Statistical methods for the analysis of the genetics of gene expression

Matthias Alexander Heinig

April 13, 2011

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Gutachter:
Prof. Dr. Martin Vingron
Prof. Dr. Norbert Hübner

1. Referent: Prof. Dr. Martin Vingron
2. Referent: Prof. Dr. Norbert Hübner

Tag der Promotion: 13. Dezember 2010

Preface

The work that led to this thesis is part of several collaborative projects in which I participated. For the sake of clarity this thesis presents all results that led to the conclusions drawn and individual contributions will be detailed here. For each project I mention publications, highlight my contributions and acknowledge the work of my collaborators.

Star project The work presented in section 3.3.1 was published in *Nature Genetics* [1] by the *Star Consortium*. It was a large EU funded project initiated by Norbert Hübner and led by Kathrin Saar aiming to assess genetic variability among rat strains and to provide haplotype and genetic maps. My contribution was the design of the strategy for the construction of the maps and its implementation and application to three different populations of rats. Moreover I performed the comparative analysis of these maps. I would like to acknowledge Herbert Schulz, Kathrin Saar, Norbert Hübner, Edwin Cuppen, Richard Mott, Domenique Gaugier and Marié-Thérèse Bihoreau for discussions and guidance and Denise Brocklebank for discussion and contributions to the implementation of the analysis. I would like to thank Oliver Hummel and Herbert Schulz for the preprocessing of genotype data.

Ephx2 project Analysis and results presented in sections 3.4.1, 4.1 and 4.2.3 were pulished in *Nature Genetics* [2]. This project was designed by Judith Fischer, Jan Monti and Norbert Hübner and carried out in the lab of Norbert Hübner mainly by Judith Fischer and Svetlana Paskas. My contribution to this project was the eQTL analysis, analysis of the genetic effect on the arachidonic acid pathway (section 4.1) and the transcription factor binding site analysis of the Ephx2 promoter (section 4.2.3). In addition I participated in the analysis of the physiological data together with Judith Fischer. I would like to thank Judith Fischer for leading the project and producing all the data. I would like to thank Norbert Hübner for discussions and giving me the opportunity to work on this data. Finally I would like to acknowledge Martin Vingron for his ideas on the pathway analysis (section 4.1). I have presented the methodology of this part of the project as a talk at the *International conference on Genetics* 2008 in Berlin.

Euratools project The work presented in sections 3.3 and 4.3 was published in *Nature* [3]. The data were generated as part of the *Euratools* consortium – an EU funded large integrated project. I contributed the eQTL analysis across seven tissues (section 3.3) and the design of the transcription factor and co-expression analysis in sections 4.3.1 and 4.3.2 which form the basis of this study. Together with Maxim Rotival and Enrico Petretto, I performed the co-expression analysis in the human eQTL data set (section 4.3.6). Analysis of the influence of the human chromosome 13 locus on the network (section 4.3.7) was carried out in parallel by myself, Anja Bauerfeind and Leonardo Bottolo. Analysis of the *EBI2 cis*-eQTL (section 4.3.7) was performed in parallel by myself and Anja Bauerfeind. Translation to human T1D

data (4.3.8) was performed in collaboration with Chris Wallace, David Clayton and John Todd with contributions of Anja Bauerfeind and myself. In particular I would like to acknowledge that Chris Wallace developed the theory for the extensions of enrichment analysis to account for confounding factors (sections 2.2.3 and 4.3.8). I am grateful to the *Euratools* consortium for providing me with a short term fellowship that got me started on this project during my stay in London. I thank Norbert Hübner, Enrico Petretto and Stuart Cook for initiating this project and finding the right collaboration partners to make it a success. I would like to thank Enrico Petretto for his constant support and advice in this project. I thank Helge Roider for his assistance with the PASTAA analysis. I am grateful to Carola Rintisch and Svetlana Paskas who performed the experiment to confirm my TFBS predictions. My special thanks go to Anja Bauerfeind for help and advice for the human eQTL analysis and genotype imputation. I would like to thank Sarah Langley who performed the cell type specificity analysis in section 4.3.3 and Leonardo Bottolo for the analysis in section 4.3.4. I would like to thank Stefan Blankeberg and his coworkers for providing us access to eQTL data from the Gutenberg Heart Study, Francois Cambien on behalf of the *Cardiogenics* consortium for access to eQTL data from the *Cardiogenics* study and John Todd for access to the human T1D data.

The analysis presented in section 4.4 has not been published yet, but a manuscript is currently being submitted. The data analysed there are the same as presented above. I have designed and implemented the analysis. I would like to thank Enrico Petretto and Mario Falchi for stimulating discussions over coffee that shaped my ideas. I would like to thank Martin Vingron for his ideas and advice on writing the manuscript. I acknowledge numerous discussions about linear models with Hughes Richard. Finally I would like to thank Norbert Hübner who prompted me to evaluate phenotype data in this context.

sTRAP project The method presented in section 4.2 was published in *Human Mutation* [4]. The project was a close collaboration between Thomas Manke and myself. We designed and performed the analysis together. I would like to thank Helge Roider for providing me the *C*-code of *TRAP* which allowed me to implement the *tRap* *R*-package and the *sTRAP* webserver.

Acknowledgements I would like to thank both my supervisors Norbert Hübner and Martin Vingron for offering me the opportunity to get the best of both worlds, experimental and computational by co-supervising my thesis. I am grateful for the many fruitful discussions, the ideas and support I have enjoyed from both during my thesis. I would like to thank Norbert Hübner for embedding me into the very stimulating environment of the *Euratools* consortium which allowed me to participate in many collaborations. I would like to thank Martin Vingron for creating an extraordinary working atmosphere in his group. I am very grateful to all members of the Vingron and Hübner groups and all PhD students of both institutes many of whom became friends. Last but not least I would like to thank my parents for their support and Kristina for her patience.

Berlin, September 2010 Matthias Heinig

Contents

1 Introduction

Phenotypic diversity results from a complex interplay of genetic and environmental factors. The goal of genetic research is to identify the heritable factors that underlie the phenotypic diversity. Ultimately, molecular genetics aims to link these factors to molecular mechanisms that explain the diversity of phenotypic traits. Since many common human diseases have a genetic component [5, 6, 7] this knowledge is highly valuable for molecular medicine since insight into the molecular basis of disease processes is the prerequisite for the development of therapies. Moreover genetic studies facilitate animal and plant breeding and allow insights to be gained into the principles of evolution.

Modern genetic research began with the studies of Mendel [8]. Based on his work on heritable traits of peas he postulated two laws: 1) the law of segregation and 2) the law of independent assortment, which were later shown to be the consequence of meiosis [9]. In his laws he postulated the existence of heritable factors – nowadays called "genes" or more general "loci". Different individuals have different variants of a gene which are called alleles. These variants constitute the basis of the genetic diversity among individuals. Physically, genes are encoded on DNA molecules that are called chromosomes in higher organisms.

Animals and plants are diploid organisms. This means, each chromosome is present in two homologous copies, each of which has been inherited from one individual of the previous generation. Sexual reproduction of animals and plants involves three major steps on the cellular level which are important for the transmission of genetic information. A special form of cell replication called meiosis takes place in the germline of the parents. It consists of two successive cell division, but only one DNA replication (Figure 1.1). In the first step (meiosis I) each of the homologous copies is replicated to form sister chromatids leading to bundles of four chromatids. Homologous recombination takes place between the non-sister chromatids which can lead to a crossover of the homologous copies. During this process different alleles of the genes can be exchanged between the homologous copies of the chromosomes. Subsequently the four chromatids are separated into haploid gametes where the second cell division (meiosis II) does not involve replication. The choice of which (recombined) parental chromosome is passed to which daughter cell is random, leading to the independent assortment observed by Mendel. Two gametes (sperm and egg) form a zygote (fertilised egg) that develops into a new diploid organism. The two mechanisms of recombination and independent assortment lead to a very large number of new combinations of genetic information.

The mechanism of independent assortment explains why genes located on different chromosomes are inherited independently of each other. Genes on the same chromosome can lead to exceptions to the law of independent assortment. Whether they are inherited together or not is a matter of how many recombinations take place between them. We speak of genetic linkage if two genes are preferentially inherited together indicating that these genes are in close proximity on the chromosome and therefore less recombinations take place. Genetic mapping

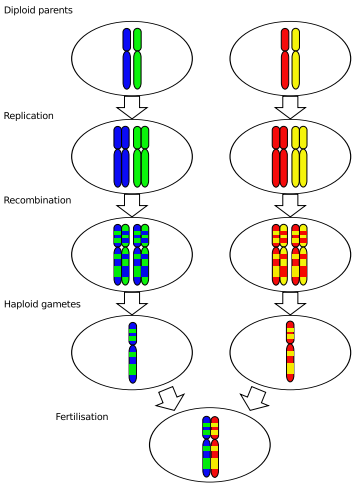

Figure 1.1: Sexual reproduction of diploid organisms.

takes advantage of these exceptions and uses the recombination rate observed in a population as distance to arrange the genes linearly on the chromosome. Since the recombination rate alone does not lead to an additive distance measure the average number of recombinations per meiosis measured in centi Morgan (cM) were proposed [10] as additive distance. However it does not generally correspond to the physical distance between genes measured in mega base pairs (Mb). Empirical data shows that genetic and physical distance tend to be linearly related with the exception of certain recombination hotspots [11] (see also section 3.3.1).

In the early days of genetics Mendel's laws were much debated [12], since they could only explain categorical phenotypes like flower colour or seed shape but did not seem to provide an explanation for quantitative phenotypes like body height. Study of continuous traits gave rise to its own field termed biometrics. The objective of this research was to establish whether certain continuous traits were heritable or not by means of segregation analysis [12]. Even if quantitative phenotypes are found to be heritable, often their distribution in the population resemble the Gaussian normal distribution instead of a bimodal distribution, which would be expected when assuming Mendelian inheritance of a single genetic factor. So, how can quantitative traits be explained by Mendel's laws? In 1918 Fisher [13] described how the observed distribution of quantitative phenotypes can arise from a mixture model of multiple genetic factors (polygenic) with small effects (see section 1.2). This reconciliation of Mendel's laws with the polygenic model of quantitative traits together with the chromosome theory of Thomas Morgan constitutes the theory of classical genetics.

Nowadays, molecular genetics provides the tools to apply this theory in an unprecedented scale. Molecular marker maps enabled the genetic mapping of quantitative trait loci (QTL) in biological model systems for human disease (section 1.2). Facilitated through recent technological advances in sequencing and genotyping millions of molecular markers in large population samples a number of systematic efforts are now underway to map human variation [14] and the genetic basis for many common diseases in genome-wide association studies (GWAS) [6, 15] (see

section 1.3).

A particular advantage of animal model systems for the study of human disease is that environmental factors can be controlled and thus the phenotypic diversity observed is to a large degree due to genetic variation between individuals. Now, with the feasibility of GWAS, some argue that the human is the new model system [16]. It is true that candidate gene identification can also take advantage of human GWAS, however functional studies on the organismal level are only possible in model systems. Therefore a translational approach that combines genetics in humans and animal models constitutes an actively pursued strategy [17] to identify loci and the associated disease causing biological mechanism.

GWAS and QTL studies typically result in a region or a set of markers associated with the trait of interest. These molecular markers are mostly non-functional variations that are in linkage disequilibrium (LD) with the unknown functional variants. Occasionally, variations reside in protein-coding regions where they can directly affect the function of the encoded protein. A classical example describes the molecular basis of sickle cell disease as a mutation of an amino acid in the β-globin chain, which causes polymerisation of haemoglobin, reduced elasticity of red blood cells, but also protection against malaria [18]. Unsurprisingly, most variations have been observed in non-coding regions since they make up most of the genome. Although the understanding of non-coding sequences is far from complete it is known that they harbour a variety of gene regulatory elements [19]. Therefore variations in non-coding sequences might alter these regulatory elements and the regulatory network. The functional consequences of regulatory variations are, however, more difficult to predict and validate, because the regulatory code is much more complex and flexible than the genetic code. For example, the dysregulation of α-globin is known to cause α thalassemia; a reduction in functional haemoglobin, and also increased protection against malaria [20].

The regulatory effects of sequence variations can be measured systematically at the level of gene expression using the transcript level of each individual gene as a quantitative trait (see section 3.1) giving rise to expression QTLs (eQTLs). If disease associated variations also show an effect on gene expression it is likely that they tag a regulatory variant. Therefore the analysis of the genetics of gene expression provides excellent means for the identification of regulatory variants. This thesis presents a set of tools and novel approaches for the analysis of the genetics of gene expression. In section 3.4 strategies will be presented for the identification of disease genes using the regulatory variation hypothesis. If refined data about additional sequence variants is available the question of which molecular mechanism is in action at the regulatory variation can be addressed (section 4.2).

Since most biological processes require the coordinated action of sets of functionally related genes the analysis of gene expression networks in segregating populations can provide new insights into disease processes [21, 22, 23] (see chapter 4). The naturally occurring genetic variation can be used to identify gene networks and the loci underlying their regulation (see section 4.3). So far, the proportion of the phenotypic variance explained by disease associated variations is rather small [24] which is evidence for a more complex mode of inheritance. In this context gene expression networks identified in biological model systems can provide a useful functional context for the interpretation of GWAS results (see section 4.3). In most gene expression network analysis [25, 26, 27, 22, 28, 21] the focus has been to identify networks of co-regulated genes, where the expression is determined by the sequence variant that has been found to be

associated with the disease phenotype. What has been neglected so far, is that the observed phenotype could also be a consequence of genotype dependent perturbations of co-expression. In section 4.4 we will introduce an approach to identify genotype dependent perturbations of the co-expression network and show how this knowledge can be used to identify more complex modes of inheritance such as epistatic interactions.

1.1 Objective and structure of the thesis

Objective The goal of this thesis is to design and apply statistical tools that allow for a functional interpretation of the results of genetic mapping experiments. These tools will be applied to analyse two eQTL data sets generated in experimental crosses of laboratory rat strains to gain insights into the mechanisms of gene regulation that underly phenotypic traits. Ultimately, the goal is to translate findings from animal models of human disease to interpret results of genetic studies in human case control cohorts.

Structure In chapter 1 we provide an introduction to the biological mechanisms underlying the heredity of phenotypic traits. Basic principles and the study design of experimental crosses for the analysis of quantitative traits will be introduced. Chapter 2 describes the statistical tools used in the studies presented here.

Chapter 3 introduces the concept of expression quantitative trait locus (eQTL) mapping. The basic problem that led to this paradigm is that genetic studies typically result in large chromosomal regions that are statistically associated with a phenotypic trait but do not provide a molecular explanation of this association. Ultimately the goal of any genetic study is the identification of genes underlying the phenotype and a mechanistic model of how causal sequence variations lead to the development of this phenotype. Here we explain how the eQTL paradigm may be used to identify candidate genes by postulating that sequence variation affects gene regulation rather than protein function since protein coding sequence constitutes only a small part of the genome. In section 3.3 the principles of eQTL mapping are illustrated using data from a set of recombinant inbred lines. The integration of eQTL data and physiological data is illustrated in two case studies. The first (section 3.4.1) resulted in the identification of a candidate gene for heart failure [2], the second (section 3.4.2) led to the identification of a candidate gene for systolic blood pressure [29].

Chapter 4 explains four approaches for the functional interpretation of eQTL data and constitutes the main contribution of this thesis. Section 4.1 shows how functional annotation can be used to identify genetic markers that influence functionally related gene expression networks. *Cis* and *trans*-acting gene regulatory mechanisms that lead to genotype dependent expression patterns (eQTLs) are investigated in the two following sections. Both are concerned with the role of transcription factors (TFs). Section 4.2 shows how information about sequence variations and a biophysical model of TF – DNA interaction can be used to identify both the most likely *cis*-regulatory elements in the promoters of eQTL transcripts and the TF that is most likely the upstream regulator of the transcript. Section 4.3 deals with the role of TFs as mediators of *trans*-acting eQTLs. Finally section 4.4 describes an approach to analyse genotype dependent perturbations of gene expression networks solely on the level of expression data.

Each section presenting analyses of experimental data is self contained with methods, results and discussion. Chapter 5 summarises the analyses presented at a more general level and puts them into broader context.

1.2 Genetic mapping of quantitative trait loci (QTL)

Most quantitative traits are the result of an interplay of (multiple) genetic and environmental factors. The genetic mapping of quantitative traits is concerned with the identification of underlying genes or loci called quantitative trait loci (QTL) by the use of molecular genetic markers. From a statistical point of view the analysis of quantitative traits in experimental crosses derived from inbred strains of model organisms and human genome wide association studies (GWAS) of unrelated individuals can be approached with the same set of standard statistical tools. First we will discuss the general setup derived from experimental crosses and show how it applies to human GWAS. However, some subtle differences remain and will be discussed in the subsections of this chapter. Details of the statistical methods are given in the next chapter. Finally the experimental crosses used in this thesis will be described.

Here we study only diploid organisms that have two homologous copies of each chromosome, but the theory can be extended to polyploid organisms. In sexual reproduction one copy of each chromosome is transmitted by the father, one by the mother (Figure 1.1). Inbred strains are populations of model organisms that are homozygous throughout the genome – i.e. the genetic information (alleles) from both parents is identical for every gene. In rodents this is achieved by recurrent brother-sister mating. Suppose that the father has alleles a_1a_1 and the mother has alleles a_2a_2. Every individual in the first generation (F1) will have both alleles a_1a_2. After the first meiosis the distribution of genotypes a_1a_1, a_1a_2, a_2a_2 in the F2 generation is $1/4, 1/2, 1/4$. After n generations of brother sister mating the fraction of heterozygous is $(1/2)^n$ which is negligible after eight to ten generations.

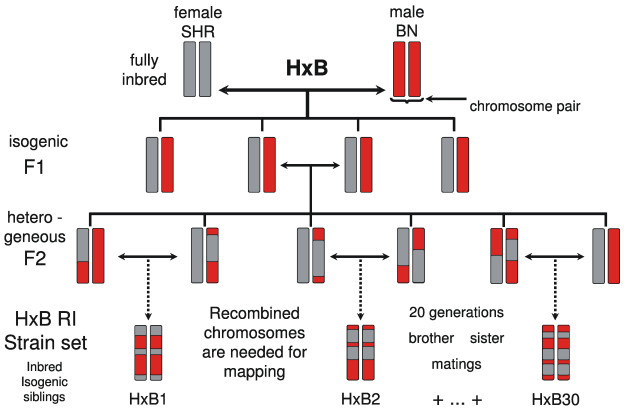

Figure 1.2: **Breeding schema of the BXH/HXB recombinant inbred strains.**

All experimental mating schemes involve the cross of two divergent inbred strains to produce a heterozygous F1 generation. For the F2 intercross, individuals from the F1 generation are mated to generate recombinations in both maternal and paternal gametes. This gives rise to the above mentioned distribution of genotypes which we will call ($A = a_1a_1, H = a_1a_2, B = a_2a_2$). Recombinant inbred (RI) strains are created by first performing a F2 intercross followed by several generations of brother sister mating to derive new inbred strains with homozygous genotypes ($A = a_1a_1, B = a_2a_2$) at almost all loci (Figure 1.2). This way one can establish a renewable biological resource of recombined offsprings of two inbred strains.

Statistically a QTL can be described as an unobserved categorical variable that affects the value of the quantitative trait. For an additive trait with one QTL, this can be formalised by a three component mixture model where each component represents the probability density function (PDF) of the trait y for one genotype:

component	genotype	proportion	PDF
1	A	$p_A = 0.25$	$\mathcal{N}(y, \mu_A, \sigma^2)$
2	H	$p_H = 0.50$	$\mathcal{N}(y, \mu_H = 0.5(\mu_A + \mu_B), \sigma^2)$
3	B	$p_B = 0.25$	$\mathcal{N}(y, \mu_B, \sigma^2)$

Here the PDF is the normal distribution

$$\mathcal{N}(y, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(y-\mu)^2}{2\sigma^2} \tag{1.1}$$

with mean μ and variance σ^2. In case of RI strains, the proportions are $p_A = 0.5, p_H = 0, p_B = 0.5$.

Genetic mapping of molecular markers allows to determine the genotypes of mostly non-functional genetic variations. Since inbred strains show a high degree of linkage disequilibrium, we can expect that there is a low recombination rate r between neighbouring loci. This also applies to known (non-functional) genetic markers and unknown functional QTL. Suppose the two parental strains have genotypes $m_1a_1//m_1a_1$ and $m_2a_2//m_2a_2$ where the // separates the chromosome pair and m denotes the marker allele and a the QTL allele respectively. In the F2 generation the recombined alleles m_1a_2 or m_2a_1 occur with probability r. Thus, if the marker is close to the QTL (r is small) then the genotype groups at the marker will clearly have different means. The basic idea of genetic mapping is to reverse the statement: if the genotype groups at the marker show clearly different means, we have evidence for a QTL close to the marker.

Whether the different genotypes of a marker indeed induce different means can be tested with analysis of variance (ANOVA) or standard regression tools [30, 31] which will be discussed in detail in section 2.1. As an example consider the model $y = \mu + \beta_i + \epsilon_j, i \in \{A, H, B\}$. The linear modelling framework can easily account for additional markers or other co-variates. If genotype data for some individuals is missing, the mixture model framework above and data from flanking markers can be used to infer the most likely genotypes [32, 33] via the expectation maximisation algorithm (EM). This technique estimates the joint probability of genotypes and phenotypes which is particularly useful when the genetic map is sparse. Then positions between markers without genotype information but with a given map distance can be imputed, which is also known as *interval mapping*.

Conversely, the alternative approach suggested by [34] assumes that the genotype probabilities at the marker are known and need not be estimated jointly. Conditional on the genotype probabilities the expected trait value is $\mu = p_A\mu_A + p_H\mu_H + p_B\mu_B$. This leads to the linear model

$$y = (p_A, p_H, p_B) \begin{pmatrix} \mu_A \\ \mu_H \\ \mu_B \end{pmatrix} + \epsilon \tag{1.2}$$

with the error term $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Assuming known genotype probabilities, the maximum likelihood estimators of the model can easily be found using the least squares method (see section 2.1.1) and is therefore called *marker regression mapping*. The main advantage over interval mapping is computation time. The EM algorithm generally needs several iteration steps for convergence whereas the regression is solved in a single step. When a dense map of markers with few missing genotypes is available this method performs equally well as interval mapping [35].

Most approaches use a likelihood based statistic to measure the strength of association. The log odds score (LOD) and the likelihood ratio statistic LRS are two closely related and frequently used statistics. They are defined [35] as

$$LOD = \log_{10} \frac{L_{full}}{L_{reduced}} \tag{1.3}$$

$$LRS = 2\log_e \frac{L_{full}}{L_{reduced}} \tag{1.4}$$

where L_{full} is the maximum likelihood under the QTL model and $L_{reduced}$ is the maximum likelihood under the null model of no QTL. Therefore their relation is obtained by simply changing the base of the logarithm

$$LRS = 2\log_e(10) \cdot LOD \approx 4.6 \cdot LOD \ . \tag{1.5}$$

Now with the tools to check single markers for the presence of a QTL at hand the strategy of genome wide QTL detection can be formulated as a model selection problem. Each marker or sets of markers constitute different competing QTL models that have to be evaluated according to certain criteria in order to select the best QTL model for a trait. The simplest strategy is the single marker analysis where the QTL likelihood profile is computed across the whole genetic map using marker regression or interval mapping and likelihood peaks that exceed a genome wide threshold of significance are selected. No analytical solutions for the genome wide distribution of test statistics are known, therefore simple permutation [36] or bootstrapping [37] strategies have been proposed. In both strategies data following the null hypothesis of no QTL is generated by randomisation of the trait data. This randomised trait data is then analysed using the genetic map many times (e.g. $n = 10^6$) and the genome wide maximum of the test statistic of each run is used to create an empirical cumulative distribution function.

Several methods like *composite interval mapping* [38, 39] or *multiple QTL mapping* [37] have been proposed to identify QTL models that include multiple markers. They adapt one of the general strategies of model selection described in the following paragraph to make them suitable for genetic mapping. In particular the correlation between genetic markers due to LD has to

be handled carefully. In the theory of linear models, there are three standard model selection strategies: forward selection, backward elimination or stepwise selection. As a selection criterion one can use the F-test to compare nested models (see section 2.1.2), the Akaike information criterion (AIC) [40], the Bayesian information criterion (BIC) [41] or the Cp criterion [42]. In the forward selection, one starts with an empty model and adds the variable that improves the model most significantly according to the criterion. Conversely, in backward elimination one starts with the full model and removes variables that contribute least to the model according to the criterion. The stepwise procedure is a mixed strategy where addition and removal of variables is performed in alternation [43].

Other model selection procedures avoid the problems of collinearity of markers by directly penalising model complexity by parameter shrinkage, like ridge regression [44] or lasso regression [45] (see also section 2.1.4). Several authors have proposed to use these methods for the inference of QTL models [46, 47, 48, 49].

1.3 Genome wide association studies

Before the sequencing of the human genome, only a limited number of genes could be identified by positional cloning and family based linkage analysis [50, 15]. Alternatively, population based association studies of candidate genes with a case – control design were used to identify genetic variants that alter the risk of disease. Power calculations of linkage and association analysis showed that the required sample size for association studies to identify common genetic variants with a allele frequency $> 5\%$ and modest effects in a genome wide screen [51] was small enough ($n > 2000$) to be feasible. The so called "common disease, common variant" hypothesis [51, 52, 50] has led to the paradigm of genome wide association studies (GWAS) and the initiation of the Hapmap project [53, 11]. Usually the experimental design of GWAS is a case – control study. The statistical association of genetic markers and the binary outcome is most frequently assessed using logistic regression [54] . However, also quantitative traits can be analysed in a genome wide association study [55]. In this case the methodology to assess the presence of a QTL close to a marker is the same as for experimental crosses (section 1.2). The genome wide strategy [51] to identify associated loci or QTL is to apply single marker tests to all markers that have been genotyped and correct for multiple testing using false discovery rate (FDR) methods [56].

1.4 The BXH/HXB recombinant inbred strains

In this work we study the BXH/HXB recombinant inbred strains which are derived from a cross of the spontaneously hypertensive rat SHR.Ola abbreviated as SHR and the brown norway rat BN.Lx/Cub abbreviated as BN [57, 58]. SHR is a widely studied model system for human hypertension and shows features of the metabolic syndrome [59, 60, 61, 62, 63] while BN represents a normotensive control strain. The BN.Lx/Cub is a congenic strain carrying the polydactyly-luxate syndrome which leads to malformations of the hindlimb [64]. As mentioned in section 1.2 the RI strains are generated by mating the two parental inbred strains in order to obtain recombined F2 animals. Each of the F2 animals carries a unique combination of maternal and

paternal genes because of the independent segregation and recombination of the homologous chromosomes during gametogenesis in the F1 generation. Pairs of F2 animals are selected randomly for inbreeding by brother sister mating for at least 20 generations. In the BXH/HXB RI strains gender reciprocal crossing was performed which provides two sets of strains with different sources of mitochondrial DNA and Y chromosomes. Strains designated by HXB are offsprings of female SHR and male BN rats and vice versa for BXH.

RI strains have several advantages over single generation intercross or backcross progeny: (1) they are homozygous across the whole genome (2) individuals of one strain are genetically identical which allows for biological replication, (3) makes phenotyping and genotyping cumulative and (4) allows to investigate different developmental stages. The phenotype data that has been accumulated and genetically mapped spans a wide range from blood pressure parameters [65, 66, 67], heart weight [65], lipid levels [68, 69], renal phenotypes [70] and metabolic parameters [62, 58]. The QTLs identified in these studies however span large chromosomal regions of sizes up to several Mb, leaving the underlying genes and mechanisms unknown. Sections 3.1 and 3.4 will present strategies for the identification of these genes by the use of gene expression data in genetic systems.

1.5 The F2 intercross between SHHF and SHRSP

In a study of heart failure we analysed the spontaneously hypertensive heart failure rat (SHHF) - an inbred, genetically homogenous rat model, which mirrors the human situation of hypertension-associated heart failure [71]. Human heart failure is an epidemiologically important disease with $> 30\%$ mortality at one year after diagnosis [72, 73, 74]. It is a complex phenotype resulting from an interplay of genetic and environmental risk factors. Most heart failure patients have impaired systolic function with a reduced ejection fraction [74]. SHHF rats not only develop heart failure late in life after high blood pressure and left ventricular hypertrophy have developed [75], but also exhibit many of the associated transcriptional and metabolic features of the human disease [76, 77, 78]. The model's genetic propensity is underscored by the fact that a closely related strain, the stroke-prone spontaneously hypertensive rat (SHRSP) does not develop heart failure despite similarly elevated blood pressures. We conducted a co-segregation analysis in F2 hybrids bred from SHHF and SHRSP, thus removing blood pressure and left ventricular hypertrophy (LVH) as confounding variables which have been hindering genetic analysis in humans. In section 3.4 we will show how an integrated analysis of physiological and gene expression data from this F2 intercross led to the identification of a candidate gene for heart failure. Section 4.2.3 describes the identification of the *cis*-regulatory element and the upstream regulator of *Ephx2*. In section 4.1.1 we show how *a priori* knowledge of functional gene sets can be used to interpret the genetics of expression data and relate the candidate gene to the biochemical pathway it is involved in.

2 Statistical tools

This chapter briefly introduces common statistical concepts and methods that will be used throughout the remainder of this thesis.

2.1 Linear models

The general linear model describes the statistical relationship between the (dependent) random variable Y on the (independent) non-random variables $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$. Often $Y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^p$. The linear model decomposes Y into a deterministic part and a random part

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \tag{2.1}$$

with the unknown parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ from the parameter space $\Omega_{\boldsymbol{\beta}}$ and the random error ϵ which has mean zero. In the general form, no assumption about the distribution of $\boldsymbol{\epsilon}$ is made.

In order to estimate the parameters of this *population model* we need to obtain observations of populations represented by the model. If we observe a sample $\{(y_i, \mathbf{x}_i), i = 1 \ldots n\}$ of size n we can write down the *sample model* as a set of n equations in matrix notation

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}, \mathrm{cov}(\boldsymbol{\epsilon}) = \Sigma. \tag{2.2}$$

which defines the general linear sample model [79].

Here we consider a subset of linear models where we assume that the error terms ϵ_i are independent and identically distributed (iid)

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\mu} = \mathbf{0}, I\sigma^2), \sigma^2 \text{ unknown} \tag{2.3}$$

with mean zero and the same unknown variance σ^2. I denotes the identity matrix.

2.1.1 Estimation of parameters

The unknown parameters $\boldsymbol{\beta}, \sigma$ are estimated by the method of maximum likelihood. Since $\mathbf{Y}$ is distributed $\mathcal{N}(X\boldsymbol{\beta}, I\sigma^2)$ the likelihood function is

$$\mathrm{L}(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, X) = (\frac{1}{2\pi\sigma^2})^{n/2} \exp(-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})'(\mathbf{y} - X\boldsymbol{\beta})). \tag{2.4}$$

We log transform equation 2.4 and obtain the partial derivates

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathrm{logL}(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, X) = \frac{1}{\sigma^2}(X'\mathbf{y} + X'\boldsymbol{\beta}) \tag{2.5}$$

$$\frac{\partial}{\partial \sigma^2} \mathrm{logL}(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, X) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - X\boldsymbol{\beta})'(\mathbf{y} - X\boldsymbol{\beta}). \tag{2.6}$$

The maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ of $\boldsymbol{\beta}$ and σ^2 are found by setting the partial derivatives to zero

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y} \tag{2.7}$$

$$\hat{\sigma}^2 = \frac{1}{n}\mathbf{y}'(I - X(X'X)^{-1}X')\mathbf{y}. \tag{2.8}$$

2.1.2 Hypothesis testing

Using the estimated parameters and the assumptions about the distribution of the error terms one can derive hypothesis tests about the parameters $\boldsymbol{\beta}$. Here we are interested in the hypothesis H_0 that $\boldsymbol{\beta}$ is constrained to a subspace ω of the parameter space Ω

$$\Omega = \{(\boldsymbol{\beta}, \sigma^2) | \boldsymbol{\beta} \in \mathbb{R}^p, \sigma^2 > 0\} \tag{2.9}$$

$$\omega = \{(\boldsymbol{\beta}, \sigma^2) | \boldsymbol{\beta} \in \mathbb{R}^p, \sigma^2 > 0, \mathbf{l}'\boldsymbol{\beta} = l_0\}. \tag{2.10}$$

where a certain linear combination of $\boldsymbol{\beta}$ is constant. We note that this is a special case of the more general hypothesis of $H\boldsymbol{\beta} = \mathbf{h}$. The two parameter spaces define nested models where the model with parameter space ω is a special case of the model with parameter space Ω. The general form of hypothesis testing in nested models is to test the reduced model $\mathbf{l}'\beta = l_0$ versus the full model using the generalised likelihood ratio test [79]. Assuming that the error terms $\epsilon_k \sim \mathcal{N}(0, \sigma)$ are independent and have the same variance σ the test statistic

$$F = \frac{(\mathbf{l}'\hat{\beta} - l_0)^2}{\hat{\sigma}(\mathbf{l}'(X'X)^{-1}\mathbf{l})}, \tag{2.11}$$

with

$$\hat{\beta} = (X'X)^{-1}X'\mathbf{y}, \tag{2.12}$$

$$\hat{\sigma} = \frac{1}{n-p}\mathbf{y}'(I - X(X'X)^{-1}X')\mathbf{y} \tag{2.13}$$

follows an F distribution with 1 and $(n-p)$ degrees of freedom. Therefore the hypothesis H_0 that $\mathbf{l}'\boldsymbol{\beta} = l_0$ is rejected at significance level α if F is larger than the $1-\alpha$ quantile of the F distribution with 1 and $(n-p)$ degrees of freedom.

2.1.3 The multiple linear regression model

As stated above, the general linear model assumes that the independent variables $\mathbf{x}$ are fixed non-random variables. In most real world problems, this assumption does not hold true and the independent variables are also random. The multiple linear regression model extends the general linear model to account for random effects in the independent variables [79]. The fixed variables $\mathbf{x}$ of length p are replaced by the random variables $\boldsymbol{\chi}$. The multiple linear regression model assumes a joint distribution of the $p+1$ random vector $\binom{Y}{\boldsymbol{\chi}}$. Then the conditional expectation of Y is modelled as a linear function of $\boldsymbol{\chi}$:

$$\mathbb{E}(Y|\boldsymbol{\chi}) = \boldsymbol{\mu}_Y(\boldsymbol{\chi}) = \boldsymbol{\chi}\boldsymbol{\beta}. \tag{2.14}$$

The multiple linear regression model is specified by

$$\mathbf{Y} = \boldsymbol{\mu}_Y(\boldsymbol{\chi}) + \epsilon, \mathbb{E}(\epsilon) = 0, \mathrm{var}(Y) = \sigma^2. \tag{2.15}$$

If the joint distribution of $\binom{Y}{\boldsymbol{\chi}}$ is assumed to be a $(p+1)$-variate normal distribution or the conditional distribution of $(Y|\boldsymbol{\chi})$ is assumed to be normal, the estimation of parameters and testing of hypothesis described for the general linear model can also be applied to the multiple linear regression model [79].

2.1.4 Model selection via lasso

In genome-wide studies the number of variables p measured for each individual is very large, often exceeding the sample size n. Therefore methods to select the relevant variables x_i to model the dependent variable Y are needed. For linear models there are a range of model selection methods like forward, backward and stepwise selection or coefficient shrinkage methods like ridge regression [44] or the lasso [45]. Both ridge regression as well as the lasso method are based on penalising the regression coefficients and therefore lead to models with fewer $\beta_i \neq 0$ which is desirable for QTL models and has been used for genetic mapping [46, 47, 48, 49]. We are using lasso regression, since its solutions tend to be more sparse than ridge regression [80]. The lasso estimate for a centred matrix X is defined as

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{lasso} &= \underset{\boldsymbol{\beta}}{\mathrm{argmin}}\ \mathbf{y} - X\boldsymbol{\beta} \\ &\quad \text{subject to } \boldsymbol{\beta}'\mathbf{1} \leq \lambda. \end{aligned} \tag{2.16}$$

This penalises the L_1-norm of $\boldsymbol{\beta}$ whereas the ridge regression would penalise the L_2-norm. If λ is larger than the L_1-norm of the least squares solution $\hat{\boldsymbol{\beta}}^{LS}$, the lasso solution is identical to the least squares solution. The solution of the constrained minimisation problem can be found by introducing Lagrange multipliers and solving the Lagrangian.

What remains open is the choice of a good shrinkage parameter λ. Toward that end cross validation [80] can be used to estimate the prediction errors for a range of possible λ values and select the one with the lowest error estimate. Hastie *et al.* propose to use the most parsimonious model within one standard cross validation error of the minimum [80].

2.2 Functional enrichment analysis

Many biological high throughput experiments such as differential gene expression analysis using DNA microarrays (section 3.2) but also expression QTL studies result in ranked lists of genes. Interpretation of these list is facilitated by functional annotation of genes. The most comprehensive such annotation is the gene ontology (GO) [81]. Other sources of functional annotation are more specific such as the Kyoto Encyclopedia of genes and genomes (Kegg) [82] which focuses mainly on metabolic pathways. Since many biological processes involve not only single genes but whole pathways (e.g. signalling cascades) the remainder of this section discusses how results of high throughput experiments can be analysed in terms of sets of functionally related genes.

2.2.1 Exact test on a contingency table

The most commonly used functional enrichment analysis [83] is based on the comparison of two sets [84]. One set is derived from the experiment, e.g. significantly differentially expressed genes from a microarray experiment represented by the binary random variable D which is one if the gene is differentially expressed. The other is given by the functional annotation represented by the random variable S. The two sets are cross tabulated in a contingency table:

	differential ($D = 1$)	non-differential ($D = 0$)	total
in gene set ($S = 1$)	n_{11}	n_{12}	n_{1+}
not in gene set ($S = 0$)	n_{21}	n_{22}	n_{2+}
total	n_{+1}	n_{+2}	n

In order to establish a functional enrichment we need to assess whether the set of differentially expressed genes contains more genes from the given functional gene set than expected by chance. Intuitively, one would ask whether the proportion of differentially expressed genes in the set n_{11}/n_{1+} is significantly different from the background ratio n_{21}/n_{2+}? To formalise this question assume that the genes are sampled form a bivariate population with the joint distribution $\mathbb{P}(D, S)$. The question is rephrased to whether D and S are independent

$$H_0 : \mathbb{P}(D = i, S = j) = [\sum_{s=0}^{1} \mathbb{P}(D = i, S = s)][\sum_{d=0}^{1} \mathbb{P}(D = d, S = j)], \forall i, j \in \{0, 1\} \tag{2.17}$$

For large sample sizes this hypothesis can be assessed using Pearson's χ^2 test on contingency tables [85]. The test statistic is defined as

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{j+}/n} \tag{2.18}$$

and follows a χ^2-distribution with two degrees of freedom in our example. In the more general case with r and c categories in each variable the degrees of freedom are defined by $(r-1)(c-1)$.

For small sample sizes *Fisher's exact test* which conditions on the marginal counts is used [85]. In the special case of a 2×2 table it is equivalent to the hypergeometric test:

$$\mathbb{P}(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{21}}}{\binom{n}{n_{+1}}}. \tag{2.19}$$

H_0 is rejected with significance level α if $\mathbb{P}(x > n_{11}) = \sum_{x>n_{ij}} \mathbb{P}(x) < \alpha$. This small sample size version is the most commonly used enrichment test.

2.2.2 Gene set enrichment analysis

The major drawback of the contingency table approach is that the result of the biological experiment has to be discretised into two sets. In the analysis of differential gene expression this is mostly achieved by setting a threshold on the significance of differential expression. Usually

the threshold is set very stringently to account for the multiple testing of thousands of genes. Biologically however, it is of interest to detect small but consistent changes in functionally related gene sets. More specifically one is interested to find a gene set S that contains genes that are more often differentially expressed than the rest of the genome or to reject the hypothesis

$$H_0 : \text{genes} \in S \text{ are as often differentially expressed as genes} \notin S. \tag{2.20}$$

In order to detect such events a method called *Gene set enrichment analysis* (GSEA) has been proposed [86]. The input to GSEA is a ranked list L of genes obtained by evaluating the experiment on a single gene level. This information is aggregated on the level of a priori defined gene sets by defining an enrichment score (ES). The ES of set S is the maximum of a weighted running sum along L which is increased whenever a gene from S is encountered and decreased otherwise. The weighting depends on the degree of differential expression. The ES is a weighted version of a Kolmogorov-Smirnov-like statistic. The significance of the ES is determined by a permutation procedure that randomises the outcome variable which was initially used to determine differential expression.

2.2.3 Extensions to gene set enrichment analysis

Various extensions to the original GSEA method have been proposed [87, 88, 89] and reviewed [90, 91]. These methods can be classified according to the null hypotheses they are testing, the statistics they are using, whether they are global tests e.g. [89] or they aggregate the single gene statistics and the way that significance is assessed.

Alternative null-hypotheses

The different possible null-hypotheses are discussed at length in [90, 91]. A brief summary is given here.

$$Q_1 : \text{genes} \in S \text{ are as often differentially expressed as genes} \notin S. \tag{2.21}$$

Q_1 is called the "competitive null hypothesis" and corresponds to the null hypothesis of the original GSEA method. It compares the differential expression of genes in the set S against all other genes. Note that the sampling unit for Q_1 are the genes and single gene statistics are fixed. On these grounds, the GSEA approach has been criticised [90] because it tests Q_1 but the sampling in the permutations is done on the individuals.

$$Q_2 : \text{no gene} \in S \text{ is differentially expressed.} \tag{2.22}$$

In contrast, the so called "self-contained null hypothesis" Q_2 only considers genes within the set. It assesses the differential expression of the genes in S compared to random outcome vectors. So the sampling unit here are the individuals and gene set membership is fixed.

$$Q_3 : FDR \text{ estimates for genes} \in S \text{ are the same as estimates for all genes.} \tag{2.23}$$

Q_3 is called the "nested null hypothesis" because it compares differential expression of genes in S to the differential expression of all genes both inside and outside of S [92, 91] .

Alternative test statistics

Of the many alternative test statistics we only discuss [87] and [93] that have been used in this work. Tian et al. have proposed a method to assess null hypothesis Q_2 [87] . In a scenario where the outcome variable is a binary classification the single gene statistic is defined as the t-score. In order to aggregate single gene statistics on the level of the set S the average is used. Since the sampling unit are the individuals, significance is based on permutations of the outcome variable. An alternative to the GSEA statistic for the test of Q_1 has been suggested by [93]. In order to test the difference of the distributions of single gene statistics within the set S and outside of the set they propose to use the Wilcoxon rank sum test. Since we have applied this statistic in section 4.3.8 we will define it here.

Suppose that a sample of N observations is ranked according to a measure of association with an experimental condition. Let $(r_1, r_2, \ldots, r_N)$ denote the ranks of the observations. Suppose we have two groups, say the funcionally related gene set S consisting of n_1 genes and the rest of the genome $\bar{S}$ consisting of $n_2 = N - n_1$ genes. Two equivalent test statistics, the Wilcoxon rank sum and the Mann-Whitney statistic are computed as follows

$$W = \sum_{i \in S} r_i \tag{2.24}$$

$$U = W - n_1(n_1 + 1)/2. \tag{2.25}$$

For large sample sizes the distribution of U can be approximated by a normal distribution with $\mu_U = \frac{n_1 n_2}{2}$ and $\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$. Note that W has the same variance and a shifted mean $\mu_W = \mu_U + n_1(n_1 + 1)/2$.

Accomodating confounding factors

In some situations other properties of a gene may coincide with its assignment to a functional set. If this property is also correlated with the experimental properties measured it might lead to false positive functional enrichment. As an example we consider an application in genetic studies where the sets of genes will be expanded to sets of SNPs which occur in (or around) the genes of a gene set. In these studies the association of a SNP to a disease is the experimental evidence used for the functional enrichment test. Minor allele frequency (MAF) could be a confounding property as common SNPs with higher MAF tend to be more often associated in GWAS. Therefore, if the set contains more SNPs with a higher MAF than the background set it might be that the observed functional enrichment is due to this selection bias.

Inverse probability weighting using propensity scores is a way to account for this selection bias [94, 95]. Propensity scores represent the conditional probability of a gene being selected in the set given other covariates, in this case MAF. Let $Z = 1$ if a gene is element of the set S and $Z = 0$ otherwise and let x denote the covariates potentially biasing the selection. Then the propensity score is defined as

$$e(x) = \mathbb{P}(Z = 1|x). \tag{2.26}$$

For discrete x the propensity scores can be found easily by counting. Continuous x either have to be modelled e.g. using logistic regression or discretised. In our application we chose to discretise

the MAF into bins of width 0.05. These propensity scores are then applied to the ranks used in the Wilcoxon test. In order to apply weights to each item Eq.2.24 is expressed equivalently as

$$W = \sum_{i \in S} \sum_{j \in \bar{S}} \mathbb{I}(i \leq j). \tag{2.27}$$

Finally the weights are incorporated to obtain

$$\bar{W} = \sum_{i \in S} \sum_{j \in \bar{S}} e(x_i)^{-1} e(x_j)^{-1} \mathbb{I}(i \leq j). \tag{2.28}$$

The distribution of $\bar{W}$ is still asymptotically normal with mean and variance given in [96].

A second source of confounding can be introduced when using different experimental platforms to measure the association of each gene. For instance several different cohorts that are usually genotyped on different SNP arrays are combined to obtain SNP genotypes for a meta analysis. This situation amounts to the combination of independent experiments into a single Wilcoxon statistic described in [97, 98]. The combined statistic W^* of K independent experiments with groups of sizes n_{k1} and $n_{k2} (k = 1 \dots K)$ is a linear combination of the statistics of the K experiments

$$W^* = \sum_{k=1}^{K} c_k W_k. \tag{2.29}$$

It has been shown [97] that the weights $c_k = (n_{k1} + n_{k2} + 1)^{-1}$ yields the highest efficiency of the test. For large sample sizes the distribution of W^* can be approximated by a normal distribution with

$$\mu_{W^*} = \sum_{k=1}^{K} \frac{1}{n_{k1}} \quad \text{and} \quad \sigma^2_{W^*} = \frac{1}{12} c_k n_{k1} n_{k2}. \tag{2.30}$$

Let C_k be the set of SNPs typed on platform k, then Eq.2.28 is plugged into Eq.2.29 in order to obtain the final adjusted test statistic

$$W' = \sum_{k=1}^{K} c_k \sum_{i \in S \cap C_k} \sum_{j \in \bar{S} \cap C_k} e(x_i)^{-1} e(x_j)^{-1} \mathbb{I}(i \leq j), \tag{2.31}$$

with $c_k = (|S \cap C_k| + |C_k \backslash S| + 1)^{-1}$.

2.2.4 The iterated hypergeometric test

Functional enrichment analysis as presented above is based on the comparison of a fixed set of functionally related genes either to (1) another fixed set derived from an experiment or (2) a quantitative ranking of genes derived from an experiment. Sometimes also the notion of functional relatedness can be made quantitative, for instance by computing transcription factor (TF) binding affinities to promoters of genes [99] to define potentially co-regulated genes. In this case two ranked lists of genes are compared. Let E denote the ranking of genes according to the experiment and F denote the ranking accoring to a functional definition. In order to show

enrichment of targets of a certain TF among the genes responding to a certain experiment the null hypothesis

$$Q_4 : E \text{ and } F \text{ are independent} \tag{2.32}$$

has to be rejected.

The iterated hypergeometric test [99] is a procedure to assess Q_4. As the name suggests it is based on the exact test discussed in section 2.2.1. In contrast to the exact test the thresholds t_E and t_F are not fixed to define sets from the ranked lists E and F. Instead a (restricted) exhaustive search for the optimal thresholds t_E^* and t_F^* is conduted. The objective function of this optimisation is the significance of the exact test based on the two sets defined by t_E and t_F. The restriction to the exhaustive search with maximum ranks is necessary because otherwise results tend to become unspecific [100]. Significance of the optimised result is obtained empirically by running the optimisation on 10^6 randomly permuted ranked lists. In addition to the statement about Q_4 the iterated hypergeometric test also yields the optimal thresholds which can be used to actually define two sets and more importantly their intersection - the experimentally and functionally related genes.

3 Expression quantitative trait loci (eQTL)

3.1 Gene expression as quantitative trait

A common feature of LD based genetic mapping – whether QTLs in animal models or human GWAS – is that it merely identifies tagging genetic variants and generally not the functional variants that are sought. Functional DNA variation can have two major types of consequences. Either a variation is coding or regulatory. Best understood are coding variations as they affect the sequence of protein coding genes leading to amino acid exchanges, premature stop codons or splice defects resulting in non-functional proteins. Less well studied variations could affect non-coding genes e.g. for microRNAs. Regulatory variations comprise variations in the proximal promoters and enhancer elements. Additionally, DNA variations that affect epigenetic factors such as histone positioning or histone modifications fall into this category. Sometimes variations could also be of both types; i.e. coding variations in a protein that lead to a feedback regulation of that gene or adaptive regulation of other genes.

Since protein coding genes only represent a small fraction of the genome (e.g. 1.2% in humans [101]) most variations are expected in the non-coding part where regulatory elements reside [19]. Although the exact molecular mechanism of regulatory variations might not be solved immediately, the consequences are visible as changes of gene expression levels. With the availability of DNA microarrays (see section 3.2) it became possible to characterise subjects of a genetic study not only for their genetic variation on the DNA level but also for their variation of genome wide gene expression levels. Thus combining genetic mapping with global gene expression profiling provides a strategy to identify genes whose transcript levels are affected by regulatory variations [35]. In this combination gene expression levels are treated as intermediate phenotypes and subjected to standard QTL mapping methodology described in section 1.2. Loci that affect the expression levels of a transcript are called expression quantitative trait loci (eQTL). Pioneering studies undertaken in yeast [102], maize, mouse and human [103] as well as rat [104] identified thousands of genes whose expression was associated to eQTLs. Though the eQTL approach is relatively new, it is worth of note that the landmark work of Jacob and Monod [105] on the *lac*-operon can be considered the first study of the genetics of gene expression – of course not on a genome-wide scale.

In the following the underlying data generating technology of DNA microarrays will be described, together with the implications for normalisation and data analysis. Then we will describe in detail the data for and steps of the eQTL analysis that we have performed in the BXH/HXB RI strains. The chapter concludes with a presentation of a strategy to identify positional candidate disease genes by integrated analysis of physiological and eQTL data which was applied in two case studies.

3.2 Measuring gene expression with microarrays

3.2.1 Gene expression microarrays

DNA microarrays are a technology that enables to quantify the expression levels of thousands of genes simultaneously. These arrays consist of small solid surfaces that carries DNA fragments called probes that are specifically hybridising to the complementary target mRNA sequences of known genes. Meanwhile there exist a variety of different technologies to manufacture these arrays [106, 107, 108, 109]. The gene expression data presented in this thesis have been generated using the Affymetrix Genechip technology [110] and the Illumina BeadChip technology [111].

Affymetrix synthesises short (25bp) oligonucleotide probe sequences directly on a silicone surface using a photolithographic technique [112]. Because of their short length, these oligonucleotide probes exhibit a certain degree of cross-hybridisation with unspecific target sequences. Therefore genes are represented by 11 probes summarised in a probe set. Additionally the arrays contain for each probe a mismatch probe, where the central base is exchanged. These can be used to estimate the level of cross hybridisation.

The Illumina BeadChip has longer (50bp) probe sequences that are attached to beads. Each probe sequence is attached to 15 beads on average. These beads also contain address tag sequences. They are distributed randomly on a microwell plate where they self-assemble. Hybridisation to the address tags is used to map the positions of the probes.

In both technologies, biotin labeled cRNA from one sample is then hybridised to one array. After washing off the excess only stably hybridised RNA is left on the array and can be stained with a fluorescent labeled biotin antibody. Finally the arrays are scanned with a confocal laser microscope which records the signal intensities for each position. In an image analysis step, the intensities are summarised for each probe location. This data is then stored in the ".cel" file format. Together with the array design information gene level expression summaries can be computed.

3.2.2 Normalisation

Direct comparison of raw intensity values can be misleading because of non-biological variation which affects all genes systematically. Reasons for non-biological variation include difference in the amount of total RNA hybridised in each sample, difference in labelling efficiency or general settings of the laser scanner. For a more exhaustive discussion see [113]. Therefore the data has to be normalised before the analysis.

A plethora of normalisation methods has been proposed for microarray analysis. A systematic overview and comparison showed that quantile normalisation yielded the best results [114]. This led to the development of the RMA-algorithm [115] which consists of (1) background correction, (2) quantile normalisation and (3) aggregation of probe signals of a probe set. This method was used in our analyses and will be presented briefly.

For the background correction, the perfect match signal for probe set n, array i and probe j is modeled as $PM_{ijn} = bg_{ijn} + s_{ijn}$, a mixture of background bg_{ijn} and real signal s_{ijn}. It is assumed that signals are following an exponential distribution while the background distribution

is normal, for each array. The background corrected intensities are defined as

$$B(PM_{ijn}) = \mathbb{E}(s_{ijn}|PM_{ijn}). \tag{3.1}$$

Background corrected intensity values are then used for quantile normalisation. The underlying assumption and also objective of the quantile normalisation method is that the overall distributions of gene expression levels should be the same for all arrays. Quantile normalisation is motivated by the $Q-Q$ plot. If two distributions are equal, then their quantiles are equal, therefore all points in the $Q-Q$ plot are lying on the diagonal. This idea can be extended to compare more than two distributions. When n distributions are equal, the points lie on the diagonal line given by the unit vector $\mathbf{d} = (\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$. For arrays with p probes, let $\mathbf{q_k} = (q_{k1}, \ldots, q_{kn})$ for $k = 1 \ldots p$ be the kth quantiles of array n. The following projection of the quantiles $\mathbf{q_k}$ onto the diagonal $\mathbf{d}$

$$\text{proj}(\mathbf{q_k}) = (\frac{1}{n}\sum_{i=1}^{n} q_{ki}, \ldots, \frac{1}{n}\sum_{i=1}^{n} q_{ki}) \tag{3.2}$$

implies a simple algorithm that computes quantiles for each array, averages them and substitutes the original values by these averages. Given a $(p \times n)$-matrix X of expression values of p probes and n arrays:

1. sort columns of X
2. substitute the values in the rows by the row average
3. arrange the columns of X back to their original order

In order to obtain summaries on the level of probe sets, the background corrected, normalised and $\log_2$ transformed PM intensities Y are described by a linear model

$$Y_{ijn} = \mu_{in} + \alpha_{jn} + \epsilon_{ijn} \tag{3.3}$$

with α_{jn} being a probe specific affinity, μ_{in} the log scale expression level on array i and ϵ_{ijn} an i.i.d. error with mean 0. The median polish method [116] is used for a robust fitting of the model. The so called robust multi array average (RMA) defined as μ_{in} is used as the final expression value of the probe set. Figure 3.1 shows the effect of RMA normalisation on the distribution of expression values for the left ventricle data set of the BXH/HXB RI strains.

In the study of rat expression QTLs we encountered an additional caveat. The data was produced in batches tissue by tissue. The first four batches were analysed using the array *rae230 a*, which is one part of the array set 230. With advances in microarray technology, it was possible to place all the probes of the two original arrays on a single array *rat230 2.0*. So later batches were analysed using this array. For the combined analysis of all tissues we had to extract the subset of probes, that were on the *rae230 a* from the data generated using the *rat230 2.0*. This allowed for the joint normalisation of all tissues. We have verified that normalising only subsets of probes gives similar results compared to the complete set of probes as demonstrated in Figure 3.2.

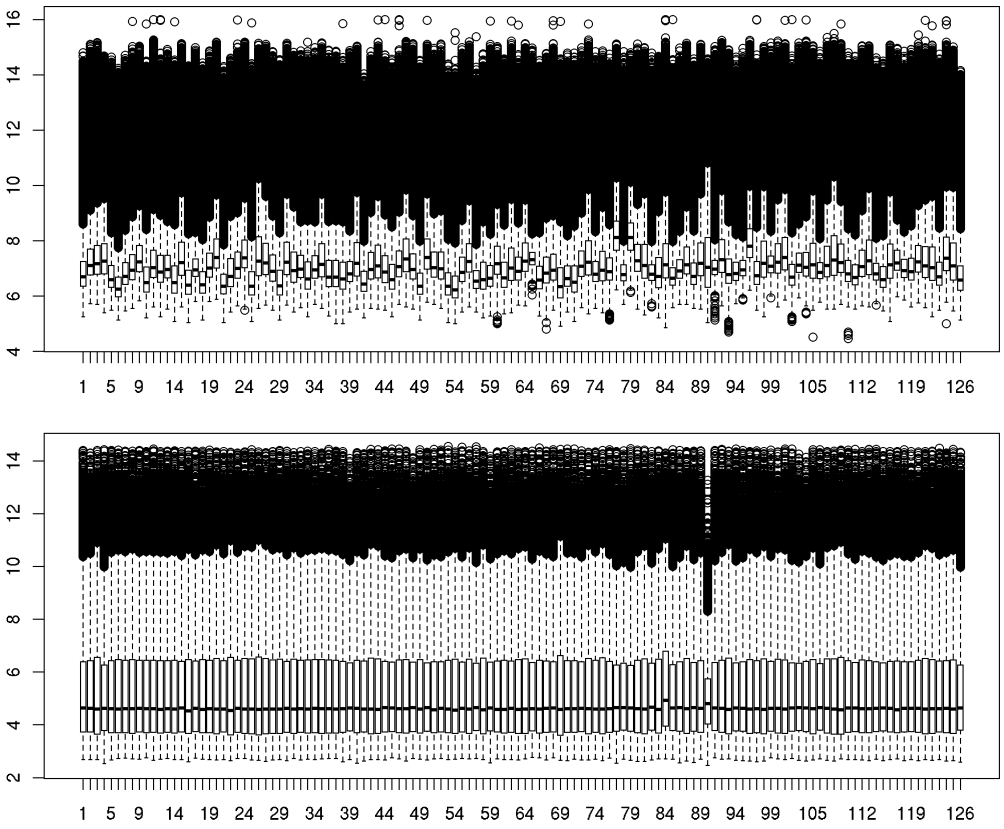

Figure 3.1: **Effect of RMA normalisation on the distribution of microarray data.** The boxplots in the top panel show the 25, 50 and 75 percentiles of the $\log_2$ transformed raw intensity values for each of the 128 microarrays measured in the left ventricle across RI strains. The whiskers indicate the 10 and 90 percentile, outliers are plotted as circles. The bottom panel shows the distributions after RMA normalisation. Note that if only quantile normalisation is performed the boxplots of all arrays should be identical. The deviation of the boxplots from identity is due to the aggregation step from probe to probe set level.

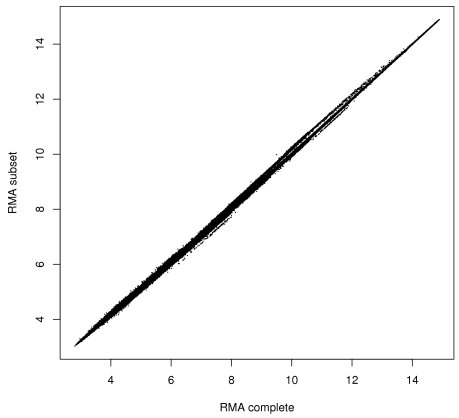

Figure 3.2: **Effect of subset extraction of microarray data.** This scatterplot shows the correlation of RMA normalised expression values using all probe sets of the array compared to using only probe sets that are available on the *rae230 a* array.

3.3 Mapping of eQTLs in the BXH/HXB RI strains

The BXH/HXB recombinant inbred strains are an excellent model system to study the genetics of cardiovascular phenotypes as explained in section 1.4. Mapping genetic determinants of gene expression can lead to the identification of functional mechanisms that underlie phenotypic diversity as discussed in section 3.1. A complete eQTL analysis of seven tissues in the BXH/HXB is described in the remainder of this section. The link between genotype, gene expression and phenotype is described in the following section 3.4 where eQTL strategies for heart failure and hypertension are presented.

3.3.1 Construction of a high density SNP map

The genetic map constitutes the basis of all genetic studies including the study of expression QTL. In order to construct a high density genetic map from single nucleotide polymorphisms (SNP) markers we have used data generated in the context of a large scale effort by the STAR consortium [1]. Accurate and complete genotypes for a subset of 20,238 SNPs across 167 distinct inbred rat strains, two rat recombinant inbred (RI) panels, and an F2-intercross were obtained in this project.

For the construction of the genetic map we have used the two independent panels of recombinant inbred (RI) strains derived from SHR and BN-Lx rats (BXH/HXB, see section 1.4) ($n = 31$), and from F344/Stm and LE/Stm rats (FXLE-LEXF) ($n = 33$), and 89 progeny of a F2 cross between BN/Par and GK/Ox rats (GKxBN) where 9,691 SNPs were typed in.

Genetic mapping in this cross and both panels of RI strains was performed using the R and R/QTL software packages [117, 118] integrating SNP genotype and physical map data resulting in 16,543 SNPs mapped. Data were initially filtered to remove markers containing genotyping errors (e.g. absence of segregation in the cohort despite apparent allele variation in the parental strains or over 10% of heterozygous genotypes in the RI strains) and blocks of adjacent SNPs with identical segregation patterns were collapsed into strain distribution patterns (SDPs). We have systematically removed markers that generated suspiciously large map distances, using criteria derived from the approximately linear relationship of genetic and physical distances. The criterion to call an interval suspicious was obtained from a linear model. It is defined by a user-specified intercept that is the minimal genetic distance at which distances are considered for removal and a slope that is computed chromosome-wise from the data. We set this threshold to 3cM. In order to determine the slope, an initial genetic map is estimated for all markers using the order defined by the physical map. Then all map distances greater than the 95% quantile are removed and the model is fitted. For each chromosome we performed the following steps: (1) compute the initial map based on the physical order of markers (2) estimate the linear model (3) while the size of the genetic map is reduced, evaluate the size of the genetic map when removing candidate markers and select the marker leading to the minimal map size.

Details of the typed markers and mapped positions are given in supplementary Table 4 of [1] and `http://www.snp-star.eu`. Strong evidence of discrepancies between the genetic map and the draft genome assembly were found (Figure 3.3). In particular, genetic mapping in all three panels identified a p11-centromeric segment of chromosome 1, which has been wrongly assembled in the p14-telomeric region of chromosome 17. Genetic mapping data suggest further

additional intra- and inter-chromosomal relocations in regions of chromosomes 2, 4, 11, 12, 14, 17. Known conflicts between rat genome assemblies, provided by BCM and Celera `http://rgd.mcw.edu/gbreport/gbrowser_error_conflicts.shtml`, indicate the relocation in the p14 region of chromosome 17 supporting the Celera assembly, and one conflict on chromosome 9 is resolved favouring the BCM assembly (not shown). The other conflicting mapping results require further independent verifications.

When we set out to construct a genetic map for the X chromosome based on the physical order of markers we detected several unlinked markers which rendered the mapping impossible. In-depth investigation of these linkage breaks revealed that they occur on contig boundaries (supplementary Table 5 of [1]). We rearranged the fragments of the chromosome resulting from splitting the contigs that were not linked ($LOD < 2$) in the order that generates the smallest average recombination fraction in the three populations. Using the resulting marker positions we constructed three genetic maps summarised in supplementary Table 4 of [1].

The genetic maps that were generated from RI panels and an F2 cross show that the draft genome sequence is largely correct, but did also reveal several regions that need further investigation. And for the purpose of this study, we provided a high-resolution map of the contribution of ancestral genomic segments for every individual strain in the BXH/HXB recombinant inbred panel.

3.3.2 Gene expression data

Expression data for eQTL analysis has been generated for seven tissues: adrenal gland, aorta, fat, kidney, left ventricle, liver and skeletal muscle using Affymetrix RAE230A and RAE230_2.0 arrays. Data from fat, kidney and heart have been described in [104, 119]. Data of the remaining tissues has been published in [3]. For each of the SHR and BN parental strains and the 30 RI strains 4-5 biological replicates were profiled. Altogether 907 expression arrays were analysed. Data was normalised using the RMA algorithm [120] (see section 3.2.2) with background correction, quantile normalisation and $\log_2$ transformation together for each set of microarrays profiled in each tissue. For each transcript and within the replicates of one strain, we removed outliers from the expression data using the Nalimov outlier test, as previously described [104, 119]. All expression data are accessible via ArrayExpress.

3.3.3 Expression QTL mapping

We have used the genetic map of the BXH/HXB RI strains [104] generated in a large scale effort by the STAR consortium [1] described in section 3.3.1. This map was derived from around 13,000 polymorphic SNP markers leading to $\sim 1,400$ unique strain distribution patterns (SDP) for the genetic analysis. The expression values of each probeset were averaged over the biological replicates of each strain and subjected to genetic mapping using the QTL reaper software [121].

This software implements the marker regression method proposed by [34] which is explained in section 1.2. Multiple testing of a single transcript against all genetic markers is accounted for by a permutation strategy to calculate genome-wide corrected P-values (P_{GW}) for each transcript. During the permutations the empirical significance of the genome-wide maximum of the LRS score of each transcript is established. The same empirical null distribution is used to assign

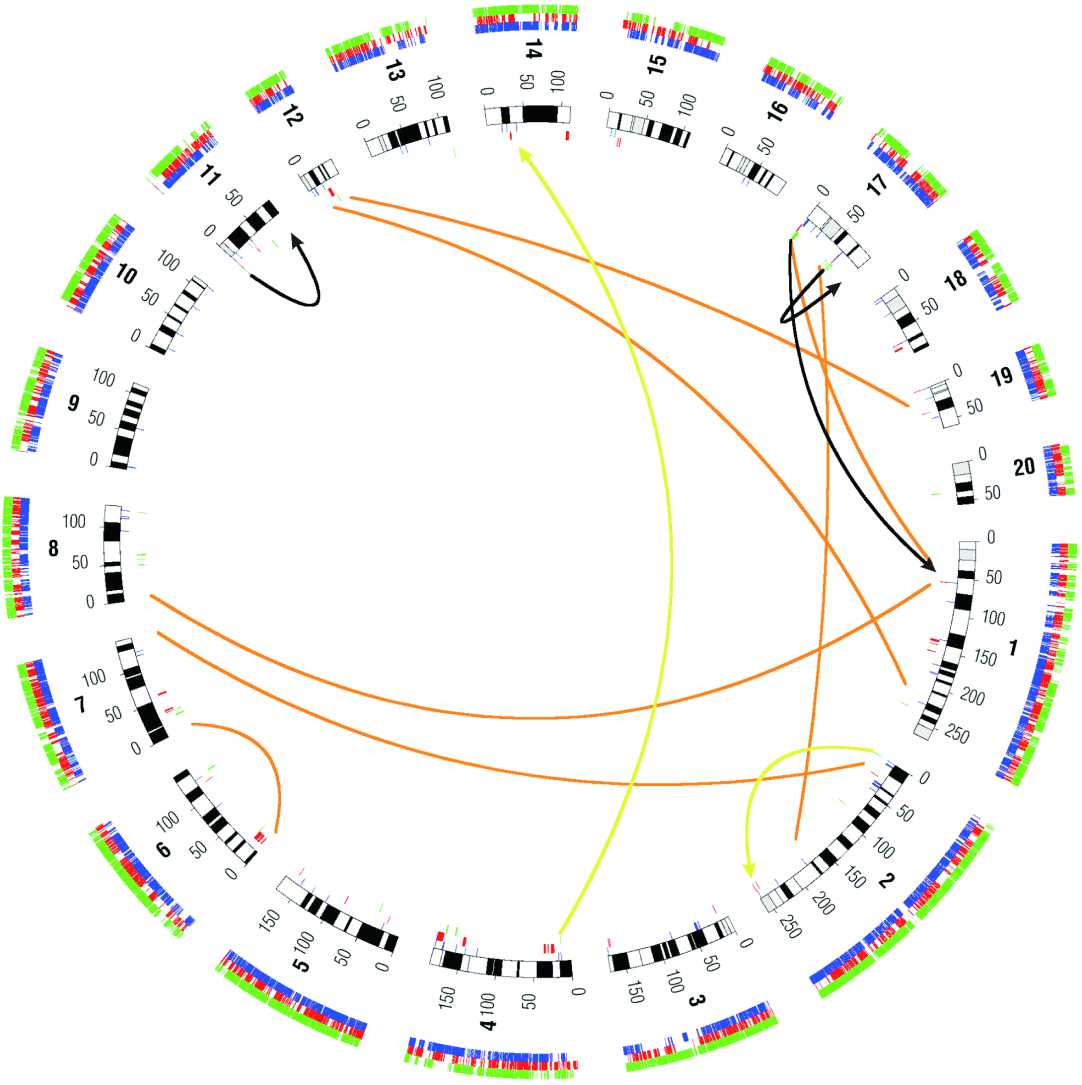

Figure 3.3: **Identified discrepancies between rat genome assembly and genetic maps.** Rearrangement of the physical map according to genetic mapping information. Data from each cohort are colour coded (red: FXLE-LEXF, green: HXB-BXH, blue: GKxBN). Black lines: all crosses support this rearrangement; lime green: HXB-BXH and F2 cross support this rearrangement. Orange lines indicate unresolved genomic conflicts. The outer circle marks positions of informative SNPs for each cohort. Arrows indicate the relocation of SNP markers that had extreme genetic distances compared to their physical distance from adjacent markers. Markers were relocated according to minimal recombination fraction. Conflicts in the genetic map are marked by bars in the inner circle.

P-values to pairs of markers and transcripts that were not at the genome-wide maximum. Here we used 1 million permutations to assess genome-wide significance of the eQTLs.

The above procedure corrects for the number of markers tested for each transcript, however we have tested all 30.000 transcripts on the microarray. Therefore we have to apply a second multiple testing correction. Since the empirical *P*-values are based on the maximum of the genome-wide *LRS* score, we have used the corresponding *P*-values of all transcripts in order to estimate the false discovery rate using Storey's *q*-values [56] as described in [104].

Genetic mapping of gene expression phenotypes also allows for a broad classification of the mode of regulation. If the genetic marker that affects gene expression of a transcript is in close vicinity of that transcript, it is very likely that the marker is a proxy for a variation in a *cis*-regulatory element. Conversely, we have evidence for *trans*-regulation if the marker is located distal from the transcript or even on a different chromosome. Following [104] we define regions of 10 Mb around the transcripts as *cis*-regulatory, while other regions are *trans*-regulatory.

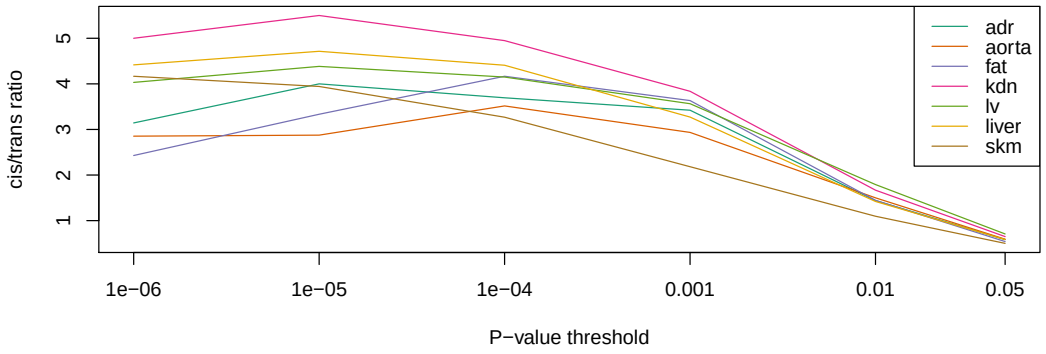

Figure 3.4: **Ratio of the number of *cis* / *trans*-eQTLs as a function of genome wide significance.**

Table 3.1 summarises the numbers of significant *cis* and *trans* eQTLs found in each of the seven tissues. Fat, kidney and adrenal gland have been profiled on the smaller *rae230 a* chip which contains only half the number of probe sets compared to *rat230 2*. This difference is also apparent in the total number of transcripts under genetic control that were detected. Fat and adrenal gland have less genetically regulated transcripts than kidney, whereas kidney has approximately half as many eQTL transcripts as the other four tissues profiled on the *rat230 2* microarray.

Table 3.1: Number of significant eqtls in the HXB RI panel for varying P-value thresholds.

tissue	P-value	q-value	*cis*	*trans*	unknown	total	transcripts	markers
adr	0.05	0.42	483	827	377	1687	1559	714
adr	0.01	0.18	344	239	158	741	724	424
adr	0.001	0.03	219	64	79	362	360	228
adr	1e-04	6.2e-03	133	36	42	211	210	144
adr	1e-05	1.2e-03	76	19	19	114	114	79
adr	1e-06	3.7e-04	22	7	7	36	36	26
aorta	0.05	0.32	1002	1692	1367	4061	3737	980
aorta	0.01	0.14	732	488	718	1938	1888	719
aorta	0.001	0.029	496	169	407	1072	1066	496
aorta	1e-04	5.3e-03	327	93	288	708	707	365
aorta	1e-05	9.8e-04	184	64	173	421	421	256
aorta	1e-06	1.9e-04	97	34	99	230	230	161
fat	0.05	0.41	464	858	336	1658	1543	667
fat	0.01	0.18	341	236	136	713	696	390
fat	0.001	0.037	218	60	71	349	348	226
fat	1e-04	6.4e-03	125	30	46	201	201	140
fat	1e-05	1.6e-03	50	15	19	84	84	57
fat	1e-06	3.8e-04	17	7	9	33	33	28
kdn	0.05	0.20	695	1070	492	2257	2077	755
kdn	0.01	0.086	497	298	220	1015	973	532
kdn	0.001	0.016	330	86	100	516	512	333
kdn	1e-04	2.9e-03	193	39	58	290	290	198
kdn	1e-05	5.1e-04	110	20	34	164	164	118
kdn	1e-06	1.3e-04	40	8	13	61	61	50
liver	0.05	0.31	1346	1902	1524	4772	4296	1052
liver	0.01	0.13	1012	565	738	2315	2227	812
liver	0.001	0.026	688	193	429	1310	1295	578
liver	1e-04	4.6e-03	444	107	263	814	812	428
liver	1e-05	9.8e-04	263	60	155	478	478	284
liver	1e-06	1.9e-04	125	31	80	236	236	169
lv	0.05	0.26	1059	1811	1297	4167	3863	1030
lv	0.01	0.11	767	540	618	1925	1860	736
lv	0.001	0.021	494	151	321	966	957	475
lv	1e-04	3.5e-03	313	71	190	574	574	327
lv	1e-05	6.8e-04	165	35	94	294	294	185
lv	1e-06	2.2e-04	53	12	27	92	92	71
skm	0.05	0.20	1308	2612	1718	5638	4922	1065
skm	0.01	0.073	992	905	869	2766	2632	820
skm	0.001	0.013	681	312	515	1508	1497	604
skm	1e-04	2.1e-03	438	134	332	904	902	434
skm	1e-05	3.4e-04	276	70	190	536	536	298
skm	1e-06	9.7e-05	100	24	70	194	194	140

In keeping with previous studies [122, 103, 123, 124] we have observed genetic markers that are associated with a large number of transcripts. We term these markers together with the set of transcripts *trans-clusters*. As previously noted the strength of association of *trans*-linked transcripts is much lower than that of *cis*-linked transcripts [103, 104, 125]. This relation is visualised in Figure 3.4. For the analysis of *trans*-clusters the thresholds on genome wide corrected P-values are usually not stringently corrected because not individual eQTLs but the clustering pattern in the genome is highly significant [103]. Figure 3.5 shows the size distribution of *trans*-clusters across the seven tissues. Large *trans* clusters with more than 50 transcripts (with $P_{GW} < 0.05$) can be observed in all seven tissues. Adrenal gland, fat and kidney only have one such *trans* cluster each, whereas they are more abundant in other tissues (Table 3.2). The maximum size of *trans* clusters also varies across tissues (Table 3.2) with a very large cluster of 645 transcripts in skeletal muscle.

Table 3.2: **Overview of large *trans* clusters in the BXH/HXB RI strains.**

	adr	aorta	fat	kdn	lv	liver	skm
nr of trans clusters > 50	1	10	1	1	10	8	23
maximum size	85	291	81	95	215	102	645

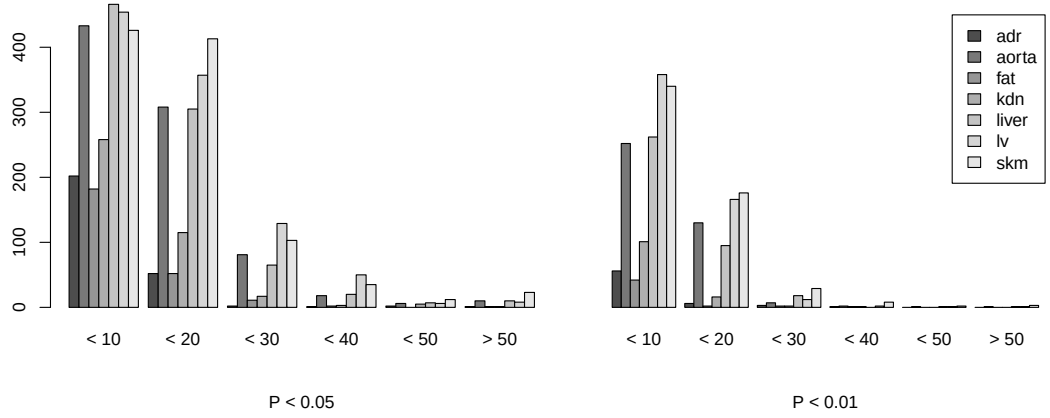

Figure 3.5: **Size distribution of *trans* clusters in the BXH/HXB RI strains.** We have counted the number of markers that are associated with at least 5 transcripts at two different thresholds of significance $P_{GW} < 0.05$ (left) and $P_{GW} < 0.01$ (right). Tissues are encoded by different shades of Gray (adr: adrenal gland, aorta:aorta, fat: peritoneal fat, kdn: kidney, liver: liver, lv: left ventricle, skm: skeletal muscle). The difference in the number of probe sets profiled for each tissue is also affecting the number of *trans* clusters identified.

3.4 Integrated analysis of eQTLs and physiological data

There are two principal genetic strategies to identify genes for phenotypic traits. The classical forward genetic strategy is based on genetic mapping in a segregating population (see section 1.2). Ideally fine mapping with ever more markers and increasing sample sizes leads to small linkage intervals containing only one gene. This strategy works well for mendelian traits [126, 127] but is rather time consuming and has not led to many results for complex traits [128]. The reverse genetic strategy investigates the relation of genes and traits by targeted perturbation of candidate genes in an experimental system. This allows to study the direct effect of the gene on the trait but requires the identification of a candidate gene first.

Combining the forward and reverse genetic strategy can speed up the identification of disease genes significantly. In particular when gene expression and physiological traits are integrated in the forward genetic strategy [128, 129, 2, 119, 130].

The knowledge about the expression of a gene represents the first step into a functional investigation (*functional genomics* was a term coined for microarray analysis). Gene expression studies have shown that correlation between physiological phenotypes and gene expression can be used to identify disease genes [131, 132]. Especially *cis*-regulated eQTLs represent strong evidence for a regulatory variation in close proximity to the gene. Therefore transcripts with *cis*-eQTLs that co-localise with QTLs for physiological traits and show correlation between gene

expression and disease phenotype represent excellent candidate genes for the trait, assuming that a regulatory variation is underlying the trait.

A common problem in microarray studies of diseases is that it is difficult to disentangle which gene expression changes are causative for the disease and which are mere adaptations to the disease state. In contrast, integrated genetic studies of gene expression and disease allow to derive three simple causal scenarios when gene expression and quantitative phenotype are both affected by the same genetic variant [129]. Either the DNA variant affects (1) gene expression which leads to disease (causal model), (2) the disease phenotype which leads to adaptive changes in gene expression (reactive model) or (3) both are affected independently. A method called likelihood based causality model selection [129] has been proposed to infer the most likely of these scenarios.

In the following, two case studies using the integrated analysis of phenotypic and gene expression data for candidate identification are presented for heart failure and blood pressure.

3.4.1 Identification of a risk factor for heart failure

We aimed to identify gene variants associated with heart failure by using a rat model of the human disease [2] (see also section 1.5). Using progeny from a F2 intercross of SHHF and SHRSP we performed physiological QTL (pQTL) analysis for left ventricular ejection fraction, the major clinical parameter used to characterise the failing heart. A locus on rat chromosome 15 centred at marker D15Rat10 showed statistically highly significant linkage ($LOD = 4.3$) . At the same locus, we further demonstrated statistically significant linkage for cardiac contractility. The SHHF allele was invariably associated with cardiac dysfunction. Grouping F2 animals according to zygosity at D15Rat10 did not show any genotype association with blood pressure or LVH, as determined by analysis of variance. This finding indicates that linkage to this locus was independent of blood pressure. Our data indicate that this QTL affects heart failure in an additive mode accounting for 15.1% of the phenotypic variance observed for the ejection fraction.

We searched for *cis*-regulated eQTLs, assuming that a regulatory variant changing gene expression in the left ventricle is responsible for the heart failure. Only 2 transcripts that were differentially expressed at a false discovery rate ($FDR < 5\%$) between the parental SHRSP and SHHF and showed a *cis*-regulated eQTL ($FDR < 5\%$) in the F2(SHHFxSHRSP) within the region of linkage to heart failure on chromosome 15. Of the two *cis*-regulated transcripts, *Ephx2* showed the strongest genetic evidence for significant allele-specific regulation in the F2(SHHFx-SHRSP) population. The linkage peak (genome-wide corrected $P < 10^{-6}$; $FDR < 10^{-4}$) for the *Ephx2* eQTL was at D15Rat10, the same marker that also defined the peak of the heart failure QTL on chromosome 15 in this cross. D15Rat10 is the nearest marker to the *Ephx2* gene and is localised within 2.7 Mb distance, based on the rat genome reference sequence v3.4. This finding suggests that the *Ephx2* transcript levels are regulated in *cis* at the *Ephx2* gene itself and coincide with the peak of linkage to the heart failure QTL. The second *cis*-acting eQTL gene in this region was *Mmp14*. Even though it is located within the 95% confidence interval encompassing the heart failure QTL, *Mmp14* is located at a distance of more than 10Mb from the heart failure QTL peak marker D15Rat10. Strongest eQTL linkage of *Mmp14* was observed with marker D15Rat83 (genome-wide corrected $P < 10^{-3}$; $FDR < 5\%$). Based on statistical significance and proximity to the heart failure peak marker we thus prioritised *Ephx2* as

candidate for further investigations.

We set out to identify potential regulatory polymorphisms and sequenced 5,000 bp upstream of the first exon. We found three SNPs and a two-nucleotide deletion in the putative *Ephx2* promoter in SHHF as compared to SHRSP. To test whether or not the allelic promoter variants influence *Ephx2* gene expression, we performed luciferase reporter assays and compared the variant promoters between SHHF, SHRSP, and WKY animals. We found a strong increase in promoter activity in the SHHF compared to the SHRSP allele [2]. The findings are consistent with the observed *cis*-regulated eQTL in the F2(SHHFxSHRSP) in which the SHHF allele is the allele that shows higher expression.

In addition we confirmed the effect of allele specific regulatory variation using allele-specific real time PCR of cDNA from heart tissue of 10 F1(SHHFxSHRSP) rats [2]. This experiment in F1 animals confirmed that the expression of *Ephx2* is regulated in cis and that the $Ephx2^{SHHF}$ allele-specific transcript levels are significantly elevated compared to the $Ephx2^{SHRSP}$ allele ($P = 8 \times 10^{-12}$).

Using computational predictions (see also section 4.2), we identified a consensus AP-1 (activator protein 1) transcription factor binding site in the *Ephx2* promoter that exactly covers the two-nucleotide deletion in the SHHF strain. To investigate whether or not the mutated AP-1 binding site affects AP-1 binding in vitro, we performed electrophoretic mobility-shift assays and found specific AP-1 binding to the SHHF promoter, while AP-1 binding in SHRSP was abolished.

In the forward genetics approach we demonstrated that *cis*-variation at *Ephx2* co-segregated with heart failure and increased transcript expression which is a consequence of a mutation in a transcription factor binding site. In addition the study showed co-segregation with increased protein expression, and enzyme activity, leading to a more rapid hydrolysis of cardioprotective epoxyeicosatrienoic acids [2]. In the backward genetic approach we confirmed our results by testing the role of *Ephx2* in heart failure using a knockout mouse model. These experiments showed that *Ephx2* gene ablation protects from pressure overload-induced heart failure and cardiac arrhythmias [2]. In addition to the study of model systems, the findings were also translated to humans by showing differential regulation of *Ephx2* in human heart failure, suggesting a cross-species role for *Ephx2* in this complex disease [2].

3.4.2 Identification of a candidate gene for systolic blood pressure

In the second case study, we aimed to identify candidate genes for systolic blood pressure in the rat. Previous studies have identified a QTL region for this trait on chromosome 1 which was confirmed by different congenic strains [133, 134, 135]. Although representing less than 1% of the genome, the region defined by genetic markers D1Rat200 and D1Rat57 is still quite large (50 Mb) and contains hundreds of genes that all could potentially influence the phenotype. So we decided to follow the strategy outlined above and to focus on regulatory variation.

Since the BXH/HXB RI strains are also a model system for hypertension we used the new SNP based genetic map (see section 3.3.1) to identify QTL for previously measured blood pressure phenotypes [65]. Using the QTL reaper software (see section 3.3.3) we identified a novel QTL for systolic blood pressure on chromosome 1, overlapping with the locus described above.

Our candidate identification strategy combines gene expression data from the congenic strains

with eQTL data and physiological data from the BXH/HXB RI strains. Again, assuming a regulatory variation, genes need to be consistently differentially expressed in the congenic strain and form *cis*-eQTL in the RI stains. Additionally, their transcript levels need to be correlated with systolic blood pressure. Genes meeting these requirements constitute priority positional candidates.

For the expression analysis of the congenic strain microarray profiling was conducted in kidney tissue from 5 animals of the parental strains WKY-1 and SHRSP, and the congenic line WKY-1.SHRSP-Mt1pa/D1Rat200 using the RGU ABC chip. Amongst the differentially expressed genes identified by ANOVA at $FDR < 5\%$, only those that mapped to the congenic region were considered. Further selection of candidate genes was based on a post hoc significance value of $P < 0.05$ as a cutoff for changed expression in SHRSP and congenics versus WKY-1. Allele dependent differential expression is expected to result in concordant direction of expression change in congenics and SHRSP versus WKY-1 since SHRSP and congenics carry the same alleles of genes in the region of interest. Therefore, concordant direction of expression change in congenics and SHRSP versus WKY-1 was used as an additional selection criterion.

Of the 455 probe sets representing the genes located in the region of interest, only eight showed significant differential expression of their corresponding transcripts with concordant direction of expression change in SHRSP and congenic animals versus WKY-1 and therefore represented potential candidate genes. When ranked according to the significance of expression changes in congenics versus WKY-1, the greatest fold change (FC) in expression was observed for the transcript detected by the microarray probe set rc_AI070448_at ($\log FC = -2.9, P = 0.0021$). The transcript detected by this probe set was the 297 bp cDNA clone UI-R-C2-mq-d-11-0-UI. Aligning the clone sequence against the entire rat genome showed that it matched with the first three exons of the gene *MrpL48* and contains further sequence that aligned with intronic sequence of *MrpL48*. This suggested that rc_AI070448_at detects an additional short transcript isoform of *MrpL48* ($MrpL48^{short}$) generated by alternative splicing. It also suggests that the variant is actually not regulatory but rather a coding variant in the gene.

Kidney gene expression data for the BXH/HXB RI strains was generated using the Rae230a chip (see section 3.3.2). However, no probe set corresponding to the target sequence of probe set rc_AI070448_at representing the $MrpL48^{short}$ transcript was included on this chip. Thus, qRT-PCR with a Taqman probe complementary to the target sequence of probe set rc_AI070448_at was conducted in kidney tissue of all 29 RI strains and $MrpL48^{short}$ expression was genetically mapped as well as the expression of all other genes encoded in the region defined by the genetic markers D1Rat200 and D1Rat57.

The eQTL analysis identified 13 genes encoded in the region of interest to be significantly *cis*-regulated ($P_{GW} < 0.05$) in the RI panel. The only candidate from the congenic analysis that was among these was the $MrpL48^{short}$ transcript. None of the other seven candidate genes constituted *cis*-eQTLs in the RI panel ($P > 0.05$). The $MrpL48^{short}$ transcript was therefore the only one from the congenic analysis that could be verified to be *cis*-regulated in the RI strain panel. The peak of linkage for the *MrpL48* *cis*-eQTL was at marker B01P1027, which is the closest marker to *MrpL48* with a distance of 1.6 Mb. The observed *cis*-regulation for $MrpL48^{short}$ suggested that genetic sequence variation(s) in the gene itself influence its own expression. The expression of $MrpL48^{short}$ transcript at the *cis*-eQTL (marker B01P1027) showed a bimodal distribution. Therefore we grouped the samples according to the presence

or absence of the expression of $MrpL48^{short}$ transcript, where the presence of this transcript is represented by Taqman $\Delta ct < 5$ and absence of the transcript is represented by Taqman $\Delta ct > 5$, respectively.

After determining significant *cis* regulation for expression of the $MrpL48^{short}$ transcript, correlation of transcript levels and blood pressure parameters was carried out for each of the candidates identified in the congenic expression analysis and for the additional genes identified to form *cis*-eQTLs in the RI panel in the region of interest. The $MrpL48^{short}$ transcript expression showed significant correlation ($P = 0.008$) with systolic blood pressure (SBP), but not with diastolic blood pressure (DBP). No significant correlation with blood pressure parameters was found for any of the other seven transcripts from the congenic strain expression analysis. Two transcripts out of the 12 additionally identified *cis*-regulated genes in the RI panel were significantly correlated with SBP (Dgat2: $r = 0.43, P = 0.021$; Ascl3: $r = -0.48, P = 0.009$). However, of all *cis*-eQTLs in the region of interest, $MrpL48^{short}$ transcript showed the strongest correlation (r=0.52, P=0.008).

Taken together, in two distinct models, *MrpL48* scored twice as the top priority candidate gene. Among the candidates identified by expression analysis in congenic strains, *MrpL48* was the most significantly and strongest differentially expressed gene and was the only one that could be verified in the RI panel as cis-regulated candidate and was the strongest correlated cis-eQTL to SBP in the RI panel. Therefore, it was prioritised for further functional analyses [29]. Currently, the reverse genetic experiments with knockout mice for Mrpl48 are underway in order to confirm the role of Mrpl48 in blood pressure regulation.

4 eQTL genes in gene expression networks

Most biological processes require the coordinated action of sets of functionally related genes. Enzymatic reactions in metabolic pathways are the most obvious example [82]. Coordinated activity also requires coordinated expression. Following this assumption, genome wide gene expression data can be used to reverse engineer sets of functionally related genes [136]. Conversely, *a priori* knowledge of functional gene sets can also be used to facilitate the analysis and interpretation of gene expression data (see section 2.2). Transcription factors (TFs) are the predominant regulators of gene expression [137]. Therefore combined analysis of gene expression data and transcription factor target predictions provides a way to connect regulatory mechanisms to functional gene expression networks.

This chapter introduces four novel approaches to interpret eQTL data in terms of gene expression networks. First we will make use of *a priori* knowledge of functional gene sets to interpret clusters of *trans*-eQTLs. Then we will shed light on the role of TFs for *cis*-regulated eQTLs and clusters of *trans*-eQTLs. Finally, we show how eQTL genes can be placed into a functional context by considering genotype induced changes of co-expression.

4.1 Extension of gene set enrichment analysis for genetic mapping

In our eQTL analysis of the BXH/HXB RI strains we have observed the existence of large *trans*-clusters (section 3.3). If a common genetic regulation of gene expression implies a common function it should be possible to detect functional enrichment of *trans*-clusters which would facilitate their interpretation. Previously we used eQTLs and defined sets of genes linked to the same locus by using a threshold on the eQTL P_{GW}-value. Subsequently we compared these sets to functional categories like GO or KEGG using the exact test described in section 2.2.1. However, for some tissues many *trans*-clusters do not contain enough annotated genes for this approach. This may be due to the fact that a threshold for genome wide significance is too stringent. It has been noted by others [103] that for most *trans*-clusters, the single transcripts hardly reach genome wide significance but the pattern of clustering in the genome is non-random.

In order to circumvent this problem we used a threshold free way of computing the enrichment of gene sets. GSEA (section 2.2.2) is a tool exactly for this purpose [86, 87, 88]. The basic idea is to use a ranked list describing the association between gene expression and genotypes and to assess whether the genes of a functional gene set are coherently found in the top (or bottom) of this list. This has been shown to work well in gene expression versus phenotype settings [86]. Small, but ubiquitous changes in expression (20%) of all members of the set could be detected. We aim to apply this method to all genetic markers in order to identify associations between functional gene sets and genetic markers.

In order to assess the association of the gene-expression on the level of a pathway we propose a novel method inspired by the idea of gene set enrichment analysis [86] and its extensions [87, 88].

We test the null hypothesis Q_2 that no gene of the pathway is associated to a given marker (see section 2.2.3). Instead of using a classification of the samples according to a phenotype, as it is usually the case in expression experiments we use a classification based on the genotype at each of the markers. Following the notation of [88] we define a pathway-association score between markers and pathways based on single gene statistics and a set summary statistic. Let genes, samples, pathways and markers be indexed by $i = 1, \ldots, B$, $j = 1, \ldots, n$, $k = 1, \ldots, K$ and $l = 1, \ldots, L$. The incidence matrix A encodes the absence or presence of a gene in a pathway:

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1B} \\ \vdots & & & \vdots \\ a_{K1} & a_{K2} & \ldots & a_{KB} \end{pmatrix} \tag{4.1}$$

where

$$a_{ki} = \begin{cases} 1 & \text{if} \quad g_i \in C_k \\ 0 & \text{if} \quad g_i \notin C_k \end{cases} \tag{4.2}$$

and Ck is the set of genes of the k-th pathway. The association of gene expression patterns and the genotypes at the L marker locations are summarised in the matrix

$$Z = \begin{pmatrix} z_{11} & z_{12} & \ldots & z_{1L} \\ \vdots & & & \vdots \\ a_{B1} & a_{B2} & \ldots & a_{BL} \end{pmatrix} \tag{4.3}$$

where z_{il} is the single gene statistic of association between gene expression of gene i and the genotypes at marker l. Here we use the $-\log_{10}(P)$ of the ANOVA test with genotypes as grouping factor. We do not use the F-statistics directly because the degrees of freedom may vary between markers depending on the number of genotyped individuals. The final matrix of pathway association scores X for all pairs of markers and pathways is then simply

$$X = AZ/\mathrm{rs}(A) \tag{4.4}$$

where $\mathrm{rs}(A)$ are the row sums of A.

We chose a permutation strategy in order to assess the statistical significance of the pathway scores while keeping the correlation structure between genetic markers intact. To compute empirical P-values for the pathway score statistic we randomise the data by permuting the labels of the samples thus breaking the genotype to phenotype relationship in the same way for all markers. The test is then applied to all pairs of markers and pathways from a data base of pathways. Finally the resulting P-values have to be corrected for multiple testing using an FDR approach [56].

4.1.1 Linking the arachidonic acid pathway to heart failure

We have applied our approach to the eQTL data generated for an F2 intercross of SHHFx-SHRSP (section 1.5). The integrated analysis of phenotype and expression data resulted in the identification of the *Ephx2* gene as a candidate for heart failure (section 3.4.1). Therefore the transcriptional regulation of pathways related to the function of *Ephx2* was of major interest.

We focused on transcripts of the arachidonic acid metabolism that include several important enzymatic reactions in the generation of epoxides (e.g. 14,15-EETs) from arachidonic acid and the enzymes that regulate arachidonic acid synthesis or beta-oxidation, according to the Kyoto Encyclopaedia of Genes and Genomes (KEGG). We ranked the genes from the arachidonic acid metabolic pathway according to their association to the heart failure locus on chromosome 15. None of these genes had a genome-wide significant eQTL at this locus when correcting for all transcripts tested. Nevertheless, we observe a concerted differential expression – in *trans* to *Ephx2* – of the top ranking genes with individual $P < 0.1$. Applying this relaxed filtering resulted in a set of genes containing six members of the CYP superfamily and three members of the PLA2 family that will be referred to as the *Ephx2* pathway (Figure 4.1). To quantify the small but consistent effect, we computed the gene set enrichment score for the association of gene expression on the level of the whole pathway with a genetic marker and applied it in a genome-scan over all genetic markers used in F2(SHHFxSHRSP). We show that the maximum association of the *Ephx2*-pathway transcripts occurred at the *Ephx2* locus and the pathway association score has genome-wide significance ($P = 2.4 \times 10^{-5}$). This analysis highlighted that increased *Ephx2* expression leading to a depletion of cardioprotective EETs results in a feedback regulation increasing the expression of the upstream pathway to compensate the EET depletion.

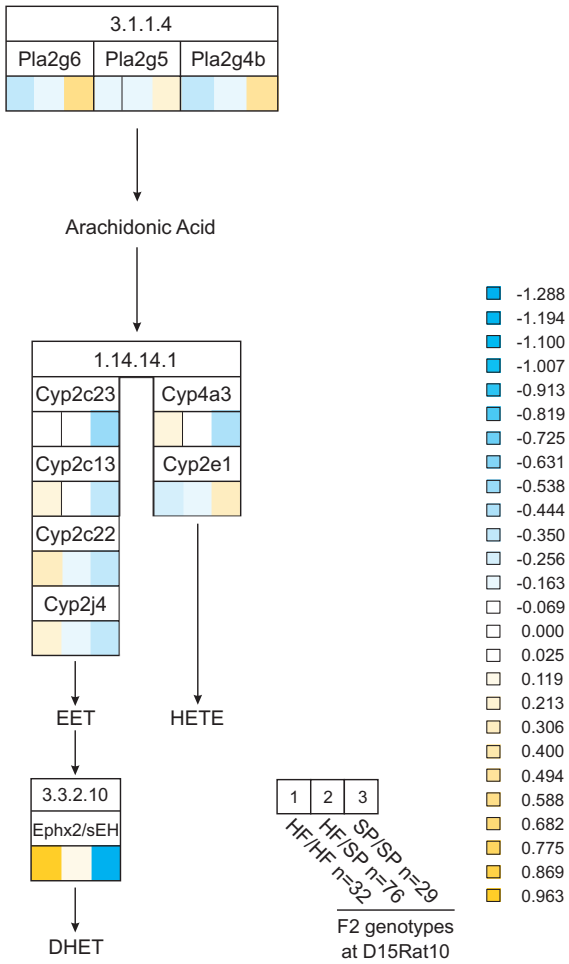

Figure 4.1: **Identification of regulatory trans eQTLs mapping to chromosome 15.** A simplified overview of the *Ephx2*-pathway is depicted. All genes analysed with the corresponding enzymatic function are listed in the boxes below. Each box is headed by an enzyme classification number describing the reaction that is catalysed. For each gene a colour code indicates the genotype specific average expression values according to zygosity (HF/HF, HF/SP and SP/SP; from left to right) at the *Ephx2* locus. The expression matrix has been centred and scaled to $\sigma = 1$ prior to visualisation. Red indicates over-expression, white baseline, and green under-expression. The CYP genes are subdivided into two columns where the left producing predominantly EETs and the right producing predominantly HETEs. The *Ephx2* pathway has its genome-wide pathway linkage peak at the *Ephx2* locus with $P = 2.4 \times 10^{-5}$.

4.2 Sequence variation in transcription factor binding sites

As outlined in the introduction, most disease associated sequence variations are found in regions that are not encoding proteins. Likely, the associated variation is not functional but tagging an unknown functional variant near by. Since the variation is not coding it might affect gene regulation. Section 3.1 introduced a systematic approach to detect consequences of potential regulatory variations on the level of target gene expression. However, the mechanism that causes the differences in gene regulation remains unknown. We shall come back to the example of the dysregulation of α-globin which is known to cause α thalassemia. Only recently experiments have provided first insights into possible molecular mechanisms, namely the creation of a novel *Gata-1* binding site and other hallmarks of regulatory activity in the upstream region of α-globin [20] as visualised in Figure 4.3. Another example of a disease causing regulatory variation is *Ephx2* (section 3.4.1). It has been identified by screening for genes which show *cis*-acting regulatory variations in an eQTL approach. Further sequence analysis of the putative promoter region showed variation between the parental strains. In section 4.2.3 we will describe how computational analysis can be used to predict that the creation of an *Ap1* binding site led to the increase of expression. Promoter assays and electro mobility shift assays confirmed the functionality of this binding site [2].

As the above example shows, computational studies can be used to prioritise SNPs and to generate hypotheses about the regulatory mechanisms. Recent work has attempted to utilise sequence conservation to assess the functionality of non-coding sequence variations [138]. These authors also noted that binding signals from transcription factors do not increase the power of regulatory SNP prediction.

Here we do not aim to predict functional or causative SNPs, but we assume that a potentially functional SNP has already been identified by other means, such as described in the *Ephx2* example. Instead we aim at a systematic approach to predict which transcription factor is most likely to be affected by a given sequence variation. To this end we developed a combined biophysical and statistical approach (sTRAP), which can predict SNP-induced changes in the binding affinity of a transcription factor. Importantly and owing to our statistical framework, we are able to compare these changes for a comprehensive set of transcription factors. We validated our approach against a set of known SNP-TF associations and find that sTRAP correctly predicts a large fraction of known SNP-TF associations at a small rate of false predictions. Finally we applied our approach to identify the variants of the *cis*-regulatory element and the upstream regulator of *Ephx2*.

4.2.1 Modelling transcription factor binding site affinities

Binding Models

An increasing number of genome-wide *in vivo* and *in vitro* studies of protein-DNA interactions [140, 141] aim to provide a comprehensive compendium of binding models for transcription factors under different conditions and in various species. For the purpose of this work we used a preliminary compendium of binding models, as available from the TRANSFAC database [142]. We used information on 202 vertebrate transcription factors which is encoded by 554 position specific weight matrices. In earlier work it has been shown how this information can be used to

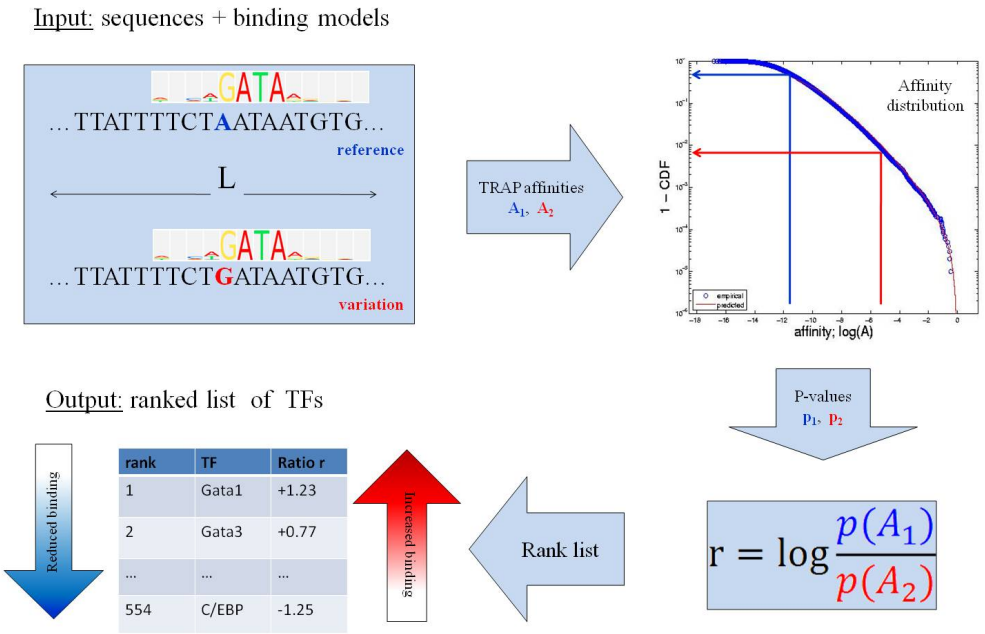

Figure 4.2: Overview of the sTRAP method. Using sequence data and a comprehensive set of transcription factor binding models as input, we predict the binding affinities of all transcription factors (TF) to the reference sequence and its variation. Thus we obtain, for every TF, two affinity values, which are normalised with the help of the affinity distribution from [139]. The log-ratio of the two p-values is used to rank all TFs according to their change in binding affinity.

predict the site-specific binding affinities of a transcription factor using a biophysical framework [143]. For the local binding affinities at sequence position l we use

$$a_l(R_0, \lambda) = \frac{R_0 \; e^{-E_l(\lambda)}}{1 + R_0 \; e^{-E_l(\lambda)}} \quad , \tag{4.5}$$

where $R_0(W) = 0.6W - 6$ and $\lambda = 0.7$ are two parameters which were fitted in [143], and W denotes the width of the motif. The local affinity predictions can be utilised in two different ways.

First we consider, for each SNP and every motif matrix, the W pairs of local affinities which are changed when comparing the reference sequence with its variation. Such a local comparison may suggest large effects on the predicted binding affinity, even if the flanking sequence contains additional binding sites which may buffer the effect of a sequence variation. Therefore we also employed a second strategy in which we calculate the regional affinity, A, of a transcription factor to a longer sequence region (i.e SNP + flanking region). This can be obtained by summing the local binding affinities, a_l, over all accessible sites. This procedure has the added benefit that, for each SNP and every matrix, we only need to compare two numbers.

Distribution of Binding Affinities

To compare binding affinities from different transcription factors we follow the statistical framework developed in [139]. There it was shown that a simple parameterisation effectively describes the distribution of affinities for most transcription factors:

$$\log A \propto P(x|a,b,c) = \exp\left(-[1 + a\frac{x-c}{b}]^{-1/a}\right). \tag{4.6}$$

The three parameters of this distribution also depend on the length of the sequence region and their value was determined in [139] for all 554 binding models from TRANSFAC [144]. While the distribution of Eq.4.6 is based on a sequence model of known human promoters, in some situations it might be preferable to use a different background model which captures the specific sequence properties in the vicinity of the SNP more accurately. In those instances we calculate the affinities of all transcription factors in a sufficiently large window around the SNP, and estimate the empirical P-value without parameters, but based on the rank statistics obtained from all windows in the neighbourhood of the SNP.

4.2.2 sTRAP: A framework to rank affinity changes

We use the quantitative framework for the computational prediction of transcription factor binding sites and TF binding affinities [143] described in section 4.2.1 and the distributions of binding affinities [139] described in section 4.2.1. Similar approaches have been studied by a number of other groups [145, 146, 147, 148].

Here we extend this approach to annotate pairs of sequences with respect to changes in their affinity. Just as we had previously ranked transcription factors with respect to their affinity for a single sequence, we now rank transcription factors with respect to affinity changes induced by sequence changes. For the purpose of this work we think of these sequence pairs as being

derived from a reference sequence and a possible mutation. This is illustrated in Figure 4.3. In particular we are interested in scoring the abolishment or creation of a binding site in one or the other sequence.

Let $A_X(S1)$ and $A_X(S2)$ denote the binding affinities of transcription factor X to two different sequences, $S1$ and $S2$. In the case under investigation, $S1$ might represent the reference sequence while $S2$ will denote the variation. For simplicity we assume that both sequences have the same length. According to Eq. 4.6 one can associate a normalised affinity (a P-value) with each affinity and define the log-ratio

$$r_X = \log_{10}\left(P_X(S1)/P_X(S2)\right) \quad . \tag{4.7}$$

Large positive values denote cases where the factor X increases its binding affinity, while for large negative values the binding affinity is decreased with respect to the reference sequence $S1$. Importantly, the ratios for different transcription factors, X, are directly comparable, because they are based on P-values, rather than absolute affinities. In logical terms, this corresponds to an exclusive disjunction (XOR), only that the score of Eq. 4.7 provides a more quantitative ranking.

Since minor changes in the binding affinity may also result in a considerable difference of the expression, one might also want to assign a high score if $S1$ or $S2$ or both show strong binding to a transcription factor, in which case one could replace Eq. 4.7 by the minimum of $p_X(S1)$ and $p_X(S2)$. However, for the purpose of this paper, we focus on the score defined in Eq. 4.7.

SNP data

In order to be able to assess the performance of our method we need a data set of known regulatory SNPs. While there is massive data on sequence variation from large-scale mapping efforts [14, 149], the regulatory potential of SNPs is badly documented and only occasionally reported. Here we study 20 known associations of regulatory SNPs with transcription factors, which were collected by [138]. These comprise SNPs which are naturally occurring or were generated by targeted mutagenesis. Moreover, for those SNPs the binding of selected transcription factors was shown to be affected. For each of these SNPs we retrieve a flanking regions of 60bp, 100bp, 500bp, and 1000bp.

sTRAP predicts many known SNP-TF associations

As a first test of our method we applied sTRAP to a list of known regulatory sequence variations and their associated transcription factors collected by Andersen *et al.* [138]. For each SNP, our method predicts a ranked list of transcription factors which is compared to the list of TFs known to be affected by the variation. As an input list we took 554 vertebrate transcription factor motifs from the TRANSFAC database [144]. A good method would predict known TFs at the top of the list. Indeed, in Figure 4.4 it is shown that a large fraction of known TFs appear top when ranked according to Eq. 4.7. We also compared this to random expectations where all TF motifs are assigned a random rank. The deviation from random expectations is clearly significant. Notice that the slight deviation of the random set from uniformity is due to the fact that some TFs have multiple motifs and we always take the best rank.

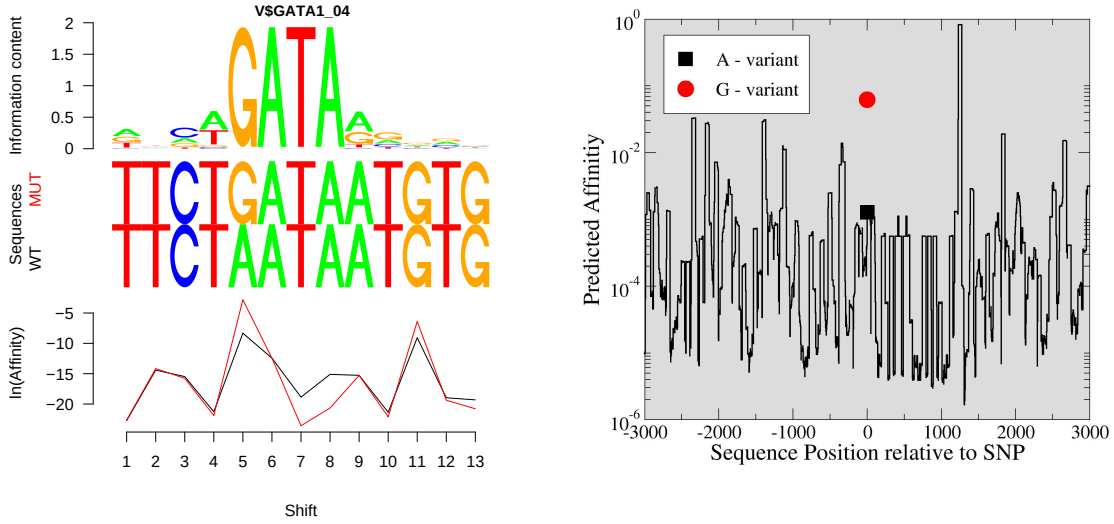

Figure 4.3: The left figure illustrates how a SNP may cause changes to the local binding affinity. Notice that the two curves at the bottom denote the raw affinity which is plotted on a logarithmic scale. This example is for a regulatory SNP in the SP1 promoter region and transcription factor NFY, whose sequence logo is shown at the top. This right figure shows a global version of the same approach where a larger window (L=60bp) was shifted across a ±3kb region around the SNP. The affinity of NFY is indeed strongly affected as evident by the large shift of the affinity at the position of the SNP. The surrounding sequence may serve to assess this change in the light of the fluctuations in affinity, but for the paper we utilised a parameterisation of affinity distribution from ref. [139].

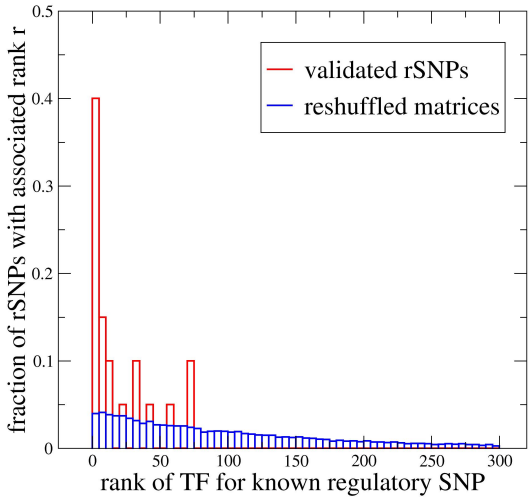

Figure 4.4: The sTRAP approach successfully predicts many known SNPs. For each known regulatory SNP and its associated transcription factor we record the rank of the corresponding matrix according to our scoring scheme. This figure shows that many known associations get a significantly high rank according to our scoring scheme (Eq.4.7). In blue we show the same histogram for a set of reshuffled matrices. The slight increase towards higher ranks is due to multiple matrices assigned to some factors.

sTRAP predictions are specific

While it is encouraging to see that many known associations of regulatory SNPs with their respective transcription factors can be detected, we now investigate more carefully the rate of false positives. Since only individual transcription factors have been tested for any given SNPs there is a generic lack of knowledge regarding the binding affinity for other factors. For our purposes we make the conservative assumption that all other transcription factors are not affected by the SNP and therefore count predicted associations of untested factors as false positives. This prescription is likely to inflate the estimated rate of false positives. In Figure 4.5 we plot the rate of true positives against the rate of false positives, when the threshold θ is changed to call a transcription factor as affected, $|r_X| > \theta$. The area under the curve is significantly larger than 0.5 as expected from random assignments. At a false positive rate of 15% we recover more than half of the known TF-SNP associations.

To assess the robustness of our method, we have compared the classification performance of different parameter settings and classifiers, using the area under the ROC-plot (AUC). As classifiers we have used different thresholds on the absolute log ratio or the minimum of the two p-values, corresponding to the two scenarios where SNPs affect the binding strongest or fall in strong binding sites without drastic changes of affinity. We evaluated different lengths of the sequence regions used to compute the binding affinities and different background models to determine the distribution of affinities. For the length of the sequence region, L, we used $L \in \{60, 100, 500, 1000\}$ bp to account for the effect of multiple binding sites, or $L = W$, where W is the motif width. The latter corresponds to the local approach described in Section 4.2.1. Furthermore, we also evaluated the classification performance for two classes of background models: (i) the GEV parameterisation and (ii) empirical p-values derived from the neighbouring sequence around the SNP of interest. In the latter case we defined different lengths of background sequences, $B \in \{1000, 5000, 10000\}$ bp. Our results are summarised in Table 4.1 which shows that the method is robust with respect to the choice of classifiers and the length of the sequence region.

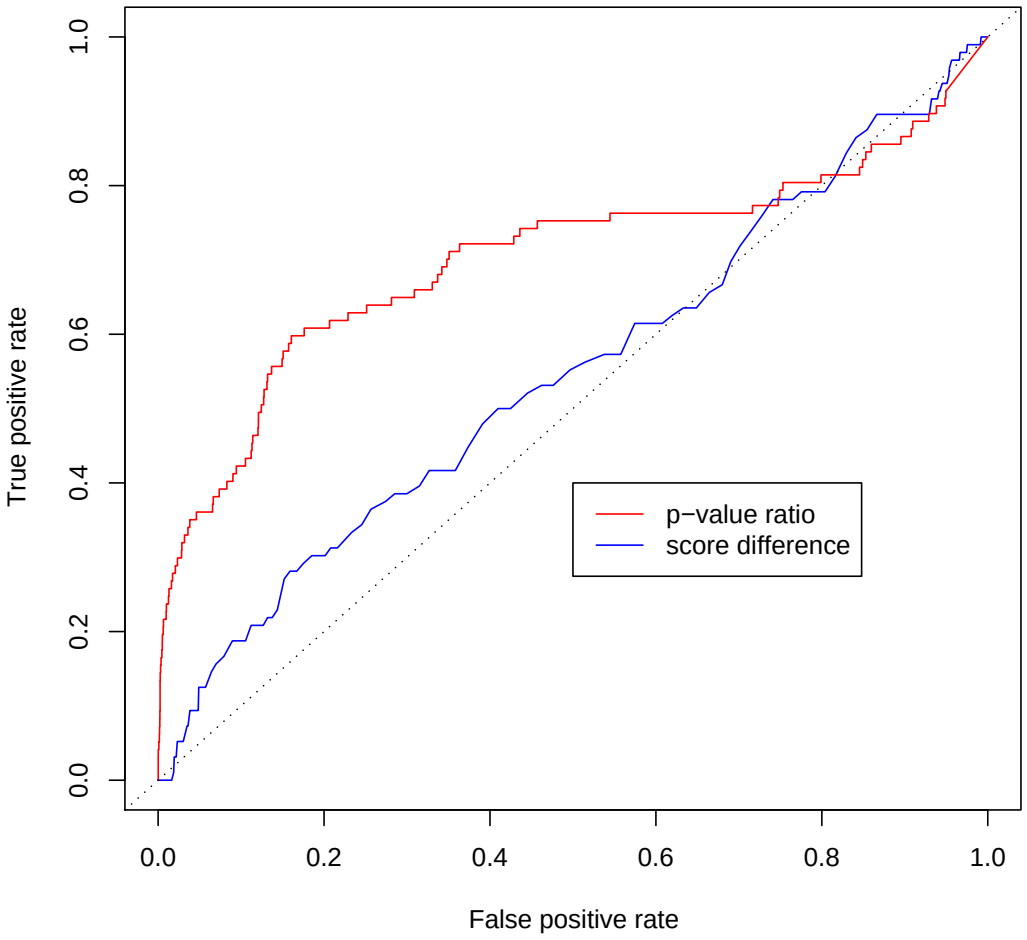

Figure 4.5: Here we show that the sTRAP approach is specific. The ROC curve is obtained by varying the threshold r_X in Eq. 4.7. At a rate of 10% false positive we recover 50% of all true positive. The area under the curve is 0.776.

Table 4.1: **Robust performance.** Here we summarize the area under the ROC curve (AUC) as a performance measure of our method. As decscribed in the main text we utilized different length of the flanking region. For the em local method, the length was set to the variable width, W, of the different transcription factors. The specific choice of L=61bp was motivated by our analysis of the data from [138]. The last column gives the AUC for an alternative approach, were the minimum of the two p-values is taken to be the score.

length	ratio	min
W	0.743	0.833
61	0.702	0.852
100	0.713	0.817
500	0.736	0.769
1000	0.749	0.767

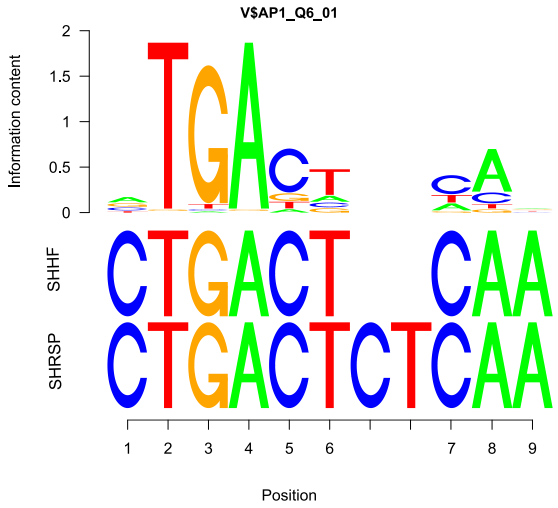

Figure 4.6: ***De novo* creation of an AP-1 binding site by a 2 bp deletion in SHHF.** The binding motif of AP-1 is aligned to the SHHF sequence which features a 2 bp deletion compared to the SHRSP sequence.

4.2.3 Identification of the regulator of Ephx2

In section 3.4.1 we described how a *cis*-regulated transcript was identified as candidate gene for heart failure. Having found the target gene of the regulatory variant we set out identifying the regulatory variant itself. Sequencing of the 5,000 bp putative promoter revealed three SNPs and a two-nucleotide deletion [2]. A difference in promoter activity could be established using a Luciferase reporter assay [2]. What remained to be elucidated was which of the four variants affected a *cis*-regulatory element and which TF bound to this element. Toward that end we have applied the sTRAP framework to rank pairs of SNPs and TFs. Table 4.2 shows the 10 Transfac matrices with largest change of affinity. AP-1 was the TF with the largest change of affinity caused by the 2 bp deletion in SHHF which created a *de novo* AP-1 binding site as shown in Figure 4.6. An electromobility shift assay (EMSA) showed an allele specific AP-1 binding [2], confirming that AP-1 is indeed the TF affected by the sequence variant.

Table 4.2: **sTRAP analysis of promoter polymorphism of the *Ephx2* gene.**

Matrix	Affinity SHHF	Affinity SHRSP	*P*-value SHHF	*P*-value SHRSP	Ratio	Log Ratio
V$AP1_Q6_01	0.232	0.009	0.001	0.100	0.006	-2.194
V$AP1_Q4_01	0.170	0.010	0.002	0.071	0.021	-1.671
V$FOXP3_Q4	0.055	0.003	0.013	0.259	0.052	-1.286
V$TCF11MAFG_01	1.020e-04	2.605e-07	0.075	0.731	0.103	-0.988
V$AP1_C	0.184	0.016	0.006	0.053	0.109	-0.962
V$AP1_Q2_01	0.041	0.002	0.048	0.382	0.125	-0.902
V$AP1_Q6	0.091	0.006	0.019	0.136	0.136	-0.867
V$STAT5A_04	0.353	0.353	2.885e-04	0.002	0.152	-0.818
V$AP1FJ_Q2	0.051	0.006	0.032	0.203	0.158	-0.801
V$SMAD4_Q6	0.040	0.146	0.023	0.004	6.138	0.788

4.2.4 Conclusions

Current GWAS result in many SNPs associated with common diseases, only a small fraction of which might be causative. Moreover, there is a lack of follow-up mechanistic studies to rationalise those predictions in molecular terms. In this work we undertook a first such step to predict *which* transcription factor is likely to be affected by a sequence variation. Our new approach (sTRAP) is based on an earlier biophysical framework (TRAP) and predicts sequence-induced changes in binding affinities. The quantitative approach relies on a proper normalisation of binding affinities and permits a robust ranking of the most affected transcription factors.

We have shown that the performance of the different classifiers is robust against the choices of parameters and the statistics used to transform affinities to P-values. Different classifiers were chosen based on different biological premises, but they have comparable performance (Table 4.1). We have implemented our method in an R-package called *tRap* and provide a public web interface, that allows the user to submit pairs of sequences corresponding to the two SNP alleles for analysis.

We believe that our method represents a valuable tool for the exploratory data analysis elucidating the mechanisms and possible consequences of regulatory SNPs. Besides its importance for understanding genetic diseases, our approach also provides clear suggestions for transcription factors that affect gene expression as we demostrated by the identification of the regulatory element and the upstream regulator of *Ephx2*. Another possible application is the study of species-specific sequence variation.

As with all sequence-based methods, our approach assumes that the binding dynamics of transcription factors to the sequence is rapid and that the equilibrium binding strength is the key parameter to control gene expression. Currently we employed a large but limited set of transcription factor motifs from the TRANSFAC database. Recent experimental advances and high-throughput data, such a protein-binding arrays [150], are likely to alleviate this limitation in the near future and permit an even more comprehensive assessment of the effect of sequence variations.

While sTRAP predictions are generically powerful as evidenced by the ROC curve of Figure 4.5, it will be a challenging task to validate individual predictions and integrate them into a molecular understanding of signalling and gene expression. The overarching goal of this project is to render computational binding predictions more quantitative. Clearly much work remains to be done. However, there is hope that the theoretical developments will increasingly be driven by technological advances and quantitative data [151], against which the models can be optimised.

4.3 The role of transcription factors in clusters of trans-eQTLs

While genome-wide association studies (GWAS) have uncovered a large number of common genetic variants associated with human diseases, the molecular mechanisms by which DNA variation affects disease risk remain poorly characterised [7]. To translate genetic association into biological function DNA variation has been correlated with gene expression to identify the genetic control points of gene networks that may be important determinants of disease aetiology [21, 22, 23]. Gene networks consist of transcripts of related biological function that are coordinately regulated by key transcription factors (TFs) although in yeast TFs are not

commonly encoded at genetic loci associated with gene networks [102, 122]. Previously, we have accumulated genetic mapping and genomic expression data in a panel of recombinant inbred (RI) rat strains derived from the spontaneously hypertensive rat (SHR) and the Brown Norway (BN) progenitor strains [104] (see section 1.4). Here we used this RI panel as a source of naturally occurring genetic variation to study TF-driven gene networks and their regulatory loci and integrated these data with human gene expression and GWAS data to identify genes, loci and pathways for human disease.

4.3.1 Identification of genetically regulated TF-networks

We combined expression quantitative trait loci (eQTLs) from three tissues (fat, kidney and heart) [104, 119] with new eQTL data in an additional four tissues (aorta, skeletal muscle, adrenal and liver) to create genome-wide eQTL datasets across seven rat tissues (see section 3.3.3). We used a two-step procedure to integrate eQTL data of TFs and TF-target genes to identify TF-driven gene networks. In the first step, we identified 147 TFs with known TFBS with a model in TRANSFAC [142] whose expression mapped to 587 eQTLs (genome-wide corrected $P_{GW} < 0.1$) across seven tissues, which were mostly ($> 90\%$), under trans-regulatory genetic control, in keeping with previous studies in yeast [122]. TFs act through transcription factor binding sites (TFBSs) in promoters and enhancers of TF-target genes. In the second step of the combined analysis, we tested for enrichment of predicted TFBSs (of TFs identified in the first step) in the putative promoter sequences of genes that mapped as trans-eQTLs.

For each TF, we retrieved a list of 1,000 top-ranked transcripts according to their Likelihood Ratio Statistics (LRS) score at the TF eQTL peak marker and subjected this ranked list to a TFBS enrichment analysis using PASTAA [99]. PASTAA compares this LRS-determined ranking to a ranking based on predicted TF binding affinities to the 200 bp proximal promoter determined by a biophysical model [143]. Promoter sequences were extracted from ENSEMBL and TSS annotation provided by P. Carnici (unpublished). The comparison is performed using an iterated hypergeometric test for the overlap of the 'top ranked' genes when varying the thresholds on both lists (see section 2.2.4). This procedure circumvents the recurring problem of setting more or less arbitrary thresholds on eQTL P-values or TFBS predictions. In addition to quantifying the overall enrichment this procedure also determines the thresholds used to define eQTLs and TF targets and moreover the set of genes satisfying both criteria which is used to define the differentially expressed targets of the TF. The statistical significance of this procedure is assessed using a null distribution generated from 1×10^6 permuted gene lists.

Out of the 13 TF-driven gene networks identified through the integrated analyses (Table 4.3) we observed the strongest TFBS enrichment ($P < 1 \times 10^{-6}$) for interferon regulatory transcription factor *Irf7*. *Irf7* TFBSs were predicted in the promoters of 23 genes, including *Irf7* itself, that all mapped to a single trans eQTL on rat chromosome 15q25 in adrenal, kidney, heart and liver (Table 4.3). We confirmed experimentally a subset of the predicted *Irf7* targets by chromatin immunoprecipitation (ChIP) and quantitative PCR that established direct interaction of *Irf7* with the promoters of these genes (Figure 4.8, [3]). Hence, *Irf7* and a group of *Irf7*-regulated transcripts are under *trans*-regulated genetic control at a single locus on rat chromosome 15 across four tissues. This provides evidence for a TF-driven regulatory cascade in which genetic variation on chromosome 15q25 directly or indirectly modulates the expression of

Irf7 (encoded on chromosome 1) with consequent effects on *Irf7* target genes expression (Figure 4.7, Figure 4.8).

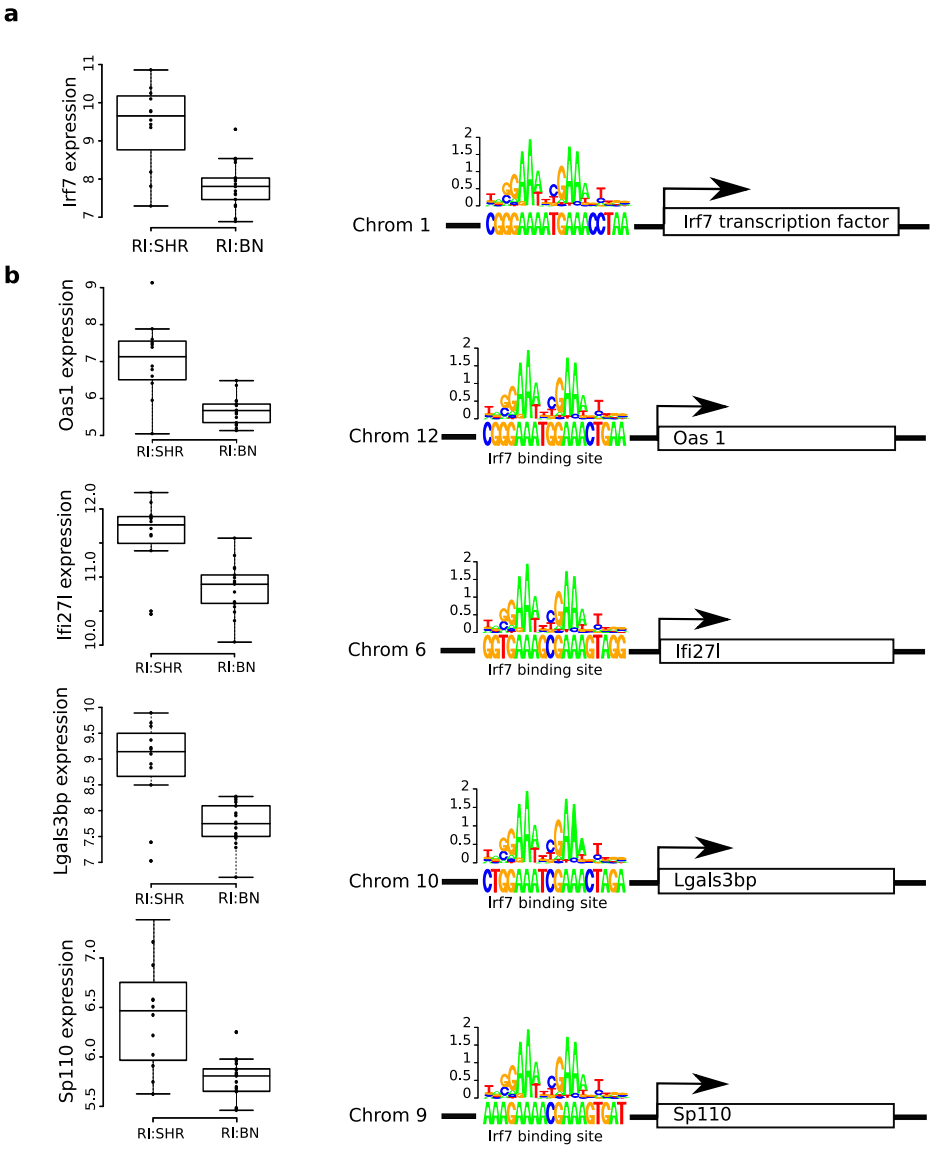

Figure 4.7: **The *Irf7*-driven inflammatory gene network (iDIN). a, Trans-regulated expression of *Irf7* and b, genes containing *Irf7* transcription factor binding sites by rat chromosome 15q25 at SNP J666808.** Left panels, gene expression in the left ventricle is shown in the recombinant inbred (RI) rat strains grouped by SHR or BN genotype at SNP J666808 (SHR allele, RI:SHR; BN allele, RI:BN). Right panels, transcription factor binding site predictions are represented for the five (out of 23 predicted) *Irf7* target genes. The chromosome (Chrom) encoding the *Irf7* target is shown to the left of the predicted *Irf7* binding sites. These data provide evidence for a regulatory cascade in which a locus on chromosome 15q25 regulates the expression of *Irf7* on chromosome 1 in an allele-dependent manner with consequent effects on *Irf7* target genes mediated through *Irf7* transcription factor binding sites.

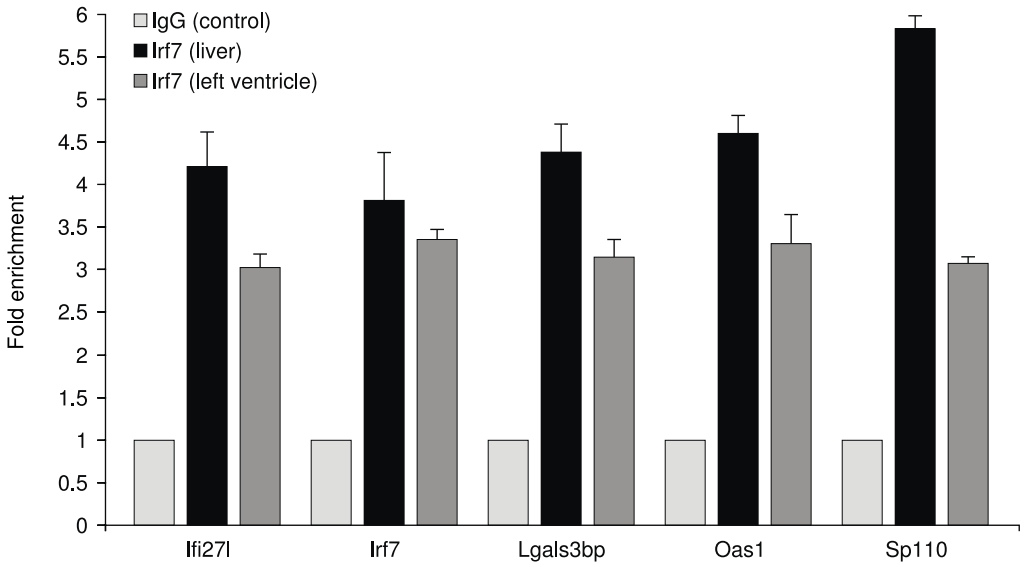

Figure 4.8: **Quantitative chromatin immuno-precipitation of predicted *Irf7* target genes.** Direct binding of *Irf7* to the promoters of the predicted targets *Ifi27l*, *Irf7*, *Lgals3bp*, *Oas1*, and *Sp110* was confirmed in liver and heart tissues. Fold enrichments are shown relative to non-immune IgG control. Our analysis predicts an auto-regulatory loop of *Irf7* acting on its own promoter, a finding that has previously been reported [152].

Table 4.3: **Genome-wide analysis of TF eQTL downstream effects.** The table contains results of the PASTAA analysis: Gene symbol and Affymetrix rat230 2.0 probeset id of the TF, transfac matrix identifier of the TF, eQTL marker of the TF, tissue in which expression was measured: (aorta = aorta, adr = adrenal, fat = peritoneal fat, kdn = kidney, lv = left ventricle, liver = liver, skm = skeletal muscle). Optimized P is the most significant P-value in the iterated hypergeometric test (see Methods) which results from an intersection of the given size when using the first "cutoff eQTL" transcripts and the first "cutoff TF" predicted transcription factor targets from a universe of genes of size "total" (see Supplementary Methods). Set size denotes the total number of genes used as potentially differentially expressed genes. The corrected P-value was obtained using permutation testing. The FDR accounts for the total number of 587 TF eQTLs tested.

TF symbol	TF Probeset	Matrix	Marker	Tissue	Optimized p	Intersect	Cutoff eQTL	Cutoff TF	Total	Set size	Corrected p	FDR
Irf7	1383564_at	V$IRF7_01	J666808	adr	1.25E-09	4	17	25	27544	893	< 1.0E-06	< 5.10E-05
Irf7	1383564_at	V$IRF7_01	J666808	kdn	1.37E-08	12	150	250	27544	883	< 1.0E-06	< 5.10E-05
Irf7	1383564_at	V$IRF7_01	J666808	liver	6.06E-12	12	110	175	27544	893	< 1.0E-06	< 5.10E-05
Irf7	1383564_at	V$IRF7_01	J666808	lv	2.60E-13	12	85	175	27544	871	< 1.0E-06	< 5.10E-05
Atf4	1367624_at	V$ATF4_Q2	J5383383	adr	8.30E-11	17	300	200	27544	868	< 1.0E-06	< 5.10E-05
Atf4	1367624_at	V$ATF4_Q2	J540732	adr	8.30E-11	17	300	200	27544	863	< 1.0E-06	< 5.10E-05
Tcf4	1368841_at	V$EBOX_Q6_01	J1268931	kdn	2.11E-07	21	600	254	27544	862	2.10E-05	8.81E-04
Stat1	1387354_at	V$STAT_Q6	gko-88g11_rp2_b1_418	lv	3.73E-07	13	78	801	27544	873	3.20E-05	1.22E-03
E2f1	1382511_at	V$E2F1_Q4	WKY-G-i-07e07_r1_602	aorta	1.37E-06	7	600	27	27544	902	1.20E-04	3.81E-03
Atf1	1389623_at	V$CREBATF_Q6	J519566	adr	1.81E-06	35	400	1002	27544	873	1.51E-04	4.53E-03
Atf1	1389623_at	V$CREBATF_Q6	J557374	adr	2.62E-06	56	800	1002	27544	867	2.04E-04	5.58E-03
Srebf2	1371979_at	V$SREBP_Q3	J489757	fat	3.23E-06	2	2	50	27544	901	2.61E-04	6.52E-03
E2f1	1382511_at	V$E2F1_Q4	WKYc86h09_s1_364	aorta	3.79E-06	7	700	27	27544	901	3.02E-04	7.10E-03

4.3.2 Extension of TF-networks by co-expression analysis

Irf7 is a master regulator of the type 1 interferon response [153] and genes directly regulated by this TF may comprise the core components of a larger gene network. To capture the broader regulatory effects of the chromosome 15 locus, we constructed a co-expression network around the *Irf7* target genes from gene expression profiles across 7 tissues and 30 strains. For this analysis we adjusted expression values of each gene for tissue effects because we were only interested in the genetic variability across strains. Assuming independence of expression values between tissues within the same strain we have used all 203 samples for pairs of transcripts where both transcripts were measured in all tissues. For pairs of transcripts where at least one transcript was present only on the Affymetrix array RAE230 2.0 microarrays we used the available 116 samples to compute pair-wise Pearson correlation coefficients. The network is formally defined as the tuple (V, E) with V being the set of nodes or vertices and E the set of edges. Here V corresponds to the set of all transcripts profiled on the Affymetrix RAE230a gene array. Since we are interested in the network neighbourhood of the predicted *Irf7* targets we define the set $I \subset V$ of predicted targets. The set of Edges E is defined as all tuples (v_1, v_2) of transcripts where one is a predicted target ($v_1 \in I$) and the other is from the rest of nodes ($v_2 \in V \backslash I$) with significant Pearson correlation coefficients. We used a false discovery rate (FDR) approach to determine significance of Pearson correlation coefficients [154]. This approach estimates the FDR using a mixture model on the correlation coefficients with an empirical parametric null-distribution and a non-parametric alternative distribution. For the estimation of the null distribution we have used 99% of the data. We considered the correlation significant with $FDR < 0.001$. Finally, we have removed unconnected nodes leading to the final set of nodes V' of the expanded *Irf7* network. For the visualisation we then re-estimated the FDR using the same approach described above using the complete correlation matrix for the set V'.

The analysis of transcripts profiled in all seven tissues revealed a large gene network of 247 genes across seven tissues, which was further expanded to 305 genes in four of the seven tissues where additional gene expression data were available (False Discovery Rate, $FDR < 0.1\%$) (Supplementary Table 2 of [3]). Groups of co-expressed genes can describe biological pathways and gene ontology (GO) analysis of the network showed enrichment for specific biological processes, including immune response ($P = 3.6 \times 10^{-19}$), response to virus ($P = 2.5 \times 10^{-7}$) and acute inflammatory response ($P = 2.6 \times 10^{-5}$) (Supplementary Table 3). Based on these findings we designated the larger network the *Irf7*-driven inflammatory gene network (iDIN) comprising 305 genes (Figure 4.9).

4.3.3 Cross species cell type enrichment analysis

We examined the cell types in which the human and mouse orthologous of iDIN genes were most highly expressed (compared to their average expression levels) using an atlas of gene expression from 65 distinct mouse and 71 human tissues and cell types.

Gene expression profiles for both mouse and human were downloaded from Novartis BioGPS (`http://biogps.gnf.org/downloads`) listed under Gene Expression Omnibus (GEO) code GSE113316 and GSE1024617; disease cell types were removed from both datasets prior to analysis.

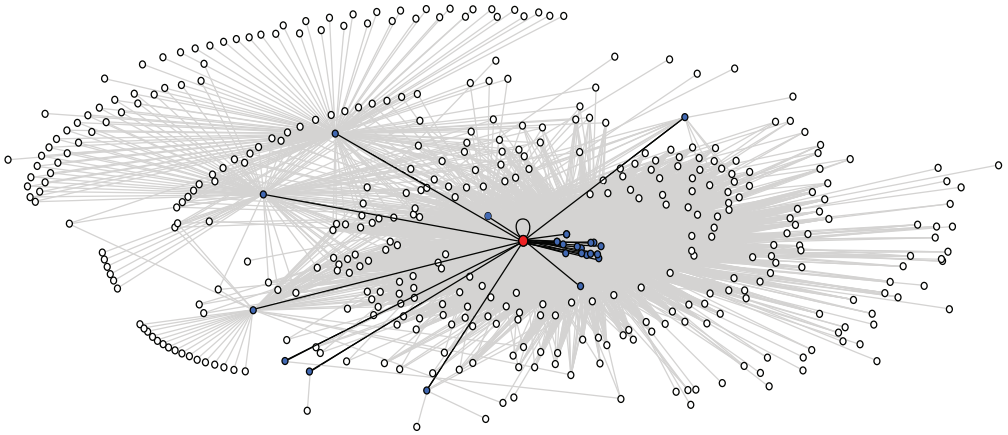

Figure 4.9: **The expanded *Irf7*-driven inflammatory gene network (iDIN).** Nodes represent genes, the node representing *Irf7* is coloured red and its predicted targets are coloured blue. Edges connect genes that are either predicted *Irf7* targets (black) or show significant Pearson correlation to one of the predicted targets (grey) ($FDR < 0.1\%$).

There were 135 genes in the rat iDIN for which there were both mouse and human orthologous and 71 human and 65 mouse cell types, respectively. We tested each gene for extreme expression in a given cell type, compared to its average expression across all cell types and tissues, using the Z-test separately in both species. Combined P-values were calculated across genes using Fisher's combined probability test where the logged P-values calculated across cell types were summed and multiplied by negative two, following equation [155, 156]:

$$X^2 = -2\sum_{i=1}^{k} \ln p_i. \tag{4.8}$$

The test statistic follows a χ^2 distribution with $2n$ degrees of freedom, with n being the number of tests performed. A Bonferroni correction was then applied to the combined P-values to correct for multiple testing.

iDIN transcripts were most enriched for expression in mouse bone marrow macrophages ($P =$

1.6×10^{-159}) and human monocytes ($P = 6.0 \times 10^{-177}$) with high levels of expression in other immune cells, including B lymphocytes (Supplementary Figure 1 of [3]).

4.3.4 Genetic mapping of the expanded TF-network

Whilst a core of 23 *Irf7* target genes mapped as trans-eQTLs to a single locus on rat chromosome 15 the overall genetic control of the iDIN remained to be determined. To investigate to what extent the iDIN was regulated by common genetic loci ('hot-spots')[124] we used sparse Bayesian regression models [157] to determine the association between expression levels of the network genes across seven tissues and genome-wide SNPs. For each tissue, we identified the major regulatory 'hot-spots' for control of the iDIN ($FDR < 1\%$). The same single locus on rat chromosome 15q25, which controlled *Irf7* and its targets in *trans*, was also associated with iDIN expression in all tissues and showed the strongest evidence for common regulation in five out of seven tissues with increased expression associated with the allele of the hypertensive strain (Figure 4.10). Since the iDIN may represent a molecular signature of macrophages that are associated with risk of common inflammatory diseases [158] and other diseases, such the autoimmune disease, type 1 diabetes [159], we characterised expression of *Cd68*, an established marker of macrophages [160], in SHR and BN hearts and the RI strains. In parental strains, *Cd68* mRNA levels were elevated in the SHR as compared to BN heart ($P = 0.01$), which reflected increased numbers of macrophages ($P = 2 \times 10^{-22}$); in the RI strains, *Cd68* and macrophage expressed genes were under trans-acting genetic control at the chromosome 15q25 locus that regulates the iDIN (Supplementary Figure 2 of [3]).

4.3.5 Identification of a candidate regulatory factor

We then analyzed genetic variation in the RI panel using SNPs [1] from the 15q25 region and determined the expression of iDIN genes in an additional seven inbred rat strains of known genotype that refined the locus to a 700kb region (Supplementary Figure 3 of [3]). This region contained seven annotated protein-coding genes and genetic variation in these genes between parental strains was characterised using the recently generated SHR genome sequence [161]. Of the genes in the region *Dock9*, *Ebi2* and *Tm9sf2* exhibited DNA variation, which was synonymous for *Dock9*, non-synonymous but not predicted to be functional for *Tm9sf2* and a 5'UTR SNP for *Ebi2* (Supplementary Table 4 of [3]). We evaluated gene expression of all transcripts in the region using RNA-Sequencing and quantitative PCR. *Ebi2* was the only differentially expressed gene between parental strains within the region ($P = 0.016$), was cis-regulated in the RI panel in heart and kidney and was enriched for expression in myeloid cell types (Supplementary Figure 3 and 4 of [3]). We assessed the effect of the *Ebi2* 5'UTR SNP by luciferase assay, the SHR allele resulted in reduced luciferase activity as compared to the BN allele (Supplementary Figure 4 of [3]).

Ebi2 (or *Gpr183*) encodes an orphan G protein-coupled receptor (GPCR) that controls B cell migration [162, 163] and, hence, we hypothesised that genotype-dependent *Ebi2* expression affects activity and migration of macrophages and underlies the iDIN regulatory effect of the chromosome 15q25 region. In the rat we localised *Ebi2* expression to $Cd68^{+ve}$ macrophages within the heart (Supplementary Figure 5 of [3]), an observation that we confirmed and extended

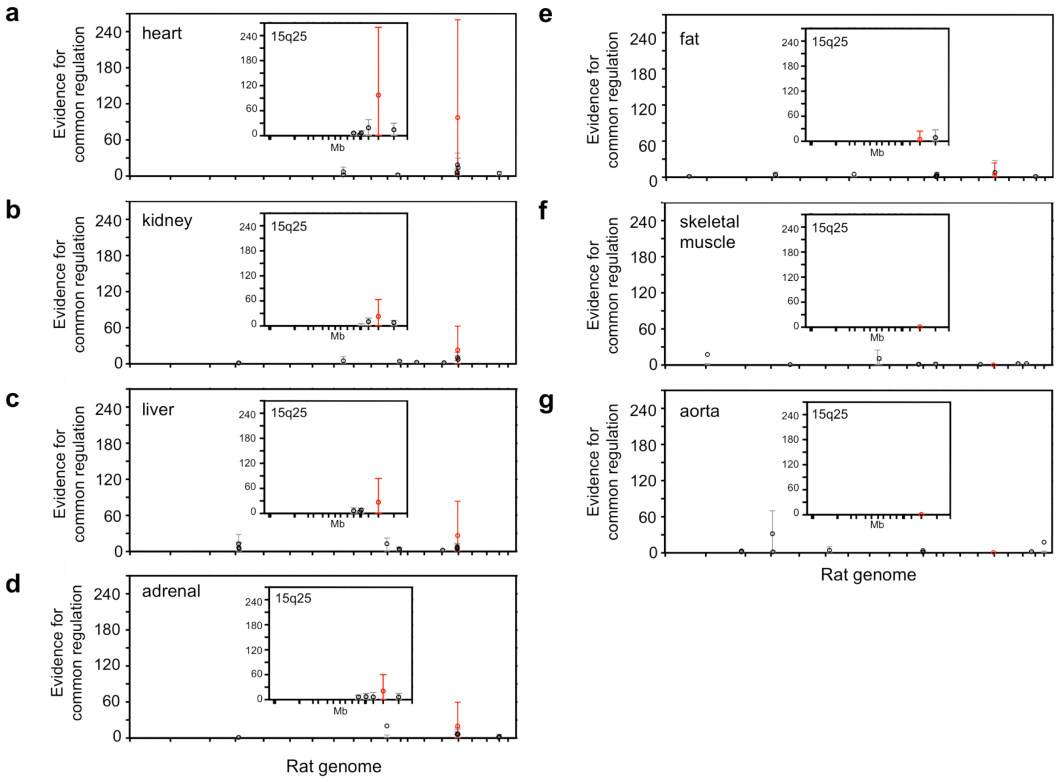

Figure 4.10: **Genetic mapping of regulatory 'hot-spots' for the iDIN.** a-g For each rat autosomal chromosome (x-axes), the strength of evidence for a SNP being a regulatory 'hot-spot' for controlling the network is measured by the average Bayes Factor (y-axes). Controlling the FDR at 1% level for each eQTL, the average Bayes Factor indicates the evidence in favour of common genetic regulation versus no genetic control, and is reported as a ratio between the strengths of these models (see Supplementary Information). For the 10 largest regulatory hot-spots the average Bayes Factors (circles) and their 90% range (5th-95th percentiles) are reported; a single SNP (J666808) that is consistently and most strongly associated with the network in 5 out of 7 tissues is highlighted in red. Inserts, average Bayes Factors and 90% range for the SNPs on rat chromosome 15q25 (87,479,238 - 108,949,015 bp). SNP positions in the region are indicated by tick marks.

across tissues (pancreas, liver, kidney and heart) in the $Ebi2^{GFP/+}$ mouse [162] (Supplementary Figure 6 of [3]). To assess whether *Ebi2* directly regulates iDIN gene expression we performed siRNA knockdown of *Ebi2* in primary cultures of rat macrophages (Supplementary Figure 7a of [3]). We show that silencing *Ebi2* increases expression of *Irf7*, the central hub of the iDIN, and of iDIN genes (Figure 4.11).

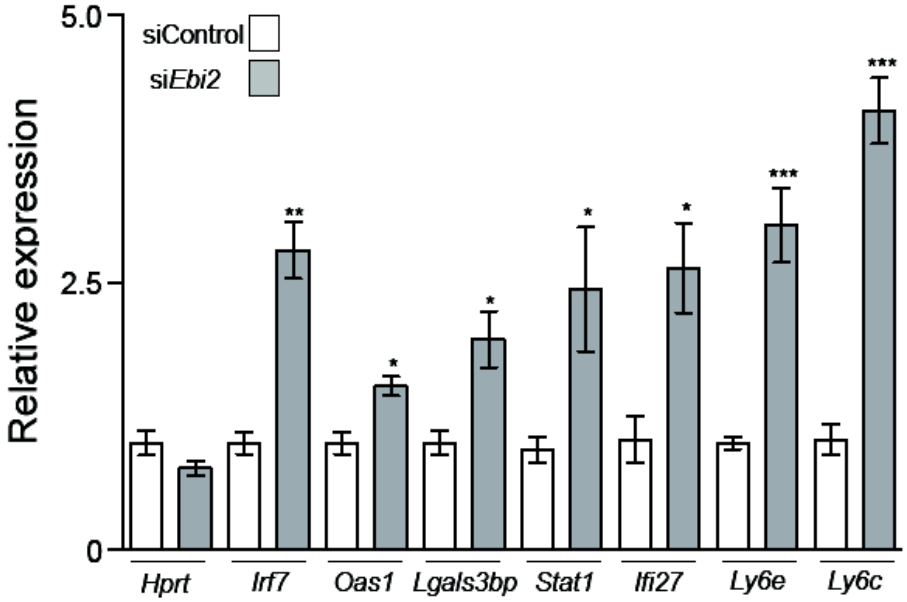

Figure 4.11: **Effect of siRNA-mediated knockdown of *Ebi2* on expression of *Irf7* and iDIN genes in rat bone marrow-derived macrophages.** As compared to control siRNA, siRNA against *Ebi2* (siEbi2) significantly down-regulated (97% inhibition) *Ebi2* mRNA expression. While siEbi2 had no effect on control gene expression (*Hprt*) there was a significantly increase in expression of *Irf7* and of iDIN genes *Oas1*, *Lgals3bp*, *Stat1*, *Ifi27*, *Ly6e* and *Ly6c*. Data, normalised to β-actin levels (see [3]), are shown as means relative to control ± SEMs. *, $P < 0.05$; ** $P < 0.01$; *** $P < 0.005$.

4.3.6 Comparative co-expression analysis with humans

To translate our findings to humans, we first tested whether the iDIN was recapitulated in human immune cells using genome-wide expression data from monocytes isolated from 1,490 individuals from the Gutenberg Heart Study (GHS) [164]. We performed TFBS enrichment analysis, analogous to that performed in the rat, and expanded the *IRF7* network by co-expression analysis (Supplementary Table 5 and see Supplementary Information). Here, we were interested in the TF - target relationships without a special focus on the genetic regulation. Therefore we ranked all potential TF targets of *IRF7* according to their co-expression with the TF, as measured by the absolute Pearson correlation coefficient. Analogously to the rat TFBS enrichment analysis (section 4.3.1) we obtained sequences 200 bp upstream of the transcription start

site (TSS) as annotated in the ENSEMBL database and ranked them according to predicted TF affinities. Subsequently we quantified the enrichment of predicted TFBS in the top 1,000 co-expressed genes using PASTAA [143]. Subsequent network expansion was performed using the same procedure as in the rat (section 4.3.2) .

In order to replicate results obtained from the GHS data, we analysed gene expression data from a distinct cohort of 758 subjects from the Cardiogenics Study [165]. We performed the same TFBS enrichment analysis as in the GHS data set. Since the cardiogenics study involves multiple centres, we have adjusted gene expression data for the center prior to the analysis. We correlated expression profiles of all genes that were analyzed for the Gutenberg Heart Study (GHS) versus the predicted *IRF7* targets. In keeping with the analysis in the GHS study, we applied a threshold of $FDR < 0.001$ to identify the significant set of adjacent genes (i.e., co-regulated genes). This threshold corresponded to an absolute correlation of at least 0.53.

The analysis of the GHS data suggests that *IRF7* regulates nine of its direct target genes (Supplementary Table 5 of [3]). Expansion of the network results in a set of 531 co-expressed genes and is most strongly enriched for Gene Ontology (GO) terms "response to virus" ($P = 1.88 \times 10^{-13}$) and "immune response" ($P = 1.28 \times 10^{-9}$).

In the Cardiogenics data we found the same set of co-regulated *IRF7* target genes and one additional target. We have identified a set of 791 genes that are co-expressed with the core set of predicted *IRF7* targets. The overlap with the GHS expression network comprises 186 genes ($P = 8.32 \times 10^{-23}$) and Gene Ontology (GO) enrichment analysis showed the strongest enrichment for "immune response" ($P = 1.31 \times 10^{-11}$) and "response to virus" ($P = 4.68 \times 10^{-11}$) categories, respectively.

In order to compare the genes in the rat and human iDIN we used ENSEMBL to derive orthologous genes between rat and human. Of all genes represented on the rat and human expression arrays we were able to identify a common set of 9,909 human genes that had a rat orthologous and were represented on both expression arrays. We defined this as the "orthologous expression set". The human iDIN in the GHS contained 508 Ensembl genes (Supplementary Table 5 of [3]) of the orthologous expression set. Out of the 305 Ensembl rat iDIN genes (Supplementary Table 2 of [3]), 248 Ensembl genes were contained in the orthologous expression set. The overlap between the two sets was 51 Ensembl genes ($P = 9.1 \times 10^{-20}$), which is reported in the supplementary information of [3].

The human *IRF7*-driven network exhibited strong cross-species overlap with orthologous of the rat iDIN ($P = 9.1 \times 10^{-20}$), and was annotated by the same gene ontology terms (immune response, inflammatory response, response to virus) (Supplementary Table 6 of [3]).

4.3.7 Analysis of the human regulatory locus

Association of human iDIN with the chromosome 13 locus

We determined whether the human chromosome 13q32 locus (spanning 1 Mb, Supplementary Table 7), which is orthologous to the critical rat chromosome 15q25 region, was associated with expression of the *IRF7* network genes in humans. For each SNP at the chromosome 13q32 locus, multivariate analysis of variance (MANOVA) [166] was performed to test the hypothesis that the mean monocyte expression levels of *IRF7* and all predicted *IRF7*-target genes are the same in

all genotype categories in the GHS and Cardiogenics cohorts. Among the different statistics that can be used to evaluate the MANOVA hypothesis, we employed the Pillai' trace, which is more robust to violations of normality and homogeneity of dispersion [167]. Significant associations between network genes and SNPs at the locus were assessed using Storey's FDR at the 5% level [56].

We also verified the validity of the two most important assumptions of MANOVA analysis, namely, multivariate normality and, conditionally to each SNP, the homogeneity of variance-covariances matrices. We performed multivariate Box-Cox transformation to protect against deviations from the normality assumption, while the homogeneity of variance-covariance matrices among groups was assessed using Box's M statistics [168]. To ensure that the MANOVA results presented here are not affected by these factors, we repeated all analyses using the rank-based Wilks' lambda MANOVA [169], which is robust to violations of both conditions. The results of the rank-based MANOVA analysis supported the original MANOVA findings in the Cardiogenics cohort for six out of seven SNPs in the region (see supplementary information of [3]).

Multivariate analysis of variance of the Cardiogenics monocyte expression and genotype data revealed that six SNPs in the 13q32 region (including rs9557217, $P = 5.0 \times 10^{-5}$; and rs9585056, $P = 1.1 \times 10^{-3}$) were associated with *trans*-regulated expression of *IRF7* and *IRF7* target genes (fourth panel of Figure 4.12). We did not, however, detect a signal for *trans*-regulation of *IRF7* or *IRF7* target genes at the 13q23 locus in the GHS cohort (bottom panel of Figure 4.12). This could be due to differences in the two datasets (different monocyte preparation protocols between the two cohorts, see supplementary information of [3]), as we demonstrate in the following analysis of *EBI2* expression.

Association of genetic variation and EBI2 expression

We then examined whether monocyte gene expression of the human gene *EBI2*, was under *cis*-regulatory control. Association between SNPs at the chromosome 13 locus and *EBI2* mRNA abundance was assessed by linear regression. The subset of SNPs sufficient to explain *EBI2*-gene-expression variation at the locus was determined by using the lasso shrinkage and selection method for linear regression [170] (see also section 2.1.4). The optimal shrinkage parameter λ was selected using ten-fold cross validation implemented in [171].

Lasso selection of *EBI2* eQTL models in GHS resulted in a set of three SNPs (rs9585056, rs9517723, rs7325697). When adding imputed SNPs, rs9517725 explains most of the variation of the *EBI2* expression ($P = 6.8 \times 10^{-13}$) at this locus. Lasso model selection in Cardiogenics yielded an overlapping set of three SNPs (rs9557217, rs9585056, rs9517725) highlighted in Figure 4.12. A formal hypothesis test [172] of a common causal genetic variant was not rejected ($P = 0.14$).

In both the GHS and Cardiogenics cohorts, *EBI2* showed evidence of *cis*-regulation at the 13q32 locus but this differed between the two cohorts (most associated SNPs: Cardiogenics, rs9585056, $P = 2.2 \times 10^{-8}$; GHS, rs9517725, $P = 6.8 \times 10^{-13}$) (Figure 4.12). Two of the five SNPs, rs9557217 and rs9585056 contained in the model explaining *EBI2* expression also exhibited a significant *trans*-effect on iDIN expression in the Cardiogenics cohort (section 4.3.7 and Figure 4.12), suggesting common regulatory control by this locus on the *IRF7* network and

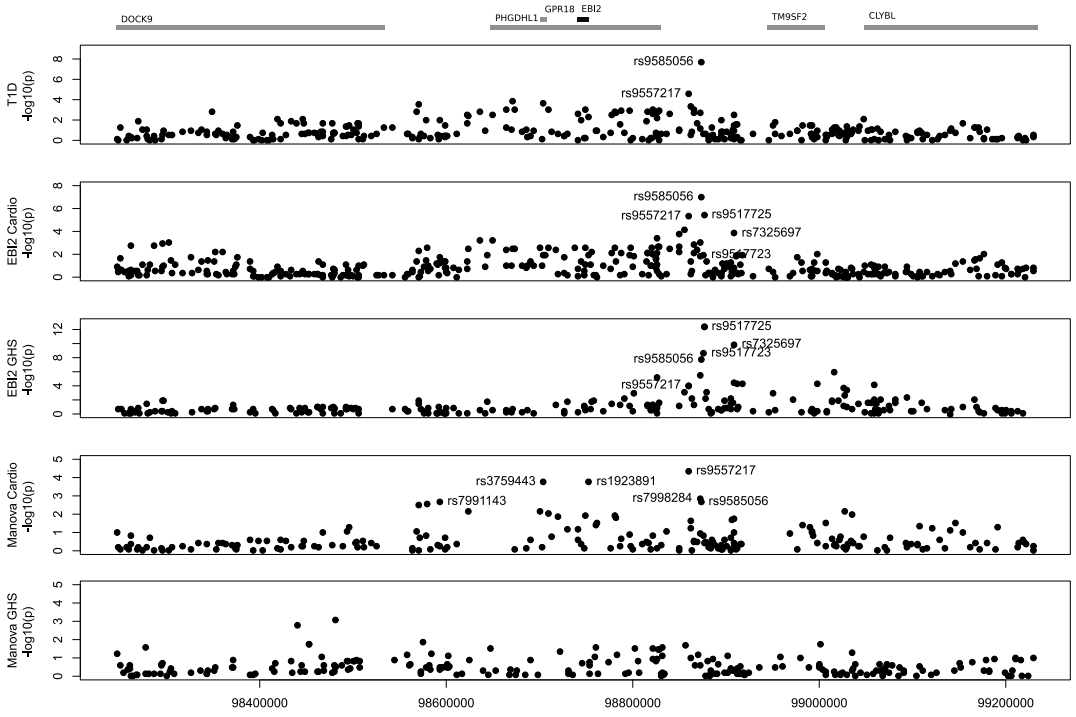

Figure 4.12: **A regulatory locus for T1D risk.** Results of *EBI2* eQTL analysis in GHS (top panel), Cardiogenics (second panel) and T1D association (third panel) at the human chromosome 13 locus that is orthologous to the 700kb rat chromosome 15q25 region. The upper panel shows the nominal -$\log_{10}$ P-values of marker regression against gene expression of *EBI2* for all SNPs in the region. The third panel shows the -$\log_{10}$ P-values of T1D association with SNPs in the region. SNP rs9585056 showed the strongest association with T1D ($P = 7.0 \times 10^{-10}$) amongst the genotyped markers.

EBI2 expression.

4.3.8 Translation to human GWAS data

Association of iDIN genes with T1D

Monocyte-derived macrophages are critical determinants of inflammatory processes important for common disease [173] and autoimmune type 1 diabetes (T1D) [174]. The iDIN, which is expressed in macrophages, is enriched for viral response genes and contained *IFIH1*, a well characterised T1D susceptibility gene [175, 176]. A link between *IRF7* and *IRF7*-regulated genes and T1D was supported by data in the non obese diabetic mouse, where blockade of the interferon alpha receptor down-regulates *IRF7* and iDIN genes' expression in immune cells and

delays and attenuates T1D development [177]. This prompted us to evaluate the association of the human orthologous of rat iDIN genes and genes in the human monocyte network (Figure 3) with T1D.

To examine whether genes in the network were generally associated with T1D, we divided SNPs genotyped in T1D genome-wide association scans into two:

A: those within 1Mb of any gene in the network, and

B: those within 1Mb of any ENSEMBL gene not in network, excluding any in A

We used a Wilcoxon rank test to examine whether the distribution of association statistics varied between sets A and B, on the basis of their one degree of freedom χ^2 test statistics. Note that this approach tests the null hypothesis that the pattern of association is the same in the two datasets (Q_1 from Eq.2.21), rather than a null that no SNPs in A are associated with T1D [87] (Q_2 from Eq.2.22). We chose to use a nonparametric approach due to the extreme right skew displayed by test statistics for T1D; χ^2 values above 300 are seen for SNPs in the strongly associated HLA region. The more commonly used gene set enrichment analysis is based around the Kolmogorov-Smirnov test, which is underpowered for detecting differences in distributions [178]. To avoid potential confounding by allele frequency, we applied inverse probability weighting to the ranks used in the Wilcoxon according to a propensity score measuring the chance of a SNP appearing in the *inside* network group, given minor allele frequency [94, 95]. The propensity score was calculated by binning minor allele frequencies into bins of width 0.05 and taking the ratio of the number of SNPs in A+B and the number in A.

Correlation between SNPs is substantial in GWA data. While the Wilcoxon test has the helpful property that the mean under the null is unaffected by correlation, correlation does lead to inflation of the standard deviation of the test statistic which we estimated by 200 permutations of the case control labels. Finally, because the GWA datasets were generated on different chips (Affymetrix and Illumina) we used a stratified Wilcoxon [97] to avoid confounding by chip.

Because we have both human- and rat-derived networks, and because HLA shows extreme association with T1D, we performed four tests:

1. genes in either rat or human networks
2. genes in both rat and human networks, restricting set B to genes with a rat orthologous
3. genes in both rat and human networks, excluding SNPs from A or B within a wide window around human MHC (chr6:29000000..34600000)
4. genes in both rat and human networks, restricting set B to genes in the rat orthologue database and excluding SNPs from A or B within a wide window around human MHC (chr6:29000000..34600000)

SNPs close to ($\leq$ 1Mb) any iDIN genes were significantly more likely to associate with T1D in large-scale GWAS than SNPs close to genes not in the network (i.e. the rest of the genome), with the strongest signal for the larger network formed by the union of genes in the rat iDIN and the human *IRF7*-driven network ($P = 2.4 \times 10^{-10}$). Since many immune genes have been associated to T1D previously, we also performed the above mentioned tests against a background set of all

genes annotated by the GO term "immune response". This established an overrepresentation of T1D associated genes in the union network ($P = 8.85 \times 10^{-6}$), indicating that the iDIN more specifically categorises T1D genes than the GO term "immune response". These data demonstrate that co-expression networks across species provide functional annotation of genes in biological processes that can be used to detect the signal of common genetic variation of small effect that is usually not reported by typical GWAS.

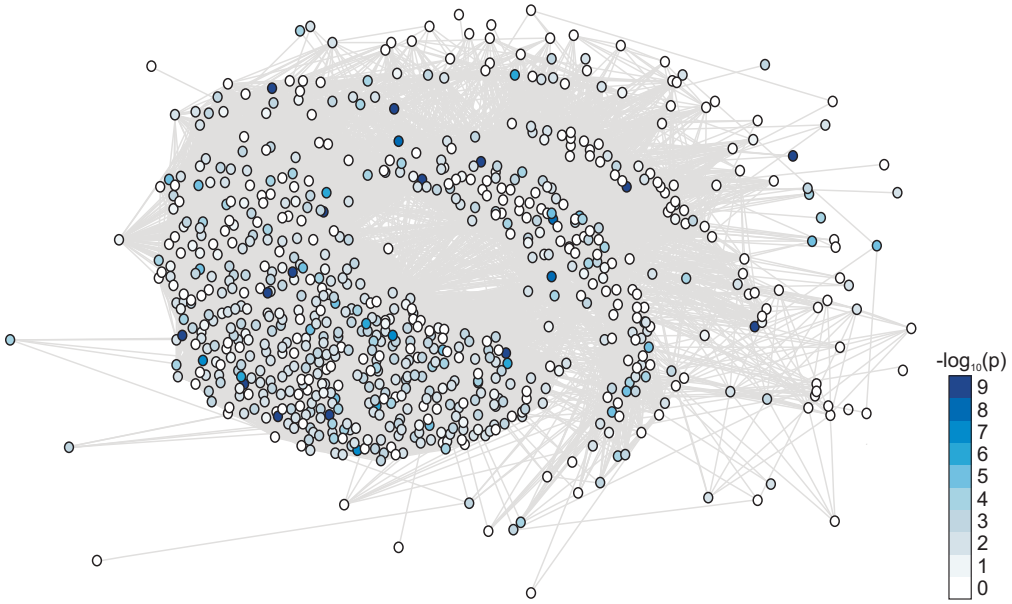

Figure 4.13: **Schematic representation of the union of *IRF7*-driven gene networks that was created using the set of human orthologous of rat iDIN genes and human iDIN genes.** A Wilcoxon rank test based gene set enrichment analysis (modified from Holden et al.) [179] showed SNPs close to iDIN genes to be significantly more likely to associate with T1D in large-scale GWAS than SNPs close to randomly selected genes ($P = 2.5 \times 10^{-10}$) and randomly selected immune response genes ($P = 8.8 \times 10^{-6}$). Nodes represent iDIN genes and the node colour indicates the P-values (-$\log_{10}$ scale) of the association with T1D (see Methods).

Association testing of rs9585056 with T1D

Logistic regression was used to test for association in case control samples, stratifying by broad UK region to control for population structure. Family data were analysed by transmission disequilibrium test, splitting multiplex families into parent offspring trios and using a pseudo-case control framework to estimate allelic effects. A score statistic was also generated, and

a score test for association in case-controls and families combined conducted by summing the scores and variances according to the method proposed by Mantel [180].

In a GWAS meta-analysis of T1D in 7,514 cases and 9,045 controls [181], we found evidence for association of the chromosome 13q32 region at SNP rs9585056 ($P = 1.3 \times 10^{-7}$) that had not been reported before (Figure 4.12). We genotyped this SNP in two independent large cohorts and increased the strength of the T1D association (combined $P = 7.0 \times 10^{-10}$, odds ratio (95% confidence interval) = 1.15 (1.09-1.21), Supplementary Table 9 of [3]). The minor C allele of SNP rs9585056 was associated with T1D risk, lower *EBI2* expression levels of iDIN genes in the Cardiogenics cohort. Although we cannot discriminate between single and multiple causal variants, overall, these results show an overlap of association signals in the same region on human chromosome 13q32 for iDIN genes, *EBI2* cis-regulation and T1D.

4.3.9 Conclusions

The immunopathology of autoimmune T1D is characterised by infiltration of the pancreas with B and T lymphocytes and macrophages with resultant beta-cell death, insulin deficiency and hyperglycaemia [159]. We have demonstrated that genes in the iDIN contribute to T1D risk and implicate the innate viral response pathway and macrophages in the aetiology of T1D. Genetic control points that perturb biological networks are thought to represent important loci for disease risk [22] and we propose that the T1D susceptibility locus that we identified regulates innate immune response genes in macrophages, as we demonstrated in the rat. *EBI2*, an orphan GPCR, which controls *IRF7*, a master regulator of the innate immune response [153], represents a candidate for *trans*-regulation of the human iDIN and for T1D risk. A role for *IRF7* in the pathogenesis of T1D is supported by functional studies of other T1D candidate genes, namely *TLR7* and *TLR8* [182], which act through *IRF7* [183]. In keeping with previous observations in lower organisms [122], our data support GPCRs as key control points of *trans*-regulated gene networks. The integrated analyses we used here highlight the power of cross-species network approaches that can be used to reveal loci, genes and biological pathways associated with human disease.

4.4 Co-expression as quantitative trait

Analysis of regulatory variants has proven to speed up the identification of genes underpinning physiological QTLs (see section 3.4). However, the proportion of the phenotypic variance explained by these loci is rather small [24] and hints at a more complex mode of inheritance. In this light it is recognised that disease genes are not acting in isolation [22], but through complex pathways that lead to the physiological endpoint. In most previous work on the identification of disease pathways from gene expression data in a segregating population [22, 28, 26, 27, 25, 3] (see also section 4.3), the focus has been to identify networks of co-regulated genes, where the expression is determined by the sequence variant that has been found to be associated to the disease phenotype. What has been neglected so far, is that the observed phenotype could also be a consequence of genotype dependent perturbations of co-expression.

Here we propose a differential network analysis method to detect genotype dependent co-expression across multiple tissues as illustrated in Figure 4.14 in an unbiased genome wide

manner inspired by [184]. We apply our method to genetic and gene expression data from a set of recombinant inbred (RI) strains called BXH/HXB (see section 1.4). Using our method we were able to identify eQTL genes and their neighbourhood of differentially co-expressed genes in a co-expression network. This allowed us to put the eQTL genes into the context of genes with which they are usually interacting, dependent on the genotype at the genetic marker: e.g. genes were co-expressed in RI strains carrying the wildtype allele but not co-expressed in the strains carrying the mutated allele or vice versa. Using a test for enrichment [185] of Gene Ontology terms [81] (see also section 2.2.1) within such a neighbourhood allowed to identify the functional context in which the eQTL gene acts. Furthermore, the topology of the resulting graph allowed us to identify pairs of eQTL, linked by sets of common neighbours. These links are opening up the opportunity to explore epistatic interactions influencing blood pressure between connected eQTL loci.

4.4.1 A linear model for the mapping of co-expression as a quantitative trait

The expression data, that we were modelling came from $N = 29$ RI strains. For each of the strains, $t = 7$ tissues have been profiled using Affymetrix chips. We assume that the gene expression in different tissues is independent within the same strain. We organised the data into the $(p \times Nt)$ expression matrix E with p genes and Nt samples. E_i denotes the expression vector of gene i.

In order to detect pairs of genes that are co-expressed and therefore co-regulated in a genotype dependent way, we designed a linear model. In contrast to eQTL mapping, where the aim is to detect pairs of transcript and marker, with the marker influencing the expression of the transcript, here we were interested in triplets consisting of two transcripts and one marker. In particular the situation we aimed to capture is that the marker influences the co-expression of the two genes across multiple tissues.

In general co-expression of two genes i and j can be detected using a linear model

$$M_{reduced} : E_{ik} = \beta_1 + \beta_2 E_{jk} + \epsilon_{ik}, \tag{4.9}$$

with $k \in 1, \ldots, Nt$. The presence of co-expression is equivalent to a non-zero slope β_2. Assuming that the error ϵ is normally distributed the hypothesis of $\beta_2 = 0$ can be tested [79]. Note that a linear model assumes that the independent variables are non-random. Therefore, strictly speaking, the model should be called a multiple regression model instead. However, with the additional assumption that the joint probability distribution of two gene expression profiles is bivariate normal the linear model framework is equivalent to the multiple regression model (see section 2.1.3). Therefore we will use the term linear model synonymously.

In order to detect the influence of a genetic marker on the co-expression we include two more parameters in the model, namely one intercept and one slope for each of the genotypes as visualised in Figure 4.14. This specifies the full model

$$M_{full} : E_{ik} = bn_{mk}(\beta_1 + \beta_3 E_{jk}) + shr_{mk}(\beta_2 + \beta_4 E_{jk}) + \epsilon_{ik}, \tag{4.10}$$

where bn_{mk} and shr_{mk} are indicator variables for the genotype of strain k at the marker m.

Suppose the genotypes of M genetic markers for the N strains are arranged in a $M \times N$ matrix G and encoded as a factor with 2 levels BN and SHR. For each marker g, we formulate the full model in terms of a design matrix

$$E'_{i,\cdot} = Y = X\boldsymbol{\beta} + \epsilon,$$

with

$$X = I^{BN}(\mathbf{1}, \mathbf{0}, E'_{j,\cdot}, \mathbf{0}) + I^{SHR}(\mathbf{0}, \mathbf{1}, \mathbf{0}, E'_{j,\cdot})$$

and

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)',$$

where

$$I^{BN}_{k,l} = \begin{cases} 1 & \text{if } k = l \wedge G_{g,k} = BN \\ 0 & \text{otherwise} \end{cases}$$

is a $(Nt \times Nt)$ diagonal matrix where the k-th diagonal element is indicating the genotype of individual k. I^{SHR} is defined analogously. β_1 is the BN specific intercept, β_2 is the SHR specific intercept, β_3 is the BN specific slope and β_4 is the SHR specific slope.

If there is no genotype dependent effect on co-expression, the reduced model and the full model fit equally well, that is $\beta_3 = \beta_4$ in the full model and also equal to β_2 in the reduced model. If the marker has influence on the co-expression, the full model will fit better. Analysis of variance of the two nested models can then be used to assess the statistical significance of the difference of fits.

Assuming that the error terms $\epsilon_k \sim \mathcal{N}(0, \sigma)$ are independent and have the same variance σ, we use the hypothesis test $\mathbf{l}\boldsymbol{\beta} = l_0$ described in section 2.1.2. The case where $\beta_3 = \beta_4$ is equivalent to $\mathbf{l} = (0, 0, 1, -1)'$ and $l_0 = 0$. The test statistic is defined in Eq.2.11 and follows an F distribution with 1 and $(Nt - 4)$ degrees of freedom. If we can reject this hypothesis we can assume genotype dependent co-expression of the two genes i and j.

In order to perform a genome scan across all genetic markers we assume independence of the markers and use these P-values obtained from our model and correction for multiple testing.

In cases where genotype dependent differences in the co-regulation were detected by the linear model, the linear relationship of the two genes was assessed using the Pearson correlation coefficient for each group of strains as a measure for the goodness of fit. We needed this additional test, in order to filter out cases where there is a difference in the estimated slopes, but the linear models are not fitting the data well. This is mostly the case when the linear relationship between the genes is not given.

We have implemented a computer program to systematically test given pairs of transcripts against all genetic markers using this linear model. The code is written in C, using lapack [186] for the model fitting and convenient to use through an interface for R [117]. Software is available on our website.

4.4.2 Construction of the co-eQTL graph

Gene expression data was normalised using the RMA algorithm [120] described in section 3.2.2. For our analysis we filtered out genes that are expressed below the 15% quantile of all gene expression values in more than 60% of the samples. Furthermore we filtered out genes with a

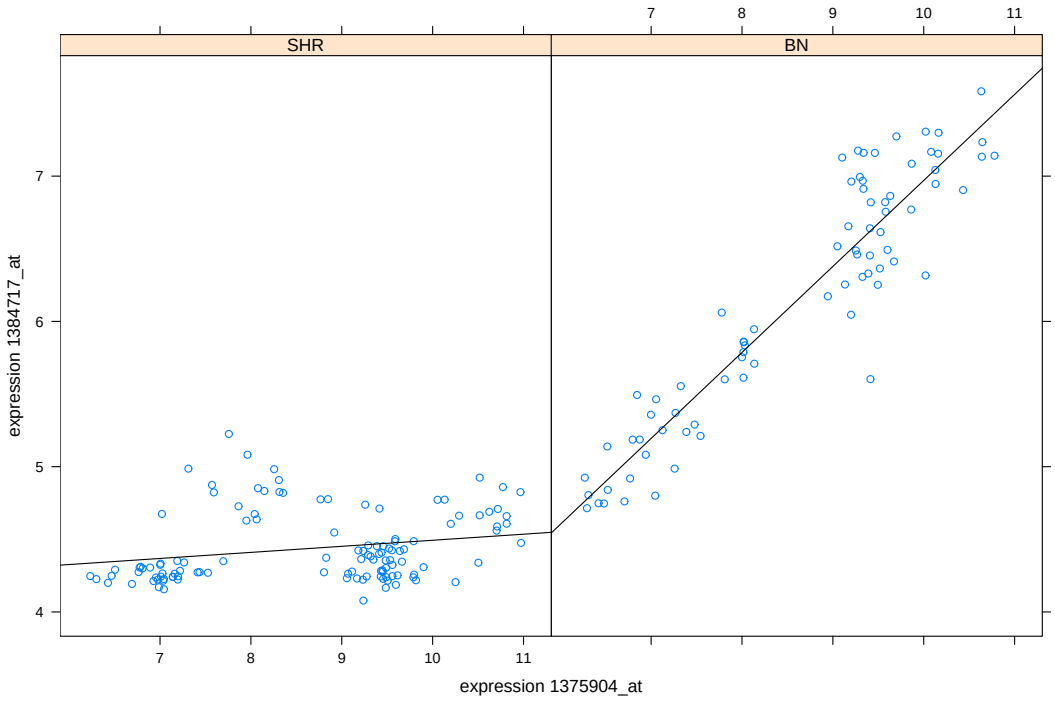

Figure 4.14: **Genotype dependent co-expression.** The left panel shows the expression of probe set 1384717_at plotted against the expression of probe set 1375904_at and the fit of the linear model across seven tissues and all RI strains with the SHR genotype at marker B07P0680. The right panel shows the same plot but for RI strains with the BN genotype at the marker B07P0680. Comparing the two slopes using a test on a linear model with an intercept and a slope for each of the groups versus a linear model with only one intercept and one slope clearly rejects the hypothesis of equal slopes ($F = 1184.35, p = 9.8 \cdot 10^{-109}$). In this case we speak of genotype dependent co-expression or a co-eQTL.

coefficient of variation below the 75% quantile of the distribution of coefficients of variation of all genes. This resulted in a set of $p = 3579$ most varying genes across seven tissues for the mapping of co-expression QTLs. These genes make up the set of nodes of the transcriptional network under investigation. The number of potential edges in this network is $(p \cdot (p-1))/2$, which have to be tested against 1400 genetic markers of the genetic map from the STAR project [1] described in section 3.3.1. Obviously this would result in a massive multiple testing problem. Therefore the number of edges for genetic mapping has been reduced by a filter. Only if the variance of the co-expression of two genes across the 29 strains is large we expected a genetic effect on the co-expression of two genes. Therefore, we have computed correlation coefficients for each pair of genes corresponding to potential edges in the network across the seven tissues

within each of the strains. We have retained 185702 edges where the difference between the minimum and maximum correlation coefficient was at least 0.5 and the variance was larger than the 90th percentile of variances for all edges. We have applied our linear model to these edges and reported the marker with the highest influence on the co-expression for each edge. All edges with $P < 10^{-12}$ and the maximum absolute value of the two correlation coefficients $> 0.7 (P < 10^{-12})$ have been used to construct the genome-wide graph of genotype dependent co-expression shown in Figure 4.15. The P-value cutoffs corresponds to a Bonferroni adjusted P-value of $P < 10^{-6}$ and were selected to guarantee stringent fits of the models to the data as can be visually verified in plots like Figure 4.14.

4.4.3 Topological analysis of the co-eQTL graph

In total there are 1546 edges representing genotype dependent co-expression between 890 genes, controlled by 24 genetic markers. Colouring edges according to the associated marker revealed subnetworks that are arranged in star topologies around central hub nodes (Figure 4.15). Hub genes are listed in Table 4.4. For further analysis we refer to these star topologies as subnetworks and call the group of genes that are differentially co-expressed with the hub co-eQTL genes.

Another feature of the co-eQTL graph is that hubs are connected to each other by several paths of length two through a set of intermediate genes. These genes are co-expressed with both hub genes dependent on two different genetic markers. For each pair of hubs we define the interface of the hubs as the intersection of the direct neighbours of the two hubs. Interestingly, there are interface genes that are connected to up to 15 hub genes.

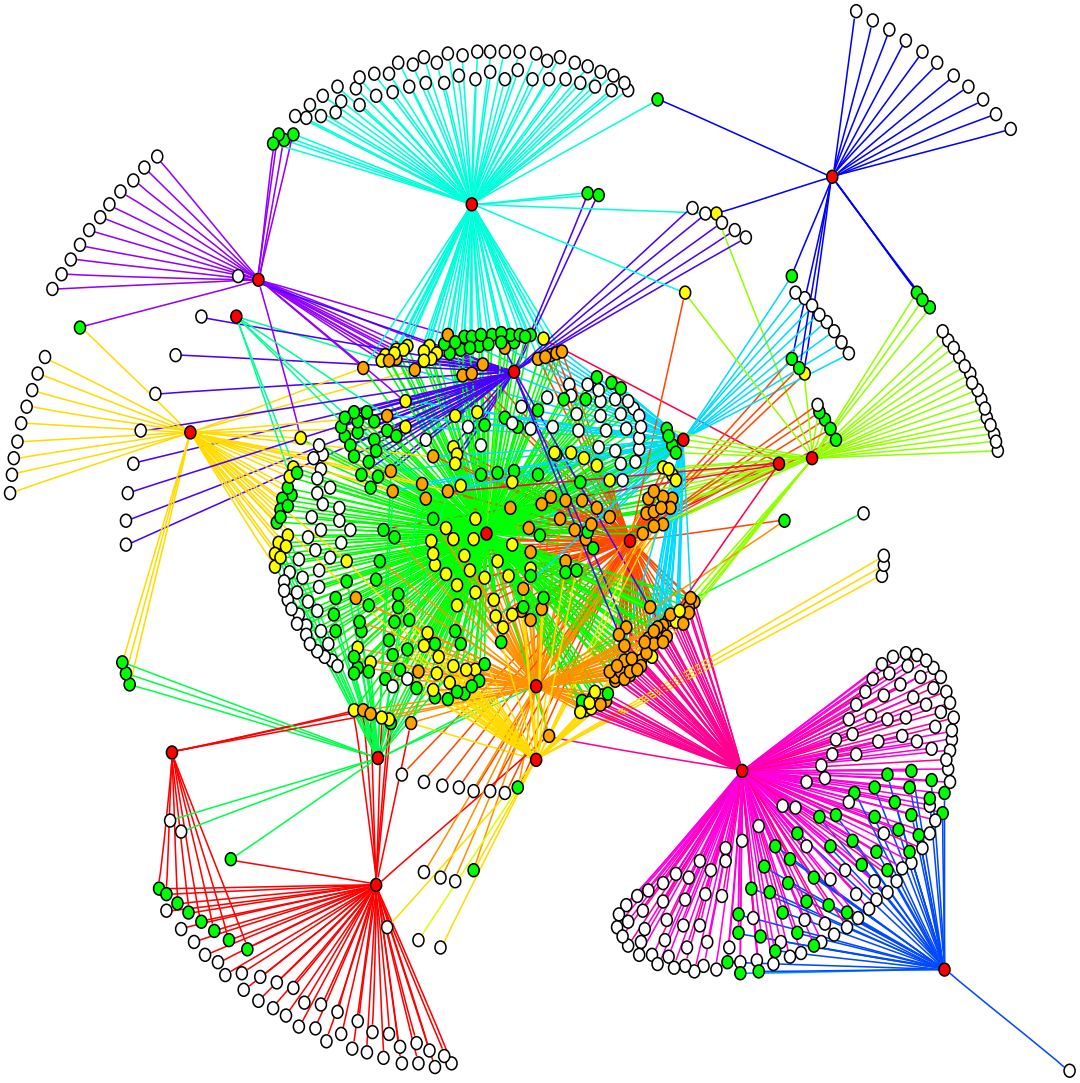

Figure 4.15: **Genome wide view of genotype dependent co-expression.** The graph shows all pairwise genotype dependent co-expression relations. The topology of the graph contains highly connected hub nodes, however the hubs are not directly connected to each other. The colours of the edges correspond to the genetic marker which induced the genotype dependent co-expression. This colouring of the graph reveals that edges linked to one genetic marker form a star topology. The hubs of these stars correspond to transcripts that have a marker dependent expression pattern (eQTL). Interestingly many hubs are connected by paths of length two.

Table 4.4: Hub transcripts

Symbol	Gene ID	Probe ID	Description
Asap1	ENSRNOG00000005739	1384717_at	Arf-GAP with SH3 domain, ANK repeat and PH domain-containing protein 1 (130 kDa phosphatidylinositol 4,5-biphosphate-dependent ARF1 GTPase-activating protein)(PIP2-dependent ARF1 GAP)(ADP-ribosylation factor-directed GTPase-activating protein 1)(ARF GTPase-activating protein 1)(Development and differentiation-enhancing factor 1)(Differentiation-enhancing factor 1)(DEF-1) [Source:UniProtKB/Swiss-Prot;Acc:Q1AAU6]
Bzw2	ENSRNOG00000005096	1377329_at	Basic leucine zipper and W2 domain-containing protein 2 (Brain development-related molecule 2) [Source:UniProtKB/Swiss-Prot;Acc:Q9WTT7]
	ENSRNOG00000032604	1373232_at	
Echdc2	ENSRNOG00000029333	1374527_at	enoyl Coenzyme A hydratase domain containing 2 [Source:RefSeq peptide;Acc:NP_001100145]
IPI00777098.1	ENSRNOG00000038701	1389990_at	
Snf1lk	ENSRNOG00000001189	1368597_at	Serine/threonine-protein kinase SIK1 (EC 2.7.11.1)(Salt-inducible protein kinase 1)(SIK-1)(Serine/threonine-protein kinase SNF1-like kinase 1)(Serine/threonine-protein kinase SNF1LK)(Protein kinase KID2) [Source:UniProtKB/Swiss-Prot;Acc:Q9R1U5]
Krt10	ENSRNOG00000030170	1373254_at	Keratin, type I cytoskeletal 10 (Cytokeratin-10)(CK-10)(Keratin-10)(K10)(Type I keratin Ka10) [Source:UniProtKB/Swiss-Prot;Acc:Q6IFW6]
RGD1311103	ENSRNOG00000025957	1389690_at	
IPI00204640.1	ENSRNOG00000004540	1377452_at	

Pik3r1	ENSRNOG00000018903	1370114_a_at	Phosphatidylinositol 3-kinase regulatory subunit alpha (PI3-kinase p85 subunit alpha)(PtdIns-3-kinase p85-alpha)(PI3K) [Source:UniProtKB/Swiss-Prot;Acc:Q63787]
IPI00777605.1	ENSRNOG00000039655	1373232_at	
	ENSRNOG00000037675	1389990_at	
NP_001099727.1	ENSRNOG00000019627	1373697_at	myosin binding protein C, fast-type [Source:RefSeq peptide;Acc:NP_001099727]
NP_001099727.1	ENSRNOG00000019627	1376968_at	myosin binding protein C, fast-type [Source:RefSeq peptide;Acc:NP_001099727]
Fam103a1	ENSRNOG00000019426	1376780_at	hypothetical protein LOC293058 [Source:RefSeq peptide;Acc:NP_001120923]
Cct6a	RGD1304763	1377006_at	chaperonin containing Tcp1, subunit 6A (zeta 1)
Ddx42	RGD1304909	1379896_at	DEAD (Asp-Glu-Ala-Asp) box polypeptide 42
Ifit1	RGD620599	1369836_at	interferon-induced protein with tetratricopeptide repeats 1 responses to dexamethasone and other inflammatory stimuli
RGD1562844	RGD1562844	1376958_at	similar to serine (or cysteine) proteinase inhibitor, clade B, member 9
Tnfaip6	RGD621359	1371194_at	tumor necrosis factor alpha induced protein 6 mouse homolog is a catalyst in the formation of the cumulus extracellular matrix and indispensable for female fertility
Znf655	RGD1309158	1376840_at	
RATVL30B		1370988_at	Rattus norvegicus transposon VL30, complete sequence
Ythdf2		1371960_at	YTH domain family, member 2
		1374583_at	EST only
		1371776_at	EST only
		1379497_at	EST only

4.4.4 Hub genes have eQTLs

Closer investigation of the hub genes revealed that their expression is controlled by the genetic marker that also controls the differential co-expression of the edges that define the subnetwork. Using the results of the eQTL analysis presented in section 3.3 we found that 50% of the hubs had eQTLs in all seven tissues, six had eQTLs in at least four tissues and the remaining six were more tissue specific with eQTLs in one or two tissues (tissue-wise eQTL $P_{GW} < 0.05$). Given that the expression of the hub gene was correlated to the genotype the expression of the co-eQTL genes can be either unchanged or also correlated with the genotype. In general the other co-eQTL genes do not show a global genetic control by the subnetwork marker. No eQTLs in more than one tissue at the very same marker could be detected for the co-eQTL genes (tissue wise eQTL $P_{GW} < 0.05$). Ten co-eQTL transcripts do show eQTLs in single tissues ($P_{GW} < 0.05$) at the subnetwork marker. Seven of these are linked to hubs with global eQTLs, such that the global pattern of genotype dependent co-expression is maintained. The remaining three do have tissue specific eQTLs in the same tissue as the hub gene and are linked to the same hub. Interestingly, these rare cases were examples where co-expression across strains and tissues was present in four tissues and the genetic marker induced co-expression across strains in a fifth tissue.

Each of the identified subnetworks corresponds to an eQTL gene (the hub) and its neighbours in a genotype dependent co-expression network. In order to determine the mode of regulation of the hub eQTLs we have determined the physical distance of the transcript and the genetic marker in the genome. If they are as close as 10 Mb, we assume a *cis*-regulatory mechanism [104], since the transcript and the mutation are in close proximity. If the transcript and the marker are on different chromosomes or further apart than 10Mb we assume *trans*-regulation. Of the 24 subnetwork hubs, 10 have *cis*-acting eQTLs and 11 have *trans*-acting eQTLs (3 of the probe sets could not be mapped uniquely to the genome).

4.4.5 Subnetworks are functionally coherent

Next we analyzed the 24 subnetworks for functional coherence. We used the standard 2×2 contingency table and functional categories from the Gene Ontology in order to find functionally enriched subnetworks [185] (see also section 2.2.1). We found 17 subnetworks enriched for at least one GO term ($P < 0.01$) and a total of 203 GO terms that were enriched at $P < 0.01$ see Table A.1. Seven of the subnetworks were enriched for signalling GO terms, 11 for terms related to ion transport, six for transcriptional regulation and seven for metabolic functions.

The majority (12) of hub genes does not have any functional annotation to gene ontology terms. For these genes, the differential co-expression with genes enriched for functional terms, can be used to place them into a functional context and infer new functional roles of these genes (see Table A.2). Of the hub genes annotated to GO terms, four were annotated to one of the enriched terms of their subnetwork, while six had differing annotations. Also for these genes, the differential co-expression can be used to infer potentially new, unknown functions or previously undescribed interactions between pathways.

Three of the markers that control the co-expression of subnetworks lie within two physiological QTL regions. A blood pressure QTL on chromosome 8 harbours the marker $SHRSPc63b01\_s1\_193$.

This marker is associated with a subnetwork that is enriched in "biogenic amine metabolism". Biogenic amines include norepinephrine, histamine and serotonine. Norepinephrine acts as a stress hormone and as such it increases heart rate, blood flow and blood pressure [187]. The hub transcript linked to $SHRSPc63b01\_s1\_193$ is annotated by Affymetrix based on EST sequences that are similar to the mouse gene histidine triad protein member 5, but does not provide any hints on the function of the gene. However, our findings may suggest a role in the regulation of blood pressure.

The other two markers are located in a QTL for heart rate on chromosome 10. The subnetwork around the gene *Ddx42* is enriched in steroid biosynthesis and lipid metabolism. In particular it contains genes for the synthesis of C-21 steroids also known as progesterones. These steroid hormones are known to affect the vascular tone [188]. However, there is no overlapping blood pressure QTL at this position. It is unclear, if this is due to statistical reasons or if there is truly no genotypic effect on blood pressure.

The subnetwork around *Krt10* is enriched for lipid metabolism and immune response pathways. Differentially co-expressed genes from the lipid metabolism are the apolipoproteins 1,2,4,5. These proteins are involved in lipid transport, such as cholesterol, which is the basis for the steroid synthesis described above. As far as the role of the immune response genes is concerned, the link between inflammation and hypertension is subject to current research [189].

4.4.6 Genes linking hubs are enriched for regulatory functions

Genes from interface sets represent links between two eQTLs which can be valuable resources for the functional interpretation of the two eQTLs and their relationship among each other. Thus we were interested if the interface genes share common functional roles at a global level. We have extracted all interface genes and subjected them to GO enrichment analysis.

Globally we have identified molecular functions involved in regulation as well as metabolic processes listed in Table A.3. Inspecting categories from the "biological process" ontology of GO, we found among others "blood pressure regulation" ($P = 2.8e-03$), "cholesterol absorption" ($P = 1.6e-03$) as well as "innate immune response" $P = 1.2e-03$. Since SHR is a model for hypertension, the blood pressure regulation genes are of special interest, because the perturbation of coordinated expression along with these genes could impact the physiological phenotypes of hypertension. The blood pressure regulating genes in the interfaces are *Adra2a*, an adrenergic receptor, *Agtr2*, an angiotensin II receptor, *Avpr2*, an arginine vasopressin receptor, *Guca2b*, *Fgb* and *Fgg*.

Following the assumption that genes which occur in more interfaces are more essential for the regulation we have computed GO enrichment for all nodes occurring in at least three or five interfaces. For genes in at least three interfaces we have found 14 GO terms with $P < 0.01$ of which ten are related to signalling, two are related to DNA binding (see Table A.4). Furthermore, we find lipid binding and chloride transport. When we investigated genes that occur in at least five interfaces we found an enrichment for cAMP mediated signalling ($P = 0.0036$).

Finally, we have investigated the functional enrichment on the level of each individual interface defined by a pair of hubs. Table A.5 summarises the results of all interfaces. We found 110 GO terms to be enriched at $P < 0.01$. Of the 28 GO terms that are found in more than one interface, 24 are related to signalling. GO terms that are enriched only in single interfaces include specific

cellular functions like lipid metabolism.

Observing an enrichment for signalling genes on the global level and especially in the highly connected interface nodes led us to hypothesise, that the genes from the interfaces represent a regulatory layer with coordinated gene expression profiles. Coordinated gene expression of eQTL transcripts is disconnected from the regulatory layer by genetic variants, that determine the expression levels of the eQTL transcripts. But can we assume that all of the interface nodes are upstream regulators of the eQTL? We consider a simple causal model, where the expression of the eQTL transcript E is dependent on at most two variables, the genotype at its marker G and the expression of a regulator R. Figure A.1 shows three possible scenarios where genotype dependent co-expression can arise. (1) the simplest one is, that G and R regulate E and one allele at G overrides the regulatory action of R. This leads to differential co-expression between E and R. (2) the two other models include a second downstream transcript T. If E is regulated by R and G, as above, but additionally R is regulating T, then we expect differential co-expression between E and R as well as between E and T. (3) if E is regulating T and we observe differential co-expression between E and T and the expression of E is dependent on G, there must exist a regulator R upstream of T that overrides the regulatory action of E. In this case there is differential co-expression between E and T, and between E and R. In case (1) and (2) we find one direct upstream regulator of E. Case (2) identifies one additional target transcript of the regulator. In case (3) we identify a regulator and a target that are not directly upstream of E but overriding the regulatory action of E. Interface nodes are connected to at least two hubs (eQTL transcripts) and therefore included in more than one instantiation of such a model. Due to this fact and the functional annotation of the interface genes as regulatory genes, it is very likely that these genes are upstream regulators of the hub genes.

4.4.7 Linked hubs reveal epistatic interactions

Interfaces not only provide insight into the functional context of the hub genes, they provide furthermore a molecular link between two genetic markers that are not expected to be linked genetically. Indeed, the average number of recombinations between two markers linked by an interface is 14.5 which is also the number expected by chance. Without any prior knowledge about potential genetic interactions, the genome-wide search for genetic interactions is underpowered. Using hub interfaces as molecular link between genetic markers, we set out to identify genetic interactions like epistasis of these markers on physiological phenotypes. In particular, we were interested in the markers linked by interfaces, containing one or more of the six genes annotated to "blood pressure regulation". These genes are present in 12 interface sets with an average size of 51 genes. The BXH/HXB cross has been phenotyped for blood pressure related traits [57], namely systolic blood pressure, diastolic blood pressure, carotid pulse pressure and mean arterial pressure.

For each pair of markers linked by such an interface and each of the blood pressure phenotypes we computed a two-way ANOVA test with interaction term. Table 4.5 lists the most significant genetic interactions. Strikingly we find significant interactions ($FDR < 0.2$) for all four traits at interfaces connecting markers on chromosome 3 and 17 and chromosome 1 and 17 respectively. For systolic blood pressure and mean arterial pressure genetic interactions were detected between loci on chromosome 1 and 3 as well as chromosome 10 and 17.

An interpretation of these findings is, that the co-regulation of two transcripts by one of the blood pressure regulating genes in the interface is perturbed. Each co-expression relationship is perturbed by mutations at a different marker. Only if the two perturbations occur at the same time and hence none of the two transcripts can be regulated by the blood pressure regulating gene, a consequence for the measured blood pressure parameters can be observed.

Table 4.5: Epistatic interactions on blood pressure phenotypes

trait	marker1	marker2	size	F	p	q
Systolic_BP	gko-90c15_rp2_b1_262	J697333	53	12.94	1.38E−03	0.06
Systolic_BP	rat109_029_j07.q1ca_518	J697333	35	10.90	3.01E−03	0.06
Mean_Arterial_Pressure	gko-90c15_rp2_b1_262	J697333	53	10.02	4.05E−03	0.06
Mean_Arterial_Pressure	rat109_029_j07.q1ca_518	J697333	35	7.70	1.05E−02	0.11
Carotid_Pulse_Pressure	rat109_029_j07.q1ca_518	J491517	75	7.49	1.15E−02	0.11
Carotid_Pulse_Pressure	gko-90c15_rp2_b1_262	J697333	53	7.05	1.36E−02	0.11
Mean_Arterial_Pressure	gnl—ti—896779106_19866866792872_234	rat110_010_c16.q1ca_212	45	5.83	2.38E−02	0.15
Systolic_BP	gnl—ti—896779106_19866866792872_234	rat110_010_c16.q1ca_212	45	5.48	2.79E−02	0.15
Diastolic_BP	rat109_029_j07.q1ca_518	J697333	35	5.31	3.01E−02	0.15
Systolic_BP	rat109_029_j07.q1ca_518	J491517	75	5.15	3.25E−02	0.15
Systolic_BP	rat109_029_j07.q1ca_518	Cpn_3002671921	27	4.94	3.60E−02	0.15
Carotid_Pulse_Pressure	gko-90c15_rp2_b1_262	Cpn_3002671921	26	4.80	3.80E−02	0.15
Diastolic_BP	gko-90c15_rp2_b1_262	J697333	53	4.24	5.01E−02	0.18
Carotid_Pulse_Pressure	rat109_029_j07.q1ca_518	J697333	35	3.97	5.77E−02	0.20
Mean_Arterial_Pressure	rat109_029_j07.q1ca_518	Cpn_3002671921	27	3.86	6.12E−02	0.20

4.4.8 Conclusions

In order to investigate consequences of genetic variations that abrogate connections in a co-expression network we have developed a linear model for the mapping of co-expression as a quantitative trait. Using this model we were able to construct a graph representing significant genotype induced perturbations of co-expression. The topological analysis of the graph revealed the presence of subnetworks, that are arranged in stars around central hub nodes. Furthermore hub interfaces can be defined as sets of genes connecting two or more hubs with each other. The hubs were eQTL transcripts where the presence of the eQTL perturbed the interaction of the hubs with the genes it interacts with under normal conditions. Functional analysis of the subnetworks revealed that the subnetworks are enriched in common GO categories. This enrichment allowed us to place the eQTL hub genes into a functional context. Moreover, in cases where annotation is missing, the functional enrichment of the subnetwork allows an educated guess about the function of the hub gene. Hub interfaces were enriched for genes of regulatory function. In particular we found six genes from the GO category "blood pressure regulation" in the interfaces. Since interfaces not only connect hub genes, but also the genetic markers that are associated with the edges, interfaces can be used as evidence for genetic interactions, such as epistasis. Normally, a genome-wide search for epistatic interactions is underpowered due to combinatorial issues. Using the molecular link provided by hub interfaces we were able to test specific genetic interactions at markers that are connected via blood pressure regulating genes. This analysis provided evidence for epistatic interactions on three major blood pressure parameters.

These results provide interesting leads for the search of genetic interactions in human GWAS. In these studies, usually 0.5 - 1 million SNPs are genotyped in large case-control cohorts. The number of possible pairwise combinations of markers makes the exhaustive search for genetic interactions impossible. However, translating the evidence for genetic interactions from animal studies, like the one presented here, makes it feasible to test only interactions of a small number of selected loci. These loci can be identified via sequence homology of the hub genes.

Genetic interactions can also be verified experimentally. There are two kinds of experiments that could be performed. The purely genetic approach would be to generate a double congenic line, which carries the disease alleles on a normotensive background, for each of the interactions. Subsequently, phenotyping of the blood pressure parameters could confirm the genetic interaction. The second approach would be a double knockout of the hub genes at these loci. Both are not in the scope of this work.

5 Discussion

5.1 Systematic profiling of regulatory variations speeds up identification of disease genes

Classical forward genetic studies that associate phenotypes to molecular markers have severe limitation for the identification of disease genes. Usually the associated regions are large chromosomal regions often containing hundreds of genes. Refinement of these regions e.g. by congenic lines is very time consuming and relies on the occurrence of recombinations around the associated region. The ultimate goal to identify causal genes or variations is rarely achieved.

These causal variations come mainly in two flavours: (1) coding variations that directly affect the protein structure and therefore function and (2) regulatory variations that affect gene expression. Since most of the genome is not occupied by protein coding genes it is instrumental to assume that the causal variation is regulatory. Especially because genetic analysis of genome wide gene expression can be used to assess the immediate consequence of regulatory variations, namely that gene expression of a certain transcript is dependent on the genotype of a genetic variant. Still, the genotyped variations are most likely not the functional variations but the targets of these regulatory variations can be identified as the eQTL transcripts.

Moreover, if gene expression and physiological traits are analysed together candidate disease genes that are targets of a regulatory variation can be identified. Section 3.4.1 and section 3.4.2 showed how this strategy was successfully applied in two rat populations. If combined with a functional study of a knockout model, the identified candidate gene can be confirmed by studying the effect of the knockout on the phenotype. In the case of *Ephx2* this strategy was successful and could even be translated to human patients.

However, there are certain limitations to this approach. Most importantly, the critical hypothesis of a regulatory causal variation might not hold. In the worst case, this hypothesis might even be misleading if one finds eQTLs in the disease associated regions that are not related to the disease. Therefore, not only the analysis of correlation between physiological phenotype and expression levels but also a careful assessment of possible causal scenarios is important [128, 129]. Other strategies when no obvious target of a regulatory variation can be identified are time consuming for instance the creation of congenic lines or sequencing of all candidate genes or they require large scale efforts as the resequencing of the associated region (using capturing) or of the complete parental genomes. However, with the current development of sequencing methods [190] it can be expected that most of the important inbred model strains will be sequenced [191].

5.2 Sequence analysis of cis-eQTL promoters identifies regulatory mechanism

A second limitation of the eQTL approach to candidate gene identification is that only the targets of regulatory variations and not the variations themselves can be identified. If a candidate gene could be identified as a potential target of a regulatory variation it is straight forward to sequence the putative proximal promoter region of that gene. As demonstrated in section 3.4.1 this approach can suffice to identify functional regulatory variations. However, mammalian gene regulation can be driven by enhancer elements which are located up to several Mb from the gene [192]. This makes it hard to apply this as a general strategy. Again, if a high resolution map of genetic variations is available such as the output of resequencing projects [193, 194, 161] it is a hard but feasible task to screen all potential regulatory variations around a gene.

Given that a regulatory variant has been identified, it is still unknown what regulatory mechanism is involved. Regulation by TFs is the best understood mechanism of transcriptional regulation. TFs bind to TFBS in the proximal promoters or enhancers of their target genes and activate or repress their transcription. Therefore genetic variations of TFBS sequences are a potential mechanism that could explain the associated expression of target genes. In order to characterise the regulatory potential of sequence variations we have developed the sTRAP tool (section 4.2). It allows to predict which transcription factor is most likely affected by a sequence variations. We have shown that known TF-SNP associations can be recovered and that these predictions are specific. In the case of *Ephx2* we could establish that the regulatory variation is most likely triggered through the *de novo* creation of an *AP1* binding site and thus shed light on the upstream regulator of this eQTL transcript.

An important limitation of our approach is that knowledge about the binding preferences of many TFs is still uncharacterised. Therefore analysis are restricted to a limited set of TFs [142, 195]. Large scale protein binding assays [150] are currently underway and likely to provide a more complete set of binding models in the future. As mentioned before, the identification of a functional regulatory variant still is a very demanding challenge [138] and has only been reported scarcely. Although a complete catalog of genetic variations is an important step for the identification of regulatory variations, more functional data like histone and DNA modification patterns as generated by the Encode project [196] are needed to discriminate functional and nonfunctional polymorphic regulatory sites. In addition cell type specific mechanisms [23] might play an important role.

5.3 Knowledge based network analysis for the interpretation of eQTL data

Most biological processes require the coordinated action of functionally related genes e.g. metabolic pathways or signalling networks. Therefore the genes of a pathway have to be expressed at the same time. This leads to the simplifying assumption that genes of a pathway are also coregulated. In eQTL studies large sets of genes that are all regulated by a common genetic variation can be observed as *trans*-clusters (section 3.3). Assuming that the coregulation is a consequence of related function *trans*-clusters might represent genetically regulated pathways.

So there are two related problems: (1) how can *trans*-clusters be interpreted in terms of function and (2) in terms of regulatory mechanisms.

Functional interpretation can be facilitated by the use of enrichment analysis (section 2.2). For large *trans*-clusters with many annotated genes this approach is very fruitful. However, in some experiments only smaller *trans*-clusters containing a small number of annotated genes could be detected. In those cases the main limitation of the classical approach is the arbitrary choice of a significance threshold to define the *trans*-clusters. In section 4.1 we have proposed a simple extension of gene set enrichment analysis for genetic mapping to overcome this problem.

The main limitation of functional enrichment analyses is the incompleteness of the functional annotation for instance in GO. If unknown pathways show up in *trans*-clusters, the analysis in terms of known pathways cannot aid the interpretation. To alleviate this limitation and provide functional annotation of previously uncharacterised genes remains one of the greatest challenges in genome research today.

Coregulated pathways also imply a common regulatory mechanism. Since regulation by TFs is the best understood mechanism of transcriptional regulation we have set out to identify *trans*-clusters that are coregulated by a common TF. In section 4.3 we have introduced a two step procedure to identify TFs that are themselves targets of regulatory variations and transmit this effect on to their direct target genes. This allowed to define a TF regulated network for *Irf7*. Additional functional analysis related this network to immune response which was also supported by an analysis of cell type specificity which localised the network to macrophages and immune related cells. Moreover, the regulatory variation could be narrowed down to a SNP in the $5'$ UTR of *Ebi2* which was shown to regulate *Irf7* expression in siRNA knockdown experiments in cell culture.

Analyses assuming a common transcription factor suffers from several limitations. First only a small number of TF-target relations are well studied. Therefore target predictions are used. However, these predictions are not very specific leading to a large number of false positive predictions. Here we tried to overcome this problem by the statistical approach of enrichment analysis. The second problem discussed in the previous section is that only a restricted number of TFs are characterised with respect to their binding preferences. Thirdly coregulation might involve several TFs where each might only regulate a small number of direct targets, making it difficult to be detectable by an enrichment approach. Finally the assumption that TFs themselves have to be genetically regulated might be too stringent. It might well be that other mechanisms than transcriptional regulation are responsible for a change of TF activity such as activation by phosphorylation or other cofactors.

5.4 Gene expression networks for the analysis of polygenic traits

The first results of GWAS [6] were reducing the initial enthusiasm about GWAS mainly because the effect sizes of associated loci were rather small [24]. For cardiovascular traits the first GWAS yielded no associations [6] and only meta analysis with very large sample sizes (34,433 and 29,136 individuals respectively) led to significant associations [197, 198] for hypertension and elevated blood pressure. These results are interpreted as a confirmation of the polygenic inheritance of common diseases [24]. Although it could also be evidence for the hypothesis that common

diseases are heterogeneous and different loci contribute in different families [199].

Using functional gene expression networks for the interpretation of GWAS results constitutes one way to characterise polygenic inheritance. Analogous to the threshold free analysis of microarray data introduced by GSEA (section 2.2.2) one can analyse the many loci of small effects in GWAS with respect to functional categories or previously characterised gene expression networks. If such a network is enriched for disease associated loci one would interpret this as evidence for a functional role of the network in the disease process. In section 4.3.8 we have applied this approach by using the *Irf7* regulated immune network from the rat in order to interpret human T1D GWAS data. We could show that our gene expression network captured the disease associated genes more specifically than just using functional annotation from GO. In addition we were also able to show disease association of the human regulatory locus that was translated from the rat regulatory locus we had identified using the TF centred analysis of *trans*-eQTLs in the rat. These studies show how the analysis of genetically driven gene expression networks in rats can be used to generate hypothesis about disease processes in humans.

Gene networks can also be used to guide analyses of more complex modes of inheritance such as epistasis. Using a network derived from genotype dependent perturbations of coexpression of genes we have inferred connections between genetic markers which were otherwise completely unconnected; i.e. there was no genetic linkage between these markers. The association of these specific marker pairs through the network was used to test for epistatic interactions affecting quantitative phenotypes. In particular we were interested in blood pressure phenotypes and identified significant epistatic interactions for these traits.

The problem with these network based approaches is the validation of hypothesis. The backward genetic strategy which works extremely well for single candidate genes is very hard to implement when multiple genes need to be perturbed. This hinders the validation of predicted epistatic interactions as well as the characterisation of essential parts of the gene expression networks. In the case of epistatic interactions this problem might be solvable because only two genes need to be perturbed simultaneously. General gene expression networks can be a useful tool for the generation of hypothesis. To be testable however these hypothesis need to state certain molecular mechanisms for which appropriate assay systems can be implemented.

5.5 Conclusions

Overall this thesis describes an extensive set of tools and strategies for the analysis of regulatory genetic variations. The starting point was the identification of target genes of potential regulatory variations as eQTL transcripts which has been described previously. We provide ways to address the following resulting questions about these genes. (1) What is the role of the eQTL transcript in the context of a disease model, (2) which is the *cis*-regulatory element affected by the genetic variant and which transcription factor is the upstream regulator of the eQTL transcript, (3) what are the *trans*-regulatory factors and how are their effects mediated to their target genes, and (4) what is the functional context that eQTL transcripts operate in? Moreover, we used gene expression networks derived from the analysis of the genetics of gene expression in rats to connect human disease association data to molecular function in an attempt to interpret the genetics of polygenic traits.

Bibliography

[1] STAR Consortium. Snp and haplotype mapping for genetic analysis in the rat. *Nat Genet*, 40(5):560–6, 2008.

[2] Jan Monti, Judith Fischer, Svetlana Paskas, Matthias Heinig, Herbert Schulz, Claudia Gosele, Robert Fischer, Cosima Schmidt, Alexander Schirdewan, Volkmar Gross, Arnd Heuser, Oliver Hummel, Henrike Maatz, Giannino Patone, Kathrin Saar, Martin Vingron, Steven M Weldon, Klaus Lindpaintner, Bruce D Hammock, Klaus Rohde, Rainer Dietz, Stuart A Cook, Wolf-Hagen Schunck, Friedrich C Luft, and Norbert Hubner. Soluble epoxide hydrolase is a susceptibility gene for heart failure in a rat model of human disease. *Nat Genet*, 40(5):529–37, 2008.

[3] Matthias Heinig, Enrico Petretto, Chris Wallace Leonardo Bottolo, Maxime Rotival, Han Lu, Yoyo Li, Rizwan Sarwar, Sarah R. Langley, Anja Bauerfeind, Oliver Hummel, Young-Ae Lee, Svetlana Paskas, Carola Rintisch, Kathrin Saar, Jason Cooper, Rachel Buchan, Elizabeth E. Gray, Jason G. Cyster, Cardiogenics Consortium, Jeanette Erdmann, Christian Hengstenberg, Seraya Maouche, Willem H. Ouwehand, Catherine M. Rice, Nilesh J Samani, Heribert Schunkert, Alison H Goodall, Herbert Schulz, Helge Roider, Martin Vingron, Stefan Blankenberg, Thomas Münzel, Tanja Zeller, Silke Symczak, Andreas Ziegler, Laurence Tiret, Deborah J. Smyth, Michal Pravenec, Timothy J. Aitman, Francois Cambien, David Clayton, John A. Todd, Norbert Hubner, and Stuart A. Cook. A conserved trans-acting regulatory locus underlies a proinflammatory gene expression network and susceptibility to autoimmune type 1 diabetes. Under review, 2010.

[4] Thomas Manke, Matthias Heinig, and Martin Vingron. Quantifying the effect of sequence variation on regulatory interactions. *Under review at Human molecular genetics*, 2009.

[5] V.A. McKusick. *Mendelian inheritance in man: a catalog of human genes and genetic disorders*. Johns Hopkins University Press, 1998.

[6] Wellcome Trust Case Control Consortium, Australo-Anglo-American Spondylitis Consortium (TASC), Paul R Burton, David G Clayton, Lon R Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P Kwiatkowski, Mark I McCarthy, Willem H Ouwehand, Nilesh J Samani, John A Todd, Peter Donnelly, Jeffrey C Barrett, Dan Davison, Doug Easton, David M Evans, Hin-Tak Leung, Jonathan L Marchini, Andrew P Morris, Chris C A Spencer, Martin D Tobin, Antony P Attwood, James P Boorman, Barbara Cant, Ursula Everson, Judith M Hussey, Jennifer D Jolley, Alexandra S Knight, Kerstin Koch, Elizabeth Meech, Sarah Nutland, Christopher V Prowse, Helen E Stevens, Niall C Taylor, Graham R Walters, Neil M Walker, Nicholas A Watkins, Thilo Winzer, Richard W Jones, Wendy L McArdle, Susan M Ring, David P Strachan, Marcus Pembrey,

Gerome Breen, David St Clair, Sian Caesar, Katharine Gordon-Smith, Lisa Jones, Christine Fraser, Elaine K Green, Detelina Grozeva, Marian L Hamshere, Peter A Holmans, Ian R Jones, George Kirov, Valentina Moskivina, Ivan Nikolov, Michael C O'Donovan, Michael J Owen, David A Collier, Amanda Elkin, Anne Farmer, Richard Williamson, Peter McGuffin, Allan H Young, I. Nicol Ferrier, Stephen G Ball, Anthony J Balmforth, Jennifer H Barrett, Timothy D Bishop, Mark M Iles, Azhar Maqbool, Nadira Yuldasheva, Alistair S Hall, Peter S Braund, Richard J Dixon, Massimo Mangino, Suzanne Stevens, John R Thompson, Francesca Bredin, Mark Tremelling, Miles Parkes, Hazel Drummond, Charles W Lees, Elaine R Nimmo, Jack Satsangi, Sheila A Fisher, Alastair Forbes, Cathryn M Lewis, Clive M Onnie, Natalie J Prescott, Jeremy Sanderson, Christopher G Matthew, Jamie Barbour, M. Khalid Mohiuddin, Catherine E Todhunter, John C Mansfield, Tariq Ahmad, Fraser R Cummings, Derek P Jewell, John Webster, Morris J Brown, Mark G Lathrop, John Connell, Anna Dominiczak, Carolina A Braga Marcano, Beverley Burke, Richard Dobson, Johannie Gungadoo, Kate L Lee, Patricia B Munroe, Stephen J Newhouse, Abiodun Onipinla, Chris Wallace, Mingzhan Xue, Mark Caulfield, Martin Farrall, Anne Barton, Biologics in RA Genetics, Genomics Study Syndicate (BRAGGS) Steering Committee, Ian N Bruce, Hannah Donovan, Steve Eyre, Paul D Gilbert, Samantha L Hilder, Anne M Hinks, Sally L John, Catherine Potter, Alan J Silman, Deborah P M Symmons, Wendy Thomson, Jane Worthington, David B Dunger, Barry Widmer, Timothy M Frayling, Rachel M Freathy, Hana Lango, John R B Perry, Beverley M Shields, Michael N Weedon, Andrew T Hattersley, Graham A Hitman, Mark Walker, Kate S Elliott, Christopher J Groves, Cecilia M Lindgren, Nigel W Rayner, Nicolas J Timpson, Eleftheria Zeggini, Melanie Newport, Giorgio Sirugo, Emily Lyons, Fredrik Vannberg, Adrian V S Hill, Linda A Bradbury, Claire Farrar, Jennifer J Pointon, Paul Wordsworth, Matthew A Brown, Jayne A Franklyn, Joanne M Heward, Matthew J Simmonds, Stephen C L Gough, Sheila Seal, Breast Cancer Susceptibility Collaboration (UK), Michael R Stratton, Nazneen Rahman, Maria Ban, An Goris, Stephen J Sawcer, Alastair Compston, David Conway, Muminatou Jallow, Melanie Newport, Giorgio Sirugo, Kirk A Rockett, Suzannah J Bumpstead, Amy Chaney, Kate Downes, Mohammed J R Ghori, Rhian Gwilliam, Sarah E Hunt, Michael Inouye, Andrew Keniry, Emma King, Ralph McGinnis, Simon Potter, Rathi Ravindrarajah, Pamela Whittaker, Claire Widden, David Withers, Niall J Cardin, Dan Davison, Teresa Ferreira, Joanne Pereira-Gale, Ingeleif B Hallgrimsdo'ttir, Bryan N Howie, Zhan Su, Yik Ying Teo, Damjan Vukcevic, David Bentley, Matthew A Brown, Alastair Compston, Martin Farrall, Alistair S Hall, Andrew T Hattersley, Adrian V S Hill, Miles Parkes, Marcus Pembrey, Michael R Stratton, Sarah L Mitchell, Paul R Newby, Oliver J Brand, Jackie Carr-Smith, Simon H S Pearce, R. McGinnis, A. Keniry, P. Deloukas, John D Reveille, Xiaodong Zhou, Anne-Marie Sims, Alison Dowling, Jacqueline Taylor, Tracy Doan, John C Davis, Laurie Savage, Michael M Ward, Thomas L Learch, Michael H Weisman, and Mathew Brown. Association scan of 14,500 nonsynonymous snps in four diseases identifies autoimmunity variants. *Nat Genet*, 39(11):1329–1337, Nov 2007.

[7] David Altshuler, Mark J Daly, and Eric S Lander. Genetic mapping in human disease. *Science*, 322(5903):881–888, Nov 2008.

[8] Georg Mendel. Versuche über pflanzen-hybriden. *Verhandlungen des Naturforschenden Vereines in Brünn*, 4:3–47, 1866.

[9] T. Strachan and Andrew P. Read. *Human molecular genetics*. Garland Science, 3 edition, 2004.

[10] JBS Haldane. The combination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet*, 8(29):309, 1919.

[11] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, Oct 2005.

[12] M. Lynch and B. Walsh. *Genetics and analysis of quantitative traits*. Sinauer Associates Sunderland, MA, 1998.

[13] R.A. Fisher et al. The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

[14] International HapMap Consortium, Kelly A Frazer, Dennis G Ballinger, David R Cox, David A Hinds, Laura L Stuve, Richard A Gibbs, John W Belmont, Andrew Boudreau, Paul Hardenbol, Suzanne M Leal, Shiran Pasternak, David A Wheeler, Thomas D Willis, Fuli Yu, Huanming Yang, Changqing Zeng, Yang Gao, Haoran Hu, Weitao Hu, Chaohua Li, Wei Lin, Siqi Liu, Hao Pan, Xiaoli Tang, Jian Wang, Wei Wang, Jun Yu, Bo Zhang, Qingrun Zhang, Hongbin Zhao, Hui Zhao, Jun Zhou, Stacey B Gabriel, Rachel Barry, Brendan Blumenstiel, Amy Camargo, Matthew Defelice, Maura Faggart, Mary Goyette, Supriya Gupta, Jamie Moore, Huy Nguyen, Robert C Onofrio, Melissa Parkin, Jessica Roy, Erich Stahl, Ellen Winchester, Liuda Ziaugra, David Altshuler, Yan Shen, Zhijian Yao, Wei Huang, Xun Chu, Yungang He, Li Jin, Yangfan Liu, Yayun Shen, Weiwei Sun, Haifeng Wang, Yi Wang, Ying Wang, Xiaoyan Xiong, Liang Xu, Mary M Y Waye, Stephen K W Tsui, Hong Xue, J. Tze-Fei Wong, Luana M Galver, Jian-Bing Fan, Kevin Gunderson, Sarah S Murray, Arnold R Oliphant, Mark S Chee, Alexandre Montpetit, Fanny Chagnon, Vincent Ferretti, Martin Leboeuf, Jean-François Olivier, Michael S Phillips, Stéphanie Roumy, Clémentine Sallée, Andrei Verner, Thomas J Hudson, Pui-Yan Kwok, Dongmei Cai, Daniel C Koboldt, Raymond D Miller, Ludmila Pawlikowska, Patricia Taillon-Miller, Ming Xiao, Lap-Chee Tsui, William Mak, You Qiang Song, Paul K H Tam, Yusuke Nakamura, Takahisa Kawaguchi, Takuya Kitamoto, Takashi Morizono, Atsushi Nagashima, Yozo Ohnishi, Akihiro Sekine, Toshihiro Tanaka, Tatsuhiko Tsunoda, Panos Deloukas, Christine P Bird, Marcos Delgado, Emmanouil T Dermitzakis, Rhian Gwilliam, Sarah Hunt, Jonathan Morrison, Don Powell, Barbara E Stranger, Pamela Whittaker, David R Bentley, Mark J Daly, Paul I W de Bakker, Jeff Barrett, Yves R Chretien, Julian Maller, Steve McCarroll, Nick Patterson, Itsik Pe'er, Alkes Price, Shaun Purcell, Daniel J Richter, Pardis Sabeti, Richa Saxena, Stephen F Schaffner, Pak C Sham, Patrick Varilly, David Altshuler, Lincoln D Stein, Lalitha Krishnan, Albert Vernon Smith, Marcela K Tello-Ruiz, Gudmundur A Thorisson, Aravinda Chakravarti, Peter E Chen, David J Cutler, Carl S Kashuk, Shin Lin, Gonçalo R Abecasis, Weihua Guan, Yun Li, Heather M Munro, Zhaohui Steve Qin, Daryl J Thomas, Gilean McVean, Adam Auton, Leonardo Bottolo,

Niall Cardin, Susana Eyheramendy, Colin Freeman, Jonathan Marchini, Simon Myers, Chris Spencer, Matthew Stephens, Peter Donnelly, Lon R Cardon, Geraldine Clarke, David M Evans, Andrew P Morris, Bruce S Weir, Tatsuhiko Tsunoda, James C Mullikin, Stephen T Sherry, Michael Feolo, Andrew Skol, Houcan Zhang, Changqing Zeng, Hui Zhao, Ichiro Matsuda, Yoshimitsu Fukushima, Darryl R Macer, Eiko Suda, Charles N Rotimi, Clement A Adebamowo, Ike Ajayi, Toyin Aniagwu, Patricia A Marshall, Chibuzor Nkwodimmah, Charmaine D M Royal, Mark F Leppert, Missy Dixon, Andy Peiffer, Renzong Qiu, Alastair Kent, Kazuto Kato, Norio Niikawa, Isaac F Adewole, Bartha M Knoppers, Morris W Foster, Ellen Wright Clayton, Jessica Watkin, Richard A Gibbs, John W Belmont, Donna Muzny, Lynne Nazareth, Erica Sodergren, George M Weinstock, David A Wheeler, Imtaz Yakub, Stacey B Gabriel, Robert C Onofrio, Daniel J Richter, Liuda Ziaugra, Bruce W Birren, Mark J Daly, David Altshuler, Richard K Wilson, Lucinda L Fulton, Jane Rogers, John Burton, Nigel P Carter, Christopher M Clee, Mark Griffiths, Matthew C Jones, Kirsten McLay, Robert W Plumb, Mark T Ross, Sarah K Sims, David L Willey, Zhu Chen, Hua Han, Le Kang, Martin Godbout, John C Wallenburg, Paul L'Archevêque, Guy Bellemare, Koji Saeki, Hongguang Wang, Daochang An, Hongbo Fu, Qing Li, Zhen Wang, Renwu Wang, Arthur L Holden, Lisa D Brooks, Jean E McEwen, Mark S Guyer, Vivian Ota Wang, Jane L Peterson, Michael Shi, Jack Spiegel, Lawrence M Sung, Lynn F Zacharia, Francis S Collins, Karen Kennedy, Ruth Jamieson, and John Stewart. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–861, Oct 2007.

[15] Leonid Kruglyak. The road to genome-wide association studies. *Nat Rev Genet*, 9(4):314–318, Apr 2008.

[16] Francis S Collins. In *Genomics of common disease*, 2007.

[17] Timothy J Aitman, John K Critser, Edwin Cuppen, Anna Dominiczak, Xose M Fernandez-Suarez, Jonathan Flint, Dominique Gauguier, Aron M Geurts, Michael Gould, Peter C Harris, Rikard Holmdahl, Norbert Hubner, Zsuzsanna Izsvák, Howard J Jacob, Takashi Kuramoto, Anne E Kwitek, Anna Marrone, Tomoji Mashimo, Carol Moreno, John Mullins, Linda Mullins, Tomas Olsson, Michal Pravenec, Lela Riley, Kathrin Saar, Tadao Serikawa, James D Shull, Claude Szpirer, Simon N Twigger, Birger Voigt, and Kim Worley. Progress and prospects in rat genetics: a community view. *Nat Genet*, 40(5):516–522, May 2008.

[18] V. M. Ingram. A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. *Nature*, 178(4537):792–794, Oct 1956.

[19] Roger P Alexander, Gang Fang, Joel Rozowsky, Michael Snyder, and Mark B Gerstein. Annotating non-coding regions of the genome. *Nat Rev Genet*, 11(8):559–571, Aug 2010.

[20] Marco De Gobbi, Vip Viprakasit, Jim R Hughes, Chris Fisher, Veronica J Buckle, Helena Ayyub, Richard J Gibbons, Douglas Vernimmen, Yuko Yoshinaga, Pieter de Jong, Jan-Fang Cheng, Edward M Rubin, William G Wood, Don Bowden, and Douglas R Higgs. A

regulatory snp causes a human genetic disease by creating a new transcriptional promoter. *Science*, 312(5777):1215–1217, May 2006.

[21] Eric E Schadt. Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218–223, Sep 2009.

[22] Yanqing Chen, Jun Zhu, Pek Yee Lum, Xia Yang, Shirly Pinto, Douglas J MacNeil, Chunsheng Zhang, John Lamb, Stephen Edwards, Solveig K Sieberts, Amy Leonardson, Lawrence W Castellini, Susanna Wang, Marie-France Champy, Bin Zhang, Valur Emilsson, Sudheer Doss, Anatole Ghazalpour, Steve Horvath, Thomas A Drake, Aldons J Lusis, and Eric E Schadt. Variations in DNA elucidate molecular networks that cause disease. *Nature*, 452(7186):429–435, Mar 2008.

[23] Antigone S Dimas, Samuel Deutsch, Barbara E Stranger, Stephen B Montgomery, Christelle Borel, Homa Attar-Cohen, Catherine Ingle, Claude Beazley, Maria Gutierrez Arcelus, Magdalena Sekowska, Marilyne Gagnebin, James Nisbett, Panos Deloukas, Emmanouil T Dermitzakis, and Stylianos E Antonarakis. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, 325(5945):1246–1250, Sep 2009.

[24] William Cookson, Liming Liang, Goncalo Abecasis, Miriam Moffatt, and Mark Lathrop. Mapping complex disease traits with global gene expression. *Nat Rev Genet*, 10(3):184–194, Mar 2009.

[25] J. Zhu, P. Y. Lum, J. Lamb, D. GuhaThakurta, S. W. Edwards, R. Thieringer, J. P. Berger, M. S. Wu, J. Thompson, A. B. Sachs, and E. E. Schadt. An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet Genome Res*, 105(2-4):363–374, 2004.

[26] Anatole Ghazalpour, Sudheer Doss, Bin Zhang, Susanna Wang, Christopher Plaisier, Ruth Castellanos, Alec Brozell, Eric E Schadt, Thomas A Drake, Aldons J Lusis, and Steve Horvath. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet*, 2(8):e130, Aug 2006.

[27] Pek Yee Lum, Yanqing Chen, Jun Zhu, John Lamb, Shlomo Melmed, Susanna Wang, Tom A Drake, Aldons J Lusis, and Eric E Schadt. Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. *J Neurochem*, 97 Suppl 1:50–62, Apr 2006.

[28] Valur Emilsson, Gudmar Thorleifsson, Bin Zhang, Amy S Leonardson, Florian Zink, Jun Zhu, Sonia Carlson, Agnar Helgason, G. Bragi Walters, Steinunn Gunnarsdottir, Magali Mouy, Valgerdur Steinthorsdottir, Gudrun H Eiriksdottir, Gyda Bjornsdottir, Inga Reynisdottir, Daniel Gudbjartsson, Anna Helgadottir, Aslaug Jonasdottir, Adalbjorg Jonasdottir, Unnur Styrkarsdottir, Solveig Gretarsdottir, Kristinn P Magnusson, Hreinn Stefansson, Ragnheidur Fossdal, Kristleifur Kristjansson, Hjortur G Gislason, Tryggvi Stefansson, Bjorn G Leifsson, Unnur Thorsteinsdottir, John R Lamb, Jeffrey R Gulcher, Marc L Reitman, Augustine Kong, Eric E Schadt, and Kari Stefansson. Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–428, Mar 2008.

[29] Henrike Maatz. *Identification and functional characterization of genetic factors underlying a blood pressure QTL on rat chromosome 1 through integration of genome wide gene expression profiling and correlation analysis.* PhD thesis, Humbold Universität zu Berlin, 2010.

[30] M. Soller, T. Brody, and A. Genizi. On the power of experimental design for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics*, 47:35–39, 1976.

[31] M. Soller and A. Genizi. The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations. *Biometrics*, 34:47–55, 1978.

[32] J. I. Weller. Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics*, 42(3):627–640, Sep 1986.

[33] E. S. Lander and D. Botstein. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121(1):185–199, Jan 1989.

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

[35] R. C. Jansen and J. P. Nap. Genetical genomics: the added value from segregation. *Trends Genet*, 17(7):388–391, Jul 2001.

[36] G. A. Churchill and R. W. Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963–971, Nov 1994.

[37] R. C. Jansen. Interval mapping of multiple quantitative trait loci. *Genetics*, 135(1):205–211, Sep 1993.

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

[39] C. H. Kao, Z. B. Zeng, and R. D. Teasdale. Multiple interval mapping for quantitative trait loci. *Genetics*, 152(3):1203–1216, Jul 1999.

[40] H. Akaike. A new look at the statistical model identification. 19(6):716–723, 1974.

[41] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[42] C.L. Mallows. Some comments on cp. *Technometrics*, 15:661–675, 1973.

[43] M.A. Efroymson. *Mathematical Methods for Digital Computers*, chapter Multiple regression analysis. Wiley, 1960.

[44] A. E. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[45] Robert Tibshirani. Regression shrinkage via the lasso. *J. Royal. Statist. Soc. B.*, 58:267–288, 1996.

[46] Nathalie Malo, Ondrej Libiger, and Nicholas J Schork. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *Am J Hum Genet*, 82(2):375–385, Feb 2008.

[47] Nengjun Yi and Shizhong Xu. Bayesian lasso for quantitative trait loci mapping. *Genetics*, 179(2):1045–1055, Jun 2008.

[48] Gustavo de los Campos, Hugo Naya, Daniel Gianola, José Crossa, Andrés Legarra, Eduardo Manfredi, Kent Weigel, and José Miguel Cotes. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182(1):375–385, May 2009.

[49] Wei Sun, Joseph G Ibriham, and Fei Zou. Genome-wide multiple loci mapping in experimental crosses by the iterative adaptive penalized regression. *Genetics*, Feb 2010.

[50] F. S. Collins, M. S. Guyer, and A. Charkravarti. Variations on a theme: cataloging human dna sequence variation. *Science*, 278(5343):1580–1581, Nov 1997.

[51] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, Sep 1996.

[52] E. S. Lander. The new genomics: global views of biology. *Science*, 274(5287):536–539, Oct 1996.

[53] International HapMap Consortium. The international hapmap project. *Nature*, 426(6968):789–796, Dec 2003.

[54] David Clayton. *Handbook of statistical genetics*, chapter Population association, pages 519 – 540. Wiley, 2001.

[55] Barbara E Stranger, Matthew S Forrest, Andrew G Clark, Mark J Minichiello, Samuel Deutsch, Robert Lyle, Sarah Hunt, Brenda Kahl, Stylianos E Antonarakis, Simon Tavar?, Panagiotis Deloukas, and Emmanouil T Dermitzakis. Genome-wide associations of gene expression variation in humans. *PLoS Genet*, 1(6):e78, Dec 2005.

[56] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–9445, Aug 2003.

[57] M. Pravenec, P. Klir, V. Kren, J. Zicha, and J. Kunes. An analysis of spontaneous hypertension in spontaneously hypertensive rats by means of new recombinant inbred strains. *J Hypertens*, 7(3):217–221, Mar 1989.

[58] Michal Pravenec, V?clav Z?dek, Alena Musilov?, Miroslava Sim?kov?, Vlastimil Kostka, Petr Mlejnek, Vladim?r Kren, Drahom?ra Krenova, Vlasta B?l?, Blanka M?kov?, Marie J?chymov?, Karel Hork?, Ludmila Kazdov?, Elizabeth St Lezin, and Theodore W Kurtz. Genetic analysis of metabolic defects in the spontaneously hypertensive rat. *Mamm Genome*, 13(5):253–258, May 2002.

[59] K. Okamoto. *Spontaneous hypertension: Its pathogenesis and complications*. Springer, 1972.

[60] R. H. Rao. Insulin resistance in spontaneously hypertensive rats. difference in interpretation based on insulin infusion rate or on plasma insulin in glucose clamp studies. *Diabetes*, 42(9):1364–1371, Sep 1993.

[61] S. Hulman, B. Falkner, and N. Freyvogel. Insulin resistance in the conscious spontaneously hypertensive rat: euglycemic hyperinsulinemic clamp study. *Metabolism*, 42(1):14–18, Jan 1993.

[62] T. J. Aitman, T. Gotoda, A. L. Evans, H. Imrie, K. E. Heath, P. M. Trembling, H. Truman, C. A. Wallace, A. Rahman, C. DorÃ©, J. Flint, V. Kren, V. Zidek, T. W. Kurtz, M. Pravenec, and J. Scott. Quantitative trait loci for cellular defects in glucose and fatty acid metabolism in hypertensive rats. *Nat Genet*, 16(2):197–201, Jun 1997.

[63] Scott M Grundy, H. Bryan Brewer, James I. Cleeman, Sidney C. Smith, Claude Lenfant, American Heart Association, and National Heart Lung and Blood Institute. Definition of metabolic syndrome: Report of the national heart, lung, and blood institute/american heart association conference on scientific issues related to definition. *Circulation*, 109(3):433–438, Jan 2004.

[64] V. Kren. Genetics of the polydactyly-luxate syndrome in the norway rat, rattus norvegicus. *Acta Univ Carol Med Monogr*, (68):1–103, 1975.

[65] M. Pravenec, D. Gauguier, J. J. Schott, J. Buard, V. Kren, V. Bila, C. Szpirer, J. Szpirer, J. M. Wang, and H. Huang. Mapping of quantitative trait loci for blood pressure and cardiac mass in the rat by genome scanning of recombinant inbred strains. *J Clin Invest*, 96(4):1973–1978, Oct 1995.

[66] M. Otsen, M. den Bieman, M. T. Kuiper, M. Pravenec, V. Kren, T. W. Kurtz, H. J. Jacob, A. Lankhorst, and B. F. van Zutphen. Use of aflp markers for gene mapping and qtl detection in the rat. *Genomics*, 37(3):289–294, Nov 1996.

[67] Rebecca L Jaworski, Martin Jirout, Shamara Closson, Laura Breen, Pamela L Flodman, M. Anne Spence, Vladimir Kren, Drahomira Krenova, Michal Pravenec, and Morton P Printz. Heart rate and blood pressure quantitative trait loci for the airpuff startle reaction. *Hypertension*, 39(2 Pt 2):348–352, Feb 2002.

[68] A. Bottger, H. A. van Lith, V. Kren, D. Krenová, V. Bílá, J. Vorlícek, V. Zídek, A. Musilová, M. Zdobinská, J. M. Wang, B. F. van Zutphen, T. W. Kurtz, and M. Pravenec. Quantitative trait loci influencing cholesterol and phospholipid phenotypes map to chromosomes that contain genes regulating blood pressure in the spontaneously hypertensive rat. *J Clin Invest*, 98(3):856–862, Aug 1996.

[69] A. Bottger, E. Lankhorst, H. A. van Lith, L. F. van Zutphen, V. Zídek, A. Musilová, M. Simáková, R. Poledne, V. Bílá, V. Kren, and M. Pravenec. A genetic and correlation

analysis of liver cholesterol concentration in rat recombinant inbred strains fed a high cholesterol diet. *Biochem Biophys Res Commun*, 246(1):272–275, May 1998.

[70] M. Pravenec, V. Zidek, A. Musilova, V. Kren, V. Bila, and R. Di Nicolantonio. Chromosomal mapping of a major quantitative trait locus regulating compensatory renal growth in the rat. *J Am Soc Nephrol*, 11(7):1261–1265, Jul 2000.

[71] S. McCune, P. B. Baker, and F. H. Stills. Shhf/mcc-cp rat: model of obesity, non-insulin-dependent diabetes, and congestive heart failure. *Ilar News*, 32:23–27, 1990.

[72] W. B. Kannel and A. J. Belanger. Epidemiology of heart failure. *Am Heart J*, 121(3 Pt 1):951–957, Mar 1991.

[73] Daniel Levy, Satish Kenchaiah, Martin G Larson, Emelia J Benjamin, Michelle J Kupka, Kalon K L Ho, Joanne M Murabito, and Ramachandran S Vasan. Long-term trends in the incidence of and survival with heart failure. *N Engl J Med*, 347(18):1397–1402, Oct 2002.

[74] Theophilus E Owan, David O Hodge, Regina M Herges, Steven J Jacobsen, Veronique L Roger, and Margaret M Redfield. Trends in prevalence and outcome of heart failure with preserved ejection fraction. *N Engl J Med*, 355(3):251–259, Jul 2006.

[75] M. N. Sack, T. A. Rader, S. Park, J. Bastin, S. A. McCune, and D. P. Kelly. Fatty acid oxidation enzyme gene expression is downregulated in the failing heart. *Circulation*, 94(11):2837–2842, Dec 1996.

[76] Jonathan R R Heyen, Eileen R Blasi, Kristen Nikula, Ricardo Rocha, Heather A Daust, Gregory Frierdich, John F Van Vleet, Pam De Ciechi, Ellen G McMahon, and Amy E Rudolph. Structural, functional, and molecular characterization of the shhf model of heart failure. *Am J Physiol Heart Circ Physiol*, 283(5):H1775–H1784, Nov 2002.

[77] B. J. Holycross, B. M. Summers, R. B. Dunn, and S. A. McCune. Plasma renin activity in heart failure-prone shhf/mcc-facp rats. *Am J Physiol*, 273(1 Pt 2):H228–H233, Jul 1997.

[78] M. R. Bergman, R. H. Kao, S. A. McCune, and B. J. Holycross. Myocardial tumor necrosis factor-alpha secretion in hypertensive and heart failure-prone rats. *Am J Physiol*, 277(2 Pt 2):H543–H550, Aug 1999.

[79] Franklin A. Graybill. *Theory and Application of the Linear Model*. Duxbury, Pacific Grove, CA, 1976.

[80] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of statistical learning*. Springer, 2001.

[81] M. Ashburner, 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, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, May 2000.

[82] Minoru Kanehisa, Susumu Goto, Masahiro Hattori, Kiyoko F Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res*, 34(Database issue):D354–D357, Jan 2006.

[83] Seung Yon Rhee, Valerie Wood, Kara Dolinski, and Sorin Draghici. Use and misuse of the gene ontology annotations. *Nat Rev Genet*, 9(7):509–515, Jul 2008.

[84] Sorin Draghici, Purvesh Khatri, Rui P Martins, G. Charles Ostermeier, and Stephen A Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, Feb 2003.

[85] Bernard W. Lindgren. *Statistical Theory*. Chapman and Hall, 1998.

[86] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, Oct 2005.

[87] Lu Tian, Steven A Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S Kohane, and Peter J Park. Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A*, 102(38):13544–13549, Sep 2005.

[88] Zhen Jiang and Robert Gentleman. Extensions to gene set enrichment. *Bioinformatics*, Nov 2006.

[89] Manuela Hummel, Reinhard Meister, and Ulrich Mansmann. Globalancova: Exploration and assessment of gene group effects. *Bioinformatics*, Nov 2007.

[90] Jelle J Goeman and Peter B?hlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, Apr 2007.

[91] Marit Ackermann and Korbinian Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10:47, 2009.

[92] Bradley Efron. Simultaneous inference: When should hypothesis testing problems be combined? *Annals of Applied Statistics*, 2008,:Vol.2,No.1,197–223, March 2008.

[93] William T Barry, Andrew B Nobel, and Fred A Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–1949, May 2005.

[94] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[95] P. R. Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82:387–394, 1987.

[96] Jingdong Xie and Carey E. Priebe. A weighted generalization of the mann-whitney-wilcoxon statistic. *Journal of Statistical Planning and Inference*, 102(2):441 – 466, 2002.

[97] P. van Elteren. On the combination of independent two sample tests of wilcoxon. *Bulletin of the International Statistical Institute*, 37:351–261, 1960.

[98] Y.D. Zhao. Sample size estimation for the van Elteren test-a stratified Wilcoxon-Mann-Whitney test. *Statistics in medicine*, 25(15):2675–2687, 2005.

[99] Helge G Roider, Thomas Manke, Sean O'Keeffe, Martin Vingron, and Stefan A Haas. Pastaa: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics*, 25(4):435–442, Feb 2009.

[100] Robert McLeay and Timothy Bailey. Motif enrichment analysis: a unified framework and an evaluation on chip data. *BMC Bioinformatics*, 11(1):165+, 2010.

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

[102] Rachel B Brem, Ga?l Yvert, Rebecca Clinton, and Leonid Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755, Apr 2002.

[103] Eric E Schadt, Stephanie A Monks, Thomas A Drake, Aldons J Lusis, Nam Che, Veronica Colinayo, Thomas G Ruff, Stephen B Milligan, John R Lamb, Guy Cavet, Peter S Linsley, Mao Mao, Roland B Stoughton, and Stephen H Friend. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302, Mar 2003.

[104] Norbert Hubner, Caroline A Wallace, Heike Zimdahl, Enrico Petretto, Herbert Schulz, Fiona Maciver, Michael Mueller, Oliver Hummel, Jan Monti, Vaclav Zidek, Alena Musilova, Vladimir Kren, Helen Causton, Laurence Game, Gabriele Born, Sabine Schmidt, Anita M?ller, Stuart A Cook, Theodore W Kurtz, John Whittaker, Michal Pravenec, and Timothy J Aitman. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet*, 37(3):243–253, Mar 2005.

[105] F. JACOB and J. MONOD. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3:318–356, Jun 1961.

[106] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, Oct 1995.

[107] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–1680, Dec 1996.

[108] Christopher Lausted, Timothy Dahl, Charles Warren, Kimberly King, Kimberly Smith, Michael Johnson, Ramsey Saleem, John Aitchison, Lee Hood, and Stephen R Lasky.

Posam: a fast, flexible, open-source, inkjet oligonucleotide synthesizer and microarrayer. *Genome Biol*, 5(8):R58, 2004.

[109] Jian-Bing Fan, Kevin L Gunderson, Marina Bibikova, Joanne M Yeakley, Jing Chen, Eliza Wickham Garcia, Lori L Lebruska, Marc Laurent, Richard Shen, and David Barker. Illumina universal bead arrays. *Methods Enzymol*, 410:57–73, 2006.

[110] Affymetrix. *GeneChip Rat Genome 230 Arrays*.

[111] *Human HT 12 v3 Expression BeadChip*.

[112] S. P. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas. Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251(4995):767–773, Feb 1991.

[113] A. J. Hartemink, David K Gifford, T. S. Jaakola, and R. A. Young. Maximum likelihood estimation of optimal scaling factors for expression array normalization. *SPIE BiOS*, 2001.

[114] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, Jan 2003.

[115] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, Apr 2003.

[116] D. Holder, R.F. Raubertas, V.B. Pikounis, V. Svetnik, and K. Soper. Statistical analysis of high density oligonucleotide arrays: a SAFER approach. In *GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data*, 2001.

[117] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[118] Karl W Broman, Hao Wu, Saunak Sen, and Gary A Churchill. R/qtl: Qtl mapping in experimental crosses. *Bioinformatics*, 19(7):889–890, May 2003.

[119] Enrico Petretto, Rizwan Sarwar, Ian Grieve, Han Lu, Mande K Kumaran, Phillip J Muckett, Jonathan Mangion, Blanche Schroen, Matthew Benson, Prakash P Punjabi, Sanjay K Prasad, Dudley J Pennell, Chris Kiesewetter, Elena S Tasheva, Lolita M Corpuz, Megan D Webb, Gary W Conrad, Theodore W Kurtz, Vladimir Kren, Judith Fischer, Norbert Hubner, Yigal M Pinto, Michal Pravenec, Timothy J Aitman, and Stuart A Cook. Integrated genomic approaches implicate osteoglycin (ogn) in the regulation of left ventricular mass. *Nat Genet*, 40(5):546–552, May 2008.

[120] Rafael A Irizarry, Benjamin M Bolstad, Francois Collin, Leslie M Cope, Bridget Hobbs, and Terence P Speed. Summaries of affymetrix genechip probe level data. *Nucleic Acids Res*, 31(4):e15, Feb 2003.

[121] Jintao Wang, Robert W Williams, and Kenneth F Manly. Webqtl: web-based complex trait analysis. *Neuroinformatics*, 1(4):299–308, 2003.

[122] Ga?l Yvert, Rachel B Brem, Jacqueline Whittle, Joshua M Akey, Eric Foss, Erin N Smith, Rachel Mackelprang, and Leonid Kruglyak. Trans-acting regulatory variation in saccharomyces cerevisiae and the role of transcription factors. *Nat Genet*, 35(1):57–64, Sep 2003.

[123] Chunlei Wu, David L Delano, Nico Mitro, Stephen V Su, Jeff Janes, Phillip McClurg, Serge Batalov, Genevieve L Welch, Jie Zhang, Anthony P Orth, John R Walker, Richard J Glynne, Michael P Cooke, Joseph S Takahashi, Kazuhiro Shimomura, Akira Kohsaka, Joseph Bass, Enrique Saez, Tim Wiltshire, and Andrew I Su. Gene set enrichment in eqtl data identifies novel annotations and pathway regulators. *PLoS Genet*, 4(5):e1000070, May 2008.

[124] Rainer Breitling, Yang Li, Bruno M Tesson, Jingyuan Fu, Chunlei Wu, Tim Wiltshire, Alice Gerrits, Leonid V Bystrykh, Gerald de Haan, Andrew I Su, and Ritsert C Jansen. Genetical genomics: spotlight on qtl hotspots. *PLoS Genet*, 4(10):e1000232, Oct 2008.

[125] Phillip McClurg, Jeff Janes, Chunlei Wu, David L Delano, John R Walker, Serge Batalov, Joseph S Takahashi, Kazuhiro Shimomura, Akira Kohsaka, Joseph Bass, Tim Wiltshire, and Andrew I Su. Genomewide association analysis in diverse inbred mice: power and population structure. *Genetics*, 176(1):675–83, May 2007.

[126] A. P. Monaco, R. L. Neve, C. Colletti-Feener, C. J. Bertelson, D. M. Kurnit, and L. M. Kunkel. Isolation of candidate cdnas for portions of the duchenne muscular dystrophy gene. *Nature*, 323(6089):646–650, 1986.

[127] B. Kerem, J. M. Rommens, J. A. Buchanan, D. Markiewicz, T. K. Cox, A. Chakravarti, M. Buchwald, and L. C. Tsui. Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245(4922):1073–1080, Sep 1989.

[128] Margarete Mehrabian, Hooman Allayee, Jirina Stockton, Pek Yee Lum, Thomas A Drake, Lawrence W Castellani, Michael Suh, Christopher Armour, Stephen Edwards, John Lamb, Aldons J Lusis, and Eric E Schadt. Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nat Genet*, 37(11):1224–1233, Nov 2005.

[129] Eric E Schadt, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj Guhathakurta, Solveig K Sieberts, Stephanie Monks, Marc Reitman, Chunsheng Zhang, Pek Yee Lum, Amy Leonardson, Rolf Thieringer, Joseph M Metzger, Liming Yang, John Castle, Haoyuan Zhu, Shera F Kash, Thomas A Drake, Alan Sachs, and Aldons J Lusis. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*, 37(7):710–717, Jul 2005.

[130] Jacques Behmoaras, Gurjeet Bhangal, Jennifer Smith, Kylie McDonald, Brenda Mutch, Ping Chin Lai, Jan Domin, Laurence Game, Alan Salama, Brian M Foxwell, Charles D Pusey, H. Terence Cook, and Timothy J Aitman. Jund is a determinant of macrophage

activation and is associated with glomerulonephritis susceptibility. *Nat Genet*, 40(5):553–559, May 2008.

[131] C. L. Karp, A. Grupe, E. Schadt, S. L. Ewart, M. Keane-Moore, P. J. Cuomo, J. Köhl, L. Wahl, D. Kuperman, S. Germer, D. Aud, G. Peltz, and M. Wills-Karp. Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. *Nat Immunol*, 1(3):221–226, Sep 2000.

[132] Iain A Eaves, Linda S Wicker, Ghassan Ghandour, Paul A Lyons, Laurence B Peterson, John A Todd, and Richard J Glynne. Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the nod model of type 1 diabetes. *Genome Res*, 12(2):232–243, Feb 2002.

[133] E. St Lezin, W. Liu, J. M. Wang, N. Wang, V. Kren, D. Krenova, A. Musilova, M. Zdobinska, V. Zidek, D. Lau, and M. Pravenec. Genetic isolation of a chromosome 1 region affecting blood pressure in the spontaneously hypertensive rat. *Hypertension*, 30(4):854–859, Oct 1997.

[134] S. A. Frantz, M. Kaiser, S. M. Gardiner, D. Gauguier, M. Vincent, J. R. Thompson, T. Bennett, and N. J. Samani. Successful isolation of a rat chromosome 1 blood pressure quantitative trait locus in reciprocal congenic strains. *Hypertension*, 32(4):639–646, Oct 1998.

[135] N. Hübner, Y. A. Lee, K. Lindpaintner, D. Ganten, and R. Kreutz. Congenic substitution mapping excludes sa as a candidate gene locus for a blood pressure quantitative trait locus on rat chromosome 1. *Hypertension*, 34(4 Pt 1):643–648, Oct 1999.

[136] Albert-L?szl? Barab?si and Zolt?n N Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, Feb 2004.

[137] Juan M Vaquerizas, Sarah K Kummerfeld, Sarah A Teichmann, and Nicholas M Luscombe. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, 10(4):252–263, Apr 2009.

[138] Malin C Andersen, Pär G Engström, Stuart Lithwick, David Arenillas, Per Eriksson, Boris Lenhard, Wyeth W Wasserman, and Jacob Odeberg. In silico detection of sequence variations modifying transcriptional regulation. *PLoS Comput Biol*, 4(1):e5, Jan 2008.

[139] Thomas Manke, Helge G Roider, and Martin Vingron. Statistical modeling of transcription factor binding affinities predicts regulatory interactions. *PLoS Comput Biol*, 4(3):e1000039, Mar 2008.

[140] Christopher T Harbison, D Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, Jean-Bosco Tagne, David B Reynolds, Jane Yoo, Ezra G Jennings, Julia Zeitlinger, Dmitry K Pokholok, Manolis Kellis, P Alex Rolfe, Ken T Takusagawa, Eric S Lander, David K Gifford, Ernest Fraenkel, and Richard A Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, Sep 2004.

[141] Sonali Mukherjee, Michael F Berger, Ghil Jona, Xun S Wang, Dale Muzzey, Michael Snyder, Richard A Young, and Martha L Bulyk. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet*, 36(12):1331–9, Dec 2004.

[142] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–D110, Jan 2006.

[143] Helge G Roider, Aditi Kanhere, Thomas Manke, and Martin Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–141, Jan 2007.

[144] V. Matys, E. Fricke, R. Geffers, E. Goessling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Münch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31(1):374–378, Jan 2003.

[145] Marko Djordjevic, Anirvan M Sengupta, and Boris I Shraiman. A biophysical approach to transcription factor binding site discovery. *Genome Res*, 13(11):2381–90, Nov 2003.

[146] Barrett C Foat, Alexandre V Morozov, and Harmen J Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, 22(14):e141–9, Jul 2006.

[147] Amos Tanay. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res*, 16(8):962–972, Aug 2006.

[148] Eran Segal, Tali Raveh-Sadka, Mark Schroeder, Ulrich Unnerstall, and Ulrike Gaul. Predicting expression patterns from regulatory sequence in drosophila segmentation. *Nature*, 451(7178):535–540, Jan 2008.

[149] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res*, 29(1):308–311, Jan 2001.

[150] Michael F Berger and Martha L Bulyk. Universal protein-binding microarrays for the comprehensive characterization of the dna-binding specificities of transcription factors. *Nat Protoc*, 4(3):393–411, 2009.

[151] Gerhard Mittler, Falk Butter, and Matthias Mann. A silac-based dna protein interaction screen that identifies candidate binding proteins to functional dna elements. *Genome Res*, 19(2):284–293, Feb 2009.

[152] Shunbin Ning, Leslie E Huye, and Joseph S Pagano. Regulation of the transcriptional activity of the irf7 promoter by a pathway independent of interferon signaling. *J Biol Chem*, 280(13):12262–12270, Apr 2005.

[153] Kenya Honda, Hideyuki Yanai, Hideo Negishi, Masataka Asagiri, Mitsuharu Sato, Tatsuaki Mizutani, Naoya Shimada, Yusuke Ohba, Akinori Takaoka, Nobuaki Yoshida, and Tadatsugu Taniguchi. Irf-7 is the master regulator of type-i interferon-dependent immune responses. *Nature*, 434(7034):772–777, Apr 2005.

[154] Korbinian Strimmer. A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9(1):303, 2008.

[155] R. A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1932.

[156] Ann Hess and Hari Iyer. Fisher's combined p-value for detecting differentially expressed genes using affymetrix expression arrays. *BMC Genomics*, 8:96, 2007.

[157] Enrico Petretto, Leonardo Bottolo, Sarah R Langley, Matthias Heinig, Chris McDermott-Roe, Rizwan Sarwar, Michal Pravenec, Norbert Hübner, Timothy J Aitman, Stuart A Cook, and Sylvia Richardson. New insights into the genetic control of gene expression using a bayesian multi-tissue approach. *PLoS Comput Biol*, 6(4):e1000737, 2010.

[158] Carl Nathan and Aihao Ding. Nonresolving inflammation. *Cell*, 140(6):871–882, Mar 2010.

[159] Décio L Eizirik, Maikel L Colli, and Fernanda Ortis. The role of inflammation in insulitis and beta-cell loss in type 1 diabetes. *Nat Rev Endocrinol*, 5(4):219–226, Apr 2009.

[160] C. L. Holness and D. L. Simmons. Molecular cloning of cd68, a human macrophage marker related to lysosomal glycoproteins. *Blood*, 81(6):1607–1613, Mar 1993.

[161] Santosh S Atanur, Inanç Birol, Victor Guryev, Martin Hirst, Oliver Hummel, Catherine Morrissey, Jacques Behmoaras, Xose M Fernandez-Suarez, Michelle D Johnson, William M McLaren, Giannino Patone, Enrico Petretto, Charles Plessy, Kathleen S Rockland, Charles Rockland, Kathrin Saar, Yongjun Zhao, Piero Carninci, Paul Flicek, Ted Kurtz, Edwin Cuppen, Michal Pravenec, Norbert Hubner, Steven J M Jones, Ewan Birney, and Timothy J Aitman. The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. *Genome Res*, Apr 2010.

[162] João P Pereira, Lisa M Kelly, Ying Xu, and Jason G Cyster. Ebi2 mediates b cell segregation between the outer and centre follicle. *Nature*, 460(7259):1122–1126, Aug 2009.

[163] Dominique Gatto, Didrik Paus, Antony Basten, Charles R Mackay, and Robert Brink. Guidance of b cells by the orphan g protein-coupled receptor ebi2 shapes humoral immune responses. *Immunity*, 31(2):259–269, Aug 2009.

[164] Stefan Blankenberg. The gutenberg heart study. Personal communication, 2010.

[165] Cardiogenics consortium. The cardiogenics study. Personal communication, 2010.

[166] D. F Morrison. *Multivariate Statistical methods*. Belmont, 2004.

[167] Chester L. Olson. Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, 69(348):894–908, 1974.

[168] G. E. P. Box. Non-normality and tests on variances. *Biometrika*, 40:318–335, 1953.

[169] Ravinder Nath and Robert Pavur. A new statistic in the one-way multivariate analysis of variance. *Computational Statistics & Data Analysis*, 2(4):297 – 315, 1985.

[170] R. Tibshirani. The lasso method for variable selection in the cox model. *Stat Med*, 16(4):385–395, Feb 1997.

[171] Trevor Hastie and Brad Efron. *lars: Least Angle Regression, Lasso and Forward Stagewise*, 2007. R package version 0.9-7.

[172] Vincent Plagnol, Deborah J Smyth, John A Todd, and David G Clayton. Statistical independence of the colocalized association signals for type 1 diabetes and rps26 gene expression on chromosome 12q13. *Biostatistics*, 10(2):327–334, Apr 2009.

[173] Michele A Rodrigues, Dawidson A Gomes, M. Fatima Leite, Wayne Grant, Lei Zhang, Wing Lam, Yung-Chi Cheng, Anton M Bennett, and Michael H Nathanson. Nucleoplasmic calcium is required for cell proliferation. *J Biol Chem*, 282(23):17061–17068, Jun 2007.

[174] Matthias von Herrath. Diabetes: A virus-gene collaboration. *Nature*, 459(7246):518–519, May 2009.

[175] Sergey Nejentsev, Neil Walker, David Riches, Michael Egholm, and John A Todd. Rare variants of ifih1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, 324(5925):387–389, Apr 2009.

[176] Deborah J Smyth, Jason D Cooper, Rebecca Bailey, Sarah Field, Oliver Burren, Luc J Smink, Cristian Guja, Constantin Ionescu-Tirgoviste, Barry Widmer, David B Dunger, David A Savage, Neil M Walker, David G Clayton, and John A Todd. A genome-wide association study of nonsynonymous snps identifies a type 1 diabetes locus in the interferon-induced helicase (ifih1) region. *Nat Genet*, 38(6):617–619, Jun 2006.

[177] Qing Li, Baohui Xu, Sara A Michie, Kathleen H Rubins, Robert D Schreriber, and Hugh O McDevitt. Interferon-alpha initiates type 1 diabetes in nonobese diabetic mice. *Proc Natl Acad Sci U S A*, 105(34):12439–12444, Aug 2008.

[178] R. A. Irizarry, C. Wang, Y. Zhou, and T. P. Speed. *Working Papers*, chapter Working Paper 185. Johns Hopkins University, 2009.

[179] Marit Holden, Shiwei Deng, Leszek Wojnowski, and Bettina Kulle. Gsea-snp: applying gene set enrichment analysis to snp data from genome-wide association studies. *Bioinformatics*, 24(23):2784–2785, Dec 2008.

[180] N Mantel. Chi-square tests with one degree of freedom: Extensions of the mantel-haenszel procedure. *J Am Stat Assoc*, pages 690–700, 1963.

[181] Jeffrey C Barrett, David G Clayton, Patrick Concannon, Beena Akolkar, Jason D Cooper, Henry A Erlich, Cécile Julier, Grant Morahan, Jørn Nerup, Concepcion Nierras, Vincent Plagnol, Flemming Pociot, Helen Schuilenburg, Deborah J Smyth, Helen Stevens, John A Todd, Neil M Walker, Stephen S Rich, and The Type 1 Diabetes Genetics Consortium. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet*, May 2009.

[182] J. D. Cooper, N. M. Walker, D. J. Smyth, K. Downes, B. C. Healy, J. A. Todd, and Type I Diabetes Genetics Consortium. Follow-up of 1715 snps from the wellcome trust case control consortium genome-wide association study in type i diabetes families. *Genes Immun*, 10 Suppl 1:S85–S94, Dec 2009.

[183] Taro Kawai, Shintaro Sato, Ken J Ishii, Cevayir Coban, Hiroaki Hemmi, Masahiro Yamamoto, Kenta Terai, Michiyuki Matsuda, Jun ichiro Inoue, Satoshi Uematsu, Osamu Takeuchi, and Shizuo Akira. Interferon-alpha induction through toll-like receptors involves a direct interaction of irf7 with myd88 and traf6. *Nat Immunol*, 5(10):1061–1068, Oct 2004.

[184] Michael R. Mehan, Juan Nunez-Iglesias, Mrinal Kalakrishnan, Michael S. Waterman, and Xianghong Jasmine Zhou. An integrative network approach to map the transcriptome to the phenome. In Martin Vingron and Limsoon Wong, editors, *Research in Computational Molecular Biology*, volume 4955 of *LNBI*, pages 232–245. Springer, 2008.

[185] S Falcon and R Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–8, 2007.

[186] E. Anderson. *Lapack Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, 1999.

[187] R. M. Berne. Effect of epinephrine and norepinephrine on coronary circulation. *Circ Res*, 6(5):644–655, Sep 1958.

[188] Julia M Orshal and Raouf A Khalil. Gender, sex hormones, and vascular tone. *Am J Physiol Regul Integr Comp Physiol*, 286(2):R233–R249, Feb 2004.

[189] Carmine Savoia and Ernesto L Schiffrin. Inflammation in hypertension. *Curr Opin Nephrol Hypertens*, 15(2):152–158, Mar 2006.

[190] Michael L Metzker. Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46, Jan 2010.

[191] Alison Abbott. Return of the rat. Online, 2009.

[192] David N Arnosti and Meghana M Kulkarni. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem*, 94(5):890–898, Apr 2005.

[193] 1000 Genomes Consortium. The 1000 genomes project. Online, 2009.

[194] Kelly A Frazer, Eleazar Eskin, Hyun Min Kang, Molly A Bogue, David A Hinds, Erica J Beilharz, Robert V Gupta, Julie Montgomery, Matt M Morenzoni, Geoffrey B Nilsen, Charit L Pethiyagoda, Laura L Stuve, Frank M Johnson, Mark J Daly, Claire M Wade, and David R Cox. A sequence-based variation map of 8.27 million snps in inbred mouse strains. *Nature*, 448(7157):1050–1053, Aug 2007.

[195] Jan Christian Bryne, Eivind Valen, Man-Hung Eric Tang, Troels Marstrand, Ole Winther, Isabelle da Piedade, Anders Krogh, Boris Lenhard, and Albin Sandelin. Jaspar, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res*, 36(Database issue):D102–D106, Jan 2008.

[196] E. N. C. O. D. E. Project Consortium. Identification and analysis of functional elements in 1genome by the encode pilot project. *Nature*, 447(7146):799–816, Jun 2007.

[197] Christopher Newton-Cheh, Toby Johnson, Vesela Gateva, Martin D Tobin, Murielle Bochud, Lachlan Coin, Samer S Najjar, Jing Hua Zhao, Simon C Heath, Susana Eyheramendy, Konstantinos Papadakis, Benjamin F Voight, Laura J Scott, Feng Zhang, Martin Farrall, Toshiko Tanaka, Chris Wallace, John C Chambers, Kay-Tee Khaw, Peter Nilsson, Pim van der Harst, Silvia Polidoro, Diederick E Grobbee, N. Charlotte Onland-Moret, Michiel L Bots, Louise V Wain, Katherine S Elliott, Alexander Teumer, Jian'an Luan, Gavin Lucas, Johanna Kuusisto, Paul R Burton, David Hadley, Wendy L McArdle, Wellcome Trust Case Control Consortium, Morris Brown, Anna Dominiczak, Stephen J Newhouse, Nilesh J Samani, John Webster, Eleftheria Zeggini, Jacques S Beckmann, Sven Bergmann, Noha Lim, Kijoung Song, Peter Vollenweider, Gerard Waeber, Dawn M Waterworth, Xin Yuan, Leif Groop, Marju Orho-Melander, Alessandra Allione, Alessandra Di Gregorio, Simonetta Guarrera, Salvatore Panico, Fulvio Ricceri, Valeria Romanazzi, Carlotta Sacerdote, Paolo Vineis, Inês Barroso, Manjinder S Sandhu, Robert N Luben, Gabriel J Crawford, Pekka Jousilahti, Markus Perola, Michael Boehnke, Lori L Bonnycastle, Francis S Collins, Anne U Jackson, Karen L Mohlke, Heather M Stringham, Timo T Valle, Cristen J Willer, Richard N Bergman, Mario A Morken, Angela Döring, Christian Gieger, Thomas Illig, Thomas Meitinger, Elin Org, Arne Pfeufer, H. Erich Wichmann, Sekar Kathiresan, Jaume Marrugat, Christopher J O'Donnell, Stephen M Schwartz, David S Siscovick, Isaac Subirana, Nelson B Freimer, Anna-Liisa Hartikainen, Mark I McCarthy, Paul F O'Reilly, Leena Peltonen, Anneli Pouta, Paul E de Jong, Harold Snieder, Wiek H van Gilst, Robert Clarke, Anuj Goel, Anders Hamsten, John F Peden, Udo Seedorf, Ann-Christine Syvänen, Giovanni Tognoni, Edward G Lakatta, Serena Sanna, Paul Scheet, David Schlessinger, Angelo Scuteri, Marcus Dörr, Florian Ernst, Stephan B Felix, Georg Homuth, Roberto Lorbeer, Thorsten Reffelmann, Rainer Rettig, Uwe Völker, Pilar Galan, Ivo G Gut, Serge Hercberg, G. Mark Lathrop, Diana Zelenika, Panos Deloukas, Nicole Soranzo, Frances M Williams, Guangju Zhai, Veikko Salomaa, Markku Laakso, Roberto Elosua, Nita G Forouhi, Henry Völzke, Cuno S Uiterwaal, Yvonne T van der Schouw, Mattijs E Numans, Giuseppe Matullo, Gerjan Navis, Göran Berglund, Sheila A Bingham, Jaspal S Kooner, John M Connell, Stefania Bandinelli, Luigi Ferrucci, Hugh

Watkins, Tim D Spector, Jaakko Tuomilehto, David Altshuler, David P Strachan, Maris Laan, Pierre Meneton, Nicholas J Wareham, Manuela Uda, Marjo-Riitta Jarvelin, Vincent Mooser, Olle Melander, Ruth Jf Loos, Paul Elliott, Gonçalo R Abecasis, Mark Caulfield, and Patricia B Munroe. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet*, May 2009.

[198] C-M. Chung, R-Y. Wang, J-W. Chen, C. S J Fann, H-B. Leu, H-Y. Ho, C-T. Ting, T-H. Lin, S-H. Sheu, W-C. Tsai, J-H. Chen, Y-S. Jong, S-J. Lin, Y-T. Chen, and W-H. Pan. A genome-wide association study identifies new loci for ace activity: potential implications for response to ace inhibitor. *Pharmacogenomics J*, Jan 2010.

[199] Hans-Hilger Ropers. New perspectives for the elucidation of genetic disorders. *Am J Hum Genet*, 81(2):199–207, Aug 2007.

A Co-expression as quantitative trait

A.1 Supplementary figures

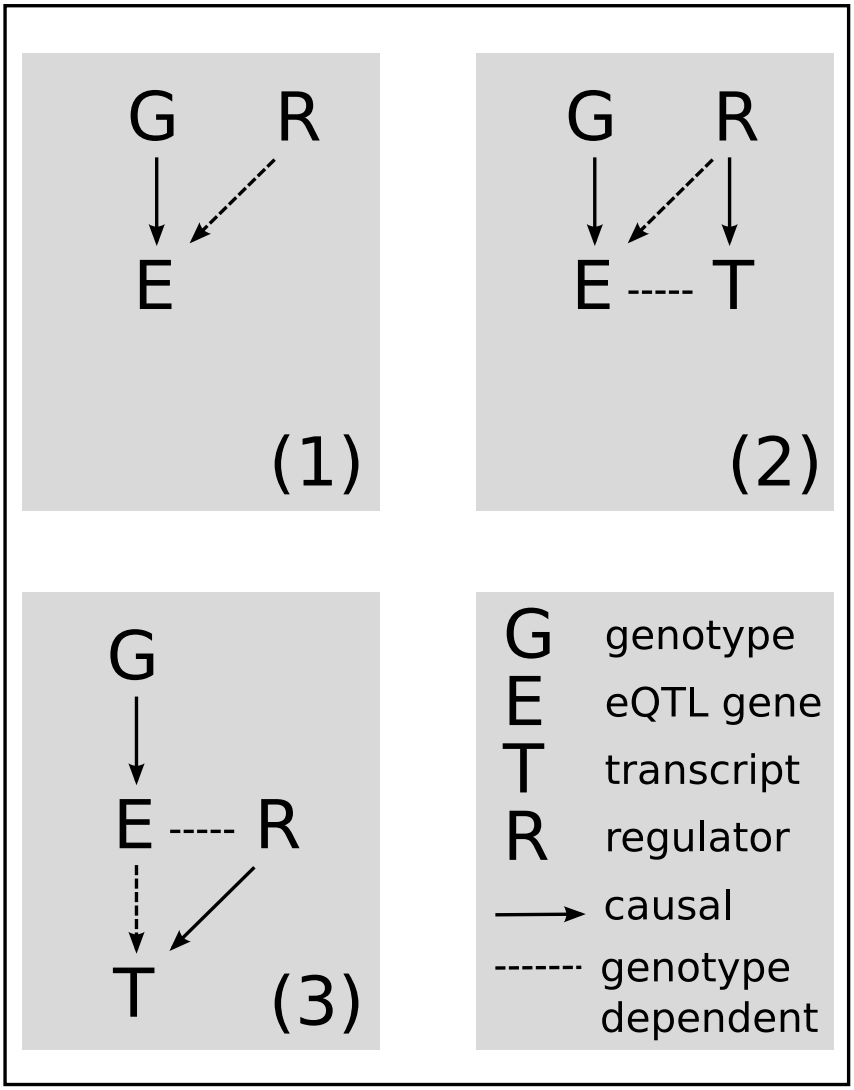

Figure A.1: **There are three possible causal scenarios where genotype dependent co-expression can arise.** (1) the simplest one is, that G and R regulate E and one allele at G overrides the regulatory action of R. This leads to differential co-expression between E and R. (2) the two other models include a second downstream transcript T. If E is regulated by R and G, as above, but additionally R is regulating T, then we expect differential co-expression between E and R as well as between E and T. (3) if E is regulating T and we observe differential co-expression between E and T and the expression of E is dependent on G, there must exist a regulator R upstream of T that overrides the regulatory action of E. In this case there is differential co-expression between E and T, and between E and R.

A.2 Supplementary tables

Table A.1: GO enrichment in subnetworks

hub	ID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1376780_at	GO:0004888	5.54E−04	3.02	5.49	15	482	transmembrane receptor activity
1376780_at	GO:0004871	5.71E−04	2.36	12.35	27	1205	signal transducer activity
1376780_at	GO:0004872	6.91E−04	2.70	7.30	18	659	receptor activity
1376780_at	GO:0003690	1.89E−03	8.54	0.54	4	44	double-stranded DNA binding
1376780_at	GO:0001584	2.51E−03	3.74	2.35	8	197	rhodopsin-like receptor activity
1376780_at	GO:0015291	2.70E−03	4.74	1.40	6	116	porter activity
1376780_at	GO:0015290	2.82E−03	4.70	1.41	6	117	electrochemical potential-driven transporter activity
1376780_at	GO:0005543	3.50E−03	7.11	0.63	4	52	phospholipid binding
1376780_at	GO:0004930	5.09E−03	3.05	3.18	9	270	G-protein coupled receptor activity
1376780_at	GO:0001614	7.23E−03	8.43	0.40	3	33	purinergic nucleotide receptor activity
1376780_at	GO:0001608	7.23E−03	8.43	0.40	3	33	nucleotide receptor activity, G-protein coupled
1376780_at	GO:0016502	7.23E−03	8.43	0.40	3	33	nucleotide receptor activity
1376780_at	GO:0045028	7.23E−03	8.43	0.40	3	33	purinergic nucleotide receptor activity, G-protein coupled
1376780_at	GO:0043566	8.19E−03	5.49	0.80	4	66	structure-specific DNA binding
1376780_at	GO:0008289	8.31E−03	3.70	1.75	6	146	lipid binding
1376780_at	GO:0008406	8.42E−04	10.84	0.44	4	36	gonad development
1376780_at	GO:0045137	8.42E−04	10.84	0.44	4	36	development of primary sexual characteristics
1376780_at	GO:0007154	2.68E−03	2.04	16.06	32	1714	cell communication
1376780_at	GO:0007548	2.69E−03	7.69	0.59	4	49	sex differentiation
1376780_at	GO:0019226	3.37E−03	3.06	3.54	10	307	transmission of nerve impulse
1376780_at	GO:0007186	3.59E−03	2.87	4.14	11	362	G-protein coupled receptor protein signaling pathway
1376780_at	GO:0007267	4.21E−03	2.58	5.40	13	481	cell-cell signaling
1376780_at	GO:0009987	5.64E−03	Inf	33.93	70	5414	cellular process
1376780_at	GO:0050877	7.47E−03	2.39	5.75	13	515	neurophysiological process
1376780_at	GO:0007165	8.60E−03	1.89	14.19	27	1464	signal transduction
1376780_at	GO:0007166	9.76E−03	2.13	7.77	16	719	cell surface receptor linked signal transduction
1389690_at	GO:0003690	2.16E−04	6.66	1.28	7	44	double-stranded DNA binding
1389690_at	GO:0043566	5.27E−04	4.87	1.92	8	66	structure-specific DNA binding
1389690_at	GO:0015280	1.15E−03	20.75	0.23	3	8	amiloride-sensitive sodium channel activity
1389690_at	GO:0015276	2.59E−03	4.16	1.92	7	66	ligand-gated ion channel activity
1389690_at	GO:0008289	2.91E−03	2.89	4.19	11	146	lipid binding
1389690_at	GO:0015268	6.66E−03	2.20	7.27	15	258	alpha-type channel activity
1389690_at	GO:0015267	6.90E−03	2.19	7.30	15	259	channel or pore class transporter activity
1389690_at	GO:0005216	7.85E−03	2.22	6.73	14	238	ion channel activity
1389690_at	GO:0046872	8.13E−03	1.59	26.39	43	1058	metal ion binding
1389690_at	GO:0043167	8.13E−03	1.59	26.39	43	1058	ion binding
1389690_at	GO:0007588	2.67E−04	7.93	0.95	6	33	excretion
1389690_at	GO:0006607	7.05E−04	26.36	0.20	3	7	NLS-bearing substrate import into nucleus
1389690_at	GO:0006821	2.98E−03	7.84	0.63	4	22	chloride transport
1389690_at	GO:0016079	3.53E−03	7.43	0.66	4	23	synaptic vesicle exocytosis
1389690_at	GO:0048489	5.49E−03	4.92	1.18	5	41	synaptic vesicle transport
1389690_at	GO:0016338	6.34E−03	9.57	0.40	3	14	calcium-independent cell-cell adhesion
1374583_at	GO:0006576	4.89E−03	9.76	0.35	3	52	biogenic amine metabolism
1369836_at	GO:0004888	5.68E−05	3.67	5.04	16	482	transmembrane receptor activity
1369836_at	GO:0008289	1.97E−04	5.69	1.61	8	146	lipid binding
1369836_at	GO:0004872	2.26E−04	3.02	6.71	18	659	receptor activity
1369836_at	GO:0001584	3.04E−04	4.74	2.15	9	197	rhodopsin-like receptor activity
1369836_at	GO:0004930	7.38E−04	3.82	2.92	10	270	G-protein coupled receptor activity
1369836_at	GO:0030594	1.10E−03	7.16	0.79	5	71	neurotransmitter receptor activity
1369836_at	GO:0003690	1.38E−03	9.35	0.49	4	44	double-stranded DNA binding
1369836_at	GO:0042165	1.59E−03	6.56	0.86	5	77	neurotransmitter binding
1369836_at	GO:0008528	2.21E−03	6.05	0.92	5	83	peptide receptor activity, G-protein coupled

1369836_at	GO:0001653	2.33E−03	5.97	0.94	5	84	peptide receptor activity
1369836_at	GO:0005543	2.57E−03	7.78	0.58	4	52	phospholipid binding
1369836_at	GO:0001614	5.71E−03	9.22	0.37	3	33	purinergic nucleotide receptor activity
1369836_at	GO:0001608	5.71E−03	9.22	0.37	3	33	nucleotide receptor activity, G-protein coupled
1369836_at	GO:0016502	5.71E−03	9.22	0.37	3	33	nucleotide receptor activity
1369836_at	GO:0045028	5.71E−03	9.22	0.37	3	33	purinergic nucleotide receptor activity, G-protein coupled
1369836_at	GO:0015276	6.09E−03	6.01	0.74	4	66	ligand-gated ion channel activity
1369836_at	GO:0043566	6.09E−03	6.01	0.74	4	66	structure-specific DNA binding
1369836_at	GO:0042277	8.90E−03	4.27	1.27	5	115	peptide binding
1369836_at	GO:0008227	9.12E−03	7.68	0.44	3	39	amine receptor activity
1369836_at	GO:0004871	9.46E−03	1.96	11.33	22	1205	signal transducer activity
1369836_at	GO:0005230	9.79E−03	7.47	0.45	3	40	extracellular ligand-gated ion channel activity
1369836_at	GO:0007154	1.53E−04	2.58	14.67	33	1714	cell communication
1369836_at	GO:0007186	4.76E−04	3.57	3.78	12	362	G-protein coupled receptor protein signaling pathway
1369836_at	GO:0007267	5.71E−04	3.18	4.93	14	481	cell-cell signaling
1369836_at	GO:0043085	8.62E−04	6.04	1.12	6	103	positive regulation of enzyme activity
1369836_at	GO:0048489	9.88E−04	10.32	0.45	4	41	synaptic vesicle transport
1369836_at	GO:0019226	1.70E−03	3.41	3.24	10	307	transmission of nerve impulse
1369836_at	GO:0016079	1.91E−03	14.12	0.25	3	23	synaptic vesicle exocytosis
1369836_at	GO:0007165	1.99E−03	2.20	12.96	27	1464	signal transduction
1369836_at	GO:0019933	2.65E−03	5.80	0.96	5	88	cAMP-mediated signaling
1369836_at	GO:0001505	2.93E−03	5.66	0.98	5	90	regulation of neurotransmitter levels
1369836_at	GO:0050877	3.37E−03	2.67	5.25	13	515	neurophysiological process
1369836_at	GO:0007269	3.60E−03	7.05	0.64	4	58	neurotransmitter secretion
1369836_at	GO:0007268	3.72E−03	3.24	3.03	9	286	synaptic transmission
1369836_at	GO:0007166	3.89E−03	2.40	7.10	16	719	cell surface receptor linked signal transduction
1369836_at	GO:0045055	4.85E−03	6.45	0.69	4	63	regulated secretory pathway
1369836_at	GO:0045761	6.97E−03	8.54	0.40	3	36	regulation of adenylate cyclase activity
1369836_at	GO:0048609	7.42E−03	5.67	0.78	4	71	reproductive organismal physiological process
1369836_at	GO:0019935	7.43E−03	4.48	1.22	5	112	cyclic-nucleotide-mediated signaling
1369836_at	GO:0007189	7.53E−03	8.29	0.41	3	37	G-protein signaling, adenylate cyclase activating pathway
1369836_at	GO:0031279	7.53E−03	8.29	0.41	3	37	regulation of cyclase activity
1369836_at	GO:0051339	7.53E−03	8.29	0.41	3	37	regulation of lyase activity
1369836_at	GO:0050876	7.79E−03	5.58	0.79	4	72	reproductive physiological process
1369836_at	GO:0007188	8.98E−03	5.34	0.82	4	75	G-protein signaling, coupled to cAMP nucleotide second messenger
1369836_at	GO:0050874	9.74E−03	1.95	11.86	23	1311	organismal physiological process
1376840_at	GO:0003677	3.37E−03	5.52	1.47	6	661	DNA binding
1376840_at	GO:0003676	4.86E−03	4.68	2.05	7	964	nucleic acid binding
1376840_at	GO:0006139	9.24E−03	4.40	2.26	7	1225	nucleobase, nucleoside, nucleotide and nucleic acid metabolism
1384717_at	GO:0003682	2.28E−04	15.61	0.31	4	39	chromatin binding
1384717_at	GO:0016563	1.83E−03	5.22	1.29	6	167	transcriptional activator activity
1384717_at	GO:0030246	9.37E−03	5.30	0.83	4	106	carbohydrate binding
1384717_at	GO:0007167	2.46E−03	4.92	1.37	6	177	enzyme linked receptor protein signaling pathway
1384717_at	GO:0006351	4.34E−03	2.78	4.74	12	664	transcription, DNA-dependent
1370988_at	GO:0003779	1.93E−03	8.66	0.54	4	114	actin binding
1370988_at	GO:0008092	2.37E−03	6.20	0.94	5	202	cytoskeletal protein binding
1370988_at	GO:0006937	2.37E−04	30.29	0.12	3	29	regulation of muscle contraction
1370988_at	GO:0042060	4.04E−04	13.62	0.36	4	84	wound healing
1370988_at	GO:0050817	1.06E−03	17.45	0.21	3	48	coagulation
1370988_at	GO:0007596	1.06E−03	17.45	0.21	3	48	blood coagulation
1370988_at	GO:0007599	1.27E−03	16.35	0.22	3	51	hemostasis
1370988_at	GO:0050878	3.29E−03	11.50	0.30	3	71	regulation of body fluids
1370988_at	GO:0007017	4.45E−03	10.27	0.34	3	79	microtubule-based process

1370988_at	GO:0006936	9.80E−03	7.62	0.44	3	105	muscle contraction
1370114_a_at	GO:0006607	3.02E−07	435.75	0.02	3	7	NLS-bearing substrate import into nucleus
1370114_a_at	GO:0006606	1.25E−04	40.26	0.10	3	46	protein import into nucleus
1370114_a_at	GO:0051170	1.25E−04	40.26	0.10	3	46	nuclear import
1370114_a_at	GO:0017038	1.60E−04	36.81	0.11	3	50	protein import
1370114_a_at	GO:0051169	4.36E−04	25.73	0.15	3	70	nuclear transport
1370114_a_at	GO:0006913	6.22E−04	22.65	0.17	3	79	nucleocytoplasmic transport
1370114_a_at	GO:0006605	2.81E−03	13.12	0.29	3	133	protein targeting
1379896_at	GO:0008202	1.62E−05	41.51	0.17	4	114	steroid metabolism
1379896_at	GO:0008610	5.20E−05	30.44	0.23	4	153	lipid biosynthesis
1379896_at	GO:0006694	6.42E−05	55.44	0.08	3	55	steroid biosynthesis
1379896_at	GO:0044255	6.72E−05	20.52	0.50	5	339	cellular lipid metabolism
1379896_at	GO:0006629	1.60E−04	16.89	0.59	5	406	lipid metabolism
1379896_at	GO:0044249	8.55E−03	7.17	0.83	4	588	cellular biosynthesis
1373232_at	GO:0008289	6.02E−03	6.18	0.73	4	146	lipid binding
1373232_at	GO:0007267	1.44E−03	4.50	2.14	8	481	cell-cell signaling
1373232_at	GO:0007268	2.00E−03	5.38	1.32	6	286	synaptic transmission
1373232_at	GO:0019226	2.86E−03	4.98	1.41	6	307	transmission of nerve impulse
1373232_at	GO:0050877	9.26E−03	3.47	2.28	7	515	neurophysiological process
1376958_at	GO:0005319	3.04E−10	19.48	0.86	11	38	lipid transporter activity
1376958_at	GO:0004497	1.15E−07	9.88	1.43	11	64	monooxygenase activity
1376958_at	GO:0016712	1.83E−07	16.99	0.68	8	30	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen
1376958_at	GO:0050381	1.83E−07	16.99	0.68	8	30	unspecific monooxygenase activity
1376958_at	GO:0016705	1.05E−06	7.68	1.77	11	79	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
1376958_at	GO:0016491	1.88E−06	3.28	9.09	26	430	oxidoreductase activity
1376958_at	GO:0008289	1.79E−05	4.67	3.23	13	146	lipid binding
1376958_at	GO:0017127	4.19E−05	135.36	0.09	3	4	cholesterol transporter activity
1376958_at	GO:0008236	9.76E−05	5.53	1.90	9	85	serine-type peptidase activity
1376958_at	GO:0015248	1.03E−04	67.67	0.11	3	5	sterol transporter activity
1376958_at	GO:0004866	1.31E−04	7.18	1.17	7	52	endopeptidase inhibitor activity
1376958_at	GO:0004857	1.70E−04	4.21	2.97	11	134	enzyme inhibitor activity
1376958_at	GO:0030414	1.88E−04	6.73	1.24	7	55	protease inhibitor activity
1376958_at	GO:0004867	2.71E−04	10.38	0.61	5	27	serine-type endopeptidase inhibitor activity
1376958_at	GO:0005507	4.53E−04	9.13	0.68	5	30	copper ion binding
1376958_at	GO:0016789	7.06E−04	5.29	1.52	7	68	carboxylic ester hydrolase activity
1376958_at	GO:0004252	9.17E−04	5.04	1.59	7	71	serine-type endopeptidase activity
1376958_at	GO:0046906	9.44E−04	7.60	0.79	5	35	tetrapyrrole binding
1376958_at	GO:0020037	9.44E−04	7.60	0.79	5	35	heme binding
1376958_at	GO:0003824	1.19E−03	1.73	37.79	69	2316	catalytic activity
1376958_at	GO:0005215	1.60E−03	1.97	15.68	30	783	transporter activity
1376958_at	GO:0046872	2.20E−03	1.83	20.36	37	1058	metal ion binding
1376958_at	GO:0043167	2.20E−03	1.83	20.36	37	1058	ion binding
1376958_at	GO:0005543	5.60E−03	4.84	1.17	5	52	phospholipid binding
1376958_at	GO:0008395	5.76E−03	9.65	0.38	3	17	steroid hydroxylase activity
1376958_at	GO:0030234	7.32E−03	2.19	7.23	15	337	enzyme regulator activity
1376958_at	GO:0030300	2.58E−08	235.54	0.13	5	6	regulation of cholesterol absorption
1376958_at	GO:0006953	7.60E−08	27.89	0.42	7	19	acute-phase response
1376958_at	GO:0044241	2.32E−07	78.48	0.18	5	8	lipid digestion
1376958_at	GO:0030299	2.32E−07	78.48	0.18	5	8	cholesterol absorption
1376958_at	GO:0006956	4.77E−07	19.67	0.53	7	24	complement activation
1376958_at	GO:0006958	6.09E−07	28.45	0.35	6	16	complement activation, classical pathway
1376958_at	GO:0050892	1.01E−06	47.07	0.22	5	10	intestinal absorption
1376958_at	GO:0009613	2.71E−06	3.84	6.04	20	287	response to pest, pathogen or parasite
1376958_at	GO:0006066	2.90E−06	4.87	3.63	15	169	alcohol metabolism
1376958_at	GO:0006629	4.53E−06	3.28	8.38	24	406	lipid metabolism
1376958_at	GO:0051707	5.08E−06	3.67	6.28	20	299	response to other organism
1376958_at	GO:0006955	5.15E−06	3.25	8.43	24	409	immune response
1376958_at	GO:0008202	5.26E−06	5.78	2.47	12	114	steroid metabolism

1376958_at	GO:0006869	1.35E−05	10.76	0.83	7	38	lipid transport
1376958_at	GO:0006952	1.37E−05	3.04	8.89	24	433	defense response
1376958_at	GO:0045087	1.47E−05	14.20	0.57	6	26	innate immune response
1376958_at	GO:0009607	3.24E−05	2.87	9.33	24	456	response to biotic stimulus
1376958_at	GO:0006950	4.03E−05	2.51	13.55	31	687	response to stress
1376958_at	GO:0050817	6.59E−05	8.12	1.05	7	48	coagulation
1376958_at	GO:0007596	6.59E−05	8.12	1.05	7	48	blood coagulation
1376958_at	GO:0042060	7.28E−05	5.77	1.83	9	84	wound healing
1376958_at	GO:0006957	9.57E−05	69.50	0.11	3	5	complement activation, alternative pathway
1376958_at	GO:0016125	9.82E−05	7.56	1.12	7	51	sterol metabolism
1376958_at	GO:0007599	9.82E−05	7.56	1.12	7	51	hemostasis
1376958_at	GO:0050878	1.30E−04	6.07	1.55	8	71	regulation of body fluids
1376958_at	GO:0016064	1.42E−04	7.08	1.18	7	54	humoral defense mechanism (sensu Vertebrata)
1376958_at	GO:0006725	2.52E−04	5.45	1.70	8	78	aromatic compound metabolism
1376958_at	GO:0042157	2.87E−04	7.65	0.94	6	43	lipoprotein metabolism
1376958_at	GO:0030301	3.24E−04	34.74	0.15	3	7	cholesterol transport
1376958_at	GO:0015918	3.24E−04	34.74	0.15	3	7	sterol transport
1376958_at	GO:0006959	3.57E−04	5.16	1.79	8	82	humoral immune response
1376958_at	GO:0009611	4.30E−04	2.90	6.06	16	288	response to wounding
1376958_at	GO:0050874	4.48E−04	1.95	23.56	45	1311	organismal physiological process
1376958_at	GO:0019752	6.75E−04	2.68	6.89	17	330	carboxylic acid metabolism
1376958_at	GO:0006082	7.22E−04	2.67	6.93	17	332	organic acid metabolism
1376958_at	GO:0050766	1.06E−03	19.84	0.22	3	10	positive regulation of phagocytosis
1376958_at	GO:0044271	1.10E−03	5.77	1.20	6	55	nitrogen compound biosynthesis
1376958_at	GO:0009309	1.10E−03	5.77	1.20	6	55	amine biosynthesis
1376958_at	GO:0031099	1.34E−03	9.81	0.51	4	23	regeneration
1376958_at	GO:0042246	1.34E−03	9.81	0.51	4	23	tissue regeneration
1376958_at	GO:0050764	1.43E−03	17.36	0.24	3	11	regulation of phagocytosis
1376958_at	GO:0050727	1.58E−03	9.31	0.53	4	24	regulation of inflammatory response
1376958_at	GO:0019748	1.58E−03	9.31	0.53	4	24	secondary metabolism
1376958_at	GO:0045807	1.88E−03	15.43	0.26	3	12	positive regulation of endocytosis
1376958_at	GO:0009605	1.99E−03	2.41	7.56	17	364	response to external stimulus
1376958_at	GO:0019216	2.15E−03	8.46	0.57	4	26	regulation of lipid metabolism
1376958_at	GO:0006879	2.41E−03	13.88	0.29	3	13	iron ion homeostasis
1376958_at	GO:0044255	2.44E−03	2.42	7.07	16	339	cellular lipid metabolism
1376958_at	GO:0006519	2.80E−03	2.77	4.70	12	221	amino acid and derivative metabolism
1376958_at	GO:0008203	3.24E−03	5.57	1.03	5	47	cholesterol metabolism
1376958_at	GO:0008652	3.26E−03	7.44	0.64	4	29	amino acid biosynthesis
1376958_at	GO:0048589	3.70E−03	7.16	0.66	4	30	developmental growth
1376958_at	GO:0009056	4.50E−03	2.33	6.85	15	328	catabolism
1376958_at	GO:0006954	5.41E−03	2.99	3.27	9	152	inflammatory response
1376958_at	GO:0007586	5.46E−03	4.87	1.16	5	53	digestion
1376958_at	GO:0050896	5.55E−03	1.70	22.33	39	1228	response to stimulus
1376958_at	GO:0006694	6.40E−03	4.67	1.20	5	55	steroid biosynthesis
1376958_at	GO:0016042	6.40E−03	4.67	1.20	5	55	lipid catabolism
1376958_at	GO:0051241	6.50E−03	6.00	0.77	4	35	negative regulation of organismal physiological process
1376958_at	GO:0006909	7.43E−03	8.67	0.42	3	19	phagocytosis
1376958_at	GO:0051239	7.74E−03	2.51	4.68	11	220	regulation of organismal physiological process
1376958_at	GO:0009064	8.60E−03	8.15	0.44	3	20	glutamine family amino acid metabolism
1376958_at	GO:0008015	9.44E−03	3.20	2.39	7	110	circulation
1376958_at	GO:0046916	9.88E−03	7.70	0.46	3	21	transition metal ion homeostasis
1376958_at	GO:0046483	9.88E−03	4.17	1.33	5	61	heterocycle metabolism
1377006_at	GO:0007389	1.24E−03	17.26	0.22	3	86	pattern specification
1379497_at	GO:0015291	4.17E−08	21.01	0.54	8	116	porter activity
1379497_at	GO:0015290	4.46E−08	20.81	0.55	8	117	electrochemical potential-driven transporter activity
1379497_at	GO:0005386	8.06E−08	15.60	0.83	9	179	carrier activity
1379497_at	GO:0015293	6.00E−07	25.25	0.32	6	68	symporter activity
1379497_at	GO:0008509	2.18E−05	18.10	0.35	5	74	anion transporter activity
1379497_at	GO:0005215	2.21E−05	5.59	3.32	13	783	transporter activity
1379497_at	GO:0015075	1.18E−04	5.98	1.93	9	433	ion transporter activity
1379497_at	GO:0005342	1.37E−04	18.14	0.27	4	57	organic acid transporter activity
1379497_at	GO:0046943	1.37E−04	18.14	0.27	4	57	carboxylic acid transporter activity
1379497_at	GO:0031402	2.29E−04	15.74	0.31	4	65	sodium ion binding
1379497_at	GO:0015103	2.34E−04	30.30	0.12	3	26	inorganic anion transporter activity

1379497_at	GO:0015370	3.61E−04	25.79	0.14	3	30	solute:sodium symporter activity
1379497_at	GO:0031420	4.25E−04	9.29	0.65	5	138	alkali metal ion binding
1379497_at	GO:0015294	6.23E−04	21.08	0.17	3	36	solute:cation symporter activity
1379497_at	GO:0015171	7.32E−04	19.87	0.18	3	38	amino acid transporter activity
1379497_at	GO:0005275	1.73E−03	14.46	0.24	3	51	amine transporter activity
1379497_at	GO:0004252	4.46E−03	10.17	0.34	3	71	serine-type endopeptidase activity
1379497_at	GO:0008236	7.37E−03	8.42	0.40	3	85	serine-type peptidase activity
1379497_at	GO:0015849	9.86E−05	19.97	0.25	4	52	organic acid transport
1379497_at	GO:0046942	9.86E−05	19.97	0.25	4	52	carboxylic acid transport
1379497_at	GO:0006820	1.97E−04	16.50	0.30	4	62	anion transport
1379497_at	GO:0006814	2.09E−04	16.21	0.30	4	63	sodium ion transport
1379497_at	GO:0006811	3.61E−04	5.66	1.77	8	391	ion transport
1379497_at	GO:0006865	4.92E−04	23.08	0.16	3	33	amino acid transport
1379497_at	GO:0015698	1.08E−03	17.28	0.21	3	43	inorganic anion transport
1379497_at	GO:0015837	1.49E−03	15.34	0.23	3	48	amine transport
1379497_at	GO:0006810	6.22E−03	2.85	5.34	13	1365	transport
1379497_at	GO:0015672	9.53E−03	5.42	0.83	4	177	monovalent inorganic cation transport
1376968_at	GO:0008324	6.63E−03	6.67	0.76	4	348	cation transporter activity
1376968_at	GO:0006936	1.98E−06	35.71	0.23	5	105	muscle contraction
1376968_at	GO:0006816	4.18E−04	26.13	0.15	3	69	calcium ion transport
1376968_at	GO:0015674	1.17E−03	18.06	0.22	3	98	di-, tri-valent inorganic cation transport
1373254_at	GO:0005319	1.79E−13	77.97	0.21	9	38	lipid transporter activity
1373254_at	GO:0008289	9.96E−06	11.47	0.78	7	146	lipid binding
1373254_at	GO:0005215	2.31E−05	5.07	3.78	14	783	transporter activity
1373254_at	GO:0004252	3.43E−05	16.21	0.38	5	71	serine-type endopeptidase activity
1373254_at	GO:0008236	8.20E−05	13.35	0.46	5	85	serine-type peptidase activity
1373254_at	GO:0004175	2.98E−03	5.80	0.98	5	186	endopeptidase activity
1373254_at	GO:0004857	5.58E−03	6.29	0.71	4	134	enzyme inhibitor activity
1373254_at	GO:0016789	5.71E−03	9.22	0.37	3	68	carboxylic ester hydrolase activity
1373254_at	GO:0042803	6.19E−03	8.94	0.38	3	70	protein homodimerization activity
1373254_at	GO:0030234	8.52E−03	3.85	1.74	6	337	enzyme regulator activity
1373254_at	GO:0006958	3.24E−08	82.09	0.10	5	16	complement activation, classical pathway
1373254_at	GO:0006953	8.51E−08	64.46	0.12	5	19	acute-phase response
1373254_at	GO:0009613	9.93E−08	9.63	1.75	12	287	response to pest, pathogen or parasite
1373254_at	GO:0051707	1.56E−07	9.20	1.82	12	299	response to other organism
1373254_at	GO:0042157	2.21E−07	30.09	0.27	6	43	lipoprotein metabolism
1373254_at	GO:0006956	3.04E−07	47.46	0.15	5	24	complement activation
1373254_at	GO:0050817	4.34E−07	26.49	0.30	6	48	coagulation
1373254_at	GO:0007596	4.34E−07	26.49	0.30	6	48	blood coagulation
1373254_at	GO:0045087	4.66E−07	42.92	0.17	5	26	innate immune response
1373254_at	GO:0006955	6.25E−07	7.38	2.44	13	409	immune response
1373254_at	GO:0007599	6.28E−07	24.71	0.32	6	51	hemostasis
1373254_at	GO:0006952	1.20E−06	6.93	2.57	13	433	defense response
1373254_at	GO:0009607	2.17E−06	6.54	2.70	13	456	response to biotic stimulus
1373254_at	GO:0006957	2.34E−06	255.35	0.03	3	5	complement activation, alternative pathway
1373254_at	GO:0006869	3.36E−06	27.26	0.24	5	38	lipid transport
1373254_at	GO:0050878	4.56E−06	17.05	0.45	6	71	regulation of body fluids
1373254_at	GO:0030300	4.65E−06	170.21	0.04	3	6	regulation of cholesterol absorption
1373254_at	GO:0006950	7.82E−06	5.19	3.92	15	687	response to stress
1373254_at	GO:0009611	7.89E−06	7.34	1.75	10	288	response to wounding
1373254_at	GO:0042060	1.22E−05	14.17	0.53	6	84	wound healing
1373254_at	GO:0044241	1.29E−05	102.09	0.05	3	8	lipid digestion
1373254_at	GO:0030299	1.29E−05	102.09	0.05	3	8	cholesterol absorption
1373254_at	GO:0016064	1.97E−05	18.31	0.34	5	54	humoral defense mechanism (sensu Vertebrata)
1373254_at	GO:0050892	2.74E−05	72.89	0.06	3	10	intestinal absorption
1373254_at	GO:0009605	6.08E−05	5.69	2.19	10	364	response to external stimulus
1373254_at	GO:0050874	1.12E−04	3.67	6.83	19	1311	organismal physiological process
1373254_at	GO:0006959	1.49E−04	11.59	0.52	5	82	humoral immune response
1373254_at	GO:0008015	5.89E−04	8.46	0.69	5	110	circulation
1373254_at	GO:0050896	5.99E−04	3.21	6.47	17	1228	response to stimulus
1373254_at	GO:0050776	2.19E−03	8.23	0.55	4	88	regulation of immune response
1373254_at	GO:0051239	2.36E−03	5.04	1.35	6	220	regulation of organismal physiological process
1373254_at	GO:0051336	2.44E−03	12.68	0.27	3	43	regulation of hydrolase activity

1373254_at	GO:0006954	2.51E−03	6.00	0.95	5	152	inflammatory response
1373254_at	GO:0006629	3.32E−03	3.74	2.42	8	406	lipid metabolism
1373254_at	GO:0007586	4.44E−03	10.13	0.34	3	53	digestion
1373254_at	GO:0016042	4.93E−03	9.74	0.35	3	55	lipid catabolism
1373254_at	GO:0051242	8.69E−03	3.14	2.81	8	476	positive regulation of cellular physiological process
1371194_at	GO:0048513	2.46E−03	14.87	0.64	4	694	organ development
1371194_at	GO:0050874	2.80E−03	17.29	1.10	5	1311	organismal physiological process
1371194_at	GO:0007275	4.10E−03	15.55	1.18	5	1421	development
1373697_at	GO:0006936	1.45E−06	113.27	0.11	4	105	muscle contraction

Table A.2: Functional coherence of hubs and their subnetwork. Hub GO terms lists the GO terms the hub transcript is annotated with, enrichment GO terms lists the GO terms that are enriched in the subnetwork of the hub. MF, BP and CC denote the minimal distance of the hub GO terms to the enrichment GO terms in each of the ontologies (MF: molecular function, BP: biological process, CC: celular compartment).

Hub	Hub GO terms	Enrichment GO terms	MF	BP	CC
1376780_at	NA	GO:0004888, GO:0004871, GO:0004872, GO:0003690, GO:0001584, GO:0015291, GO:0015290, GO:0005543, GO:0004930, GO:0001614, GO:0001608, GO:0016502, GO:0045028, GO:0043566, GO:0008289, GO:0008406, GO:0045137, GO:0007154, GO:0007548, GO:0019226, GO:0007186, GO:0007267, GO:0009987, GO:0050877, GO:0007165, GO:0007166	NA	NA	NA
1389690_at	NA	GO:0003690, GO:0043566, GO:0015280, GO:0015276, GO:0008289, GO:0015268, GO:0015267, GO:0005216, GO:0046872, GO:0043167, GO:0007588, GO:0006607, GO:0006821, GO:0016079, GO:0048489, GO:0016338	NA	NA	NA
1374583_at	NA	GO:0006576	NA	NA	NA
1369836_at	GO:0005737	GO:0004888, GO:0008289, GO:0004872, GO:0001584, GO:0004930, GO:0030594, GO:0003690, GO:0042165, GO:0008528, GO:0001653, GO:0005543, GO:0001614, GO:0001608, GO:0016502, GO:0045028, GO:0015276, GO:0043566, GO:0042277, GO:0008227, GO:0004871, GO:0005230, GO:0007154, GO:0007186, GO:0007267, GO:0043085, GO:0048489, GO:0019226, GO:0016079, GO:0007165, GO:0019933, GO:0001505, GO:0050877, GO:0007269, GO:0007268, GO:0007166, GO:0045055, GO:0045761, GO:0048609, GO:0019935, GO:0007189, GO:0031279, GO:0051339, GO:0050876, GO:0007188, GO:0050874	Inf	Inf	Inf
1376840_at	NA	GO:0003677, GO:0003676, GO:0006139	NA	NA	NA
1371776_at	NA		NA	NA	NA
1384717_at	GO:0003677, GO:0003700, GO:0005515, GO:0005634, GO:0006355	GO:0003682, GO:0016563, GO:0030246, GO:0007167, GO:0006351	2	1	Inf
1370988_at	NA	GO:0003779, GO:0008092, GO:0006937, GO:0042060, GO:0050817, GO:0007596, GO:0007599, GO:0050878, GO:0007017, GO:0006936	NA	NA	NA
1371960_at	GO:0006959		Inf	Inf	Inf
1370114_a_at	GO:0005158, GO:0005159, GO:0005515, GO:0005942, GO:0005942, GO:0006468, GO:0008022, GO:0008286, GO:0016303, GO:0016303, GO:0030183, GO:0035014, GO:0035030, GO:0043066, GO:0043560, GO:0046854, GO:0048009	GO:0006607, GO:0006606, GO:0051170, GO:0017038, GO:0051169, GO:0006913, GO:0006605	Inf	7	Inf
1379896_at	GO:0003676, GO:0004386, GO:0005524, GO:0008026	GO:0008202, GO:0008610, GO:0006694, GO:0044255, GO:0006629, GO:0044249	Inf	Inf	Inf
1373232_at	NA	GO:0008289, GO:0007267, GO:0007268, GO:0019226, GO:0050877	NA	NA	NA

1368597_at	GO:0000122, GO:0000166, GO:0000287, GO:0004674, GO:0004674, GO:0005524, GO:0005634, GO:0005737, GO:0005829, GO:0006468, GO:0006468, GO:0007049, GO:0007243, GO:0007346, GO:0016564, GO:0016740, GO:0045595		Inf	Inf	Inf
1376958_at	NA	GO:0005319, GO:0004497, GO:0016712, GO:0050381, GO:0016705, GO:0016491, GO:0008289, GO:0017127, GO:0008236, GO:0015248, GO:0004866, GO:0004857, GO:0030414, GO:0004867, GO:0005507, GO:0016789, GO:0004252, GO:0046906, GO:0020037, GO:0003824, GO:0005215, GO:0046872, GO:0043167, GO:0005543, GO:0008395, GO:0030234, GO:0030300, GO:0006953, GO:0044241, GO:0030299, GO:0006956, GO:0006958, GO:0050892, GO:0009613, GO:0006066, GO:0006629, GO:0051707, GO:0006955, GO:0008202, GO:0006869, GO:0006952, GO:0045087, GO:0009607, GO:0006950, GO:0050817, GO:0007596, GO:0042060, GO:0006957, GO:0016125, GO:0007599, GO:0050878, GO:0016064, GO:0006725, GO:0042157, GO:0030301, GO:0015918, GO:0006959, GO:0009611, GO:0050874, GO:0019752, GO:0006082, GO:0050766, GO:0044271, GO:0009309, GO:0031099, GO:0042246, GO:0050764, GO:0050727, GO:0019748, GO:0045807, GO:0009605, GO:0019216, GO:0006879, GO:0044255, GO:0006519, GO:0008203, GO:0008652, GO:0048589, GO:0009056, GO:0006954, GO:0007586, GO:0050896, GO:0006694, GO:0016042, GO:0051241, GO:0006909, GO:0051239, GO:0009064, GO:0008015, GO:0046916, GO:0046483	NA	NA	NA
1374527_at	NA		NA	NA	NA
1377006_at	GO:0005737, GO:0006457, GO:0051082	GO:0007389	Inf	8	Inf
1377452_at	GO:0001501, GO:0005509, GO:0005529, GO:0005578, GO:0005615		Inf	Inf	Inf
1377329_at	GO:0003743, GO:0006446		Inf	Inf	Inf
1389990_at	NA		NA	NA	NA
1379497_at	NA	GO:0015291, GO:0015290, GO:0005386, GO:0015293, GO:0008509, GO:0005215, GO:0015075, GO:0005342, GO:0046943, GO:0031402, GO:0015103, GO:0015370, GO:0031420, GO:0015294, GO:0015171, GO:0005275, GO:0004252, GO:0008236, GO:0015849, GO:0046942, GO:0006820, GO:0006814, GO:0006811, GO:0006865, GO:0015698, GO:0015837, GO:0006810, GO:0015672	NA	NA	NA
1376968_at	GO:0005200, GO:0005856, GO:0006936, GO:0008307	GO:0008324, GO:0006936, GO:0006816, GO:0015674	5	0	Inf

1373254_at	GO:0005198, GO:0005882, GO:0008544, GO:0045095	GO:0005319, GO:0008289, GO:0005215, GO:0004252, GO:0008236, GO:0004175, GO:0004857, GO:0016789, GO:0042803, GO:0030234, GO:0006958, GO:0006953, GO:0009613, GO:0051707, GO:0042157, GO:0006956, GO:0050817, GO:0007596, GO:0045087, GO:0006955, GO:0007599, GO:0006952, GO:0009607, GO:0006957, GO:0006869, GO:0050878, GO:0030300, GO:0006950, GO:0009611, GO:0042060, GO:0044241, GO:0030299, GO:0016064, GO:0050892, GO:0009605, GO:0050874, GO:0006959, GO:0008015, GO:0050896, GO:0050776, GO:0051239, GO:0051336, GO:0006954, GO:0006629, GO:0007586, GO:0016042, GO:0051242	2	5	Inf
1371194_at	NA	GO:0048513, GO:0050874, GO:0007275	NA	NA	NA
1373697_at	GO:0005200, GO:0005856, GO:0006936, GO:0008307	GO:0006936	Inf	0	Inf

Table A.3: GO enrichment in all interface genes

GO term ID	p value	Odds ratio	Expected counts	Counts	Size	Term
GO:0008289	6.30E-07	4.6	4.66	18	146	lipid binding
GO:0005319	2.00E-06	9.9	1.24	9	38	lipid transporter activity
GO:0003690	4.10E-04	6	1.43	7	44	double-stranded DNA binding
GO:0008236	1.40E-03	3.7	2.74	9	85	serine-type peptidase activity
GO:0004252	1.70E-03	4	2.3	8	71	serine-type endopeptidase activity
GO:0005215	2.30E-03	1.8	22.62	39	783	transporter activity
GO:0001614	3.50E-03	5.6	1.07	5	33	purinergic nucleotide receptor activity
GO:0001608	3.50E-03	5.6	1.07	5	33	nucleotide receptor activity, G-protein coupled
GO:0016502	3.50E-03	5.6	1.07	5	33	nucleotide receptor activity
GO:0045028	3.50E-03	5.6	1.07	5	33	purinergic nucleotide receptor activity, G-protein coupled
GO:0043566	4.60E-03	3.7	2.14	7	66	structure-specific DNA binding
GO:0004175	5.90E-03	2.4	5.9	13	186	endopeptidase activity
GO:0008034	7.10E-03	9.3	0.42	3	13	lipoprotein binding
GO:0004888	9.30E-03	1.8	14.58	25	482	transmembrane receptor activity
GO:0004857	9.80E-03	2.5	4.29	10	134	enzyme inhibitor activity
GO:0030234	1.00E-02	1.9	10.43	19	337	enzyme regulator activity
GO:0007599	2.70E-05	6.8	1.67	9	51	hemostasis
GO:0050878	7.30E-05	5.2	2.31	10	71	regulation of body fluids
GO:0006958	1.00E-04	14.1	0.53	5	16	complement activation, classical pathway
GO:0050817	1.20E-04	6.3	1.57	8	48	coagulation
GO:0007596	1.20E-04	6.3	1.57	8	48	blood coagulation
GO:0006953	2.50E-04	11.1	0.62	5	19	acute-phase response
GO:0006957	3.10E-04	46.2	0.16	3	5	complement activation, alternative pathway
GO:0042060	3.10E-04	4.3	2.73	10	84	wound healing
GO:0030300	6.00E-04	30.8	0.2	3	6	regulation of cholesterol absorption
GO:0006956	8.10E-04	8.2	0.79	5	24	complement activation
GO:0006607	1.00E-03	23.1	0.23	3	7	NLS-bearing substrate import into nucleus
GO:0045087	1.20E-03	7.4	0.85	5	26	innate immune response
GO:0044241	1.60E-03	18.5	0.26	3	8	lipid digestion
GO:0030299	1.60E-03	18.5	0.26	3	8	cholesterol absorption
GO:0042157	2.20E-03	5	1.41	6	43	lipoprotein metabolism
GO:0008015	2.50E-03	3.1	3.56	10	110	circulation
GO:0008217	2.80E-03	4.8	1.47	6	45	blood pressure regulation
GO:0050892	3.30E-03	13.2	0.33	3	10	intestinal absorption
GO:0050874	3.40E-03	1.6	35.08	58	1311	organismal physiological process
GO:0006821	4.70E-03	6.9	0.72	4	22	chloride transport
GO:0006869	6.70E-03	4.7	1.24	5	38	lipid transport

Table A.4: GO enrichment in interface genes occuring in at least three interfaces

GO term ID	P value	Odds ratio	Expected count	Count	Size	Term
GO:0003690	7.40E-04	7.9	0.74	5	44	double-stranded DNA binding
GO:0008289	2.77E-03	3.6	2.4	8	146	lipid binding
GO:0043566	4.55E-03	5.1	1.1	5	66	structure-specific DNA binding
GO:0030594	6.22E-03	4.7	1.18	5	71	neurotransmitter receptor activity
GO:0000149	6.95E-03	8.7	0.4	3	24	SNARE binding
GO:0042165	8.73E-03	4.3	1.28	5	77	neurotransmitter binding
GO:0007267	1.36E-03	2.5	7.2	17	481	cell-cell signaling
GO:0007154	1.74E-03	1.9	21.4	41	1714	cell communication
GO:0007165	4.81E-03	1.8	18.92	35	1464	signal transduction
GO:0006821	4.87E-03	10	0.36	3	22	chloride transport
GO:0016079	5.53E-03	9.5	0.37	3	23	synaptic vesicle exocytosis
GO:0007167	6.99E-03	3.1	2.78	8	177	enzyme linked receptor protein signaling pathway
GO:0019226	9.25E-03	2.5	4.73	11	307	transmission of nerve impulse
GO:0030518	9.66E-03	7.6	0.45	3	28	steroid hormone receptor signaling pathway

Table A.5: GO enrichment in interfaces

hub1	hub2	ID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1376780_at	1389690_at	GO:0003690	1.03E−03	10.15	0.46	4	44	double-stranded DNA binding
1376780_at	1389690_at	GO:0005543	1.94E−03	8.45	0.54	4	52	phospholipid binding
1376780_at	1389690_at	GO:0004871	3.40E−03	2.20	10.49	22	1205	signal transducer activity
1376780_at	1389690_at	GO:0008289	3.78E−03	4.43	1.49	6	146	lipid binding
1376780_at	1389690_at	GO:0001584	3.97E−03	3.84	1.99	7	197	rhodopsin-like receptor activity
1376780_at	1389690_at	GO:0043566	4.64E−03	6.52	0.68	4	66	structure-specific DNA binding
1376780_at	1389690_at	GO:0030594	6.01E−03	6.03	0.73	4	71	neurotransmitter receptor activity
1376780_at	1389690_at	GO:0004930	6.25E−03	3.20	2.70	8	270	G-protein coupled receptor activity
1376780_at	1389690_at	GO:0004872	6.35E−03	2.38	6.21	14	659	receptor activity
1376780_at	1389690_at	GO:0008227	7.39E−03	8.32	0.40	3	39	amine receptor activity
1376780_at	1389690_at	GO:0004222	7.93E−03	8.10	0.42	3	40	metalloendopeptidase activity
1376780_at	1389690_at	GO:0042165	8.00E−03	5.53	0.79	4	77	neurotransmitter binding
1376780_at	1389690_at	GO:0008237	8.37E−03	5.46	0.80	4	78	metallopeptidase activity
1376780_at	1389690_at	GO:0004888	9.62E−03	2.50	4.66	11	482	transmembrane receptor activity
1376780_at	1389690_at	GO:0008406	4.11E−04	13.28	0.36	4	36	gonad development
1376780_at	1389690_at	GO:0045137	4.11E−04	13.28	0.36	4	36	development of primary sexual characteristics
1376780_at	1389690_at	GO:0007548	1.34E−03	9.42	0.49	4	49	sex differentiation
1376780_at	1389690_at	GO:0007267	2.32E−03	2.95	4.46	12	481	cell-cell signaling
1376780_at	1389690_at	GO:0007186	2.70E−03	3.21	3.42	10	362	G-protein coupled receptor protein signaling pathway
1376780_at	1389690_at	GO:0019226	3.02E−03	3.37	2.93	9	307	transmission of nerve impulse
1376780_at	1389690_at	GO:0007154	4.12E−03	2.11	13.29	27	1714	cell communication
1376780_at	1374583_at	GO:0019226	2.48E−03	6.51	0.95	5	307	transmission of nerve impulse
1376780_at	1374583_at	GO:0050877	4.62E−03	4.80	1.55	6	515	neurophysiological process
1376780_at	1369836_at	GO:0004888	3.63E−05	4.68	3.40	13	482	transmembrane receptor activity
1376780_at	1369836_at	GO:0004872	5.72E−05	4.06	4.52	15	659	receptor activity
1376780_at	1369836_at	GO:0001584	9.41E−05	6.52	1.45	8	197	rhodopsin-like receptor activity
1376780_at	1369836_at	GO:0004930	1.50E−04	5.39	1.97	9	270	G-protein coupled receptor activity
1376780_at	1369836_at	GO:0005543	5.91E−04	11.90	0.39	4	52	phospholipid binding
1376780_at	1369836_at	GO:0004871	6.44E−04	2.88	7.64	19	1205	signal transducer activity
1376780_at	1369836_at	GO:0001614	1.88E−03	13.99	0.25	3	33	purinergic nucleotide receptor activity
1376780_at	1369836_at	GO:0001608	1.88E−03	13.99	0.25	3	33	nucleotide receptor activity, G-protein coupled
1376780_at	1369836_at	GO:0016502	1.88E−03	13.99	0.25	3	33	nucleotide receptor activity
1376780_at	1369836_at	GO:0045028	1.88E−03	13.99	0.25	3	33	purinergic nucleotide receptor activity, G-protein coupled
1376780_at	1369836_at	GO:0030594	1.91E−03	8.50	0.53	4	71	neurotransmitter receptor activity
1376780_at	1369836_at	GO:0042165	2.58E−03	7.79	0.58	4	77	neurotransmitter binding
1376780_at	1369836_at	GO:0008227	3.04E−03	11.65	0.29	3	39	amine receptor activity
1376780_at	1369836_at	GO:0008528	3.38E−03	7.19	0.62	4	83	peptide receptor activity, G-protein coupled
1376780_at	1369836_at	GO:0001653	3.53E−03	7.10	0.63	4	84	peptide receptor activity
1376780_at	1369836_at	GO:0008289	4.59E−03	5.11	1.08	5	146	lipid binding
1376780_at	1369836_at	GO:0007186	3.40E−05	5.49	2.47	11	362	G-protein coupled receptor protein signaling pathway
1376780_at	1369836_at	GO:0007267	4.37E−04	4.01	3.22	11	481	cell-cell signaling
1376780_at	1369836_at	GO:0007154	4.99E−04	2.93	9.60	23	1714	cell communication
1376780_at	1369836_at	GO:0007166	1.14E−03	3.22	4.64	13	719	cell surface receptor linked signal transduction
1376780_at	1369836_at	GO:0019226	1.33E−03	4.32	2.12	8	307	transmission of nerve impulse
1376780_at	1369836_at	GO:0050877	2.93E−03	3.27	3.43	10	515	neurophysiological process
1376780_at	1369836_at	GO:0019933	3.51E−03	7.14	0.63	4	88	cAMP-mediated signaling
1376780_at	1369836_at	GO:0007165	3.52E−03	2.48	8.48	19	1464	signal transduction
1376780_at	1369836_at	GO:0007268	3.91E−03	3.95	1.98	7	286	synaptic transmission
1376780_at	1369836_at	GO:0043085	6.16E−03	6.05	0.73	4	103	positive regulation of enzyme activity
1376780_at	1369836_at	GO:0050874	6.66E−03	2.36	7.76	17	1311	organismal physiological process
1376780_at	1369836_at	GO:0007269	8.16E−03	8.01	0.42	3	58	neurotransmitter secretion
1376780_at	1369836_at	GO:0019935	8.25E−03	5.53	0.80	4	112	cyclic-nucleotide-mediated signaling
1376780_at	1373232_at	GO:0015291	6.97E−03	8.79	0.39	3	116	porter activity
1376780_at	1373232_at	GO:0015290	7.13E−03	8.71	0.40	3	117	electrochemical potential-driven transporter activity

1376780_at	1373232_at	GO:0007267	3.28E−03	5.18	1.45	6	481	cell-cell signaling
1376780_at	1376958_at	GO:0004930	5.68E−03	6.72	0.72	4	270	G-protein coupled receptor activity
1376780_at	1376958_at	GO:0004872	6.80E−03	4.53	1.66	6	659	receptor activity
1376780_at	1376958_at	GO:0008289	7.17E−03	8.89	0.40	3	146	lipid binding
1376780_at	1376958_at	GO:0004888	8.45E−03	4.89	1.25	5	482	transmembrane receptor activity
1376780_at	1376958_at	GO:0019933	1.33E−03	16.84	0.22	3	88	cAMP-mediated signaling
1376780_at	1376958_at	GO:0007010	2.04E−03	9.29	0.55	4	223	cytoskeleton organization and biogenesis
1376780_at	1376958_at	GO:0019935	2.66E−03	13.08	0.28	3	112	cyclic-nucleotide-mediated signaling
1376780_at	1376958_at	GO:0007165	4.30E−03	4.49	3.02	9	1464	signal transduction
1389690_at	1374583_at	GO:0006576	3.03E−03	11.72	0.29	3	52	biogenic amine metabolism
1389690_at	1374583_at	GO:0050877	6.72E−03	3.34	2.69	8	515	neurophysiological process
1389690_at	1374583_at	GO:0019226	6.75E−03	4.06	1.66	6	307	transmission of nerve impulse
1389690_at	1374583_at	GO:0006575	9.77E−03	7.52	0.44	3	79	amino acid derivative metabolism
1389690_at	1369836_at	GO:0008289	1.43E−04	5.99	1.54	8	146	lipid binding
1389690_at	1369836_at	GO:0004888	4.41E−04	3.26	4.81	14	482	transmembrane receptor activity
1389690_at	1369836_at	GO:0030594	8.96E−04	7.53	0.76	5	71	neurotransmitter receptor activity
1389690_at	1369836_at	GO:0001584	1.08E−03	4.34	2.06	8	197	rhodopsin-like receptor activity
1389690_at	1369836_at	GO:0003690	1.16E−03	9.81	0.47	4	44	double-stranded DNA binding
1389690_at	1369836_at	GO:0004872	1.18E−03	2.73	6.41	16	659	receptor activity
1389690_at	1369836_at	GO:0042165	1.29E−03	6.89	0.82	5	77	neurotransmitter binding
1389690_at	1369836_at	GO:0004930	2.08E−03	3.55	2.79	9	270	G-protein coupled receptor activity
1389690_at	1369836_at	GO:0005543	2.18E−03	8.17	0.56	4	52	phospholipid binding
1389690_at	1369836_at	GO:0015276	5.18E−03	6.31	0.70	4	66	ligand-gated ion channel activity
1389690_at	1369836_at	GO:0043566	5.18E−03	6.31	0.70	4	66	structure-specific DNA binding
1389690_at	1369836_at	GO:0008227	8.06E−03	8.05	0.42	3	39	amine receptor activity
1389690_at	1369836_at	GO:0005230	8.65E−03	7.83	0.43	3	40	extracellular ligand-gated ion channel activity
1389690_at	1369836_at	GO:0007154	3.38E−04	2.51	13.98	31	1714	cell communication
1389690_at	1369836_at	GO:0007267	3.38E−04	3.38	4.70	14	481	cell-cell signaling
1389690_at	1369836_at	GO:0048489	8.24E−04	10.87	0.43	4	41	synaptic vesicle transport
1389690_at	1369836_at	GO:0007186	1.14E−03	3.39	3.60	11	362	G-protein coupled receptor protein signaling pathway
1389690_at	1369836_at	GO:0019226	1.17E−03	3.61	3.08	10	307	transmission of nerve impulse
1389690_at	1369836_at	GO:0016079	1.67E−03	14.86	0.24	3	23	synaptic vesicle exocytosis
1389690_at	1369836_at	GO:0019933	2.14E−03	6.11	0.92	5	88	cAMP-mediated signaling
1389690_at	1369836_at	GO:0001505	2.37E−03	5.97	0.94	5	90	regulation of neurotransmitter levels
1389690_at	1369836_at	GO:0007268	2.66E−03	3.43	2.88	9	286	synaptic transmission
1389690_at	1369836_at	GO:0007269	3.02E−03	7.42	0.61	4	58	neurotransmitter secretion
1389690_at	1369836_at	GO:0045055	4.08E−03	6.79	0.66	4	63	regulated secretory pathway
1389690_at	1369836_at	GO:0043085	4.25E−03	5.16	1.07	5	103	positive regulation of enzyme activity
1389690_at	1369836_at	GO:0007165	4.56E−03	2.09	12.35	25	1464	signal transduction
1389690_at	1369836_at	GO:0019935	6.06E−03	4.72	1.16	5	112	cyclic-nucleotide-mediated signaling
1389690_at	1369836_at	GO:0045761	6.10E−03	8.99	0.38	3	36	regulation of adenylate cyclase activity
1389690_at	1369836_at	GO:0050877	6.25E−03	2.56	5.00	12	515	neurophysiological process
1389690_at	1369836_at	GO:0007189	6.59E−03	8.72	0.39	3	37	G-protein signaling, adenylate cyclase activating pathway
1389690_at	1369836_at	GO:0031279	6.59E−03	8.72	0.39	3	37	regulation of cyclase activity
1389690_at	1369836_at	GO:0051339	6.59E−03	8.72	0.39	3	37	regulation of lyase activity
1389690_at	1369836_at	GO:0007188	7.59E−03	5.63	0.78	4	75	G-protein signaling, coupled to cAMP nucleotide second messenger
1389690_at	1369836_at	GO:0006887	9.10E−03	5.32	0.82	4	79	exocytosis
1389690_at	1376840_at	GO:0003677	8.29E−04	8.28	1.18	6	661	DNA binding
1389690_at	1376840_at	GO:0003676	9.25E−04	7.49	1.64	7	964	nucleic acid binding
1389690_at	1384717_at	GO:0048468	5.85E−03	6.60	0.73	4	245	cell development
1389690_at	1373232_at	GO:0007242	3.37E−03	5.52	1.47	6	632	intracellular signaling cascade
1389690_at	1373232_at	GO:0007267	5.61E−03	5.61	1.15	5	481	cell-cell signaling
1389690_at	1376958_at	GO:0004871	5.57E−03	3.40	3.64	10	1205	signal transducer activity
1389690_at	1376958_at	GO:0004888	5.93E−03	4.41	1.62	6	482	transmembrane receptor activity
1389690_at	1376958_at	GO:0004872	6.68E−03	3.88	2.15	7	659	receptor activity

1389690_at	1376958_at	GO:0007165	9.61E−04	4.50	4.03	12	1464	signal transduction
1389690_at	1376958_at	GO:0019933	3.15E−03	11.88	0.30	3	88	cAMP-mediated signaling
1389690_at	1376958_at	GO:0001505	3.36E−03	11.60	0.31	3	90	regulation of neurotransmitter levels
1389690_at	1376958_at	GO:0007154	4.24E−03	3.62	4.56	12	1714	cell communication
1389690_at	1376958_at	GO:0043085	4.91E−03	10.07	0.35	3	103	positive regulation of enzyme activity
1389690_at	1376958_at	GO:0019935	6.21E−03	9.23	0.38	3	112	cyclic-nucleotide-mediated signaling
1389690_at	1376958_at	GO:0007010	6.24E−03	6.38	0.74	4	223	cytoskeleton organization and biogenesis
1389690_at	1376958_at	GO:0008284	9.57E−03	7.83	0.44	3	131	positive regulation of cell proliferation
1374583_at	1369836_at	GO:0019226	1.40E−04	15.22	0.55	5	307	transmission of nerve impulse
1374583_at	1369836_at	GO:0050877	1.45E−04	12.51	0.89	6	515	neurophysiological process
1374583_at	1369836_at	GO:0007267	1.14E−03	9.35	0.84	5	481	cell-cell signaling
1374583_at	1369836_at	GO:0007268	1.43E−03	11.21	0.52	4	286	synaptic transmission
1369836_at	1373232_at	GO:0007267	1.70E−04	18.71	0.61	5	481	cell-cell signaling
1369836_at	1373232_at	GO:0007268	5.45E−03	11.74	0.37	3	286	synaptic transmission
1369836_at	1373232_at	GO:0019226	6.65E−03	10.88	0.40	3	307	transmission of nerve impulse
1369836_at	1376958_at	GO:0004888	6.41E−04	7.84	1.10	6	482	transmembrane receptor activity
1369836_at	1376958_at	GO:0004872	3.32E−03	5.54	1.47	6	659	receptor activity
1369836_at	1376958_at	GO:0004930	3.50E−03	7.94	0.64	4	270	G-protein coupled receptor activity
1369836_at	1376958_at	GO:0019933	8.53E−04	20.22	0.19	3	88	cAMP-mediated signaling
1369836_at	1376958_at	GO:0001505	9.11E−04	19.75	0.20	3	90	regulation of neurotransmitter levels
1369836_at	1376958_at	GO:0043085	1.35E−03	17.14	0.23	3	103	positive regulation of enzyme activity
1369836_at	1376958_at	GO:0019935	1.72E−03	15.70	0.25	3	112	cyclic-nucleotide-mediated signaling
1369836_at	1376958_at	GO:0007165	5.79E−03	4.79	2.62	8	1464	signal transduction
1369836_at	1376958_at	GO:0007186	6.69E−03	6.77	0.76	4	362	G-protein coupled receptor protein signaling pathway
1369836_at	1376958_at	GO:0050790	7.01E−03	9.34	0.40	3	184	regulation of enzyme activity
1369836_at	1376958_at	GO:0045045	7.12E−03	9.28	0.40	3	185	secretory pathway
1369836_at	1376958_at	GO:0019932	8.35E−03	8.74	0.42	3	196	second-messenger-mediated signaling
1376840_at	1384717_at	GO:0003677	4.68E−03	24.76	0.39	3	661	DNA binding
1384717_at	1377006_at	GO:0007389	2.45E−04	34.55	0.13	3	86	pattern specification
1373232_at	1376958_at	GO:0007010	5.53E−05	Inf	0.11	3	223	cytoskeleton organization and biogenesis
1373232_at	1376958_at	GO:0006996	4.13E−04	Inf	0.21	3	435	organelle organization and biogenesis
1373232_at	1376958_at	GO:0007242	1.27E−03	Inf	0.29	3	632	intracellular signaling cascade
1373232_at	1376958_at	GO:0016043	4.85E−03	Inf	0.43	3	987	cell organization and biogenesis
1376958_at	1373254_at	GO:0005319	9.44E−14	85.09	0.19	9	38	lipid transporter activity
1376958_at	1373254_at	GO:0008289	6.38E−06	12.43	0.73	7	146	lipid binding
1376958_at	1373254_at	GO:0004252	2.50E−05	17.47	0.36	5	71	serine-type endopeptidase activity
1376958_at	1373254_at	GO:0005215	5.26E−05	4.96	3.55	13	783	transporter activity
1376958_at	1373254_at	GO:0008236	6.00E−05	14.38	0.43	5	85	serine-type peptidase activity
1376958_at	1373254_at	GO:0004175	2.24E−03	6.25	0.92	5	186	endopeptidase activity
1376958_at	1373254_at	GO:0004857	4.44E−03	6.76	0.67	4	134	enzyme inhibitor activity
1376958_at	1373254_at	GO:0016789	4.78E−03	9.88	0.34	3	68	carboxylic ester hydrolase activity
1376958_at	1373254_at	GO:0042803	5.19E−03	9.59	0.35	3	70	protein homodimerization activity
1376958_at	1373254_at	GO:0030234	6.22E−03	4.15	1.63	6	337	enzyme regulator activity
1376958_at	1373254_at	GO:0006958	2.42E−08	87.59	0.10	5	16	complement activation, classical pathway
1376958_at	1373254_at	GO:0009613	4.89E−08	10.47	1.65	12	287	response to pest, pathogen or parasite
1376958_at	1373254_at	GO:0006953	6.36E−08	68.79	0.11	5	19	acute-phase response
1376958_at	1373254_at	GO:0051707	7.71E−08	10.01	1.72	12	299	response to other organism
1376958_at	1373254_at	GO:0042157	1.56E−07	32.18	0.26	6	43	lipoprotein metabolism
1376958_at	1373254_at	GO:0006956	2.28E−07	50.64	0.14	5	24	complement activation
1376958_at	1373254_at	GO:0006955	2.95E−07	8.05	2.31	13	409	immune response
1376958_at	1373254_at	GO:0050817	3.07E−07	28.33	0.29	6	48	coagulation
1376958_at	1373254_at	GO:0007596	3.07E−07	28.33	0.29	6	48	blood coagulation
1376958_at	1373254_at	GO:0045087	3.49E−07	45.80	0.16	5	26	innate immune response
1376958_at	1373254_at	GO:0007599	4.45E−07	26.42	0.31	6	51	hemostasis
1376958_at	1373254_at	GO:0006952	5.73E−07	7.56	2.43	13	433	defense response
1376958_at	1373254_at	GO:0009607	1.04E−06	7.13	2.55	13	456	response to biotic stimulus
1376958_at	1373254_at	GO:0006957	1.97E−06	271.41	0.03	3	5	complement activation, alternative pathway
1376958_at	1373254_at	GO:0006869	2.53E−06	29.09	0.23	5	38	lipid transport

1376958_at	1373254_at	GO:0050878	3.24E−06	18.23	0.42	6	71	regulation of body fluids
1376958_at	1373254_at	GO:0006950	3.41E−06	5.71	3.71	15	687	response to stress
1376958_at	1373254_at	GO:0030300	3.92E−06	180.91	0.04	3	6	regulation of cholesterol absorption
1376958_at	1373254_at	GO:0009611	4.54E−06	7.93	1.66	10	288	response to wounding
1376958_at	1373254_at	GO:0042060	8.71E−06	15.16	0.50	6	84	wound healing
1376958_at	1373254_at	GO:0044241	1.09E−05	108.51	0.05	3	8	lipid digestion
1376958_at	1373254_at	GO:0030299	1.09E−05	108.51	0.05	3	8	cholesterol absorption
1376958_at	1373254_at	GO:0016064	1.49E−05	19.53	0.32	5	54	humoral defense mechanism (sensu Vertebrata)
1376958_at	1373254_at	GO:0050892	2.32E−05	77.48	0.06	3	10	intestinal absorption
1376958_at	1373254_at	GO:0009605	3.58E−05	6.14	2.07	10	364	response to external stimulus
1376958_at	1373254_at	GO:0050874	4.11E−05	4.14	6.46	19	1311	organismal physiological process
1376958_at	1373254_at	GO:0006959	1.14E−04	12.37	0.49	5	82	humoral immune response
1376958_at	1373254_at	GO:0050896	2.63E−04	3.57	6.12	17	1228	response to stimulus
1376958_at	1373254_at	GO:0008015	4.52E−04	9.03	0.65	5	110	circulation
1376958_at	1373254_at	GO:0051239	1.75E−03	5.39	1.28	6	220	regulation of organismal physiological process
1376958_at	1373254_at	GO:0050776	1.78E−03	8.77	0.52	4	88	regulation of immune response
1376958_at	1373254_at	GO:0006954	1.95E−03	6.40	0.90	5	152	inflammatory response
1376958_at	1373254_at	GO:0051336	2.08E−03	13.48	0.26	3	43	regulation of hydrolase activity
1376958_at	1373254_at	GO:0006629	2.29E−03	4.02	2.29	8	406	lipid metabolism
1376958_at	1373254_at	GO:0007586	3.79E−03	10.77	0.32	3	53	digestion
1376958_at	1373254_at	GO:0016042	4.21E−03	10.35	0.33	3	55	lipid catabolism
1376958_at	1373254_at	GO:0051242	6.12E−03	3.37	2.66	8	476	positive regulation of cellular physiological process
1376958_at	1373254_at	GO:0043119	8.29E−03	3.18	2.79	8	501	positive regulation of physiological process

B Zusammenfassung

Aktuelle genomweite Assoziationsstudien (GWAS) und Analysen quantitativer Trait Loci (QTL) resultieren typischerweise in grossen chromosomalen Regionen, die durch (Krankheits-) assoziierte polymorphe Marker repräsentiert werden. Diese Marker sind zumeist nicht-funktionell, aber befinden sich in Kopplungsungleichgewicht mit unbekannten funktionellen Varianten. Angesichts der Tatsache, dass der Großteil des Genoms aus nicht-kodierenden Sequenzen besteht, ist es nicht überraschend, dass sich die meisten Varianten in diesen Regionen befinden. Obwohl nicht-kodierenden Sequenzen nicht vollständig verstanden sind ist bekannt, dass sie eine Vielzahl regulatorischer Elemente enthalten. Daher können Varianten in diesen Regionen regulatorische Elemente und damit das regulatorische Netzwerk beeinflussen. Deshalb ist es wahrscheinlich, dass Varianten, die auch die Expression eines Genes beeinflussen, regulatorische Elemente betreffen. Dies macht die Analyse der Genetik der Genexpression zu einem exzellentes Mittel um regulatorische Varianten zu identifizieren. Regulatorische Effekte von Sequenzvarianten können systematisch auf der Ebene der Genexpression gemessen werden indem Transkriptniveaus als quantitative Traits betrachtet und als Expressions-QTLs (eQTLs) kartiert werden. eQTL Studien stellen somit einen ersten Versuch dar, eine Verbindung zwischen genetischer Variation und molekularer Funktion herzustellen. Allerdings waren die meisten eQTL Studien bisher nicht in der Lage Hypothesen über den molekularen Mechanismus, der den eQTLs zugrunde liegt, zu generieren.

Insgesamt beschreibt diese Arbeit ein umfangreiches Gruppe von Werkzeugen und Strategien für die Analyse von regulatorischer genetischer Variation. Der Ausgangspunkt ist die schon früher beschrieben Identifikation von Zielgenen der potentiellen regulatorischen Variation als eQTL Transkripte. Darüber hinaus werden Ansätze beschrieben, um folgende Fragen bezüglich dieser eQTL Transkripte zu beantworten. (1) Welche Rolle spielt das eQTL Transkript für ein Krankheits-Modell, (2) welches *cis*-regulatorische Element ist von der Sequenzvariante betroffen und welcher Transkriptionsfaktor ist der Regulator des eQTL Transkripts, (3) welches sind die *trans*-regulatorischen Faktoren und wie werden ihrer Effekte zu ihren Zielgenen übertragen, und (4) welches ist der funktionale Zusammenhang in dem das eQTL Transkript agiert? Darüber hinaus werden Genexpressions-Netzwerke die im Rattenmodell erstellt wurden dazu verwendet, Ergebnisse von GWAS Studien am Menschen mit molekularen Funktionen in Verbindung zu bringen und damit die Genetik polygener Erkrankungen zu interpretieren.

Zuerst werden bestehende Ansätze zur Identifizierung von Krankheitsgenen diskutiert und deren Anwendung in Fallstudien demonstriert, die zur Identifizierung eines Kandidatengens für Herzinsuffizienz und eines für systolischen Blutdruck führten.

Als nächstes zeigen wir, dass funktionelle Annotation benutzt werden kann um genetische Marker zu identifizieren, die die Expression ganzer Netzwerke von Genen ähnlicher Funktion beeinflussen. Dafür betrachten wir wohlbekannte funktionell annotierte Sets von Genen. Darauf aufbauend definieren wir ein statistisches Mass für die Assoziation eines Gen-Sets zu einem

Marker und evaluieren diese für alle Paare von Sets und genetischen Markern. Dadurch können Gen-Sets, die unter der genetischer Kontrolle eines Markers stehen, identifiziert werden. Diese Methode wurde auf einen Datensatz aus einem F2 Intercross angewendet, wobei sich unsere Analyse auf den Arachidonsäure Stoffwechsel (*Ephx2*-Pathway) aus der Kyoto Encyclopaedia of Genes and Genomes (KEGG) konzentrierte. Wir konnten zeigen, dass die maximale Assoziation des *Ephx2*-Pathway am *Ephx2* locus auftritt. Dies zeigt, dass erhöhte *Ephx2* Expression, die zu einer Verringerung von kardioprotektiven Epoxiden führt, eine Rückkopplungsregulation auslöst, die die Expression des *Ephx2*-pathway erhöht um die Epoxid Verringerung zu kompensieren.

Wir demonstrieren wie Informationen über Sequenzvariationen und ein biophysikalisches Modell der TF – DNA Interaktion dazu verwendet werden kann sowohl das wahrscheinlichste *cis*-regulatorische Element im Promoter des eQTL Transkripts, als auch den wahrscheinlichsten TF zu bestimmen, der als Regulator des eQTL Transkripts in Frage kommt. Dazu entwenden wir eine Methode (sTRAP), um mögliche Konsequenzen von Sequenzvariationen auf das regulatorische Netzwerk zu bewerten. Für alle TFs mit bekanntem Bindemodell sagen wir quantitative Änderungen der Bindungsstärke voraus. Bekannten Assoziationen zwischen SNPs und deren regulatorischen Auswirkungen dienten dabei der Evaluation. Unsere Vorhersagen sind robust in Bezug auf verschieden Parameter und Modelannahmen. Angesichts der guten Leistung unserer Methode, haben wir ein Webseite veröffentlich, die als Startpunkt für Routineanalysen von Krankheits-assoziierten Sequenz-Regionen dienen kann.

Darüber hinaus analysieren wir die Rolle von TFs als Mediatoren von *trans*-acting eQTLs und identifizieren dadurch Gen-Netzwerke und deren regulatorische Loci. Wir zeigen, wie diese Gen-Netzwerke dazu verwendet werden können, Hypothesen über Funktion von Krankheits-assoziierten Regionen zu liefern, die über die Ergebnisse von typischen humanen GWAS hinausgehen. Eine integrierte Analyse von Expressionsdaten und TF Bindestellen wurde verwendet um das Interferon regulatory factor 7 (*Irf7*) - *driven inflammatory network* (iDIN) zu definieren. Dieses war statistisch mit Genen der Kategorie "Antwort auf Virus" angereichert und stellt einen molekularen Biomarker für Makrophagen dar. Das iDIN wird in mehreren Geweben von einem Locus auf Chromosom 15q25 reguliert. Die Analyse von Expressionsdaten priorisierte das Epstein-Barr Virus induzierte Gen 2 (*Ebi2* oder *Gpr183*) als Kandidaten für die Regulation des iDIN, was wir experimentell bestätigen konnten. *Ebi2* liegt im Chromosom 15q25 Locus und es ist bekannt, dass es die Migration von B-Lymphozyten kontrolliert und in Makrophagen exprimiert ist. Der orthologe Locus in Menschen liegt auf Chromosom 13q32 und kontrolliert das humane Äquivalent des iDIN, das aus Expressionsdaten von Monozyten identifiziert wurde. Mit Hilfe einer Anreicherungs-Analyse konnten wir zeigen, dass iDIN Gene mit grösserer Wahrscheinlichkeit mit Typ 1 Diabetes (T1D) – einer Autoimmunerkrankung die mit Makrophagen in Verbindung steht – assoziiert sind als zufällig gewählte Gene der Immunãntw̃ort. Auch der humane Locus, der die Expression des iDIN kontrolliert, ist mit Risiko für T1D am SNP rs9585056 assoziiert. Dieser SNP ist auch einer der fünf SNPs in der Region die mit *EBI2* Expression assoziiert sind. Diese Daten implizieren das *IRF7*-Netzwerk und seinen regulatorischen Locus in der Pathogenese von T1D.

Schlussendlich beschreiben wir einen Ansatz um Allel-abhängige Störungen des Expressions-Netzwerkes auf Basis von Expressionsdaten zu analysieren. Unser Ansatz zielt darauf ab, Verbindungen im Expressions-Netzwerk zu identifizieren, die durch genetische Variation gestört sind. Dies erlaubt uns den funktionalen Zusammenhang, in dem eQTL Transkripte agieren,

zu identifizieren. Die Anwendung unserer Methode auf eQTL Daten aus der Ratte führte zur Konstruktion eines Koexpressions-Netzwerkes. Wir verwenden topologische Eigenschaften des Netzwerks, um eine Gruppe von Genen die wir Interface-Gene nennen zu definieren. Dies erlaubt es zwei anderweitig vollständig unabhängige eQTL Transkripte miteinander in Beziehung zu setzen. Die Gruppe der Interface-Gene sind angereichert mit regulatorischen Genen. Eine Analyse von Interface-Genen, die in der Blutdruckregulation tätig sind, zeigte genetische Interaktionen zwischen Sequenzvariationen, die den Blutdruck beeinflussen. Nur wenn zwei bestimmte Allele an verschiedenen Stellen des Genoms zugleich vorhanden sind, können Änderungen im Blutdruck beobachtet werden. Zuvor war die Analyse solcher Interaktionen auf Grund der kombinatorischen Vielzahl nicht möglich. Unser Ansatz erlaubte eine gezielte Analyse spezifischer Varianten, die mit Hilfe des Netzwerkes identifiziert wurden. Daher glauben wir, dass unsere Methode eine natürliche Ergänzung zu bestehenden eQTL und Netzwerk-Analyse-Methoden darstellt und auch in anderen Spezies und experimentellen Kreuzungsexperimenten verwendet werden kann.

C Summary

Current genome wide association studies (GWAS) and quantitative trait locus (QTL) studies typically result in large chromosomal regions represented by sets of polymorphic markers associated with a (disease) phenotype of interest. These molecular markers are mostly non-functional variations that are in linkage disequilibrium (LD) with unknown functional variants. Unsurprisingly, most variations have been observed in non-coding regions since they make up most of the genome. Although the understanding of non-coding sequences is far from complete it is known that they harbour a variety of gene regulatory elements. Therefore variations in non-coding sequences might alter these regulatory elements and the regulatory network. Thus, if disease associated variations also effect gene expression it is likely that they tag regulatory variants and the analysis of the genetics of gene expression provides excellent means for the identification of these regulatory variants. The regulatory effects of sequence variations can be measured systematically at the level of gene expression using the transcript level of each individual gene as a quantitative trait giving rise to expression QTLs (eQTLs). eQTL studies represent a first attempt to link genetic variation to molecular function. However, most previous eQTL studies were not able to suggest molecular regulatory mechanisms underlying the eQTLs.

Overall this thesis describes an extensive set of tools and strategies for the analysis of regulatory genetic variations. The starting point is the identification of target genes of potential regulatory variations as eQTL transcripts which has been described previously. We provide ways to address the following resulting questions about these genes. (1) What is the role of the eQTL transcript in the context of a disease model, (2) which is the *cis*-regulatory element affected by the genetic variant and which transcription factor is the upstream regulator of the eQTL transcript, (3) what are the *trans*-regulatory factors and how are their effects mediated to their target genes, and (4) what is the functional context that eQTL transcripts operate in? Moreover, we used a translational approach where gene expression networks derived from the analysis of the genetics of gene expression in a model organism are used to connect human disease association data to molecular function in an attempt to interpret the genetics of polygenic traits.

First we discuss previously described strategies for the identification of disease genes and demonstrate how their application in two case studies resulted in the identification of a candidate gene for heart failure and a candidate gene for systolic blood pressure.

Next we show that functional annotation can be used to identify genetic markers that influence functionally related gene expression networks. To that end we consider well annotated functional sets of genes i.e. pathways. Subsequently, an association statistic for gene sets is defined and evaluated for all pairs of gene sets and genomic markers. Finally, we can identify pathways that are under genetic control of a DNA variant by assessing the significance through permutations and correction for multiple testing. We applied our method to a data set from an F2 intercross. We focused our analysis on transcripts of arachidonic acid metabolism (*Ephx2*-pathway) from

the Kyoto Encyclopaedia of Genes and Genomes (KEGG) and we showed that the maximum association of the *Ephx2*-pathway transcripts occurred at the *Ephx2* locus. This highlighted that increased *Ephx2* expression leading to a depletion of cardioprotective EETs results in a feedback regulation increasing the expression of the upstream pathway to compensate the EET depletion.

We demonstrate how information about sequence variations and a biophysical model of TF – DNA interaction can be used to identify both the most likely *cis*-regulatory elements in the promoters of eQTL transcripts and the TF that is most likely the upstream regulator of the transcript. Towards that end we introduce a new computational framework to suggest possible consequences of sequence variations on regulatory networks. Our method, called sTRAP, analyses variations in the DNA sequence and predicts quantitative changes to the binding strength of any transcription factor for which there is a binding model. We have tested the method against a set of known associations between SNPs and their regulatory consequences. Our predictions are robust with respect to different parameters and model assumptions. Importantly we set an objective and quantifiable benchmark against which future improvements can be compared. Given the good performance of our method, we developed a publicly available tool which can serve as an important starting point for routine analysis of disease-associated sequence regions.

Moreover we analyse the role of TFs as mediators of *trans*-acting eQTLs and identify gene networks and the loci underlying their regulation. Importantly, we show how this gene expression network can be used to suggest a functional hypothesis beyond the results of typical human GWAS alone. Combined expression and transcription factor binding site analysis was used to define an interferon regulatory factor 7 (*Irf7*) -driven inflammatory network (iDIN) enriched for viral response genes. It represents a molecular biomarker for macrophages and was regulated in multiple tissues by a locus on rat chromosome 15q25. Computational analysis prioritised Epstein-Barr virus induced gene 2 (*Ebi2* or *Gpr183*) as the candidate regulator, which we confirm experimentally. It lies at the chromosome 15q25 locus and has been shown to control B lymphocyte migration and it is expressed in macrophages. The human orthologous locus on chromosome 13q32 controlled the human equivalent of iDIN, which was identified from monocyte expression data. Using an enrichment approach we show that iDIN genes are more likely to associate with Type 1 Diabetes (T1D) – a macrophage-associated autoimmune disease – susceptibility than randomly selected immune response genes. The human locus controlling the iDIN, was associated with the risk of T1D at SNP rs9585056, which was one of five SNPs in this region associated with *EBI2* expression. These data implicate the *IRF7* network genes and their regulatory locus in the pathogenesis of T1D.

Finally we describe an approach to analyse genotype dependent perturbations of gene expression networks solely on the level of expression data. Our approach aims to identify connections in gene co-expression networks that are perturbed by genetic variations. This allows to identify the functional context in which eQTL genes usually act. Applying our method to a data set from an experimental cross of normotensive Brown Norway rats and the Spontaneously Hypertensive Rat, we reconstructed a co-expression network. We used topological features of the network to define a set of genes that we call interface genes. This allows to link two otherwise completely independent eQTL genes. The set of interface genes is enriched in regulatory genes. Analysis of interface genes involved in blood pressure regulation revealed genetic interactions between variations influencing blood pressure phenotypes. Only if two different variants in two distinct

genomic locations are present at the same time, changes in blood pressure can be observed. Analysis of interactions was previously hindered by the combinatorial explosion when comparing pairs of variants. Our approach allowed to target this analysis to specific variants identified from the network. We believe that our method is a natural complement to existing eQTL and network analysis methods used in the analysis of the genetics of gene expression and can be applied to other experimental crosses and other species.

D Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Berlin, September 2010 Matthias Heinig