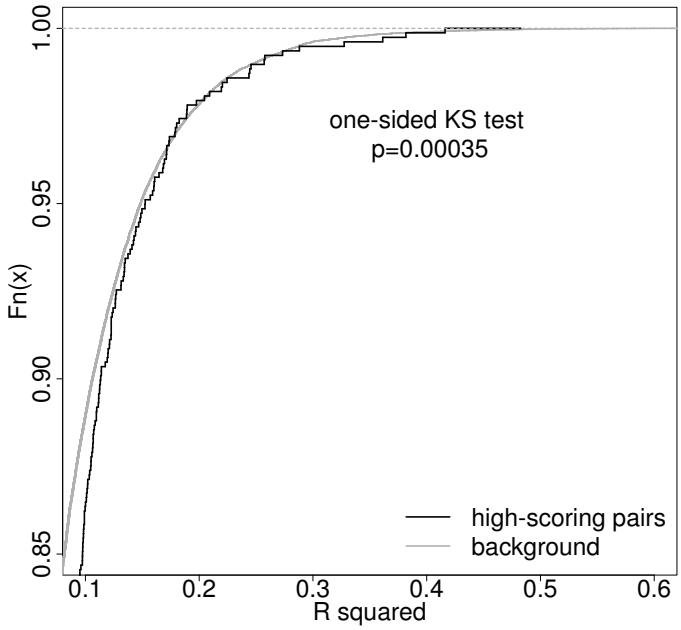


Figure 3.8: Distant LD in RIL with respect to ImAP. Cumulative distribution function of the overall distant linkage disequilibrium in the RIL (grey) and RIL marker pairs with ImAP p-value ≤ 0.0005 (black).

(p = 0.013), anterior/posterior axis specification (p = 0.021) or determination of left/right symmetry (p = 0.032)). This analysis shows that markers participating in interactions are enriched for relevant GO categories. One might also expect that pairs of interacting markers share similar functions. However, we did not observe that interacting markers share GO categories more often than expected by chance (data not shown).

3.3.6 Comparison of interaction profiles

Epistatic interactions affecting the viability of an organism often bridge parallel biological pathways (Kelley and Ideker, 2005; Beyer et al., 2007). A pathway can be regarded as a set of genes or their protein products, which act together in a concerted way in order to effect a determined biological function. The assumption underlying the between-pathway model is the existence of functional redundancy among pathways. A decrease in functionality of only one of two genes operating in two redundant pathways still allows for regulation of the downstream process through the second alternative pathway. However, if both genes are dysfunctional, both pathways will be disrupted, which may lead to a severe phenotype (i.e. an epistatic interaction between the two genes). Therefore, two genes in the same

pathway should share some of their interaction partners, namely those in a functionally similar pathway (Roguev et al., 2008). Thus, the interaction profiles of genes in the same pathway should be correlated (Figure 3.9A).

Here, we are interested in markers having a significant number of common interactors. In order to find such groups of markers with similar interaction profiles, we compared the marker interaction profiles from the ImAP analysis using the congruence score (Ye et al., 2005). It is calculated as the negative $\log_{10}$ transformed p-value of a hypergeometric test for the number of shared interaction partners (see Section A.5 for details). Thus, the score relates the number of interactions shared between two markers to the total number of interactions each single marker participates in (Ye et al., 2005).

Since here we are analyzing interaction profiles (i.e. all interactions of a given marker rather than single interactions) we chose a less stringent cutoff value for interacting block pairs ($p < 0.001$). Even though using the more stringent cutoff of 0.0001 also yielded more correlated interaction pairs in the real data than in the simulations, choosing a higher cutoff increases the difference between real and simulated data. The fraction of block pairs with congruence scores > 1 and > 2 is higher in the original data than in the five simulations (Figure 3.10). This difference between the proportions is significant in four out of five cases for a significant congruence score (> 2, one-sided χ^2 test p-values $< 10^{-5}$, 0.239, $< 10^{-15}$, < 0.0001 and $< 10^{-15}$). Thus, interaction profiles are more consistent in the real data compared to our simulations.

3.3.7 Combining ImAP scores with expression data

An important and nontrivial step in any genetic mapping study is to identify the causal genes encoded in the significant loci. Additional, independent genomic information has been widely used to prioritize genes in a genetic region of interest (Suthram et al., 2008; Lage et al., 2007; Lee et al., 2009).

Here, we are using expression data for prioritizing candidate genes at interesting loci. It is likely that several of the allele incompatibilities are caused through functionally relevant changes of gene expression between the minor and major alleles at the two loci (Mehrabian et al., 2005). We used expression data from three tissues (lung, liver, hippocampus) measured in a subset of the HS mice (257, 273 and 468 individuals, respectively). For each marker we considered all genes encoded in the region defined by the flanking markers. We then filtered for genes showing significant expression differences between individuals carrying the major versus minor alleles. This analysis was performed independently for each marker using one-way analysis of variance (ANOVA) with the three possible genotypes as levels. Each genotype had to be observed in at least 5 individuals.

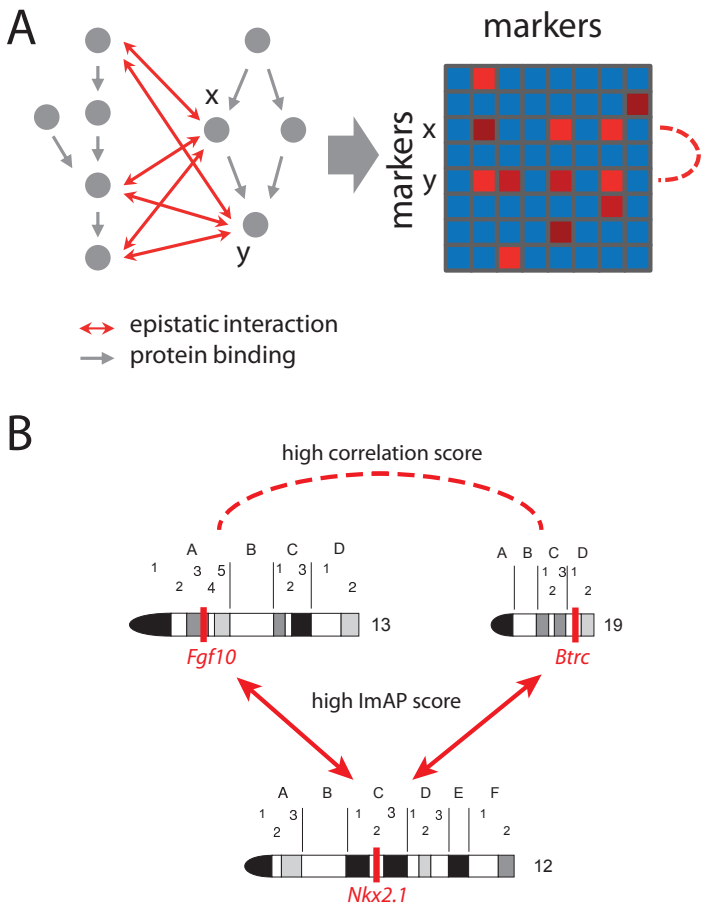


Figure 3.9: Correlated interaction profiles. A Schematic showing relationship between epistatic interactions and molecular pathways. The genes x and y share three allele incompatibilities with genes from a parallel pathway. In the schematic interaction matrix on the right these shared interactions (red color shading) lead to correlated interaction profiles (between rows, indicated by a dashed line). **B** Example of two loci on chromosomes 13 and 19 sharing a common interacting locus on chromosome 12. The position of the loci on the chromosomes is indicated by red bars. The putatively causal genes are written below the loci. Arrows indicate interactions with ImAP p-values < 0.0005, the dashed line indicates a high congruence score (> 2).

Among the $1,464$ top scoring ImAP pairs, we found 204, 113 and 122 pairs where each locus contained at least one differentially expressed gene (p-value < 0.05) in the hippocampus, liver and lung data sets, respectively. 23 locus pairs were associated with the same differentially expressed genes in all three tissues.

Among the 525 marker pairs with a congruence score greater than 2 there are 68, 25 and 43 locus pairs containing at least one differentially expressed gene in the hippocampus, lung and

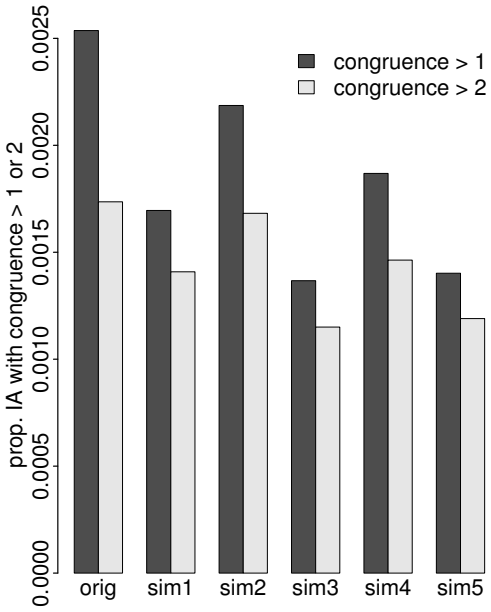


Figure 3.10: Congruence scores of original data versus simulations. Fraction of congruence scores > 1 and > 2 for interaction profiles in original data and five simulations.

liver data, respectively. Figure 3.9B shows an example of such a marker pair. The putatively causal genes *Fgf10* and *Btrc* showed differential expression (ANOVA p-value $< 10^{-6}$) in the hippocampus. The two genes are critically involved in the development of several tissues such as lung, mammary gland, tooth or telencephalon (Maeda et al., 2007; Kudo et al., 2004; Pedchenko and Imagawa, 2000; Miletich et al., 2005; Pispa et al., 1999). This is consistent with the GO terms we found to be enriched among the top scoring ImAP pairs (Table S1). *Btrc* is an inhibitor of Sonic Hedgehog (*Shh*) signaling, which is involved in the development of the lung and the telencephalon (Gulacsi, 2006). Both, *Fgf10* and *Shh* signaling are involved in development of anatomical structures and are known to influence each other (Hébert, 2005).

According to our gene expression analysis, the minor allele of *Fgf10* leads to a reduced expression of this gene while individuals carrying the minor allele of *Btrc* show a higher expression than individuals with the major allele. Since *Btrc* is an inhibitor of *Shh* signaling, this implies that both minor alleles reduce Hedgehog signaling.

The *Btrc* and *Fgf10* loci share 13 ImAP interactions. One of them involves a locus on chromosome 12 containing, among others, the homeobox transcription factor *Nkx2.1*, which is indispensable for lung and telencephalon development. Depending on the cell type and developmental stage *Nkx2.1* either interacts with the *Fgf10* and *Shh* pathway (Gulacsi,

2006; Sakiyama, 2003) or it independently acts in parallel (Minoo, 1999). Thus, the reduced activity of Hedgehog signaling in carriers of the minor *Btrc* or *Fgf10* alleles may be rescued by a fully functional *Nkx2.1*. The ImAP analysis suggests that the combination of the minor allele at the *Nkx2.1* locus together with minor alleles at either the *Btrc* or *Fgf10* locus leads to an embryonic lethal phenotype, presumably due the loss of the buffering effect of *Nkx2.1*.

3.4 Discussion of ImAP and related issues

We present a new approach to infer epistatic interactions on a genome-wide scale in family data using sequence information only. The method scans all marker pairs in the genome for deviation from the expected allele pair frequencies resulting in a list of putative pairs featuring an allele incompatibility. Relying on sequence data only is an advantage compared to existing methods for the inference of gene-gene interactions, since the approach can readily be applied to existing SNP data. There is no need for resource- and cost-intensive phenotype measurements.

Regression and χ^2 methods have been proposed in the past for the identification of epistatic interactions (Spielman et al., 1993; Cordell, 2002; Cordell et al., 2004; Cordell, 2009; Liu et al., 2011; Wang et al., 2010) and the two approaches have been shown to be interconvertible (Agresti, 2002). We chose a χ^2-based approach, which makes the fewest assumptions about the underlying genetic model (Zheng et al., 2009). Which ever way the detection of allele incompatibilities is performed, the key notion is to implement means for accounting for the confounding factors and to remove single-marker effects (e.g. leading to a deviation from Hardy-Weinberg equilibrium). Only after considering these confounding factors we got an appreciable number of significant allele incompatibilities.

We identified substantially more interacting loci than expected by chance, which is first evidence that we detect true 'signal'. Further, we could show that interacting marker pairs are enriched for genes involved in developmental processes and a significant number of interactions could be validated using independent external data. Due to the very large number of pairs tested, finding a large number of interactions with low p-values even in the simulations is expected. However, at low p-values we observed significantly more interactions in the original data than in any of the simulations; e.g. at $p \leq 0.0001$ we found at least 26 interactions more than in any of the simulations. Considering that so far virtually no allele incompatibilities between mouse strains were reported, this is a surprisingly large number. Suitable statistical tools for the detection of allele incompatibilities at a genomic scale did not exist so far. Hence, this study presents first evidence about the extent of allele incompatibilities in model populations such as the HS. Although the number of interactions we identified might not seem immense, it partly explains the difficulties faced when breeding recombinant inbred

lines (Williams et al., 2001). For example, during the generation of the Collaborative Cross, a multiparental recombinant inbred strain panel, 198 of the 650 initial lines were lost during the first three to five generations of inbreeding (Chesler et al., 2008). ImAP helps to better understand these issues and it can reveal potential biases in the breeding process that might be introduced due to allele incompatibilities.

Future work should also include haplotype information. Local haplotypes have been inferred for the HS population in terms of probability of inheritance from any of the eight founder strains (Mott et al., 2000). That means haplotypes are expressed as 8-dimensional vectors of probabilities. Consideration of these haplotypes would remarkably increase the complexity of the analysis (thereby also increasing the number of tested hypotheses), but it might further improve the accuracy.

An epistatic interaction is always defined with respect to a specific phenotype. In this study the phenotype is implicit, hidden. Indeed, looking for allele pairs that are under-represented in the HS population reveals the genotype of the non-existing individuals. Therefore, the hidden phenotypes should relate to any biological processes affecting the fertilization, the development or the viability of an individual and thus prevent its appearance in the population. Interestingly, top scoring marker pairs are enriched for genes involved in these expected phenotypes.

It is not immediately obvious how our findings translate to human populations (Stearns et al., 2010; Kosova et al., 2010). Although we are working with outbred mice, they were derived from eight genetically distinct inbred strains. These founder strains differ at at least $311,647$ genomic positions (SNPs and structural variations, Keane et al., 2011). It is likely that many of the incompatibilities that we see in the HS developed in the inbred founder strains used for establishing the HS. Even though allele incompatibilities cannot evolve in mixing populations, also human races have been in isolation for more than 100 generations (de la Chapelle, 1993; Li and Durbin, 2011; Gutenkunst et al., 2009). Hence, it is possible that an appreciable number of incompatibilities exist in the human species. Anderson et al. (2010) have shown that incompatibilities in yeast can manifest already after relatively few (approximately 500) generations. Again, also that finding is not easily transfered to mammals, as the speed of such process will depend on several factors, including recombination- and mutation rates. As the number of family trios that is being fully sequenced increases (Durbin et al., 2010; Roach et al., 2010), we expect that our framework will be applicable to human populations within the next years to address these questions.

Chapter 4

Dynamic eQTL

The work presented in this chapter is related to the following publication:

Ackermann, **Marit**, M. Clément-Ziza, J. J. Michaelson, and A. Beyer (2012). Teamwork: improved eQTL mapping using combinations of machine learning methods. *PLoS ONE 7*(7), e40916.

Moreover, the following manuscripts are in preparation:

Ackermann, **Marit**, W. Sikora-Wohlfeld, and A. Beyer (2012). Impact of natural genetic variation on gene expression dynamics. *submitted*.
Sikora-Wohlfeld, W., **M. Ackermann**, E. Christodoulou, and A. Beyer (2012). Transcription factor target gene identification based on ChIP-seq data. *submitted*.

4.1 Introduction to dynamic eQTL mapping

Natural genetic variation affects gene expression levels and thereby impacts on molecular and physiological phenotypes such as protein levels, cell morphology or disease phenotypes. In this respect, gene expression has proven instrumental as an intermediate phenotype from which conclusions about the emergence of high level traits can be drawn. A genetic locus containing a sequence variant that affects transcript levels of a gene is called an *expression quantitative trait locus* (eQTL). Studying eQTL has demonstrated its value for revealing the molecular mechanisms underlying disease associated SNPs, that were previously identified e.g. through genome wide association studies (GWAS) (Dermitzakis, 2008; Altshuler et al., 2008). Moreover, it has been shown that eQTL SNPs are more likely to be disease causing than random genetic loci (Zhong et al., 2010) and can thus be used to prioritize genetic markers in GWAS.

Differences in mRNA expression levels caused by natural genetic variation can manifest themselves between individuals, populations, environments and, very importantly, between cell types and tissues (see Dimas et al. (2009) and Nica et al. (2011) and references therein). Since cells forming different tissues must have very different morphology, organization and

function, distinct patterns of gene expression are required for each cell type. This variation of gene expression between cell types is under the influence of natural genetic variation. A number of studies (summarized in Table 4.1) compared eQTL across different cell types and tissues in mouse and human samples and report that 5% to 94% of the eQTL are cell type-specific. Potential reasons for the seemingly divergent outcomes of these studies are the different levels of relatedness of tissues under study and the different sample sizes of the studies. The last point is especially important in that cell type specificity is probably over-estimated due to low power of eQTL studies (Dimas et al., 2009; Lohmueller et al., 2003). Nevertheless, there is clear evidence for cross-tissue differences in genetic variation influencing transcript levels. This raises the question whether conclusions drawn from an eQTL study in one cell type or even a cell line translate to other cell types, a problem which is highly relevant when explaining disease mechanisms with eQTL studies that are conducted in tissues which are different from the disease tissue or when several cell types are involved in the disease etiology.

Another layer of complexity is added when considering dynamic processes such as cellular differentiation or response to internal or external stimuli. These situations go along with drastic changes of the cell's morphology or molecular state being induced through the adaptation of gene expression patterns. Here we propose to not only compare eQTL observed in individual cell types (at steady state), but to additionally map the expression changes measured during the cell state transitions as traits. More specifically, we do not only distinguish static and non-static eQTL, as has been done before (Table 4.1). Instead, we further divide the group of non-static eQTL into eQTL observed for specific cell states/types and eQTL resulting from mapping expression changes/differences as traits. We show that these classes of eQTL represent different sets of eQTL corresponding to different modes of expression variation.

The main goal of the present study is to provide a functional classification of eQTL reflecting the spectrum of genetic contributions to gene expression variation over a range of dynamically changing cell states. A well-studied model for a dynamic process, being accompanied by substantial gene expression changes, is the differentiation of hematopoietic stem cells (HSCs) into the different lineages of mature blood cells (Gerrits et al., 2008). We decided to use this system to investigate eQTL based on three different categories of expression-based traits: (i) eQTL that are observed across all cellular states (*static eQTL*), (ii) eQTL being specific to one or a subset of cell states (*conditional eQTL*) and (iii) eQTL affecting changes of transcript levels during differentiation (*dynamic eQTL*). Although our scheme can serve to classify eQTL across very generic cell *states*, we will use the term cell *type* in the remainder of this thesis, referring to the application to hematopoietic cell types.

In Section 4.2 we introduce these three classes of eQTL and propose strategies to map

eQTL in the different classes. By applying our proposed approaches to gene expression data from mouse inbred lines taken from four different hematopoietic cell types (Gerrits et al., 2009) in Section 4.3, we demonstrate that eQTL from the above three classes, although based on the analysis of the same set of expression and genotype data, comprise different sets of regulatory loci having to be inferred from separate mappings. In particular, we show that basic cellular processes and state and differentiation specific functions are regulated by different eQTL categories. The choice of the eQTL mapping procedure has considerable influence on the outcome of the study. Section 4.4 summarizes and discusses the most important aspects of the analysis. The R code to reproduce the analysis can be found in Appendix E.

Table 4.1: eQTL tissue specificity. Proportion of tissue-specific eQTL reported in different studies in mouse and human. We report the tissues/cell types that were analyzed, whether only local (i.e. *cis*) eQTL or both local and distant eQTL were inferred. The last column describes whether eQTL mapping was conducted separately in each cell type or by including a tissue factor into the analysis.

tissues	**proportion of specific eQTL**	**ref**	***cis/trans*-eQTL**	**mapping strategy**
liver, muscle, SAT, VAT, peripheral blood	93.6%	Fu et al. (2012)	*cis*	separate mappings, meta-analysis on non-blood tissues
blood and LCL	44%	Powell et al. (2011)	*cis*	separate heritability analyses
T helper and regulatory T cells	37.5% (*cis*), 3% (*trans*)	Alberts et al. (2011)	*cis* and *trans*	separate mappings
blood and adipose tissue	50%	Price et al. (2011)	*cis*	single- and cross-tissue heritability estimates
LCL, skin and fat	29%	Nica et al. (2011)	*cis*	separate mappings
normal, uninvolved and lesional psoriatic skin	$1-5\%$	Ding et al. (2010)	*cis*	separate mappings
liver, omental adipose and subcutaneous adipose tissue	$19\%-28\%$	Zhong et al. (2010)	*cis* and *trans*	separate mappings; specificity defined as the fraction of eQTL occuring in at most 2 out of 3 tissues
LCL, heart, kidney, liver, lung, testes	48%	Bullaughey et al. (2009)	*cis*	separate mappings; eQTL were selected to have a strong effect
fibroblasts, LCL and T cells	79.5%	Dimas et al. (2009)	*cis*	separate mappings
HSC, myeloid progenitor cells, erythroid cells and myeloid cells	78%	Gerrits et al. (2009)	*cis* and *trans*	separate mappings, ANOVA including cell type and interaction effects
PBMC and cortical brain tissue	74%	Heinzen et al. (2008)	*cis*	separate mappings
blood and adipose tissue	50%	Emilsson et al. (2008)	*cis*	separate mappings

4.2 Methods

4.2.1 eQTL classification

We distinguish static, conditional and dynamic eQTL. A static eQTL affects a gene's expression in all conditions under consideration (Figure 4.1A). It is independent of the cell type and will thus be detected in all cell types. In contrast, a conditional eQTL is only active in one or a subset of the conditions under consideration (Figure 4.1B). Dynamic eQTL drive changes in mRNA levels during the transition from one cell type to another (Figure 4.1C).

In this respect our definition of dynamic eQTL differs from definitions used in the literature. For example, Gerrits et al. (2009) define a dynamic eQTL as an eQTL that is present in one condition but not in another. We refer to those eQTL as conditional. A concept very similar to dynamic eQTL has been introduced in the context of studying transcriptional regulation in different growth conditions in yeast (Smith and Kruglyak, 2008). The authors define eQTL affecting expression changes between conditions as gene-environment interaction eQTL (gxeQTL). A similar study has been conducted on differential expression in two different temperatures in worms (Li et al., 2006).

Different computational means can be used to detect the three eQTL types defined above. Dynamic eQTL require mapping of the expression changes (fold changes, slopes) observed at the transition from one type to another (Smith and Kruglyak, 2008; Li et al., 2006). Conditional eQTL may be detected through independently mapping eQTL in the various cell types and then identifying such eQTL that were found in some, but not all conditions. Such an approach requires defining two thresholds: first a significance threshold (e.g. maximum p-value) for calling eQTL that are active in one cell type and second, an insignificance threshold (e.g. minimum p-value) for deciding that the same eQTL is not active in other cell types. Note that both thresholds are required and that they have to be sufficiently different. Using just one threshold would lead to a situation where all eQTL that are just above the threshold in one cell type and just below the threshold in other cell types would be called 'conditional' although the eQTL scores are very similar across all conditions.

Here we propose a different approach that we termed 'simultaneous mapping', because it simultaneously identifies static and conditional eQTL and because it simultaneously uses the expression data from all conditions (Table 4.2, Section 4.2.2). We inferred dynamic eQTL by mapping expression differences between pairs of cell types (Table 4.2, Section 4.2.4).

4.2.2 Simultaneous eQTL mapping

The goal of simultaneous eQTL mapping is to infer eQTL that are specific for each of the cell types $k = 1, \ldots, c$ (conditional eQTL) as well as static eQTL in one single analysis. To

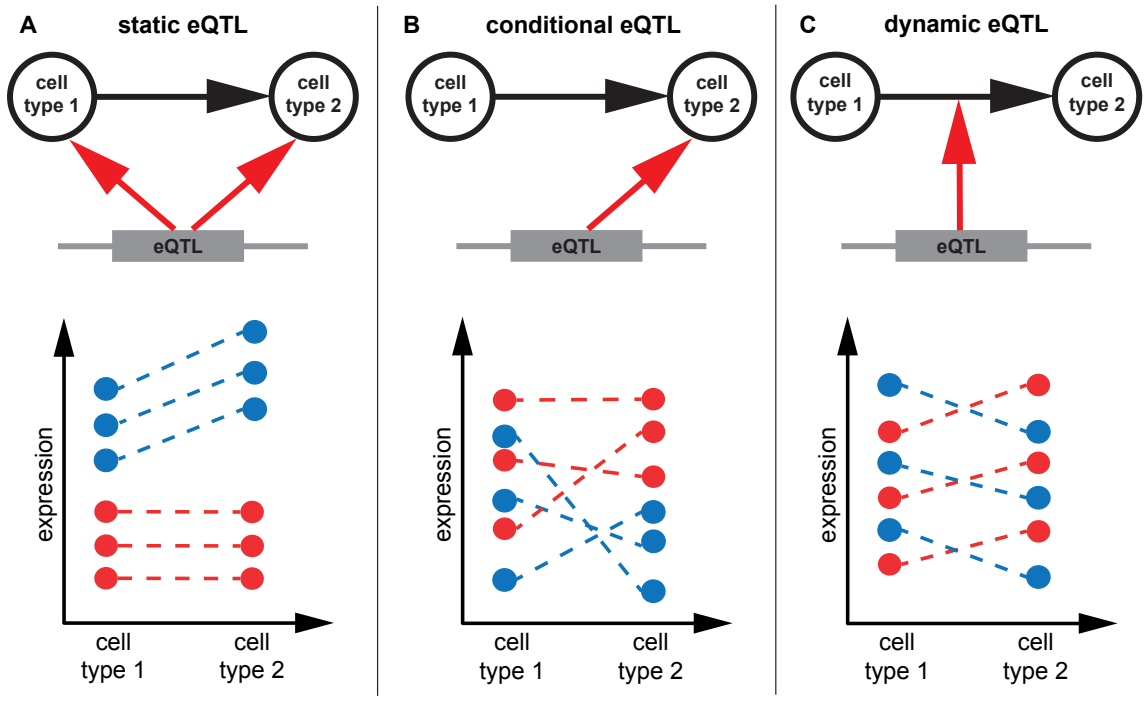


Figure 4.1: eQTL classification. Schematic representation of static, conditional and dynamic eQTL. For the sake of simplicity only two conditions are considered, but the concept is extensible to any number of cell types. The top part of each panel shows in which condition the eQTL influences a gene's expression or if it affects expression changes between cell types. The lower parts of the panels show exemplary mRNA expression profiles of the gene in six samples. The genotype of the eQTL in each sample is indicated by the color, assuming homozygous diallelic markers. **A** A static eQTL impacts expression in all cell types. The ranking of gene expressions per genotype is the same in all conditions, the slope of expression change between cell types can be similar or different between genotypes. **B** A conditional eQTL influences gene expression in only one of the two conditions. Thus, gene expression is a function of genotype in one cell type but not in the other. The slopes of expression changes may or may not be dependent on the genotype at the eQTL locus. **C** A dynamic eQTL drives expression changes between cell types. This implies that the slopes of expression changes between conditions are dependent on the genotype at the eQTL.

this end, the expression vectors $\mathbf{y}_{jk}$ of each gene j ($j = 1, \ldots, N$) from all conditions k are concatenated to form a new trait vector $\mathbf{y}_j = \left[\mathbf{y}_{j1}^T, \ldots, \mathbf{y}_{jc}^T\right]^T$ (Figure 4.2). Note that this vector might contain several entries for the same strain, each from a different cell type. In order to get a matching genotype matrix, we replicate the genotype matrix as many times as there are cell types. Because not all individuals (mouse lines) were measured under all conditions, we subset the genotype matrices to the samples for which gene expression data is

Table 4.2: Overview of eQTL mapping methods. Overview of the traits and predictors of the eQTL mapping methods applied in this paper.

Mapping method	Trait	Predictors
simultaneous mapping	concatenated gene expression over all cell types	genotypes + cell type indicators
single cell type mapping	gene expression in one specific cell type	genotypes
dynamic eQTL mapping	gene expression differences between a pair of cell types	genotypes

available. The resulting matrices $\mathbf{X}_1$ to $\mathbf{X}_c$ are concatenated in order to obtain a predictor matrix matching $\mathbf{y}$. Since we would like to distinguish static and conditional eQTL, we need to add additional predictors indicating whether a sample was measured in a certain cell type or not. Therefore, $\mathbf{X}$ is extended by as many dummy variables as there are cell types, where the entry for a given sample in a given cell type indicator is 1 if the sample was obtained from this cell type and 0 otherwise (Figure 4.2).

This new predictor matrix $\mathbf{X}$ allows to detect genetic loci affecting a gene's expression either in the same way in all cell types or differently or exclusively in a subset of cell types. We apply Random Forests (ImAP, Breiman, 2001) for mapping eQTL. RF is a machine learning approach based on an ensemble of decision trees (for details, see Section A.2). We have previously shown that multivariate eQTL mapping methods and in particular ensemble approaches such as RF outperform traditional approaches for eQTL mapping (Ackermann et al., 2012; Michaelson et al., 2010). Apart from the predictions of the response, RF returns measures for the average importance of each marker on the prediction. We use the selection frequency (SF) of each genotype marker as a measure of its importance for predicting mRNA levels. A marker that is used more often than expected by chance is an eQTL of the corresponding gene. Significance is assessed using a permutation approach (Section 4.2.6).

4.2.3 Discrimination of static and conditional eQTL

For each significant eQTL - target gene pair emerging from the simultaneous mapping (we used FDR < 0.1 as a significance threshold for the hematopoiesis data), we fit two linear models to the target's expression levels: a full model containing the eQTL genotype $\mathbf{x}_m$, a cell type factor variable $\mathbf{z}$ with as many levels as there are cell types, and an interaction

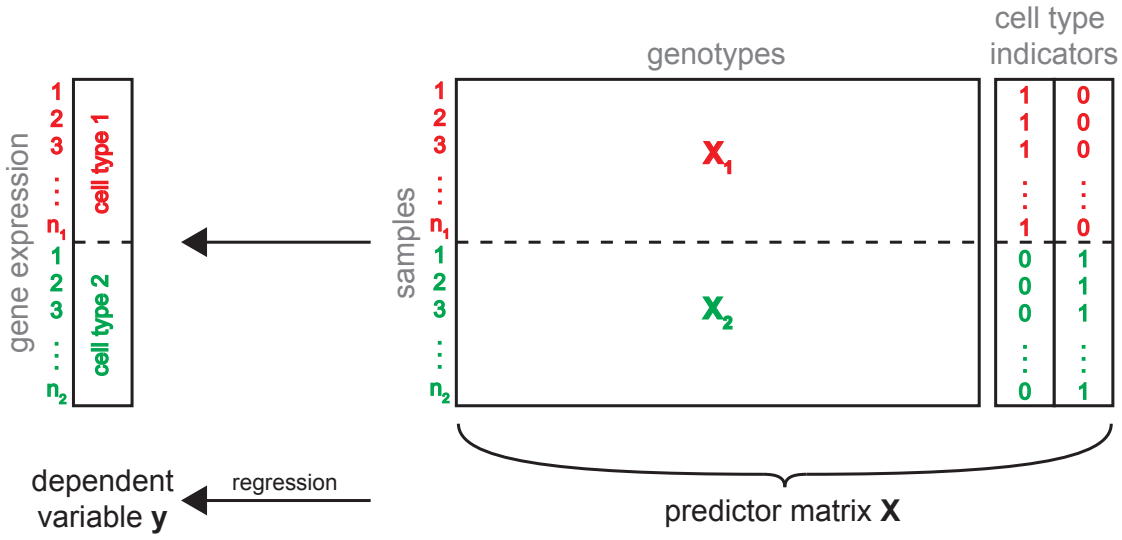


Figure 4.2: Simultaneous eQTL mapping. Schematic of simultaneous eQTL mapping for two cell types. This approach combines the available information from the two cell types (red and green) in one eQTL analysis. To this end, the gene expressions measured in the different conditions are combined into one vector $\mathbf{y}$. Similarly, for each condition the genotype matrix is subset to all samples for which there are expression measurements in this cell type. The resulting two submatrices $\mathbf{X_1}$ and $\mathbf{X_2}$ are concatenated into one genotype matrix. In order to discriminate static and conditional eQTL, two additional predictors indicating the cell type from which a sample was derived, are added to the predictor matrix. The combined genotype and cell type indicator matrix is used to find the model which best predicts gene expression simultaneously in all conditions.

term ($\mathbf{x}_m\mathbf{z}$) between the two variables:

$$\mathbf{y}_j = \mu + \beta\mathbf{x}_m + \alpha\mathbf{z} + \gamma(\mathbf{x}_m\mathbf{z}) + \mathbf{e}; \tag{4.1}$$

and a reduced model containing only the two main effects without their interaction:

$$\mathbf{y}_j = \mu + \beta\mathbf{x}_m + \alpha\mathbf{z} + \mathbf{e}. \tag{4.2}$$

In both models, μ is the mean expression over all cell types, β denotes the parameter estimate of the marker effect, α the parameter estimate of the cell type effect, γ the estimate of the interaction effect and $\mathbf{e}$ is a vector of normally distributed errors. By applying an

analysis of variance (ANOVA) on both models, we test whether the full model explains the gene expression significantly better than the reduced one (see Section A.3 for details). If this is the case, i.e. if the ANOVA FDR < 0.1, we call the eQTL 'conditional'. The cell types in which the eQTL is active are found with post-hoc Wald tests (Section A.4). The resulting p-values for each eQTL - target gene pair are corrected for multiple hypothesis testing using the stringent Bonferroni correction (Shaffer, 1995).

4.2.4 Dynamic eQTL mapping

For mapping genetic loci driving expression dynamics between two cell types, we create a new trait vector $\mathbf{y}_{k_1k_2}^{\text{diff}}$ containing the sample-wise expression differences of a given gene between two conditions k_1 and k_2:

$$\mathbf{y}_{k_1k_2}^{\text{diff}} = \mathbf{y}_{k_1} - \mathbf{y}_{k_2}. \tag{4.3}$$

We then apply RF and a permutation scheme to obtain p-values as described in Sections 4.2.2 and 4.2.6 to conduct the eQTL mapping using $\mathbf{y}_{k_1k_2}^{\text{diff}}$ as the quantitative trait. As opposed to the simultaneous mapping, the predictor matrix now only contains the marker genotype vectors of each sample and no cell type variables.

4.2.5 Mean eQTL mapping

Similar to using expression differences as a quantitative trait for dynamic eQTL mapping, we employed mean expression levels across cell types as a means to map static eQTL. Before calculating the average expression, we centered the data across strains, because otherwise mean expressions might be too much influenced by cell types in which the mRNA levels are higher compared to the remaining types. Hence, the mean expression $\bar{y}_{ij}$ of gene j in strain i is given as

$$\bar{y}_{ij} = \frac{1}{c}\sum_{k=1}^{c}\left(y_{ijk} - \frac{1}{n}\sum_{i=1}^{n} y_{ijk}\right), \tag{4.4}$$

where c is the number of conditions in which expression levels are measured and n is the sample size (assumed to be the same for each condition).

If available, one could incorporate additional information on the quality of or certainty about each of the measurements into the calculation of mean expressions. The data of Gerrits et al. (2009) were measured on Illumina Sentrix Mouse-6 BeadChips. Apart from the raw sig-

nal intensities, the Illumina image analysis software returns quality scores for the raw signal intensities for each of the probes on the expression microarray. They are given as p-values reflecting the probability of achieving the given intensity by chance. We use the negative $\log_{10}$ transformed p-values as weights for the calculation of average expressions across cell types, thereby assigning higher weights to probes that were measured with high confidence. To this end, weighted mean probe intensities across cell types $\bar{y}_{ij}^{\text{probe}}$ are calculated as

$$\bar{y}_{ij}^{\text{probe}} = \frac{1}{\sum_{k=1}^{c} w_{ijk}^{\text{probe}}} \sum_{k=1}^{c} \left[w_{ijk}^{\text{probe}} \left(y_{ijk}^{\text{probe}} - \frac{1}{n} \sum_{i=1}^{n} y_{ijk}^{\text{probe}} \right) \right], \tag{4.5}$$

where w_{ijk}^{probe} is the weight derived from the quality score for probe intensity y_{ijk}^{probe}. Subsequently, we use the median to summarize mean probe levels for each gene.

4.2.6 p-value calculation

We use the RF SF as a measure of the impact of each genetic locus on gene expression. It has previously been shown that this importance measure outperforms classic measures like the permutation importance in eQTL mapping (Michaelson et al., 2010). However, the raw SF itself is not an absolute indicator of the importance of each predictor since its distribution depends on the nature of the predictors in at least two respects. Firstly, quantitative predictors will be selected more often than binary variables (Strobl et al., 2007). Obviously, this is not a problem for eQTL mapping where all predictors can be assumed to have the same number of levels (two in the case of the panel of inbred strains derived from C57BL/6J (B) and DBA/2J (D) (BXD) mice, the B and the D allele). Secondly, groups of correlated predictors will generally achieve smaller SFs than a predictor which is not correlated with any other one (Strobl et al., 2008). This causes problems for eQTL mapping since a block of markers in LD, which contains a causal regulatory locus, might not be detected because all markers in the block are almost equally correlated with gene expression and thus have to share the locus's importance among each other.

A simple solution to this problem is the calculation of p-values based on a permutation approach, where the expression vector is permuted many times. For each permutation, the eQTL mapping with the calculation of SFs is repeated. We assume that under the null hypothesis of no correlation between a given marker m and a gene's expression, the distribution of SFs of that marker is the same for all genes. Hence, we pool SFs of each marker over all genes and all permutations in order to obtain an empirical null distribution of SFs for this marker. Finally, the p-values of an eQTL - target gene pair (marker m and

gene j) can be calculated from the SF null distribution of the corresponding marker m:

$$p_{mj} = \frac{1}{NP} \sum_{r=1}^{NP} \mathbb{1}\{\mathrm{sf}_m^r \geq \mathrm{sf}_{mj}^{\mathrm{obs}}\}, \tag{4.6}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function, N is the number of genes and P the number of permutations. sf_r^m is the SF of marker m in one permutation r out of all permutations across all genes, and $\mathrm{sf}_{\mathrm{obs}}^{mj}$ is the observed SF for marker m for the unpermuted expression vector of gene j.

The bottleneck of this approach is the run time of the RF, strongly restricting the number of permutations, which in turn results in a rather low resolution of the eQTL p-values, even after pooling SFs over genes. In order to overcome this problem, we decided to combine the permutation procedure with an analytical p-value calculation. After pooling SFs over a small number of permutations (10 in all our eQTL mappings), we fit an exponential function to the top one percent of the SF density. This is done by minimizing the squared difference between the observed SF density and the fitted density function with respect to its parameter λ:

$$\lambda_{\mathrm{fit}} = \underset{\lambda}{\mathrm{argmin}} \left(f(\mathrm{sf}) - \lambda \exp(-\lambda\ \mathrm{sf})\right)^2 \mathbb{1}\{\mathrm{sf} > \tilde{\mathrm{sf}}_{0.99}\},$$

where λ_{fit} is the fitted parameter value of the exponential distribution, $f(\mathrm{sf})$ is the observed density of the SF and $\tilde{\mathrm{sf}}_{0.99}$ is the 99% quantile of this density. The parameter optimization was conducted by applying the Nelder-Mead algorithm (Nelder and Mead, 1965) using the R function `optim` (R Development Core Team, 2011). We chose to fit an exponential distribution because all SFs are always greater than or equal to zero, where we expect most SFs to be very close to zero. This behavior is nicely mirrored by an exponential function. We also tested other distribution functions, for example the log-normal and the inverse gamma distribution, but found all of them to result in less accurate fits than the exponential distribution.

Although the density is computed based on all observed SFs, we only penalize fitting errors on the tail of the distribution. This is the part of the distribution we would like to fit most accurately since it contains the putative eQTL - target gene pairs. Consequently, this fitted exponential distribution is used to calculate more precise p-values for the top one percent of observed SFs. The remaining 99% of the p-values are still obtained from the empirical SF distribution as described before. The false discovery rate (FDR, Benjamini and Hochberg, 1995) is calculated with the procedure of Benjamini and Hochberg (Benjamini and Hochberg, 1995).

4.2.7 GO enrichment analysis

We tested for enrichment of certain biological functions among eQTL regions and target genes. We used Gene Ontology (GO, Ashburner et al. (2000)) Biological Process gene annotation, which we retrieved from the Ensembl database release 66 (`www.ensembl.org`) via the `biomaRt` (Durinck et al., 2009) interface of `R` (R Development Core Team, 2011). eQTL loci were annotated with the functions of all genes encoded in the locus or being closer to this locus than to any other (if not more than 1 cM away from it). This approach ensures a conservative evaluation of functional enrichment and prevents a bias in the results due to clusters of functionally related genes within a locus. The GO enrichment testing was conducted within the `topGO` framework (Alexa et al., 2006) implemented in the `R` package `topGO` (Alexa and Rahnenführer, 2010). We applied a Fisher test for over-representation of GO biological processes among significant eQTL loci or targets within each specific eQTL analysis or class. `topGO` offers several procedures to correct for the nested structure of the GO tree, which might inflate the significance of enrichment test results. We used the 'weight' algorithm of the `topGO` package. Although this correction also implicitly accounts for multiple hypothesis testing, we further calculated an empirical FDR for each term based on a shuffled gene/eQTL region to GO term assignment, preserving the number of terms assigned to each gene/region. We call all terms with a FDR < 0.01 significant.

4.3 Application to mouse hematopoiesis study

4.3.1 Mouse hematopoiesis data

HSC differentiation is a prominent example of a dynamic process that is heavily genetically regulated (Shivdasani and Orkin, 1996; Gerrits et al., 2008; Orkin and Zon, 2008; Iwasaki and Akashi, 2007; Swiers et al., 2006). This has been shown, among others, by analyzing natural genetic variation between mouse recombinant inbred lines exhibiting very different hematopoietic phenotypes (Müller-Sieburg et al., 2000; Van Zant et al., 1983). One of the best studied examples is the panel of BXD recombinant inbred lines that were derived from crossing the C57BL/6 and DBA/2 lines. We are using published genome-wide mRNA expression levels measured in 25 BXD strains in four cell types of HSC differentiation with varying degrees of lineage commitment: multipotent HSC with the potential for self-renewal, lineage restricted erythroid-myeloid progenitor cells, and lineage committed erythroid as well as myeloid cells (Figure 4.3, Gerrits et al. (2009)).

Gene expression data of Gerrits et al. (2009) were downloaded from GeneNetwork (Wang et al. (2003), `http://www.genenetwork.org`, accession numbers GN144-151). The data

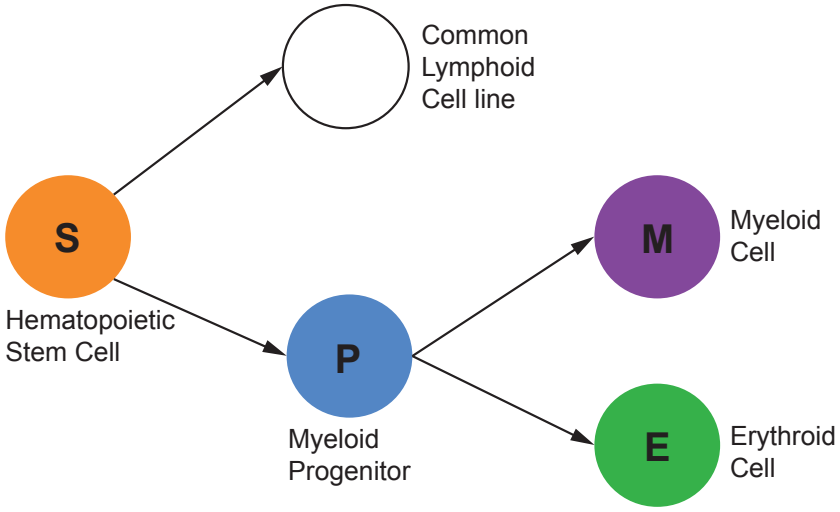


Figure 4.3: Hematopoietic stem cell differentiation. Schematic representation of hematopoietic stem cell (HSC) differentiation focusing on cell types that were analyzed in this work. Multipotent HSC with the capacity to self-renew differentiate into pluripotent progenitor cells. Progenitors are committed to the lymphoid or myeloid cell line. Our analysis focusses on the myeloid cell line, in which the progenitors can differentiate into either myeloid or erythroid cells.

were already preprocessed by the original authors. Their preprocessing consisted of $\log_2$ transformation and subsequent joint quantile normalization of expression data from all four cell types (stem cells, myeloid progenitors, erythroid and myeloid cells) as well as a batch correction. We mapped Illumina probe IDs to Ensembl gene IDs using mapping information from GeneNetwork and the R `biomaRt` package (Durinck et al., 2009) and summarized expression measurements for each gene by calculating the median expression profile over all its probes. Finally, we discarded all genes with a standard deviation of less than 0.1 in all four cell types, resulting in expression measurements of 14,724 genes on 22 to 24 BXD strains, depending on the cell type.

Genotype information of the BXD strains was also downloaded from GeneNetwork (Wang et al., 2003). Since we had expression information on only 25 strains, some neighboring genetic markers in the genotype matrix contained identical information (i.e. they were perfectly correlated). Because it is impossible to distinguish these markers with respect to their association to gene expression traits in the eQTL mapping, we merged neighboring markers

with identical genotype profiles across strains, which resulted in genotype information on 849 distinct markers or marker intervals across the mouse genome with a median interval size of 1.5 Mb (min: 4.6 kb, max: 32.1 Mb).

We applied the above eQTL classification scheme to systematically search for genetic regions causing gene expression dynamics during hematopoiesis as well as the static and conditional variation of expression in the different cell types. Using the data from Gerrits et al. (2009), we focused on three cell type transitions during HSC differentiation: from stem to progenitor cells (S-P), from progenitor to erythroid cells (P-E) as well as from progenitor to myeloid cells (P-M) (Figure 4.3).

4.3.2 Frequencies of eQTL types

Our simultaneous eQTL mapping detected $3,916$ significant eQTL target gene pairs at an FDR of 0.1. Among those, $2,729$ eQTL did not show a significant interaction with the cell type indicator and thus constitute the class of static eQTL. We also found $1,187$ conditional eQTL. These eQTL have to fulfill three conditions: (i) simultaneous mapping FDR < 0.1, (ii) FDR for interaction between marker and cell type indicator < 0.1 and (iii) the contrast test(s) for specific cell types have to be significant (Bonferroni corrected p-value < 0.005).

The number of eQTL being active under several conditions decreases with increasing number of conditions. While 643 eQTL are active in one cell type, 357, 124 and 63 have a significant contrast test result in two, three or four conditions, respectively (Figure 4.4). eQTL with four significant cell type interactions arise if an eQTL is active in all cell types, but with changing effect sizes. In this case, the ANOVA detects the dependence of the effect size of the eQTL marker on the cell type as an interaction between the marker and the cell type variable. The subsequent contrast tests then only test whether the eQTL is absent or present regardless of effect size, which is true for all cell types. In this respect, these eQTL represent a special class of conditional eQTL.

Around 9% of the static eQTL (244) are local eQTL, i.e. the target gene is encoded at or nearby the eQTL locus (left-hand side of Figure 4.5). It is assumed that local eQTL are mostly caused by mutations in *cis* and thus they are commonly referred to as *cis*-eQTL. As opposed to that, an eQTL affecting a distant gene is called a *trans*-eQTL. It is noteworthy that the number of static and conditional *cis*-eQTL is relatively similar, whereas we find substantially more static than conditional *trans*-eQTL (Figure 4.5). (Thirty-one percent (363) of all conditional eQTL are *cis* effects.) The statistical power for detecting static eQTL is much higher than the power for detecting conditional eQTL, because of the additional tests needed for detecting significant differences between the cell types, which are based on only a subset of all samples. Thus, the total number of *cis*-eQTL might be relatively limited

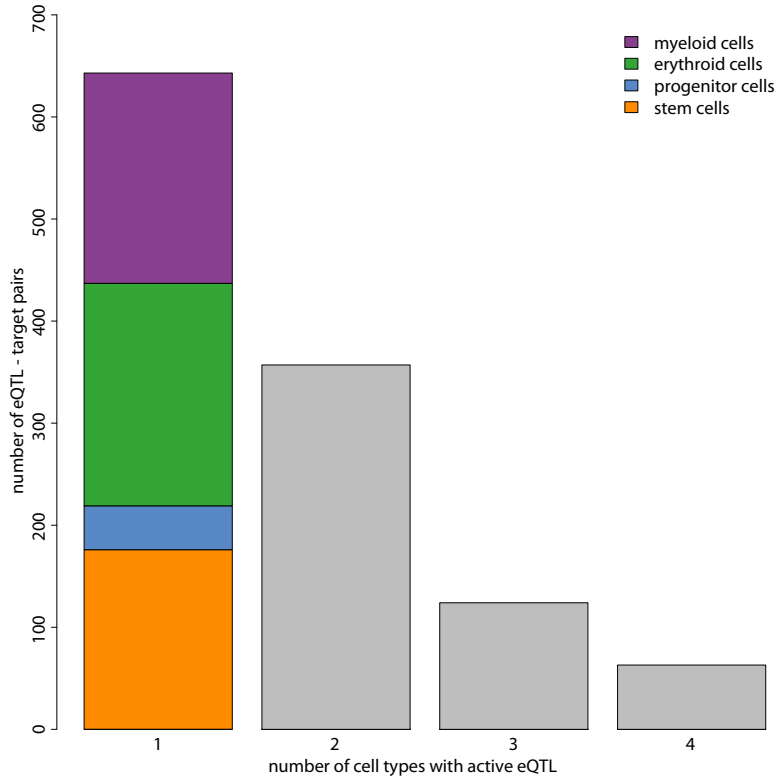


Figure 4.4: Number of cell types in which eQTL are active. The bars show the number of eQTL conditional in one, two, three or four cell types. Results are obtained from post-hoc Wald tests in the linear model comprising the eQTL marker, the cell type and their interaction. Only models with a significant marker - cell type interaction are considered. eQTL that are conditionally active in exactly one cell type are further classified by cell type (stem, progenitor, erythroid and myeloid cells).

and lower power is needed for detecting them (Petretto et al., 2006). Increasing the power by considering more samples may therefore not further increase the number of detectable *cis*-eQTL. We confirmed this interpretation by varying the number of samples considered in the analysis, which showed that increasing the number of samples increased the number of detectable *trans*-eQTL more than the number of detectable *cis*-eQTL.

Most of the eQTL that are conditional in exactly one cell type ("cell type-specific") occur in the more committed lineages (218 in erythroid cells, 206 in myeloid cells, Figure 4.4). We find less eQTL in the multipotent stem cells (176). Only 43 eQTL are significant in the progenitor cells. This might be caused by the fact that the sorting of progenitor cells is rather difficult so that this group of cells might in fact be a mixture of different cell types,

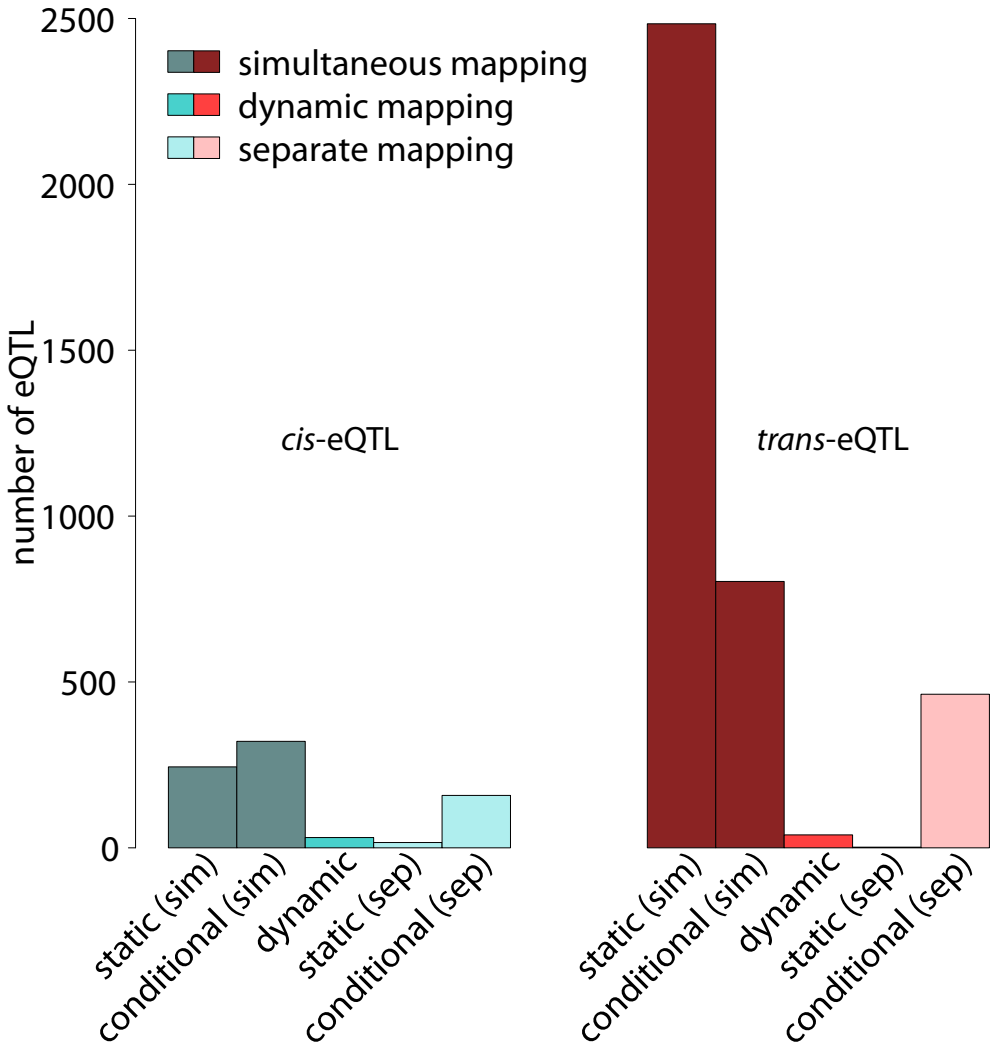


Figure 4.5: Number of *cis*- and *trans*-eQTL in different eQTL classes. Numbers of significant eQTL with FDR < 0.1 shown separately for cis-eQTL (left) and trans-eQTL (right). Static, non-static, and dynamic eQTL are distinguished (see labels at the bottom). Static eQTL are detectable in all four cell types, whereas non-static eQTL are insignificant (absent) in at least one of the four cell types tested. Further, the figure distinguishes simultaneous and separate eQTL mappings, which represent alternative ways for distinguishing static and non-static eQTL. Simultaneous mapping increases the statistical power leading to substantially more eQTL significant at the same level (FDR < 0.1). Even though both, *cis*- and *trans*-eQTL are increased when performing simultaneous mapping, *trans*-eQTL benefit more from the increase in power. See main text for exact definitions of the various eQTL types.

leading to a somewhat noisy expression signal. Depending on the cell type, $14-23\%$ of these cell type-specific eQTL act in *cis*.

In contrast to the large number of static and conditional eQTL, we detected very few dynamic eQTL. At an FDR of 0.1 there were six eQTL driving gene expression changes during the transition from progenitor to erythroid cells and 66 eQTL for the transition from progenitor to myeloid cells. Two of the eQTL in these two groups are identical, i.e the same loci (both in *cis*) affect the same target genes during both, the P-E and the P-M transition. These targets are *Gadd45gip1* (see Section 4.3.5 and Figure 4.12D) and *Lrrc51*. We were not able to find any dynamic eQTL in the transition from stem to progenitor cells. Dynamic eQTL comprise a much larger fraction of *cis*-eQTL compared to simultaneous eQTL (44%, Figure 4.5). This is not surprising considering the fact that dynamic eQTL depend on gene expression measurements in two cell types at a time. They are thus more vulnerable to noise, but at the same time they have to be inferred from only one fourth of the samples available for the simultaneous mapping. Hence, we might only catch the strongest effects here, which are often found in *cis* (Petretto et al., 2006).

From the above considerations, it also becomes clear that dynamic eQTL might overlap with conditional and static eQTL. These overlaps are shown in Figure 4.6 after summarizing results from different subsets of cell types or transitions within each eQTL class. To facilitate comparison of conditional eQTL obtained with different mapping approaches (Section 4.3.3), eQTL that are detected in exactly one cell type are shown as a subgroup of conditional eQTL. By definition, there is no overlap between conditional and static eQTL. Intriguingly, none of the 70 dynamic eQTL are static, while 45 were found to be conditional. Finally, 25 loci that influence the dynamics of gene expression during the transition from one cell type to another could not have been detected by the simultaneous mapping. This fraction of 36% eQTL that are exclusively found in the dynamic mapping is in line with the findings of dynamic (gene-environment interaction) eQTL in two growth conditions in yeast (Smith and Kruglyak, 2008), where 38% of the dynamic eQTL did not meet the genome-wide significance level in any of the tested conditions.

4.3.3 Comparison between separate and simultaneous eQTL mapping

Comparative eQTL studies have so far mostly mapped eQTL separately in each cell type, subsequently classifying eQTL as 'static' if they are significant in all mappings, otherwise as 'cell type-specific' (Table 4.1). This approach leads to a situation very different from our simultaneous mapping: in separate mappings an eQTL has to be significant independently in each cell type in order to be classified as static. In other words, large power is needed to detect static eQTL. As opposed to that, in our approach the eQTL has to be significantly

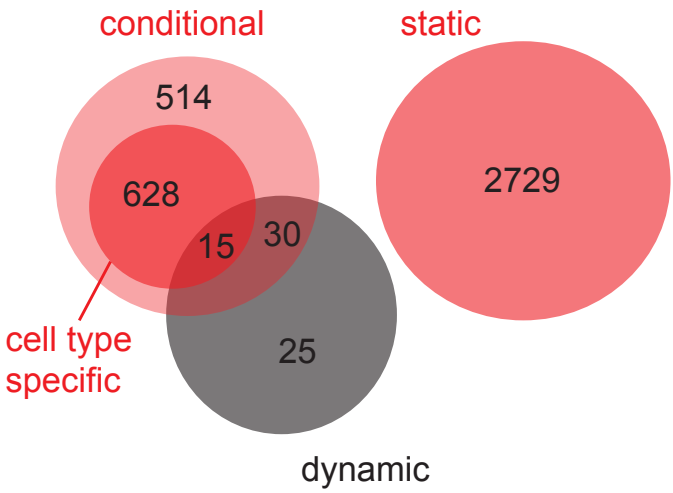


Figure 4.6: Venn diagram for the overlap between static, conditional and dynamic eQTL. Static and conditional eQTL were obtained from the simultaneous eQTL mapping (red circles). eQTL that are detected in exactly one cell type are shown as a subgroup of conditional eQTL (dark red circle). The dynamic eQTL were derived from mapping expression differences between pairs of cell types (black circles). The results are summarized over the three cell type transitions that were analyzed (S-P, P-E, P-M).

dependent on the cell type in order to be classified as conditional. Therefore, simultaneous mapping is more conservative with respect to calling conditional eQTL.

Consequently, eQTL obtained with these two mapping strategies overlap only partially (Figure 4.7), which is mostly owed to the fact that simultaneous eQTL mapping detects many more significant eQTL, the largest fraction of which are static. The increased statistical power of simultaneous mapping is expected due to the increased sample size for mapping static eQTL compared to the separate mappings. This superiority is especially pronounced for *trans*-eQTL (right hand side of Figure 4.5). Figure 4.8 confirms this notion: with increasing sample size, simultaneous mapping calls more eQTL in total, and a substantially larger fraction of static compared to conditional eQTL. Interestingly, the proportion of *cis*-eQTL decreases with increasing sample size, suggesting smaller effect sizes for *trans*-eQTL.

4.3.4 Static eQTL mapping based on mean expression

In the same way as dynamic eQTL are mapped using expression differences as the quantitative trait, the mean of mRNA expression levels can be considered as a trait for static

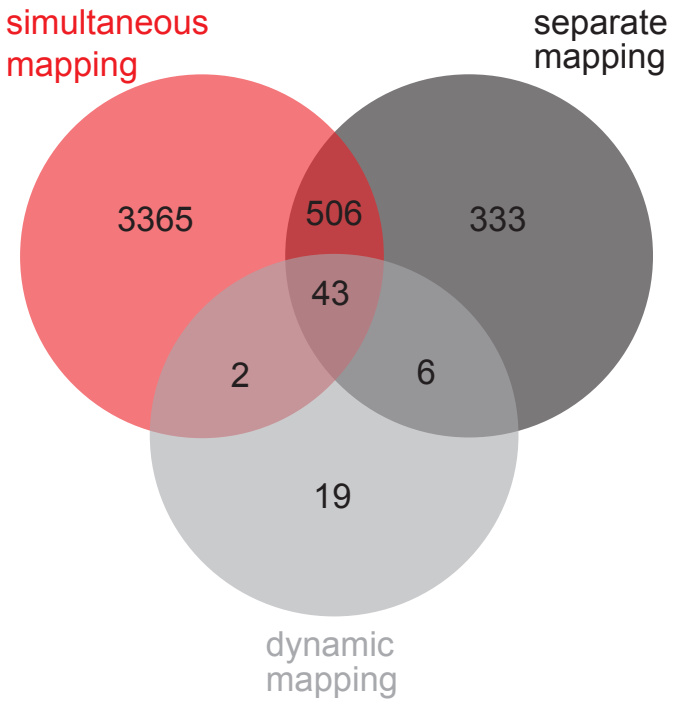
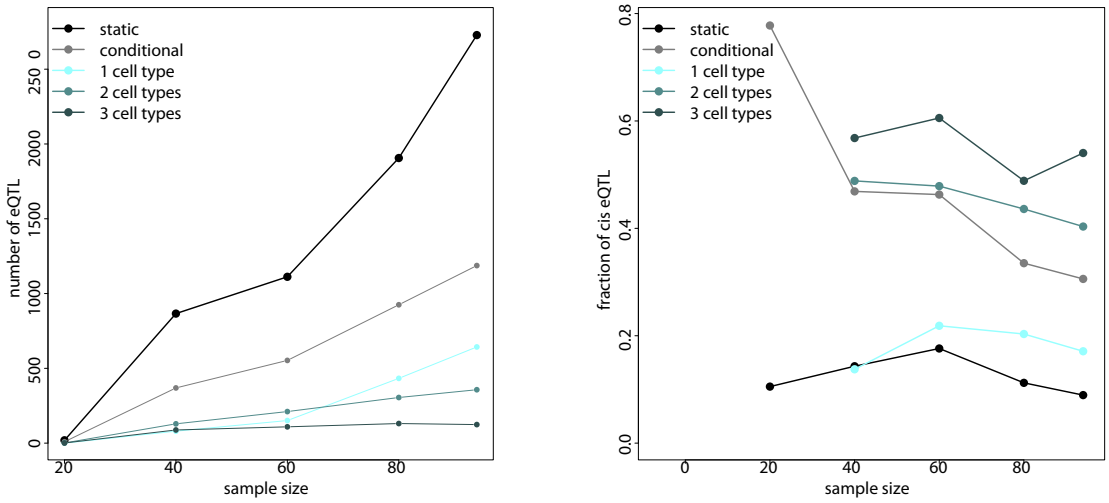


Figure 4.7: Comparison of different strategies for finding eQTL. We compared the outcomes of three eQTL mapping approaches that are eligible to all or a subset of the eQTL classes. The Venn diagram shows the overlap between all the eQTL that were called significant in any of the mappings we used the method for. In particular, simultaneous eQTL are all eQTL with an FDR < 0.1 in the simultaneous mapping regardless of the ANOVA result. Dynamic eQTL had to be significant in at least one of the three cell type transitions (S-P, P-E, P-M) while cell type-specific eQTL were required to have an FDR of 0.1 in at least one of the four cell types.

eQTL mapping. In order to avoid mean expression traits to be dominated by one cell type in which the transcript abundance is on a higher level than in the remaining types, we centered mRNA expressions of every gene with respect to the mean across strains in each cell type before calculating the average expression per gene and strain across types (Section 4.2.5).

mRNA expression is measured with varying accuracy on different arrays, i.e. in each strain and cell type. Since the quality of each measurement is usually returned as a quality score from the microarray image analysis software, it can easily be taken into account when calculating the average mRNA levels across conditions. We used the negative $\log_{10}$ transformed p-values of the quality scores from the data of Gerrits et al. (2009) as weights for the calculation of mean gene expressions (Section 4.2.5). Alternative measures of uncertainty, which can be transformed into weights, can be thought of, e.g. the variance of expression lev-

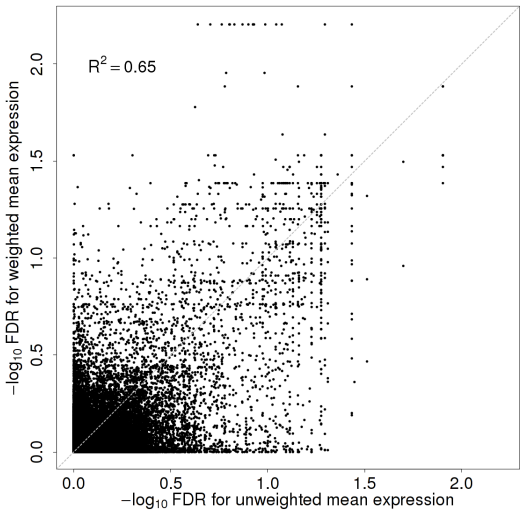


A Number of eQTL **B** Fraction of *cis*-eQTL

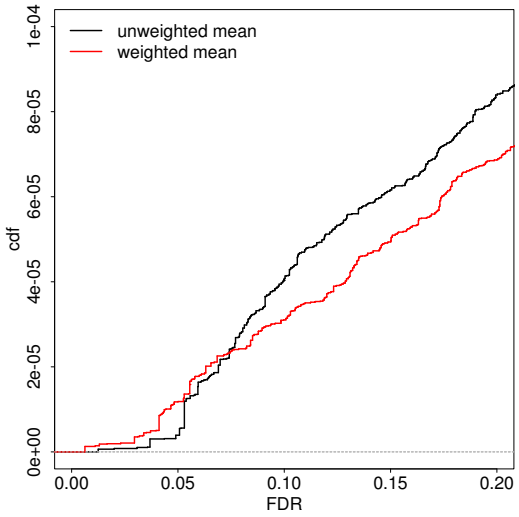
Figure 4.8: Number of eQTL and proportion of *cis*-eQTL as a function of sample size. We subsampled different numbers of strains in the simultaneous mapping (keeping ratios between cell types constant) and repeated the eQTL mapping. Panel A shows the number of eQTL in different classes as a function of sample size, while panel B shows the fraction of *cis*-eQTL among these. In order to detect any cell type-specific eQTL a minimum sample size larger than 20 is required. The proportion of *cis*-eQTL decreases with increasing sample size and is smallest for static eQTL, suggesting larger effect sizes for *cis*-eQTL compared to *trans*-eQTL.

els across technical replicates. We found that the unweighted and weighted approaches did not give identical results, but were well correlated (Figure 4.9A). Using the weighted means resulted in slightly lower FDR values for the most significant static eQTL (Figure 4.9B). Therefore, only results obtained from weighted mean expression mapping will be shown in the remainder of this section.

Because they are based on different expression traits, static eQTL obtained from simultaneous, separate and mean expression mapping did not coincide completely. As described in Section 4.3.3, we find many more static eQTL with the simultaneous mapping than when intersecting results from separate mappings in each cell type (Figure 4.10). The number of eQTL obtained from mapping mean expressions lies in between. We find $2,729$ static eQTL in the simultaneous, 18 in the separate and 241 in the mean expression mapping. The fractions of *cis*-eQTL are $9\%, 89\%$ and 40%, respectively. Eight of the static eQTL that are found in all four separate mappings are also detected as static in the simultaneous mapping (the remaining ten belong to the group of conditional eQTL with four cell type interactions). However, only 35% of the mean expression eQTL overlap with static simultaneous eQTL,



A Correlation between unweighted and weighted mean static eQTL



B FDR of unweighted and weighted mean static eQTL

Figure 4.9: Comparison of static eQTL derived from mapping unweighted and weighted mean expressions. A Although the negative $\log_{10}$ transformed FDRs of static eQTL linked to unweighted (on the x-axis) and weighted (on the y-axis) mean expressions are not identical, their correlation is quite high ($R^2 = 0.65$). eQTL that are significant when mapping unweighted mean expressions also achieve a low FDR in the weighted mapping. The weighted mapping also finds some additional eQTL, which are not detected with the ordinary mean. **B** Zoom into the cumulative distribution function of the FDR of static eQTL obtained from mapping unweighted (black) and weighted (red) mean expressions. Using the weighted mean results in lower FDRs for the most significant eQTL.

while six of the 18 eQTL from the separate mappings overlap with mean expression eQTL (Figure 4.10).

The differences in the numbers of eQTL might be explained by the power differences between the studies. As explored in Section 4.3.3, because of the larger number of informative samples, the power of the simultaneous mapping is much higher than that of both, the separate and the mean expression eQTL mappings. In contrast, the advantage of the mean expression approach compared to the other two methods is the reduction in measurement noise that results from the averaging of mRNA measurements across cell types. This might be one reason why we are able to find more static eQTL with this method than with separate mappings. Another reason is of course the larger number of tests that have to be significant in the separate mappings in order to define an eQTL as static (four compared to one in the other approaches).

Apart from the differences in power, the limited accordance among the three approaches

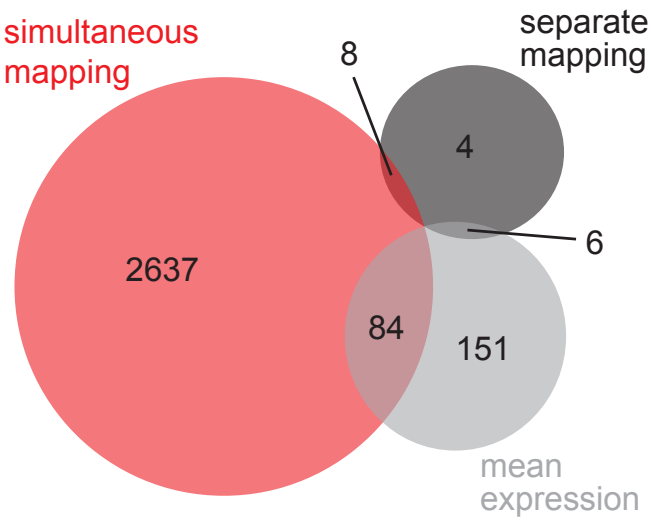


Figure 4.10: Overlap between static eQTL mapping methods. Eight eQTL that are jointly significant in all four separate mappings (black circle) are also detected as static eQTL in the simultaneous mapping (red circle). More than 60% of the eQTL obtained from using mean expression across cell types as a trait (grey circle) are not found with the other two approaches.

might be due to the fact that they represent slightly different definitions of static eQTL. While in the separate mappings a static eQTL is very stringently defined as an eQTL reaching a given significance threshold in all four cell types, it has to fulfill less severe requirements in the other approaches. Especially the idea of an impact of genetic variation on mean expression levels does not necessarily entail a significant eQTL in all cell types separately. Rather, a closer look at some examples of mean expression eQTL reveals that they might result from very small effects in the single cell types (effects that cannot be caught in each type separately) or from a very strong cell type-specific eQTL, even after centering the expressions per cell type (Figure 4.11). The latter case does not correspond to our definition of a static eQTL and can thus be regarded as a false positive finding. The interpretation of static eQTL in the simultaneous mapping, although less stringent in terms of significance thresholds, is similar to the one in the separate mappings. Here, we detect eQTL that control the expression of a gene in the same way in all four cell types.

In conclusion, both the simultaneous as well as the separate mappings in each cell type provide a means to detect static eQTL. Simultaneous mapping clearly outperforms separate mappings due to the increased power. In contrast, mapping mean expression levels of a gene across cell types seems to be less appropriate to detect static eQTL in the way we defined them in Section 4.2.1.

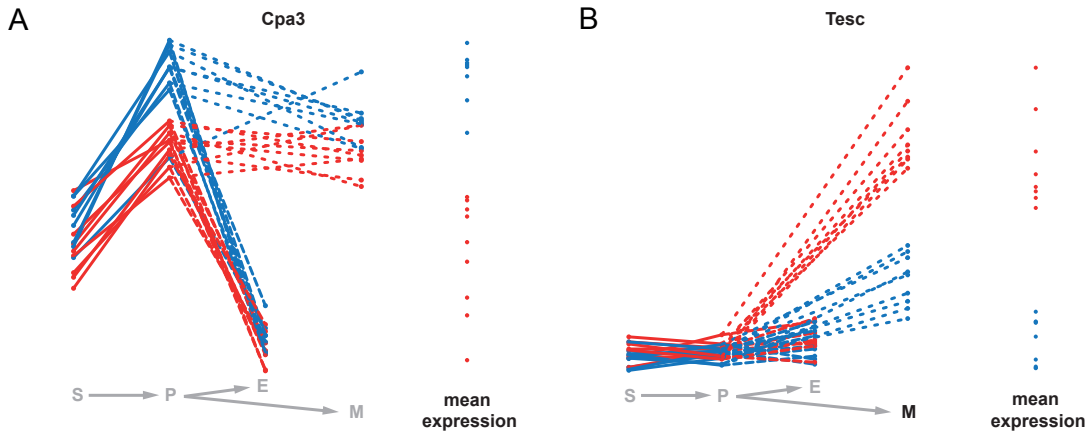


Figure 4.11: Examples of mean expression eQTL. Schematic mRNA expression profiles of two genes with a significant mean expression eQTL (FDR < 0.1) over the four hematopoietic cell types (S - stem cells, P - myeloid progenitor cells, E - erythroid cells, M - myeloid cells). The rightmost points in each panel show the mean expression across cell types. The colors represent the genotype at the eQTL marker (blue - B allele, red - D allele). Significant conditional eQTL are indicated by the black color of the respective cell type letter.
A *Cpa3* has a weak effect eQTL, which neither reached the significance level in any of the separate mappings nor in the simultaneous mapping. However, the reduction of noise due to the averaging of mRNA levels across cell types enabled us to detect the eQTL using weighted average gene expression as a quantitative trait. **B** Although *Tesc* is controlled by a conditional eQTL in the myeloid cells only, the strong effect of the eQTL genotype on gene expression levels in this cell type propagates itself to the mean expression. Therefore, we find the same eQTL in the mean expression mapping, but would not call this a static eQTL.

4.3.5 Examples for the different eQTL classes

Static eQTL affect a gene's expression in all cell types. An example of such a static eQTL is an eQTL impacting on the expression of Peroxiredoxin-2 (*Prdx2*) (Figure 4.12A), a gene involved in the response to and protection of erythrocytes against oxidative stress (Lee et al., 2003). It is one of the most abundant proteins in erythrocytes (Johnson et al., 2010). Moreover, *Prdx2* plays a role in T cell differentiation and might inhibit immune cell responsiveness (Moon et al., 2004, 2006). The importance of *Prdx2* in erythroid cells is well reflected by the fact that it is more highly expressed in erythroid cells than in any of the other cell types. Nevertheless, regulation of *Prdx2* is important in every hematopoietic cell to prevent damage from oxidative stress, which would severely impact hematopoietic cell homeostasis (Ghaffari, 2008). Since *Prdx2* is encoded at the same locus as the eQTL itself, the expression differences between the eQTL alleles are probably due to a mutation in the

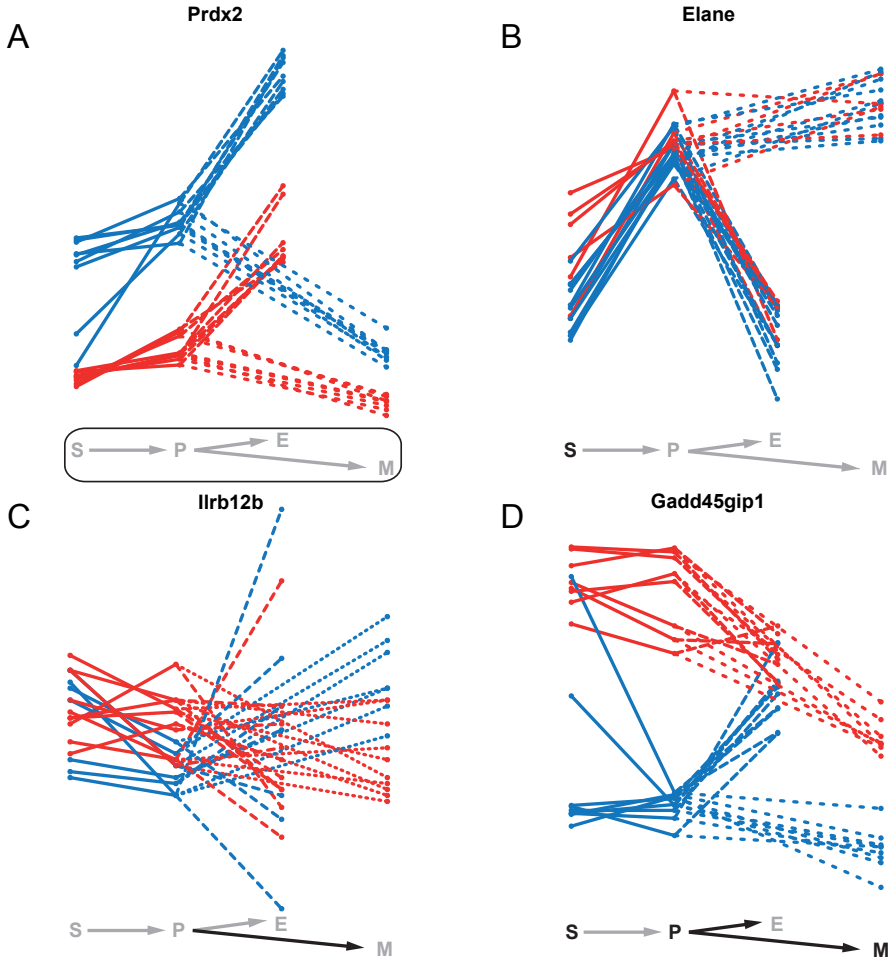


Figure 4.12: Examples of static, conditional and dynamic eQTL. mRNA expression profiles of four exemplary genes over the four hematopoietic cell types (S - stem cells, P - myeloid progenitor cells, E - erythroid cells, M - myeloid cells). The colors represent the genotype at the eQTL marker (blue - B allele, red - D allele). Significant static eQTL are shown by a rectangle around the differentiation scheme, significant conditional and dynamic eQTL by the black color of the respective cell type letter or transition arrow.
A, *Prdx2* is affected by a static eQTL in all four cell types. **B**, *Elane* is influenced by a conditional eQTL in the stem cells. **C**, the transition of *Il12rb2* expression from progenitor to myeloid cells is driven by a dynamic eQTL. The expression of *Il12rb2* increases in samples carrying the B allele at the eQTL locus, while it remains constant in samples carrying the D allele. **D**, the expression of *Gadd45gip1* is conditionally affected in three of the four cell types (S, P and M) by an eQTL which at the same time also influences the gene's expression changes during the differentiation from progenitors to the erythroid and myeloid lineages.

gene itself or in a *cis*-regulatory region.

Figure 4.12B shows *Elane* as an example of a gene being target of a conditional eQTL. *Elane*'s expression is strongly correlated with the alleles at the eQTL in hematopoietic stem cells, but not the other cell types. This neutrophil elastase is released upon activation of neutrophils and important for the immune response to degenerative and inflammatory diseases (Weinrauch et al., 2002). In line with that, its expression profile shows that it is highly expressed in myeloid cells, which are neutrophil precursors, while it is markedly down-regulated in erythroid cells. Apart from its direct immune function, *Elane* plays a crucial role in the mobilization of stem cells in the bone marrow (Lévesque et al., 2001).

Il12rb2 is an example of a gene being affected by a dynamic eQTL. The gene encodes for a transmembrane protein constituting one subunit of the Interleukin 12 receptor complex. It is known that the gene is upregulated in T helper cells and that Interleukin 12 signaling plays a role in T helper cell activation upon immune response to pathogens (Trinchieri, 2003). Apart from that, several studies have revealed that together with other colony-stimulating factors Interleukin 12 is also involved in myelo- as well as erythropoiesis (Jacobsen et al., 1993; Dybedal et al., 1995). We find a dynamic eQTL for *Il12rb2* in the differentiation from progenitor to myeloid cells, which is characterized by almost constant expression levels for strains carrying the D allele at the eQTL locus while mRNA levels increase for individuals carrying the B allele. The expression profiles of *Il12rb2* in progenitor and myeloid cells indicate that the eQTL might actually be conditional in both cell types with very small and opposite effects. The example therefore demonstrates that dynamic eQTL mapping might in special cases (such as switching allelic effects) have increased power compared to conditional mappings.

Intuitively, one expects that a significant allele-dependent expression change from one to another cell type (i.e. a dynamic eQTL) will coincide with significant, allele-dependent expression in at least one of the two cell types involved in the transition (i.e. a conditional eQTL). We often observed such co-incidence (Figure 4.6) and the cell cycle inhibitor *Gadd45gip1* (Chung et al., 2003) is a particularly interesting example (Figure 4.12D). *Gadd45gip1* is one of only two genes for which we found a dynamic eQTL affecting the transition to both, erythroid and myeloid cells. The protein encoded by this gene physically interacts with *Gadd45b*, which is involved in cell growth arrest during myeloid cell differentiation (Chung et al., 2003; Abdollahi et al., 1991). *Gadd45gip1* might support this function and arrest cell cycle in a particular phase in myeloid precursor cells, a prerequisite for differentiation (Yen and Albright, 1984). It is up-regulated in stem and progenitor cells in samples carrying the D allele at the eQTL locus (Figure 4.12D). The eQTL is in *cis*, suggesting that a mutation in the *Gadd45gip1* gene itself or in its promoter region leads to decreased expression of the gene in individuals carrying the B allele. Accordingly, down-regulation of

Gadd45gip1 in the transition to myeloid cells only occurs in samples carrying the D allele. This leads to a dynamic eQTL from progenitor to myeloid cells. Interestingly, individuals having high *Gadd45gip1* levels in progenitor cells show a down-regulation of its expression during the transition to erythroid cells, while the gene is up-regulated in individuals with low *Gadd45gip1* levels in progenitor cells. This leads to an expression equilibration in erythroid cells. Thus, (i) compensatory feedback mechanisms can 'revert' the effect of an eQTL and (ii) there seems to be a need to tightly control *Gadd45gip1* expression in erythroid cells.

4.3.6 Cell type-specific eQTL transbands

The visualization of all cell type-specific and static eQTL in an eQTL map (Figure 4.13) reveals some cell type-specific eQTL transbands, i.e. eQTL being associated with a large number of target genes in a given condition. An example of such a hotspot is a transband on chromosome 19 ($52.3 - 55.2$ Mb) affecting 31 stem cell-specific and 59 static target genes. Even though only one third of the eQTL in this locus meet the significance threshold of a stem cell-specific eQTL, there is a clear tendency towards stem cell specificity for most of them (Figure 4.14A). The eQTL locus contains the gene *Shoc2* for which we also find a *cis*-eQTL. We have previously shown that *trans* effects are often caused by genes being themselves effected through a *cis* effect (Loguercio et al., 2010), which makes *Shoc2* a putative causal gene in the region. The protein encoded by this gene is a scaffold for a *Ras*/*Raf* interaction (Sieburth et al., 1998). The *Ras* pathway is important for hematopoietic differentiation processes and frequently activated in hematopoietic malignancies (Reuter et al., 2000). However, we did not find any direct links between *Shoc2* and its putative target genes.

We found a second cell type-specific transband on chromosome 2 ($168.3 - 169.7$ Mb), whose eQTL - target gene pairs show a tendency to be myeloid, and to a lesser extent also stem cell-specific (Figure 4.14B). One possible regulator gene in this locus is *Nfatc2* (nuclear factor of activated T cells), which is gradually down-regulated during some intermediate stages of the differentiation of myeloid progenitors into megakaryocytes and neutrophils (Kiani et al., 2007). We find evidence of functional interaction between *Nfatc2* and some of its target genes in the protein interaction network STRING (Szklarczyk et al., 2011). Many of these genes (e.g. *Ccdc99*, *Cdk2*, *Cdca8*, *Birc5*) are involved in cell cycle control. Indeed, it is known that *Nfatc2* negatively regulates the expression of *Cdk4*, which controls the entry and progression of a cell in the cell cycle (Baksh et al., 2002). In line with that, *Cdk4* links *Nfatc2* and its target genes in the STRING network. Although it has been shown that *Nfatc2* is not required to block cell cycle entry, it is likely that it prevents HSC from differentiation into neutrophils and megakaryocytes via an effect on their proliferation (Kiani et al., 2007; Kiani,

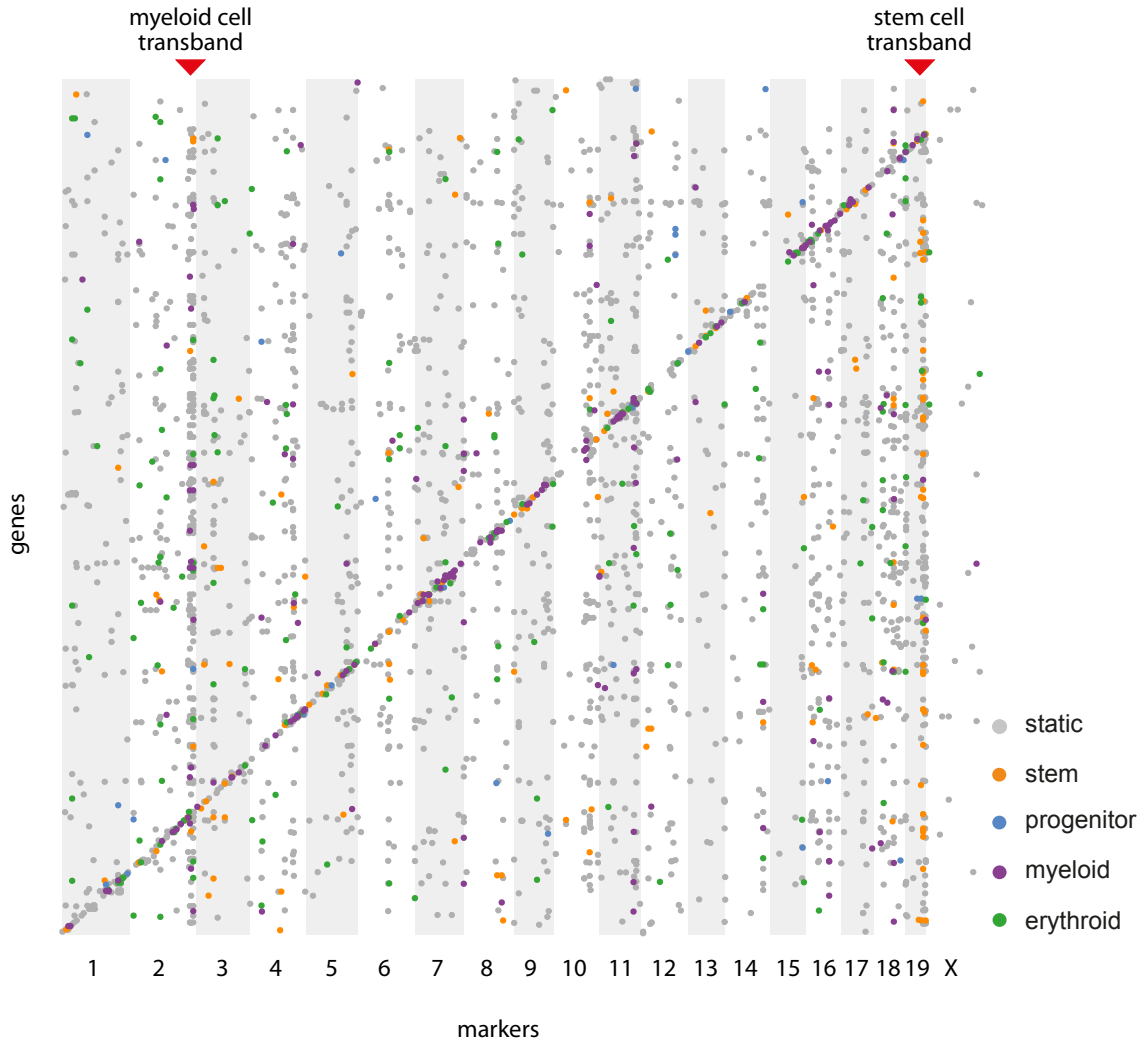
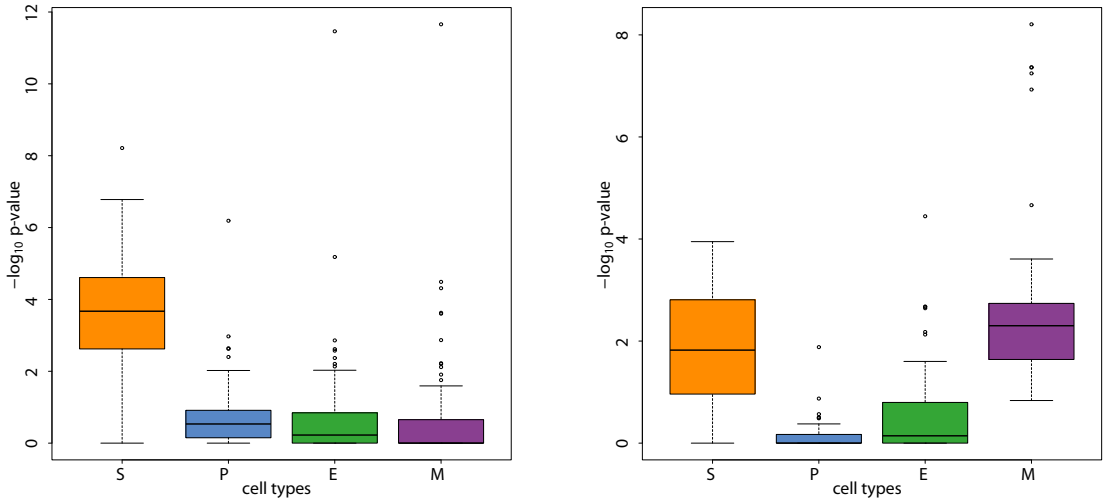


Figure 4.13: Simultaneous eQTL map. Each dot represents an eQTL - target gene pair, where physical marker positions are shown on the x-axis, gene positions on the y-axis. Significant static eQTL (FDR < 0.1) are shown in grey, cell type-specific eQTL (Bonferroni corrected p-value < 0.005 in exactly one cell type) are shown in the color scheme of Figure 4.3. Red triangles indicate two cell type-specific transbands.

2004). The importance of *Nfatc2* for both the HSC and the myeloid cells is reflected by the lower cell type specificity p-values of its targets in both types (Figure 4.14B) and corresponds well to *Nfatc2* expression levels that have been found to be high at the beginning of myeloid differentiation, go down during differentiation and finally increase again (Kiani et al., 2007).



A Stem cell-specific transband (chr. 19) **B** Myeloid cell-specific transband (chr. 2)

Figure 4.14: Distribution of contrast test p-values for cell type-specific eQTL hotspots. eQTL hotspots might affect cell type-specific processes. This is shown for two transbands on chromosomes 19 (**A**) and 2 (**B**), respectively. Colors indicate hematopoietic cell types as in Figure 4.3. Overall, the stem (in **A**) and myeloid cell (in **B**) contrast test p-values are much smaller than those for the other three cell types, indicating that the marker locus is associated with the expression of genes involved in processes specific for the given cell type (p-values are shown in $-\log_{10}$ scale on the y-axis).

4.3.7 Dynamic eQTL affect cell type-specific functions

Cell type-independent (i.e. static) eQTL might affect genes that are less specific for the processes being studied than conditional and dynamic eQTL. In order to test this notion we assessed the enrichment of functional categories among genes causing eQTL and among genes being affected by eQTL using gene annotations obtained from GO Biological Process (Ashburner et al., 2000). Such GO enrichment analysis is non-trivial for genetic regions causing eQTL, because they typically contain multiple genes and it is usually unknown which of them is the true causal gene (Rockman and Kruglyak, 2006). Therefore, we decided to annotate each region with the GO terms of all associated genes (Section 4.2.7). This rigorous solution has the following advantages. If there is a true enrichment of GO terms among causal genes, this will be 'diluted' by our approach. Thus, the procedure will lead to a conservative estimation of functional enrichment. At the same time, this strategy also avoids a bias in GO enrichment due to local clusters of functionally related genes. The enrichment testing was conducted with the R package `topGO` (Alexa and Rahnenführer, 2010), which corrects for the nested structure of GO. The top 10 significantly enriched GO terms for each eQTL

mapping can be found in Supplementary Tables C.1 to C.12.

Figure 4.15 shows exemplary results of the enrichment distinguishing cell type-specific, dynamic and static eQTL. Static eQTL are enriched for very generic functional categories such as translation, transcription and cell cycle regulation. As opposed to that, conditional eQTL are enriched for hematopoiesis-related functions: For example, stem cell eQTL targets are enriched for the term "cell migration involved in sprouting angiogenesis", in which HSC play an important role (Takakura et al., 2000). Myeloid progenitor cell eQTL are enriched for the generic immune term "myeloid leukocyte mediated immunity", while conditional eQTL in myeloid cells are enriched for very specific immune response terms like "defense response to Gram-negative bacterium". We found several GO terms related to MAP kinases enriched among eQTL in erythroid and myeloid cells. This family of serine/threonine kinases plays a crucial role in maintenance and differentiation of HSC, especially during erythropoiesis (Geest and Coffer, 2009).

Dynamic progenitor-myeloid eQTL are specifically enriched for categories related to T cell selection. This could be an indirect effect related to the role of macrophages and dendritic cells, which belong to the myeloid lineage, in adaptive immunity. These cells are involved in presenting antigens bound to the major histocompatibility complex (MHC) to naive T cells in order to activate or suppress these cells (Alberts, 2002). Accordingly, we find MHC coding genes among the dynamic eQTL targets. Since we found only six significant dynamic eQTL for the differentiation towards erythroid cells, the corresponding enriched GO categories contained very few genes or loci that are involved in significant eQTL. Therefore, we did not consider the results of the GO enrichment for this mapping.

4.4 Discussion of dynamic eQTL mapping results

The difference between static and non-static eQTL was very striking in our analyis. Due to the increased statistical power resulting from the simultaneous mapping we could identify substantially more static than non-static eQTL. Further, static and non-static eQTL differed substantially with respect to the functions of the involved genes, regarding both regulators (i.e. loci) and target genes. Whereas static eQTL involve mostly genes with generic, unspecific functions, non-static eQTL affect more cell type-specific pathways.

We found relatively few dynamic eQTL, ranging from zero (stem to progenitor cells) to 60 (progenitor to myeloid cells) per cell type transition. This is not very surprising given the fact that expression differences are prone to increased variance since they "inherit" the independent errors of expression experiments in two different conditions (Ideker and Krogan, 2012). We would also expect a large overlap between conditional and dynamic eQTL. If there is a dependency between gene expression levels and genotype in one but not another

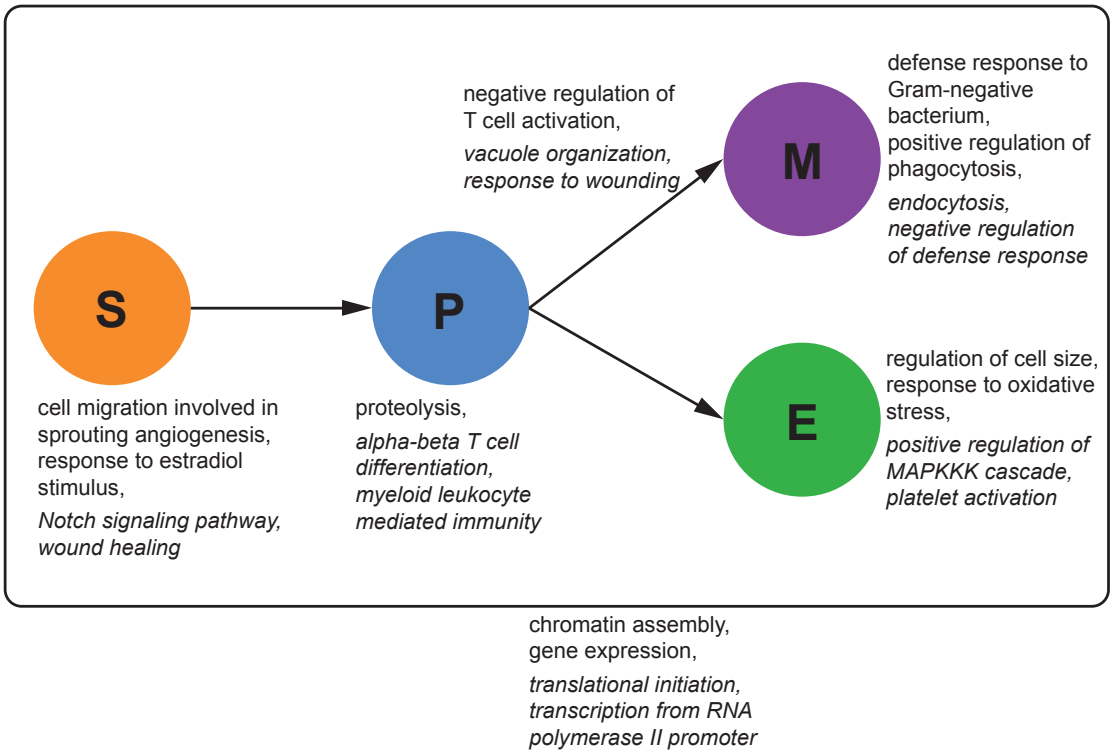


Figure 4.15: GO enrichment for eQTL classes. We tested for the enrichment of GO categories among eQTL loci and target genes in the different eQTL classes, separately for different cell types and transitions. Examples of enriched functional categories for cell type-specific and dynamic eQTL are shown next to the corresponding cell types or cell type transitions. Important GO categories that were enriched in static eQTL and their targets are shown outside the box. Terms that are significantly enriched ($p < 0.01$) among eQTL loci are shown in italic, GO categories enriched among eQTL targets in regular font. See Supplementary Tables C.1 to C.12 for a list of the top significant GO terms of each mapping.

cell type, then the magnitude of expression change between these cell types (i.e. the slope) should be genotype-dependent as well. However, we only find 45 eQTL as belonging to both, the conditional and the dynamic class, while 1,142 and 25 eQTL are exclusively conditional and dynamic, respectively.

One reason for this observation is the reduced power of the dynamic mapping leading to a failure to replicate conditional eQTL. Intriguingly, we also detect dynamic eQTL that we do not find among the conditional eQTL. Thus, there are modes of expression variation that are detectable with higher power when mapping expression differences instead of absolute expression levels. For example, we find eQTL with swapping effects on transcript levels

(such as *Il12rb2*, Figure 4.12C) among 10 out of the 25 eQTL-target gene pairs that are unique in the dynamic class. This emphasizes the need to include different expression traits (like expression differences) into a comprehensive eQTL analysis in order to detect a wide spectrum of eQTL.

Another notable feature of dynamic eQTL mapping is its ability to mitigate systematic measurement errors affecting all cell types in a similar way. In this respect, a score for relative expression change can still be meaningful even though the absolute expression levels were not (Ideker and Krogan, 2012).

The approach we proposed for mapping different classes of eQTL is only one of a palette of possible strategies. Since the focus of the present work was on the introduction of a functional eQTL classification, in particular the discussion of each classes' characteristics and its implications on biological interpretation of eQTL results, we did not comprehensively compare different approaches for eQTL class mapping. However, we still tested several variants, in particular the aggregation of static and conditional eQTL from separate mappings in every condition, which is the most widely used approach for comparative eQTL studies in the literature (see references in Table 4.1). Importantly, single cell type mappings lack power compared to the simultaneous approach, where samples from all conditions can be exploited in one comprehensive mapping. This has considerable consequences especially for the detection of static eQTL, which have to be significant in all separate mappings. Hence, the number of static eQTL that we detected with this strategy is far smaller than the number we find with the simultaneous eQTL mapping (18 compared to $2,782$).

The strategy we followed for mapping dynamic eQTL has an obvious counterpart for static eQTL, namely the mapping of mean expression levels over all conditions. However, when applying this approach to the four hematopoietic cell types, we noticed that a large fraction of the resulting static eQTL were in fact conditional eQTL in one or several types. The erroneous classification resulted from the fact that a strong cell type-specific effect can bias mean expression levels towards a significant genotype-dependent expression pattern. Thus, this approach is prone to detect false positive static eQTL and in our opinion is not well suited to classify static eQTL.

In principle, the second step of the simultaneous eQTL mapping, the distinction between conditional and static eQTL, could be directly resolved in the primary eQTL mapping step. The Random Forests framework allows to extract epistatic interactions between predictors directly from the trees (Yoshida and Koike, 2011; Bureau et al., 2005; Dutkowski and Ideker, 2011; Sakoparnig et al., 2012). However, this requires a large enough sample size in order to grow deep trees where different combinations of variables will be used for splitting in the same branch. When trying this line of action on the hematopoiesis data, it became clear that the small sample size (22 to 24 samples per cell type) is prohibitive for this step,

leading to rather unstable results. Hence, we used the remedy of applying an ANOVA to filter the conditional eQTL out of the set of simultaneous eQTL. We believe that with the improvements made on costs and quality of large sequencing studies and the further increase in computing power this approach will become feasible very soon.

The fact that we find 30% of all simultaneous eQTL to be conditional for one or several cell types emphasizes the condition specificity of many regulatory relationships, even if the conditions under study are very related. This has also been pointed out by other groups (see references in Table 4.1), who reported proportions of 5% to 94% condition or tissue-specific eQTL. Moreover, Powell et al. (2011) showed that the genetic correlation across the genome between whole blood samples and lymphoblastoid cell lines is close to zero for most genes, implying that only very few genes are regulated by the same causal locus in different tissues. Apart from that, we find that the number of conditional eQTL differs between cell types, partly due to differences in sample size and tissue impurity, but maybe also due to functional differences. These findings call for due caution when drawing conclusions about regulatory mechanisms in one condition based on results from another condition. A typical example for such a propagation of results would be the use of molecular mechanisms derived from eQTL studies in blood samples to explain disease mechanisms in other tissues like the brain or the nervous system. The use of eQTL results for the elucidation of disease etiology is further complicated by the fact that the onset of complex diseases typically involves pathways in several tissues.

On the other hand, we also have to emphasize that the ratio of conditional to static eQTL depends on the power of the given study. Simultaneous eQTL mapping has high power to detect static eQTL while the power for calling conditional eQTL is decreased. Hence, it provides a conservative estimate for the fraction of conditional eQTL. At the same time, the tissue specificities reported so far are probably an over-estimation of the true tissue specificity, owing to the failure to reproduce static eQTL in different conditions, especially if they only have a small effect on gene expression. This phenomenon is also known as the 'winner's curse' (Dimas et al., 2009; Lohmueller et al., 2003). Hence, we suspect that the number of eQTL that can be replicated in several tissues will increase with the growing amount of expression and deep-sequencing data that is becoming available in a wide range of mammalian tissues.

Chapter 5

Summary and discussion

Virtually every molecular and physiological trait emerges from the cooperation of a set of genetic and environmental factors. Deciphering the interactions between these factors is key to the understanding of organism development and function and, very importantly, disease etiology. The projects presented in this thesis both contribute methodologies for unraveling genetic interactions underlying complex traits.

First, I proposed a test for the detection of epistatic interactions with a severe impact on fitness phenotypes in parent-offspring genotype data ("Imbalanced Allele Pair frequencies", ImAP). I applied the test to an outbred population of mice with known family structure and found more such allele incompatibilities than expected by chance. We validated a large number of these interactions on external data and showed that epistatic loci are enriched for genes functioning in development, and hence being essential during the very early stages of life. The method might therefore present a major step forward in solving a problem that has not yet been tackled: detecting allele incompatibilities on the genome-scale in mammalian species. The approach is readily applicable to any genotype data set, given information about the parents of the samples, and does not require additional phenotyping. Hence, we expect the test to prove its value on the growing amount of sequencing data on human families by discovering interactions causing severe developmental phenotypes.

Second, I presented a systematic classification of genetic variation affecting gene expression during dynamic cellular processes. I defined and compared static, conditional and dynamic genetic impact on transcript levels and proposed a new strategy for mapping expression quantitative trait loci (eQTL) across conditions, simultaneously exploiting all available mRNA expression information. We applied our framework on a data set of genome-wide gene expression profiles of an inbred mouse population across a range of hematopoietic differentiation stages, thereby detecting loci inducing the immense modifications of the gene expression landscape underlying blood cell development. We were able to show that different classes of eQTL as well as different mapping approaches result in different sets of eQTL, albeit being based on the same set of expression data. The proposed framework can be applied to any eQTL study, in which expression is measured across several time points or conditions, e.g.

during development or before and after treatment.

Our intention was to provide a guideline to support data analysts in making the right decision about the appropriate mapping strategy depending on the kind of regulatory relationships she/he is interested in. Moreover, we wanted to increase the awareness about the meaning and limitations of the analysis of different variants of the expression trait. Our work also has implications for the interpretation of QTL and genome-wide association studies using information from eQTL studies: Filling the gap between genotype and disease phenotype with gene-regulatory network information should be done carefully and, if possible, only with results derived from eQTL studies in similar cell types or cellular conditions.

The two presented projects, although being seemingly quite different in their implementation, share a common biological question as well as the data they are based on. Both studies help in understanding the molecular mechanisms underlying complex traits and disease pathogenesis on the genome-level, and both rely on high-throughput data of natural genetic variation. Consequently, both studies also share some virtues and limitations, in particular on the level of data.

The use of natural genetic variation data has a number of advantages compared to the traditional approaches of genetic perturbation experiments, in which the influence of an induced mutation in only one or very few genetic loci on the expression levels of all other genes or some higher level traits is observed. The impact of natural genetic variations can be observed without the need to actively perturb the system, i.e. without any extra experiments that might be expensive or difficult to conduct, especially in more complex organisms, and which might induce side effects in the behavior of the system. Moreover, a naturally occurring genetic perturbation provides a much more realistic picture of the consequences of genetic mutations than a gene knock-out/-down study causing non-physiological over- or under-expression of genes. These data also reflect multifactorial perturbations underlying complex traits much better than an experimental perturbation of a few genes (Rockman, 2008). Not less important, large-scale genotype data contain a lot of 'hidden' replications of each allele, which allows to explore a large space of variations and their interactions (Rockman, 2008) and they are nowadays widely available for many organisms, tissues and conditions. Their amount and quality will even grow in the coming years. For example, next-generation sequencing data of mRNA samples at the same time provide a more dense measurement of genetic information as well as more precise transcript level measurements. Intriguingly, our group has found many more eQTL using mRNA sequencing data than what is normally obtained from SNP array data in similar conditions (Picotti et al., 2012).

Notwithstanding, this type of data also has its limitations. For example, the resolution and structure of naturally occurring perturbations depends on the population. While crosses

of inbred lines allow to detect large-effect mutations since they are often derived from crosses of genetically very divergent parental lines, they contain a limited number of recombinations. Moreover, the variation observed in these crosses might differ from polymorphisms segregating in an outbred population. On the other hand, variation in natural populations will only be observed if it is not linked to a locus that is subject to evolutionary selection. Hence, genetic variation is correlated with recombination rates and is less likely to contribute to the variation of a quantitative trait if it occurs in regions of low recombination (Rockman, 2008).

Since the resolution of genotype data obtained from mouse crosses is limited, the genetic loci found to be implicated in epistasis or linked to quantitative traits usually contain many genes. This makes the detection of the actual causal genes a difficult problem, which can often only be solved by integrating external information or complementary data sets, if available. Moreover, also the small sample size, especially of the hematopoietic data set, limited the power of our study to detect genetic loci driving the dynamics of gene expression. This problem might be alleviated by restricting the number of tested genes or genetic markers, e.g. by pre-filtering potentially interesting loci. On the longer run, this issue will hopefully be solved by the improvement of experimental methods and the reduction of their costs.

Apart from the genotype data, high-throughput gene expression data also have some drawbacks that need to be kept in mind. The power to detect eQTL depends on the accuracy of the mRNA expression measurements, which in turn depends on transcript abundance. In other words, there is more power to detect eQTL of highly transcribed genes compared to genes that are expressed only slightly above the threshold of detection (Rockman, 2008). Technical replicates could alleviate this issue by allowing to derive an error model for each transcript. In this thesis, quality scores of the microarray image analysis software were used to weight mRNA level measurements from different cell types in their contribution to the mean expression trait. A related problem is the impurity of the samples. Single cells or even tissues might not provide enough material to measure genome-wide mRNA levels. Therefore, experimenters are often forced to pool material across samples or tissues. For example, each single mRNA sample in the hematopoietic data of Gerrits et al. (2009) is in fact a pooled cell extract from three mice. Consequently, the transcript abundance represents the characteristics of the cell mixture and might hide inter-individual differences or slight nuances in expression patterns between tissues (Rockman, 2008).

Both projects also share the problems arising from the use of a permutation approach for p-value calculation. Although this approach is very powerful in providing a significance level for a test statistic in cases where its distribution is unknown or difficult to obtain analytically, it becomes prohibitive in large data sets due to its computational costs. Hence, the approaches presented in this work are either limited in the resolution of the resulting

p-values or they still require some assumptions on the test-score distributions. Two possible routes circumventing these issues are (i) the development of a more closed analytical p-value calculation (possibly together with modifications of the test statistics, for example the Random Forests (ImAP, Breiman, 2001) importance measure) or (ii) by incrementing the parallelization and computational resources for the calculation of the permutations.

In this work, we have validated some of our results using external data of the same type or additional, complementary information (e.g. functional gene annotation). The merits of data integration have been shown and discussed in the literature over the past decade (Carlborg and Haley, 2004; Suthram et al., 2008). The potential of this data integration approach can be further exploited in the continuation of both projects presented here. For example, it suggests itself to overlap the ImAP interactions with different kinds of genetic or protein-protein interactions networks. There exists a plethora of publicly available databases containing either manually curated or computationally inferred interaction networks, created from experimental or text mining evidence, orthology mapping etc. (Szklarczyk et al., 2011; Stark et al., 2011; Rhodes et al., 2005; McDermott et al., 2005). Since these databases provide complementary and only partly redundant interaction networks, we have proposed approaches to combine multiple interactions networks and provide a large-scale map of predicted physical and functional protein interactions (`http://www.print-db.org`, Elefsinioti et al., 2009, 2011).

We already tested high-scoring ImAP loci for enrichment of interactions from some of these databases (Szklarczyk et al., 2011; Stark et al., 2011), however with limited success. Possible reasons for the lack of accordance could be the different nature of the inferred interactions (e.g. genetic versus physical protein binding), the low coverage of known genetic interactions in mice together with possible differences between interaction networks among species and the problem of mapping high-scoring ImAP pairs to the true causal genes. Since the quality and coverage of these databases will increase rapidly in the coming years due to the advances of experimental as well as statistical methods, integration of the emerging data with the results from ImAP should be repeated and refined. This might allow to pinpoint causal genes, to explain the molecular mechanisms underlying the interactions as well as to relate them to phenotypes they might act upon. The latter point can also be achieved by combining our results with disease SNPs discovered in genome-wide association studies.

In the continuation of the data analysis described in Chapter 4, we have already begun an extensive data integration. The identification of genetic loci influencing transcript abundance is only one piece in the puzzle that needs to be solved in order to completely understand the genetic regulation of hematopoiesis. Other factors that need to be investigated are the activity of relevant transcription factors (TFs) and how this is regulated on the genetic level,

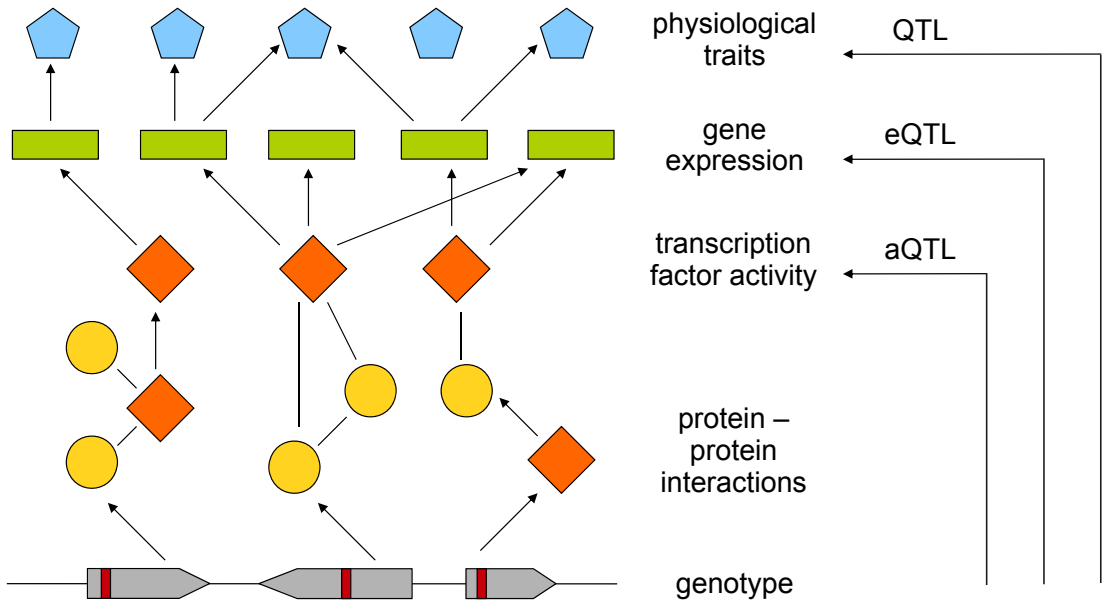


Figure 5.1: Overview of the different levels of genetic regulation of hematopoiesis. The activity and interaction of TFs (orange squares) and other regulatory proteins (yellow circles) is influenced by the genotype of the genetic loci (grey pentagons) where they are encoded. In turn, these proteins regulate the expression levels of genes (green rectangles) whose products finally change the physiological phenotype of the cell (blue pentagon) and consequently the cell state. The different approaches to detect the direct and indirect influence of genetic regulatory loci on TF activity (aQTL), gene expression (eQTL) and physiological phenotypes (QTL) is indicated with arrows connecting the involved data types. Figure courtesy of Weronika Sikora-Wohlfeld (adapted).

the interactions between these TFs and other regulatory proteins and the consequences of gene expression changes on hematopoiesis related phenotypes (Figure 5.1). Since it is known that hematopoiesis is mainly controlled by the interplay of a number of TFs, we are working on statistically inferring their activity from TF binding and gene expression data (Sikora-Wohlfeld et al., 2012). Subsequently, TF activity can be used as a quantitative trait in a new flavor of QTL mapping, "activity QTL" (aQTL, Lee and Bussemaker, 2010; Stegle et al., 2010). Moreover, we also dispose of QTL data on a number of hematopoiesis related phenotypes for the same mouse strains. The combination of data characterizing the impact of genetic variation on the three hierarchically organized layers of TF activity, gene expression and physiological phenotype (Figure 5.1) will deepen our knowledge on the systems biology of blood cell differentiation.

In principle, it could also be insightful to combine the ImAP results with the dynamic

eQTL data. More specifically, it can be expected that genetic loci do not affect transcript levels independently. Rather, they interact in pairs or even networks. Although the RF inherently takes these interactions into account when building the trees and will assign higher scores to predictors with interaction effects than to predictors that are uncorrelated with the response, it does not automatically return information about the interactions themselves. We already developed some extensions of the RF methodology in order to extract this additional information. For example, we applied it to find conditional eQTL, i.e. markers that are interacting with cell type indicators (see discussion in Chapter 4). However, the application of this approach is still hampered by the very small sample size of the BXD expression data and the computing power needed to grow enough trees to obtain stable results.
While the development of a fully functional and computationally feasible method for the detection of epistatic interactions with RF is still a long-term goal, existing genetic interaction data like the top-scoring ImAP pairs as well as external gene-gene interaction data might provide a simple filter for epistatic loci among the eQTL detected in the dynamic eQTL mapping. The only problem to be solved before ImAP pairs can be used to filter for eQTL interactions, is the mapping of the different sets of markers in the two mouse populations on each other. Otherwise, the pre-filtering of putative epistatic loci would allow to test the loci pairs one by one for a deviation from an additive effect on the quantitative trait, e.g. using a linear model or an ANOVA framework.

There is growing awareness of the fact that most complex traits are influenced by a large number of very small effect genetic variants as well as epigenetic and environmental effects (Maher, 2008; Sumazin et al., 2011). Of course, these factors do not act independently, but interact on different levels that are not yet completely understood. Since more and more of these factors can now be reliably measured on a large scale basis, there is an increasing need for appropriate statistical methods handling a large number of possibly diverse predictors and their interactions. Moreover, these approaches have to be steadily adapted to the nature of the emerging data. For example, we expect that whole genome sequencing and comprehensive measurements of molecular and physiological phenotypes will be carried out on a growing number of human families. The work described in this thesis is thus just a tiny contribution to the large challenges scientists in the field are facing, and there is a lot of room to extend and complement them in the future.

Bibliography

Abdollahi, A., K. A. Lord, B. Hoffman-Liebermann, and D. A. Liebermann (1991). Sequence and expression of a cDNA encoding MyD118: a novel myeloid differentiation primary response gene induced by multiple cytokines. *Oncogene 6*(1), 165–167.

Ackermann, M. and A. Beyer (2012). Systematic detection of epistatic interactions based on allele pair frequencies. *PLoS Genetics 8*(2), e1002463.

Ackermann, M., M. Clément-Ziza, J. J. Michaelson, and A. Beyer (2012). Teamwork: improved eQTL mapping using combinations of machine learning methods. *PLoS ONE, accepted*.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley-Interscience.

Alberts, B. (2002). *Molecular biology of the cell*. New York: Garland Science.

Alberts, B. (2004). *Essential cell biology*. New York: Garland Science.

Alberts, R., H. Chen, C. Pommerenke, A. B. Smit, S. Spijker, R. W. Williams, R. Geffers, D. Bruder, and K. Schughart (2011). Expression QTL mapping in regulatory and helper T cells from the BXD family of strains reveals novel cell-specific genes, gene-gene interactions and candidate genes for auto-immune disease. *BMC Genomics 12*, 610.

Alexa, A. and J. Rahnenführer (2010). topGO: enrichment analysis for gene ontology.

Alexa, A., J. Rahnenführer, and T. Lengauer (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics 22*(13), 1600–1607.

Altshuler, D., M. J. Daly, and E. S. Lander (2008). Genetic mapping in human disease. *Science 322*(5903), 881–888.

An, P., O. Mukherjee, P. Chanda, L. Yao, C. D. Engelman, C. Huang, T. Zheng, I. P. Kovac, M. Dubé, X. Liang, J. Li, M. de Andrade, R. Culverhouse, D. Malzahn, A. K. Manning *et al.* (2009). The challenge of detecting epistasis (GxG interactions): Genetic analysis workshop 16. *Genetic Epidemiology 33*(S1), S58–S67.

Anderson, J. B., J. Funt, D. A. Thompson, S. Prabhu, A. Socha, C. Sirjusingh, J. R. Dettman, L. Parreiras, D. S. Guttman, and A. Regev (2010). Determinants of divergent adaptation and Dobzhansky-Müller interaction in experimental yeast populations. *Current Biology 20*(15), 1383–1388.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics 25*(1), 25–29.

Baksh, S., H. R. Widlund, A. A. Frazer-Abel, J. Du, S. Fosmire, D. E. Fisher, J. A. DeCaprio, J. F. Modiano, and S. J. Burakoff (2002). NFATc2-mediated repression of cyclin-dependent kinase 4 expression. *Molecular Cell 10*(5), 1071–1081.

Balding, D. J., M. J. Bishop, and C. Cannings (2007). *Handbook of statistical genetics* (3rd ed.). Chichester; New York: Wiley.

Bateson, W. (1909). *Mendel's principles of heredity.* Cambridge: Cambridge University Press.

Beltrao, P., G. Cagney, and N. J. Krogan (2010). Quantitative genetic interactions reveal biological modularity. *Cell 141*(5), 739–745.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc 57*(1), 289–300.

Beyer, A., S. Bandyopadhyay, and T. Ideker (2007). Integrating physical and genetic maps: from genomes to interaction networks. *Nature Reviews. Genetics 8*(9), 699–710.

Bomblies, K. and D. Weigel (2007). Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. *Nature Reviews Genetics 8*(5), 382–393.

Breiman, L. (2001). Random forests. In *Machine Learning*, Volume 45, pp. 5–32.

Broman, K. W. and T. P. Speed (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(4), 641–656.

Bullaughey, K., C. I. Chavarria, G. Coop, and Y. Gilad (2009). Expression quantitative trait loci detected in cell lines are often present in primary tissues. *Human Molecular Genetics 18*(22), 4296–4303.

Bureau, A., J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and P. Van Eerdewegh (2005). Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology 28*(2), 171–182.

Carlborg, O. and C. S. Haley (2004). Opinion: Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics 5*(8), 618–625.

Chesler, E. J., D. R. Miller, L. R. Branstetter, L. D. Galloway, B. L. Jackson, V. M. Philip, B. H. Voy, C. T. Culiat, D. W. Threadgill, R. W. Williams, G. A. Churchill, D. K. Johnson, and K. F. Manly (2008). The collaborative cross at oak ridge national laboratory: developing a powerful resource for systems genetics. *Mammalian Genome 19*, 382–389.

Chun, H. and S. Keleş (2009). Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics 182*(1), 79–90.

Chung, H. K., Y. Yi, N. Jung, D. Kim, J. M. Suh, H. Kim, K. C. Park, J. H. Song, D. W. Kim, E. S. Hwang, S. Yoon, Y. Bae, J. M. Kim, I. Bae, and M. Shong (2003). CR6-interacting factor 1 interacts with gadd45 family proteins and modulates the cell cycle. *The Journal of Biological Chemistry 278*(30), 28079–28088.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics 11*(20), 2463.

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics 10*(6), 392–404.

Cordell, H. J., B. J. Barratt, and D. G. Clayton (2004). Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genetic Epidemiology 26*(3), 167–185.

Costanzo, M., A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Koh, K. Toufighi, S. Mostafavi, J. Prinz, R. P. St. Onge, B. VanderSluis, T. Makhnevych, F. J. Vizeacoumar *et al.* (2010). The genetic landscape of a cell. *Science 327*(5964), 425–431.

de la Chapelle, A. (1993). Disease gene mapping in isolated human populations: the example of Finland. *Journal of Medical Genetics 30*(10), 857–865.

Dermitzakis, E. T. (2008). From gene expression to disease risk. *Nature Genetics 40*(5), 492–493.

Devlin, B., K. Roeder, and L. Wasserman (2001). Genomic control, a new approach to genetic-based association studies. *Theoretical population biology 60*(3), 155–166.

Dimas, A. S., S. Deutsch, B. E. Stranger, S. B. Montgomery, C. Borel, H. Attar-Cohen, C. Ingle, C. Beazley, M. G. Arcelus, M. Sekowska, M. Gagnebin, J. Nisbett, P. Deloukas, E. T. Dermitzakis, and S. E. Antonarakis (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science 325*, 1246–1250.

Ding, J., J. E. Gudjonsson, L. Liang, P. E. Stuart, Y. Li, W. Chen, M. Weichenthal, E. Ellinghaus, A. Franke, W. Cookson, R. P. Nair, J. T. Elder, and G. R. Abecasis (2010). Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. *The American Journal of Human Genetics 87*(6), 779–789.

Durbin, R. M., D. L. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, F. S. Collins, F. M. De La Vega, P. Donnelly, M. Egholm, P. Flicek, S. B. Gabriel, R. A. Gibbs, B. M. Knoppers *et al.* (2010). A map of human genome variation from population-scale sequencing. *Nature 467*(7319), 1061–1073.

Durinck, S., P. T. Spellman, E. Birney, and W. Huber (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols 4*(8), 1184–1191.

Dutkowski, J. and T. Ideker (2011). Protein networks as logic functions in development and cancer. *PLoS Computational Biology 7*(9), e1002180.

Dybedal, I., S. Larsen, and S. E. Jacobsen (1995). IL-12 directly enhances in vitro murine erythropoiesis in combination with IL-4 and stem cell factor. *The Journal of Immunology 154*(10), 4950–4955.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*, 407–499.

Elefsinioti, A., M. Ackermann, and A. Beyer (2009). Accounting for redundancy when integrating gene interaction databases. *PloS One 4*(10), e7492.

Elefsinioti, A., O. S. Saraç, A. Hegele, C. Plake, N. C. Hubner, I. Poser, M. Sarov, A. Hyman, M. Mann, M. Schroeder, U. Stelzl, and A. Beyer (2011). Large-scale de novo prediction of physical protein-protein association. *Molecular & Cellular Proteomics: MCP 10*(11), M111.010629.

Elston, R. C. (1998). Linkage and association. *Genetic Epidemiology 15*(6), 565–576.

Emilsson, V., G. Thorleifsson, B. Zhang, A. S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G. B. Walters, S. Gunnarsdottir, M. Mouy, V. Steinthorsdottir, G. H. Eiriksdottir, G. Bjornsdottir, I. Reynisdottir *et al.* (2008). Genetics of gene expression and its effect on disease. *Nature 452*(7186), 423–428.

Faraway, J. J. (2006). *Extending the Linear Model With R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC Press.

Fisher, R. A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh 52*, 399–433.

Fu, J., M. G. M. Wolfs, P. Deelen, H. Westra, R. S. N. Fehrmann, G. J. Te Meerman, W. A. Buurman, S. S. M. Rensen, H. J. M. Groen, R. K. Weersma, L. H. van den Berg, J. Veldink, R. A. Ophoff, H. Snieder, D. van Heel *et al.* (2012). Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genetics 8*(1), e1002431.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward *et al.* (2002). The structure of haplotype blocks in the human genome. *Science 296*(5576), 2225–2229.

Geest, C. R. and P. J. Coffer (2009). MAPK signaling pathways in the regulation of hematopoiesis. *Journal of Leukocyte Biology 86*(2), 237–250.

Gerrits, A., B. Dykstra, M. Otten, L. Bystrykh, and G. Haan (2008). Combining transcriptional profiling and genetic linkage analysis to uncover gene networks operating in hematopoietic stem cells and their progeny. *Immunogenetics 60*, 411–422.

Gerrits, A., Y. Li, B. M. Tesson, L. V. Bystrykh, E. Weersing, A. Ausema, B. Dontje, X. Wang, R. Breitling, R. C. Jansen, and G. de Haan (2009). Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS Genetics 5*, e1000692.

Ghaffari, S. (2008). Oxidative stress in the regulation of normal and neoplastic hematopoiesis. *Antioxidants & Redox Signaling 10*(11), 1923–1940.

Griffiths, A. (2000). *An introduction to genetic analysis* (7th ed.). New York: W.H. Freeman.

Gulacsi, A. (2006). Shh maintains Nkx2.1 in the MGE by a Gli3-independent mechanism. *Cerebral Cortex 16*, i89–i95.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics 5*(10), e1000695.

Haley, C. S. and S. A. Knott (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity 69*(4), 315–324.

Harrell, F. E. (2001). *Regression modeling strategies : with applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Hastie, T. (2009). *The elements of statistical learning : data mining, inference, and prediction* (2nd ed.). New York: Springer.

Hébert, J. M. (2005). Unraveling the molecular pathways that regulate early telencephalon development. *Current Topics in Developmental Biology 69*, 17–37.

Heinzen, E. L., D. Ge, K. D. Cronin, J. M. Maia, K. V. Shianna, W. N. Gabriel, K. A. Welsh-Bohmer, C. M. Hulette, T. N. Denny, and D. B. Goldstein (2008). Tissue-specific genetic control of splicing: Implications for the study of complex traits. *PLoS Biology 6*(12), e1.

Hitzemann, B., K. Dains, S. Kanes, and R. Hitzemann (1994). Further studies on the relationship between dopamine cell density and haloperidol-induced catalepsy. *The Journal of Pharmacology and Experimental Therapeutics 271*(2), 969–976.

Hoh, J. and J. Ott (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics 4*(9), 701–709.

Ickstadt, K., M. Schäfer, A. Fritsch, H. Schwender, J. Abel, H. M. Bolt, T. Brüning, Y. Ko, H. Vetter, and V. Harth (2008). Statistical methods for detecting genetic interactions: a head and neck squamous-cell cancer study. *Journal of Toxicology and Environmental Health. Part A 71*(11-12), 803–815.

Ideker, T. and N. J. Krogan (2012). Differential network biology. *Molecular Systems Biology 8*(1).

Iwasaki, H. and K. Akashi (2007). Myeloid lineage commitment from the hematopoietic stem cell. *Immunity 26*(6), 726–740.

Jacobsen, S. E., O. P. Veiby, and E. B. Smeland (1993). Cytotoxic lymphocyte maturation factor (interleukin 12) is a synergistic growth factor for hematopoietic stem cells. *The Journal of Experimental Medicine 178*(2), 413–418.

Jansen, R. (2001). Genetical genomics: the added value from segregation. *Trends in Genetics 17*(7), 388–391.

Johnson, R. M., Y. Ho, D. Yu, F. A. Kuypers, Y. Ravindranath, and G. W. Goyette (2010). The effects of disruption of genes for peroxiredoxin-2, glutathione peroxidase-1, and catalase on erythrocyte oxidative metabolism. *Free Radical Biology & Medicine 48*(4), 519–525.

Keane, T. M., L. Goodstadt, P. Danecek, M. A. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, M. Goodson, N. A. Furlotte, E. Eskin, C. Nellaker, H. Whitley, J. Cleak *et al.* (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature 477*, 289–294.

Kelley, R. and T. Ideker (2005). Systematic interpretation of genetic interactions using protein networks. *Nature Biotechnology 23*(5), 561–566.

Kiani, A. (2004). Expression and regulation of NFAT (nuclear factors of activated t cells) in human CD34+ cells: down-regulation upon myeloid differentiation. *Journal of Leukocyte Biology 76*(5), 1057–1065.

Kiani, A., H. Kuithan, F. Kuithan, S. Kyttälä, I. Habermann, A. Temme, M. Bornhäuser, and G. Ehninger (2007). Expression analysis of nuclear factor of activated t cells (NFAT) during myeloid differentiation of CD34+ cells: regulation of fas ligand gene expression in megakaryocytes. *Experimental hematology 35*(5), 757–770.

Kim, Y., R. Wojciechowski, H. Sung, R. A. Mathias, L. Wang, A. P. Klein, R. K. Lenroot, J. Malley, and J. E. Bailey-Wilson (2009). Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects. *BMC Proceedings 3*(Suppl 7), S64.

Kosova, G., M. Abney, and C. Ober (2010). Colloquium papers: Heritability of reproductive fitness traits in a human population. *Proceedings of the National Academy of Sciences of the United States of America 107 Suppl 1*, 1772–1778.

Kudo, Y., D. Guardavaccaro, P. G. Santamaria, R. Koyama-Nasu, E. Latres, R. Bronson, L. Yamasaki, and M. Pagano (2004). Role of f-box protein betaTrcp1 in mammary gland development and tumorigenesis. *Molecular and Cellular Biology 24*(18), 8184–8194.

Lage, K., E. O. Karlberg, Z. M. Størling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tümer, F. Pociot, N. Tommerup, Y. Moreau, and S. Brunak (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology 25*(3), 309–316.

Lawrence, R., A. G. Day-Williams, R. Mott, J. Broxholme, L. R. Cardon, and E. Zeggini (2009). GLIDERS–a web-based search engine for genome-wide linkage disequilibrium between HapMap SNPs. *BMC Bioinformatics 10*, 367.

Lee, E. and H. J. Bussemaker (2010). Identifying the genetic determinants of transcription factor activity. *Molecular Systems Biology 6*, 412.

Lee, S., A. M. Dudley, D. Drubin, P. A. Silver, N. J. Krogan, D. Pe'er, and D. Koller (2009). Learning a prior on regulatory potential from eQTL data. *PLoS Genetics 5*(1), e1000358.

Lee, T., S. Kim, S. Yu, S. H. Kim, D. S. Park, H. Moon, S. H. Dho, K. Kwon, H. J. Kwon, Y. Han, S. Jeong, S. W. Kang, H. Shin, K. Lee, S. G. Rhee *et al.* (2003). Peroxiredoxin II is essential for sustaining life span of erythrocytes in mice. *Blood 101*(12), 5033–5038.

Lévesque, J. P., Y. Takamatsu, S. K. Nilsson, D. N. Haylock, and P. J. Simmons (2001). Vascular cell adhesion molecule-1 (CD106) is cleaved by neutrophil proteases in the bone marrow following hematopoietic progenitor cell mobilization by granulocyte colony-stimulating factor. *Blood 98*(5), 1289–1297.

Lewontin, R. C. (1964). The interaction of selection and linkage. I. general considerations; heterotic models. *Genetics 49*(1), 49–67.

Li, H. and R. Durbin (2011). Inference of human population history from individual whole-genome sequences. *Nature 475*(7357), 493–496.

Li, Y., O. A. Àlvarez, E. W. Gutteling, M. Tijsterman, J. Fu, J. A. G. Riksen, E. Hazendonk, P. Prins, R. H. A. Plasterk, R. C. Jansen, R. Breitling, and J. E. Kammenga (2006). Mapping determinants of gene expression plasticity by genetical genomics in c. elegans. *PLoS Genetics 2*(12), e222.

Li, Qing, Louis, Thomas A., Fallin, M. Daniele, and Ruczinski, Ingo (2009). Trio logic regression - detection of SNP - SNP interactions in case-parent trios. *Johns Hopkins University, Dept. of Biostatistics Working Papers 194*.

Liu, T., A. Thalamuthu, J. Liu, C. Chen, Z. Wang, and R. Wu (2011). Asymptotic distribution for epistatic tests in case-control studies. *Genomics 98*, 145–151.

Loguercio, S., R. W. Overall, J. J. Michaelson, T. Wiltshire, M. T. Pletcher, B. H. Miller, J. R. Walker, G. Kempermann, A. I. Su, and A. Beyer (2010). Integrative analysis of low- and High-Resolution eQTL. *PLoS ONE 5*, e13920.

Lohmueller, K. E., C. L. Pearce, M. Pike, E. S. Lander, and J. N. Hirschhorn (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics 33*(2), 177–182.

Maeda, Y., V. Dave, and J. A. Whitsett (2007). Transcriptional control of lung morphogenesis. *Physiological Reviews 87*, 219–244.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature 456*(7218), 18–21.

Marchini, J., P. Donnelly, and L. R. Cardon (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics 37*(4), 413–417.

McDermott, J., M. Guerquin, Z. Frazier, A. N. Chang, and R. Samudrala (2005). BIOVERSE: enhancements to the framework for structural, functional and contextual modeling of proteins and proteomes. *Nucleic Acids Research 33*(Web Server), W324–W325.

McLean, J. R., C. J. Merrill, P. A. Powers, and B. Ganetzky (1994). Functional identification of the segregation distorter locus of drosophila melanogaster by germline transformation. *Genetics 137*(1), 201–209.

Mehrabian, M., H. Allayee, J. Stockton, P. Y. Lum, T. A. Drake, L. W. Castellani, M. Suh, C. Armour, S. Edwards, J. Lamb, A. J. Lusis, and E. E. Schadt (2005). Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nature Genetics 37*(11), 1224–1233.

Michaelson, J. J., R. Alberts, K. Schughart, and A. Beyer (2010). Data-driven assessment of eQTL mapping methods. *BMC Genomics 11*(1), 502.

Miletich, I., M. T. Cobourne, M. Abdeen, and P. T. Sharpe (2005). Expression of the hedgehog antagonists rab23 and Slimb/betaTrCP during mouse tooth development. *Archives of Oral Biology 50*(2), 147–151.

Minoo, P. (1999). Defects in tracheoesophageal and lung morphogenesis in Nkx2.1(-/-) mouse embryos. *Developmental Biology 209*, 60–71.

Montagutelli, X., R. Turner, and J. H. Nadeau (1996). Epistatic control of non-Mendelian inheritance in mouse interspecific crosses. *Genetics 143*(4), 1739.

Moon, E., Y. H. Han, D. Lee, Y. Han, and D. Yu (2004). Reactive oxygen species induced by the deletion of peroxiredoxin II (PrxII) increases the number of thymocytes resulting in the enlargement of PrxII-null thymus. *European Journal of Immunology 34*(8), 2119–2128.

Moon, E., Y. Noh, Y. Han, S. Kim, J. Kim, D. Yu, and J. Lim (2006). T lymphocytes and dendritic cells are activated by the deletion of peroxiredoxin II (Prx II) gene. *Immunology Letters 102*(2), 184–190.

Mott, R., C. J. Talbot, M. G. Turri, A. C. Collins, and J. Flint (2000). A method for fine mapping quantitative trait loci in outbred animal stocks. *Proceedings of the National Academy of Sciences of the United States of America 97*(23), 12649–12654.

Müller-Sieburg, C. E., R. H. Cho, H. B. Sieburg, S. Kupriyanov, and R. Riblet (2000). Genetic control of hematopoietic stem cell frequency in mice is mostly cell autonomous. *Blood 95*(7), 2446–2448.

Nelder, J. and R. Mead (1965). A simplex algorithm for function minimization. *Computer Journal 7*, 308–313.

Nica, A. C., L. Parts, D. Glass, J. Nisbet, A. Barrett, M. Sekowska, M. Travers, S. Potter, E. Grundberg, K. Small, Ã. K. Hedman, V. Bataille, J. Tzenova Bell, G. Surdulescu, A. S. Dimas *et al.* (2011). The architecture of gene regulatory variation across multiple human tissues: The MuTHER study. *PLoS Genetics 7*, e1002003.

Orkin, S. H. and L. I. Zon (2008). Hematopoiesis: An evolving paradigm for stem cell biology. *Cell 132*(4), 631–644.

Orr, H. A. (1996). Dobzhansky, Bateson, and the genetics of speciation. *Genetics 144*(4), 1331–1335.

Payseur, B. A. and M. Place (2007). Searching the genomes of inbred mouse strains for incompatibilities that reproductively isolate their wild relatives. *Journal of Heredity 98*(2), 115–122.

Pedchenko, V. K. and W. Imagawa (2000). Pattern of expression of the KGF receptor and its ligands KGF and FGF-10 during postnatal mouse mammary gland development. *Molecular Reproduction and Development 56*(4), 441–447.

Petretto, E., J. Mangion, N. J. Dickens, S. A. Cook, M. K. Kumaran, H. Lu, J. Fischer, H. Maatz, V. Kren, M. Pravenec, N. Hubner, and T. J. Aitman (2006). Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genetics 2*(10), e172.

Phillips, P. C. (2008). Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics 9*(11), 855–867.

Picotti, P., M. Clément-Ziza, H. Lam, D. S. Campbell, E. W. Deutsch, H. Röst, Z. Sun, J. J. Michaelson, O. Rinner, A. Schmidt, Q. Shen, A. Frei, B. Wollscheid, A. Beyer, and R. Aebersold (2012). A mass spectrometric map for the analysis of the yeast proteome and its application to quantitative trait analysis. *in preparation*.

Pispa, J., H. S. Jung, J. Jernvall, P. Kettunen, T. Mustonen, M. J. Tabata, J. Kere, and I. Thesleff (1999). Cusp patterning defect in tabby mouse teeth and its partial rescue by FGF. *Developmental Biology 216*(2), 521–534.

Powell, J. E., A. K. Henders, A. F. McRae, M. J. Wright, N. G. Martin, E. T. Dermitzakis, G. W. Montgomery, and P. M. Visscher (2011). Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. *Genome Research 22*(3), 456–466.

Price, A. L., A. Helgason, G. Thorleifsson, S. A. McCarroll, A. Kong, and K. Stefansson (2011). Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genetics 7*(2), e1001317.

Prill, R. J., D. Marbach, J. Saez-Rodriguez, P. K. Sorger, L. G. Alexopoulos, X. Xue, N. D. Clarke, G. Altan-Bonnet, and G. Stolovitzky (2010). Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PloS One 5*(2), e9202.

R Development Core Team (2011). R: A language and environment for statistical computing.

Rabinowitz, D. and N. Laird (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human Heredity 50*(4), 211–223.

Reuter, C. W. M., M. A. Morgan, and L. Bergmann (2000). Targeting the Ras signaling pathway: A rational, mechanism-based treatment for hematologic malignancies? *Blood 96*(5), 1655–1669.

Rhodes, D. R., S. A. Tomlins, S. Varambally, V. Mahavisno, T. Barrette, S. Kalyana-Sundaram, D. Ghosh, A. Pandey, and A. M. Chinnaiyan (2005). Probabilistic model of the human protein-protein interaction network. *Nature Biotechnology 23*(8), 951–959.

Ritchie, M. D., L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics 69*(1), 138–147.

Roach, J. C., G. Glusman, A. F. A. Smit, C. D. Huff, R. Hubley, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, M. Bamshad, J. Shendure, R. Drmanac, L. B. Jorde, L. Hood, and D. J. Galas (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science 328*(5978), 636–639.

Roberts, A., F. Pardo-Manuel de Villena, W. Wang, L. McMillan, and D. W. Threadgill (2007). The polymorphism architecture of mouse genetic resources elucidated using genome-wide resequencing data: implications for QTL discovery and systems genetics. *Mammalian Genome 18*(6-7), 473–481.

Rockman, M. V. (2008). Reverse engineering the genotype-phenotype map with natural genetic variation. *Nature 456*(7223), 738–744.

Rockman, M. V. and L. Kruglyak (2006). Genetics of global gene expression. *Nature Reviews Genetics 7*(11), 862–872.

Roguev, A., S. Bandyopadhyay, M. Zofall, K. Zhang, T. Fischer, S. R. Collins, H. Qu, M. Shales, H. Park, J. Hayles, K. Hoe, D. Kim, T. Ideker, S. I. Grewal, J. S. Weissman *et al.* (2008). Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science 322*(5900), 405–410.

Sakiyama, J.-i. (2003). Tbx4-Fgf10 system controls lung bud formation during chicken embryonic development. *Development 130*, 1225–1234.

Sakoparnig, T., T. Kockmann, R. Paro, C. Beisel, and N. Beerenwinkel (2012). Binding profiles of chromatin-modifying proteins are predictive for transcriptional activity and promoter-proximal pausing. *Journal of Computational Biology 19*(2), 126–138.

Schaid, D. J. (1999). Case-parents design for gene-environment interaction. *Genetic Epidemiology 16*(3), 261–273.

Scheet, P. and M. Stephens (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics 78*(4), 629–644.

Schuldiner, M., S. R. Collins, N. J. Thompson, V. Denic, A. Bhamidipati, T. Punna, J. Ihmels, B. Andrews, C. Boone, J. F. Greenblatt, J. S. Weissman, and N. J. Krogan (2005). Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell 123*(3), 507–519.

Schwarz, D. F., I. R. König, and A. Ziegler (2010). On safari to random jungle: A fast implementation of random forests for high-dimensional data. *Bioinformatics 26*(14), 1752–1758.

Schwender, H. (2011). Testing SNPs and sets of SNPs for importance in association studies. *Biostatistics 12*(1), 18–32.

Schwender, H. and K. Ickstadt (2008). Identification of SNP interactions using logic regression. *Biostatistics 9*(1), 187.

Schwender, H., Q. Li, and I. Ruczinski (2012). trio: Detection of disease-associated SNP interactions in case-parent trio data.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology 46*(1), 561–584.

Shifman, S., J. T. Bell, R. R. Copley, M. S. Taylor, R. W. Williams, R. Mott, and J. Flint (2006). A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *PLoS Biology 4*(12), e395.

Shivdasani, R. A. and S. H. Orkin (1996). The transcriptional control of hematopoiesis. *Blood 87*(10), 4025–4039.

Sieburth, D. S., Q. Sun, and M. Han (1998). SUR-8, a conserved Ras-binding protein with leucine-rich repeats, positively regulates Ras-mediated signaling in c. elegans. *Cell 94*(1), 119–130.

Sikora-Wohlfeld, W., M. Ackermann, E. Christodoulou, and A. Beyer (2012). Transcription factor target gene identification based on ChIP-seq data. *submitted*.

Smith, E. N. and L. Kruglyak (2008). Gene-environment interaction in yeast gene expression. *PLoS Biology 6*(4), e83.

Spielman, R. S., R. E. McGinnis, and W. J. Ewens (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics 52*(3), 506–516.

Stark, C., B. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski, and M. Tyers (2011). The BioGRID interaction database: 2011 update. *Nucleic Acids Research 39*(Database issue), D698–704.

Stearns, S. C., S. G. Byars, D. R. Govindaraju, and D. Ewbank (2010). Measuring selection in contemporary human populations. *Nature Reviews. Genetics 11*(9), 611–622.

Stegle, O., L. Parts, R. Durbin, and J. Winn (2010). A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology 6*(5), e1000770.

Strobl, C., A. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008). Conditional variable importance for random forests. *BMC Bioinformatics 9*, 307.

Strobl, C., A. Boulesteix, A. Zeileis, and T. Hothorn (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics 8*, 25.

Sumazin, P., X. Yang, H. Chiu, W. Chung, A. Iyer, D. Llobet-Navas, P. Rajbhandari, M. Bansal, P. Guarnieri, J. Silva, and A. Califano (2011). An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell 147*(2), 370–381.

Suthram, S., A. Beyer, R. M. Karp, Y. Eldar, and T. Ideker (2008). eQED: an efficient method for interpreting eQTL associations using protein networks. *Molecular Systems Biology 4*, 162.

Swiers, G., R. Patient, and M. Loose (2006). Genetic regulatory networks programming hematopoietic stem cells and erythroid lineage specification. *Developmental Biology 294*(2), 525–540.

Szklarczyk, D., A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research 39*(Database issue), D561–568.

Takakura, N., T. Watanabe, S. Suenobu, Y. Yamada, T. Noda, Y. Ito, M. Satake, and T. Suda (2000). A role for hematopoietic stem cells in promoting angiogenesis. *Cell 102*(2), 199–209.

Thomas, D. C. (2004). *Statistical Methods in Genetic Epidemiology*. Oxford University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society 58*, 267–288.

Tong, A. H., M. Evangelista, A. B. Parsons, H. Xu, G. D. Bader, N. Pagé, M. Robinson, S. Raghibizadeh, C. W. Hogue, H. Bussey, B. Andrews, M. Tyers, and C. Boone (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science 294*(5550), 2364–2368.

Trinchieri, G. (2003). Interleukin-12 and the regulation of innate resistance and adaptive immunity. *Nature Reviews Immunology 3*(2), 133–146.

Van Zant, G., P. W. Eldridge, R. R. Behringer, and M. J. Dewey (1983). Genetic control of hematopoietic kinetics revealed by analyses of allophenic mice and stem cell suicide. *Cell 35*(3), 639–645.

Wall, J. D. and J. K. Pritchard (2003). Assessing the performance of the haplotype block model of linkage disequilibrium. *American Journal of Human Genetics 73*(3), 502–515.

Wang, J., R. W.Williams, and K. F. Manly (2003). WebQTL: web-based complex trait analysis. *Neuroinformatics 1*(4), 299–308.

Wang, Z., T. Liu, Z. Lin, J. Hegarty, W. A. Koltun, and R. Wu (2010). A general model for multilocus epistatic interactions in case-control studies. *PLoS ONE 5*, e11384.

Weinrauch, Y., D. Drujan, S. D. Shapiro, J. Weiss, and A. Zychlinsky (2002). Neutrophil elastase targets virulence factors of enterobacteria. *Nature 417*(6884), 91–94.

Williams, R. W., J. Gu, S. Qi, and L. Lu (2001). The genetic structure of recombinant inbred mice: high-resolution consensus maps for complex trait analysis. *Genome Biol 2*(11), 10046.

Ye, P., B. D. Peyser, X. Pan, J. D. Boeke, F. A. Spencer, and J. S. Bader (2005). Gene function prediction from congruent synthetic lethal interactions in yeast. *Molecular Systems Biology 1*(1), E1–E12.

Yen, A. and K. L. Albright (1984). Evidence for cell cycle phase-specific initiation of a program of HL-60 cell myeloid differentiation mediated by inducer uptake. *Cancer Research 44*(6), 2511–2515.

Yi, N. (2011). Statistical analysis of genetic interactions. *Genetics Research 92*(5-6), 443–459.

Yoshida, M. and A. Koike (2011). SNPInterForest: a new method for detecting epistatic interactions. *BMC Bioinformatics 12*, 469.

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics 136*(4), 1457–1468.

Zheng, G., J. Joo, and Y. Yang (2009). Pearson's test, trend test, and MAX are all trend tests with different types of scores. *Annals of Human Genetics 73*(2), 133–140.

Zhong, H., J. Beaulaurier, P. Y. Lum, C. Molony, X. Yang, D. J. Macneil, D. T. Weingarth, B. Zhang, D. Greenawalt, R. Dobrin, K. Hao, S. Woo, C. Fabre-Suver, S. Qian, M. R. Tota *et al.* (2010). Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genetics 6*(5), e1000932.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*, 301–320.

Appendix A

Statistical methods

A.1 Lewontin's D'

Suppose a two locus, two allele model where loci A and B each have two possible alleles 1 and 2. Let $x_{A_1B_1}$ be the observed frequency of the haplotype A_1B_1 in the population, p_1 and q_1 the expected frequencies of A_1 and B_1, respectively. (The frequencies of the alternative alleles are defined accordingly.) Then, the measure D of allelic association is defined as

$$D = x_{11} - p_1 q_1.$$

The two alleles A_1 and B_1 are said to be "in LD" if $D \neq 0$.
Lewontin (Lewontin, 1964) proposed to normalise D, because in its original form it depends on the allele frequencies. The normalised measure D' is given as

$$D' = \frac{D}{D_{\max}}, \text{ where}$$
$$D_{\max} = \begin{cases} \min(p_1 q_1, p_2 q_2), & \text{if } D < 0 \\ \min(p_1 q_2, p_2 q_1), & \text{if } D > 0 \end{cases}$$

A.2 Random Forests

Random Forests (RF) is a machine learning approach based on an ensemble of decision trees, which is widely used for non-parametric regression and classification problems. The following description of the Random Forests methodology is adapted from Hastie (2009).
Let $\mathbf{X}$ be a $(n \times p)$ dimensional matrix of p predictors and n samples and $\mathbf{y}$ the corresponding response vector of length n. RF creates an ensemble of decision trees, in which each tree is based on a different bootstrap sample of the the data, i.e. n pairs $(\mathbf{x}_i, y_i),\ i = 1, \ldots, n$, are drawn with replacement from the data and used to grow the tree. At each node of the tree,

the predictor that best separates the response $\mathbf{y}$ into two subsets being as homogenous as possible is chosen among a random subset of m predictors. More specifically, the algorithm seeks the predictor $\mathbf{x}_j$ and the split point s that solve

$$\min_{j,s}\left[\min_{c_1}\sum_{i|x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{i|x_i\in R_2(j,s)}(y_i-c_2)^2\right].$$

$R_1(j,s)=\{X|X_j\leq s\}$ and $R_2(j,s)=\{X|X_j>s\}$ is the pair of half-planes in which the data are separated.
Regardless of j and s this minimization is solved for $c_1=\frac{1}{|R_1(j,s)|}\sum_{i|x_i\in R_1(j,s)}y_i$ and $c_2=\frac{1}{|R_2(j,s)|}\sum_{i|x_i\in R_2(j,s)}y_i$, which are the average response values in each of the two subsets R_1 and R_2. An overview of the complete RF algorithm is given below.

Algorithm Random Forests

---

1. For b = 1 to B:
   (a) Draw a bootstrap sample $(\mathbf{X}^*,\mathbf{y}^*)$ of size n from the training data.
   (b) Grow a RF tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $\text{ns}_{\min}$ is reached.
      (i) Select m variables at random from the p variables.
      (ii) Pick the best variable/split-point among the m predictors.
      (iii) Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

   To make a prediction at a new point x: $\hat{f}_{\text{rf}}^B(x)=\frac{1}{B}\sum_{b=1}^{B}T_b(x)$.

The "randomness" of RF is two-fold: First, the same regression tree is fit to a number of random instances of the data (bootstrap samples). Second, at each split in each tree a randomly selected subset of predictors is combed through in order to find the best predictor at this node. Both strategies have the same aim, the reduction of the variance of the regression model. While the bootstrap sampling ensures that we can build identically distributed, but still varying trees (similar to the idea of cross-validation), the random sampling of predictors diminishes the correlation between the regression trees, which in turn increases the potential of variance reduction for the ensemble predictor (Hastie, 2009).

RF provides several measures for the average importance of each of the variables on the prediction. These are the permutation importance (the loss of prediction accuracy on the test data of each tree after permuting the values of the predictor among strains), the residual sum of squares importance (the total decrease in node impurities from splitting on the predictor) and the selection frequency (the number of times a variable has been selected for prediction over all trees). Throughout this thesis, we use the selection frequency as a measure of variable importance.

A.3 *F* statistic for model comparison

Consider two linear models M_1 and M_2 with independent and identically normally distributed errors. Further assume that the parameters in M_2 can be represented as a linear restriction of the parameters in M_1, e.g. if they are a subset of the parameters in M_1. The number of parameters in the larger model M_1 is p, the corresponding number in the smaller model is q. The null hypothesis that the restricted model M_2 is correct (and not the full model M_1) can be tested using the F statistic:

$$F = \frac{(RSS_2 - RSS_1)/(p-q)}{RSS_1/(n-p)},$$

where RSS is the residual sum of squares of the given model and n is the size of the data set used for estimating the models. The F statistic follows an F distribution with $(p-q)$ and $(n-p)$ degrees of freedom (Faraway, 2006).

A.4 Wald test

In a linear model of the form $\mathbf{y} = \mathbf{X}\beta + \epsilon$ let $\mathbf{y}$ be a continuous response vector of length n, $\mathbf{X}$ the $(n \times (p+1))$-dimensional matrix of p categorical or continuous predictors and an intercept, β the $(p+1)$-dimensional parameter vector describing the influence of each predictor on the outcome and ϵ an error term, where individual errors are assumed to identically and independently follow a standard normal distribution. Denote $\hat{\beta}$ the ordinary least squares estimate of β and $\hat{\Sigma}$ the estimated covariance matrix of β.
The Wald test can be used to test whether the parameter vector β or a linear combination thereof, $c^T\beta$ with variance $c^T\Sigma c$, is different from 0 (Harrell, 2001). In the latter case, the Wald test statistic is given as:

$$W = \frac{c^T \hat{\beta}}{c^T \hat{\Sigma} c}.$$

In a linear model with normally distributed error terms, W follows a Student's t-distribution with $n - p - 1$ degrees of freedom.

A.5 Congruence score

Let m_1 and m_2 be two nodes within a network of p nodes. Further assume that m_1 and m_2 each have r and s interactions respectively with other nodes in the network and that overall T pairwise interactions between nodes could be observed. The congruence score is a measure of the probability of observing at least t shared interactors between m_1 and m_2. It can be derived from the hypergeometric distribution as:

$$P(x \geq t) = \sum_{x=t}^{min(r,s)} C(T - r, s - x)/C(T, s),$$

where

$$C(a, b) = \frac{a!}{b!(a - b)!}$$

is the combinatorial factor. The congruence score is defined as the negative $\log_{10}$ transformation of $P(x \geq t)$.

Appendix B

GO enrichment of significant ImAP loci

Table B.1: GO enrichment of top ranking marker pairs in the original data. All genes between the flanking markers are considered.

GO ID	Term	weighting p-value
GO:0060592	mammary gland formation	< 0.00001
GO:0060487	lung epithelial cell differentiation	< 0.00001
GO:0060441	branching involved in lung morphogenesis	< 0.00001
GO:0021879	forebrain neuron differentiation	0.000046
GO:0032438	melanosome organization	0.000046
GO:0030878	thyroid gland development	0.000079
GO:0008593	regulation of Notch signaling pathway	0.000081
GO:0090130	tissue migration	0.000081
GO:0007034	vacuolar transport	0.00021
GO:0051345	positive regulation of hydrolase activity	0.00022
GO:0060740	prostate gland epithelium morphogenesis	0.00031
GO:0032496	response to lipopolysaccharide	0.00035
GO:0060788	ectodermal placode formation	0.00082
GO:0009880	embryonic pattern specification	0.00115
GO:0046638	positive regulation of alpha-beta T cell differentiation	0.00124
GO:0022600	digestive system process	0.00151
GO:0045931	positive regulation of mitotic cell cycle	0.00151
GO:0048839	inner ear development	0.0017
GO:0050821	protein stabilization	0.0018
GO:0021983	pituitary gland development	0.00193
GO:0046579	positive regulation of Ras protein signaling	0.00251
GO:0042593	glucose homeostasis	0.00306
GO:0060606	tube closure	0.00306
GO:0042246	tissue regeneration	0.00338
GO:0021761	limbic system development	0.00376
GO:0048762	mesenchymal cell differentiation	0.00632
GO:0006829	zinc ion transport	0.00722
GO:0031128	developmental induction	0.00722

GO:0008033	tRNA processing	0.00778
GO:0042326	negative regulation of phosphorylation	0.00778
GO:0034613	cellular protein localization	0.00809
GO:0019882	antigen processing and presentation	0.00895
GO:0048730	epidermis morphogenesis	0.00895
GO:0006338	chromatin remodeling	0.00913
GO:0007050	cell cycle arrest	0.00913
GO:0048546	digestive tract morphogenesis	0.00913
GO:0007205	activation of protein kinase C activity by G-protein coupled receptor protein signaling pathway	0.01
GO:0009268	response to pH	0.01
GO:0010948	negative regulation of cell cycle process	0.01
GO:0045737	positive regulation of cyclin-dependent protein kinase activity	0.01
GO:0048565	gut development	0.01063
GO:0001667	ameboidal cell migration	0.01229
GO:0019827	stem cell maintenance	0.01329
GO:0043616	keratinocyte proliferation	0.01329
GO:0046148	pigment biosynthetic process	0.01329
GO:0048146	positive regulation of fibroblast proliferation	0.01329
GO:0050654	chondroitin sulfate proteoglycan metabolism	0.01329
GO:0051145	smooth muscle cell differentiation	0.01329
GO:0090263	positive regulation of Wnt receptor signaling	0.01329
GO:0042476	odontogenesis	0.01412
GO:0051091	positive regulation of sequence-specific DNA binding transcription factor activity	0.01459
GO:0050921	positive regulation of chemotaxis	0.0157
GO:0008589	regulation of smoothened signaling pathway	0.01712
GO:0018149	peptide cross-linking	0.01712
GO:0045666	positive regulation of neuron differentiation	0.01831
GO:0009948	anterior/posterior axis specification	0.0215
GO:0030512	negative regulation of transforming growth factor beta receptor signaling pathway	0.0215
GO:0032312	regulation of ARF GTPase activity	0.0215
GO:0042990	regulation of transcription factor import into nucleus	0.0215
GO:0048010	vascular endothelial growth factor receptor signaling pathway	0.0215
GO:0070374	positive regulation of ERK1 and ERK2 cascade	0.0215
GO:0030539	male genitalia development	0.02645
GO:0045740	positive regulation of DNA replication	0.02874
GO:0031016	pancreas development	0.02902
GO:0007368	determination of left/right symmetry	0.03194

GO:0031076	embryonic camera-type eye development	0.03194
GO:0034976	response to endoplasmic reticulum stress	0.03194
GO:0048286	lung alveolus development	0.03194
GO:0060560	developmental growth involved in morphogenesis	0.03194
GO:0060571	morphogenesis of an epithelial fold	0.03194
GO:0060603	mammary gland duct morphogenesis	0.03194
GO:0050878	regulation of body fluid levels	0.03245
GO:0043627	response to estrogen stimulus	0.0348
GO:0001823	mesonephros development	0.038
GO:0030318	melanocyte differentiation	0.038
GO:0048009	insulin-like growth factor receptor signaling	0.038
GO:0048538	thymus development	0.038
GO:0050918	positive chemotaxis	0.038
GO:0060324	face development	0.04459
GO:0060445	branching involved in salivary gland morphogenesis	0.04459
GO:0060993	kidney morphogenesis	0.04459
GO:0007492	endoderm development	0.04677

Table B.2: GO enrichment of top ranking marker pairs in the simulated data. All genes between the flanking markers are considered.

GO ID	Term	weighting p-value
GO:0009581	detection of external stimulus	< 0.00001
GO:0010761	fibroblast migration	0.000041
GO:0002474	antigen processing and presentation of peptide antigen via MHC class I	0.000084
GO:0009582	detection of abiotic stimulus	0.00029
GO:0046058	cAMP metabolic process	0.00055
GO:0051320	S phase	0.00102
GO:0048585	negative regulation of response to stimulus	0.00104
GO:0006813	potassium ion transport	0.00193
GO:0007286	spermatid development	0.00298
GO:0006195	purine nucleotide catabolic process	0.00355
GO:0055085	transmembrane transport	0.00388
GO:0009266	response to temperature stimulus	0.00443
GO:0001541	ovarian follicle development	0.00504
GO:0001910	regulation of leukocyte mediated cytotoxicity	0.00512
GO:0006997	nucleus organization	0.00512
GO:0007613	memory	0.00604
GO:0030521	androgen receptor signaling pathway	0.00604
GO:0009207	purine ribonucleoside triphosphate catabolic process	0.00659
GO:0016525	negative regulation of angiogenesis	0.0089
GO:0007018	microtubule-based movement	0.01085
GO:0002707	negative regulation of lymphocyte mediated immunity	0.01253
GO:0030048	actin filament-based movement	0.01253
GO:0045582	positive regulation of T cell differentiation	0.01253
GO:0009416	response to light stimulus	0.01424
GO:0071706	tumor necrosis factor superfamily cytokine production	0.01446
GO:0030198	extracellular matrix organization	0.01471
GO:0006096	glycolysis	0.01526
GO:0030335	positive regulation of cell migration	0.01609
GO:0045333	cellular respiration	0.01789
GO:0002366	leukocyte activation during immune response	0.02345
GO:0010948	negative regulation of cell cycle process	0.02345
GO:0034613	cellular protein localization	0.02435
GO:0006939	smooth muscle contraction	0.02843
GO:0019048	virus-host interaction	0.02858
GO:0055114	oxidation reduction	0.0324
GO:0015833	peptide transport	0.03488
GO:0002064	epithelial cell development	0.03578
GO:0006986	response to unfolded protein	0.03578
GO:0034754	cellular hormone metabolic process	0.04031
GO:0031343	positive regulation of cell killing	0.04196
GO:0042439	ethanolamine and derivative metabolic process	0.04196
GO:0042446	hormone biosynthetic process	0.04394
GO:0048741	skeletal muscle fiber development	0.04394
GO:0034504	protein localization in nucleus	0.04722

Appendix C

GO enrichment of static, conditional and dynamic eQTL

C.1 Functional enrichment of eQTL targets

Table C.1: Stem cell specific eQTL targets.

GO.ID	Term	p-value	FDR
GO:0001763	morphogenesis of a branching structure	0.00043	0.00028
GO:0002042	cell migration involved in sprouting angiogenesis	0.00043	0.00028
GO:0042036	negative regulation of cytokine biosynthetic process	0.00078	0.00055
GO:0032355	response to estradiol stimulus	0.00214	0.00111
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	0.00214	0.00111
GO:0061039	ovum-producing ovary development	0.00495	0.00166
GO:0043406	positive regulation of MAP kinase activity	0.00541	0.00194
GO:0043112	receptor metabolic process	0.00616	0.00250
GO:0002688	regulation of leukocyte chemotaxis	0.00816	0.00333

Table C.2: Progenitor cell specific eQTL targets.

GO.ID	Term	p-value	FDR
GO:0006508	proteolysis	0.0152	0.00056

Table C.3: Erythroid cell specific eQTL targets.

GO.ID	Term	p-value	FDR
GO:0045744	negative regulation of G-protein coupled receptor protein signaling pathway	0.0017	0.00028
GO:0034341	response to interferon-gamma	0.0022	0.00056
GO:0006024	glycosaminoglycan biosynthetic process	0.0031	0.00056
GO:0045926	negative regulation of growth	0.0032	0.00083
GO:0006081	cellular aldehyde metabolic process	0.0041	0.00111
GO:0015807	L-amino acid transport	0.0041	0.00111
GO:0030166	proteoglycan biosynthetic process	0.0046	0.00111
GO:0008361	regulation of cell size	0.0070	0.00139
GO:0006979	response to oxidative stress	0.0088	0.00278
GO:0043066	negative regulation of apoptosis	0.0101	0.00305

Table C.4: Myeloid cell specific eQTL targets.

GO.ID	Term	p-value	FDR
GO:0070838	divalent metal ion transport	0.00034	0.00083
GO:0050829	defense response to Gram-negative bacterium	0.00083	0.00139
GO:0050830	defense response to Gram-positive bacterium	0.00136	0.00194
GO:0030318	melanocyte differentiation	0.00171	0.00222
GO:0009132	nucleoside diphosphate metabolic process	0.00197	0.00222
GO:0050730	regulation of peptidyl-tyrosine phosphorylation	0.00266	0.00222
GO:0006611	protein export from nucleus	0.00384	0.00250
GO:0006886	intracellular protein transport	0.00419	0.00250
GO:0050766	positive regulation of phagocytosis	0.00480	0.00250
GO:0009166	nucleotide catabolic process	0.00497	0.00250

Table C.5: Dynamic progenitor to myeloid differentation eQTL targets.

GO.ID	Term	p-value	FDR
GO:0003229	ventricular cardiac muscle tissue development	0.00036	0.00000
GO:0003208	cardiac ventricle morphogenesis	0.00108	0.00000
GO:0050868	negative regulation of T cell activation	0.00164	0.00000
GO:0018108	peptidyl-tyrosine phosphorylation	0.02375	0.00083
GO:0006954	inflammatory response	0.02852	0.00083

Table C.6: Static eQTL targets.

GO.ID	Term	p-value	FDR
GO:0034645	cellular macromolecule biosynthetic process	< 0.00001	0.00000
GO:0034728	nucleosome organization	< 0.00001	0.00000
GO:0032774	RNA biosynthetic process	< 0.00001	0.00000
GO:0031497	chromatin assembly	< 0.00001	0.00000
GO:0010467	gene expression	0.00016	0.00000
GO:0006338	chromatin remodeling	0.00038	0.00000
GO:0016071	mRNA metabolic process	0.00049	0.00000
GO:0007034	vacuolar transport	0.00101	0.00027
GO:0000184	nuclear-transcribed mRNA catabolic process	0.00133	0.00082
GO:0002224	toll-like receptor signaling pathway	0.00133	0.00082

C.2 Functional enrichment of eQTL markers

Table C.7: Stem cell specific eQTL markers.

GO.ID	Term	p-value	FDR
GO:0007219	Notch signaling pathway	< 0.00001	0.00000
GO:0046394	carboxylic acid biosynthetic process	0.00003	0.00026
GO:0030641	regulation of cellular pH	0.00026	0.00184
GO:0006024	glycosaminoglycan biosynthetic process	0.00031	0.00184
GO:0042060	wound healing	0.00034	0.00184
GO:0070232	regulation of T cell apoptosis	0.00059	0.00236
GO:0060603	mammary gland duct morphogenesis	0.00066	0.00263
GO:0035108	limb morphogenesis	0.00073	0.00289
GO:0050817	coagulation	0.00076	0.00289
GO:0000266	mitochondrial fission	0.00076	0.00289

Table C.8: Progenitor cell specific eQTL markers.

GO.ID	Term	p-value	FDR
GO:0046632	alpha-beta T cell differentiation	0.00029	0.00079
GO:0002444	myeloid leukocyte mediated immunity	0.00030	0.00079
GO:0032102	negative regulation of response to external stimulus	0.00077	0.00131
GO:0045604	regulation of epidermal cell differentiation	0.00078	0.00158
GO:0002456	T cell mediated immunity	0.00085	0.00158
GO:0048041	focal adhesion assembly	0.00091	0.00158
GO:0030334	regulation of cell migration	0.00096	0.00158
GO:0045807	positive regulation of endocytosis	0.00110	0.00158
GO:0051797	regulation of hair follicle development	0.00167	0.00210
GO:0006909	phagocytosis	0.00203	0.00263

Table C.9: Erythroid cell specific eQTL markers.

GO.ID	Term	p-value	FDR
GO:0051091	positive regulation of sequence-specific DNA-binding transcription factor activity	0.00001	0.00053
GO:0043410	positive regulation of MAPKKK cascade	0.00002	0.00053
GO:0010883	regulation of lipid storage	0.00002	0.00053
GO:0071901	negative regulation of protein serine/threonine kinase activity	0.00007	0.00158
GO:0090207	regulation of triglyceride metabolic process	0.00007	0.00158
GO:0030168	platelet activation	0.00008	0.00184
GO:0031331	positive regulation of cellular catabolic process	0.00009	0.00236
GO:0043409	negative regulation of MAPKKK cascade	0.00010	0.00236
GO:0048610	cellular process involved in reproduction	0.00012	0.00263
GO:0051346	negative regulation of hydrolase activity	0.00012	0.00263

Table C.10: Myeloid cell specific eQTL markers.

GO.ID	Term	p-value	FDR
GO:0006897	endocytosis	< 0.00001	0.00000
GO:0090207	regulation of triglyceride metabolic process	< 0.00001	0.00053
GO:0051222	positive regulation of protein transport	< 0.00001	0.00184
GO:0051052	regulation of DNA metabolic process	< 0.00001	0.00184
GO:0043407	negative regulation of MAP kinase activity	< 0.00001	0.00263
GO:0046889	positive regulation of lipid biosynthetic process	< 0.00001	0.00263
GO:0032102	negative regulation of response to external stimulus	< 0.00001	0.00263
GO:0031348	negative regulation of defense response	0.00002	0.00315
GO:0043409	negative regulation of MAPKKK cascade	0.00002	0.00420
GO:0043623	cellular protein complex assembly	0.00002	0.00420

Table C.11: Dynamic progenitor to myeloid differentiation specific eQTL markers.

GO.ID	Term	p-value	FDR
GO:0007033	vacuole organization	< 0.00001	0.00000
GO:0006644	phospholipid metabolic process	< 0.00001	0.00026
GO:0046677	response to antibiotic	0.00001	0.00079
GO:0045061	thymic T cell selection	0.00002	0.00079
GO:0016049	cell growth	0.00003	0.00105
GO:0017156	calcium ion-dependent exocytosis	0.00006	0.00184
GO:0009611	response to wounding	0.00006	0.00184
GO:0048813	dendrite morphogenesis	0.00006	0.00184
GO:0071229	cellular response to acid	0.00006	0.00184
GO:0071418	cellular response to amine stimulus	0.00006	0.00184

Table C.12: Static eQTL markers.

GO.ID	Term	p-value	FDR
GO:0032269	negative regulation of cellular protein metabolic process	< 0.00001	0.00026
GO:0006413	translational initiation	< 0.00001	0.00026
GO:0016311	dephosphorylation	< 0.00001	0.00026
GO:0009101	glycoprotein biosynthetic process	0.00001	0.00026
GO:0006417	regulation of translation	0.00002	0.00026
GO:0014070	response to organic cyclic compound	0.00002	0.00026
GO:0006366	transcription from RNA polymerase II promoter	0.00005	0.00131
GO:0009890	negative regulation of biosynthetic process	0.00007	0.00158
GO:0060070	canonical Wnt receptor signaling pathway	0.00010	0.00236
GO:0001892	embryonic placenta development	0.00011	0.00289

Appendix D

ImAP R code

D.1 Observed and expected allele frequencies

```
#########################################################################
## calculation of observed and expected genotype frequencies of
   single markers
#########################################################################

## the wrapper function 'array_calc' needs as input:
# chrom: a chromosome identifier
# data: genotype data of the trios on this chromosome in HAPPY
   file format (http://www.well.ox.ac.uk/happy/formats.shtml)
# snp: a table of the observed genotypes for each marker (one row
   could look like this: "AA" "AG" "GG")

array_calc <- function(chrom, data, snp){

## for which individuals do we have genotype datas for them and
   their parents:
all.nas <- apply(data[ , -(1:6)], 1, function(x) sum(is.na(x)))
fathers <- as.character(data[ , 3])
mothers <- as.character(data[ , 4])
fa.nas <- apply(data[fathers, -(1:6)], 1, function(x) sum(is.na(x)
   ))
names(fa.nas) <- rownames(data)
mo.nas <- apply(data[mothers, -(1:6)], 1, function(x) sum(is.na(x)
   ))
names(mo.nas) <- rownames(data)

ancestors <- which(((fa.nas == length(data[1, -(1:6)])) | (mo.nas
   == length(data[1, -(1:6)]))) & (all.nas < length(data[1, -(1:6)
   ])))
unknowns <- which(all.nas == length(data[1, -(1:6)]))
```

```
good.inds <- 1:nrow(data)
good.inds <- good.inds[-c(unknowns, ancestors)]

## create a table that contains the index of father and mother for
    all individuals:

parents.ind <- matrix(nrow=dim(data)[1], ncol=2)
for (i in 1:(dim(data)[1])){
    parents.ind[i,1] <- ifelse(data[i,3] != "0", which(data[,2] ==
        data[i,3]), NA)
    parents.ind[i,2] <- ifelse(data[i,4] != "0", which(data[,2] ==
        data[i,4]), NA)
}


## calculation of observed and expected frequencies of genotypes
   on single markers:

obs.freq <- array(0, dim=c(dim(data)[1], (dim(data)[2]-6)/2, 3))
exp.freq <- array(0, dim=c(dim(data)[1], (dim(data)[2]-6)/2, 3))

for (ind in good.inds){
    obs_exp.markers <- t(sapply(1:((dim(data)[2]-6)/2), all.geno,
       data, ind, snp, parents.ind))

    obs.markers <- obs_exp.markers[, 1:3]
    colnames(obs.markers) <- c("AA", "Aa", "aa")
    obs.freq[ind, , ] <- obs.markers

    exp.markers <- obs_exp.markers[, 4:6]
    colnames(exp.markers) <- c("AA", "Aa", "aa")
    exp.freq[ind, , ] <- exp.markers
}

## save data:

obs.chr1 <- obs.freq
exp.chr1 <- exp.freq

## store them separately since the expected arrays are needed for
   the permutations as well:
save(obs.chr1, file=paste("arrays", chrom, ".RData", sep=""))
save(exp.chr1, file=paste("exp_arrays", chrom, ".RData", sep=""))
}
```

```
######################
## internal functions:
######################

## function to find observed and expected genotype of an
   individual on one marker:

all.geno <- function(marker, data, ind, snp, parents.ind){
    geno.vec <- numeric(3)
    names(geno.vec) <- snp[marker, ]
    exp.alleles <- numeric(3)

    if (!any(is.na(data[ind, c(marker*2-1, marker*2) + 6]))){
        mother <- parents.ind[ind,2]
        father <- parents.ind[ind,1]
        mat.alleles <- as.character(data[mother, c(marker*2-1,
           marker*2)+6])
        pat.alleles <- as.character(data[father, c(marker*2-1,
           marker*2)+6])
        if (!any(is.na(mat.alleles)) & !any(is.na(pat.alleles))){
            exp.alleles <- prob.allele(mat.alleles, pat.alleles,
               snp[marker,])
            ind.geno <- paste(sort(as.character(data[ind, c(marker
               *2-1, marker*2) +6])), collapse="")
            geno.vec[ind.geno] <- 1
        }
    }
    return(c(geno.vec, exp.alleles))
}

## function for calculation of genotype frequencies on one marker:
prob.allele <- function(mat.alleles, pat.alleles, genotypes){
    allele.combis <- mapply(internal, rep(sort(mat.alleles), each
       =2), rep(sort(pat.alleles), 2))
    combi.freqs <- numeric(3)
    names(combi.freqs) <- genotypes
    combi.freqs[names(table(allele.combis))] <- table(
       allele.combis)/4
    return(combi.freqs)
}
```

```
## function used for allele combi creation:
internal <- function(x,y){
    res <- paste(sort(c(x,y)), collapse="")
    return(res)
}

#####################################################################
## application of the function to obtain observed and expected
   genotype matrices of original data:
## (should possibly be parallelized on a cluster of CPUs with the
   R package snow)
#####################################################################

# for (chrom in 1:19){
#   out = array_calc(chrom, data, snp)
# }
```

D.2 χ^2 statistic

```
#####################################################################
## calculation of ImAP test statistic
## function to calculate the chi-squared like test statistic based
    on the arrays of observed and expected genotype frequencies
   per marker:
## needs a input:
# chroms: a chromosome pair indicator
#####################################################################

perm.part <- function(chroms){
    load(paste("arrays", chroms[1], ".RData", sep=""))
    load(paste("exp_arrays", chroms[1], ".RData", sep=""))
    obs1 <- obs.chr1
    exp1 <- exp.chr1

    load(paste("arrays", chroms[2], ".RData", sep=""))
    load(paste("exp_arrays", chroms[2], ".RData", sep=""))
    obs2 <- obs.chr1
    exp2 <- exp.chr1

    ##### adjustment against selection pressure #####

    ## calculate observed and expected values for each marker over
        all individuals;
    ## for each genotype calculate ratio of observed and expected
```

```
   genotypes

obs.allinds1 <- apply(obs1, c(2,3), sum)
exp.allinds1 <- apply(exp1, c(2,3), sum)

ratio1 <- obs.allinds1/exp.allinds1
ratio1[is.na(ratio1)] <- 0

obs.allinds2 <- apply(obs2, c(2,3), sum)
exp.allinds2 <- apply(exp2, c(2,3), sum)

ratio2 <- obs.allinds2/exp.allinds2
ratio2[is.na(ratio2)] <- 0

## adjust expected frequencies with ratios:

exp1.adj <- exp1
exp2.adj <- exp2
for (i in 1:(dim(exp1.adj)[1])){
exp1.adj[i, , ] <- exp1.adj[i, , ]*ratio1
exp2.adj[i, , ] <- exp2.adj[i, , ]*ratio2
}

## normalise so that each individual's expected value over all
    genotypes on each marker is one:
corr.fac1 <- apply(exp1.adj, c(1,2), sum)
corr.fac2 <- apply(exp2.adj, c(1,2), sum)

exp.adj.prime1 <- exp1.adj
exp.adj.prime2 <- exp2.adj
for (i in 1:3){
exp.adj.prime1[ , , i] <- exp.adj.prime1[ , , i]/corr.fac1
exp.adj.prime2[ , , i] <- exp.adj.prime2[ , , i]/corr.fac2
}

exp.adj.prime1[which(exp.adj.prime1 == "NaN")] <- 0
exp.adj.prime2[which(exp.adj.prime2 == "NaN")] <- 0

##### calculation of observed and expected number of genotype
   combinations for each marker pair #####
##### including correction against allelic drift #####

obs.marker.geno <- array(0, dim=c(dim(obs1)[2], dim(obs2)[2],
   9))
```

```
    exp.marker.geno.adj <- array(0, dim=c(dim(exp1)[2], dim(exp2)
      [2], 9))

    slice <- 1
    for (i in 1:3){
    for (j in 1:3){
        obs.marker.geno[,,slice] <- t(obs1[,,i]) %*% obs2[,,j]
        exp.marker.geno.adj[,,slice] <- t(exp.adj.prime1[,,i]) %*%
            exp.adj.prime2[,,j]
        slice <- slice+1
    }
    }

    # dim of matrices: (nr. marker chrom 1) x (nr. marker chrom 2)
        x 9

    ##### calculation of chi-squared statistic for each marker
      combi #####

    ## alleles with zero expectation get zero contribution to the
      score:
    imp.combis.adj <- which(exp.marker.geno.adj == 0, arr.ind =
      TRUE)
    exp.marker.denom.adj <- exp.marker.geno.adj
    exp.marker.denom.adj[imp.combis.adj] <- 1/10000

    ## calculation of chi-squared statistic:
    chi2.single.adj <- (obs.marker.geno - exp.marker.geno.adj)^2/
      exp.marker.denom.adj
    chi2.scores.adj <- apply(chi2.single.adj, c(1:2), sum, na.rm=
      FALSE)

    save(chi2.scores.adj, file=paste("chi2scores", chroms[1], "vs"
      , chroms[2], ".RData", sep=""))

}

#########################################################################
## application of the function to calculate ImAP test statistic of
    original data:
## (should possibly be parallelized on a cluster of CPUs with the
   R package snow)
#########################################################################
```

```
# pairs <- cbind(rep(1:19, times=19:1), c(1:19, 2:19, 3:19, 4:19,
   5:19, 6:19, 7:19, 8:19, 9:19, 10:19, 11:19, 12:19, 13:19,
   14:19, 15:19, 16:19, 17:19, 18:19, 19))
#
# for (i in 1:nrow(pairs)){
#   out = perm.part(pairs[i,])
# }
```

D.3 Data preparation with trio package

```
#######################################################################
## script to infer all possible genotypes that could have been
   inherited to each child from trio package
#######################################################################

#######################################################################
## wrapper to use trio functions on HAPPY format genotype data
## this function requires the following input:
# chrom: chromosome indicator
# data: genotype matrix in HAPPY format
# maf: a matrix containing alleles and minor allele frequencies
   for all markers. Each row corresponds to a marker and would
   have entries like :min="G", maj="A", maf="0.2".
#######################################################################

require(trio)

trioData <- function(chrom, data, maf){

## convert markers into major allele (1) and minor allele (2):
for(x in 1:(nrow(maf))){
    data[, 6+((2*x-1):(2*x))][data[, 6+((2*x-1):(2*x))] == maf[x,"
       maj"]] <- 1
    data[, 6+((2*x-1):(2*x))][data[, 6+((2*x-1):(2*x))] == maf[x,"
       min"]] <- 2
}

## phenotype for all individuals is affected=2
data[,6] <- 2

for(i in 7:(ncol(data))) data[, i] <- as.integer(data[, i])

## what is the coding for missing values in trio? --> 0 !
data[is.na(data)] <- 0
```

```
## create trio data:
# Need to set replace=TRUE, since there might be genotyping errors
    or NAs.
foo <- trio.check(data, replace=TRUE)

## prepare data for logic regression, i.e. make pseudo controls
trio.bin = trio.prepare(trio.dat=foo)

save(data, foo, trio.bin, file=paste("trio_data", chrom, ".RData",
    sep=""))

}

############################
## application of function:
############################

# for (chrom in 1:19) {
#    out = trioData(chrom, data, maf)
# }
```

D.4 Generation of pseudo-controls

```
#########################################################################
## script to simulate pseudo-control data for the permutation p-
   value calculation of ImAP
#########################################################################

#########################################################################
## function to restructure haplotype data to spare one step in the
    observed matrix preparation
## (Has to be done only once, then the data are just loaded.)
## this function requires the following input:
# chrom: a chromosome indicator
# snp: a table of the observed genotypes for each marker (one row
   could look like this: "AA" "AG" "GG")
# maf: a matrix containing alleles and minor allele frequencies
   for all markers. Each row corresponds to a marker and would
   have entries like :min="G", maj="A", maf="0.2".
#########################################################################

pseudoControls <- function(chrom, maf, snp){
```

```
## transform outcome of trio function:
load(paste("trio_data", chrom, ".RData", sep=""))

quads <- trio.bin$bin[, -1]

for(x in 1:nrow(maf)){
    quads[, ((2*x-1):(2*x))][quads[, ((2*x-1):(2*x))] == 0] <-
       maf[x,"maj"]
    quads[, ((2*x-1):(2*x))][quads[, ((2*x-1):(2*x))] == 1] <-
       maf[x,"min"]
}

offspring <- rownames(foo$trio)[seq(3, nrow(foo$trio), 3)]
theo.geno <- vector(mode="list", length=length(offspring))
# some of the offspring names have a number appended that we
   dont want (since they are also parents...)
offspring <- sapply(offspring, substr, start=1, stop=12)
names(theo.geno) <- offspring

for (i in 1:length(offspring)) theo.geno[[i]] <- quads[(4*i-3)
   :(4*i), ]

## functions to adapt output of the trio package to the HAPPY
   format:
## the functions need as input:
# marker: vector of marker names for which the analysis should
   be done (possibly filtered for some quality criteria etc.)
# data: genotype matrix in HAPPY format
# haplos: a matrix containing the four possible genotypes of the
    an individual could have inherited from its parents in the
   coding from the R package trio. i.e. each column of haplo is
   a vector of length 4 giving the four possible codings of the
   dominant or recessive allele coding of a snp, the number of
   columns is twice the number of markers.
# theo.geno: a list of "haplos" matrices, one for each child.
# snp: a table of the observed genotypes for each marker (one
   row could look like this: "AA" "AG" "GG")

trafo.geno <- function(marker, haplos, snp){
    geno.vec <- numeric(3)
    names(geno.vec) <- snp[marker, ]
    ind.geno <- paste(sort(as.character(haplos[c(marker*2-1,
       marker*2)])), collapse="")
    geno.vec[ind.geno] <- 1
```

```
        return(geno.vec)
    }

    haplos.trafo <- lapply(theo.geno, function(ii) {
            tmp <- apply(ii, 1, function(pp) {
                    sapply(1:((dim(data)[2]-6)/2), trafo.geno, pp,
                      snp)
                    })
            out <- array(0, dim=c(4,(dim(data)[2]-6)/2, 3))
            for (i in 1:4){
                    out[i,,] <- matrix(tmp[,i], ncol=3, byrow=TRUE)
            }
    return(out)
    })

    save(haplos.trafo, file=paste("haplo_data", chrom, ".RData", sep
      =""))
}

################################
## application of the function
################################

# for (chrom in 1:19){
#   out = pseudoControls(chrom, maf, snp)
# }

#################################################
## simulation of offspring data from trio data
## this function requires the following input:
# perm: number of the permutation/pseudo-control
# chrom: chromosome indicator
# data: genotype matrix in HAPPY format
# snp: a table of the observed genotypes for each marker (one row
   could look like this: "AA" "AG" "GG")
#################################################

array_calc_pseudo <- function(perm, chrom, data, snp){
  # load pseudo-control genotypes and expected genotype matrix
  load(paste("haplo_data", chrom, ".RData", sep=""))
  load(paste("exp_arrays", chrom, ".RData", sep=""))

  ## draw random number fixing the haplotype vector for each
     individual:
```

```
  rand <- sample(1:4, size=length(haplos.trafo), replace=TRUE)
  names(rand) <- names(haplos.trafo)


  ## aggregate observed genotype array:

  # indicator about which parent marker info is available for each
      individual:
  pInfo <- apply(exp.chr1, c(1,2), sum)
  rownames(pInfo) <- rownames(data)

  obs.freq <- array(0, dim=c(dim(data)[1], (dim(data)[2]-6)/2, 3))

  for (ind in names(haplos.trafo)){
      obs.markers <- haplos.trafo[[ind]][rand[ind],,]
      colnames(obs.markers) <- c("AA", "Aa", "aa")
      # delete markers where we have no info about parents, i.e.
         expected frequencies:
      obs.markers[pInfo[ind,]==0,] <- c(0,0,0)
      obs.freq[which(rownames(data) == ind), , ] <- obs.markers
  }

  ## save data:

  obs.chr1 <- obs.freq
  save(obs.chr1, file=paste("sim", perm, "_arrays", chrom, ".RData
     ", sep=""))

}

#########################################################################
## application of the function to obtain observed and expected
   genotype matrices for nperms pseudo-controls:
## (should possibly be parallelized on a cluster of CPUs with the
   R package snow)
#########################################################################

# nperms <- 10000
#
# for (chrom in 1:19){
#   out = sapply(1:nperms, array_calc_pseudo, chrom, data, snp)
# }
```

D.5 p-values

```
##############################################
## calculation of ImAP permutation p-value
##############################################

perm.part <- function(pperm, chrom1, chrom2, chi2.orig){
    load(paste("sim", pperm, "_arrays", chrom1, ".RData", sep=""))
    load(paste("exp_arrays", chrom1, ".RData", sep=""))
    obs1 <- obs.chr1
    exp1 <- exp.chr1

    load(paste("sim", pperm, "_arrays", chrom2, ".RData", sep=""))
    load(paste("exp_arrays", chrom2, ".RData", sep=""))
    obs2 <- obs.chr1
    exp2 <- exp.chr1

    ##### adjustment against selection pressure #####

    ## calculate observed and expected values for each marker over
        all inds;
    ## for each genotype calculate ratio of observed and expected

    obs.allinds1 <- apply(obs1, c(2,3), sum)
    exp.allinds1 <- apply(exp1, c(2,3), sum)

    ratio1 <- obs.allinds1/exp.allinds1
    ratio1[is.na(ratio1)] <- 0

    obs.allinds2 <- apply(obs2, c(2,3), sum)
    exp.allinds2 <- apply(exp2, c(2,3), sum)

    ratio2 <- obs.allinds2/exp.allinds2
    ratio2[is.na(ratio2)] <- 0

    ## adjust expected frequencies with ratios:

    exp1.adj <- exp1
    exp2.adj <- exp2
    for (i in 1:(dim(exp1.adj)[1])){
    exp1.adj[i, , ] <- exp1.adj[i, , ]*ratio1
    exp2.adj[i, , ] <- exp2.adj[i, , ]*ratio2
    }
```

```
## normalise so that each individual's expected value over all
    genotypes on each marker is one:
corr.fac1 <- apply(exp1.adj, c(1,2), sum)
corr.fac2 <- apply(exp2.adj, c(1,2), sum)

exp.adj.prime1 <- exp1.adj
exp.adj.prime2 <- exp2.adj
for (i in 1:3){
exp.adj.prime1[ , , i] <- exp.adj.prime1[ , , i]/corr.fac1
exp.adj.prime2[ , , i] <- exp.adj.prime2[ , , i]/corr.fac2
}

exp.adj.prime1[which(exp.adj.prime1 == "NaN")] <- 0
exp.adj.prime2[which(exp.adj.prime2 == "NaN")] <- 0

##### calculation of observed and expected number of genotypes
    combinations for each marker pair #####
##### including correction against allelic drift #####

obs.marker.geno <- array(0, dim=c(dim(obs1)[2], dim(obs2)[2],
   9))
exp.marker.geno.adj <- array(0, dim=c(dim(exp1)[2], dim(exp2)
   [2], 9))

slice <- 1
for (i in 1:3){
for (j in 1:3){
    obs.marker.geno[,,slice] <- t(obs1[,,i]) %*% obs2[,,j]
    exp.marker.geno.adj[,,slice] <- t(exp.adj.prime1[,,i]) %*%
        exp.adj.prime2[,,j]
   slice <- slice+1
}
}

# dim of matrices: (nr. marker chrom 1) x (nr. marker chrom 2)
    x 9

##### calculation of chi2 statistic for each marker combi
   #####


## alleles with zero expectation get zero contribution to the
   score:
imp.combis.adj <- which(exp.marker.geno.adj == 0, arr.ind =
```

```
      TRUE)
    exp.marker.denom.adj <- exp.marker.geno.adj
    exp.marker.denom.adj[imp.combis.adj] <- 1/10000


    ## calculation of chi2 statistics:
    chi2.single.adj <- (obs.marker.geno - exp.marker.geno.adj)^2/
      exp.marker.denom.adj
    chi2.scores.adj <- apply(chi2.single.adj, c(1:2), sum, na.rm=
      FALSE)

    ##### 5. count scores exceeding the original score #####
    counter <- (chi2.scores.adj >= chi2.orig)
    return(counter)
}

#########################################################################
## application of the function
## (should possibly be parallelized on a cluster of CPUs with the
   R package snow)
#########################################################################

# nperms = 2000
#
# for (chrom1 in c(1:19)){
#     for (chrom2 in chrom1:19){
#
#       ## load original chi2 scores:
#       load(paste("chi2scores", chrom1, "vs", chrom2, ".RData",
   sep=""))
#       chi2.orig <- chi2.scores.adj
#
#       ## calculation of p-values
#       all.counts <- sapply(1:nperms, perm.part, chrom1, chrom2,
   chi2.orig)
#       pvals <- matrix(apply(all.counts, 1, sum)/nperms, nrow=
   nrow(chi2.orig), ncol=ncol(chi2.orig))
#
#       ## save results:
#       save(pvals, file=paste("perm_pvalues", chrom1, "vs",
   chrom2, ".RData", sep=""))
#
#     }
# }
```

Appendix E

Dynamic eQTL R code

E.1 Data preparation

```
##########################################
## load and prepare data for eQTL mappings
##########################################

## load expression data:
# all matrices are assumed to have the same number of columns (
    corresponding to strains) in the same order

stem <- read.table("expression_data/stem_norm_gene_GN_
    notimputed.txt", header = TRUE, sep = "\t", stringsAsFactors=
    FALSE)
progen <- read.table("expression_data/progen_norm_gene_GN_
    notimputed.txt", header = TRUE, sep = "\t", stringsAsFactors=
    FALSE)
mye <- read.table("expression_data/mye_norm_gene_GN_notimputed.txt
    ", header = TRUE, sep = "\t", stringsAsFactors=FALSE)
ery <- read.table("expression_data/ery_norm_gene_GN_notimputed.txt
    ", header = TRUE, sep = "\t", stringsAsFactors=FALSE)

stem <- data.matrix(stem)
progen <- data.matrix(progen)
ery <- data.matrix(ery)
mye <- data.matrix(mye)

## expression differences:
diffSP <- stem - progen
diffSP <- diffSP[, which(apply(diffSP, 2, function(x) sum(is.na(x)
    )) == 0)]
diffSE <- stem - ery
diffSE <- diffSE[, which(apply(diffSE, 2, function(x) sum(is.na(x)
    )) == 0)]
```

```
diffSM <- stem - mye
diffSM <- diffSM[, which(apply(diffSM, 2, function(x) sum(is.na(x)
  )) == 0)]
diffPE <- progen - ery
diffPE <- diffPE[, which(apply(diffPE, 2, function(x) sum(is.na(x)
  )) == 0)]
diffPM <- progen - mye
diffPM <- diffPM[, which(apply(diffPM, 2, function(x) sum(is.na(x)
  )) == 0)]
diffEM <- ery - mye
diffEM <- diffEM[, which(apply(diffEM, 2, function(x) sum(is.na(x)
  )) == 0)]


## concatenated gene expressions over conditions for simultaneous
  eQTL mapping:
allTypes <- cbind(apply(stem, 2, function(y) y-rowMeans(stem,
  na.rm = TRUE)), apply(progen, 2, function(y) y-rowMeans(progen,
   na.rm = TRUE)), apply(ery, 2, function(y) y-rowMeans(ery,
  na.rm = TRUE)), apply(mye, 2, function(y) y-rowMeans(mye, na.rm
   = TRUE)))
allTypes <- allTypes[, which(apply(allTypes, 2, function(x) sum(
  is.na(x))) == 0)]

## unweighted mean expression:
stem <- apply(stem, 2, function(y) y-rowMeans(stem, na.rm = TRUE))
progen <- apply(progen, 2, function(y) y-rowMeans(progen, na.rm =
  TRUE))
ery <- apply(ery, 2, function(y) y-rowMeans(ery, na.rm = TRUE))
mye <- apply(mye, 2, function(y) y-rowMeans(mye, na.rm = TRUE))

meanAll <- 0.25*(stem + progen + ery + mye)
meanAll <- meanAll[, which(apply(meanAll, 2, function(x) sum(is.na
  (x))) == 0)]

## weighted mean mean expression:
meanExprGene <- read.table(file="expression_data/logweighted_
  centered_mean_expr_allstrains.txt", header= TRUE, sep = "\t",
  stringsAsFactors=FALSE)

### load genotype matrix x:
load("R_files/BXD_genotypes_filtered.RData")
```

```
## preparation of predictor matrix for simultaneous eQTL mapping (
   Figure 4.2 in thesis)
## i.e. multiplication of genotypes according to number of
   conditions/cell states and addition of cell state indicators
xCTInd <- x[, colnames(allTypes)]

aa <- sum(apply(stem, 2, function(x) sum(is.na(x))) == 0)
bb <- sum(apply(progen, 2, function(x) sum(is.na(x))) == 0)
cc <- sum(apply(ery, 2, function(x) sum(is.na(x))) == 0)
dd <- sum(apply(mye, 2, function(x) sum(is.na(x))) == 0)

xCTInd <- rbind(xCTInd, c(rep(1, aa), rep(0, bb+cc+dd)), c(rep(0,
   aa), rep(1, bb), rep(0, cc+dd)), c(rep(0, aa+bb), rep(1, cc),
   rep(0, dd)), c(rep(0, aa+bb+cc), rep(1, dd)))
```

E.2 eQTL mapping

```
#####################################################
## Random Forest for eQTL mapping
## y=gene expression vector
## x=genotype matrix
## ntree, mtry, nodesize as in randomForest package

rf.ff <- function(y, x, ntree, mtry=floor(ncol(x)/3), nodesize=5)
   {
  require(randomForest)
  rf = randomForest(y = y, x = x, ntree = ntree, mtry=mtry,
     nodesize=nodesize)
  sf = rfsf(rf)
}

#############################################################
## function for extracting selection frequencies from an RF
## rf=the RF from which selection frequencies are desired

rfsf = function(rf){
 vu = varUsed(rf)
 sf = vu/sum(vu)
 names(sf) = rownames(rf$importance)
 return(sf)
}
```

```
#####################################
## eQTL mapping
## (exemplary for simultaneous eQTL)
#####################################

## cluster setup
require(snow)

ncpus=as.integer(Sys.getenv(c("RMPI_NCPUS")))
nslaves=ncpus-1
cl <- makeCluster(nslaves, type = "MPI")

clusterExport(cl, "rf.ff")
clusterExport(cl, "rfsf")
clusterEvalQ(cl, library(randomForest))


## RF mapping:
outSim <- parApply(cl, allTypes, 1, function(y, xs) {rf.ff(y=
   as.numeric(y),x=xs, ntree=20000, mtry=70, nodesize=3)}, xs=t(
   xCTInd))


## randomisations for p-value calculation:

randSim <- vector(mode='list', length=10)

for (i in 1:length(randConc_CTInd)){
  randSim <- parApply(cl, allTypes[, sample(1:ncol(allTypes),
     replace=FALSE)], 1, function(y, xs) {rf.ff(y=y,x=xs, ntree
     =20000, mtry=70, nodesize=3, keep.forest=TRUE, geneName=NULL)
     }, xs=t(xCTInd))
}

stopCluster(cl)
mpi.quit()


## p-value and FDR calculation:
## (based on functions in Section "Functions for p-value
   calculation")

pSim <- pvalWrapper(orig=outSim, rand=randSim, rcl = FALSE, ccl =
   FALSE, rfit = TRUE, cfit = FALSE, use.method="mix", perc1=0.99,
```

```
    perc2=0.99, combine=FALSE, nnodes=2, clustmeth="kmeans",
  nclust=NULL, correct=TRUE)

pSim[pSim == 0] <- min(pSim[pSim != 0])
fdrSim <- matrix(p.adjust(pSim, method="BH"), nrow=nrow(pSim))
```

E.3 Detection of conditional eQTL

```
#####################################################################
## take the significant results from simultaneous eQTL mapping to
## distinguish static and conditional eQTL
#####################################################################

eqtl_fdr <- which(fdrSim[1:849,] < 0.1, arr.ind=TRUE)

## filter out markers that are in LD and regulate the same gene

#############################################################
## function to filter significant eQTL - target pairs
## for markers in LD
## eqtltab=2-column matrix containing positions of
##         significant eQTL - target gene pairs
## eqtlMat=eQTL matrix
## posInfo=matrix with 4 columns containing position
##         infos of each genotype marker
##         (chromosome, start, end, center of marker region)
#############################################################

LDfilterEQTL <- function(eqtltab, eqtlMat, posInfo){
  geneList <- split(eqtltab[,1], eqtltab[,2])
  tmp <- lapply(1:length(geneList), function(i, gl){
    # for each gene:
    if (length(geneList[[i]]) > 1){
      msort <- sort(geneList[[i]])
      dists <- outer(msort, msort, '-')
      dists <- dists[cbind(2:length(msort), 1:(length(msort)-1))]
      breaks <- which(dists > 2)
      int <- cbind(c(1,breaks+1), c(breaks, length(msort)))
      # check if the intervals span chromosome boundaries:
      int <- lapply(as.data.frame(t(int)), function(x){
        if (posInfo[msort[x[1]], 1] != posInfo[msort[x[2]], 1]){
          if (x[2] == (x[1] + 1)){
            out <- rbind(c(x[1], x[1]), c(x[2], x[2]))
          } else{
```

```
            tmp <- ((x[1]+1):(x[2]-1))[which(posInfo[msort[(x
              [1]+1):(x[2]-1)],1] != posInfo[msort[x[1]], 1])[1]]
            out <- rbind(c(x[1], tmp-1), c(tmp, x[2]))
          }
          return(out)
        } else {return(x)}
      })
      int <- do.call(rbind, int)
      mReduced <- apply(int,1, function(x) msort[x[1]:x[2]][
        which.min(eqtlMat[msort[x[1]:x[2]], as.numeric(names(gl)[
        i])])])
      return(mReduced)
    } else {
      return(geneList[[i]])
    }
  }, gl=geneList)
  eqtltab_red <- cbind(as.integer(do.call('c', tmp)), as.integer(
    rep(names(geneList), times=sapply(tmp, length))))
  return(eqtltab_red)
}

## application of the function to eQTL results:
eqtl_fdr_filtered <- LDfilterEQTL(eqtl_fdr, fdrSim, mSubInfo)
rm(eqtl_fdr)

## define genes for which anova is done
eQTLgenes <- unique(eqtl_fdr_filtered[,2])

############################################################
## conduct Anova on significant eQTL - target gene pairs
############################################################

############################################################
## function to fit one model including the marker and
## all cell types for one particular eQTL target pair
## pair=vector of marker and gene position in eQTL matrix
## exprs=gene expression matrix
## predictor=predictor matrix
##           (genotypes + cell state indicators)
## parents=rows in predictor matrix
##         containing the cell state indicators
############################################################
```

```
epiLMcomb <- function(pair, exprs, predictor, parents){
    ctvar = rep(1, length(predictor[parents[1],]))
    ctvar[predictor[parents[2],] == 1] <- 2
    ctvar[predictor[parents[3],] == 1] <- 3
    ctvar[predictor[parents[4],] == 1] <- 4
    data=data.frame(e = exprs[pair[2],], m = as.factor(predictor[
       pair[1], ]), ct= as.factor(ctvar))
    full <- lm(e ~ m + ct + m:ct, data=data)
    red <- lm(e ~ m + ct, data=data)
    out <- anova(red,full)[2,6]
    return(out)
}

## applicatoin of function to all significant eQTL - target gene
   pairs, FDR calculation:
anovaComb <- apply(eqtl_fdr_filtered, 1, epiLMcomb, allTypes,
   xCTInd, parents=850:853)
anovaFDR <- p.adjust(anovaComb, method='BH')

########################
## Post-hoc Wald tests:
########################

#################################################################
## function for Wald test on one particular eQTL target pair
#################################################################

findIAwald <- function(pair, exprs, predictor, parents){
    require(contrast)
    ctvar = rep(1, length(predictor[parents[1],]))
    ctvar[predictor[parents[2],] == 1] <- 2
    ctvar[predictor[parents[3],] == 1] <- 3
    ctvar[predictor[parents[4],] == 1] <- 4
    out <- which(predictor[pair[1], ] == 2)
    if (length(out) > 0){
     data=data.frame(e = exprs[pair[2],-out], m = as.factor(
        predictor[pair[1], -out]), ct= as.factor(ctvar[-out]))
    } else {
     data=data.frame(e = exprs[pair[2],], m = as.factor(predictor[
        pair[1], ]), ct= as.factor(ctvar))
    }
   full <- lm(e ~ m + ct + m:ct, data=data)
   eachMarkerEffect <- contrast(full,
   list(ct = levels(data$ct), m = "3"),
```

```
  list(ct = levels(data$ct), m = "1"))$Pvalue
  return(p.adjust(eachMarkerEffect, 'bonferroni'))
}

## run the test on the 'significant' pairs:
signA <- which(anovaFDR < 0.5)
contrTestWald <- t(apply(eqtl_fdr_filtered[signA,], 1, findIAwald,
    exprs=allTypes, predictor=xCTInd, parents=850:853))

## aggregation of results:
tmp_wald <- rowSums(contrTestWald < 0.005)
signSim <- cbind(eqtl_fdr_filtered, fdrSim[eqtl_fdr_filtered],
  anovaFDR, matrix(NA, nrow=nrow(eqtl_fdr_filtered), ncol=5))
colnames(signSim) <- c('marker', 'gene', 'sim_fdr', 'anova_fdr', '
  #ctIA', 'stem', 'progen', 'ery', 'mye')
signSim[signA, '#ctIA'] <- tmp_wald
signSim[is.na(signSim[, '#ctIA']), '#ctIA'] <- 0
signSim[signA, 6:9] <- contrTestWald
```

E.4 Functions for p-value calculation

```
#########################################################
## Functions needed to calculate different versions
## of p-values from RF selection frequencies
## options:
## - clustering of genes or markers
##   based on quantiles of SF
## - fitting a function to tail of SF distribution
## - empirical/fitted/mixed p-value calculation
#########################################################

###############################################
## function to cluster selection frequencies
## based on their distributions:
###############################################

## Input:
## rand: a list of randomized SF matrices (probably from
   randomising the samples/strains and then repeating the RF)
## nnodes: the number of nodes that can be used in parallel for
   the clustering
## clcut: height at which to cut hierarchical clustering tree in
   order to get row clusters
```

```
## Output:
## rowclust: the clustering of rows

selfreq_cl <- function(rand, nnodes=2, clustmeth="kmeans", nclust=
   NULL){
        ## combine the random data in columns:
        randMat <- do.call('cbind', rand)
        if (is.null(nclust)) nclust <- nrow(randMat)/4

        ######################################
        ## cluster rows by their
        ## quantile distribution:
        ## (use amap for parallel clustering)

        require(amap)
        rowQ <- t(apply(randMat, 1, quantile, seq(0,1,0.01)))
        if (clustmeth == "kmeans"){
                rowQcl <- kmeans(rowQ, centers=nclust)
                rowCl <- rowQcl$cluster
        } else{
                rowQcl <- hcluster(rowQ, method="euclidean", link=
                   "complete", nbproc=nnodes)
                rowCl <- cutree(rowQcl, k=nclust)
        }
        return(rowCl)
}

############################################
## function to fit distribution to the tail
## of the random selection frequencies:
############################################

## Input:
## rand: a list of randomized SF matrices (probably from
   randomising the samples/strains and then repeating the RF)
## clMemb: a vector giving the cluster membership for each row/
   column of the SF matrix
## perc: a numeric value between 0 and 1 which defines the top
   percent of the data to which the function is fitted

## Output:
## rowfit: the parameters of the inverse gamma distribution fitted
    to each row (cluster)
```

```
selfreq_fit <- function(rand, clMemb, perc1=0.99){
    ## combine the random data in rows and columns:
    randMat <- do.call('cbind', rand)

    ## function to be optimised during fitting, depends on the
       percentage of data we fit on:
    objfExp <- function(p,x,y){
        e = (y-dexp(x+p[1],rate=p[2]))
        sum(e^2)
    }

    f2exp_row_cl <- t(sapply(sort(unique(clMemb)), function(cl) {
        tmp <- as.numeric(randMat[cl,])
        histPoints <- hist(tmp, breaks="fd", plot=FALSE)
        xhist <- histPoints$mids[histPoints$mids > quantile(tmp,
           perc1)]
        yhist <- histPoints$density[histPoints$mids > quantile(tmp
           , perc1)]
        p1 <- 0.5
        p2 <- 15
        out <- optim(p = c(p1, p2), objfExp, gr = NULL, x=xhist, y
          =yhist, method = "Nelder-Mead", control=list(reltol=1e
          -15, trace=0))$par
        if (out[2] < 500){
          p2 <- 20
          out <- optim(p = c(p1, p2), objfExp, gr = NULL, x=xhist,
              y=yhist, method = "Nelder-Mead", control=list(reltol
            =1e-15, trace=0))$par
        }
        return(out)
        }
    ))
return(rowfit=f2exp_row_cl)

}

##########################################################
## function to calculate row and column p-values
## from both the fitted and empirical distribution
## based on fitting only the tail of the distribution
## and switching from empirical to analytical
## p-value at a defined quantile of the data:
##########################################################
```

```
## Input:
## selfreq: a matrix of SF
## rand: a list of random SF matrices to be pasted together, the p
   -value calculation will always be done on the rows!
## clMemb: a vector giving the cluster membership for each row/
   column of the SF matrix
## paraFit: the matrix of fitted parameters for each row/column (
   cluster)
## perc: a numeric value between 0 and 1 which defines the top
   percent of the data for which the p-value is calculated from
   the fitted distribution

## Output:
## rowPcl: matrix with p-values

mixP <- function(selfreq, rand, clMemb, paraFit, perc2){
    randMat <- do.call('cbind', rand)
    rowPcl <- matrix(NA, ncol=ncol(selfreq), nrow=nrow(selfreq))

    for (cl in sort(unique(clMemb))){

        low <- which(selfreq[which(clMemb == cl), ] <= quantile(
           as.numeric(selfreq[which(clMemb == cl), ]), perc2))
        high <- which(selfreq[which(clMemb == cl), ] > quantile(
           as.numeric(selfreq[which(clMemb == cl), ]), perc2))

        rowPcl[which(clMemb == cl),][low] <- 1 - ecdf(randMat[
           which(clMemb == cl), ])(selfreq[which(clMemb == cl), ][
           low])

        rowPcl[which(clMemb == cl),][high] <- 1 - pexp(selfreq[
           which(clMemb == cl), ][high]+paraFit[cl, 1], rate=
           paraFit[cl, 2])
    }
    return(rowPcl)
}

##################################################
## function to calculate row and column p-values
## from the fitted distribution only:
##################################################

## Input:
## selfreq: a matrix of SF
```

```
## clMemb: a vector giving the cluster membership for each row/
   column of the SF matrix
## paraFit: the matrix of fitted parameters for each row/column (
   cluster)

## Output:
## rowPcl: matrix with p-values

## function for row p-values:
fitP <- function(selfreq, clMemb, paraFit){
        require(pscl)
        rowPcl <- matrix(NA, ncol=ncol(selfreq), nrow=nrow(selfreq
           ))

        for (cl in sort(unique(clMemb))){
                rowPcl[which(clMemb == cl),] <- 1 - pigamma(
                   selfreq[cl, ], alpha=paraFit[cl, 1], beta=
                   paraFit[cl, 2])
        }
        return(rowPcl)
}

#################################################
## function to calculate row and column p-values
## from the empirical distribution only:
#################################################

## Input:
        ## selfreq: a matrix of SF
        ## rand: a list of random SF matrices to be pasted
           together, the p-value calculation will always be done
           on the rows!
        ## clMemb: a vector giving the cluster membership for each
            row/column of the SF matrix
## Output:
        ## rowPcl: matrix with p-values
## function for row p-values:
empP <- function(selfreq, rand, clMemb){
        randMat <- do.call('cbind', rand)
        rowPcl <- matrix(NA, ncol=ncol(selfreq), nrow=nrow(selfreq
           ))

        for (cl in sort(unique(clMemb))){
                rowPcl[which(clMemb == cl),] <- 1 - ecdf(randMat[
```

```
                    which(clMemb == cl), ])(selfreq[which(clMemb ==
                      cl), ])
                # set 0 p-values to the minimum:
                rowPcl[which(clMemb == cl),][rowPcl[which(clMemb
                  == cl),] == 0] <- 1/length(randMat[which(clMemb
                    == cl),])
        }
        return(rowPcl)
}

##########################################
## function to calculate combined p-values
## from row and column p-values:
##########################################

## Input:
## rowP, colP: matrices of row/column p-values

## Output:
## pComb: matrix with combined p-values

combiP <- function(rowP, colP){

        ## combine p-values with Fisher:
        pComb <- 1 - pchisq(-2*(log(rowP) + log(colP)), df=4)
        colnames(pComb) <- colnames(rowP)
        rownames(pComb) <- rownames(rowP)
        return(pComb)
}

########################################
## function to correct the p-values for
## the background (random) distribution:
## this is done in order to transform the
## background distribution into a uniform
########################################

## Input:
## origP: Matrix with p-values that are to be corrected
## randP: list of matrics with p-values from random SF. These are
   used to calculate the background distribution

## Output:
## pCor: matrix of corrected p-values
```

```
trafoP <- function(origP, randP){
        randCDF <- ecdf(unlist(randP))
        pCor <- matrix(randCDF(origP), nrow=nrow(origP))
        colnames(pCor) <- colnames(origP)
        rownames(pCor) <- rownames(origP)
        return(pCor)
}

#####################################################
## wrapper function that integrates the
## different approaches for p-value calculation:
#####################################################

## Input:
## orig: a matrix of SF for which p-values are to be calculated
## rand: a list of randomized SF matrices (probably from
   randomising the samples/strains and then repeating the RF)
## rcl, ccl: logical, if TRUE the rows/columns of the combined SF
   matrix will be clustered based on their quantiles in order to
   pool values for the fitted/empirical cdf
## rfit, cfit: logical, if TRUE calculation are done row/column-
   wise (also for clusters)
## use.method: one of "fit", "emp", "mix"
## combine: logical, if TRUE the p-values from row- and column-
   wise calculations will be combined with the "Fisher method" (
   without applying the chi-squared distribution)
## nnodes: the number of nodes that can be used in parallel for
   the clustering
## clcutRow/clcutCol: height at which to cut hierarchical
   clustering tree in order to get row/column clusters. Default
   parameters were chosen so that the rows and columns in the
   example used for exploration had comparable cluster sizes.
   However, the cutoffs might differ considerably for different
   data sets.

## Output:
## pMatCor: final p-value matrix

pvalWrapper <- function(orig, rand, rcl = TRUE, ccl = FALSE, rfit
   = TRUE, cfit = FALSE, use.method="fit", perc1=0.99, perc2=0.95,
    combine=FALSE, nnodes=2, clustmeth="kmeans", nclust=NULL,
   correct=TRUE){
```

```
## clustering:
if (rcl){
        rowCl <- selfreq_cl(rand, nnodes=nnodes, clustmeth
          =clustmeth, nclust=nclust)
} else {rowCl <- 1:nrow(orig)}

if (ccl){
        colCl <- selfreq_cl(lapply(rand, t), nnodes=nnodes
          , clustmeth=clustmeth, nclust=nclust)
} else {colCl <- 1:ncol(orig)}

## fitting an empirical distribution to the clusters:
if (rfit & use.method!="emp") rowFit <- selfreq_fit(rand,
  rowCl, perc=perc1)
if (cfit & use.method!="emp") colFit <- selfreq_fit(lapply
  (rand, t), colCl, perc=perc1)

## calculation of original and random p-values:
if (use.method=="fit" & rfit==TRUE){
        pRow <- fitP(orig, clMemb=rowCl, paraFit=rowFit)
        pRowRand <- lapply(rand, fitP, clMemb=rowCl,
          paraFit=rowFit)
}
if (use.method=="fit" & cfit==TRUE){
        pCol <- t(fitP(t(orig), clMemb=colCl, paraFit=
          colFit))
        pColRand <- lapply(lapply(rand, t), fitP, clMemb=
          colCl, paraFit=colFit)
        pColRand <- lapply(pColRand, t)
}

if (use.method=="emp" & rfit==TRUE){
        pRow <- empP(orig, rand, clMemb=rowCl)
        pRowRand <- lapply(rand, empP, rand=rand, clMemb=
          rowCl)
}
if (use.method=="emp" & cfit==TRUE){
        pCol <- t(empP(t(orig), lapply(rand, t), clMemb=
          colCl))
        pColRand <- lapply(lapply(rand, t), empP, rand=
          lapply(rand, t), clMemb=colCl)
        pColRand <- lapply(pColRand, t)
}
```

```
        if (use.method=="mix" & rfit==TRUE){
                pRow <- mixP(orig, rand, clMemb=rowCl, paraFit=
                    rowFit, perc=perc2)
                pRowRand <- lapply(rand, mixP, rand=rand, clMemb=
                    rowCl, paraFit=rowFit, perc=perc2)
        }
        if (use.method=="mix" & cfit==TRUE){
                pCol <- t(mixP(t(orig), lapply(rand, t), clMemb=
                    colCl, paraFit=colFit, perc=perc2))
                pColRand <- lapply(lapply(rand, t), mixP, rand=
                    lapply(rand, t), clMemb=colCl, paraFit=colFit,
                    perc=perc2)
                pColRand <- lapply(pColRand, t)
        }

        if (combine==TRUE){
                pMat <- combiP(rowP=pRow, colP=pCol)
                pMatRand <- combiP(rowP=pRowRand, colP=pColRand)
        }
        if (combine==FALSE & rfit==TRUE & cfit==FALSE){
                pMat <- pRow
                pMatRand <- pRowRand
        }
        if (combine==FALSE & rfit==FALSE & cfit==TRUE){
                pMat <- pCol
                pMatRand <- pColRand
        }

        ## correction for background p-value distribution:
        if (correct){
                pMatCor <- trafoP(origP=pMat, randP=pMatRand)
        } else {
                pMatCor <- pMat
        }
        return(pMatCor)
}

#########################################################
## function to calculate "exact p-values"
## based on a larger number of randomisations
## without any pooling over predictors or traits and
## without any fitting of analytical distributions:
#########################################################
```

```
## Input:
## orig: a matrix of SF for which p-values are to be calculated
## rand: a list of randomized SF matrices (probably from
   randomising the samples/strains and then repeating the RF)

## Output:
## pMat: final p-value matrix

exactP <- function(orig, rand){
        randMat <- array(NA, dim=c(nrow(orig), ncol(orig), length(
           rand)))
        for (i in 1:length(rand)) randMat[,,i] <- rand[[i]]
        pMat <- apply(randMat, c(1,2), function(x) x >= orig)/
           length(rand)
        return(pMat)
}
```