From the Institute of Human Genetics,
Helmholtz Zentrum München,
Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH)
Head: Prof. Dr. Thomas Meitinger

Gene expression studies:

From case-control to multiple-population-based studies

Thesis

Submitted for a Doctoral Degree in Natural Sciences
at the Faculty of Medicine,
Ludwig-Maximilians-Universität München

Katharina Schramm

Dachau, Germany

2016

With approval of the Faculty of Medicine
Ludwig-Maximilians-Universität München

Supervisor/Examiner:	Prof. Dr. Thomas Illig
Co-Examiners:	Prof. Dr. Roland Kappler
Dean:	Prof. Dr. med. dent. Reinhard Hickel
Date of oral examination:	22.12.2016

Dedicated to my family.

Abstract

Recent technological developments allow genome-wide scans of gene expression levels. The reduction of costs and increasing parallelization of processing enable the quantification of 47,000 transcripts in up to twelve samples on a single microarray. Thereby the data collection of large population-based studies was improved.
During my PhD, I first developed a workflow for the statistical analyses of case-control studies of up to 50 samples. With large population-based data sets generated I established a pipeline for quality control, data preprocessing and correction for confounders, which resulted in substantially improved data. In total, I processed more than 3,000 genome-wide expression profiles using the generated pipeline. With 993 whole blood samples from the population-based KORA (Cooperative Health Research in the Region of Augsburg) study we established one of the largest population-based resource.
Using this data set we contributed to a number of transcriptome-wide association studies within national (MetaXpress) and international (CHARGE) consortia. Here I will focus on three studies with main contributions:
I) Association study of gene expression levels with blood pressure related phenotypes.
II) Association study investigating changes of gene expression levels associated with aging.
III) Analysis of the impact of genetic variation on the gene expression levels.
National and international collaborations substantially increased the power of the studies and ensured independent replication. Within the German consortium we developed protocols for meta-analyses and optimized preprocessing of diverse data sets.
Whole blood is particularly useful because of its easy sampling. Especially, we could show that the impact of genetic variation is very robust and replicable within heterogeneous population-based studies.

Zusammenfassung

Moderne technologische Entwicklungen erlauben einen genomweiten Einblick in die Expression der Gene. Die Kostenreduzierung und die Möglichkeit der Parallelisierung bei der Probenvorbereitung erlaubt es 47.000 Transkripte in bis zu zwölf Proben mit einem Microarray gleichzeitig zu quantifizieren. Dadurch wird die Datenerhebung von größeren populations-basierten Studien erleichtert.
Während meiner PhD Zeit entwickelte ich zunächst einen Arbeitsablauf für die statistische Analyse von Fall-Kontroll-Studien mit weniger als 50 Proben. Mit der Generierung von populations-basierten Datensätzen etablierte ich eine Pipeline für die Qualitätskontrolle, die Vorbereitung der Daten und die Korrektur für Störfaktoren, was zu einem deutlich verbesserten Datensatz führte. Ingesamt habe ich mit dieser Pipeline mehr als 3.000 genomweite Expressionsprofile für die Auswertung vorbereitet. Mit 993 Proben aus Vollblut von Probanden der populations-basierten KORA-Studie (Kooperative Gesundheitsforschung in der Region Augsburg) haben wir dabei eine der größten populations-basierten Ressourcen geschaffen.
Mit diesem Datensatz haben wir zu zahlreichen transkriptom-weiten Assoziationsstudien in nationalen (MetaXpress) und internationalen (CHARGE) Konsortien beigetragen. In dieser Arbeit werde ich mich auf drei Studien, an denen wir maßgeblich beteiligt waren, fokusieren:
I) Eine Assoziationsstudie der Genexpressionslevel mit Phänotypen, die im Zusammenhang mit Blutdruck stehen.
II) Eine Assoziationsstudie, die die Veränderung der Genexpression im Alter untersuchte.
III) Eine Studie über den Einfluss der genetischen Variation auf die Genexpressionslevel.
Nationale und internationale Kollaborationen haben die Aussagekraft dieser Studien wesentlich erhöht und konnten unabhängige Replikationen sicher stellen. Im Rahmen des deutschen Konsortiums entwickelten wir Protokolle für Meta-Analysen und optimierten die Vorbereitung von verschiedenen Datensätzen.
Dabei erwiesen sich Proben aus Vollblut wegen der einfachen Gewinnung als besonders hilfreich. Außerdem konnten wir zeigen, dass vor allem der Einfluss der genetischen Variation sehr robust und replizierbar innerhalb heterogener populations-basierter Studien ist.

Acknowledgements

Of course it is not possible to write a PhD thesis without any help. There are lots of people who are involved in many steps of preparing this thesis and I owe them all a debt of gratitude.

First of all I want to thank my doctoral thesis supervisor Prof. Dr. Thomas Illig for his support and help during the last years. Especially for his straightforwardness and his patience.

I would also like to thank Prof. Dr. Thomas Meitinger and Dr. Holger Prokisch for allowing me to work in the Institute of Human Genetics and providing me all the data.

During the PhD project we had lots of collaborations and I would especially like to thank Alexander Teumer and Claudia Schurmann for lots of fruitful discussions per mail, on the phone, and in personal.
I got help with the analysis of genotype data from Christian Gieger, Eva Albrecht and Janina Ried.
For the analysis of the methylation data we had a very close and intensive collaboration with Melanie Waldenberger, Eva Reischl, Anja Kretschmer, Petra Wolf, Nazanin Karbalai, Brigitte Kühnel, and Carola Marzi. Without their help and support no results would have been produced.
Additionally I would like to thank all the members of the MetaXpress and CHARGE consortium for a great collaboration. And last, but not least I am indebted to all participants of the KORA study. Without them no data analysis would be possible.

A special thank goes to all the colleagues at the Institute of Human Genetics who always listened to all the problems that appeared during my Phd student time. We had and still have a great time and they helped that the time flew. Especially I would like to thank Caro, Carola, Konzi, and Martina for spending a lot of time reading this thesis, Thomas S. for answering my Latex-questions, and Thomas W. for being my personal bioinformatician.

I gratefully dedicate this thesis to my parents, my sisters, my patient husband, and my little sunshines Annika and Adrian.

Contents

List of Tables

List of Figures

1. Introduction

In the beginning of gene expression era statistical analyses were quite simple because gene expression levels were determined only for a few candidate genes and could be compared graphically or by using simple statistical tests.
The development of microarrays wherein thousands of measurements for one single sample are conducted simultaneously led to the generation of large amounts of data making statistical analyses more complex and time-consuming. These experiments started with very small sample sizes and were mainly designed to compare cases with controls especially in humans and mice.
The Institute of Human Genetics of the Helmholtz Center Munich was one of the first institutions that established a larger genome-wide data set of more than 300 healthy individuals from a population-based study using whole blood. This data set provided us with an opportunity to assess gene activity across the whole genome in a hypothesis-free approach.
However soon it became clear that data analyses together with interpretation of multiple significant hits was no longer trivial, as sample sizes were continuously growing due to the possibility to analyze the activity of the whole genome in a short time for less money (Ramasamy et al., 2008).

Nowadays, the analysis of high-dimensional data is no longer an exception but rather normal and all-round. There are lots of population-based studies with large sample sizes which are analyzed together to identify even very small effects.
This thesis reflects the development of gene expression studies from case-control studies with small sample sizes ($n<50$) to studies with large sample sizes ($n>7{,}000$) using data from different populations.
The challenges and improvements of quality control and analysis of data are shown over time and for increasing sample sizes.
This thesis can be regarded as a guideline to analyze gene expression data obtained mainly from whole blood but also from other tissues. On the one hand it supports statisticians who have not worked with genetic data so far and on the other hand biologists who are not familiar with statistical analyses of gene expression data. Therefore, both the genetic and statistical background is given so that all analysis steps can be understood and reproduced. Especially for the reproducibility the required R codes are provided.
The population-based phenotypes, gene expression and genetic data used in this thesis can be obtained from KORA-PASST (**p**roject **a**pplication **s**elf-**s**ervice **t**ool) on `http://epi.helmholtz-muenchen.de/`. The expression data can be downloaded (without any phenotypes due to protection of data privacy) from ArrayExpress using the project number E-MTAB-1708 (`https://www.ebi.ac.uk/arrayexpress/`).

1.1. Genetic background - From DNA to gene expression

All genetic information of each human being (and of all other living organisms) is stored in the deoxyribonucleic acid (DNA) which is unique for each individual. The structure of the DNA was firstly described by James Watson and Francis Crick in 1953 in a Nature publication: "Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid" (WATSON and CRICK, 1953). They introduced the double-helix structure of two strands of nucleotides (see Figure 1.1). Each nucleotide consists of three components:

1. One base: adenine (A), cytosine (C), guanine (G) or thymine (T)
2. One sugar (deoxyribose)
3. One phosphate

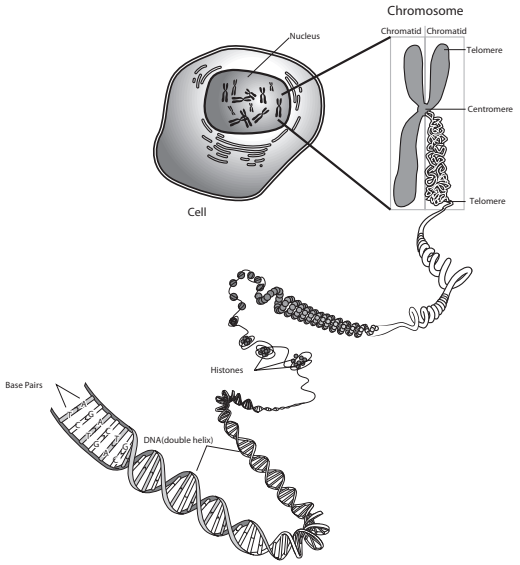

Figure 1.1.: Structural organization of DNA in the cell nucleus:
Base pairs build up the DNA in form of a double-helix. DNA is ordered and structured by histones and form the chromosomes. Thereby each chromosome consists of two identical chromatids and its intersection called centromere. The telomeres at each end of the chromosome have a protective function (McClintock, 1941). Humans have 22 pairs of chromosomes and additionally two gender-specific chromosomes (females have two X-chromosomes, males have one X- and one Y-chromosome).

Thereby the sugar and the phosphate serve as the so-called backbone of the DNA and the bases are attached with hydrogen bonds. In this process an A can only bond to T and a C only to G.
The DNA can be found in the nucleus of each single cell. The different shapes and functions of a cell are due to the different activity of genes. A gene is a defined part (ranging from a

few kilobases to several megabases) of the DNA and consists of a certain number of coding (exons) and non-coding (introns) regions.

The activity of the genes is called gene expression and this process consists of two steps:

1. Transcription: Generation of a copy of the gene, the so-called messenger RNA (mRNA). In contrast to the DNA, the Ribonucleic Acid (RNA) is single-stranded and contains ribose instead of deoxyribose. Additionally, the base thymine is replaced by uracil.

2. Translation: The information of the mRNA is translated to amino acids that form proteins.

If variation in the DNA sequence in comparison to the reference sequence is located in the coding region of a gene, an exchange of an amino acid could be the consequence. If one of these changes with only one affected base pair occurs in more than 1% of the population it is called a Single Nucleotide Polymorphism (SNP) (Wrba et al., 2007). In humans these variants can be found on average every 500 to 1,000 base pairs and normally they are not disease-relevant. There are three different states (genotypes) possible for each SNP. Either an individual is homozygote for the major allele, meaning that both chromosomes carry the same allele that is most frequent in the normal population at this locus, an individual is homozygote for the minor allele or an individual is heterozygote, meaning that both chromosomes are carrying different alleles (Ziegler et al., 2010).

The tissue specific activation and inactivation of genes can be regulated by DNA methylation which is one of the main epigenetic regulatory mechanisms (Portela and Esteller, 2010). Epigenetics is the study of heritable changes in gene function that occur without a change in the DNA sequence. Epigenetic mechanisms include histone modification, DNA methylation, and RNA interference (Nikolova and Hariri, 2015).
DNA methylation is the process in which a methyl group is added to the DNA, most commonly to cytosine if it is directly followed by guanine (Tollefsbol, 2010). Thus cytosine can occur "normally" or in a methylated version, i.e. with an attached methyl group. The regions in the DNA where the bases cytosine and guanine are only separated by a phosphate are called Cytosine-phosphate-guanine (CpG) sites (Miller et al., 1974). CpG islands are regions in the genome with a high frequency of CpG sites (Figure 1.2).

Figure 1.2.: Categories of CpG sites:
Genomic regions with a high frequency of CpG sites are called CpG islands. The regions 2 kb up- and downstream of CpG islands are called North and South Shore, respectively and the flanking regions are called North and South Shelves, respectively (Bibikova et al., 2011).

In humans it is assumed that methylation of CpG sites or CpG islands close to the transcription start site of a gene can repress gene expression (see Figure 1.3) but methylation within the gene body might not interfere with gene expression (Jones, 2012).

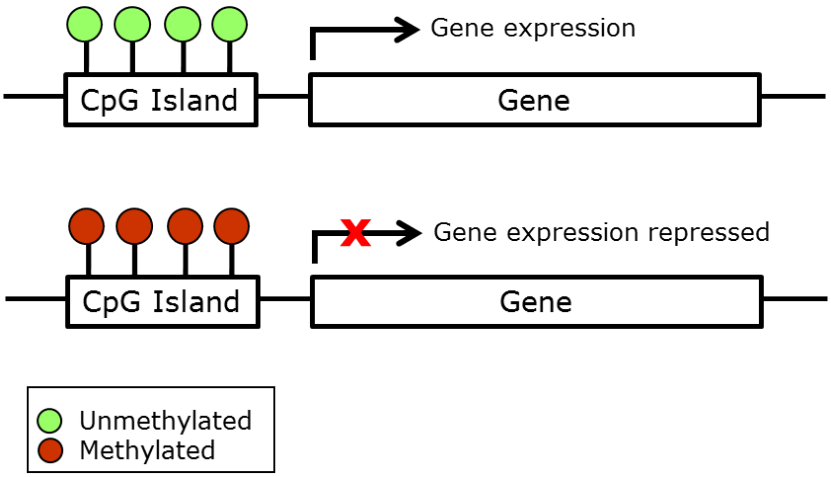

Figure 1.3.: Expected correlation of CpG sites and gene expression:
A methylated CpG site or a CpG island (genomic regions enriched in CpG sites) might repress the expression of a neighboring gene (Nikolova and Hariri, 2015).

1.2. Gene expression studies

The principle of unidirectional information flow of genetic information was established in the mid-1960s (Struhl, 1999) and thereby closed the gap between DNA in the nucleus and proteins in the cytoplasm (O'Connor and Adams, 2010). Volkin and Astrachan (1956) discovered that genetic information is transported from DNA and translated into proteins by RNA.

The DNA in the nucleus consists of exons and introns and is transcribed to RNA molecules by enzymes called RNA polymerases (see Figure 1.4). The RNA is single-stranded and complementary to one strand of the DNA. The non-coding regions are removed and the remaining exons are spliced together. Next, the messenger RNA is exported to the cytoplasm and the transcripts are translated to chains of amino acids which finally form the proteins (O'Connor and Adams, 2010).
This highly complex process including transcription of gene and translation to protein is called gene expression. Gene regulation determines the amount and time point of specific gene products and can occur in each step (Maston et al., 2006). The mechanisms for regulation of gene expression levels occur mostly at the level of transcription (Holstege and Young, 1999).
A high proportion of gene regulation is a result of an interaction between DNA sites (binding sites) and proteins that bind to these sites, so-called transcription factors. They can either bind to promoters (to initiate the transcription), enhancers (to enhance the transcription) or to silencers (to repress the transcription). Promoters, enhancers, and silencers are short parts

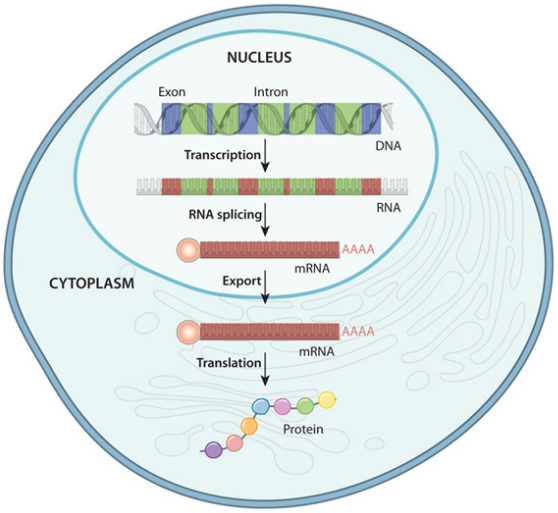

Figure 1.4.: Shematic diagram of gene expression:
The DNA in the nucleus, divided in exons (blue) and introns (green), is transcribed to a pre-mRNA molecule (red regions align with blue exons and green regions align with introns of DNA). Next, mRNA is build by removing introns and splicing of exons. The newly formed mRNA is exported from the nucleus to the cytoplasm and translated to proteins which consists of different amino acids (here illustrated by different colors) (O'Connor and Adams, 2010). Figure is taken from `www.nature.com/scitable`.

of DNA and are normally located close to the transcription start site of the regulated gene (Pennacchio et al., 2013; Blackwood and Kadonaga, 1998; Maston et al., 2006). Mostly, transcription factors interact with other proteins (coactivator and corepressor) to result in an up- or down-regulation of a gene, meaning the rate of transcription is increased or decreased, respectively.
After transcription of DNA to mRNA there can be some post-transcriptional regulation to determine the amount of mRNA that is translated to proteins. Post-transcriptional regulation also includes mechanisms to manipulate RNA transport or stability.
Finally, genes can be regulated in translational and post-translational steps. This includes all stages of protein biosynthesis and chemical modifications of proteins (Mehta, 2009).

Microarrays measure the current level of a transcript and provide information about gene activity.
Variation in gene expression levels does not necessarily result in defined clinic symptoms but may lead to complex diseases or influence quantitative traits like BMI or height. In contrast, diseases like Mendelian disorders represent an extreme consequence of genetic variation (Cheung et al., 2003).

Most gene expression studies were performed to investigate differences between special conditions in gene expression levels. Although there are studies analyzing only a set of genes the focus of this thesis is the analysis of genome-wide microarray data, of up to 30,000 transcripts.

Generation and quantification of gene expression levels are not as robust as genome data. In various tissues (Petretto et al., 2006) and cellular states (Gerrits et al., 2009) different genes are expressed and measurable. The RNA levels are more susceptible to experimental design of the study, to technical and biological variables, and to environmental factors. The variability of gene expression data can be used to identify differentially expressed genes that are altered by traits or diseases.
The analysis of the transcriptome aims to include the complete set of mRNA molecules at a given time point from a defined cell type or tissue (Cornelis and Hu, 2013). Mostly, these transcriptome-wide analyses are conducted in whole blood as it is easily available.

1.2.1. Case-control studies

Case-control studies are a type of observational studies in which samples from two different groups are compared for a better understanding of underlying biology and pathomechanism and to identify biomarkers. The "cases" are usually patients suffering from a disease or individuals with a particular condition (for example high BMI), whereas the "controls" are the individuals that are healthy (or are at least not affected by the disease of the cases) or do not have the same conditions (for example low BMI). In comparison to population-based studies, in which the general population is investigated, in case-control studies the controls are accurately chosen (for example they fit to the cases in age, gender and other sociodemographic variables) to optimize the power for the identification of small differences between the two groups. In this way it is possible to investigate even rare diseases with a small sample size.
In gene expression studies using whole blood usually the expected effects in gene expression levels between the groups are quite small as there is not always a direct connection between blood and the investigated disease and the genes of interest are not differentially expressed in blood. There are some possibilities to avoid this problem:

- Analysis of the affected tissue.
  Lots of genes are not expressed in every tissue and especially in whole blood. When analyzing rare diseases sample sizes are usually small and therefore it is possible to investigate the affected tissue.

- High contrast between cases and controls.
  To increase the contrast between the groups it is necessary to have homogenous cases and controls. In the optimal situation each case gets one or more matched control based on variables that are expected to be confounders (Rose and Laan, 2009).

- Analysis of selected genes.
  To reduce the number of tests it could be useful to restrict the analysis on a predefined set of genes.

In practice it is not always possible to meet all these criteria. Mostly, the limitation is to investigate the affected tissue as there are high ethical standards that do not allow to conduct, for example, biopsies on healthy individuals. Therefore the selection of cases and controls or the exclusion of samples with bad quality is even more important.

1.2.2. Population-based gene expression studies

Large cohort studies are usually based on individuals who represent the normal population. In these cohorts there are no specific tissues available and due to ethical reasons most of the time only blood is taken -as easily available sample- from the voluntary study participants. However, it could be shown that up to 80% of all genes are expressed in whole blood (Liew et al., 2006). Therefore population-based transcriptome-wide association studies are mostly conducted on whole blood samples using phenotypes that are accessible (e.g. BMI, height, age, blood pressure,...) or studying diseases that are quite common in the normal population (e.g. diabetes, hypertension,...). As usually large sample sizes and appropriate replication cohorts are available there is high power to detect small effects even when not having the optimal tissue.

1.2.3. Genome-wide association studies

Just like the measurement of gene expression levels, it is also possible to determine the genotypes of hundred of thousands SNPs on microarrays. The development of this technique was the beginning of genome-wide association studies (GWAS). In a GWAS genetic variation is analyzed genome-wide to identify genetic loci that are associated with the trait of interest. The starting point of this era was the publication from Klein et al. (2005) who investigated 103,611 SNPs from 96 AMD (age-related macular degeneration) patients and 50 controls followed by the Wellcome Trust Case Control Consortium in 2007 which analyzed almost 400,000 SNPs from 16,000 samples suffering from seven different diseases namely bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes (Consortium, 2007; Visscher et al., 2012). Since then thousands of GWAS have been conducted and thousands of associations with diseases or phenotypes have been identified.
GWAS are performed to determine a so-called quantitative trait locus (QTL), a genetic locus that is associated with any quantitative phenotype or variable. Normally, the result of a GWAS is a list of SNPs which are associated with the disease (for example diabetes or cardiovascular diseases) or the trait (for example BMI or height) of interest. Most of the SNPs are located in intronic or even in intergenic regions and the function is not clear (Mehta et al., 2012).

1.2.4. eQTL studies

Jansen and Nap (2001) first introduced the concept of the genome-wide analysis of genetic and gene expression data (Li and Deng, 2010). As the molecular mechanism that is responsible for the association between the genetic locus and the phenotype is often not clear, gene expression data can help to understand the functional consequences of SNPs. This became more important because of the constantly increasing number of GWAS identifying SNPs in non-coding regions. So far, more than three thousand loci have been identified to be associated with a disease or a trait. These loci are all summarized in the GWAS catalog (Hindorff et al., 2009).
With the development of expression-microarrays it is now possible to investigate the impact

of genetic variation on gene expression levels systematically and independent from GWAS results. Gene expression levels are treated as quantitative phenotypes and analyzed genome- and transcriptome-wide to identify expression QTLs (eQTLs).
An eQTL is a locus in the DNA that influences the expression level of one or more genes (Albert and Kruglyak, 2015) and is mainly identified in population-based studies. eQTL studies provide the opportunity to detect transcriptional regulatory relationships on a genome-wide level (Fehrmann et al., 2011).
There are two different kinds of associations between gene expression levels and genetic variation that were described first by Haldane et al. (1941) and were mostly classified according to their physical distance (Gilad et al., 2008).

- *cis*-eQTLs:
  According to the original definition *cis*-acting elements have an influence on allele-specific gene expression. This includes for example promoter regions, silencer or enhancers (Gilad et al., 2008). *cis*-eQTLs are often synonymous with local eQTLs (Albert and Kruglyak, 2015), meaning that the SNP is close to the regulated gene (Michaelson et al., 2009).

- *trans*-eQTLs:
  trans-acting elements regulate the expression of both alleles (Gilad et al., 2008). Mostly, the locus is located far away from the regulated gene (distant eQTL), either on the same or on a different chromosome.

When using microarrays it is not possible to measure allele-specific gene expression. Therefore it is not possible to definitely distinguish between mode of action of *cis*- and *trans*-eQTLs. Consequently, in eQTL studies it has been common to name all local eQTLs *cis*-eQTLs and all distant eQTLs *trans*-eQTLs (Albert and Kruglyak, 2015).
So far, there is no common threshold to separate *cis*- and *trans*-eQTLs. It varies from 100 kb (Dixon et al., 2007) to 1 Mb (Hao et al., 2012). We defined "close" between SNP and regulated gene as less than 500 kb upstream and downstream of the transcription start and end site, respectively (see Figure 1.5).

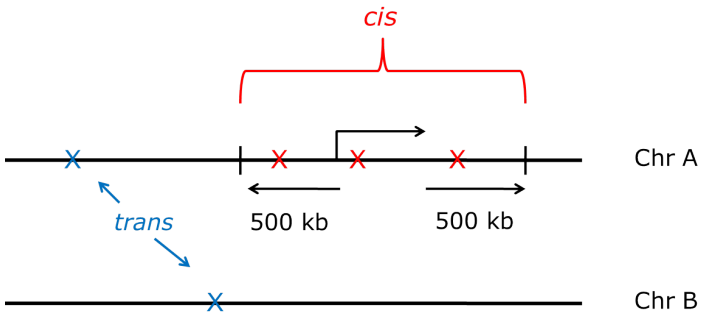

Figure 1.5.: Definition of effects of SNPs on gene expression levels in *cis* and *trans*:
If the SNP is located less than 500 kb (or any other small predefined distance) from the expression probe on the same chromosome the eQTL is called *cis*-eQTL. If the SNP is located further away on the same chromosome or on an other chromosome than the expression probe, the eQTL is called *trans*-eQTL.

1.2.5. Statistical processing of gene expression data

Gene expression data are the result of several complex sample preparation and measurement steps followed by the computational translation of signals from microarrays into data points (Hartemink et al., 2001). Therefore there are lots of possibilities where variability of data could be caused by technical instead of biological factors of interest. As a consequence, experiments have to be prepared carefully to remove or at least reduce the technical variability and increase the probability to detect real biological relevant signals. There are many challenges in the analysis of gene expression data and the research on finding the optimal way for the preprocessing and analysis is still ongoing.

One characteristic of gene expression analysis is the $p > n$ situation, meaning that the number of investigated features p is much larger than the sample size n. When using microarrays, up to 50,000 data points can be measured per sample. A study cohort only consists of a few hundred to thousand individuals. Not all statistical methods can handle this problem and multivariate models have to be applied carefully. Especially when analyzing gene expression data from studies with a small sample size the study design has to be considered with caution to get a homogenous study population.

After choosing the optimal set of samples, problems can occur during the measurements in the laboratory. As mentioned before, gene expression levels are influenced by several environmental and technical factors especially batch effects. To make expression data comparable across different studies, it is mandatory to standardize all experimental steps and the following data preprocessing steps should include variance stabilization and normalization methods. Lastly, suitable statistical models are needed that are able to exclude variance due to technical known and unknown factors.

Additional variance in the data set can be the result of outliers. A sample can be an outlier due to several reasons. For example a bad quality of RNA can lead to overall lower expression levels and thereby decrease the number of detectable genes. Another possibility is sample mixing in each step of the experiments. Especially when men and women samples are mixed problems can occur in any gender-specific analysis because of highly different expression patterns on sex-specific chromosomes. Therefore the data set can be improved by identification and removal of outliers and potential mixed samples before analysis to harmonize the data set and reduce the variance.

Gene expression analyses are often performed to identify differentially expressed genes between two or more different groups of samples. These differentially expressed genes can be used to obtain insights in diseases or serve as clinical biomarkers. In this approach all measured transcripts are investigated one by one and this results in are large number of statistical tests. To avoid false positive hits multiple testing correction methods have to be applied.

Gene expression data are biological data that usually do not follow the assumptions of the distributions of standard statistical models (Du et al., 2010). This has to be kept in mind when applying parametric approaches and could maybe be a reason for nonparametric models.

In order to avoid type I error, results have to be confirmed in an independent data set of comparable population structure. Alternatively, data can be analyzed in a meta-analysis with the additional advantage that this increases the sample size and thus the power to detect small differences or effects. Meta-analyses or analysis in large data sets can result in large lists of potentially interesting genes. Statistical or bioinformatical tools are able to identify relations between these results or identify enriched biological pathways (Krumsiek et al., 2011).

1.3. Outline of this thesis

The main focus of this thesis is the analysis of whole genome expression data. An overview of the outline of this thesis is given in Figure 1.6. Within the framework of this thesis six studies are described which display various different characteristics.

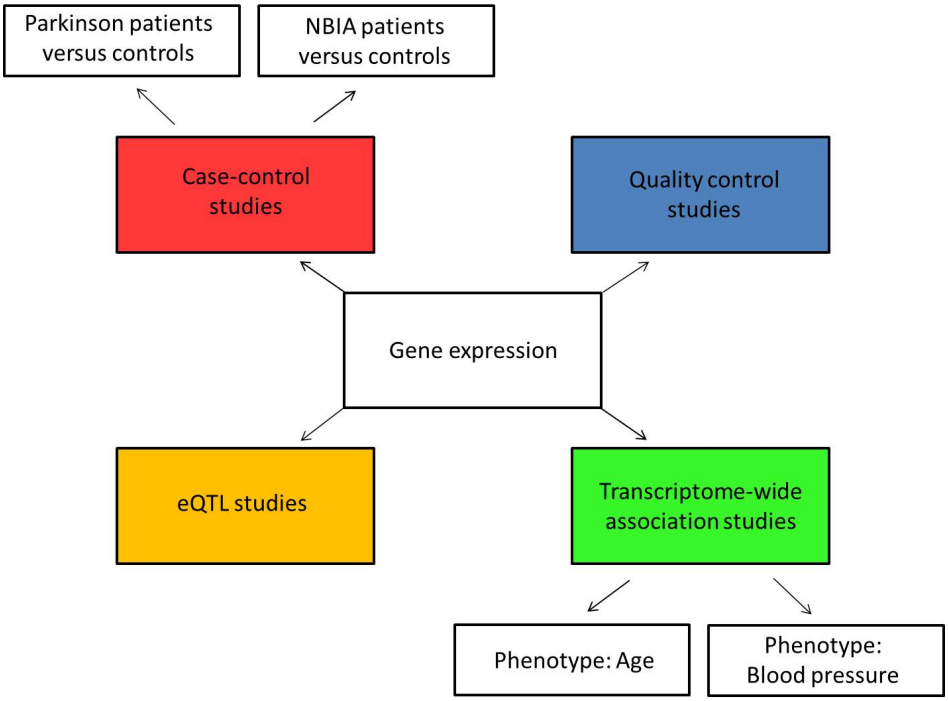

Figure 1.6.: Overview of conducted studies

1. They were conducted using different study designs: case-control studies (Chapter 3), quality control studies (Chapter 4), transcriptome-wide association studies (Chapter 5), and eQTL studies (Chapter 6).

2. The expression of genes was measured in different samples/populations: a Parkinson case-sontrol sample (Section 3.2), a NBIA (Neurodegeneration with Brain Iron Accumulation) case-control sample (Section 3.3), and population-based studies (Chapter 4, 5, and 6).

3. They aimed to address different research questions such as: a) Identification of differentially expressed genes in case-control studies (Chapter 3), b) Optimization and standardization of quality control and analysis (Chapter 4), and c) Identification of associations using population-based data (Chapter 5 and 6).

The six studies conducted within the framework of this thesis are described in detail in the following section. An overview of these studies is given in Table 1.3.

Case-control studies

In Chapter 3 two case-control studies with small sample sizes are introduced to show arising problems in analysis of gene expression data sets with small sample sizes. In the first case-control study (Section 3.2), we identified genes that were differentially expressed in single cells from brain in Parkinson patients and age-matched controls. Results were followed up in young controls to describe Parkinson's disease as accelerated aging. In the second case-control study (Section 3.3) we compared patients suffering from NBIA to controls using a different expression array system than in all other projects described in this thesis (i.e. arrays from Affymetrix instead of those from Illumina). The aims of both studies were:

- The development of an optimal preprocessing pipeline including the selection of the samples to obtain a homogenous data set.
  For this purpose, expression data were normalized and clustered to identify and remove outliers.
- The identification of differentially expressed genes between patients and controls.
  Expression levels of selected genes (Parkinson study) or transcriptome-wide expression levels (NBIA study) were compared to identify differentially expressed genes.
- The identification of molecular mechanisms.
  Differentially expressed genes were used to identify pathways which explain relations between the underlying genes and the disease phenotype.

Quality control studies

Chapter 4 can be seen as the basis for all further analyses in population-based data sets. It describes the development of quality control steps from a manual quality control in a small cohort (N=381) to a common data preparation and analysis strategy in a large consortium consisting of three population-based studies with altogether more than 3,000 individuals. The overall aims were:

- To explore robustness and variability of gene expression data by analyzing biological and technical replicates.
  Therefore, blood was taken from three voluntary individuals three times at three different time points. Gene expression was measured to compare the inter- and intra-variability between all three samples.
- To optimize preparation of samples in the laboratory.
  Experimental processing for a larger data set of more than 3,000 samples was standardized and different amounts of RNA were compared to obtain expression data with comparable intensity levels.

- To optimize preparation of data.
  With increasing sample size we were able to exclude sample outliers based on different criteria, like cluster outliers, mixed samples or samples with bad RNA quality.

- To identify technical and clinical variables that influence gene expression levels.
  In order to create three comparable data sets the most important pre-requisite for a joint consortium analysis of population-based studies was the standardization of preprocessing steps. Two variance stabilization methods were compared (log2 versus variance stabilization transformation). Using different approaches, we tried to uncover main technical influences on variation of gene expression levels. We investigated the effects of SNPs in probe sequences by calculating the association between gene expression levels of transcripts having a SNP in the probe sequence and the corresponding SNP.

In summary, all steps were conducted with the objective to reduce disturbing variability in the data in order to create comparable data sets. Combined analysis in consortia are now easy to conduct due to the development of harmonized data sets.

Transcriptome-wide association studies

In Chapter 5 two different phenotypes (age and blood pressure related phenotypes) were analyzed on a transcriptome-wide scale. Each phenotype was analyzed in three different populations.
The first population-based association study (Section 5.1) was conducted to identify genes which are associated with phenotypes related to blood pressure such as systolic and diastolic blood pressure, pulse pressure, and hypertension.
The second population-based study (Section 5.2) was focused on age-related gene expression. The power to identify age-related expression patterns is limited in a population with 1000 samples who are older than 60. However, we contributed with our data to a meta-analysis of more than 7,000 samples with a large range of age (20 years to 100 years).

eQTL studies

In Chapter 6, we investigated the impact of genetic variation on gene expression levels in an eQTL study in two different data sets of 322 and 890 samples. The aims of the studies were:

- To identify genetic determinants of gene-expression in *cis* and *trans* (i.e. *cis*- and *trans*-eQTLs).
  We identified *cis*- and *trans*-eQTLs on transcriptome- and genome-wide scales. For *trans*-eQTLs we also searched for what we called "master regulatory sites". These loci simultaneously influence the activity of several genes.

- To analyze the robustness and reproducibility of eQTLs in whole blood across different studies.
  All eQTLs that have been identified in the eQTL studies described in this thesis were tested for replication in all together three different independent data sets. As one of

these data sets showed differences in the preprocessing of gene expression samples it was of particular interest to explore the robustness of eQTLs.

- To validate if whole blood can be used as a surrogate tissue.
  We compared the identified *cis*-eQTLs in whole blood to publicly available eQTL results observed in other tissues (liver, lymphoblastoid cell lines, monocytes, b-cells, and lung tissue, respectively).

Taken together, this thesis shows the analysis of whole-transcriptome expression data in different settings (different study design and various different study populations) with the aim to explore various important research questions.
One major limitation in whole genome expression data are small effect sizes that require large sample sizes in order to detect effects on a genome-wide scale.
We started with about 20 samples in case-control studies and ended up with several thousand samples in transcriptome-wide association and eQTL studies, respectively (see Figure 1.7).
The development of increasing sample sizes required improved methods for data preprocessing and analysis which are summarized in this thesis.

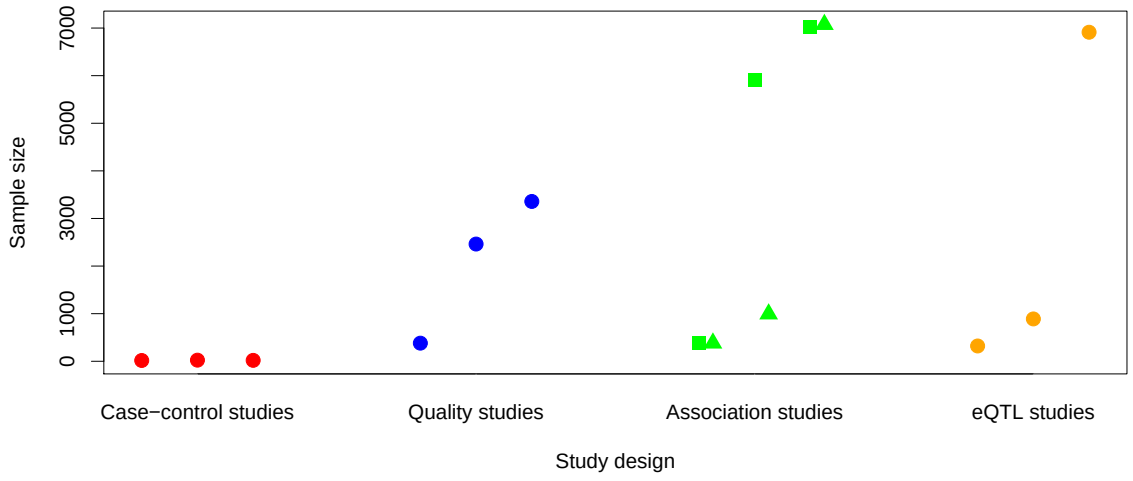

Figure 1.7.: Sample sizes in the gene expression studies described in this thesis:
Each study was conducted in three different populations with increasing sample sizes. We performed two different association studies with two different phenotypes: blood pressure related phenotypes (indicated with squares) and aging (indicated with triangles).

N	Study design	Tissue	Study population	Aim	Trait	Chapter	Reference
17	Case-control	Single cells from brain	Parkinson patients, old controls	Disease-specific identification		3.2.3	Elstner et al. (2009)
24	Case-control	Single cells from brain	Parkinson patients, young/old controls	of biomarkers and	disease	3.2.4	Elstner et al. (2011)
20	Case-control	whole blood	NBIA patients, controls	targets		3.3	Hartig et al. (2011)
381	Quality control	whole blood	KORA F3	Development of		4.1	
2,463	Quality control	whole blood	KORA S4, F4, F4OGTT	quality controlled	quality	4.2	
3,358	Quality control	whole blood, monocytes	MetaXpress	preprocessing and analysis strategy	variables	4.3	Schurmann et al. (2012)
377/989	Association	whole blood	KORA F3/F4	Identification of genes		5.1.1	
5,907	Association	whole blood, monocytes	MetaXpress, MESA	associated with blood pressure	blood pressure	5.1.2	Mueller et al. (2014)
7,017	Association	whole blood	CHARGE	related phenotypes		5.1.3	Huan et al. (2015)
381	Association		KORA F3	Identification of		5.2.1	Mehta (2009)
993	Association	whole blood	KORA F4	age-related	age	5.2.2	
7,074	Association		CHARGE	phenotypes		5.2.3	Peters et al. (2015)
322	Association		KORA F3	Identification	genetic variation	6.1	Mehta et al. (2012)
890	Association	whole blood	KORA F4	of *cis*- and		6.2	Schramm et al. (2014)
6,913	Association		CHARGE	*trans*-eQTLs		6.3	Westra et al. (2013)

Table 1.1.: Overview of conducted studies:
In all studies gene expression data were analyzed transcriptome-wide with different traits of interest.
KORA is the abbreviation for Kooperative Gesundheitsforschung in der Region Augsburg (Cooperative Health Research in the Region of Augsburg). It is a population-based study consisting of several surveys such as F3, S4 or F4. MetaXpress is a consortium consisting of three population-based studies with available gene expression data measured in whole blood and monocytes, respectively. MESA is the abbreviation for US Multi-Ethnic Study of Atherosclerosis and is a population-based study. CHARGE (The Cohorts for Heart and Aging Research in Genomic Epidemiology) is a large international consortium and consists of several European, American, and Australian population-based studies.

2. Material and methods

2.1. Study population and consortia

2.1.1. KORA

KORA (Kooperative Gesundheitsforschung in der Region Augsburg - Cooperative Health Research in the Region of Augsburg) exists since 1996 in the region of Augsburg in the south of Germany and builds on MONICA (Monitoring of trends and determinants in cardiovascular disease) (Holle et al., 2005). It is a regional research platform for population-based surveys and follow-up studies and a cohort of more than 18,000 subjects are actively followed up to date.
Four cross-sectional healthy surveys S1 to S4 have been performed at five years intervals each containing independent random samples with German nationality from Augsburg city and sixteen communities from the adjacent counties. All study participants were asked for sociodemographic variables, risk factors (smoking, alcohol consumption, physical activity, etc.), medical history and family history of chronic diseases and some more clinical parameters.

The KORA F3 (follow-up study of KORA S3) samples were collected between 2003 and 2004 and 2,974 individuals were included. Additionally the genotypes of 1,388 samples were collected.

The measurement of gene expression levels in KORA S4 and F4 was a project in collaboration with the DDZ (Deutsches Diabetes Zentrum - German Diabetes Center Düsseldorf) and therefore the aim of the study was the early diagnosis and the prediction of diabetes. Testing for pre-diabetes in a large population is only possible with a time-consuming oral glucose tolerance test (OGTT). Normal and easy fasting blood glucose monitoring methods miss about 40% of the undetected diabetics.
Blood was taken from fasting subjects and the fasting blood glucose levels were determined. Then 75g of dextrose was drunk, two hours later blood was taken and the blood glucose level was determined again. So the reaction of the body to supply of glucose could be identified. Altogether three blood probes of each sample were available: a baseline measurement (KORA S4), a follow-up measurement around eight years later (KORA F4) and the measurement after the oral glucose tolerance test (KORA F4 OGTT).
Additionally there are several clinical and sociodemographic variables available and the genetic variation.

2.1.2. SHIP-TREND

SHIP is a population-based project in West Pomerania, a region in the northeast of Germany. For all projects in this thesis samples from the SHIP-TREND study were used. Baseline examinations for this study started in 2008. From the total population of West Pomerania com-

prising approximately 210,000 inhabitants, a stratified random sample of 8,016 adults was drawn. Stratification variables were age, sex, and city/county of residence. By the end of 2012, 4,420 samples have been examined. The detailed study design and sampling methods are described by Völzke et al. (2010). All analyses in the SHIP-TREND study were conducted by Dr. Claudia Schurmann or Dr. Alexander Teumer.

2.1.3. ECGUT

The Biobank of the Estonian Genome Center of the University of Tartu (EGCUT) is based on a population-based study which collected data of more than 50,000 individuals. In comparison to the KORA F4 data the age distribution of EGCUT reflects the age distribution of the adult Estonian population. We worked together with Eva Reinmaa who conducted the analysis in the EGCUT data set.

2.1.4. GHS

The Gutenberg Health Study (GHS) is designed as a community-based, prospective, observational, single-center cohort study in the Rhine-Main area of Western Germany (Wild et al., 2010). All participants (50% males) live in Mainz and the district of Mainz-Bingen and are between 35 and 74 years of age. Baseline examinations of 15,000 study participants were performed between 2007 and 2012. All analyses were conducted by Arne Schillert and Christian Müller.

2.1.5. MetaXpress

In 2011 the MetaXpress Consortium was founded within the DZHK - **D**eutsches **Z**entrum für **H**erz-**K**reislauf-Forschung (German Center for Cardiovascular Research). MetaXpress consists of three large German study cohorts with available gene expression data:

- GHS (Mainz, Lübeck, Hamburg)
  Described in Section 2.1.4
- KORA F4 (Munich)
  Described in Section 2.1.1
- SHIP-TREND (Greifswald)
  Described in Section 2.1.2

Aim of the project is the common analysis of gene expression data associated with cardiovascular phenotypes, like obesity (BMI, Waist-to-hip ratio), hypertension (systolic and diastolic blood pressure, pulse pressure), and diabetes (fasting glucose, 2h-glucose).
The experimental procedure was almost identical for KORA F4 and SHIP-TREND and slightly different for GHS (see Figure 2.1). One reason for this is that samples from KORA F4 and SHIP-TREND were both proceeded in Munich using the same protocol. Additionally for both cohorts gene expression was measured in whole blood while GHS used monocytes.

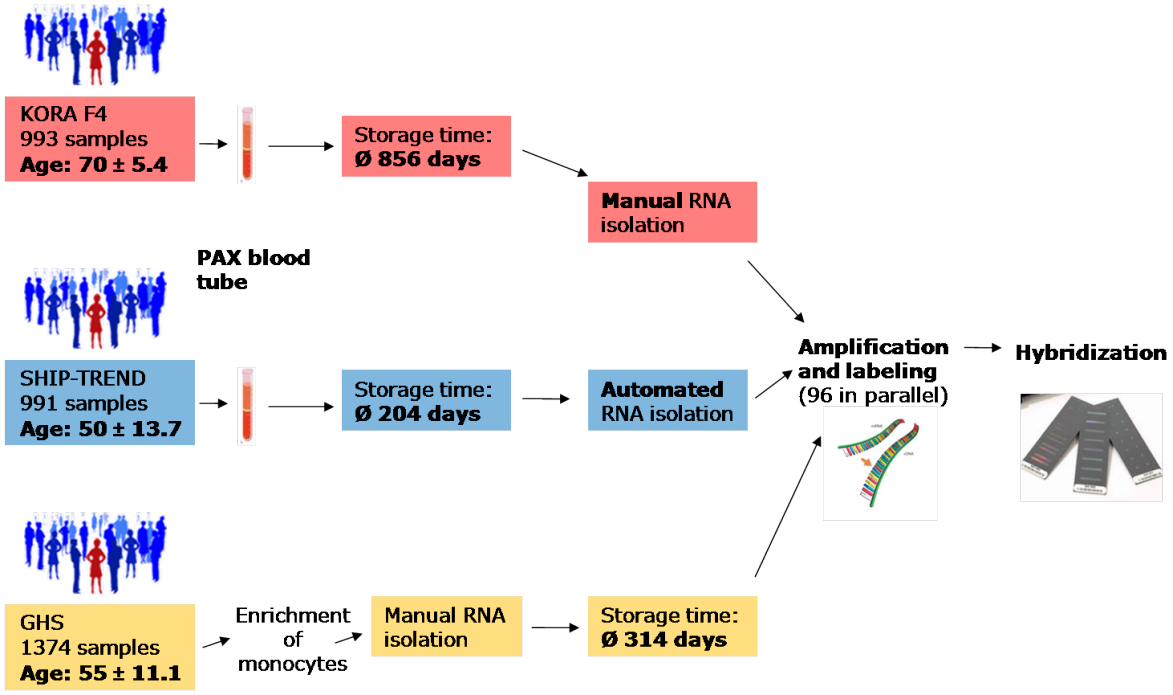

Figure 2.1.: MetaXpress - Experimental procedure

2.1.6. CHARGE

CHARGE is the abbreviation for "The **C**ohorts for **H**eart and **A**ging **R**esearch in **G**enomic **E**pidemiology" Consortium. It consists of several European, American, and Australian population-based studies. With the expression data of KORA F4 we contribute to the working group "Gene Expression".

2.2. Useful software for the analysis and visualization of gene expression data

2.2.1. GenomeStudio

The GenomeStudio software from Illumina is useful to convert the scanned data to expression values and to show first quality parameters. It can also be used to do some basic analyses and graphics. But with an increasing number of samples a high amount of RAM is required to upload all samples in one project.
One quality score is the detection p-value which indicates the probability that the observed expression value of a transcript is significant higher than the background noise. All samples (if they are from the same tissue) should have an comparable number of detected transcripts. A low number could be an indication for a poorly processed sample and therefore the number of detected transcripts is a good overall quality score (Illumina, 2007).

2.2.2. R and Bioconductor

R is a open source software environment for statistical computing and it is available for download at `http://www.r-project.org/`. It can be used on UNIX, Windows and MacOS. It is especially helpful to handle large data sets but could also be used as a calculator. Problem specific code-packages are available on an internet platform (cran) which are made available by other R users. Specifically the Bioconductor (`www.bioconductor.org`) is an extension of R and was developed to analyze biological data sets (Huber et al., 2015). The following two lines are necessary to install Bioconductor with the basic packages:

Listing 2.1: Installing Bioconductor with basic packages

```
source("http://bioconductor.org/biocLite.R")
biocLite()
```

The following R packages were used for the analysis in this thesis:

- `lumi`
  This Bioconductor package was used for the normalization of the gene expression data.
- `affy`
  This Bioconductor package was used for the preprocessing the data obtained from the expression chip of Affymetrix.
- `cluster`
  This package was used to cluster the gene expression data.
- `qqman` (Turner, 2014)
  A user-friendly function to create Manhattan plots.
- `nlme`
  This package was used to calculate linear mixed models.

2.2.3. PLINK

PLINK is a toolset for the analysis of genome-wide associations and the simple handling of large genetic data. It can be downloaded for free from `http://pngu.mgh.harvard.edu/~purcell/plink/`. A variety of questions could be addressed but in this thesis it was mainly used to filter SNPs according to different criteria and to calculate associations between SNPs and gene expression levels.
PLINK is a command line program and needs the genotypes to be in two files.

- MAP-file
  This file contains the chromosome, the rs-number for the SNP, the genetic distance (this is set to zero in the KORA data), and the position of the SNP (in bp units).
- PED-file
  This file is in the same order as the MAP file and contains the genotypes for each sample. The first and the second column contain the family and the sample ID (identical for KORA), the third and fourth column contain the paternal and maternal ID (set to zero

in KORA data), the fifth column contains the sex information (1=male, 2=female), and the sixth column contains a phenotype if necessary (set to -9 in KORA data). The following columns contain the genotypes in two columns per sample indicating the two different alleles.

To calculate the association between any kind of phenotype and the genotypes the following code could be used:

Listing 2.2: Calculation of genome-wide association using PLINK

```
plink
--file genotypes
--assoc
--pheno phenotype.txt
--recode
```

2.2.4. SNAP

SNAP (**SN**P **A**nnotation and **P**roxy Search) is a web-based tool to calculate the linkage disequilibrium (LD) between two SNPs or to search for proxy SNPs[1] (Johnson et al., 2008). The tool can be found on `https://www.broadinstitute.org/mpg/snap/`.
The LD is calculated from real data obtained from the HapMap project.
Proxy SNPs are indicated due to the LD structure, the localization on the genome and the availability on commercial genotype platforms. This tool is very helpful if data from two different genotype platforms should be compared because usually the genotyped SNPs are not identical.

2.2.5. GWAS catalog

The NHGRI GWAS catalog on `https://www.ebi.ac.uk/gwas/` was initially founded by the NHGRI (National Human Genome Research Institute) (Hindorff et al., 2009) and later improved by the European Bioinformatics Institute (EMBL-EBI) (Welter et al., 2014). The aim was to collect the results from all published GWAS and make them available online. On August, 25th 2015 the catalog contains

- 2,269 studies
- 15,020 SNPs
- 16,831 SNP-trait associations

All associations with p-value $< 1.0 * 10^{-5}$ published until May 2014 are shown in Figure 2.2.

[1] A proxy SNP is a SNP that could replace another SNP and is highly correlated with this SNP (correlation coefficient is greater than 0.8)

Figure 2.2.: Published SNP-trait association from the GWAS catalog from May 2014: The distribution of all associations from published GWAS with p-value $< 1.0 * 10^{-5}$ on the chromosomes is shown.

2.2.6. Ingenuity Pathway Analysis Software

Ingenuity pathway analysis software (IPA) is a commercial software and can be downloaded from `http://www.ingenuity.com/`. It is based on an internal library of pathways and is calculating a p-value for each canonical pathway by using the probability that the pathway occurs by chance. Therefore the right-tailed Fisher´s exact test with an optional Benjamini-Hochberg correction is applied.
Lists of genes can be uploaded to identify enriched pathways. The disadvantage of this software is the cost intensiveness but it is justified by the fact that the data base is always up-to-date.

2.2.7. Circos

Circos is a perl-based command line program to create plots that show connections between data points. Because of its circular layout it is optimal to show connections across the genome. The software can be downloaded free of charge from `http://circos.ca/`.

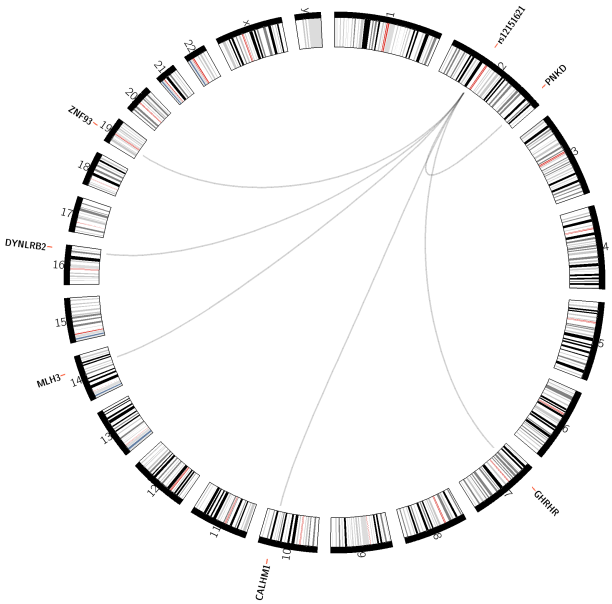

Figure 2.3.: Example circos plot

Figure 2.3 shows an example for a circos plot created with Circos. Here a SNP (rs12151621) is significantly associated with six different genes (*PNKD*, *CALHM1*, *DYNLRB2*, *ZNF93*, *GHRHR*, *MLH3*). The code for creating this figure is as follows:

Listing 2.3: Creating a circos plot

```
<colors>
<<include etc/colors.conf>>
<<include etc/brewer.conf>>
</colors>

<fonts>
<<include etc/fonts.conf>>
</fonts>

<<include ideogram.conf>>
<<include ticks.conf>>

<image>
<<include etc/image.conf>>
</image>
karyotype   = data/karyotype/karyotype.human.hg19.txt
chromosomes_units = 1000000
chromosomes_display_default = yes

# Links (are defined in <links> blocks)
<links>
z             = 0
radius        = 0.975r
bezier_radius = 0.2r

<link segdup>
show          = yes
color         = black_a5
thickness     = 5
file          = links.txt
record_limit = 5000
</link>
</links>

##Adding Probe Gene Labels
<plots>
<plot>
type             = text
color            = black
file             = gene.labels.txt
r0 = 1.07r
r1 = 1.5r
show_links     = yes
link_dims      = 4p,4p,8p,4p,4p
```

```
link_thickness = 4p
link_color     = red
label_size   = 25p
label_font   = condensed
padding  = 0p
rpadding = 0p
</plot>
</rules>
<<include etc/housekeeping.conf>>
restrict_parameter_names* = no
```

The files "etc/colors.conf", "brewer.conf", "fonts.conf", "ideogram.conf", "ticks.conf", "image.conf", "housekeeping.conf" and "data/karyotype/karyotype.human.hg19.txt" are included within the software. The files "links.txt" and "gene.labels.txt" have to be created.

The file "links.txt" contains the information about the two data points that should be connected. The two data points need one common identifier (column 1 in Table 2.1). The second column shows the chromosome, then the start and the end position of the gene or the SNP, respectively (for SNPs the start and end position is identical).

eQTL1	hs2	85934498	85934498
eQTL1	hs2	219187917	219211515
eQTL2	hs2	85934498	85934498
eQTL2	hs10	105213143	105218648
eQTL3	hs2	85934498	85934498
eQTL3	hs16	80574853	80584539
eQTL4	hs2	85934498	85934498
eQTL4	hs19	20011786	20045765
eQTL5	hs2	85934498	85934498
eQTL5	hs7	31003635	31019141
eQTL6	hs2	85934498	85934498
eQTL6	hs14	75480466	75518235

Table 2.1.: Example file for creating a circos plot: "links.txt"

The file "gene.labels.txt" (Table 2.2) contains information about every single data point: chromosome (column 1), start position (column 2), end position (column 3), and the label (column 4).

hs2	85934498	85934498	rs12151621
hs2	219187917	219211515	PNKD
hs10	105213143	105218648	CALHM1
hs16	80574853	80584539	DYNLRB2
hs19	20011786	20045765	ZNF93
hs7	31003635	31019141	GHRHR
hs14	75480466	75518235	MLH3

Table 2.2.: Example file for creating a circos plot: "gene.labels.txt"

After preparing these two files the circos plot will be plotted by typing:

```
perl bin/circos -conf circos_hotspot1.conf outputdir "/..."
-outputfile plot.png
```

2.3. Statistical methods

2.3.1. Linear regression models

Linear models are usually applied to determine the influence of one or more covariables $(x_1, ..., x_k)$ on a response variable (y). Normally the basic model looks like this:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i, \qquad i = 1, .., n$$

where $\epsilon_1, ..\epsilon_n$ are independent and identical distributed (i.i.d) unobservable errors with

$$\mathrm{E}(\epsilon_i) = 0, \qquad \mathrm{Var}(\epsilon_i) = \sigma^2$$

The response variables y and the errors ϵ can be summarized in the vectors y and ϵ and all the covariables in a matrix to get the matrix notation:

$$y = X\beta + \epsilon$$

Now the β_is (summarized in the vector β) can be estimated by the least-square method:

$$\widehat{\beta} = (X^T X)^{-1} X^T y$$

This works only if the matrix $X^T X$ can be inverted, meaning that the inverse A^{-1} of a matrix A exists if and only if A is regular ($|A| \neq 0$) and has the property $AA^{-1} = A^{-1}A = I$.
The result is an unbiased ($E(\widehat{\beta}) = \beta$) estimator with minimal variance. But numerically the calculation of the least-square estimator is inappropriate therefore the QR-decomposition is used. It is the decomposition of a matrix into an orthogonal[2] and a right triangular matrix.

To test whether one of the covariables x_j has an impact on the dependent variable y the following hypotheses are used:

$$H_0 : \beta_j = 0, \qquad H_1 : \beta_j \neq 0$$

If H_0 is true the covariable x_j has no impact on the dependent variable. The test statistic for this hypothesis is

$$T_j = \frac{\widehat{\beta}_j}{\widehat{\sigma}_j}$$

where $\widehat{\sigma_j}$ is the estimated standard derivation of $\widehat{\beta}$. The null hypothesis is rejected if

$$|T_j| > t_{1-\alpha/2}(n - p - 1)$$

In R the command for the linear regression is

[2] A square matrix A is orthogonal if $AA^T = A^T = I$

Listing 2.4: Calculation of a linear regression using R

```
lm(outcome ~ variable1 + variable2 + ...)
```

In the genetic field for the decision of a test problem only the so-called p-values are used (Dastani et al., 2012) . The p-value is the probability of obtaining a value of the test statistic at least as extreme as the one that was actually observed, given that the null hypothesis is true. Because p-values are probabilities the values are always between 0 and 1. So the advantage of using p-values is that comparing different test statistics is possible (Fahrmeir et al., 2003).

When analyzing gene expression data the linear regression is applied when the influence of different phenotypes (for example age, sex, BMI, and so on) on the gene expression levels is investigated. But it is also used to analyze the effect of SNPs on gene expression levels. Here an additive effect is assumed, meaning that the effect of two mutations is twofold in comparison to one mutation (see Figure 2.4). This is called an additive genetic model (Gieger et al., 2008).

If additional to the fixed effects β a random effect should be considered a linear mixed model has to be calculated. An main example for the application is the repeated measurement of the same individual. Here the subject effect is included in the model as a random effect. In R the linear mixed model is implemented in the package `nlme`:

Listing 2.5: Calculation of a linear mixed model with a random subject effect using R

```
library(nlme)
lme(outcome ~ variable1 + variable2 + ..., random = ~ 1|subject)
```

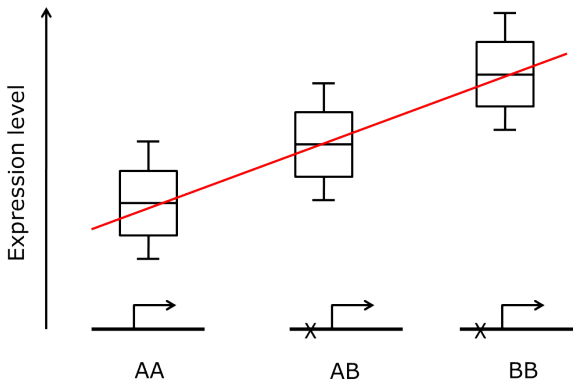

Figure 2.4.: Additive effect of polymorphisms on gene expression levels

If the independent variable is binary, the linear regression is identical to the standard t-test that compares two means from two independent groups (see Section 2.3.3).

2.3.2. Sobel test

The Sobel test is used to calculate if the effect of an independent variable A on the dependent variable B is mediated by a third variable, the so-called mediator (see Figure 2.5).
The test statistic is given by (Sobel, 1982)

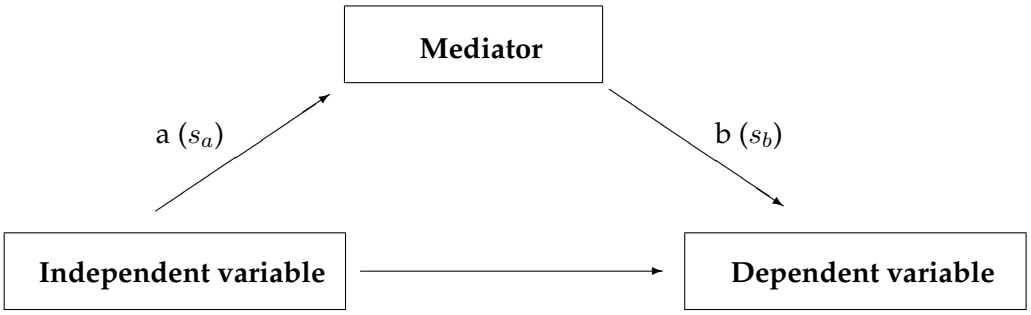

Figure 2.5.: Mediation scheme:
a is the regression coefficient for the association between the independent variable and the mediator and s_a is the standard error of this association. *b* is the coefficient for the association between the mediator and the dependent variable, adjusted for the independent variable and s_b the corresponding standard error.

$$t_{Sobel} = \frac{a * b}{\sqrt{b^2 * s_a^2 + a^2 * s_b^2}}$$

Thereby *a* is the regression coefficient for the association between the independent variable and the mediator and s_a the corresponding standard error of the linear regression model. *b* is the coefficient for the association between the mediator and the dependent variable, adjusted for the independent variable and s_b the corresponding standard error.

When calculating the Sobel test with data from a meta-analysis and using the z-scores the test statistic is

$$Z_{Sobel} = \frac{Z_1 Z_2}{\sqrt{Z_1^2 + Z_2^2}}$$

with Z_1 is equal to *a* and Z_2 is equal to *b*.

In R the Sobel test could be calculated using

Listing 2.6: Calculation of Sobel test using R

```
mediation.test(mediator.var, independent.var, dependent.var)
```

2.3.3. Analysis of variance (ANOVA)

The analysis of variance (ANOVA) is the generalization of the t-test and could be used to compare more than two groups. The test statistic of a t-test is

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{S^2(1/n + 1/m)}}$$

where $S^2 = ((n-1)S_X^2 + (m-1)S_Y^2)/(n+m-2)$ is the variance of the two independent groups. The ANOVA takes also the variance between the groups into account which results in the following test statistic:

$$F = \frac{variance\,between\,groups}{variance\,within\,groups} = \frac{\sum_{i=1}^{I} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2/(I-1)}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})/(n-I)}$$

where I is the number of groups and n_i the number of samples in group i (Fahrmeir et al., 2003). The p-value of the ANOVA only indicates that there is an overall difference between the groups but not between which groups. Therefore it is necessary to compare post-hoc all groups pairwise.

In R the ANOVA could be calculated with the function

Listing 2.7: Calculation of an ANOVA using R

```
aov(outcome ~ group.variable)
```

2.3.4. Fisher´s exact test

The Fisher´s exact test is used to analyze the association between two groups and two different characteristics which can be displayed in the following contingency table:

	group 1	group 2	row sum
trait 1	a	b	a+b
trait 2	c	d	c+d
column sum	a+c	b+d	a+b+c+d=n

If there is no association between the groups the probability for any kind of values is determined by the hypergeometric distribution:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

In R the p-value could be determined using:

Listing 2.8: Calculation of Fisher´s exact test using R

```
fisher.test(rbind( c(a,b), c(c,d) ) )
```

The Fisher´s exact test could be used for example to test if a population is in Hardy-Weinberg-Equilibrium (Wigginton et al., 2005).
A population is in Hardy-Weinberg-Equilibrium if the distribution of the genotypes is the same across several generations (without mutation, migration and with randomly pairing) (Victor, 2007). Assuming two alleles A and B the probability for genotype AA, genotype BB and genotype AB respectively is

$$P(AA) = P(A)^2 \qquad P(BB) = P(B)^2 \qquad P(AB) = 2P(A)P(B)$$

so that

$$P(AA) + P(AB) + P(BB) = P(A)^2 + 2P(A)P(B) + P(B)^2 = [P(A) + P(B)]^2 = 1$$

2.3.5. Multiple testing problem

When analyzing gene expression data, often each gene is analyzed separately and therefore a high number of statistical tests is necessary and performed. Doing *k* independent tests the probability α^* for getting a false positive result is

$$\alpha^* = 1 - (1 - \alpha)^k$$

When choosing an α of 0.05 and doing 1000 tests the chance for a false positive result is

$$\alpha^* = 1 - (1 - 0.05)^{1000} = 1 - 5.29 * 10^{-23} \approx 1$$

To handle the problem of an increasing false positive rate there are different commonly used methods:

- Controlling the family-wise error rate (FWER)
  In genetics mostly the Bonferroni correction is used because it is easy to apply. Either the observed p-value is multiplied by the number of performed tests or the significance level is divided by the number of performed tests.

- Controlling the false discovery rate (FDR)
  This method is less conservative than controlling the FWER. Table 2.3 shows the summary of a multiple testing situation.

	# declared non-significant	# declared significant	$\sum$
# true null hypotheses	U	V	m_0
# non-true null hypotheses	T	S	$m - m_0$
$\sum$	$m - R$	R	m

Table 2.3.: Number of errors committed when testing m null hypotheses (from Benjamini and Hochberg (1995)). Hereby only R is an observable variable, the others are random.

The false discovery rate Q_e is the expectation of the random variable $Q = V/(V + S)$ (proportion of rejected null hypotheses which are wrongly rejected)[3]:

$$Q_e = E(Q) = E\{V/(V + S)\} = E(V/R)$$

In case all null hypotheses are true ($s = 0$ and $v = r$), the FDR equals to the FWER. Here $Q = 0$ if $v = 0$ and $Q = 1$ if $v > 0$. Then you get $P(V \geq 1) = E(Q) = Q_e$. So controlling the FDR means controlling the FWER in a weak sense.
Benjamini and Hochberg (Benjamini and Hochberg, 1995) mentioned the procedure to control the FDR in the following way (Dudoit et al., 2002):
Assuming that it is planned to test m null hypotheses $H_1, H_2, ..., H_m$ with p-values

[3]Definition: if $V + S = 0$, Q is defined to be 0

$p_1, p_2, ..., p_m$, where $p_{(1)} \leq p_{(2)} ... \leq p_{(m)}$ are the ordered p-values and $H_{(j)}$ is the according hypothesis to $p_{(j)}$. To control the FDR (at significance level α) it has to be defined that

$$j^* = \max \left\{ j : p_j \leq \frac{j}{m}\alpha \right\}$$

Then all hypotheses H_j for $j = 1, ...j^*$ will be rejected.
Finally the adjusted equivalent p-values are

$$\widetilde{p}_j = \min_{k=j,...,m} \left\{ \min(\frac{m}{k} p_k, 1) \right\}$$

- Permutation
  Churchill and Doerge (1994) introduced the idea of permutations for microarray experiments. It can be used to estimate the significance level for each gene separately. The disadvantage is that it is very computer-intensive, especially for large sample sizes when the p-values have to be calculated for all possible permutations. The concept for the permutation test is:
  1. Choose any statistical test
  2. Analyze the gene of interest and calculate the test statistic
  3. Permute the samples and calculate the test statistics for each permutation (normally about 1,000 permutations)
  4. Compute the percentage of events where the permuted test statistic is higher than the "real" test statistic to obtain the p-value

2.3.6. Pearson's and Spearman's correlation coefficient

The correlation coefficient is a measurement for the relation between two traits. The Pearson's correlation coefficient is defined by

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

where $-1 \leq r \leq 1$.

This coefficient was used to determine the correlation between two SNPs which are coded by 0, 1, 2 or when using the imputed data[4], the values range between 0 and 2.

The formula of the Spearman's correlation coefficient is identical to those of the Pearson's correlation coefficient but applied on the ranks of the data:

$$r = \frac{\sum_{i=1}^{n}(rg(x_i) - \overline{rg}_X)(rg(y_i) - \overline{rg}_Y)}{\sqrt{\sum_{i=1}^{n}(rg(x_i) - \overline{rg}_X)^2 \sum_{i=1}^{n}(rg(y_i) - \overline{rg}_Y)^2}}$$

where

$$\overline{rg}_X = \frac{1}{n}\sum_{i=1}^{n} rg(x_i) = \frac{1}{n}\sum_{i=1}^{n} i = (n+1)/2$$

[4]When genotype data are imputed the determined SNPs are used to impute the missing SNPs. So for each imputed SNP the probability for each genotype is given.

$$\overline{rg}_Y = \frac{1}{n}\sum_{i=1}^{n} rg(y_i) = \frac{1}{n}\sum_{i=1}^{n} i = (n+1)/2$$

The Spearman's rank correlation test is based on the correlation coefficient and is a non-parametric alternative to the linear regression. In R it could be calculated by

Listing 2.9: Calculating Spearman's rank correlation test using R

```
cor.test(x,y, method = ''Spearman'')
```

2.3.7. Agglomerative clustering

The aim of clustering gene expression data is to identify outliers that show a different expression pattern than all other samples. Here the agglomerative clustering was used. This means that the number of clusters is not predefined and in the beginning each sample belongs to one small cluster. This small clusters were added together until no further similarities between two clusters could be identified. The distance between two samples could be calculated in different ways. The most commonly used distance is the Euclidean distance:

$$dist(x,y) = \sqrt{(x_1 - y_1)^2 + ... + (x_n - y_n)^2}$$

The result of the clustering is normally shown in a dendrogram (Rahnenführer, 2004) to visualize the following information:

- Associated clusters are next to each other and connected by a line.
- The length of the connecting line is the average distance between the observations in the two clusters.
- The variation in the clusters on the left side is smaller than the variation in the clusters on the right side of the dendrogram.

The R package `cluster` and more precisely the function `agnes()` with default settings was used to cluster all data.

2.3.8. Principal component analysis

The principal component analysis (PCA) is used to reduce the dimension of large data sets without loosing information. Therefore a set of unrelated and orthogonal variables, the so-called principle components (PC), were determined. Principal components are linear combinations of the original variables. The first PC is explaining most of the variance of the original data. The remaining PCs are explaining less variance in a descending order. For genetic data the PCA (especially the first two PCSs) is sometimes used to identify outliers or clusters in the data.

If $\mathbf{Y}^t = (Y_1, ..., Y_m)$ is a m-dimensional vector of random variables with expectation μ and covariance matrix Σ (for expression data Y_1 represents the expression values of probe i and so on) the aim is to identify new uncorrelated variables $Z_1, ..., Z_m$ in which

$$Z_j = a_{1j}Y_1 + a_{2j}Y_2 + ... + a_{mj}Y_m = \mathbf{a}_j^t\mathbf{Y},$$

where $a_j^t = (a_{1j}, a_{2j}, ..., a_{mj})$ is a vector of constant variables. To avoid arbitrary scales it is normalized to $a_j^t a_j = \sum_{k=1}^{m} a_{kj}^2 = 1$. The most important condition is the maximization of $Var(Z_j) = Var(a_j^t X)$.
The so-called loadings a_{ij} are determined by applying a singular value decomposition where **X** is decomposed to

$$\mathbf{Y} = \mathbf{UDV}^T$$

where the matrices **U** and **V** are column orthogonal so that $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$ and **D** is a diagonal matrix.

The PCA is implemented in the R function `prcomp()`. For all analyses in this thesis the following setups which returns the Eigen-vectors (so-called Eigen-genes when working with gene expression data) were used :

Listing 2.10: Calculating a PCA using R

```
pc <- prcomp(data, retx = TRUE, center = TRUE, scale. = TRUE)
ev <- as.data.frame(pc$rotation)
```

- retx = TRUE
  The rotated variables are returned.
- center = TRUE
  The variables are shifted to be zero centered.
- scale. = TRUE
  The variables are scaled to have unit variance before the analysis.

The principal component analysis was used to determine the influence of technical variables on the gene expression levels and to reduce the variance in the data by adjusting for a certain number of Eigen-genes. It is also the basic principle for the Eigen-R^2 algorithm (Chen and Storey, 2008). The Eigen-R^2 can be calculated for each kind of variable to determine the proportion of variance that is explained by this variable. This algorithm consists of four steps:

1. **Y** is an *m* x *n* expression matrix with *m* probes and *n* samples. With the singular value decomposition this matrix is decomposed to $Y = UDV^T$. **U** and **V** are column orthogonal and **D** is a diagonal matrix.
2. Each column of **V** is denoted by v_i and the user can specify a model of v_i on the independent variable **z** for getting fitted values $\widehat{v}_i$, where i = 1,2,...,n. Then the R^2 is calculated with the following formula:

$$R^2_{\widehat{v}_i} = \frac{\widehat{\sigma}^2_{\widehat{v}_i}}{\widehat{\sigma}^2_{v_i}} = \frac{\sum_{j=1}^{n}(\widehat{v}_{ij} - \overline{\widehat{v}_i})^2}{\sum_{j=1}^{n}(v_{ij} - \overline{v}_i)^2}$$

3. The proportion of variation that is explained by v_i is calculated with the formula

$$\pi_i = \frac{d_i^2}{\sum_{l=1}^{L} d_l^2}$$

   Here d_i represent the Eigen-value of the *i*-th Eigen-vector.

4. At last the overall Eigen-R^2 is calculated by:

$$Eigen - R^2 = \sum_{i=1}^{n} \pi_i R^2_{\widehat{v_i}}$$

The algorithm is implemented in the R-package `eigenR2` and was originally available on the Biconductor homepage. Now it can only be downloaded from the author´s homepage: `www.genomine.org/eigenr2/`. The default settings were used which means that the models were fitted by least squares.

2.4. Genotyping of KORA F3 and F4 samples

The genotypes of KORA F3 and F4 were determined using so called SNP chips, a microarray-based technology. The genotyping of KORA F3 was performed using the Affymetrix 500K array set while KORA F4 was performed using the Affymetrix 6.0 chip.
The genotype data were quality controlled and imputed using IMPUTE (Howie et al., 2009) in the Institute of Epidemiology at the Helmholtz Center Munich.

2.4.1. Filtering of SNPs in KORA F3

The SNPs were filtered using the same criteria as Döring et al. (2008). In total there were 500,568 SNPs and PLINK was used to filter out all SNPs that pass the following criteria:

1. Not on X-chromosome: 490,032 autosomal SNPs are remaining.
2. Genotyping efficiency $< 95\%$: 49,325 SNPs have more than 5% missing values.
3. Minor allele frequency $< 5\%$: 10,1323 SNPs were excluded because the frequency of the less frequent allele is less than 5%.
4. Deviation from Hardy Weinberg equilibrium tested with Fisher´s exact test: 4,232 SNPs were excluded.

All together 335,152 SNPs were selected for the subsequent analysis.

The used PLINK code was:

Listing 2.11: Filtering of SNPs using PLINK

```
--file KORA_Genotypes
      --geno 0.05
      --maf 0.05
      --hwe 0.000001
      --fisher
      --model
      --recode
      --out KORA_Genotypes_snps_filtered
```

2.4.2. SNP selection in KORA F4

Genotype data of 993 samples from KORA S4 and F4 are available. PLINK was used to filter the SNPs according to common filter criteria that were also used for the KORA F3 data:

- Starting with 692,637 SNPs
- 136 markers to be excluded based on HWE test ($p \leq 1e^{-6}$)
- 56 SNPs failed missingness test ($GENO > 0.05$)
- 75,555 SNPs failed frequency test ($MAF < 0.05$)
- Resulted in 616,941 SNPs

The positions of the SNPs were checked in a database and updated.
The major and minor alleles and the minor allele frequency of each SNP were determined with the PLINK command `--freq`.
SNPs with more than one hit in the database or with hits on the mitochondrial-chromosome (n = 26) were excluded from further analysis (all together n = 2,790 SNPs).

2.5. Measuring of gene expression levels

The development of microarrays allows to measure the expression levels of thousands genes simultaneously. The two largest providers of these arrays are Illumina and Affymetrix, two American companies that develop and sell products for the analysis of genetic information. With only one exception the expression arrays of Illumina are used and therefore the work flow for Illumina expression data is described here more in detail.

2.5.1. Experimental protocol for measuring gene expression using Illumina arrays

The blood was collected in PAX tubes at the KORA study center in Augsburg and immediately transported to the Institute of Human Genetics of the Helmholtz Center in Munich. The PAX tubes were stored overnight at room temperature according to the manufacturers instructions and then further stored at 4°C until required.

The quality of the RNA was measured with the Agilent Bioanalyzer. The Bioanalyzer is an automated bio-analytical device to perform the quality control of DNA and RNA samples (Mueller et al., 2000). The most informative output is the RNA integrity number (RIN) (Schroeder et al., 2006) which ranges from 1-10. A RIN number of 1 indicates that the RNA is destroyed by enzymes (degraded RNA) while a RIN number of 10 indicates an intact RNA sample.
The amount of RNA was determined using the Invitrogen Ribogreen kit.

The RNA obtained from whole blood is usually not enough for a microarray experiment and furthermore it is not labeled. Therefore, a step of amplification combined with reverse transcription and labeling with Biotin is required before the sample can be processed on the

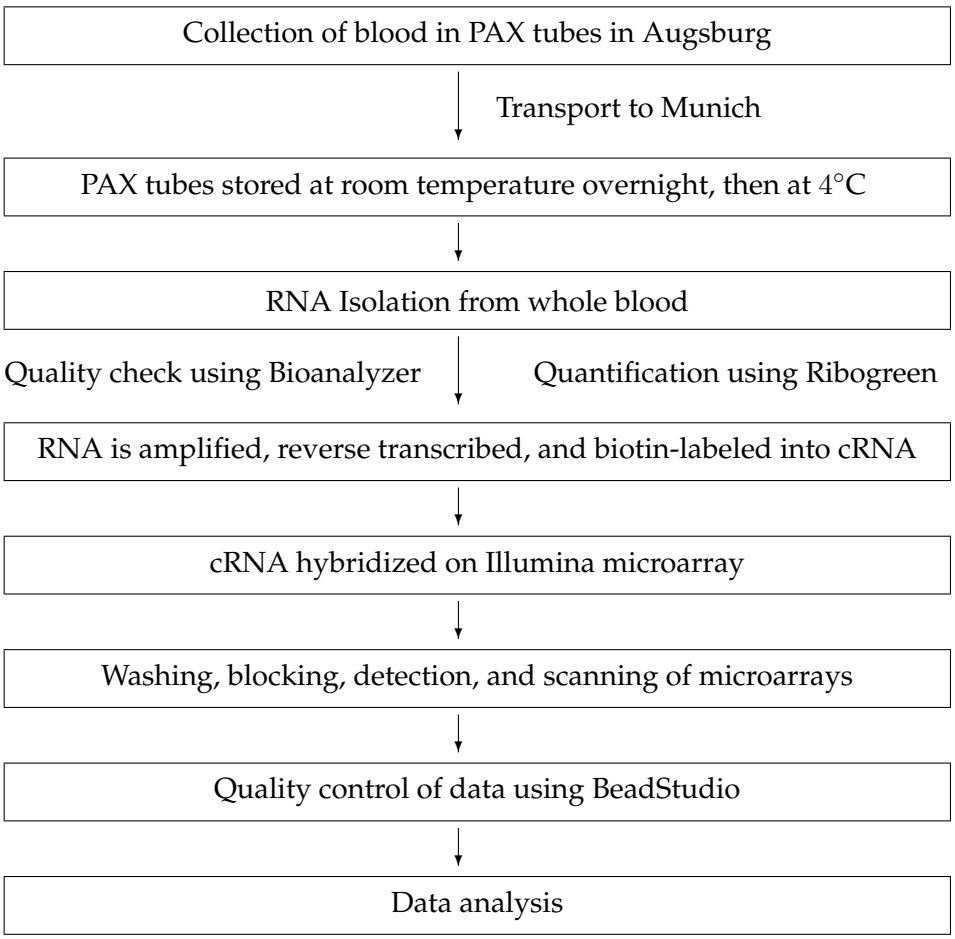

Figure 2.6.: Experimental workflow for measuring gene expression using an Illumina expression array in KORA F3

microarray. The Illumina Total Prep RNA amplification kit was used to generate biotinylated, amplified cRNA out of $500ng$ RNA for hybridization with Illumina arrays.

The Illumina system uses a direct hybridization assay, whereby gene-specific probes are used to detect labeled RNA. Each bead in the array contains a sequence-specific probe of 50 nucleotides. Illumina offers three whole-genome formats: 6-sample (used for KORA F3), 8-sample and 12-sample (used for KORA F4). Each array in the matrix consists of thousands to tens of thousands of different oligonucleotide (short, single-stranded parts of RNA) probe sequences. Multiple copies of each bead type are present in the array, on average ≈ 30 copies per probe.
In KORA F3 $1500ng$ and in KORA F4 $3000ng$ of cRNA was used for the hybridization on the Illumina HumanWG-6 v2 (KORA F3) or the Illumina HumanHT-12 v3 (KORA S4 and F4) Expression BeadChip.

The Illumina Bead Array reader was used to image the Bead chips. After scanning, the raw data was imported from the Illumina BeadStudio (used for KORA F3 data) software which was replaced in 2009 by the GenomeStudio (used for KORA F4 data).

The main difference between KORA F3 and F4 was that the samples in KORA F3 were used directly after the blood draw. The samples were transported immediately to the Helmholtz Center and the RNA was isolated. For the S4 and F4 samples the blood was also collected in PAX tubes, but then they were frozen until the RNA isolation. The S4 samples were partly frozen up to ten years. Another difference is that the F3 samples were processed individually however the F4 samples were processed in groups of 96.

2.5.2. Analyzing gene expression data using Affymetrix arrays

Blood samples (19ml of peripheral blood) were taken from patients and their healthy siblings under fasting conditions after informed consent. 2.5ml of the blood were collected directly in PAXgene Blood RNA tubes (PreAnalytiX) and were stored for six months at $-70°$C. The RNA extraction was done using the PAXgene Blood RNA Kit (Qiagen).
1μg of RNA of each sample was used to reduce the globin with the Ambion GLOBINclearTM Kit. The following steps were done with both probes of each sample.
RNA and cRNA quality control was carried out using the Bioanalyzer (Agilent) and quantification using Ribogreen (Invitrogen). The concentration was measured with the NanoDrop. 200ng of RNA were reverse transcribed into cDNA and biotin-UTP-labeled the RNA using the GeneChip WT Terminal Labeling Kit from Ambion.

This was the only project where the cRNA was hybridized to the Affymetrix GeneChip Human Gene 1.0 ST Array. Washing steps were carried out in accordance with the Affymetrix protocol. The quality of the expression data was checked with the Affymetrix Software Expression Console and due to bad quality of the RNA two samples were removed from further analysis and 19 samples passed all quality criteria. The statistical analysis was done using the Bioconductor package `affy`. The expression values were calculated with the function `rma` (Irizarry et al., 2003) and were by default logarithmized and normalized with the quantile normalization (see Figures 2.7 and 2.8). The R code for the transformation of the raw

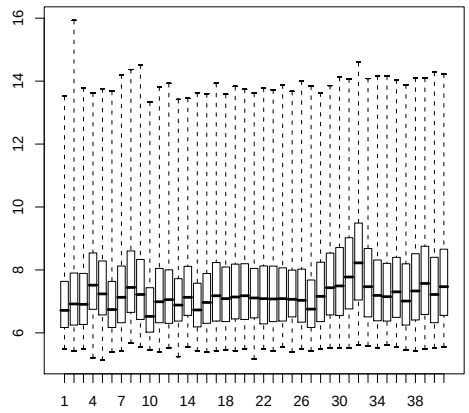

Figure 2.7.: Not normalized samples

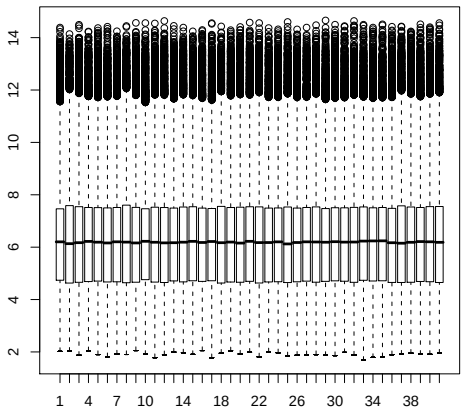

Figure 2.8.: Normalized samples

expression data to a data frame with normalized values with the affy package is as follows:

Listing 2.12: R code for using the Bioconductor package `affy`

```
library(affy)
pd <- read.AnnotatedDataFrame("Info.txt", header=T, row.names=1)
raw.data <- ReadAffy(filenames=rownames(pData(pd)), phenoData=pd)
rma.data <- rma(raw.data)
expr.data <- exprs(rma.data)
```

2.5.3. Normalization of microarry data

Microarray experiments are very susceptible to small changes in the protocol. Therefore the optimal way would be to perform all measurements on the same day by the same person under exactly the same conditions. Of course this is normally not possible especially for large sample sizes and even if samples are processed on the same day on the same amplification plate there might be differences in the signal intensities (see Figure 2.9).
One possibility to ensure the comparability of several samples is to normalize the expression values before further analyses.

2.5.3.1. LOESS normalization

Cleveland (1979) introduced the LOESS normalization which is the abbreviation for LOcal regrESSion.
The non-parametric local regression is applied on logarithmized and MA-transformed data (Dudoit et al., 2002) where M and A are calculated by:

$$M = sample1 - sample2$$

$$A = 1/2(sample1 + sample2)$$

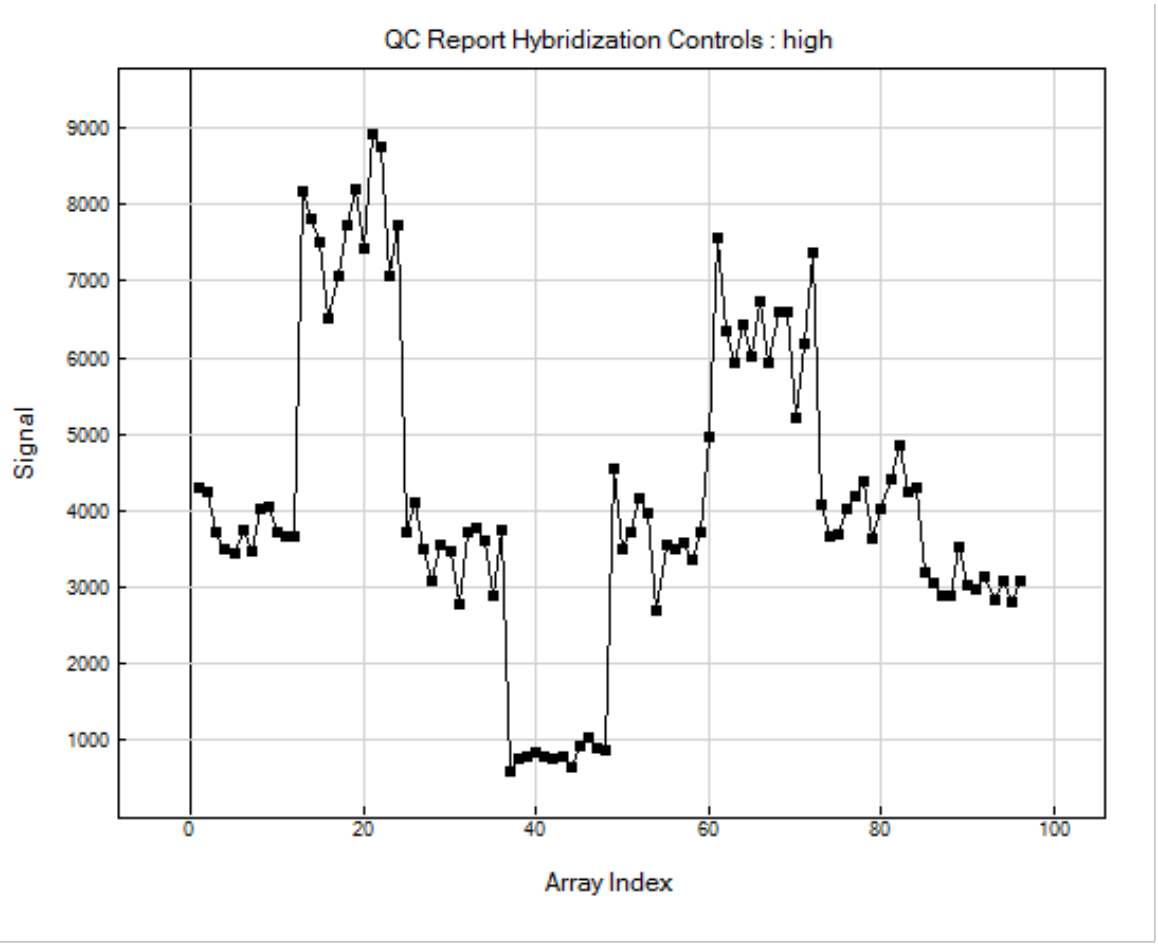

Figure 2.9.: Quality plot to illustrate the necessity of normalization:
The quality plot from the GenomeStudio shows eight arrays that were processed at the same time on the same amplification plate. Nevertheless, the eight different arrays differs substantially when comparing the average signal of the hybridization controls.

It is assumed that the response variable y_i is explained by a smooth function $f(z_i)+\epsilon_i$ (where $i = 1, ..., n$). To identify the $f(z)$ the following algorithm is used (Fahrmeir et al., 2007):

1. The k nearest neighbors of z are defined by $N(z)$ where the neighborhood $N(z) = i : d_i \in d_{(1)}, ..., d_{(k)}$ and $d_{(1)}, ...d_{(k)}$ are the ordered distances $d_i = |z_i - z|$
2. Define the largest distance of two data points by

$$\Delta(z) = \max_{i,j \in N(z)} |z_i - z_j|$$

3. Weights are defined by

$$w_{\Delta(z)}(z, z_i) = K(\frac{|z - z_i|}{\Delta(z)}),$$

   where K is a tricube core function

$$K(u) = \begin{cases} (1 - |u|^3)^3, & \text{if } |x| < 1 \\ 0, & \text{otherwise} \end{cases}$$

4. $\hat{f}(z)$ is calculated by (weighted) least squares based on the data points in the neighborhood $N(z)$.

The normalization was performed using the R function `loess()` with default settings. The smoothing parameter which defines the size of the neighborhood was set to 0.1. The disadvantage of this normalization is that one sample has to be chosen as reference sample. Depending which sample is chosen the results could differ appreciably.

2.5.3.2. Quantile normalization

The aim of the quantile normalization is to equalize the distribution of each sample (Bolstad et al., 2003).
The concept of the normalization is shown here with random numbers:

1. Start with the raw data:

$$\begin{array}{c} \\ Gene1 \\ Gene2 \\ Gene3 \end{array} \begin{array}{c} \begin{array}{ccc} Sample1 & Sample2 & Sample3 \end{array} \\ \begin{pmatrix} 10 & 300 & 1100 \\ 100 & 200 & 15 \\ 1000 & 5 & 300 \end{pmatrix} \end{array}$$

2. Order the data within each sample and calculate the mean for each row:

$$\begin{array}{c} \\ Quantile1 \\ Quantile2 \\ Quantile3 \end{array} \begin{array}{c} \begin{array}{cccc} Sample1 & Sample2 & Sample3 & Mean \end{array} \\ \begin{pmatrix} 10 & 5 & 15 & 10 \\ 100 & 200 & 300 & 200 \\ 1000 & 300 & 1100 & 800 \end{pmatrix} \end{array}$$

3. Replace the raw data by the mean:

$$\begin{array}{c} \\ Quantile1 \\ Quantile2 \\ Quantile3 \end{array} \begin{array}{c} \begin{array}{cccc} Sample1 & Sample2 & Sample3 & Mean \end{array} \\ \begin{pmatrix} 10 & 10 & 10 & 10 \\ 200 & 200 & 200 & 200 \\ 800 & 800 & 800 & 800 \end{pmatrix} \end{array}$$

4. Restore the order of the original data:

$$\begin{array}{c} \\ Gene1 \\ Gene2 \\ Gene3 \end{array} \begin{array}{c} \begin{array}{ccc} Sample1 & Sample2 & Sample3 \end{array} \\ \begin{pmatrix} 10 & 800 & 800 \\ 200 & 200 & 10 \\ 800 & 10 & 200 \end{pmatrix} \end{array}$$

In R the quantile normalization is implemented in the package `lumi` which was developed for the analysis of Illumina data but could also be used for any kind of expression data. The normalization function is called `lumiN()` where different normalization methods can be applied. The quantile normalization is performed by using

Listing 2.13: Normalize expression data using R

```
lumiN(data, method = ''quantile'')
```

2.5.4. Preparation of gene expression data for eQTL studies

2.5.4.1. KORA F3

Altogether there are 48,701 probes on the Illumina HumanWG-6 v2 expression array. For the first analyses the number of probes was reduced by using only these that are significant detected in more than 5% of the samples. In KORA F3 this results in 13,767 probes.
Later it was decided to keep also probes with low expression levels because replication cohorts were available to verify also effects in low-expressed probes. Therefore the 50bp sequences of each probe were mapped to the human reference sequence hg18 and all probes were kept that mapped uniquely or had only up to two mismatches per probe. 7,292 probes did not map uniquely, so 41,409 probes remained and were used for analysis. Of these probes, 27,623 mapped to annotated transcripts and 13,786 mapped to intergenic regions.

2.5.4.2. KORA F4

On the Illumina HumanHT-12 v3 array are 48,803 expression probes. For the analyses the probes were mapped by Alexander Teumer to the available mRNA sequences of the UCSC genome annotation database (hg19). 28,691 probes could be perfectly mapped to an unique mRNA or to known transcripts or annotated RefSeq genes. These probes map to 18,606 different RefSeq genes.
For further analyses probes that could not be mapped or mapped to intergenic regions were removed.

For the eQTL analysis it was decided to reduce the technical variance in the data by adjusting for a determined number of Eigen-genes (this is described in detail in Section 4.3).

First, a principal component analysis was conducted and different numbers of Eigen-genes (five to 100 in steps of five) were removed from the data by keeping the residuals in a linear model with expression as dependent and Eigen-genes as independent variables[5]. Additionally the uncorrected and for age and gender corrected data were used.
PLINK was used to test systematically the association of all SNP-probe combinations for all different data sets. PLINK calculates linear regression models with additive effects of SNPs where the direction of the regression coefficient represents the effect of each extra minor allele (i.e. a positive regression coefficient means that the minor allele increases the expression levels). Due to a limitation of memory capacity only combinations with a p-value below $1 * 10^{-7}$ were stored.

Listing 2.14: Calculation of association between gene expression levels and genotypes using PLINK

```
--file KORA_Genotypes_snps_filtered
--assoc
--pheno expression_adjusted.txt
--all-pheno
--pfilter 1e-7
```

The number of significant probe-SNP-combinations and transcripts (p-value threshold = $6.02 * 10^{-9}$ and $2.81 * 10^{-12}$ for *cis*- and *trans*-eQTL, respectively) were plotted (Figures 2.10 and 2.11). Also mean standard error, beta, and explained variance R^2. The optimal number of Eigen-genes was graphically determined.

For the *cis*-analysis the highest number of significant eQTLs was observed when correcting for 55 Eigen-genes. Correcting for more Eigen-genes led only to marginal more significant hits and the mean standard error did not get smaller any longer.
For the *trans*-analysis the most hits were obtained when correcting for 25 Eigen-genes. Using more Eigen-genes even worsened the result. These results were similar to the results from Fehrmann et al. (2011). They removed 50 Eigen-genes for *cis*- and 25 Eigen-genes for *trans*-analysis, respectively.

2.6. Comparison of *cis*-eQTL results in KORA F4 with published *cis*-eQTLs

We compared the *cis*-eQTLs from whole blood with already published *cis*-eQTLs in different tissues. The following publications were considered:

1. Fairfax et al. (2012) - eQTLs in monocytes and b-cells
   Total RNA from monocytes and b-cells from 283 healthy volunteer was quantified using the Illumina HumanHT-12 v4 BeadChip. eQTLs were calculated for 29,022 expression probes and 651,210 SNPs (from Illumina Human OmniExpress-12v1.0 BeadChips) using linear and Spearman´s rank models (because of little differences in both models

[5] In the original publication the Eigen-genes were misleadingly called principle components. That is the reason for slightly different figures in the original publication and this thesis.

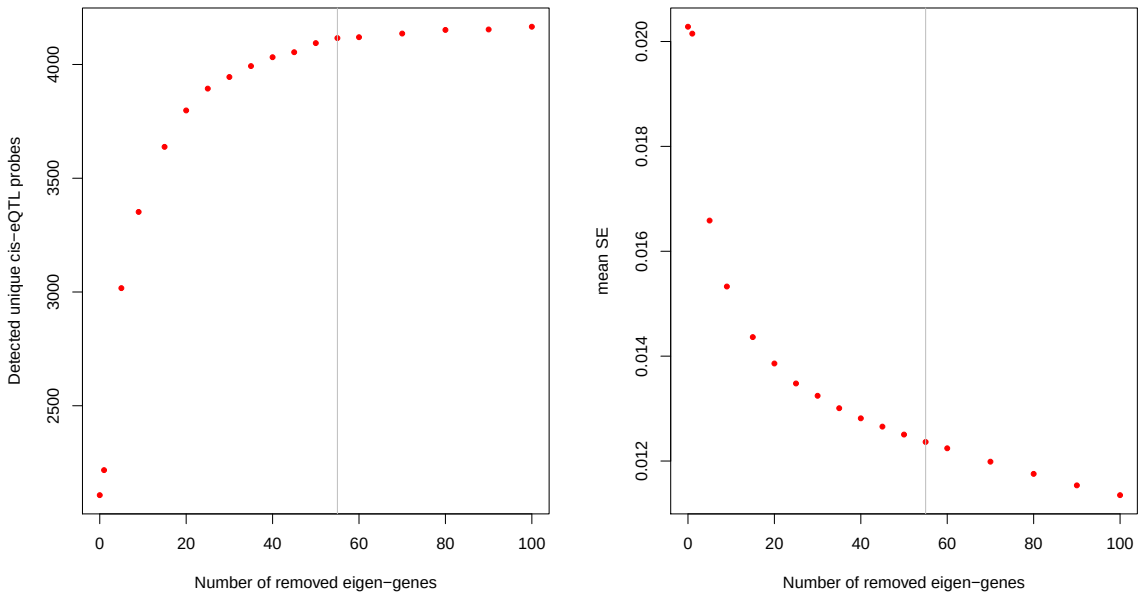

Figure 2.10.: Number of unique significant *cis*-eQTL probes for different numbers of removed Eigen-genes in KORA F4:
The expression levels were adjusted for different numbers of Eigen-genes and the association between adjusted expression levels and genotypes were calculated. The number of unique significant *cis*-eQTL with p-value $< 6.02 * 10^{-9}$ is plotted.

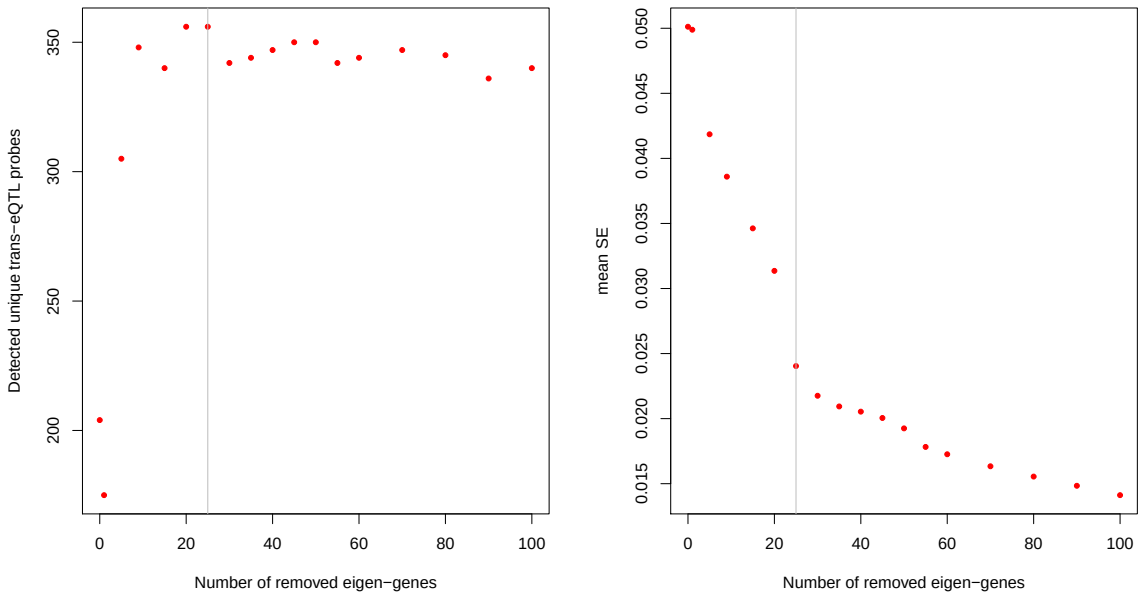

Figure 2.11.: Number of unique significant *trans*-eQTL probes for different numbers of removed Eigen-genes in KORA F4:
The expression levels were adjusted for different numbers of Eigen-genes and the association between adjusted expression levels and genotypes were calculated. The number of unique significant *trans*-eQTL with p-value $< 2.81 * 10^{-12}$ is plotted.

only eQTLs that reached threshold in both analyses were carried forward). *cis* was defined as a 2.5 Mb-interval on either side of probe with a p-value threshold of $1 * 10^{-3}$. To compare the results Table S1 from the original publication was downloaded which consists of 82,346 SNP-probe interactions.

2. Fehrmann et al. (2011) - eQTLs in whole blood
   RNA from whole blood from 1,469 European samples was investigated. Illumina Human HT-12 and H8v2 arrays were used which resulted in 52,061 unique probes were analyzed, representing 19,609 unique genes. Additionally 289,044 common SNPs (Illumina HumanHap300 platform) were analyzed. eQTLs were calculated using the non-parametric Spearman´s rank correlation. The FDR was controlled at 0.05 by permuting expression phenotypes 100 times. An eQTL was defined as *cis*-eQTL if the distance from probe midpoint to the SNP was $\leq$ 250 kb. Supplement Table S1 with 65,535 entries was downloaded to compare the results.

3. Zeller et al. (2010) - eQTLs in monocytes
   Monocyte cells from 1,490 European samples were analyzed using 675,350 SNPs from the Affymetrix Genome-Wide Human SNP Array 6.0 and 12,808 well characterized detected genes from the Illumina HT-12 v3 BeadChip. The associations were calculated using the ANOVA and were checked with the Kruskal-Wallis test[6]. For 2,477 genes a *cis*-eQTL (distance between SNP and gene is less than 1 Mb) with a p-value less than $5.78 * 10^{-12}$ (Bonferroni threshold) could be observed and they are summarized in Supplement File S1.

4. Schadt et al. (2008) - eQTLs in liver
   Liver samples of 427 Caucasian samples were analyzed using an Agilent expression array which consists of 34,266 known and predicted genes of which 39,280 probes. The Kruskal-Wallis test was used for analyzing association between expression traits and 782,476 SNPs (Affymetrix 500K genotyping array plus Illumina 650Y panel) and the FDR (10%) was used for multiple testing correction. *cis*-eQTLs were defined as interactions where the distance between probe and SNP is $\leq$ 1 Mb. As the comparison of *cis*-eQTLs was already done by Zeller et al. (2010) the table was downloaded from Zeller et al. (2010) (Supplement File S4).

5. Stranger et al. (2007b) - eQTLs in LCLs
   The Eppstein-Barr virus-transformed lymphoblastoid cell lines of 270 individuals from HapMap consortium (30 Caucasian trios of northern and western European origin, 45 unrelated Chinese individuals from Beijing, 45 unrelated Japanese individuals from Tokyo, 30 Yoruba trios from Ibadan, Nigeria) were anaylzed using 14,925 expression probes from the Illumina WG-6 v1 BeadChip and about 290,000 SNPs (HapMap). The Spearman's Rank Correlation test was performed. *cis*-eQTLs were defined as the distance from probe genomic midpoint to SNP genomic location was $\leq 1Mb$ and the significance threshold was determined using 10,000 permutations of expression phenotypes. As the comparison of *cis*-eQTLs was already done by Zeller et al. (2010) the table was downloaded from Zeller et al. (2010) (Supplement File S2).

6. Innocenti et al. (2011) - eQTLs in liver
   The study consists of three independent sample collections:

[6]The Kruskal-Wallis test is the non-parametric version of the ANOVA. It considers the ranks of the values.

- discovery set: primary liver tissues at University of Chicago (n=206)
- replication set: primary liver tissues at University of Washington (n=60)
- replication set: published set of liver eQTL data (Schadt; n= 266)

The genotyping was performed on Illumina SNP arrays (Illumina quad-610 and 55k, consisting of more than 500,000 SNPs) and gene expression was measured on Agilent and Illumina expression arrays (14,703 genes were surveyed in the reference study, 11,245 RefSeq genes in all three studies). An eQTL was indicated as *cis*-eQTL when the distance between SNP and TSS was less than 250 kb and eQTL were calculated using a Bayesian regression. Supplement Table S1 was downloaded with eQTLs for all genes. According to the publication all genes with Bayes Factors > 5 (1,173 genes) were selected for comparison with KORA eQTLs.

7. Sasayama et al. (2013) - eQTLs in whole blood - Japanese samples
RNA from whole blood from 76 Japanese samples was analyzed using 534,404 autosomal SNPs (Illumina HumanOmnil-Quad BeadChip) and 30,465 expression probes (Agilent Human Genome 4x44 K array). The Spearman´s rank correlation test was used to calculate the association between gene expression and genotype and an eQTL was indicated as *cis*-eQTL when the SNP was within 1 Mb upstream or downstream of the gene. The significance threshold was $3.1 * 10^{-12}$ (Bonferroni correction). Supplement Table S4 was downloaded which includes 3,883 SNP-probe pairs.

8. Hao et al. (2012) - eQTLs in lung
Lung samples from 1,111 individuals (409 samples from Laval, 363 samples from Groningen, 339 samples from UBC) were analyzed using 51,627 expression probes (custom Affymetrix array) and all SNPs from the Illumina Human1M-Duo BeadChip array. A *cis*-eQTL was defined as an association between an expression probe and the SNP in which the SNP is located within 1 Mb distance of the probe. eQTLs were calculated using the Kruskal-Wallis-Test with applying a FDR correction of 10%. Results were downloaded from Supplement Table S2a which consists of 17,049 entries.

9. Dixon et al. (2007) - eQTLs in LCLs
The Eppstein-Barr virus-transformed lymphoblastoid cell lines from 400 children from families recruited through a proband with asthma were collected. 408,273 SNPs (Illumina Sentrix Human-1 Genotyping BeadChip and Illumina HumanHap300 Genotyping BeadChip) and 54,675 transcripts (Affymetrix U144 Plus 2.0 GeneChip) were analyzed. Associations were calculated using a linear model with a Bonferroni correction which leads to a significance threshold of $1.2 * 10^{-7}$. The maximum distance for *cis*-eQTLs was 100 kb upstream or downstream of the gene. As the comparison of *cis*-eQTLs was already done by Zeller et al. (2010) the table was downloaded from Zeller et al. (2010) (Supplement File S3).

10. Göring et al. (2007) - eQTLs in lymphocytes
Lymphocytes of 1,240 samples from San Antonio Family Heart Study using 20,413 expression probes (Illumina WG-6 v1 BeadChip) and 432 SNPs from the Human MapPairs Genome-Wide Screening Set Version 6 and 8 from Research Genetics were analyzed. A linear model was used to test for association between expression levels and SNPs and an FDR of 5%. As the comparison of *cis*-eQTLs was already done by Zeller et al. (2010) the table was downloaded from Zeller et al. (2010) (Supplement File S5).

3. Analysis of gene expression data in case-control studies

3.1. Introduction to case-control studies

Case-control studies are a type of observational studies in which samples from two different groups are compared. The "cases" could be patients suffering from a disease or individuals with a particular condition (for example high BMI), whereas the "controls" are the individuals that are healthy (or are at least not affected by the disease of the cases) or do not have the same conditions. In comparison to population-based studies in which the general population is investigated the case-control studies have lower sample sizes and the controls should be accurately chosen to ensure identification of even very small differences between the two groups.
The advantage of case-control studies is that it is possible to study rare diseases because the patients are already affected by the disease when they are included in the study. The better the controls fit to the cases (for example in age, gender and other sociodemographic variables) the higher the power to detect also small differences even when having small sample sizes.
Case-control studies in which genome-wide gene expression levels should be compared between two or more groups have to be designed carefully. As the expected effects between the groups and the sample sizes are usually quite small the selection of cases and controls have to be quite homogenous. Additionally it has to be considered that some samples have to be excluded due to bad quality. But the advantage is that it is possible to measure the gene expression levels in the affected tissues, at least for the patients. If it is not possible to obtain the right tissue from the patients or due to ethical reasons it is not possible to get the relevant tissue of the controls the expression is measured in whole blood because it is easily accessible and cheap.
In the following section two different case-control studies are presented (see Table 3.1). In both studies, patients with a neurodegenerative disorder are compared. In the first study patients with Parkinson's disease are compared to young and old controls and in the second study patients with two different subgroups of NBIA are compared to controls. In both studies gene expression was measured genome-wide using two different expression chips from Illumina and Affymetrix respectively. For the Parkinson study it was possible to use the affected tissue while for the NBIA project whole blood was used.

	Parkinson study	NBIA study
Controls	young and old controls	one control group
Patients	Parkinson patients	two different subgroups of NBIA
Analyzed tissue	single cells from brain	whole blood
Used expression platform	Illumina	Affymetrix

Table 3.1.: Comparison of two case-control studies

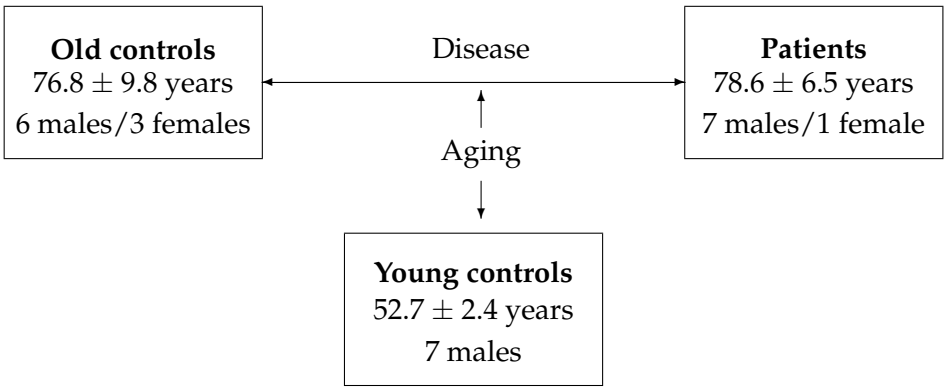

Figure 3.1.: Experimental design of Parkinson study.
The age is depicted as mean ± standard deviation.

3.2. Neurodegeneration and aging

The aim of the study was the identification of differentially expressed genes between patients and controls using genome-wide expression data obtained from single cells from brain that were isolated postmortem by Dr. Matthias Elstner in Newcastle, UK. Additionally the effect of aging should be analyzed by also comparing the patients to a second control group consisting of younger controls, as well as both control groups, one against each other.

3.2.1. Parkinson‘s disease and aging

Parkinson‘s disease (PD) is a neurodegenerative disease and starts with problems of the movement: shaking, slowness of movement, rigidity and difficulty with walking. It is assumed that the reason for Parkinson‘s disease is an interaction of genetic and environmental factors (Sulzer, 2007). The trigger for the disease is the degeneration of dopaminergic neurons of the substantia nigra. This is a normal process of aging but in patients with Parkinson’s disease this process is accelerated. The reason for the massive loss of the dopaminergic neurons is so far not clarified. A multi-causal development seems likely with genetic and environmental factors or severe medical conditions (like stroke or tumors).
To identify differentially expressed genes altogether three groups were analyzed: Parkinson patients, age-matched controls and younger controls (see Figure 3.1). The young controls were used to detect differences between normal aging processes and processes that could be explained by the disease.

As most changes in gene expression levels should occur in the brain of the patients neurons were extracted from the substantia nigra and RNA was isolated from 100 neurons per sample. The patients were clinically well-documented Parkinson patients and the age-

matched controls are without any neurological disease. In Figure 3.2 the experimental design from case selection over RNA isolation to the data analysis is shown.

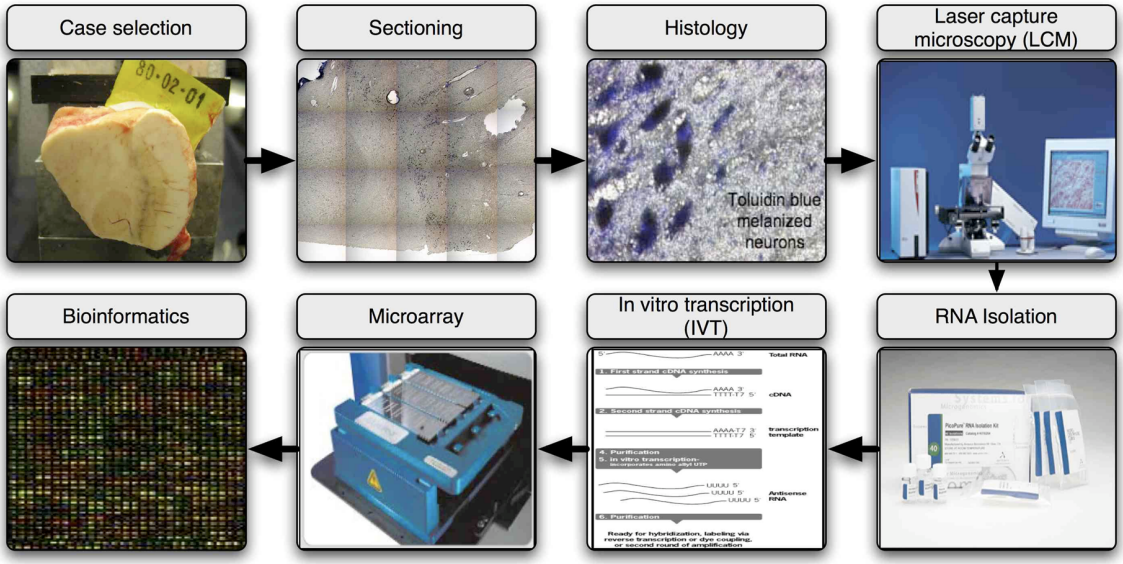

Figure 3.2.: Workflow for genome-wide expression profiles from single cells:
The frozen midbrains were sectioned at 20μm thickness, neurons were identified by toluidine stain and neuromelanin pigment, RNA was extracted from 100 neurons, two rounds of In vitro transcription yielded in $> 3\mu g$ cRNA and was used for hybridization on Illumina WG-6 v1 expression chips.

3.2.2. Data preparation and analysis

The data set consisted of gene expression data from the Illumina WG-6 v1 BeadChip of 48 probes from 28 different samples (eleven Parkinson patients, seven young controls and ten old controls). The raw expression values were extracted from the BeadStudio to R, logarithmized and different normalization methods were applied. The LOESS normalization was graphically determined to be the optimal normalization method for these data (for details see (Heim, 2008)). Next, all probes were clustered using the R function `agnes()`. The dendrogram is shown in Figure 3.3. The twelve samples that clustered separately were excluded from further analyses due to possible quality problems. For all remaining samples that were measured twice the mean expression value was calculated and used for the analysis. Finally the data set consisted of eight PD samples, seven young controls and nine old controls.

Due to the small sample size the power to detect any differentially expressed genes was quite low. One possibility was to reduce the number of performed tests to increase the significance threshold and reduce the multiple testing problem. Therefore the number of expression probes was reduced before the analysis by choosing only 8,491 out of 47,312 probes that were significantly detected in all samples, meaning that the expression level of these probes is significantly higher than the background noise (detection p-value below 0.05).

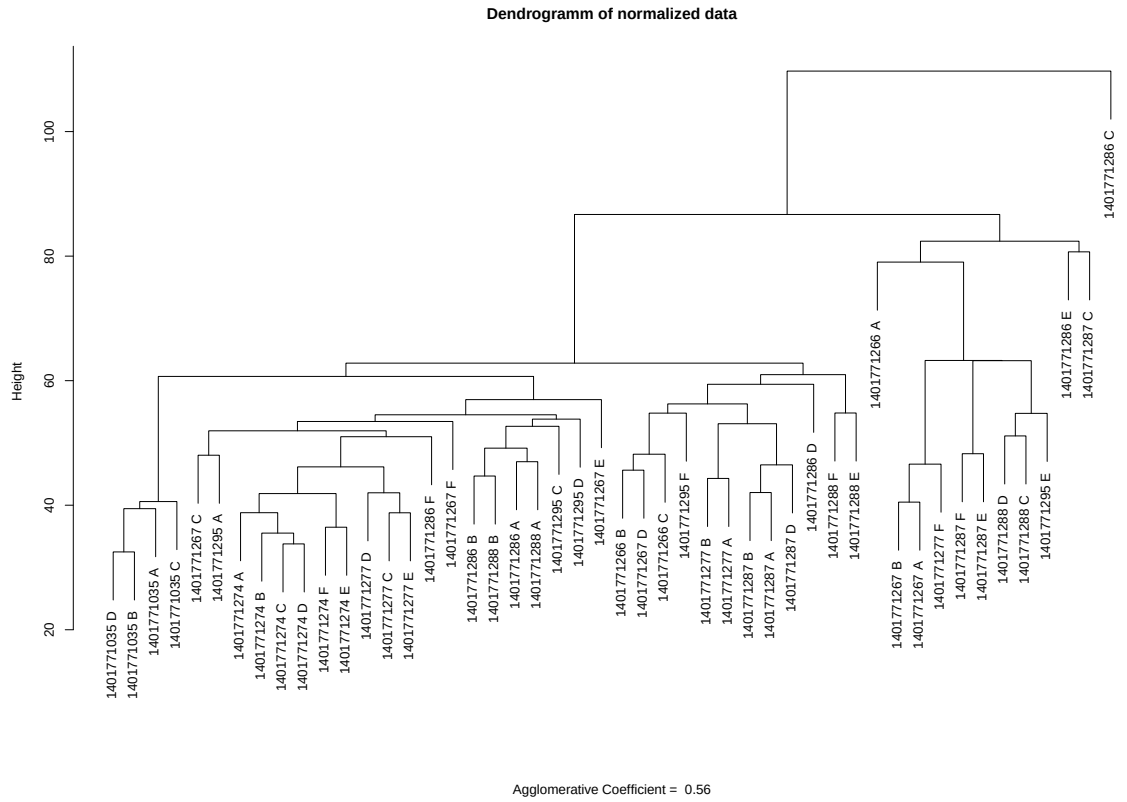

Figure 3.3.: Dendrogram of all samples from the Parkinson study:
All twelve samples that clustered in the right branch were excluded for further analyses due to potential quality problems.

3.2.3. Identification of a new risk gene for Parkinson's disease

To identify differentially expressed genes between Parkinson patients and controls, all nine old control samples were compared to the eight patient samples using the t-test (Elstner et al., 2009). When applying a Bonferroni correction only four probes remained significant: *MTND2*, *PDXK*, *SRGPA3* and *TRAPPC4* (see Table 3.2 and Figure 3.4). The upregulation of the genes *MTND2* and *PDXK* could be confirmed by real-time Polymerase Chain Reaction (PCR). This technology is usually used to amplify and quantify a DNA molecule. The results of the real-time PCR are shown in the upper two figures of Figure 3.4.

Gene	Definition	Fold Change	p-value	Biological process
MTND2	Homo sapiens NADH[1] dehydrogenase, subunit 2 (complex I)	up 1.70	$1.14 * 10^{-7}$	ATP[2] synthesis coupled electron transport
PDXK	Homo sapiens pyridoxal (pyridoxine, vitamin B6) kinase	up 1.32	$3.27 * 10^{-6}$	Pyridoxine biosynthetic process
SRGPA3	Homo sapiens SLIT-ROBO Rho GTPase activating protein 3	up 1.23	$5.65 * 10^{-6}$	Signal transduction
TRAPPC4	Homo sapiens trafficking protein particle complex 4	down 1.69	$5.8 * 10^{-6}$	ER[3] to Golgi vesicle-mediated transport

Table 3.2.: Differentially expressed genes between patients and old controls: [1]Nicotinamide-Adenine-Dinucleotide-Hydrid, [2]Adenosine triphosphate, [3]endoplasmic reticulum

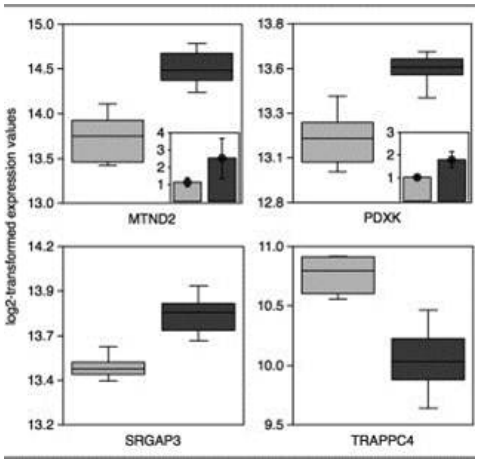

Figure 3.4.: Differentially expressed genes between Parkinson patients and old controls: The light gray boxes represent the PD samples and the black boxes the controls. The small boxes in the upper two figures show the result of the gene expression measurement by real-time PCR.

3.2.4. Differences in expression patterns for Parkinson's disease and aging

To identify genes that are influenced not only by Parkinson's disease but also by normal aging the seven young controls were taken into account (Elstner et al., 2011). The young controls were compared to the old controls and the Parkinson patients by applying an ANOVA. The Benjamini-Hochberg correction was used to control the false discovery rate. This resulted in 409 probes with p-value <0.01, 1,661 probes with p-value <0.05 and 2,953 probes with p-value <0.1. The final threshold was set to 0.05. The significant probes could be mapped to 1,608 different genes using the Ingenuity Pathway Analysis Software (IPA).

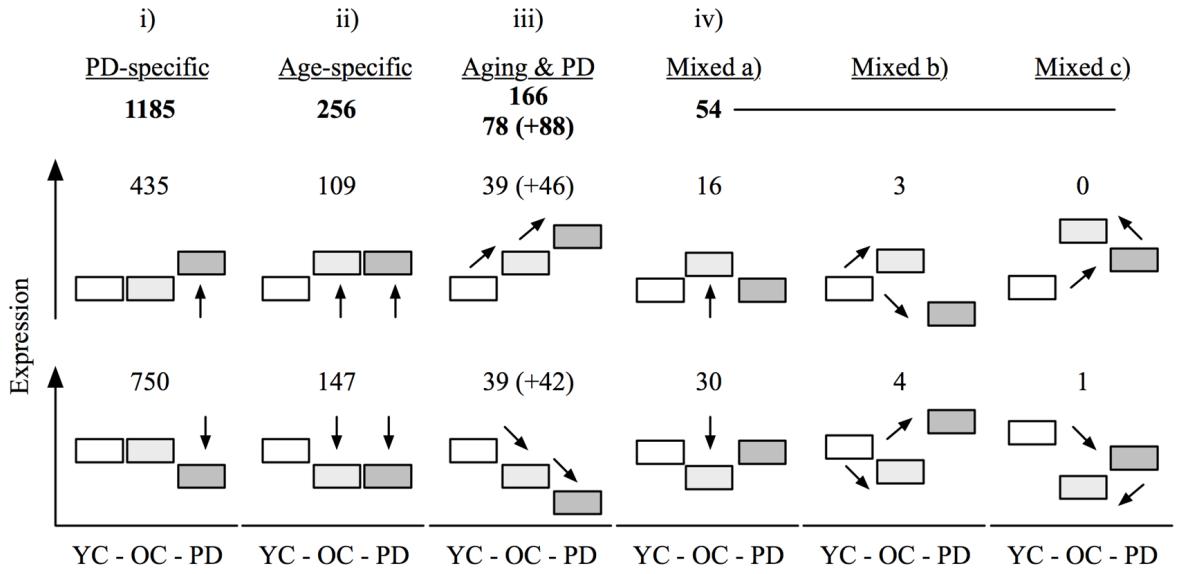

Figure 3.5.: Parkinson's disease as accelerated aging:
Following ANOVA analysis (FDR 5%), a total of 1,661 transcripts were significantly altered between the three groups: young controls (YC), age-matched (old) controls (OC) and PD. We mainly found three patterns of alterations: i) PD-specific: 1,185 transcripts were differentially expressed in Parkinson patients (435 up-regulated in PD and 750 down-regulated in PD, respectively) ii) Age-specific: 256 transcripts were differentially expressed only in young controls (147 up- and 109 down-regulated, respectively) iii) PD as accelerated aging: 78 transcripts were nominally significant ($p < 0.05$) differentially expressed between YC, OC and Parkinson patients, but the effects were going in the same direction. The numbers in brackets indicate the number of transcripts with p-value > 0.05. Only 54 transcripts were showing a different pattern.

To distinguish between expression patterns that are PD-specific or age-specific a t-test was conducted between young and old controls, young controls and patients and old controls and patients with a nominal p-value threshold of 0.05. Remarkably, all hits were not distributed equally across all possible patterns. The most common pattern was PD-specific, meaning that the expression level was almost identical in young and old controls and decreased or rather increased in the patients. In the second common pattern the expression levels differed only in young controls (age-specific pattern). The third common pattern de-

scribed Parkinson's disease as accelerated aging because the expression level increased/decreased in the old controls and even more in the patients. All exact numbers and all other possible patterns can be seen in Figure 3.5.
The age- (n=256) and PD-specific (n=1185) gene lists were used for a pathway analysis with IPA (Table 3.3).

Canonical pathway	p-value	↑ / ↓	main direction
Specific for aging			
cAMP[1]-mediated signaling	2.95E-03	5/2	↑
GABA[2] receptor signaling	3.16E-03	1/3	↓
Sphingolipid metabolism	9.33E-03	2/3	↓
Aminophosphonate metabolism	1.26E-02	0/3	↓
RAR[3] activation	1.66E-02	5/1	↑
Cardiac-adrenergic signaling	1.86E-02	3/2	↑
Methionine metabolism	2.63E-02	0/3	↓
Selenoamino acid metabolism	2.88E-02	0/3	↓
Glucocorticoid receptor signaling	3.39E-02	6/1	↑
G-Protein coupled receptor signaling	4.07E-02	5/1	↑
IL[4]-22 signaling	4.07E-02	2/0	↑
Estrogen receptor signaling	4.07E-02	2/2	↔
TNFR2[5] signaling	4.37E-02	2/0	↑
Glycosaminoglycan degradation	4.68E-02	2/1	↑
Aging and PD			
Nicotinate and nicotinamide metabolism	1.62E-02	3/1	↑
Agrin interactions at neuromuscular junction	1.86E-02	2/1	↑
Regulation of actin-based motility by rho	3.39E-02	1/2	↓
PAK[6] signaling	3.55E-02	1/2	↓
Purine metabolism	3.63E-02	2/5	↓
Huntington's disease signaling	3.80E-02	2/3	↓
Pantothenate and CoA[7] biosynthesis	4.37E-02	0/2	↓

Table 3.3.: Canonical pathways specific for aging and for Parkinson's disease:
For the pathway analysis 1,185 PD-specific transcripts and 256 age-specific transcripts were used as input for the Ingenuity Pathway Analysis Software.
[1]Cyclic adenosine monophosphate, [2]gamma-Aminobutyric acid, [3]Retinoic acid receptor, [4]Interleukin 22, [5]Tumor necrosis factor receptor 2, [6]p21 activated kinase, [7]Coenzyme A

.

3.3. Gene expression in patients with mitochondrial disorders

As the main focus of my working group is the genetic of mitochondrial disorders, there were performed some gene expression measurements on patients with mitochondrial disorders. The mitochondria are the so-called power-plants of the cells. Defective mitochondria may cause failure of metabolism. The most common and best studied mitochondrial defects are

defects of the respiratory chain. Defects of the respiratory chain cause a cellular energy deficiency and lead to neurodegenerative disorders. Most of the time the central nervous system, the skeletal muscle or the heart are affected because they have a high energy rate. Mitochondrial disorders are a very rare disease with a prevalence of 1:50,000.

This project was a collaboration with Dr. Monika Hartig and Dr. Arcangela Iuso from the Institute of Human Genetics of the Technical University in Munich (Hartig et al., 2011) on NBIA.

Neurodegeneration with **B**rain **I**ron **A**ccumulation (NBIA) is a neurodegenerative disorder. It comprises a very homogeneous group of neurodegenerative disorders with various combinations of symptoms. The only commonality are the abnormal high levels of brain iron which can be diagnosed with MRI (Magnetic Resonance Imaging)(TIRCON). There are ten subgroups of NBIA (see Figure 3.6) and the investigated patients are from the following two groups:

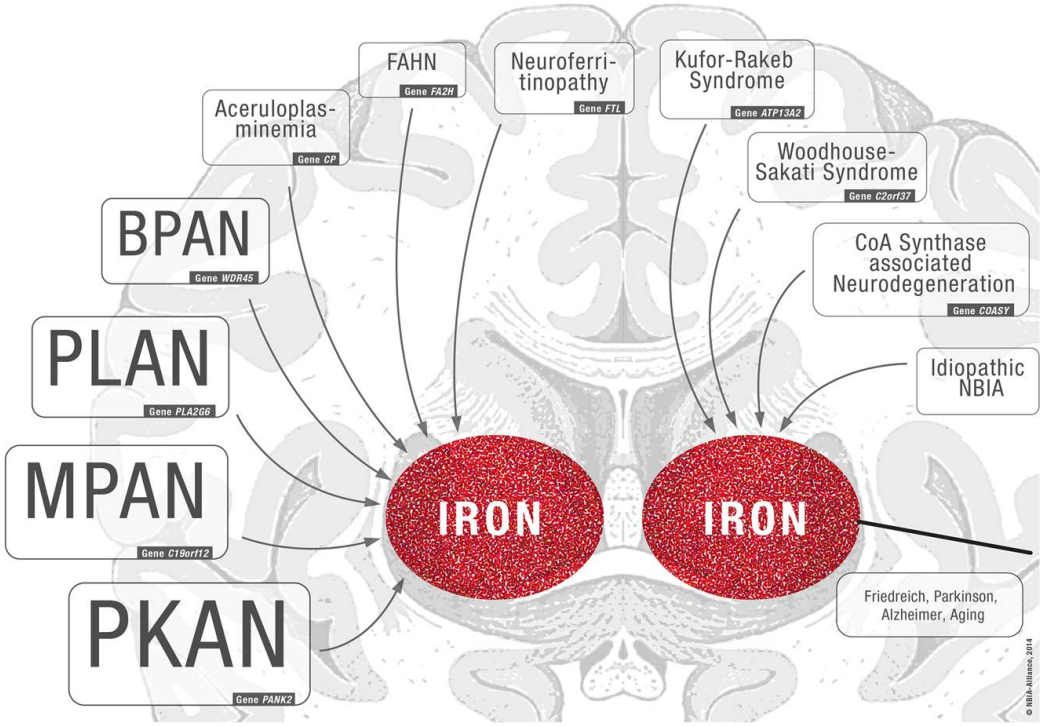

Figure 3.6.: Subgroups of NBIA:
In addition to the different subgroups of NBIA (displayed with the causal gene) also diseases like Parkinson's disease or Alzheimer or aging can cause iron accumulation in the brain.
Adopted from `http://Tircon.eu`.

- The most common form is PKAN (**p**antothenate **k**inase **a**ssociated **n**eurodegeneration) which is caused by mutations in the *PANK2* gene. An indication for PKAN is the so-called "Eye of the tiger"-sign in the MRI.
- MPAN (**m**itochondrial membrane **p**rotein-**a**ssociated **n**eurodegeneration) is caused by

mutations in the *C19orf12* gene.

There is an early and a late onset form of both subgroups, whereas the early onset form is the classical type and is characterized by a rapid progression and a life expectancy of less than 20 years.
In this study blood samples were taken from children suspected to suffer from one of the two different subgroups of NBIA and their healthy siblings and expression was measured using the Affymetrix GeneChip Human Gene 1.0 ST Array (see Section 2.5.2). Each sample was measured twice whereas in one measurement globin was reduced. The expression values were calculated in R using the Bioconductor package `affy` with the function `rma` (Irizarry et al., 2003) and were by default logarithmized and normalized with the quantile normalization. Afterwards all samples were clustered using the `agnes` function from R-package `cluster`. Three outliers could be identified and were removed from further analysis (right branch of the dendrogram in Figure 3.7).

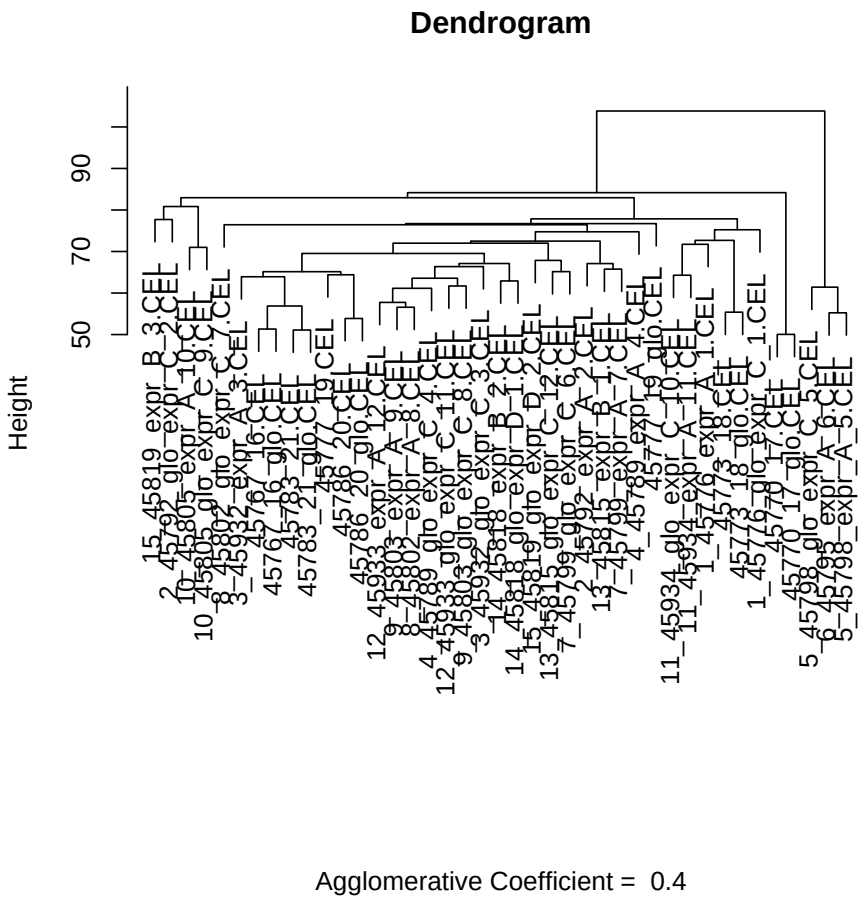

Figure 3.7.: Dendrogram of NBIA samples:
The expression levels of 41 samples were clustered and three outliers (in the right branch of the cluster) were identified and excluded for the further analysis.

To also take globin-reduced samples into account we calculated linear mixed models. By

this we compared the expression levels of patients having a mutation in *C19orf12* (N = 7) with the seven controls and patients having a mutation in *PANK2* (N = 6) with the controls. The expression probes were sorted by p-value and the top 500 probes with lowest p-values were used for a pathway analysis with Ingenuity. Additionally we conducted a pathway analysis in *C19orf12*-coregulated genes identified in a healthy population (KORA; N=381). This population-based study is presented in detail in 2.1.1. In all three analyses pathways that are related to mitochondrial functions are identified (see Table 3.4).

C19orf12 co-regulation in controls (n=381)	Diff. expression in MPAN (n=6) versus controls (n=7)	Diff. expression in PKAN (n=7) versus controls (n=7)
Fatty Acid Biosynthesis (4.64E-06)	Natural Killer Cell Signaling (2.49E-08)	Cholecystokinin/Gastrin-mediated Signaling (1.03E-03)
Valine, Leucine and Isoleucine Degradation (5.86E-04)	Prolactin Signaling (1.36E-05)	**Mitochondrial Dysfuntion (2.7E-03)**
Protein Ubiquitination Pathway (4.11E-03)	Fcg Receptor-mediated Phagocytosis in Macrophages and Monocytes (1.62E-05)	Oxidative Phosphorylation (3.72E-03)
Propanoate Metabolism (5.33E-03)	Growth Hormone Signaling (6.25E-05)	Natural Killer Cell Signaling (4.92E-03)
Fatty Acid Elongation in Mitochondria (1.18E-02)	**Mitochondrial Dysfunction (2.5E-03)**	Folate Biosynthesis (7.33E-03)

Table 3.4.: Results of pathway analysis for NBIA patients versus controls:
Pathway analyses were conducted using Ingenuity Pathway Analysis Software using three different gene lists. The first column shows the result from a healthy population-based study (KORA) where genes that were co-regulated with *C19orf12* were identified. For the second column six MPAN patients were compared to seven controls and for the third column seven PKAN patients were compared to the same controls.

3.4. Summary and discussion

The difficulty when analyzing gene expression data from case-control studies is the relatively small sample size in comparison to the large number of measured gene expression probes. To avoid this problem a few possibilities are available and were applied on the two data sets.
First of all a homogeneous data set is important because one extreme outlier could distort the whole result and lead to wrong conclusions. For both data sets we used clustering to identify samples with potentially bad quality and removed them from further analysis even if this reduced the sample size rigorously (in the Illumina data twelve out of 48 samples were removed as they were cluster outliers).
To decrease the multiple testing problem the expression probes in the Parkinson project were limited by using only expression probes that are above the detection threshold. For the NBIA project this was not possible because Affymetrix does not provide a score like the detection

p-value that indicates if the expression level is significant above the background. Therefore all probes were used. Due to the fact that no gene was differentially expressed when correcting for multiple testing, the 500 top hits were selected for a pathway analysis and this yielded in pathways that are related to mitochondrial functions.
However the sample size was very small in both studies, the obtained results of both studies were the beginning of future work.
The biggest advantage of the Parkinson study is that the expression levels were measured in the affected tissue and that there are two different kinds of control groups. We identified a differential expressed gene between old controls and patients namely *PDXK*. This gene converts vitamin b6 from the nutrition in its active form. Epidemiological studies showed a decreased risk for Parkinson's disease in individuals with a high vitamin b6 level (de Lau et al., 2006). With our data we confirmed that the intake of vitamin b6 could decrease the risk for Parkinson's disease .
With the first results from the NBIA project it was possible to justify that a larger study is necessary and helpful. It was the precursor experiment of a larger project with a larger sample size, planned in the near future (see Chapter 7).

4. Improvement and development of quality control of gene expression data

The development of microarray technologies, the associated reduction of costs, the improvement of quality and reduced time requirement for experiments enables now the determination of gene expression levels in population-based studies with larger sample sizes. When we started to establish a genome-wide expression data set in a population-based study, for comparison only data sets with few samples were available. Therefore the experimental settings have to be established first before measuring all samples. We started with KORA F3, improved the protocols in KORA F4 and finally standardized all procedures and protocols in the German gene expression consortium MetaXpress.

4.1. Biological and technical replicates and manual quality control: KORA F3

Divya Mehta measured the gene expression of 381 samples of the KORA F3 cohort using the Illumina HumanWG-6 v2 BeadChip in the context of her dissertation (Mehta, 2009). The KORA F3 cohort was among the first cohorts with genome-wide expression data. There were no comparable data sets available at this time so that a quality control protocol had to be established and validated.

4.1.1. Biological and technical replicates

The first question to address was how robust and reproducible gene expression levels from whole blood samples are. In the beginning it was unknown to which extent the expression levels differ from day to day and from individual to individual. Before the expression levels were measured in the KORA F3 samples the experimental protocol was validated by measuring three different samples (1-3) at three different time-points (a-c). Three healthy voluntary males from the Institute of Human Genetics were recruited and blood was taken under fasting conditions three times once a week. The expression values were normalized, Pearson's correlation coefficient was calculated and all samples were plotted pairwise (Figure 4.1). All correlation coefficients were greater than 0.95 however not all nine samples could be measured on the same array as one array can only be loaded with six samples in parallel.
As differences between the three measurements of one individual are negligible the expression levels seems to provide a good overview of the gene activity and it was started to measure all samples with available PAX tubes from KORA F3.

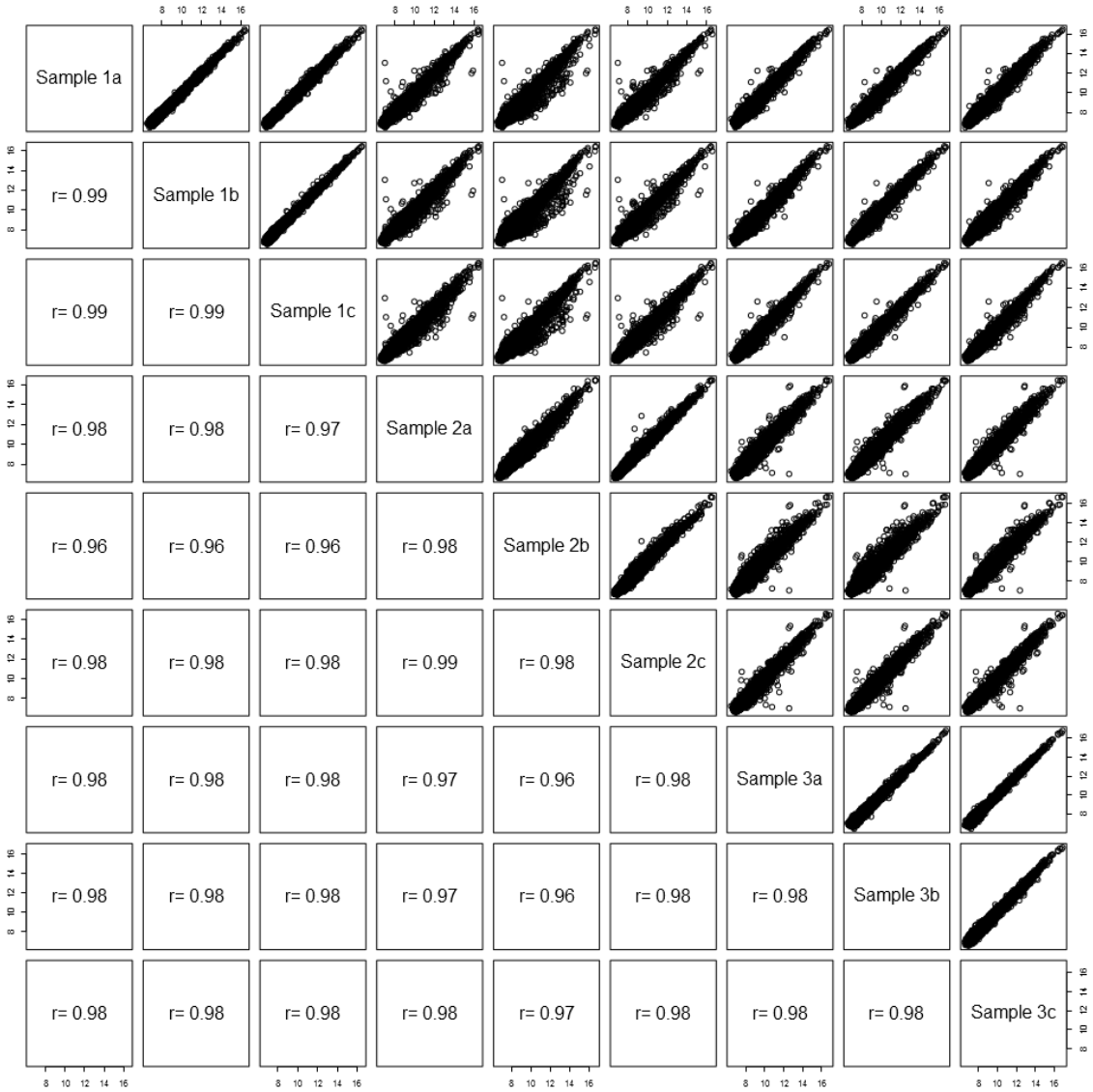

Figure 4.1.: Scatterplot matrix of three technical and biological replicates:
The upper panels show the Pearson's correlation coefficient between technical and biological replicates, which is always greater than 0.95, indicating a very high correlation between all technical and biological replicates.

4.1.2. Quality control

The first quality score that can be obtained for each sample after RNA isolation is the RNA integrity number (RIN). The RIN can reach values between 1 and 10 in which 1 means totally degraded[1] RNA and 10 nondegraded RNA (Schroeder et al., 2006). The RIN in the analyzed KORA F3 samples ranged from 2.3 to 8.4. One reason for this large range of the RIN might be the badly organized management of the PAX tubes. All PAX tubes from one day were collected and transported via cab from the study center in Augsburg (where the blood was taken) to the Institute of Human Genetics (where the RNA was isolated). The storage time and temperature of the PAX tubes was not standardized and varied a lot.

It could be assumed that samples with high quality, indicated by a high RIN, have a high number of detected genes. The number of detected genes per sample could be obtained from the GenomeStudio (the software from Illumina that allows the quality control and some basic analyses). It indicates the number of genes that are significantly higher expressed than the background noise. To verify the assumption that samples with a higher RIN also have more detected genes, a linear regression model was calculated. Testing the effect of the RIN on the number of detected genes a direct significant relationship could be observed with a p-value of $2.99 * 10^{-18}$ (see Figure 4.2). Surprisingly, there are also samples with a very small RIN (< 5) but more than 8,000 significant detected genes.
However, it is difficult to make a statement on low expressed genes (gene level is less than the background noise). We kept also samples with lots of low expressed genes for the reason that power to detect significant changes in the expression levels increases with sample size.

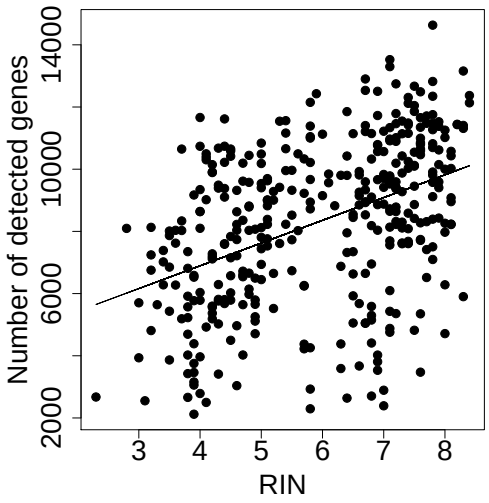

Figure 4.2.: RIN versus number of detected genes in KORA F3:
A higher RIN leads to significant more detected genes (p-value of linear regression = $2.99 * 10^{-18}$).

[1]RNA in cells is not very stable to allow a continuous degeneration. This process is done with enzymes (RNasen).

When processing several hundred samples in parallel there is a risk for sample mixing. Since gender of the probands was given, one way to identify mixed samples is to take advantage of the sex-specific expression pattern. Naturally women should have low expression values for all Y-genes. Therefore the expression pattern of all genes on the Y-chromosome was inspected graphically and three samples with high expression levels however they are indicated to be female or the other way round were excluded.

4.1.3. Processing of the data

To ensure the comparability of gene expression levels of samples that were measured on different days and on different arrays, all 381 samples were normalized trying three different normalization methods (rank invariant, LOESS and VSN normalization). The samples were compared graphically to determine the LOESS normalization as the optimal normalization method (for more details see (Heim, 2008)).

The data set consisted now of 381 samples and the expression levels of 48,701 probes per sample. When analyzing all probes separately the number of performed tests was obviously quite high and the number of false positive hits increased (multiple testing problem). To face the multiple testing problem we reduced the number of probes. One way was to use only probes that are significantly detected in more than 5% of all samples. "Significantly detected" means in this case that the detection p-value was less than 0.01 in more than 19 out of 381 samples. Using this threshold the number of probes was reduced from 48,701 to 13,767. Applying a Bonferroni correction the threshold for significance changed just from 0.05/48,701 = $1.03 * 10^{-6}$ to 0.05/13,767 = $3.6 * 10^{-6}$ but the number of performed tests is reduced by a third and speeded up the analysis, respectively.

4.2. Validation of new technology: KORA F4

The experiences gained from the first experiments with KORA F3 were used to establish an even larger whole-genome gene expression data set. PAX tubes were available for more than 1,000 samples from KORA F4. Additionally there were samples from baseline survey S4 which were frozen for several years. For most of participants who were non diabetics there were blood samples collected after an oral glucose tolerance test (OGTT). Expected quality for the S4 samples was low after the long freezing time and therefore the focus was led on the KORA F4 samples as better comparability to other expression data was expected.

4.2.1. One or two arrays per sample?

The KORA F3 samples were measured using the Illumina HumanWG-6 v2 Bead Chip which contains on average about 30 beads per probe. This array was no longer available when starting the measurement of the KORA F4 data. Another array was introduced by Illumina (Illumina HumanHT-12 v3) and due to the lower price and the advantage that 12 samples instead of 6 could be measured in parallel the HumanHT-12 v3 array was used for the following experiments. The advantage was that two of this new arrays were cheaper than one

of HumanWG-6 v2 arrays. The disadvantage of the HT-12 v3 array was that it was not clear in the beginning if the quality of the data could be worse due to the fact that it contains only half amount of beads (n = 15) per probe compared to the HumanWG-6 v2 array.

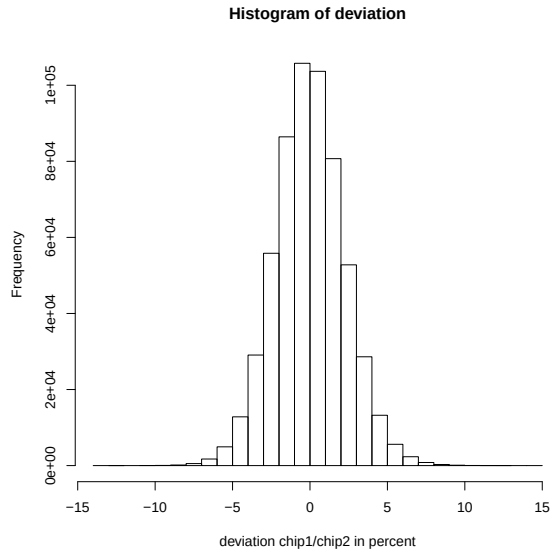

Figure 4.3.: Histogram of deviations between expression values when using two different arrays per sample:
Twelve samples were measured on two different arrays. Weighted means were calculated for each probe of each sample using the number of beads as weights. One array has on average 15 beads per probe and two arrays ∼30 beads accordingly. The largest deviation from the weighted mean is 12% and 99.9% of the deviations are less than 5%.

To figure out if two arrays are necessary to get the same result as with using only one array the first twelve samples were measured on two different arrays. For each of the twelve samples the deviation of expression values was calculated for every probe on the array. For 97% of the probes the deviation was less than 5%. Further the probe specific weighted mean was calculated by using the number of beads per probe as weight. The largest deviation from the weighted mean was 12% and 99.9% of the deviations are less than 5%. (see histogram in Figure 4.3). Double measurement did not improve the quality but would duplicate the work. Therefore we stayed with only one array per sample.

4.2.2. Amount of cRNA, scanner regulation, and amount of RNA

To establish and optimize the protocol for gene expression measurement four samples were measured three times and we used the "number of detected genes" as quality criterion.

- Different starting amounts of RNA were tested (100ng and 200ng RNA)
  $\Rightarrow$ 200ng RNA led to higher number of detected genes.
- Different amounts of cRNA for hybridization were tested (750ng cRNA, 1.5μg cRNA,

and 3μg cRNA)
$\Rightarrow$ 3μg cRNA showed the highest amount of detected genes.

- Different scanner settings were tested (default: gain1, gain2, and gain4)
  $\Rightarrow$ No difference was observed between all settings and therefore the default regulation gain1 was chosen.

4.2.3. Establishment of a comprehensive quality-controlled data set

In total, RNA was isolated from 3,301 PAX tubes from KORA S4, F4 and F4 after the oral glucose tolerance test (F4 OGTT). For 391 PAX tubes the amount or quality of RNA did not meet the quality criterion and was therefore excluded. Bad quality was defined by a RIN < 6 (especially the samples from KORA S4 had low RINs (< 3)) to ensure a more homogenous data set. In addition we defined a threshold for the number of detected genes $>$ 6,000 which further excluded 25 samples. Due to the limited number of detected genes ($<$ 6,000) in the first run, we measured 327 probes two times. Ten probes were measured three times and four probes four times. Sixteen probes (corresponding to nine different samples) were not in the KORA data base and were not allowed to be used for further analysis.
Altogether 341 probes were measured more than one time: 135 F4 samples, 126 F4 after OGTT samples and 80 S4 samples. We could not observe that samples with an impaired glucose have to be measured more often than samples with a normal glucose.
Finally we ended up with a data set consisting of expression levels of 2,509 samples with more than 6,000 detected genes.
For further processing all expression values were log2-transformed and normalized using the quantile normalization. To identify outliers all samples were clustered using the whole expression profile and thereby three outliers (three S4 samples) were identified and were removed from the data set (the dendrogram is shown in Appendix A.1).

In a second step we controlled for sample mixing by determine the gender of a sample by clustering of samples due to the expression of probes that are located on the sex chromosomes. Female samples should have very low expressed probes on the Y- and males should have low expressed probes on the X-chromosome. To spot mixed samples, all three cohorts were clustered using only probes that were located on X- and Y-chromosome. 35 probes of all three cohorts were deleted due to wrong classification (see Table 4.1). More than half of the wrongly classified samples were from the S4 cohort (57%). In KORA F4 only 5 out of 998 samples (0.5%) have been mixed or contained unusual X-Y gene expression levels.

cohort	males	wrong males	females	wrong females
S4	360	11	317	9
F4	504	1	494	4
F4 OGTT	403	6	420	4

Table 4.1.: Gender classification in KORA S4, F4, and F4 OGTT:
The gender of each sample was determined by clustering all expression probes that are located on X- and Y-chromosomes.

After removing all outliers and mixed samples there were altogether 657 S4 samples, 993 F4 samples and 813 F4 OGTT samples which were amplified on 32 different amplification plates. For about half of the samples (N = 460) three measurements from the three different time points were available (see Figure 4.4).

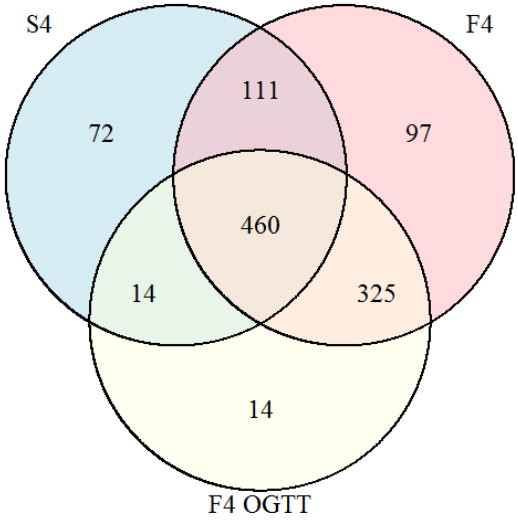

Figure 4.4.: Overlap of samples from KORA S4, F4, and F4 OGTT

4.3. Common quality controlled preprocessing and analysis strategy: MetaXpress

In 2011 the MetaXpress Consortium was founded within the DZHK - German Center for Cardiovascular Research (**D**eutsches **Z**entrum für **H**erz-**K**reislauf-Forschung). MetaXpress consists of three large German study cohorts with available gene expression data:

- KORA F4 (Munich)
  Described in Section 2.1.1.
- GHS (Mainz, Lübeck, Hamburg)
  Described in Section 2.1.4.
- SHIP-Trend (Greifswald)
  Described in Section 2.1.2.

Firstly the methodological aspects of the data preprocessing should be standardized in all three participating cohorts to ensure the comparability. Since there was no gold standard for expression data preprocessing, we compared available methods to stabilize the variation of the expression data and defined a set of confounders with impact on this variation.

In MetaXpress, we analyzed gene expression levels in terms of specific mRNA abundances in whole blood (SHIP-TREND and KORA F4) or blood monocyte samples (GHS). The descriptive statistics of the participants and parameters analyzed in the studies are provided

in Table 4.2. Despite higher age of the participants of KORA F4 (mean age = 70.4 years) than in SHIP-TREND (mean age = 50.1 years) and GHS (mean age = 54.7 years) and the storage time of the samples was longer in KORA F4 (855 days) than in SHIP-TREND (204 days) and GHS (314 days), there are no large differences between the three cohorts.

Variable (mean/SD)	SHIP-TREND	KORA F4	GHS
Sample size	991	993	1374
Storage time*	204.0±153.8	855.5±179.4	314.4±91.6
RNA integrity number (RIN)	8.56±0.50	8.68±0.61	9.36±0.43
Females (%)	555 (56.0)	493 (49.6)	622 (48.4)
Age [years]	50.1±13.7	70.4±5.4	54.7±11.0
Body height [cm]	169.8±9.0	165.3±8.8	171.0±9.3
Body weight [kg]	79.0±15.1	78.9±13.7	79.1±15.5
Body mass index [kg/m^2]	27.3±4.6	28.9±4.5	27.0±4.6
Hip circumference [cm]	101.3±9.6	107.8±9.3	100.5±9.6
Waist circumference [cm]	88.0±12.9	98.6±12.1	93.5±13.4
Waist-to-hip ratio	0.87±0.09	0.91±0.08	0.93±0.09
White blood cell count [Gpt/l]	5.72±1.48	6.00±1.80	7.04±3.81
Red blood cell count [Tpt/l]	4.63±0.39	4.50±0.40	4.69±0.41
Hematocrit	0.42±0.03	0.41±0.03	0.42±0.03
Hemoglobin [mmol/l]	8.62±0.74	8.69±0.75	9.10±0.74
Platelets [Gpt/l]	225.7±50.3	244.7±65.1	271.5±67.9
Serum C-reactive protein [mg/l]	NA	3.05±6.27	3.78±4.92
High density lipoprotein [mmol/l]	1.48±0.37	1.43±0.36	1.47±0.40
Serum triglycerides [mmol/l]	1.42±0.85	1.50±0.84	1.46±0.91
Active smokers [%]	214 (22.0)	66 (6.7)	239 (18.6)
Systolic blood pressure [mmHg]	124.4±16.9	128.7±20.0	132.2±17.8
Diastolic blood pressure [mmHg]	76.6±9.8	74.0±10.1	83.5±9.68

Table 4.2.: Descriptive statistics of MetaXpress cohorts:
*Storage time: Time between blood sampling and RNA isolation (SHIP-TREND and KORA F4) or time between RNA isolation and RNA amplification (GHS) [days].
Serum C-reactive protein was not available in SHIP-TREND.

All statistical analyses were performed in each cohort separately. Dr. Claudia Schurmann and Dr. Alexander Teumer were responsible for the SHIP-TREND analysis, Arne Schillert and Christian Müller for the GHS analysis and I performed all analysis on the KORA F4 data set.

4.3.1. Variance stabilization transformation versus log2 transformation

Gene expression levels can reach values between zero (not expressed) and infinite (very high expressed). Thereby most of the expression levels are between zero and 100 which is identical to or below the background level. For any kind of parametric statistic a symmetrical distribution is assumed and therefore the values are usually logarithmized before the ana-

lysis. It is irrelevant if the log_{10}, log_e or log_2 is used. Biologists mostly prefer the log_2 scale because these values are easier to understand. The whole blood data sets from KORA F4 and SHIP-TREND were already on log2 scale whereas the monocyte data set from GHS was prepared using the variance stabilization transformation (VST). For the log2 transformation the average signal per probe (calculated from the about 15 beads per probe) is used. In comparison to the log2 transformation the VST (for details see Appendix A.2.1) takes all single measurements of each probe for each bead on the Illumina array into account.
Comparing the log2 expression values to the VST expression values the average values per probe are almost identical for large intensity values ($> 2^9$) in all three cohorts, but the log2 values are recognizably smaller for low intensity values (see upper left part of Figure 4.5). This is consistent with already published results (Lin et al., 2008). To compare the effects of the two different transformations we created a random normal distributed phenotype (N(0,1)) and calculated the associations for all expression probes with this uncorrelated phenotype by using a linear regression model. Further we used the body mass index (BMI) which is known to be highly correlated with gene expression levels in monocytes (Zeller et al., 2010) and whole blood (Xu et al., 2011). For both phenotypes the absolute effect sizes and the standard errors (SE) from the linear model were smaller when applying the VST in low intensity values (see Figure 4.5 for BMI and Figure A.2 for the random phenotype). This resulted in highly correlated association p-values ($R^2 = 0.9956$ in KORA F4). Therefore we concluded that there is no relevant difference between both transformation methods. As the log2 transformation is most often used, easier to apply and to interpret all three data sets were log2-transformed.

4.3.2. Determination of factors influencing gene expression

To identify technical variables with influence on gene expression levels principal component analysis was conducted in all three data sets. More than 96% of the variance in the expression data is explained by the first Eigen-vector (the so called Eigen-gene)[2] in all three cohorts. This result was independent of the applied variance stabilization method and the used tissue (see Figure 4.6).

In the SHIP-TREND data set it was additionally tested if more or less variance is explained by the first Eigen-vectors when excluding all probes with low intensity values (detection p-value less than 0.01 in at least 50% of the samples). No difference could be observed between the log2 transformation and the VST even when excluding the probes with low intensity values (see two upper figures of Figure 4.7). More than 95% of the variation in the expression levels is explained by the first Eigen-vectors.
Xu et al. (2011) also analyzed the impact of the first principal components on the variation of the gene expression levels. He investigated 24 whole blood samples measured on the Affymetrix GeneChip. In comparison to our results in this data set only 28.2% of the variation in the expression levels was explained by the first principal component. To test if this difference was due to the different sample sizes we selected a random sample of 24 individuals and calculated how much of the variation is explained by the first Eigen-vectors. For our

[2]In the original publication the Eigen-vectors were misleadingly called principal components however the Eigen-vectors were meant.

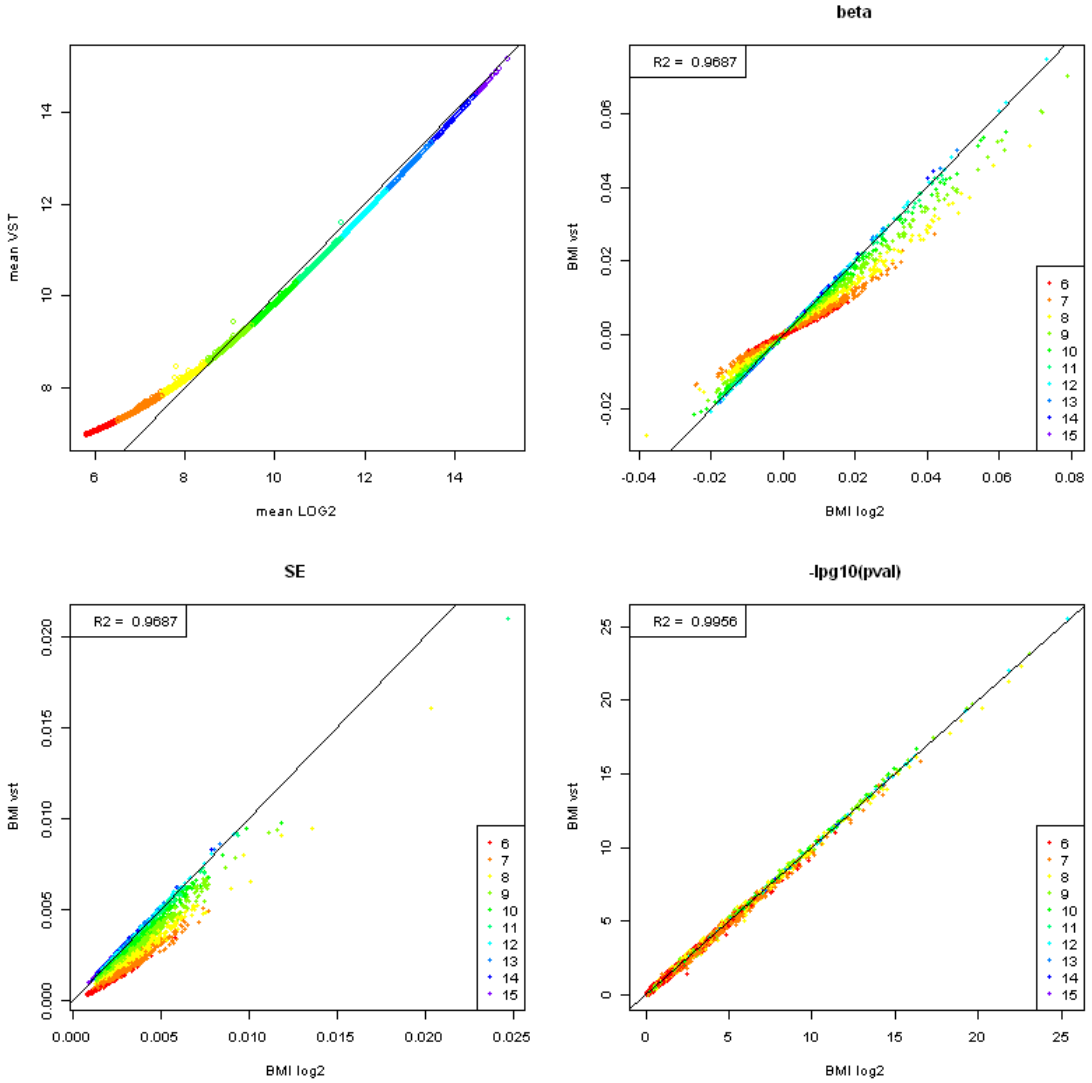

Figure 4.5.: Comparison of VST- and log2-transformed expression values in association with BMI in KORA F4:
The figures show the difference between the VST- and log2-transformed expression values in association with BMI. Each dot represents one probe and the color code is given in the legend of the plots. Differences are observed for the mean intensity values for low intensity values, for betas, and for standard errors of betas. Despite those differences the p-values are highly correlated (r^2=0.9956).

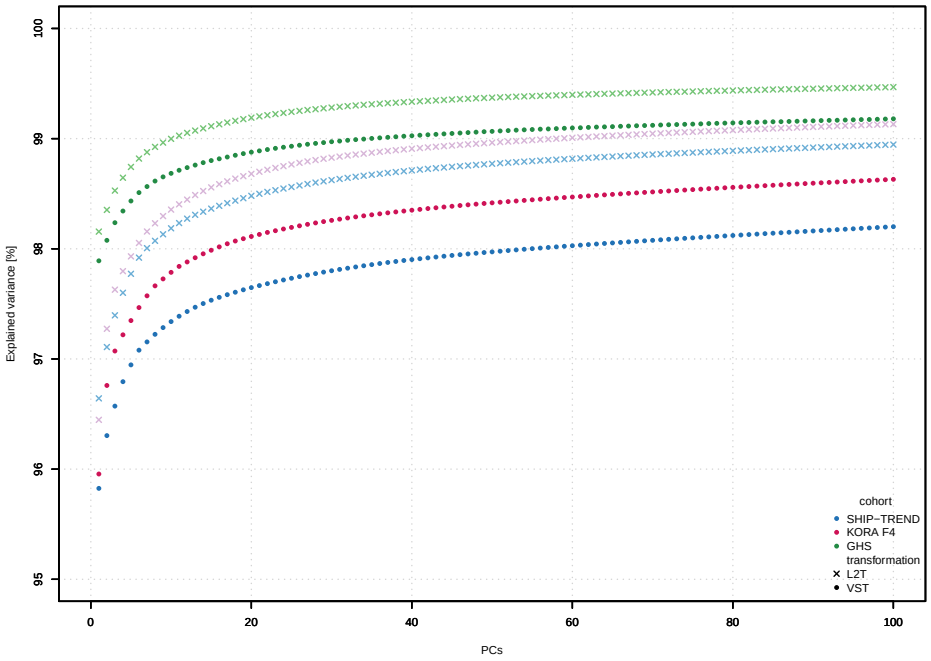

Figure 4.6.: Explained variance by the Eigen-genes in KORA F4, GHS, and SHIP-TREND: The explained variance is depicted for the first 100 Eigen-genes in all three cohorts for the two different transformations (log2 transformation and VST). More than 95% of the variance is explained by the first Eigen-gene in all three cohorts, independent from the transformation.

data we could not see a difference between the original and the reduced sample size. In the reduced subset of 24 samples the first Eigen-vector explains 97.3% of the variation in the expression data (see figure in the left lower corner of Figure 4.7). To investigate whether technical issues due to the different expression platforms (Illumina versus Affymetrix) could cause this large differences the whole SHIP-TREND data set was adjusted for technical factors (*Amplification batch*, *sample storage time* and *RIN*). Now the first Eigen-vector just explained 5.9% of the variation of gene expression (see figure in the right lower corner of Figure 4.7). This result suggested that the Eigen-vectors are highly correlated with technical factors and the differences to the data set from Xu et al. (2011) might be explained by the sample preprocessing (96 samples in parallel on one amplification plate) or the Illumina array (twelve samples on one array).

To identify the variables with the highest impact on the variation of expression levels the Eigen-R^2 was calculated for 25 technical, biological and clinical variables which were available in all three cohorts (and the Serum C-reactive protein as it is known to be associated with expression levels of some genes in whole blood). The Eigen-R^2 is a measurement for the explained variance of predefined variables. The Eigen-R^2 are shown in Table 4.3 for all three cohorts separated in technical and non-technical factors.

Most of the variance of expression levels in all three cohorts can be explained by the Illumina chip design (twelve samples per array). The highest value of 48.18% in KORA F4 can be explained by the higher number of samples that were processed. In KORA the expression was measured in S4, F4 and F4 OGTT at once and therefore the factor representing the Illumina chip number has more levels. In GHS only 26.55% of the variation was explained by the Illumina chip. One reason for this is the most optimal preprocessing of the samples. Since all samples were prepared on the same day the variation due to technical variables was smaller than in SHIP-TREND and KORA.
Beside those observations all Eigen-R^2 values were similar in all three cohorts. Only for blood cell-related factors (white and red blood cell count, hematocrit, and hemoglobin) we observed differences between the whole blood and the monocyte data.

Additionally to the calculation of the Eigen-R^2 the Eigen-genes were correlated with the same technical, biological, and clinical variables. Figure 4.8 shows the results for KORA F4 and Figure 4.9 for all three cohorts, respectively. For most technical variables the highest correlation was observed with one of the first five Eigen-genes but they are correlated with almost all Eigen-genes. The highest correlation was observed for the plate design which is highly correlated with the Illumina chip due to the distribution of 96 samples from one plate on eight arrays (twelve samples per array).

4.3.3. Reduction of unexplained variance by adjustment for covariates

To reduce the unexplained variance in the linear regression models the same phenotypes as in Section 4.3.1 were used (BMI and the random phenotype). We first included systematically up to 100 Eigen-genes in steps of five to the linear model, saved the adjusted R^2 values and calculated the means for each cohort separately. We could reduce the unexplained variance by about 30% when adding 50 Eigen-genes to the linear regression model (see Table 4.4 and

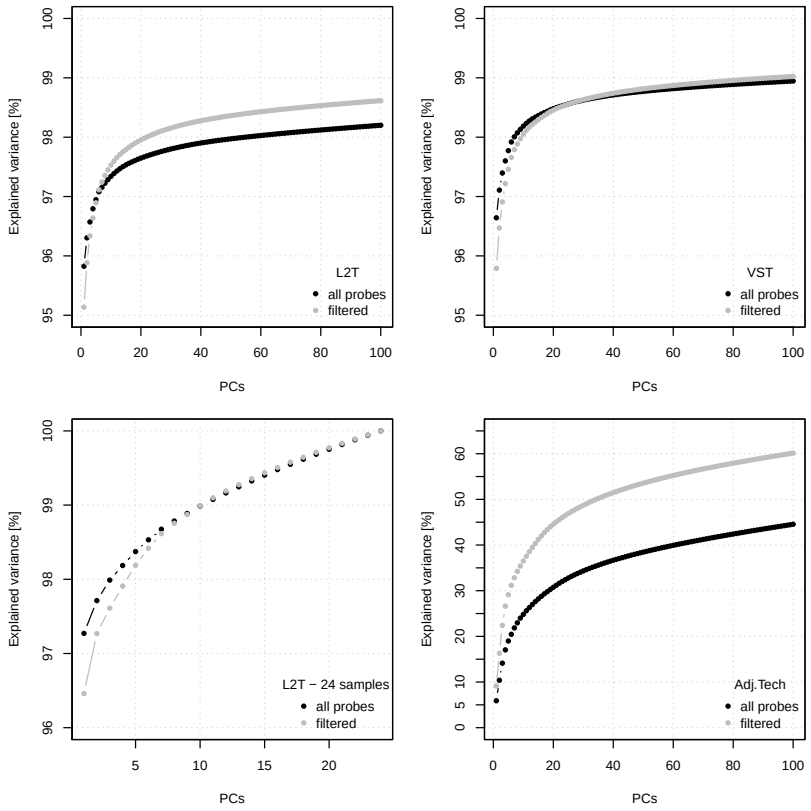

Figure 4.7.: Explained variance by the first 100 Eigen-genes for different filter methods in SHIP-TREND:
In SHIP-TREND it was tested if different filtering methods reduced the explained variance by the Eigen-genes. The upper two figures show the difference between using all expression probes that are available on the array (all probes) and using only the significantly detected probes (filtered) for the log2 transformation and VST, respectively.
The figure at the bottom left shows the difference between all and filtered probes, but using only 24 random samples. Only in the figure at the bottom right the explained variance by the first Eigen-gene is reduced to less than 10%. These data were adjusted for the technical variables (amplification plate, RIN, and sample storage time). The filtering of probes shows the same result.

Parameter	SHIP-TREND	KORA F4	GHS
Illumina Chip (12 samples per array)	33.75%	48.18%	26.55%
RNA amplification batch (96 well plate)	20.18%	24.30%	12.44%
Storage time [days]*	2.86%	1.60%	1.70%
Month of blood sampling	18.72%	3.31%	8.11%
Time of blood sampling [h]	0.20%	0.41%	0.61%
RNA integrity number	1.36%	0.77%	0.29%
Sex	0.95%	0.87%	1.51%
Age [years]	0.58%	0.45%	0.30%
Body height [cm]	0.54%	0.48%	0.82%
Body weight [kg]	0.59%	0.60%	0.51%
Body mass index [kg/m2]	0.68%	0.54%	0.35%
Hip circumference [cm]	0.60%	0.41%	0.27%
Waist circumference [cm]	0.77%	0.67%	0.52%
Waist to hip ratio	0.65%	0.70%	0.82%
White blood cell count [Gpt/l]	0.89%	0.74%	0.23%
Red blood cell count [Tpt/l]	0.38%	0.35%	0.65%
Hematocrit	0.47%	0.46%	0.83%
Hemoglobin [mmol/l]	0.50%	0.42%	1.03%
Platelets [Gpt/l]	0.32%	0.27%	0.63%
High density lipoprotein [mmol/l]	0.49%	0.48%	0.48%
Serum triglycerides [mmol/l]	0.68%	0.87%	0.23%
Active smokers [%]	0.36%	0.23%	0.26%
Systolic blood pressure [mmHg]	0.41%	0.15%	0.26%
Diastolic blood pressure [mmHg]	0.37%	0.14%	0.19%
Serum C-reactive protein [mg/l]	NA	0.30%	0.26%

Table 4.3.: Eigen-R^2 values for technical and non-technical variables in KORA F4, SHIP-TREND, and GHS:
*Storage time means the time between blood sampling and RNA isolation in KORA F4 and SHIP-TREND and between RNA isolation and amplification in GHS. Serum C-reactive protein was not available in SHIP-TREND.

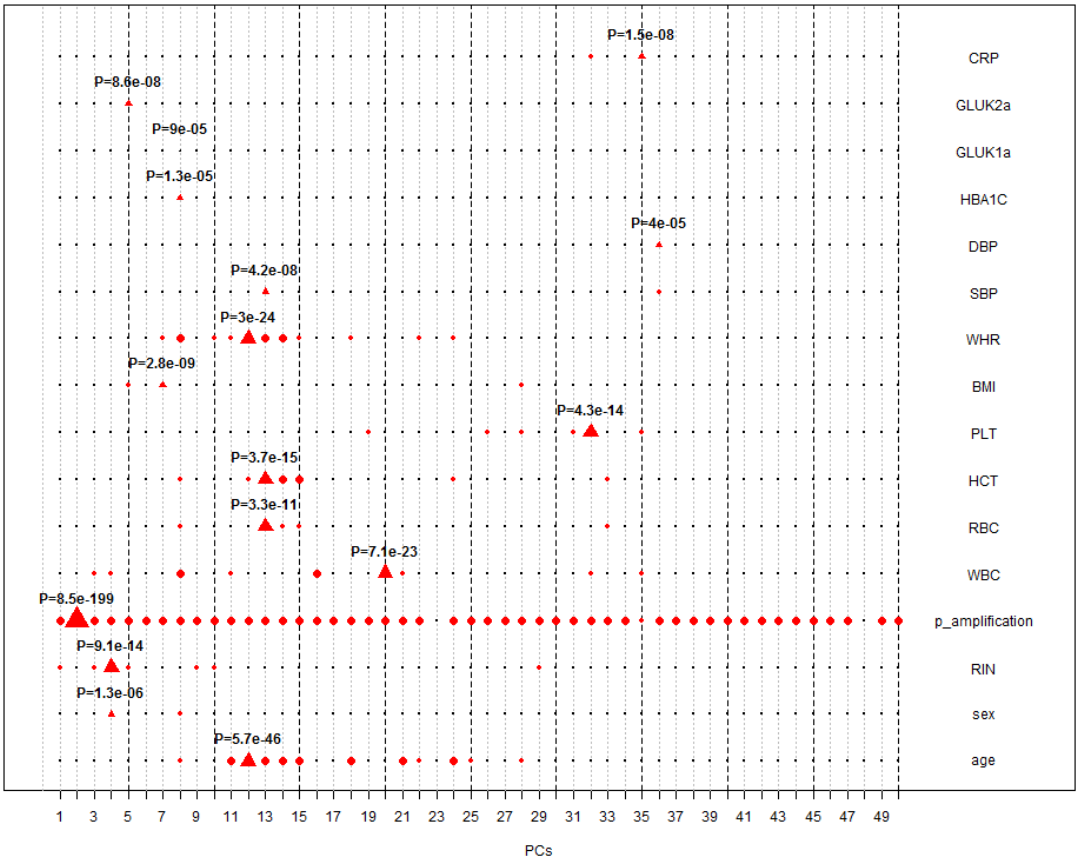

Figure 4.8.: Correlation of covariables with the first 50 Eigen-genes in KORA F4:
The smaller the p-value the larger the dots. The smallest p-value per covariable is indicated with a triangle. (CRP: Serum C-reactive protein, GLUK2a: 2h-glucose level (after glucose tolerance test), GLUK1a: fasting glucose level, HABA1C: glycated hemoglobin, DBP: diastolic blood pressure, SBP: systolic blood pressure, WHR: waist-hip-ratio, PLT: platelets, HCT: hematocrit, RBC: red blood cell count, WBC: white blood cell count)

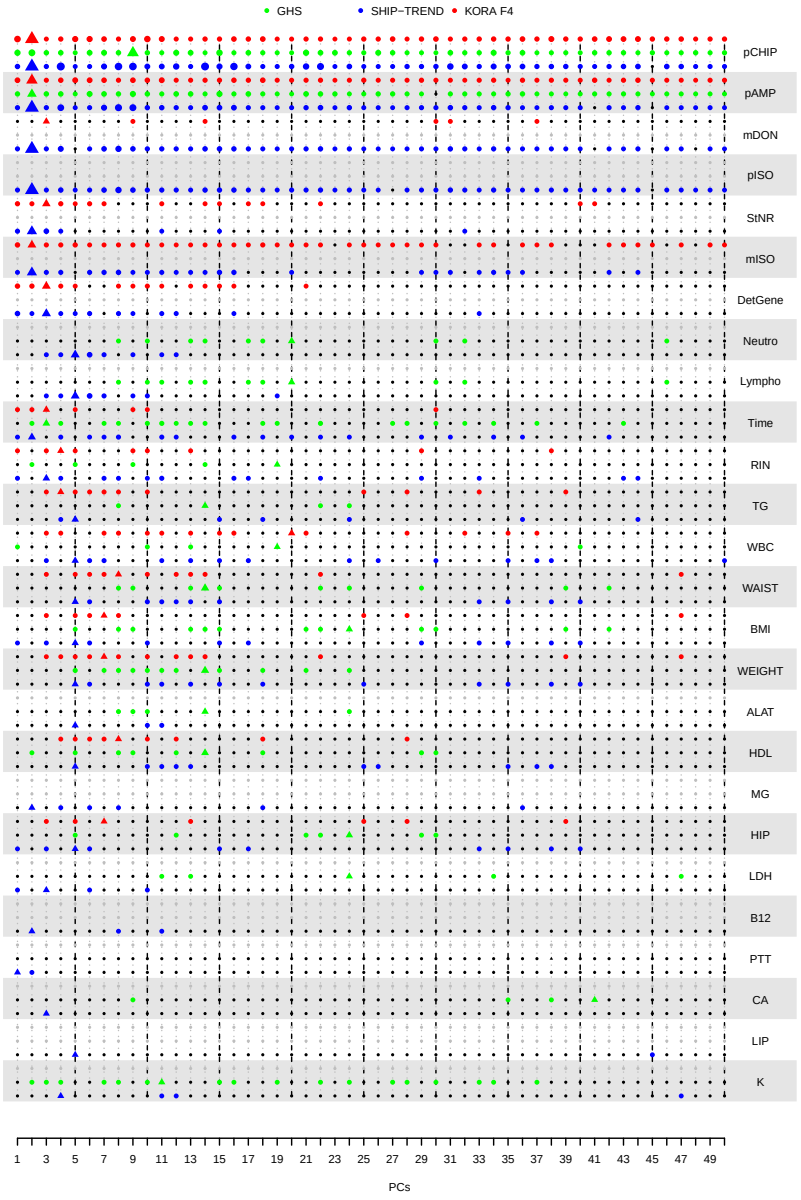

Figure 4.9.: Correlation of Eigen-genes with several factors in KORA F4, GHS, and SHIP-TREND:
The smaller the p-value the larger the dots. The smallest p-value per variable is indicated with a triangle. Grey dots show a missing trait in the cohort.
On the y-axis are depicted: Illumina chip (pCHIP), RNA amplification batch (pAMP), month of blood sampling (mDON), RNA isolation batch (96 well plate) (pISO), signal-to-noise ratio (StNR), month of RNA isolation (mISO), number of detected genes (DetGene), percentage of neutrophils (Neutro), percentage of lymphocytes (Lympho), storage time (Time), RNA integrity number (RIN), serum triglyceride concentrations (TG), white blood cell count (WBC), waist circumference (WAIST), body mass index (BMI), body weight (WEIGHT), alanine aminotransferase concentrations (ALAT), high density lipoprotein concentrations (HDL), serum magnesium concentration (MG), hip circumference (HIP), lactate dehydrogenase concentrations (LDH), vitamin B12 concentrations (B12), partial thromboplastin time (PTT), serum calcium concentrations (CA), serum lipase concentrations (LIP) and serum potassium concentrations (K).

Figure 4.10). Adding more than 50 Eigen-genes to the model did not further decreased the unexplained variance.

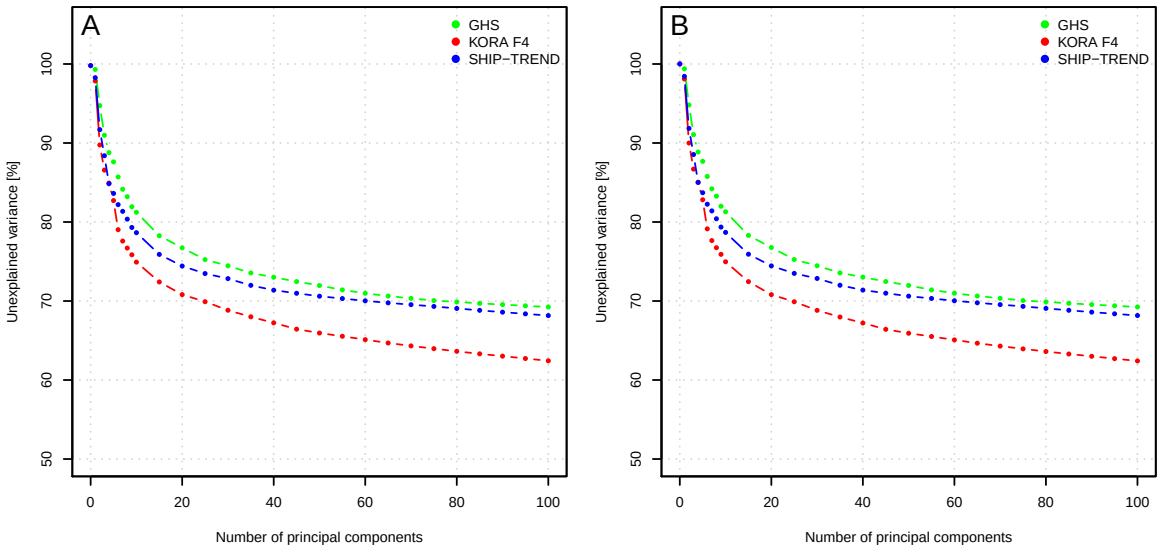

Figure 4.10.: Mean unexplained variance in KORA F4, SHIP-TREND, and GHS for BMI (A) and the random phenotype (B):
A linear regression was conducted in all three cohorts with expression levels as dependent variable and BMI or the random phenotype as independent variable. Systematically more Eigen-genes were added to the model, the adjusted R^2 values were saved and the mean was calculated for each number of Eigen-genes.

In addition we compared the effect sizes, standard errors, and p-values when adding different covariates to the linear model. According to the results that we obtained from the Eigen-R^2 calculation we used mainly technical variables (because they explain most of the variation in the expression data) to further reduce the residual variance. We therefore compared ten different models for BMI and the random phenotype:
gene expression $\sim$ BMI/random phenotype +

- nothing
- age + sex
- age + sex + technical variables (amplification plate, RIN, sample storage time)
- technical variables (amplification plate, RIN, sample storage time)
- technical variables + first Eigen-gene
- technical variables + number of significant detected genes
- technical variables + Signal-to-noise-ratio (comparison of measured signal with background level)

removed Eigen-genes	BMI	Random Phenotype
0	1.00	1.00
1	0.98	0.98
2	0.90	0.90
3	0.87	0.87
4	0.85	0.85
5	0.83	0.83
6	0.79	0.79
7	0.78	0.78
8	0.77	0.77
9	0.76	0.76
10	0.75	0.75
15	0.72	0.72
20	0.71	0.71
25	0.70	0.70
30	0.69	0.69
35	0.68	0.68
40	0.67	0.67
45	0.66	0.66
50	0.66	0.66
55	0.66	0.66
60	0.65	0.65
65	0.65	0.65
70	0.64	0.64
75	0.64	0.64
80	0.64	0.64
85	0.63	0.63
90	0.63	0.63
95	0.63	0.63
100	0.62	0.62

Table 4.4.: Mean unexplained variance for BMI and the random phenotype in KORA F4: A linear regression with expression levels as dependent variable and BMI or the random phenotype as independent variables was conducted. Systematically different numbers of Eigen-genes were added to the model, the adjusted R^2 values were saved, and the mean was calculated for each number of Eigen-genes.

- 50 Eigen-genes
- age + sex + technical variables + cell types
- technical variables + all non technical variables (white blood cell count, red blood cell count, hematocrit, and mean platelet in KORA F4)

The mean standard errors for all models were calculated for all three cohorts (Table 4.5). The lowest mean standard error could be observed for the 50 Eigen-genes for the random phenotype in all three cohorts. The standard errors were reduced by 21%, 27%, and 25% compared to the unadjusted models in SHIP-TREND, KORA F4, and GHS, respectively.

Phenotype	additional covariates (besides phenotype)	Mean SE		
		SHIP-TREND	KORA F4	GHS
Random phenotype	none	0.0060256	0.007050737	0.005558932
	age+sex	0.0060040	0.006928487	0.005541641
	age+sex+technical	0.0054934	0.006401871	0.005288458
	technical	0.0055128	0.006413871	0.005305218
	technical+Eigen-gene1	0.0054879	0.00637871	0.005004323
	technical+detected genes	0.0054451	0.006270553	-
	technical+*signal-to-noise ratio*	0.0054482	0.006290344	-
	50 Eigen-genes	0.0047421	0.005124327	0.004193441
	age+sex+technical+cell types	0.0054243	-	-
	technical+all non technical	0.0055731	-	-
BMI	none	0.0013035	0.001547339	0.001149232
	age+sex	0.0013500	0.001547741	0.001172341
	age+sex+technical	0.0012342	0.001425894	0.001121818
	technical	0.0011932	0.001425156	0.001099153
	technical+Eigen-gene1	0.0011921	0.001416861	0.001094797
	technical+detected genes	0.0011854	0.001409976	-
	techCov+*signal-to-noise ratio*	0.0011949	0.001413946	-
	50 Eigen-genes	0.0012536	0.001264766	0.001055827
	age+sex+technical+cell types	0.00123254	-	-
	technical+all non technical	0.01305295	-	-

Table 4.5.: Mean standard errors after different covariate adjustments:
The mean standard errors (for all probes) were calculated in each cohort for linear models with gene expression as dependent and BMI or the random phenotype as independent variable. Different covariates were added to the model. Missing covariates in cohorts are indicated with a dash.
cell types: percentage of lymphocytes, neutrophils, monocytes, eosinophils and basophils, detected genes: number of genes with detection p-value less than 0.01, technical: RNA amplification batch, RIN, and sample storage time, non-technical: parameters with an Eigen-$R^2 > 0.3\%$ in SHIP-TREND, *signal-to-noise* ratio: Comparison of measured signal to background level.

For BMI the mean standard error decreased with adjusting for 50 Eigen-genes. Three Eigen-genes (5, 7 and 28) are correlated with BMI (see Figure 4.8) and further all significant associations were lost if we adjusted for these three Eigen-genes. For almost all available

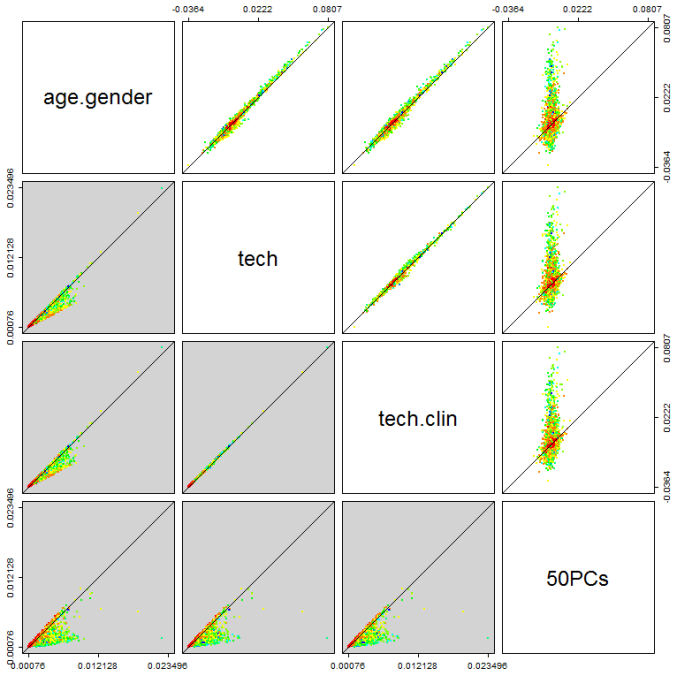

Figure 4.11.: Comparison of different adjustments in KORA F4 for BMI:
Linear models with gene expression as dependent and BMI as independent variable were calculated. The models were adjusted for age + gender (age.gender), amplification plate + RIN + sample storage time (tech), amplification plate + RIN + sample storage time + white blood cell count + red blood cell count + hematocrit + mean platelet (tech.clin) and, 50 Eigen-genes (50PCs), respectively. In the gray boxes the standard errors and in the white boxes the effect sizes of these different models are shown.

covariates significant associations with at least one Eigen-gene were observed and therefore we do not recommend to include them in the linear model.
Since adjustment for the technical variables RNA amplification batch, RIN, and the sample storage time (time between blood sampling and RNA isolation in SHIP-TREND and KORA F4 and time between RNA isolation and amplification in GHS) decreased the standard errors for both phenotypes by about 8% we decided to use those three variables for all further analyses.
The mean standard error increased when we adjusted for sex and age in the BMI regression model in comparison to the unadjusted model. This might be explained by the correlation between BMI and sex (p-value of t-test = 0.01699 in KORA F4).

For further analyses we wanted to establish a KORA F4 data set that was already adjusted for the three relevant technical variables (RNA amplification batch, RIN, and the sample storage time). Firstly, I adjusted the expression values for KORA S4, F4, and F4 OGTT data together for these variables by calculating a linear model with gene expression as dependent and the technical variables as independent variables and kept the residuals from these models. To test whether the explained variance was reduced in the residuals I calculated the Eigen-R^2 for the amplification plate as this was the variable with the highest explained variance in KORA F4 (Eigen-R^2 = 24.30%). In the residuals the Eigen-R^2 in KORA F4 for the amplification plate was still 3.83% and 11.61% in S4. This might be due to the unbalanced distribution of the samples from the three studies on the amplification plates (see Table 4.6). Especially the S4 samples were distributed unequally.
Therefore I secondly split the data set in the three studies (S4, F4, and F4 OGTT) and then adjusted for the technical variables. This approach reduced the explained variance by the amplification plate to $2.06 * 10^{-16}$% in S4 and $2.41 * 10^{-16}$% in F4, respectively.
Applying the above described method to the three cohorts separately reduced the explained variance of the amplification plate more efficiently. The residuals of this second approach were regarded as the best optimized expression data set and was used for all following projects.

4.3.4. SNPs in probes

SNPs that are located within a probe sequence on the array could cause a decrease in the hybridization efficiency and reduce the signal intensities due to a lower binding efficiency. This could cause false positive results when analyzing these probes. Therefore 8,898 probes that cover exactly one exon and could be mapped uniquely to an annotated UCSC transcript ((Kent et al., 2002), (Dreszer et al., 2012)) were analyzed systematically to test whether the SNP influences the expression level. Of these transcripts 3,376 (38%) contain at least one SNP according to the 1000 Genomes[3] reference panel (Altshuler et al., 2010).
In SHIP-TREND for 986 individuals the genotyping information was known. 24% of the probes contained a polymorphic SNP with a minor allele frequency greater than 0.01 (7%) and 0.05 (4%) respectively. For 1,561 probes containing 2,128 SNPs the effect of the SNP on the expression level was analyzed using linear regression adjusted for sex, age, and the first 50 Eigen-genes. Out of these, 55% of the SNPs were associated with an decreased signal

[3]The 1000 Genomes Project was the first project with the aim to sequence a large number of individuals to provide public available data to analyze the genomic variation in humans.

Plate	all probes	used probes	S4	F4	F4 OGTT
Plate01	87	55	7	28	20
Plate02	60	60	18	24	18
Plate03	108	81	10	36	35
Plate04	111	80	24	28	28
Plate05	100	83	20	33	30
Plate06	90	85	32	28	25
Plate07	96	88	22	34	32
Plate08	96	1	0	0	1
Plate09	36	32	11	10	11
Plate09a	51	46	15	18	13
Plate10	79	72	10	27	35
Plate11	100	94	38	35	21
Plate12	98	84	22	33	29
Plate13	97	92	52	20	20
Plate14	96	92	0	53	39
Plate15	99	96	0	56	40
Plate16	97	79	2	44	33
Plate17	70	55	2	36	17
Plate18	96	94	0	51	43
Plate19	98	86	2	45	39
Plate20	121	78	11	40	27
Plate21	92	91	0	52	39
Plate22	100	96	0	56	40
Plate23	86	64	2	38	24
Plate24	96	94	9	48	37
Plate25	67	48	11	20	17
Plate26	97	70	8	29	33
Plate27	91	76	27	26	23
Plate28	96	81	66	9	6
Plate29	96	88	88	0	0
Plate30	76	66	0	29	37
Plate31	96	79	79	0	0

Table 4.6.: Distribution of KORA samples on amplification plates:
The distribution of all KORA samples having gene expression data available on the amplification plates is shown. The numbers of “all probes” and “used probes” differ because of the excluding of some samples due to bad quality.

intensity. Surprisingly, for 45% of the SNPs the effect was in the opposite direction than expected.
KORA F4 individuals were genotyped with different arrays. Only 70 of the 2,128 SNPs were investigated. But it was seen that the effects are in both directions, negative (55%) and positive (see Figure 4.12).
In summary, the results look more randomly than systematically and we cannot exclude a decrease in the hybridization efficiency without seeing a systematic scheme. Conclusively, we recommend not to exclude probes including SNPs within their sequence but point towards an additional investigation when this probe shows significant further results.

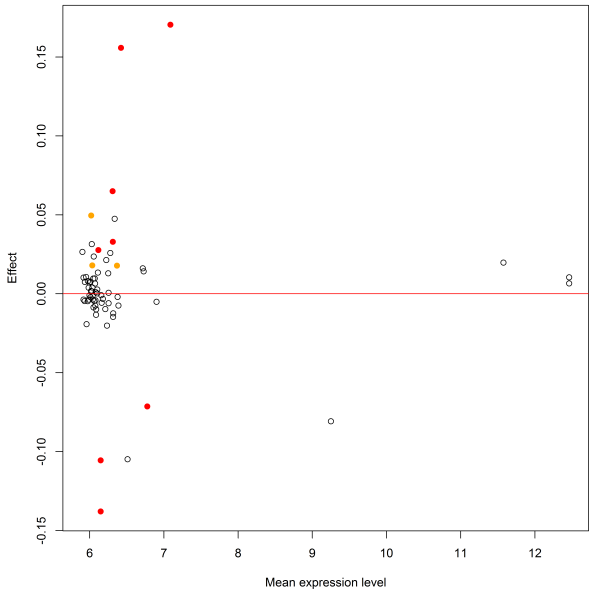

Figure 4.12.: Effect of SNPs within a probe sequence on expression levels in KORA F4: The mean expression levels of all probes containing a SNP (66 probes) in the KORA F4 data set is plotted against the effect sizes of the linear model with expression level as dependent variable and SNP as independent variable. Each spot displays the effect of a SNP on a probe. Associations with significant p-values after Bonferroni correction ($p< 2.3 * 10^{-5}$) are colored in red and with p-values below 0.05 are colored in orange. For 55% the effect of the SNPs was negative, meaning that the SNP was associated with a decreased signal intensity.

4.3.5. Annotation

Illumina arrays contain both, probes that map to a well annotated transcript and probes which map to a potential transcript. The probe annotation from Illumina is only updated when a new array is introduced and it is normally not up to date. Therefore Dr. Alexander Teumer mapped all probe sequences to the available mRNA sequences of the UCSC genome annotation database (version 12/06/2009, February 2009 assembly of the human genome, HG19). Out of all probes, 28,691 probes could be perfectly mapped to known transcripts or annotated RefSeq genes.

The updated annotation file provides the date genome positions of all mappable probes and is now in use for all analyses.

4.4. Summary and discussion

In the beginning of my PhD there were no other large population-based studies with expression data and no standards for the quality control. One aim of my thesis was to optimize expression data for analysis. Since our research is in a very active and new field some of the methods we tried in the beginning are no longer up to date, for example the exclusion of expression probes to reduce the number of tests. In KORA F3 probes that are not detected significantly in more than 5% of the samples were excluded from further analysis. This approach is nowadays no longer recommended because even low signals reflect low expression values and several studies have shown that results based on low signal probes contain valuable information and can be replicated.

The improvement in experimental processing steps from KORA F3 to KORA F4 are shown in Figure 4.13. The variation in the detected genes was smaller in KORA F4 although it consists of more samples (standard deviation in KORA F3 is 2,587 in KORA F4 922). In KORA F3 each sample was transported separately to the Helmholtz Center and not all samples were treated identically. The KORA F4 samples were all frozen in the study center in Augsburg and then transported together to the Helmholtz Center. This is reflected in the quality of the RNA. The KORA F4 data set is more homogenous than the F3 data set as the selection criteria for samples was more stringent (RIN $>$ 6 and number of detected genes $>$ 6,000).

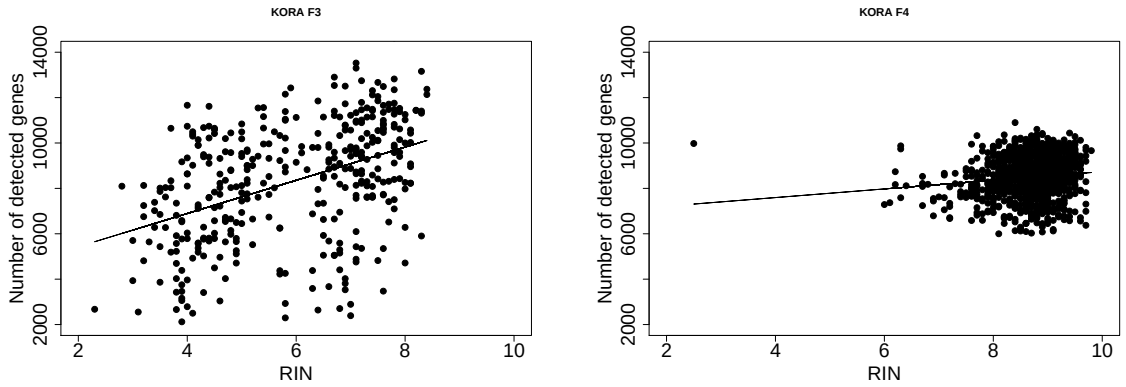

Figure 4.13.: RIN versus number of detected genes in KORA F3 and KORA F4:
The correlation between RIN and detected genes is significant for both cohorts (p-value = $2.99 * 10^{-18}$ for KORA F3 and p-value = $7.63 * 10^{-5}$ for KORA F4), but if the RIN increases by one in KORA F3 the average number of detected genes increases by 731, in KORA F4 just by 190. The low RIN in KORA F4 was a measurement error from the Bioanalyzer and therefore this sample was also used for further analysis.

Currently the most frequently used normalization method for large data sets is the quan-

tile normalization. In 2010, when the KORA F4 samples were ready to analyze 1,240 samples with expression data were published by Dubois et al. (2010). They used the same array and suggested the quantile normalization because it works without any assumptions.

When joining the MetaXpress consortium it was necessary to agree on a common protocol for all quality steps. As SHIP-TREND samples were measured in Munich experimental settings of these two data sets were almost identical. When SHIP-TREND samples were measured in Munich, KORA data were already ready-to-use and therefore they adopted most of our quality steps to their data set. The exclusion of outliers and potentially mixed-up samples was done in each cohort separately and no common strategy for this procedure was developed.
The normalization was the same for each cohort. The only difference was that both other cohorts first normalized the data and then logarithmized it. As this does not make a difference for all further analysis we changed the order for the KORA F4 data to ensure better comparability.
Since no guideline for quality control and preprocessing of expression data measured with the Illumina HumanHT12 BeadChip was available it was developed during several meetings with the data analysts working in the MetaXpress consortium.

1. Installation of the GenomeStudio Software from Illumina is necessary to obtain the expression levels from the scanned data.
2. Create a new project and select all samples that should be analyzed. If the number of samples is very high this step has to be done several times.
3. Impute all missing values (normally less than ten values per sample).
4. Ensure the quality of each sample by excluding samples with less than 6,000 significantly detected genes (detection p-value < 0.01).
5. Export the expression values with an unique identifier (e.g. Illumina ProbeID) as csv-file.
6. Install R from `http://www.r-project.org/` and install the Bioconductor package `lumi` with the following R command:

```
source("http://bioconductor.org/biocLite.R")
biocLite("lumi")
```

7. Load the expression data to R by typing:

```
data <- read.csv("ExpressionData.csv")
```

8. Normalize the data (quantile normalization) and perform a log2 transformation by using the `lumi` package:

```
library(lumi)
norm.data <- lumiN(data, method = "quantile")
norm.data.log <- lumiT(norm.data, method = "log2")
```

9. Use the expression levels of genes on sex-specific chromosomes to identify mixed samples.

10. When analyzing associations between expression levels and phenotypes always consider RNA amplification batch, RIN, and the sample storage time as covariables. For eQTL studies use the Eigen-vectors, the so called Eigen-genes to remove most of the technical variation from the data and increase the number of significant eQTLs.

11. Use an updated annotation file.

12. As probes containing SNPs are not excluded, have a closer look at these probes after analysis of the data.

5. Association studies

5.1. Gene expression and blood pressure related phenotypes

High blood pressure is a major risk factor for several cardiovascular diseases. The phenotypes "systolic" and "diastolic blood pressure", as well as "pulse pressure" (difference between diastolic and systolic blood pressure) were chosen to be analyzed within the MetaXpress Consortium. On August, 19th 2015, 51 studies identifying 416 associations concerning several blood pressure traits were found in the GWAS catalog (for example (Newton-Cheh et al., 2009), (Levy et al., 2009), (Ehret et al., 2011)). In spite of a lot of known loci the proportion of explained variance is very small. The International Consortium of Blood Pressure (ICBP) calculated an explained variance for all 29 identified and reported loci of about one percent in the general population (Ehret et al., 2011).
We hypothesized that an association study between gene expression and blood pressure might help to close the gap between estimated and explained variance for blood pressure.
All of the so far published studies analyzing gene expression and blood pressure did not replicate their results or had small sample sizes:

- Leonardson et al. (2010) analyzed 40 Caucasian males from the greater Reykjavik area in Iceland and identified 896 significant associations with systolic and 3,329 with diastolic blood pressure using a p-value threshold of 0.0001. No replication of these hits was performed.
- Zeller et al. (2010) analyzed 1,490 monocyte samples from Germany (GHS study, see Section 2.1.4) and identified 48 expression traits that were associated with systolic blood pressure and 18 with diastolic blood pressure, respectively. These results were also not replicated.
- Bull et al. (2004) compared gene expression levels between 15 patients suffering pulmonary arterial hypertension and 6 healthy controls. Of all analyzed genes, 28 had a p-value below 0.01. Two genes could be verified using qPCR.
- Korkor et al. (2011) compared three patients suffering hypertension to three healthy controls. By this, 49 differentially expressed genes were identified. Ten of these genes could be verified using qPCR.

5.1.1. Results from KORA F3/F4

In KORA F3 and F4 the association between gene expression levels and systolic and diastolic blood pressure was analyzed. As individuals taking anti-hypertensive drugs could falsify the results, the analyses were also repeated without these individuals (N=173 in KORA F3 and N=571 in KORA F4). In KORA F3 377 samples and in KORA F4 989 samples were analyzed. The linear regression model for both cohorts was the following:

$$expression\,level \sim\ phenotype + age + sex + BMI + technical\,variables$$

with the technical variables being RIN, sample storage time, and amplification plate in KORA F4 and RIN in KORA F3.
When analyzing all KORA F4 samples only one gene was significantly associated with systolic (p-value = $2.84 * 10^{-08}$) and diastolic (p-value = $4.22 * 10^{-08}$) blood pressure. The gene is called *FOSB* and is located on chromosome 19. In 990 SHIP-TREND samples p-values for this gene are $1.51 * 10^{-05}$ (diastolic blood pressure) and $1.47 * 10^{-05}$ (systolic blood pressure) respectively.
When including only subjects that did not take anti-hypertensive drugs no gene was significantly associated with systolic or diastolic blood pressure. The p-values for *FOSB* were $2.32 * 10^{-04}$ (systolic blood pressure) and $2.95 * 10^{-04}$ (diastolic blood pressure), respectively.

Cohort	KORA F3	KORA F4
Sample size	377	989
diastolic BP (mm Hg)	83.2	74.0
systolic BP (mm Hg)	136.7	128.7
Samples taking drugs	173	571
diastolic BP (mm Hg)(without drugs)	84.97	75.4
systolic BP (mm Hg)(without drugs)	136.6	128.5

Table 5.1.: Study description of KORA F3/F4 for blood pressure related phenotypes

5.1.2. Results from MetaXpress

Altogether, results of the association studies on blood pressure phenotypes in whole blood of KORA F4 participants were not promising. In contrast, several significant hits could be identified in the monocyte sample of the GHS. Therefore, monocyte samples plus samples from the US Multi-Ethnic Study of Atherosclerosis (MESA) cohort were used as discovery panel (N= 2,549) and whole blood samples from KORA F4 and SHIP-TREND were used as replication cohorts (Mueller et al., 2014). All four studies used the same linear regression model for the three phenotypes diastolic and systolic blood pressure, as well as pulse pressure:

$$expression\,level \sim\ phenotype + age + sex + BMI + technical\,variables$$

Results from GHS and MESA were meta-analyzed and an association was called significant if the Benjamini-Hochberg adjusted p-value was below 0.05. P-values of these associations were assessed in KORA F4 and SHIP-TREND. If the p-values in both studies were below 0.05 and the direction of the effect was consistent, the gene was chosen for validation in a clinical trial (TEAMSTA[1]). A qPCR was performed to measure the expression level of each of the eight candidate genes in 613 hypertension patients before and after a six-month treatment with anti-hypertensive drugs. For the analysis the patients were divided in responders (systolic blood pressure decreased $\geq$ 10mmHg) and non-responders (systolic blood pressure decreased $\leq$ 2mmHg) and in both groups both time-points were compared separately.
All eight genes showed significant differences between start and end point of the clinical trial

[1] "This was a multicenter, multinational, 8-week randomized, double-blind, parallel-group study that evaluated the efficacy and safety of two SPCs of telmisartan/amlodipine (T/A) compared with amlodipine monotherapy in patients with uncontrolled hypertension." (Neldam et al., 2011)

(p-value $< 0.05/8=0.00625$) . These eight genes were *CEBPA*, *CRIP1*, *F12*, *LMNA*, *MYADM*, *TIPARP*, *TPPP3*, and *TSC22D3*.

5.1.3. Results from CHARGE consortium

The blood pressure project in CHARGE included 7,017 samples from six population-based studies, namely Framingham Heart Study (FHS), EGCUT, Rotterdam Study (RS), InCHIANTI, SHIP-TREND and KORA F4 (Table 5.2).
Gene expression levels were associated with three phenotypes, namely systolic (SBP) and diastolic (DBP) blood pressure, and hypertension (HTN) with hypertension being defined as SBP $\geq$ 140 mm Hg or DBP $\geq$ 90 mm Hg. For each expression probe a linear model adjusted for age, sex, BMI, cell counts (if available) and technical covariates (RIN, sample storage time, and amplification plate in KORA) was applied. FHS used a mixed model to adjust for family structure.

	FHS	EGCUT	RS	InCHIANTI	KORA F4	SHIP-TREND
Sample size	3,679	972	604	597	565	600
Age	51±12	36±14	58±8	71±16	72±5	46±13
SBP (mm Hg)	118±15	122±16	132±20	132±20	129±21	120±15
DBP (mm Hg)	74±9	76±10	82±11	78±10	73±11	75±9
Hypertension	11%	19%	35%	45%	26%	12%

Table 5.2.: Characteristics of the six study cohorts included in meta-analysis on blood pressure related phenotypes:
Individuals receiving anti-hypertensive treatment were excluded from the analysis. Hypertension was defined as SBP $\geq$ 140 mm Hg or DBP $\geq$ 90 mm Hg.

As gene expression levels of all six cohorts were measured using two different platforms (FHS used Affymetrix, while all others used Illumina) the analysis was conducted in two steps (Figure 5.1). At first, panels using different platforms were analyzed separately and the significant hits were replicated in the panel using the other platform and secondly, both panels were meta-analyzed (This was possible for an intersecting set of 7,717 genes that were measured with both platforms). Results are summarized in Table 5.3.

The meta-analysis of all six cohorts resulted in 34 significant differentially expressed genes (Table 5.4) associated with either SBP (21), DBP (20), or HTN (5), whereat expression levels of ten genes are associated with more than one phenotype. For 33 of these genes a *cis*- and for 26 a *trans*-eQTL in whole blood was identified in an eQTL study (Westra et al., 2013).

Gene	CHR	FHS p-value	Illumina p-value	Meta p-value
SBP Signature genes				
SLC31A2	9	1.2E-13	9.9E-11	<1E-16
MYADM	19	2.2E-14	2.2E-12	<1E-16
DUSP1	5	1.1E-08	3.7E-07	2.0E-14
TAGLN2	1	1.0E-06	1.3E-06	5.8E-12
CD97	19	1.4E-07	1.6E-05	1.0E-11
BHLHE40	3	4.3E-06	6.4E-07	1.2E-11

MCL1	1	7.5E-07	1.5E-06	1.4E-11
PRF1	10	2.5E-09	1.0E-03	1.6E-11
GPR56	16	3.5E-09	3.0E-03	3.9E-11
PPP1R15A	19	1.7E-09	2.8E-05	1.5E-08
FGFBP2	4	5.8E-06	1.5E-03	3.3E-08
GNLY	2	3.6E-05	3.0E-04	4.0E-08
FOS	14	1.6E-11	3.6E-05	4.8E-08
NKG7	19	1.9E-05	8.8E-03	9.4E-07
GRAMD1A	19	2.1E-05	1.8E-02	1.1E-06
GLRX5	14	1.3E-05	3.5E-02	1.5E-06
TMEM43	3	3.0E-04	2.4E-03	2.3E-06
TIPARP	3	1.3E-07	3.3E-04	2.6E-06
AHNAK	11	4.1E-04	4.0E-03	5.2E-06
PIGB	15	5.3E-04	1.9E-03	6.1E-06
TAGAP	6	5.7E-12	7.1E-04	6.4E-06
DBP Signature genes				
BHLHE40	3	2.3E-06	2.8E-06	2.7E-11
ANXA1	9	1.2E-09	6.3E-03	6.5E-11
PRF1	10	3.2E-07	5.7E-04	6.7E-10
KCNJ2	17	3.9E-06	2.6E-04	4.9E-09
CLC	19	2.6E-06	5.7E-04	5.8E-09
CD97	19	1.6E-06	1.1E-03	7.4E-09
IL2RB	22	3.0E-06	2.4E-03	2.5E-08
S100A10	1	2.4E-07	9.9E-03	4.0E-08
GPR56	16	1.1E-06	1.7E-02	5.5E-08
TIPARP	3	1.3E-04	2.8E-04	1.4E-07
HAVCR2	5	3.8E-04	1.8E-04	2.4E-07
PTGS2	1	2.2E-05	9.0E-03	1.0E-06
MYADM	19	1.7E-08	8.6E-05	1.1E-06
ANTXR2	4	5.2E-06	5.5E-02	1.7E-06
OBFC2A	2	7.2E-06	3.8E-02	1.8E-06
GRAMD1A	19	1.4E-05	7.8E-02	2.8E-06
ARHGAP15	2	1.1E-03	1.5E-03	5.2E-06
FBXL5	4	2.1E-05	5.5E-02	5.3E-06
SLC31A2	9	1.0E-08	2.6E-03	5.4E-06
VIM	10	5.5E-06	2.0E-01	6.2E-06
HTN Signature genes				
SLC31A2	9	1.9E-05	2.1E-06	1.8E-10
MYADM	19	1.2E-08	6.2E-04	3.0E-07
TAGAP	6	3.2E-05	5.3E-03	7.3E-07
GZMB	14	1.1E-11	9.6E-04	1.4E-06
KCNJ2	17	8.4E-04	5.5E-04	1.7E-06

Table 5.4.: Significantly associated genes for blood pressure related phenotypes: Linear models were calculated for three different phenotypes (SBP, DBP, and hypertension) in FHS, all cohorts that used Illumina arrays, and in meta-analysis of all cohorts.

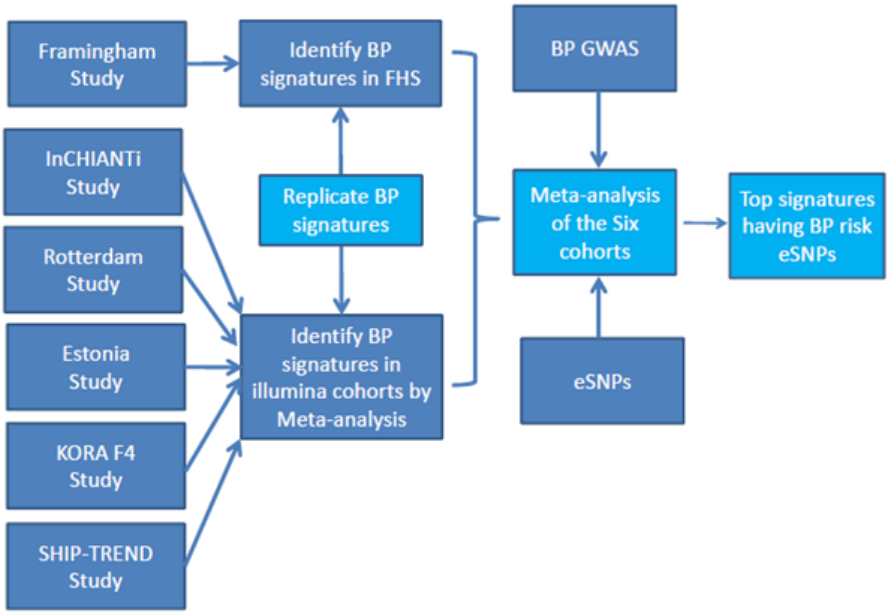

Figure 5.1.: Analysis framework for gene expression study on blood pressure related phenotypes

	SBP	DBP	HTN
Illumina cohorts	6	1	1
replicated	6	1	1
Affymetrix cohort	73	31	8
replicated	10 (of 55)	5 (of 22)	2 (of 8)
Meta-analysis	21	20	5

Table 5.3.: Results from gene expression study on blood pressure related phenotypes: Number of significant genes for all three phenotypes using two different expression platforms. "Significant" was defined as Bonferroni corrected p-value below 0.05. Not all significant genes from the Affymetrix panel were available on the Illumina array. The number of available genes is displayed in brackets.

5.2. Gene expression and aging

As demonstrated in Section 3.2 aging is one of the main factors for age-related diseases like Parkinson's disease, Alzheimer or cardiac infarction. Therefore, many of the studies listed in the GWAS catalog (n=98) are linked to aging (August, 10th 2015). In comparison, only 52, 36, and 9 studies are listed for BMI, height, and hair color, respectively. But as genetic is just a static investigation, it is very interesting to analyze gene expression levels as an indicator for gene activity.

5.2.1. Results from KORA F3

In the beginning of the gene expression era assumptions were not very stringent. The first published results from KORA F3 (Mehta, 2009) on the association between gene expression and aging did not consider any confounders. The model was:

$$expression\,level \sim age$$

This model was applied to the 13,767 probes that were significantly detected in more than 5% of the samples. This resulted in eleven probes significantly associated with aging after Bonferroni correction (see Table 5.5).

Gene	CHR	ProbeId	old gene name from Illumina	p-value
LRRN3	CHR7	$ILMN\_1773650$	*LRRN3*	1.33E-08
SGK223	CHR8	$ILMN\_1766236$	*DKFZP761P0423*	2.92E-07
GPR18	CHR13	$ILMN\_1780368$	*GPR18*	4.07E-07
CD248	CHR11	$ILMN\_1726589$	*CD248*	4.16E-07
ANXA2R	CHR5	$ILMN\_1675465$	*C5ORF39*	9.14E-07
CCR7	CHR17	$ILMN\_1715131$	*CCR7*	1.53E-06
OCIAD2	CHR4	$ILMN\_1700306$	*OCIAD2*	1.76E-06
		$ILMN\_1674983$	*LOC387841*	2.79E-06
FBL	CHR19	$ILMN\_1719205$	*FBL*	3.36E-06
PCED1B	CHR12	$ILMN\_1712431$	*FAM113B*	3.37E-06
	CHR6	$ILMN\_1804935$	*VNN3*	3.49E-06

Table 5.5.: Significantly associated genes in KORA F3 with aging

Expression levels of these eleven probes were used to create a prediction model for age. All significantly associated probes were simply included in one linear model:

$$\begin{aligned} age = \beta_1 * LRRN3 + \beta_2 * DKFZP761P1 + \beta_3 * GPR18 + \beta_4 * CD248 + \beta_5 * LOC389289 \\ + \beta_6 * CCR7 + \beta_7 * LOC387841 + \beta_8 * OCIAD2 + \beta_9 * VNN3 + \beta_{10} * LY9 \\ + \beta_{11} * FAM113B + \epsilon \end{aligned}$$

For 25% of the samples the difference between chronological and predicted age was less than 2.5 years. For 50% the difference was between 2.5 and 8 years and for the remaining

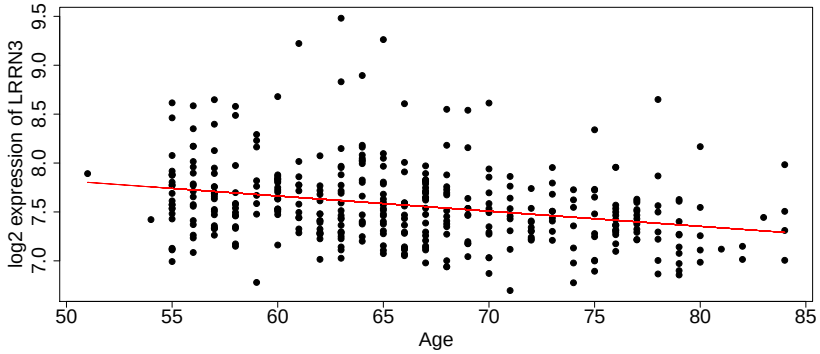

Figure 5.2.: Age-specific gene expression in KORA F3:
The age is plotted versus the expression level of the most significant gene: *LRRN3*

samples it was more than 8 years.

For a better comparison of these results the analysis was repeated using all available probes on the array and using sex and RIN (as technical variable for KORA F3) as covariables. This resulted in five significant associations between gene expression and age (*CD248*, *SGK223*, *GPR18*, *LRRN3*, and *NELL2*).

5.2.2. Results from KORA F4

KORA F4 expression data are highly affected by batch effects. Nevertheless, the associations between gene expression and aging were calculated identically as in KORA F3, to show that missing covariates that influence gene expression levels could lead to false positive associations. Therefore, the linear model was calculated once without any covariables and once adjusted for sex and technical covariables (RIN, sample storage time, and amplification plate). To compare the results from F4 to F3 the models were also calculated in a random subset of KORA F4 with the same sample size as in F3 (N=381). Results of these comparisons are shown in Table 5.6.

	Unadjusted	Adjusted	Overlap
N=993	370	194	83
N=381	74	45	13
Overlap	64	31	

Table 5.6.: Number of genes significantly associated with age in KORA F4

Due to the known strong batch effects in KORA F4 results of the unadjusted model might be false-positive findings and not directly comparable to the results of KORA F3. This demonstrated that the consideration of technical variables bisected the number of significant hits. It also shows that in the KORA F4 data more probes are associated with age, although the age distribution in both data sets is similar. There is even a wider age range in KORA F3

compared to KORA F4 (see Figures 5.3 and 5.4).

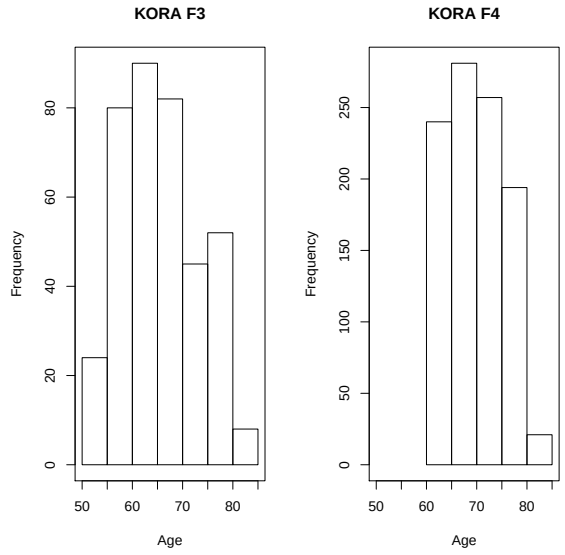

Figure 5.3.: Histogram of age distribution in KORA F3 and F4

When using the model adjusted for sex and technical variables in KORA F4, 194 expression probes were significantly associated with age. Four of these loci were also identified in KORA F3 (*CD248*, *GPR18*, *LRRN3* and *NELL2*). For *SGK223* the p-value was marginally not significant (p = $1.67 * 10^{-6}$). A Bonferroni threshold of $1.02 * 10^{-6}$ was applied.

5.2.3. Results from CHARGE consortium

The association between gene expression and aging was the first project of the gene expression working group of the CHARGE consortium. Six independent cohorts with expression levels measured in whole blood were included in the analyses resulting in a total of 7,074 samples:

- EGCUT: N=1,086 (described in 2.1.3)
- FHS-2nd generation: N=2,446 (Framingham Heart Study, consists of three generations of participants, in the 2nd generation (1971) the offspring of the first generation from 1948 were investigated)
- InCHIANTI: N=698 (Invecchiare in Chianti - aging in the Chianti area, population-based prospective study in Italy)
- KORA F4: N=993 (described in 2.1.1)
- RS III: N=881 (Rotterdam Study III, population-based prospective study, all samples are older than 45)
- SHIP-TREND: N=970 (described in 2.1.2)

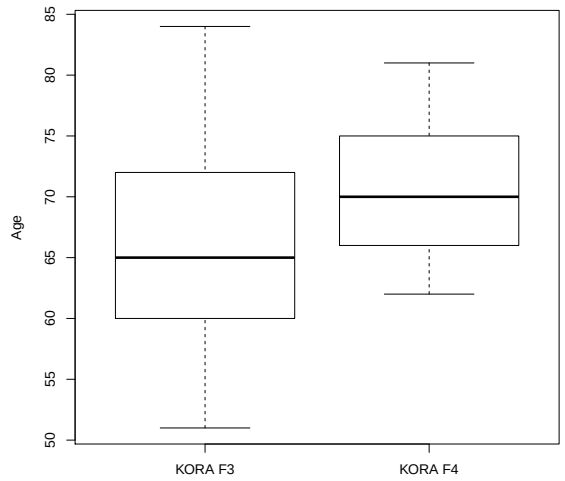

Figure 5.4.: Boxplot of age in KORA F3 and F4

5.2.3.1. Association between gene expression and age

Out of 11,908 genes with significant expression levels, 2,228 were significantly associated with age in the discovery stage with p-value $< 4.2 * 10^{-6}$. The association was tested using linear models adjusted for technical variables, cell counts (not available in KORA F4), smoking, and fasting status (in KORA F4 there were eight non-fasting samples). For replication, 7,909 whole-blood samples from seven different cohorts were available and 1,497 genes could be replicated with p-value $< 2.2 * 10^{-5}$.
For the replicated genes it was tested whether they were also significantly associated with age in samples with different ancestry or in different tissues (Table 5.7). Due to the low number of samples for African Americans (n=359) and two brain tissues (n=394) the number of genes associated with age was quite low (27%, 19%, and 26%, respectively). But for the larger samples of Native (n=1,457) and Hispanic (n=1,244) Americans the rates were quite high (71% and 74%, respectively). To investigate whether the effects of age on gene expression levels are also existent in populations with different ancestry or in other tissues higher sample sizes comparable to that of discovery cohorts will be required.

5.2.3.2. Age prediction

The prediction of age using expression levels results in the so-called "transcriptomic age". The difference between the chronological and the transcriptomic age (here called Δage) could be explained by the biological age.
For the age prediction in all participating CHARGE cohorts a formula was applied for each cohort separately by using all expressed genes and calculate the meta-analysis without the relevant cohort. That way, an own prediction formula was generated for each individual of each cohort:

$$Z = \sum x_{v(i)} \hat{b}_{R(i)}$$

Ancestry/tissue	sample size	expressed genes	genes with p<0.05
Native Americans	1,457	95%	1,005 (71%)
Hispanic Americans	1,244	40%	440 (74%)
African Americans	359	99%	392 (27%)
Cerebellum brain tissue	394	58%	163 (19%)
Frontal cortex brain tissue	394	58%	229 (26%)

Table 5.7.: Significantly associated genes with aging in different tissues:
The expression levels of all genes that could be replicated in 7,909 samples were tested for association with age in different tissues and in samples with different ancestry. Due to low sample sizes in the samples from different tissues, the number of replicated genes was quite low (<30%).

where $x_{v(i)}$ is the expression level of the i-th probe in the cohort and $\hat{b}_R = (\mathbf{R} + \mathbf{I}\lambda/n)^{-1}\hat{b}$. Here, **R** is the correlation matrix between expression probes in a reference sample, **I** is the identity matrix, λ is a parameter that was optimized in the BSGS cohort, n is the sample size and $\hat{b}$ is defined as

$$\hat{b} = \frac{z}{\sqrt{n + z^2}}$$

with z being the test statistic from the meta-analysis. In a last step the predicted age was scaled:

$$Z_S = \mu_{age} + (Z - \mu_Z) * \frac{\sigma_{age}}{\sigma_Z}$$

with μ_{age} and σ_{age} being the mean and standard deviation of the chronological age and μ_Z and σ_Z the mean and the standard deviation of the predicted age Z.

The correlation between the chronological and the transcriptomic age was significant in all cohorts with p-values $< 2 * 10^{-29}$ (in KORA the p-value was $1.71 * 10^{-29}$) and the average difference was 7.8 years (in KORA 4.84 years). The small difference in KORA F4 shows no indication for a very remarkable prediction because the standard deviation of age is only 5.4. The results from KORA F4 are shown in Figure 5.5.

For all samples the difference between the chronological and the transcriptomic age was calculated. Samples having a positive Δage were predicted to be older than they are. This could be a hint that they age faster than other individuals. Therefore, it was tested whether the Δage was associated with biological phenotypes known to be correlated with the chronological age. The p-values for the correlation adjusted for chronological age of the meta-analysis and of KORA F4 are shown in Table 5.8. When assuming the Bonferroni threshold of 0.05/12 = $4.17 * 10^{-3}$ as significance threshold, only fasting glucose levels were positively correlated with Δage in KORA F4, meaning that individuals that have a higher Δage also have higher fasting glucose levels which could be an indication for type 2 diabetes. In the meta-analysis also systolic and diastolic blood pressure, total and HDL cholesterol levels, BMI, and waist-hip-ratio were positively correlated with Δage (see Table 5.8).

5.2.3.3. Analysis of gene expression, methylation, and chronological age

It is known that not only gene expression but also methylation changes along with higher age (Richardson, 2003). Therefore, methylation data were additionally analyzed. Methy-

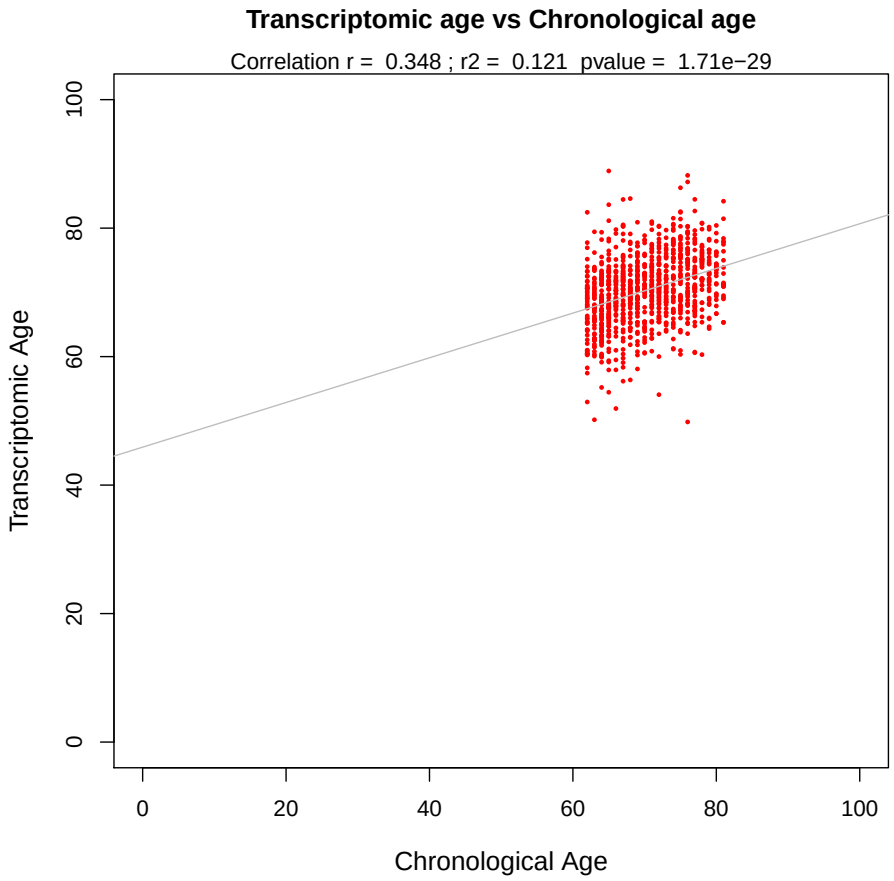

Figure 5.5.: Chronological versus predicted age in KORA F4

	KORA F4			meta-analysis			
Phenotype of Interest	Effect	P	n	Zscore	P	Direction	n
Sex (0=male, 1=female)	0.0007	8.33E-01	984	-27.610	5.76E-03	−+−+-+	8,836
Systolic BP (mmHg)	0.2752	3.03E-02	983	98.510	6.78E-23	++++++++	8,571
Diastolic BP (mmHg)	0.0187	7.64E-01	983	77.200	1.16E-14	++++++++	8,568
Total cholesterol (mmol/L)	-0.0031	6.38E-01	984	54.190	5.99E-08	++++++-++	8,688
HDL cholesterol (mmol/L)	-0.0018	4.28E-01	984	44.630	8.07E-06	++++++-+-	8,687
Fasting glucose (mmol/L)	0.4725	1.56E-03	984	69.330	4.11E-12	++++++??	7,330
Body Mass Index (kg/m2)	0.0804	5.18E-03	984	53.860	7.21E-08	++++++++	8,829
Waist Hip Ratio	0.0011	2.52E-02	984	33.800	7.25E-04	++??++++	4,837
Hand grip strength (kg)	na	na	na	-15.120	1.31E-01	++?-????	3,651
Renal function	na	na	na	0.8740	3.82E-01	+++-+?-?	7,317
Mini Mental State Exam	na	na	na	-13.130	1.89E-01	−??????	1,492
Current smoking	0.0025	1.07E-01	984	55.100	3.59E-08	+-?+++−	7,379

Table 5.8.: Association between transcriptomic age and age-related phenotypes:
The association between transcriptomic age and age-related phenotypes was calculated using different subgroups of the meta-analysis having the according phenotype. The p-values are displayed for KORA F4 and the meta-analysis.

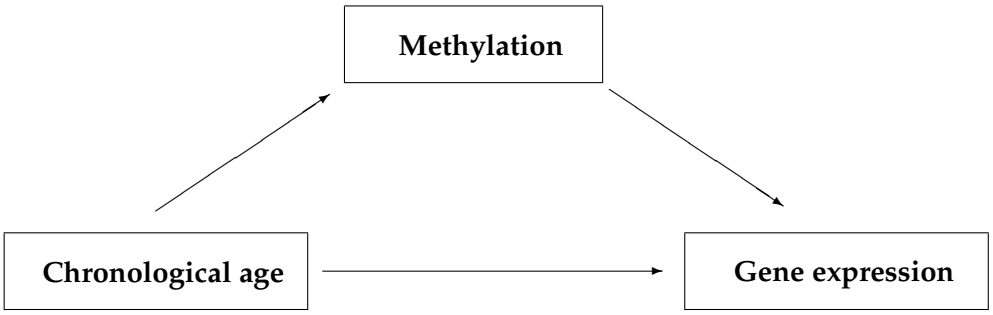

Figure 5.6.: Mediation of age-expression relationship by methylation

lation data from the Illumina 450K array were available for seven cohorts including KORA F4. Altogether there were 3,073 samples with KORA F4 being the largest cohort with 735 samples.
Only 135,230 CpG sites located in a 250kb window around all 1,497 genes significantly associated with the chronological age were analyzed. The window size was chosen based on a prior publication of Bonder et al. (2014) which showed that gene expression is most likely influenced by methylation of a CpG site within this distance.

Three models were calculated in each cohort for all CpG sites:

$$Model\ 1 : methylation \sim\ age + covariates$$

$$Model\ 2 : expression \sim\ methylation + covariates$$

$$Model\ 3 : expression \sim\ methylation + age + covariates$$

The covariates in all three models were sex, fasting and smoking status, cell counts (not available in KORA F4) and technical variables. The results of all models were meta-analyzed and the significance threshold was defined as $0.05/135{,}230 = 3.7 * 10^{-7}$.
For 31,331 CpG sites significant association between methylation and the chronological age (model 1) could be observed and 12,280 CpG sites were correlated with expression levels of a nearby gene (model 2).

To test whether the association between gene expression and chronological age was mediated by methylation (see Figure 5.6) the Sobel test was applied (Sobel, 1982) where the Sobel z-score was calculated by

$$Z_{Sobel} = \frac{Z_1 Z_2}{\sqrt{Z_1^2 + Z_2^2}}$$

with Z_1 being the Z-score from model 1 and Z_2 being the Z-score from model 3.

83% of the age-associated genes (1,248 out of 1,497) had at least one mediating CpG site whereas each gene was mediated by one to 154 CpG sites.

5.3. Summary and discussion

In the last chapter I presented two different association studies with gene expression data in whole blood. Initially, we investigated the impact of phenotypes related to blood pressure

on gene expression levels and afterward we analyzed the effect of aging on gene expression levels. For both association studies we started with the small KORA F3 data set, then analyzed the KORA F4 data set and finally contributed with the KORA F4 data set to a meta analysis of the CHARGE consortium. The blood pressure phenotypes were additionally analyzed in the MetaXpress consortium.

The impact of blood pressure on gene expression levels (or vice versa) seems to be quite small, at least in whole blood. This could be a reason why no significant association was found in 377 KORA F3 samples and just one significant association was detected in 989 KORA F4 samples. But due to the fact that the association between systolic/diastolic blood pressure and the expression level of *FOSB* was not identified in the MetaXpress and the CHARGE consortium, the relevance of *FOSB* has yet to be validated.
At least two of the eight genes that were identified in the MetaXpress study could be confirmed in the CHARGE data: *MYADM* and *TIPARP* with *TIPARP* being linked to blood pressure at least partly via SNP rs3184504 (Levy et al., 2009). This SNP was identified in a GWAS to be associated with blood pressure and was also a *trans*-eQTL for *TIPARP* and five other genes that were associated with blood pressure (*FOS*, *PP1R15A*, *TAGAP*, *S100A10*, *FGBP2*).
Functional characterization of the eight candidate genes that were identified in the monocyte samples, replicated in the whole blood samples, and validated in the clinical trial is still ongoing.

The association of gene expression and age was analyzed in three different studies. Firstly, in the smallest study KORA F3 with 381 samples, secondly in the larger KORA F4 study with 993 samples, and lastly in a consortium of 7,074 samples. As can be seen in Figure 5.7 the number of significant hits increased with the sample size.
When using the adjusted model (for sex and technical variables) in 993 KORA F4 samples 194 expression probes were significantly associated with age. Four of these were also identified in 381 KORA F3 samples (*CD248*, *GPR18*, *LRRN3* and *NELL2*). For *SGK223* the p-value was only marginally not significant ($1.67 * 10^{-6}$). The Bonferroni threshold was $1.02 * 10^{-6}$.
Out of 194 significant genes in KORA F4, 150 were analyzed in 7,074 samples of the CHARGE consortium and 130 of them were also significant in the discovery cohort and 105 additionally in the replication cohort.

The age prediction in KORA was very difficult and biased because of the small age range in the data set. In KORA F3, the age ranged from 51 to 84, in KORA F4 from 62 to 81. As can be seen in Figure 5.8 the age distribution in the other cohorts was wider and therefore, the results of the age prediction were more precise when using the correlation between predicted age and chronological age as a marker for a "good" prediction.

To improve the age prediction the methylation data could be taken into account as it was shown that 31,331 CpG sites were correlated with chronological age and 83% of the age-associated genes had at least one mediating CpG site. For KORA F4 and the Rotterdam Study two epigenetic age predictors ((Horvath, 2013) and (Hannum et al., 2013)) were available and were compared to the transcriptomic predictor. Both prediction methods were positively correlated with the transcriptomic predictor. But the epigenetic predictor was associated with different clinical phenotypes (transcriptomic predictor was associated with systolic blood pressure, waist-hip-ratio, and smoking, Horvath predictor only with waist-

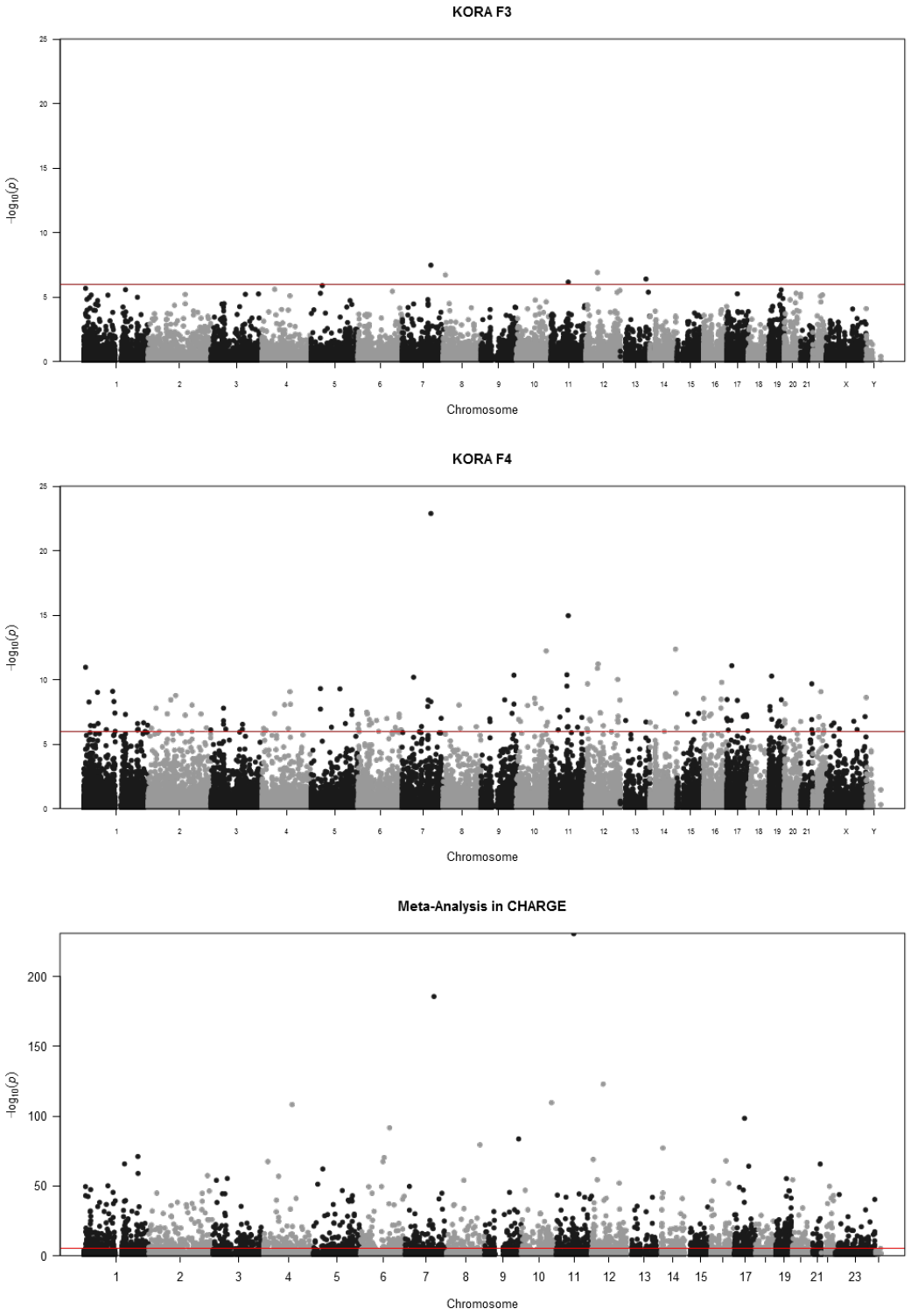

Figure 5.7.: Manhattan plots of results from association study between gene expression and aging in KORA F3, F4, and CHARGE:
The genomic location of each expression probe (meaning the TSS of the transcript to which the probe was mapped) is plotted against the -log10(p-value) of the result from the association between gene expression level and age. The red lines indicate the Bonferroni significance threshold.

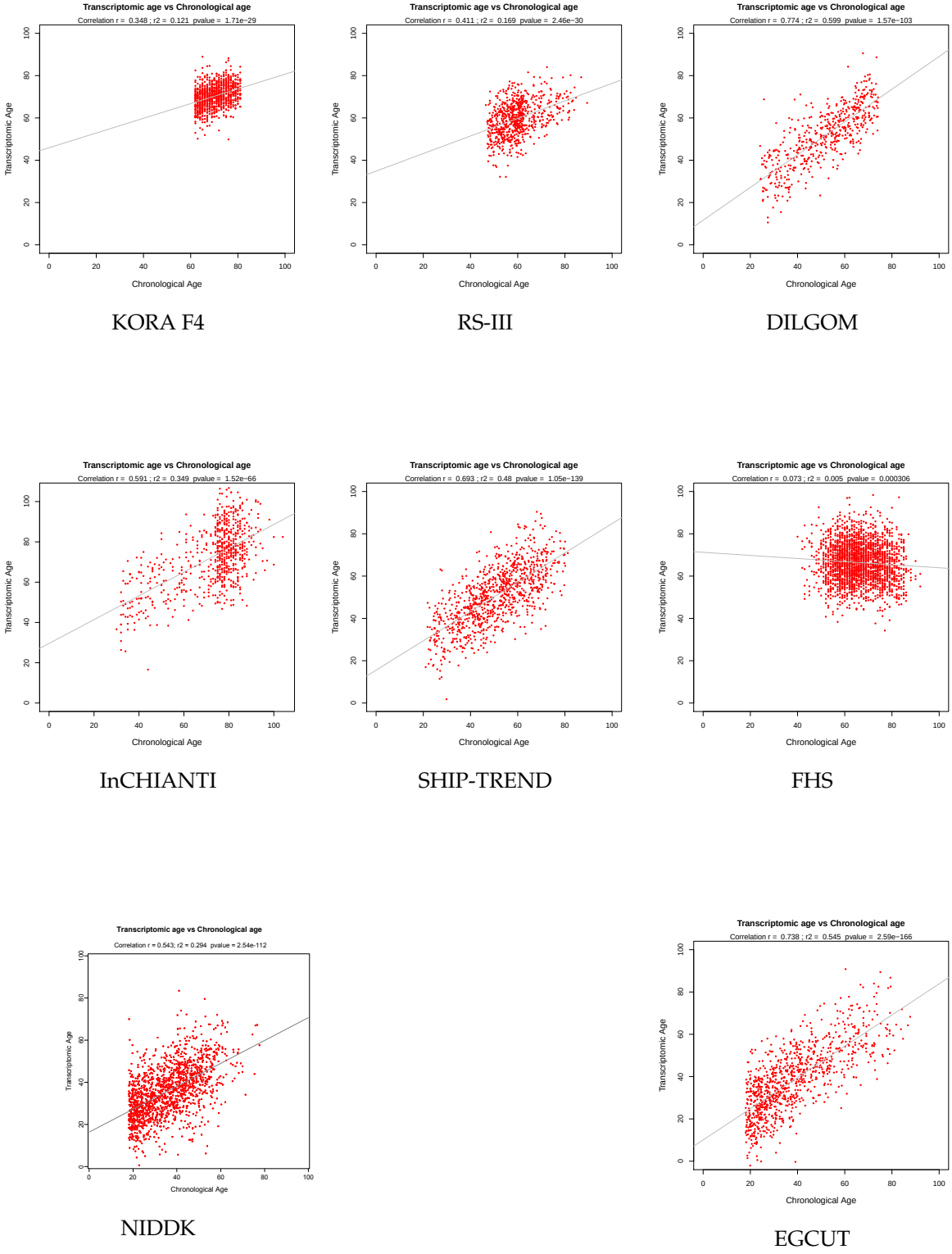

Figure 5.8.: Chronological age versus transcriptomic age in all participating cohorts of gene expression working group in CHARGE

hip-ratio, Hannun predictor with fasting glucose, waist-hip-ratio, and smoking)[2]. Therefore, a combination of both prediction methods (epigenetic and transcriptomic prediction) would make sense and should be tested in a larger sample size.
In conclusion, a large list of age-associated genes was identified in the meta-analysis with 7,074 samples. The gene expression levels from whole blood could be used as biomarker to predict the biological age. The transcriptomic age prediction in combination with the epigenetic age prediction could be used to identify subjects that have a higher biological age and therefore a higher risk to suffer any age-related disease.

[2]These models were adjusted for chronological age, sex, and BMI.

6. Power issues in eQTL studies

To identify the influence of genetic variation on gene expression levels eQTL studies are performed to identify SNPs which affect expression levels of nearby or distant genes. Ideally eQTLs help to detect molecular mechanism underlying a SNP-phenotype association identified in a GWAS (Gilad et al., 2008).
The first eQTL studies were conducted in less than 100 samples ((Stranger et al., 2005), (Cheung et al., 2005)). However, sample sizes increased fast to nowadays more than 5,000 samples (Westra et al., 2013) and will soon increase further to more than 20,000 samples (ongoing project within the CHARGE consortium).
We started the first eQTL study using 322 samples from the KORA F3 cohort (Mehta et al., 2012), continued with 890 samples from the KORA F4 cohort (Schramm et al., 2014) and participated as replication cohort with 740 KORA F4 samples in a large eQTL consortium (Westra et al., 2013). The following chapter describes the development and differences in eQTL studies with increased sample sizes.

6.1. Mapping of whole-blood *cis*- and *trans*-eQTLs in KORA F3

Aims of this first eQTL study were:

- The identification of eQTLs in whole blood of human samples.
- Analysis of the robustness and reproducibility across different studies.
- Exploration whether whole blood eQTLs allow the identification of functional variants observed in GWAS.

6.1.1. Identification of *cis*- and *trans*-eQTLs

The analysis of 41,409 expression probes and 335,152 SNPs yielded in 4,802,373 SNP-probe combinations with the SNP being located within $+/-500$kb from the transcription start and end site. The 500kb window was determined by comparing the number of significant associations between SNPs and expression probes for different window sizes (see Figure 6.1).
Linear models using log-transformed and normalized expression levels as dependent and SNP, age and gender as independent variables were calculated for each SNP-probe pair in 322 KORA F3 samples with expression and genotype data. An additive model was assumed where the homozygous major allele is coded with 0 and the homozygous minor allele is coded with 2 (see Figure 2.4).
When using the stringent Bonferroni correction (threshold = $1.03 * 10^{-8}$), 2,149 significant SNP-probe pairs corresponding to 363 different eQTLs were identified. These associations were distributed equally across the genome (see Figure 6.2). The SNPs with the lowest p-values are located close to the transcription start site (Figure 6.3) which supports the threshold of 500kb for *cis*-eQTLs.

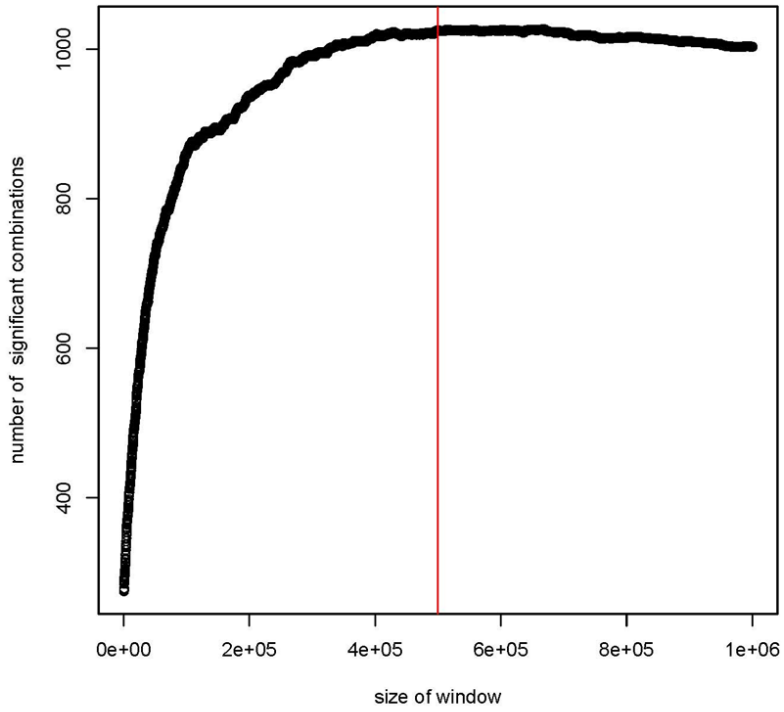

Figure 6.1.: Number of significant SNP-probe combinations for different window sizes in KORA F3:
The distance between gene and SNP (in bp) is plotted against the number of significant *cis*-associations to determine the optimal window size. Using a window size of more than 500kb leads to less significant hits due to the higher number of tests and the lower Bonferroni threshold of significance.

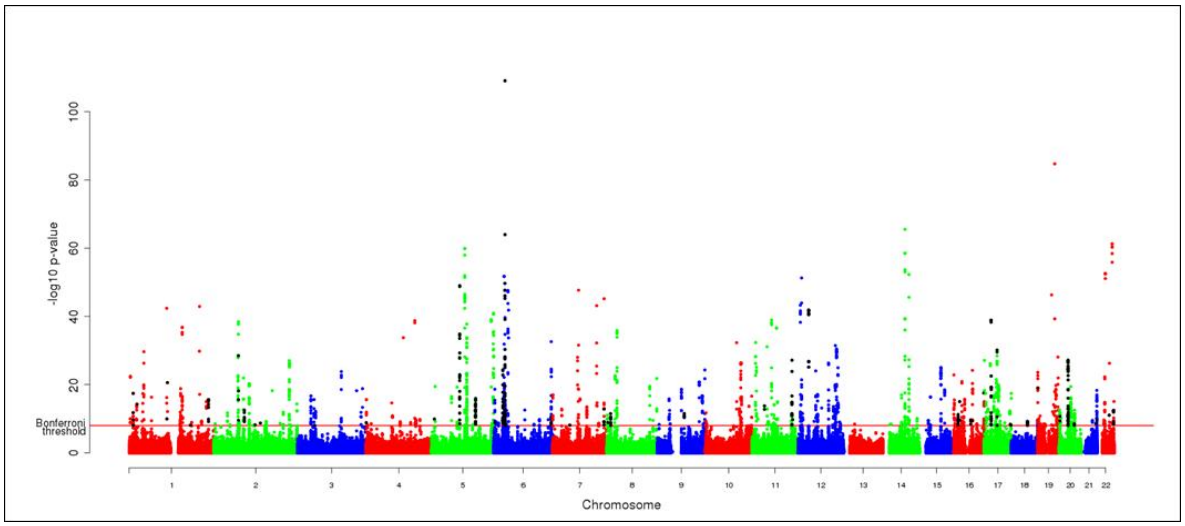

Figure 6.2.: Manhattan plot of significant *cis*-eQTLs in the KORA F3 discovery cohort: The genomic position of each SNP is plotted against the -log10(p-value). The red line indicates the Bonferroni threshold of significance.

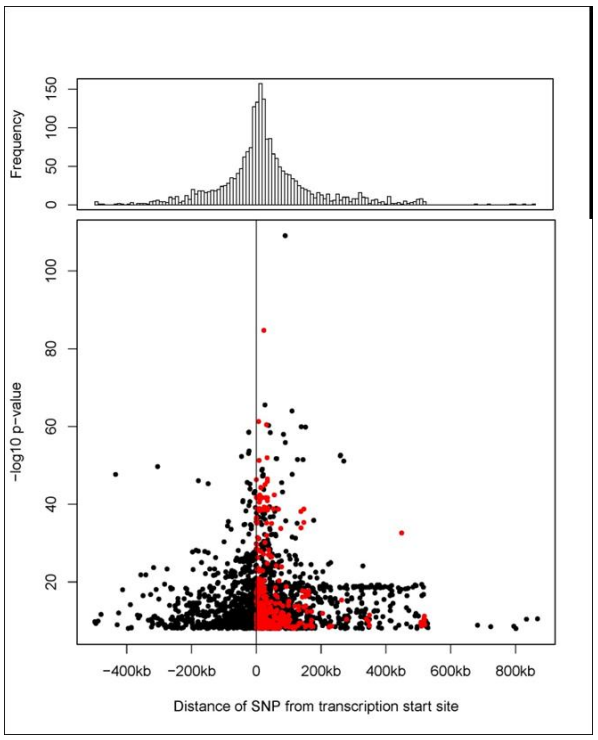

Figure 6.3.: Plot of distance between SNP and transcription start site in KORA F3: The difference between the genomic position of the SNP and the TSS is plotted against the -log10(p-values). The red dots show the associations where the SNP is located within the transcripts. The histogram additionally shows the distribution of the associations.

Furthermore, 50 of the 363 probes contained a SNP annotated within its probe sequence by the Illumina chip using the re-annotation pipeline from Barbosa-Morais et al. (2010). Using SNAP, we determined the correlation between the SNP within the probe and the identified eQTL SNP. For ten eQTLs R^2 was higher than 0.7 and we could not exclude that the SNP within the probe is responsible for the changes of the expression level. For all other eQTLs R^2 was either lower or could not be determined because of a low minor allele frequency ($< 1\%$ for 20 SNPs) of the SNP within the probe. We did not exclude probes with SNPs from the analysis which is consistent with other published studies (Murphy et al., 2010; Emilsson et al., 2008).

The *trans*-analysis was conducted using the software PLINK and like for the *cis*-analysis the linear model was also adjusted for age and gender. Due to large linkage disequilibrium blocks *trans*-hits were limited to SNPs being on a different chromosome than the probe or the distance between probe and SNP being greater than 4 Mb. The Bonferroni threshold of significance was $3.6 * 10^{-12}$. With these stringent criteria we identified 37 SNP-probe pairs corresponding to 8 eQTLs. Because of this small number of *trans*-eQTLs we could not observe any master regulators or eQTL hotspots.

6.1.2. Adjusting for possible confounders in the KORA F3 discovery cohort

Due to the fact that whole blood is used to measure gene expression levels it is sometimes recommended to adjust for different cell types. Therefore we also included the number of white and red blood cell counts to the linear model. 352 (97%) of the 363 eQTLs remained significant.

6.1.3. Replication of whole-blood eQTLs in two independent cohorts

For replication we used the independent cohorts KORA F4 and SHIP-TREND.
All significant associations were tested in 740 KORA F4 samples and 653 SHIP-TREND samples. The summary of the replication is shown in Figure 6.4.

It was decided to use two different thresholds of significance: The most stringent variant by using the threshold from the discovery cohort and the less stringent threshold of $p < 0.05$. This way we were able to replicate 98.6% of *cis*- and 40% of *trans*-eQTLs using $p < 0.05$ and 81.8% of *cis*- and 20% of *trans*-eQTLs when applying the discovery Bonferroni threshold of significance.

6.1.4. Comparison of results with published peripheral blood eQTLs

In addition to replications in KORA F4 and SHIP-TREND the eQTLs were compared to eQTLs detected in whole blood in 1,469 individuals and available online (Fehrmann et al., 2011). They also used the Illumina HumanHT-12 chip but a different genotyping platform (Illumina HumanHap 300). We only compared identical SNP-probe pairs. Due to the different platforms, the overlap contained about 50,000 SNPs. In spite these technical limitations 32% (117 out of 363) of the *cis*- and 14% (one out of seven) of the *trans*-eQTLs were identi-

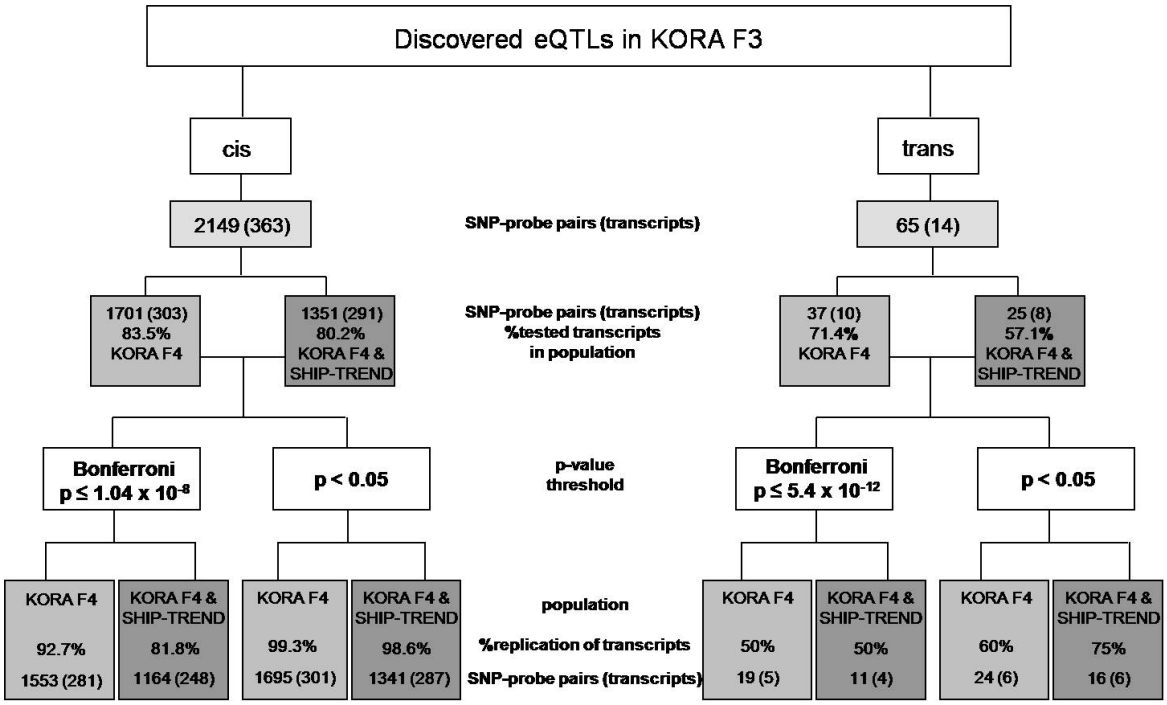

Figure 6.4.: Flowchart of the number of eQTLs from KORA F3 discovery cohort tested and replicated in KORA F4 and SHIP-TREND

cal in both studies. This additional replication indicates a high reproducibility of eQTLs in whole blood.

6.1.5. eQTL mapping of complex trait-associated variants

In a next step all GWAS hits published so far were downloaded from the GWAS catalog. In October 2011, 1,058 GWAS studies for 566 different traits were available and we could test 7,995 unique SNP-probe pairs in KORA F3. For this analysis imputed genotypes were used to maximize the number of SNPs.
Of the 3,699 unique SNPs, 639 SNPs were nominally associated with the expression level of the reported transcript. After Bonferroni correction (threshold of significance = $6.25 * 10^{-6}$) 79 SNPs remained significant. 43 of these SNPs were already described to be associated with the expression level of the relevant transcript. 28 SNPs are located on chromosome 6 in the HLA-region and because of the high LD structure of this region their biological significance remains unclear.
Eight of the SNPs were not reported to be an eQTL in any tissue previously. These eight novel findings are summarized in Table 6.1.

Especially in cases where the SNP is reported to be correlated with several genes the expression results could allow prioritization of this transcript compared to other candidates that are located close to the SNP. This is remarkable because blood is not related to all of these traits.

GWAS locus	GWAS trait	SNP	Gene	ProbeID	p-value	adjusted R^2	beta	First author
1	Beta thalassemia	rs2071348	*HBE1*	6520176	0.283414	-0.00503376	-0.18484	Nuinoon M.
	hemoglobin E disease	rs2071348	HBG2	6400079	5.38E-08	0.09514532	0.52015	
		rs2071348	*HBG2*	6620605	0.030536	0.00723574	-0.01432	
		rs2071348	*HBG2*	940181	0.365623	-0.00315586	0.006403	
		rs2071348	*HBG1*	4150187	1.94E-06	0.07676854	0.443254	
		rs2071348	*HBD*	6250037	0.195357	-0.00360737	-0.11191	
		rs2071348	*HBBP1*	7100747	0.881443	-0.00768138	-0.00186	
2	Crohn's disease	rs2058660	IL18RAP	6770424	0.154442	-0.00297075	-0.01054	Franke A.
		rs2058660	*IL18RAP*	5130475	2.68E-15	0.18068679	-0.44984	
		rs2058660	*IL12RL2*	NA	NA	NA	NA	
		rs2058660	*IL18R1*	1500328	0.698045	-0.00654905	-0.01298	
		rs2058660	*IL1RL1*	670411	0.909656	0.00344823	-0.00091	
		rs2058660	*IL1RL1*	3870753	0.349253	-0.00614557	-0.00746	
3	Graves' disease	rs9355610	*RNASET2*	5310131	7.05E-19	0.21555351	0.27063	Chu X.
		rs9355610	*FGFR1OP*	6580446	0.424956	-0.00652282	-0.00825	
4	Body mass index	rs7359397	*SH2B1*	6620092	0.133159	0.00166276	0.032315	Speliotes E.K.
		rs7359397	*APOB48R*	2070044	0.927515	0.0020211	0.002928	
		rs7359397	*SULT1A2*	1740113	0.095138	0.01485394	-0.01268	
		rs7359397	*SULT1A2*	1980554	0.316768	-0.00467373	0.010131	
		rs7359397	*AC138894.2*	NA	NA	NA	NA	
		rs7359397	*ATXN2L*	990524	0.21315	0.00104259	0.007658	
		rs7359397	*ATXN2L*	1300541	0.227558	-0.001122	0.007915	
		rs7359397	*ATXN2L*	5720435	0.668993	-0.00590007	0.005323	
		rs7359397	*TUFM*	6270735	4.89E-10	0.10666734	0.139009	
5	Systemic lupus	rs131654	*HIC2*	7050673	0.210732	-0.00152262	-0.01773	Han J.W.
	erythematosus	rs131654	*UBE2L3*	770523	0.179421	0.00400821	0.011055	
		rs131654	*UBE2L3*	1050360	1.24E-06	0.06346902	-0.15129	
6	Asthma	rs11078927	*GSDMB*	6620170	4.02E-18	0.20898095	-0.22407	Torgerson D.G.
7	Alzheimer's disease	rs6859	*PVRL2*	2570544	7.03E-07	0.07723456	-0.18595	Abraham R.
	Alzheimer's disease	rs6859	*TOMM40*	3400747	0.462314	0.00170593	0.013467	Naj A.C.
	(late onset)	rs6859	*APOE*	4150338	0.686794	-0.00123172	0.002674	
8	Alcohol dependence	rs8062326	*SYT17*	730725	5.72E-07	0.0710058	0.798277	Lydall G.J.
		rs8062326	*ITPRIPL2*	2710551	0.998	0.003417	0.001	

Table 6.1.: List of eight novel GWAS catalog eSNPs (SNP that is associated with gene expression level) significantly associated with expression levels of the reported transcript in KORA F3.

6.1.6. Summary of eQTLs in KORA F3

We performed a genome-wide eQTL mapping study in 322 KORA F3 samples and used 740 KORA F4 and 653 SHIP-TREND samples as replication cohorts. Due to the small sample size of the discovery cohort the power to detect eQTLs was relatively small. Nevertheless, we identified 363 *cis*- and 8 *trans*-associations. Using a p-value threshold of 0.05 we could replicate 98.6% of *cis*- and 40% of *trans*-eQTLs. Additionally we found eight novel GWAS catalog eSNPs that are significantly associated with expression levels of the reported transcript in the KORA F3 data set.

6.2. eQTL study in KORA F4

In the last chapter the eQTL analysis that was conducted in 322 KORA F3 subjects was described. A high number of *cis*-eQTLs could be replicated, but an identification of *trans*-eQTLs was difficult due to the problem of multiple testing and the relatively small sample size. Therefore, this analysis was repeated in 890 KORA F4 individuals with available gene expression levels and genotypes. The expectation was to increase the number of significant eQTLs because of the higher sample size and an improved quality of expression data.

6.2.1. Discovery of *cis*- and *trans*-eQTLs for KORA F4

Altogether, 4,210 eQTLs reached genome-wide significance when applying the most conservative Bonferroni correction (p-value threshold = $6.02 * 10^{-9}$ and $2.81 * 10^{-12}$ for *cis*- and

trans-eQTLs, respectively). 4,116 of these eQTLs were defined as *cis*-eQTLs (p-value range = $6.1 * 10^{-299} - 6.0 * 10^{-9}$) when using the same window size as for the KORA F3 data (500kb). The 4,116 *cis*-eQTLs corresponded to 3,449 RefSeq genes (HG19). The remaining 161 SNP-probe associations were defined as *trans*-eQTLs. After removing SNPs that are in high LD ($R^2 > 0.7$) using SNAP, 94 genomic loci out of the 161 SNP-probe associations with an impact on several distant genes remained (p-value range = $2.8 * 10^{-248} - 2.8 * 10^{-12}$).

6.2.1.1. Detailed description of *cis*-results

The definition of a *cis*-eQTL as being in a 500kb window around the transcription unit resulted in 8,308,092 possible SNP-probe combinations in the data set, consisting of 616,941 different SNPs and 28,691 expression probes. The linear model with expression probe (adjusted for 55 Eigen-genes) as dependent and SNP as independent variable was calculated for each combination. After applying Bonferroni correction 55,593 SNP-probe combinations had a p-value below $6.02 * 10^{-9}$ and were identified to be significantly associated. The distribution of p-values of all calculated SNP-probe combinations across the chromosomes is shown in Figure 6.5.

Lots of probes are associated with SNPs that are located very closely (Figure 6.5). As many

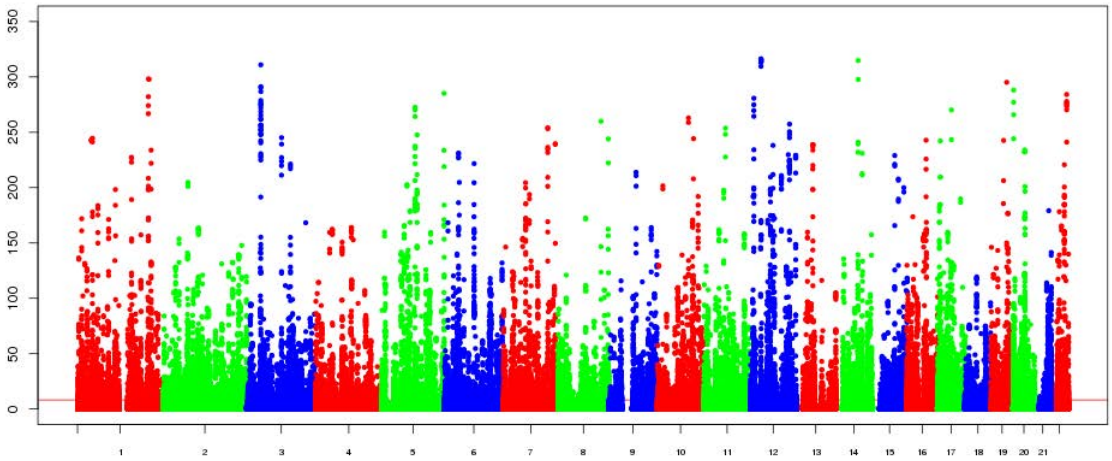

Figure 6.5.: Manhattan plot of *cis*-results in KORA F4:
The genomic position of the SNP is plotted against the -log10(p-value). The red line indicates the Bonferroni threshold of significance.

of these SNPs could be in high LD they might not describe different eQTLs. Hence, we restricted the results to the top-SNP per probe and kept only the SNPs with the lowest p-value.

eQTLs with low p-values are located in close proximity to the transcription start site (see Figure 6.6). This finding replicates the observation in KORA F3. Furthermore, it supports the restriction of *cis*-effects to a 500 kb window as a larger window size would not result in more meaningful hits.

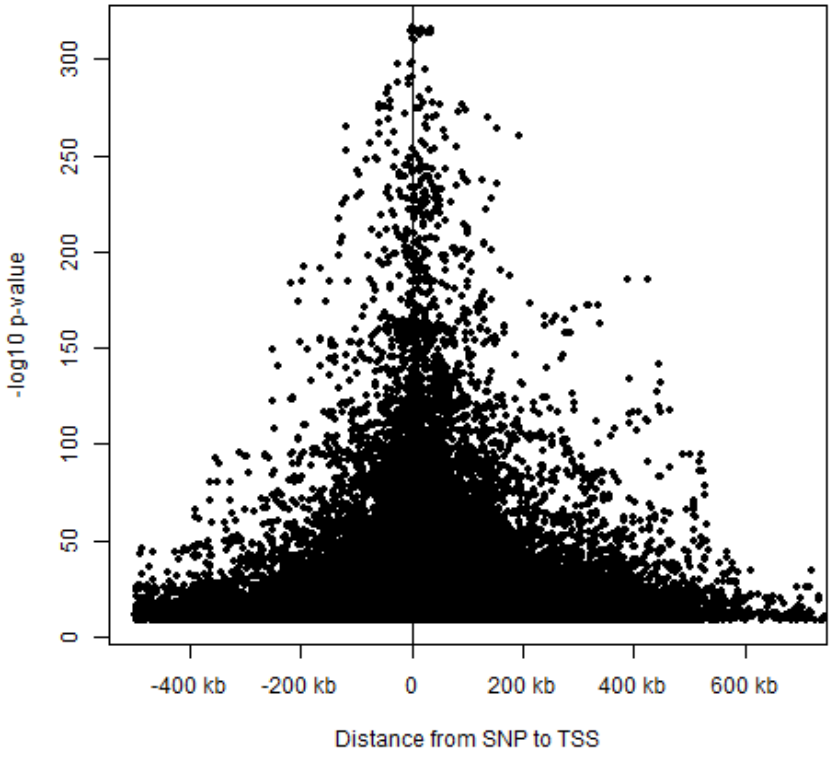

Figure 6.6.: Distance of the SNP to transcription start site for significant *cis*-eQTLs in KORA F4:
The distance between the TSS and the SNP is plotted against the -log10(p-value).

6.2.1.2. Detailed description of *trans*-results

All expression probes that were associated with a SNP that is located in a distance of more than 500 kb we tested whether the probe was also associated with a SNP in *cis*. Only if these two SNPs were not in the same LD block the association was called a *trans*-hit. For 77 expression probes the *trans*-SNP was highly correlated with the *cis*-SNP ($R^2 > 0.5$, tested with SNAP). Additionally, the LD was assessed for all SNPs residing on the same chromosome that were associated with the same transcription probe. This resulted in 94 remaining different *trans*-eQTLs (see circos plot in Figure 6.7 and a list of all *trans*-eQTLs in Appendix A.1).

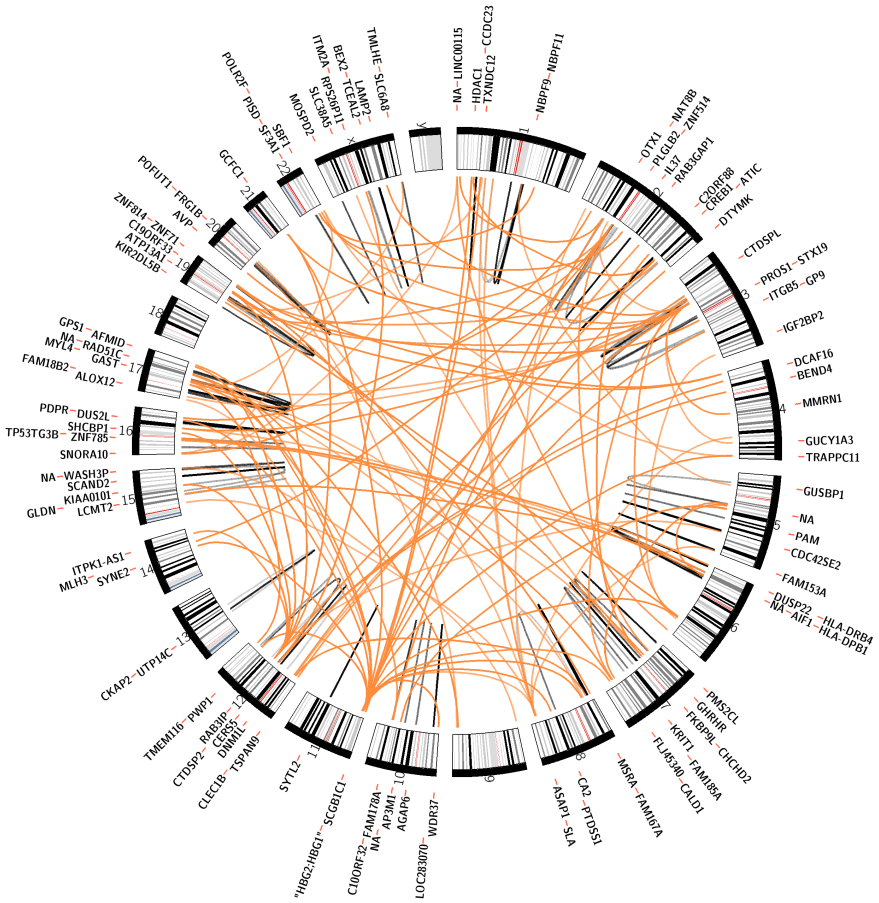

Figure 6.7.: Circos plot of *trans*-hits in KORA F4:
All significant *trans*-associations on all chromosomes in KORA F4 are shown. The orange lines indicate hits where the SNP and the expression probe are on different chromosomes while the black lines show associations on the same chromosome.

Altogether 74 loci are associated with just one probe, seven with two probes, seven with three probes, two with four probes, one with five probes, one with six probes, one with seven probes, one with 13 probes and one with 14 probes.

6.2.2. Replication of *cis*- and *trans*-eQTLs in two independent studies

The SHIP-TREND study from Greifswald and the EGCUT study from Estonia were used as replication cohorts using an R script prepared by Claudia Schurmann.
The SHIP-TREND samples were prepared almost identical as the KORA samples. RNA was isolated in Greifswald and then sent to the Helmholtz Center for the remaining steps. In contrast, the EGCUT study is based on samples of non-fasting individuals and blood was taken during the whole day (see Figure 6.8). Additionally, they used a different blood storage system (Tempus tubes instead of PAX tubes). It is already known that usage of PAX tubes and Tempus tubes results in heterogeneous measurements and it is recommended not to use data from these two systems in a joint analysis (Menke et al., 2012).
The preparation of the 890 KORA F4, 976 SHIP-TREND, and 842 EGCUT samples is described in Table 6.2 to indicate differences and similarities between the three cohorts.

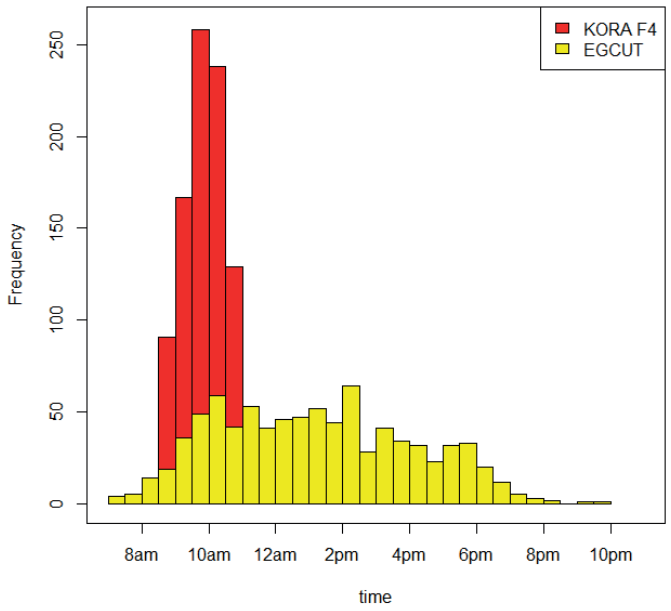

Figure 6.8.: Comparison of times of blood collection in KORA F4 and EGCUT:
While the KORA F4 blood samples were all taken in the morning before eleven o´clock in EGCUT the blood was taken the whole day.

All significant probe-SNP combinations were taken forward for replication in both replication studies and for that purpose either the identical or a proxy SNP (if the identical SNP was not available the proxy SNP was determined using SNAP) was used. Based on previous data assessment studies the replication studies decided to adjust their linear models for 50 (*cis*) and 25 (*trans*) Eigen-genes. We indicated an eQTL as significantly replicated if the

p-value was below $1.21 * 10^{-5}$ for *cis* and $3.1 * 10^{-4}$ for *trans*, respectively (0.05 divided by the number of conducted tests).

Study	N	Age	Gender	Fasting status	RNA collection	RNA isolation	Expression Chip	Genotyping	Imputation
KORA F4	890	70.57 ±5.42	448 males, 442 females	fasting (8 non-fasting samples)	PAX tubes	PAXgene Blood miRNA Kit	Illumina HumanHT-12 v3	Affymetrix 6.0	
SHIP-TREND	976	50.12 ±13.74	428 males, 548 females	all fasting	PAX tubes	PAXgene Blood miRNA Kit	Illumina HumanHT-12 v3	Illumina HumanOmni2.5-Quad	IMPUTE v2.1.2.3
EGCUT	842	37.16 ±15.60	415 males, 427 females	non-fasting	Tempus tubes	Tempus Spin RNA Isolation Kit	Illumina HumanHT-12 v3	Illumina Human370CNV	IMPUTE v2.2.3

Table 6.2.: Study description of KORA F4, SHIP-TREND, and EGCUT

Of all 4,116 significant eQTLs we were able to replicate 3,847 (91%) in at least one of the studies. As expected from the previous results the number of replicated *cis*-eQTLs was higher than the number of replicated *trans*-eQTLs (92% versus 84%). A comparison of p-values is shown in Figure 6.9.

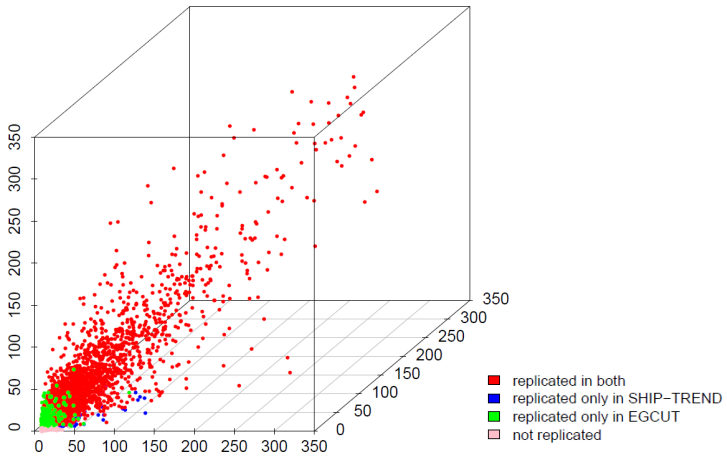

Figure 6.9.: Comparison of p-values for eQTL replication between KORA F4, SHIP-TREND, and EGCUT:
The p-values of all significant associations in KORA F4 are plotted against the p-values from SHIP-TREND and EGCUT. Red dots represent all hits that are replicated in both replication cohorts, blue dots are hits that are only replicated in SHIP-TREND, and green dots only replicated in EGCUT, respectively. The pink dots show all hits that could not be replicated.

When we compared the single replication rates separately in SHIP-TREND and EGCUT we observed a slightly higher replication rate in EGCUT (82% - 85% *cis*- and 72% *trans*-eQTLs) than in SHIP-TREND (78% - 81% *cis*- and 75% *trans*-eQTLs). This was in contrast to the a priori hypothesis of a higher replication rate in SHIP-TREND due to more similar study characteristics and data handling. This indicates a high robustness of whole blood eQTLs in contrast to whole blood gene expression profiles.

6.2.3. Correlation of mitochondrial SNPs with expression probes

Out of all identified SNPs, 26 SNPs were located within the mitochondrial DNA. The mitochondrial DNA (mtDNA) is double-stranded, circular and located in the mitochondria. It consists of 16,569 basepairs (`http://www.ncbi.nlm.nih.gov/nuccore/251831106`) and 37 genes. For one of the SNPs all samples had the same genotype and therefore this SNP was excluded from the analysis. Because it was not clear a priori if only *cis*-effects could be observed, 25 Eigen-genes were used as covariates in the linear model.
Thirteen SNPs were associated with at least one probe when applying a Bonferroni correction (p-value $< 3.94 * 10^{-8}$). This corresponded to six different probes two of which did not match to the genome and to two different chromosomes respectively. *LOC653566* and *RNF113A* were associated with seven SNPs. The expression levels of these two genes were correlated (Pearson's correlation coefficient r^2 = 0.37). Four SNPs (rs3915952, rs28359172, rs28359178 and rs3088309) were correlated with expression levels of a probe that could not be matched to the mRNA. The results are shown in Table A.2.
The expression levels of 834 probes were associated with at least one SNP with a p-value below $1 * 10^{-3}$. These probes were used for a pathway analysis. Only 538 of the 834 probes were annotated in the Ingenuity database and were included in the pathway analysis. When applying a Benjamini-Hochberg correction for the Fisher´s Exact p-value no pathway was significant. The top pathway "Role of NFAT (nuclear factor of activated T-cells) in Cardiac Hypertrophy" had a corrected p-value of 0.656. Our finding is in line with a few publications about a relation between mitochondria and cardiac hypertrophy (Rosca et al., 2013).

6.2.4. Comparison of results with published *cis*-eQTLs from different tissues

It is known that gene expression levels vary across different cell types. However, as whole blood cells are easily to obtain and analyze, it is interesting to investigate if it is a suitable surrogate tissue and also reflect transcriptional relationships of other tissues. Therefore, we compared our results to already published eQTLs in different cell lines and tissues. Unfortunately, databases provided by other studies and publicly available were mostly restricted to *cis*-eQTLs ((Fehrmann et al., 2011), (Sasayama et al., 2013), (Fairfax et al., 2012), (Zeller et al., 2010), (Schadt et al., 2008), (Stranger et al., 2007a), (Innocenti et al., 2011), (Dixon et al., 2007), (Hao et al., 2012)). We therefore limited our comparison to the *cis*-results.

Zeller et al. (2010) already did the comparison of their eQTLs in monocytes with published data. They used datasets of eQTLs in liver (Schadt et al., 2008), lymphoblastoid cell lines ((Dixon et al., 2007) and (Stranger et al., 2007b)), and lymphocytes (Göring et al., 2007). We additionally compared our eQTLs to eQTLs in whole blood ((Fehrmann et al., 2011) and (Sasayama et al., 2013)), in monocytes ((Zeller et al., 2010), (Fairfax et al., 2012)), in b-cells (Fairfax et al., 2012), in lung tissue (Hao et al., 2012) and in liver tissue (Innocenti et al., 2011).
All studies are shortly presented in Section 2.6 and results of the *cis*-eQTL comparison are summarized in Table 6.3.

The comparison of eQTLs across different tissues was not easily to address. Usage of different expression platforms and different versions of expression arrays made comparison challenging. Therefore, only genes with identical names were included.
We observed an overlap of 65%-68% with two more *cis*-eQTL studies conducted in whole

First author	Tissue	Year	N	*cis* eQTLs	available	identical
Schramm	whole blood	2014	890	3,449		
Fairfax	monocytes b-cells	2011	283	7,468 6,831	5,694 5,073	1,764 1,354
Zeller	monocytes	2011	1,490	2,477	2,323	1,620
Schadt	liver	2008	427	1,525	1,401	642
Stranger	lymphoblastoid cell lines	2007	90	412	378	229
Fehrmann	whole blood	2011	1,469	5,928	4,194	2,250
Innocenti	liver	2011	206	1,173	1,057	487
Dixon	lymphoblastoid cell lines	2007	400	727	659	373
Göring	lymphocytes	2007	1,240	6,684	6,163	1,768
Sasayama	whole blood	2013	76	308	252	171
Hao	lung	2012	1,111	9,138	7,556	2,252

Table 6.3.: Comparison of *cis*-eQTLs to *cis*-eQTLs in different tissues:
We compared the replicated *cis*-hits to online availabe *cis*-eQTLs in different tissues. Available means here that we analyzed the same gene and identical means it was significant in our data.

blood and 51%-70% with *cis*-eQTL studies conducted in primary monocytes and lymphocytes from whole blood or blood-derived lymphoblastoid cell lines (LCLs) (see Table 6.3). Additionaly, we observed major cross-tissue similarity when comparing our results to those of eQTL studies conducted in b-cells, lung, and liver tissue (40-70%, Table 6.3).

6.2.5. Functional properties of significant whole blood *cis*- and *trans*-eQTLs

The Ingenuity Pathway Analysis (IPA) software was used to relate whole blood eQTLs in *cis* and *trans* to known pathways. Only the replicated eQTLs were included and most of them were annotated in the Ingenuity data base (3,720 out of 3,768 *cis*-eQTLs and 139 out of 144 *trans*-eQTLs, respectively). When applying the Benjamini-Hochberg correction two significant canonical pathways with p-values below $2.28 * 10^{-2}$ and ten with p-values below $3.45 * 10^{-2}$ could be identified for the *cis*- and *trans*-eQTLs, respectively (see Table 6.4).

6.2.6. Master regulatory loci

In the significant *trans*-list we identified 21 eQTL-SNPs that were significantly associated with expression levels of two or more genes. Four of these SNPs had an impact on the expression level of five or more genes and were therefore called "master regulatory loci" (Table 6.5). These loci are:

- Hotspot 1 on chromosome 2: rs12151621 (2:85934499) upstream of *ATOH8*.
  Genes: *PNKD*, *CALHM1*, *DYNLRB2*, *ZNF93*, *GHRHR*, *MLH3*
  Displayed in Figure 6.10.

Transcripts associated with *cis*-eQTL		
Significant Canonical Pathways	**-log(B-H p-value)**	**Molecules**
NAD Salvage Pathway II	1.89	NT5C3B,NT5C3A,ACP2,NMRK1,NT5E,NT5M, ACP1,NT5C2,ACP5,NMNAT3,ACPL2,ACPP
Glutathione Redox Reactions I	1.64	GSR,GSTT1,GPX3,MGST1,MGST2, GPX4, GPX7,PRDX6,MGST3,GSTK1
Transcripts associated with *trans*-eQTL		
Significant Canonical Pathways	**-log(B-H p-value)**	**Molecules**
Allograft Rejection Signaling	4.38	HLA-DRB4,HLA-A,HLA-DRB3,HLA-DQA1,HLA-DQB1, HLA-DPB1
Cytotoxic T Lymphocyte-mediated Apoptosis of Target Cells	4.38	HLA-DRB4,HLA-A,HLA-DRB3,HLA-DQA1,HLA-DQB1, HLA-DPB1
OX40 Signaling Pathway	4.30	HLA-DRB4,HLA-A,HLA-DRB3,HLA-DQA1,HLA-DQB1, HLA-DPB1
Antigen Presentation Pathway	3.93	HLA-DRB4,HLA-A,HLA-DRB3,HLA-DQA1,HLA-DPB1
Cdc42 Signaling	3.33	HLA-DRB4,HLA-A,HLA-DRB3,HLA-DQA1,MYL4, HLA-DQB1,HLA-DPB1
Dendritic Cell Maturation	2.64	HLA-DRB4,HLA-A,HLA-DRB3,CREB1,HLA-DQA1,HLA-DQB1,IL37
Graft-versus-Host Disease Signaling	2.46	HLA-A,HLA-DQA1,HLA-DQB1,IL37
Crosstalk between Dendritic Cells and Natural Killer Cells	2.36	KIR2DL5B,HLA-DRB4,KIR2DL5A,HLA-A,HLA-DRB3
Communication between Innate and Adaptive Immune Cells	1.54	HLA-DRB4,HLA-A,HLA-DRB3,IL37
Autoimmune Thyroid Disease Signaling	1.46	HLA-A,HLA-DQA1,HLA-DQB1

Table 6.4.: Results of pathway analysis for *cis*- and *trans*-eQTLs in KORA F4:
3,720 *cis*-eQTL genes were annotated in IPA and two significant pathways with p-value $< 2.28 * 10^{-2}$ were identified. For *trans*-results, 139 genes were annotated and ten significant pathways with p-value $< 3.45 * 10^{-2}$ were identified.

- Hotspot 2 on chromosome 3: rs1344142 (3:56857433) in *ARHGEF3* involved in osteoporosis and rs12485738 (3:56865776, p14.3) in *ARHGEF3*.
  Genes: *MMRN1*, *ITGB5*, *PROS1*, *NAT8B*, *ALOX12*, *TSPAN9*, *CALD1*, *GP9*, *CLEC1B*, *PARVB*, *GUCY1A3*, *CTDSPL*, *ITGB3*
  Displayed in Figure 6.11.
  This locus was already reported by Meisinger et al. (2009) and Fairfax et al. (2012).

- Hotspot 3 on chromosome 11: rs10742583 (11:5248641) and rs12786766 (11:5225505) in *HBB*.
  Genes: *ADCK2*, *PTDSS1*, *ASAP1*, *RAD51C*, *DTYMK*, *HDAC1*, *PWP1*, *TRAPPC11*, *SYNE2*, *CNBP*, *WDR59*, *GPS1*, *WDR37*
  Displayed in Figure 6.12.

- Hotspot 4 on chromosome 12: rs10784774, rs2231700, rs11177577, rs2603089, rs11177644 upstream of *LYZ*.
  Genes: *EID2B*, *AFMID*, *CDKN2AIPNL*, *ITPK1-AS1*, *SHCBP1*, *CREB1*, *KIAA0101*
  Displayed in Figure 6.13.
  This locus was also already reported by Fairfax et al. (2012).

All master regulatory loci could be replicated in both replication studies. Surprisingly, all these eQTLs were not found to be significant *cis*-eQTLs, although one of the loci (on chromosome 12) was already identified to be a *cis*- and *trans*-eQTL in monocytes (Fairfax et al., 2012). The *cis*-association was observed in a much smaller sample size of 283 individuals indicating a monocyte specific effect or at least a very small effect in whole blood.

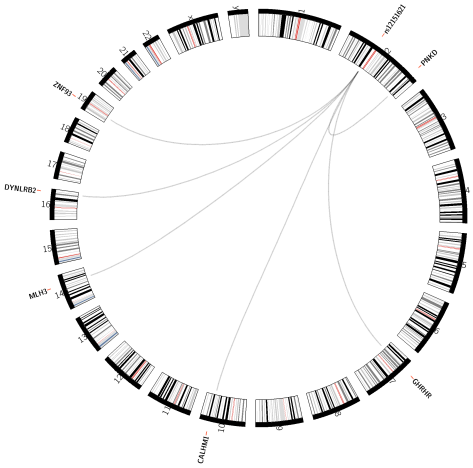

Figure 6.10.: *Trans* hotspot 1

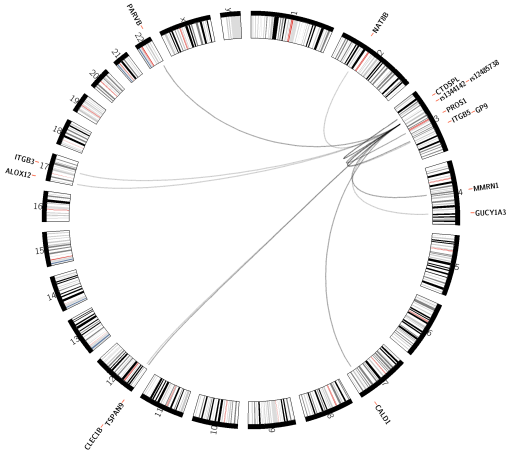

Figure 6.11.: *Trans* hotspot 2

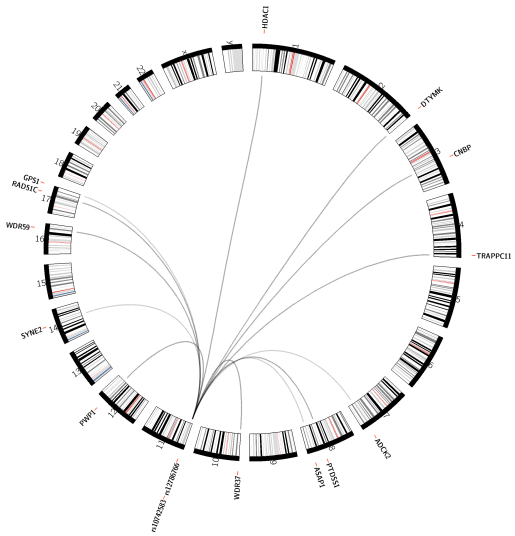

Figure 6.12.: *Trans* hotspot 3

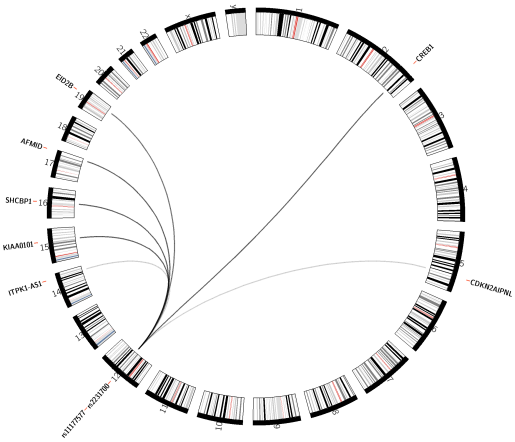

Figure 6.13.: *Trans* hotspot 4

SNP*	Chr SNP	Gene of SNP	Probe_Id	Gene of probe	Chr gene	n	BETA	SE	Pvalue
rs12151621	2	N/A	ILMN_2282282	*MLH3*	14	888	0.1904	0.01396	1.312e-38
	2	N/A	ILMN_1679130	*CALHM1*	10	888	0.1424	0.01326	2.244e-25
	2	N/A	ILMN_1697317	*DYNLRB2*	16	888	0.1389	0.01285	1.124e-25
	2	N/A	ILMN_1652161	*PNKD*	2	888	0.1407	0.01227	1.892e-28
	2	N/A	ILMN_1724158	*ZNF93*	19	888	0.09093	0.0125	7.624e-13
	2	N/A	ILMN_1740186	*GHRHR*	7	888	0.07777	0.00913	6.929e-17
rs12485738	3	*ARHGEF3*	ILMN_1787919	*PARVB*	22	890	-0.1362	0.01922	2.769e-12
	3	*ARHGEF3*	ILMN_1729453	*TSPAN9*	12	890	-0.1682	0.02039	5.695e-16
	3	*ARHGEF3*	ILMN_1668374	*ITGB5*	3	890	-0.1388	0.01918	9.885e-13
	3	*ARHGEF3*	ILMN_2392189	*CTDSPL*	3	890	-0.1302	0.0172	9.16e-14
	3	*ARHGEF3*	ILMN_1743290	*GP9*	3	890	-0.1793	0.02214	1.823e-15
	3	*ARHGEF3*	ILMN_1730487	*CALD1*	7	890	-0.08874	0.01179	1.263e-13
	3	*ARHGEF3*	ILMN_1713731	*ALOX12*	17	890	-0.1091	0.01538	2.607e-12
	3	*ARHGEF3*	ILMN_1691264	*NAT8B*	2	890	-0.1413	0.01871	1.055e-13
	3	*ARHGEF3*	ILMN_1671928	*PROS1*	3	890	-0.1616	0.01783	7.819e-19
	3	*ARHGEF3*	ILMN_1808590	*GUCY1A3*	4	890	-0.1073	0.01453	3.579e-13
	3	*ARHGEF3*	ILMN_1660114	*MMRN1*	4	890	-0.08856	0.01129	1.231e-14
	3	*ARHGEF3*	ILMN_1745103	*CLEC1B*	12	890	-0.1498	0.02055	6.809e-13
rs10784774	12	N/A	ILMN_2334242	*CREB1*	2	889	0.1903	0.0123	5.329e-48
	12	N/A	ILMN_2182482	*SHCBP1*	16	889	0.189	0.009911	3.288e-68
	12	N/A	ILMN_2095653	*AFMID*	17	889	-0.1173	0.009898	3.367e-30
	12	N/A	ILMN_2412521	*KIAA0101*	15	889	0.1283	0.009883	2.183e-35
	12	N/A	ILMN_2134381	*ITPK1-AS1*	14	889	0.09329	0.01308	2.021e-12
	12	N/A	ILMN_2051900	*EID2B*	19	889	0.1495	0.01267	5.998e-30
	12	N/A	ILMN_2130078	*CDKN2AIPNL*	5	889	0.07231	0.009745	2.737e-13
rs10742583	11	N/A	ILMN_1743049	*PWP1*	12	890	-0.1424	0.01076	1.198e-36
	11	N/A	ILMN_1688753	*PTDSS1*	8	890	0.2389	0.01192	5.786e-74
	11	N/A	ILMN_1769319	*CNBP*	3	890	0.1136	0.01028	1.015e-26
	11	N/A	ILMN_1752086	*TRAPPC11*	4	890	-0.07452	0.009111	9.901e-16
	11	N/A	ILMN_1727458	*HDAC1*	1	890	0.1139	0.01002	4.303e-28
	11	N/A	ILMN_1795428	*WDR59*	16	890	-0.0939	0.009298	8.944e-23
	11	N/A	ILMN_1795876	*GPS1*	17	890	0.06541	0.007896	4.36e-16
	11	N/A	ILMN_1690963	*ASAP1*	8	890	0.1099	0.01438	5.594e-14
	11	N/A	ILMN_1663132	*ADCK2*	7	890	0.0939	0.01212	2.552e-14
	11	N/A	ILMN_1716445	*DTYMK*	2	890	0.09812	0.01061	1.688e-19
	11	N/A	ILMN_1796464	*WDR37*	10	890	0.09877	0.009316	8.254e-25
	11	N/A	ILMN_1754579	*SYNE2*	14	890	0.06739	0.008484	5.915e-15
	11	N/A	ILMN_1695386	*RAD51C*	17	890	0.09094	0.01153	8.906e-15

Table 6.5.: Master regulatory loci:
We defined master regulatory loci as eQTLs with simultaneous impact on the expression of at least five genes. Only the SNP which displayed strongest associations in the region is displayed.

6.2.7. Comparison of *cis*- and *trans*-results with the published GWAS catalog

On July, 18th 2012 the GWAS catalog consisted of 1,310 publications including 6,603 SNPs (according to the information on the homepage). The whole catalog was downloaded and consisted of 10,421 SNP-gene combinations (10,099 unique). 214 were deleted because no SNP and no gene were reported (NR). The SNP-gene combinations that were reported by more than one study were merged and altogether 8,566 unique combinations were systematically compared to the *cis*- and *trans*-results. There were 4,471 unique genes of which 3,508 could be found in our annotation file.
In the first step the overlap between the reported genes and the eQTL genes was determined and for the corresponding SNPs all proxy SNPs with $R^2 > 0.7$ were selected using SNAP. 181 gene-SNP combinations of the GWAS catalog could also be found in the list of significant *cis*-hits, either with the exactly same SNP as reported or with a SNP that was in high LD ($R^2 > 0.7$) with the reported SNP. Two SNPs as well as four genes could additionally be found in the list of significant *trans*-hits.
In addition, when we looked up GWAS hits in our hits we identified 565 reported genes in the *cis*-list, 16 of which were also in the *trans*-list. Three SNPs of this list are also in the *trans*-list.
Of the 6,966 gene-SNP combinations that were not in the *cis*-list, ten genes and 13 SNPs were in the *trans*-list.

We focused on the 746 genes that were both eQTL-genes and reported to be associated with a clinical or any other phenotype. These findings could be helpful to improve the knowledge about the reported associations and to identify mechanisms underlying the SNP-phenotype association.

As this project was a collaboration with the Diabetes Center in Düsseldorf we were particulary interested in genes that were reported to be associated with diabetes-relevant phenotypes. One SNP (rs592423, located on chromosme 6) of our list was reported to be associated with adiponectin (low levels of adiponectin could increase the risk for diabetes, while high levels lead to a protection against diabetes) (Dastani et al., 2012). We identified a *trans*-association with the expression level of a type 2 diabetes gene called *IGF2BP2* (*insulin-like growth factor 2 mRNA-binding protein 2*) on chromosome 3 (p-value $= 1.2 * 10^{-13}$). As the adiponectin levels two hours after an oral glucose tolerance test were available for 738 KORA F4 samples, we tested if there was a correlation between this level and the gene expression level. Because the p-value for this association was below 0.05 (p=0.025) we hypothesized that there might be a possible effect of rs592423 on adiponectin level via the expression of *IGF2BP2* (see Figure 6.14).

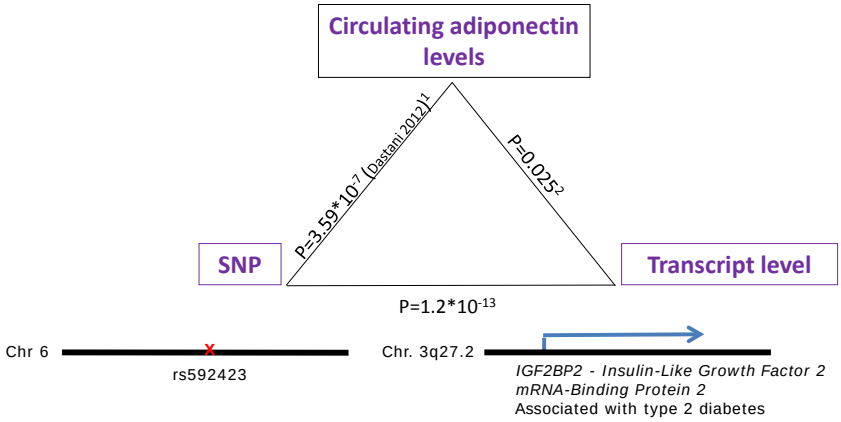

Figure 6.14.: Triangular relationship between eQTL-SNP rs592423, gene expression level of *IGF2BP2* in *trans*, and adiponectin

In the GWAS catalog several SNPs are reported to be associated with different clinical phenotypes (with so-called pleiotropic associations). In our data we identified two SNPs that on the one hand had an impact on several gene expression levels and on the other hand showed such pleiotropic effects:

- SNP rs10784774, which was published to affect height (Gudbjartsson et al., 2008), pulmonary function decline (Imboden et al., 2012), and response to diuretic therapy (Turner et al., 2010) in GWAS.
- SNP rs12485738 with reported associations for creatinine levels (Chambers et al., 2010), blood pressure (Wain et al., 2011), chronic kidney disease (Köttgen et al., 2010), and mean platelet volume (Meisinger et al., 2009; Soranzo et al., 2009).

For the SNP rs12485738 we could provide evidence for a triangular relationship between the SNP, mean platelet volume (measured in 889 KORA F4 samples) and gene expression activity of nine out of the twelve annotated genes.

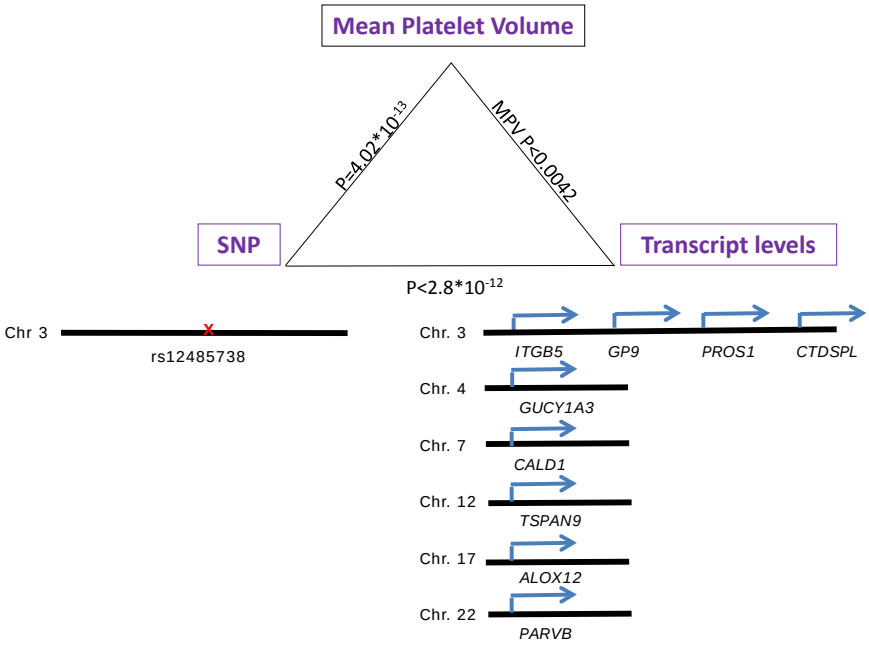

Figure 6.15.: Triangular relationship between eQTL-SNP rs12485738, gene expression levels in *trans* and mean platelet volume:
For nine out of twelve significant *trans*-eQTL genes for rs12485738 we could identify significant triangular relationships between SNP, mean platelet volume and gene expression levels.

6.2.8. Comparison of *cis*-eQTLs with metQTLs

In the KORA F4 samples also metabolite concentrations are measured in whole blood. And due to the fact that it has already been observed that metabolic phenotypes could combine genetic and environmental factors to explain complex disorders (Suhre and Gieger, 2012) the *cis*-eQTLs were compared to results obtained from a GWAS of metabolic traits. This GWAS was conducted in 1,809 samples from KORA F4 and the results were replicated in 422 samples of the TwinsUK study. Thereby, 163 metabolic traits that were measured in whole blood

were analyzed. Illig et al. (2010) identified 18 SNPs that were associated with several metabolites or the ratio of two metabolites, so-called metabolomic quantitative trait loci (metQTLs). These results are available in the Supplement Table 2 of the original publication.
Of these 18 SNPs six SNPs (rs174547, rs211718, rs8396, rs541503, rs272889, and rs964184) were also *cis*-eQTLs in our data set (see Table 6.6). To identify triangular relationships between transcriptomics, genomics, and metabolomics the missing association between gene expression and the metabolite concentration or the ratio of two metabolites was calculated. Additionally, p-values for the SNP-metabolite association and the p-values for the SNP - gene expression association were determined in the smaller subset. For these analyses metabolomic and transcriptomic data of 717 KORA F4 samples were available.

In this effort we identified three significant triangular relationships:

1. SNP rs541503 which is located upstream of *PHGDH* (*Phosphoglycerate Dehydrogenase*) was associated with the expression level of *PHGDH* (p-value $= 4.14 * 10^{-11}$) and with the metabolite Serine (Ser, p-value $= 2.59 * 10^{-3}$). The expression level of *PHGDH* was also associated with the metabolite concentration (p-value $= 9.38 * 10^{-4}$). The triangular relationship is illustrated in Figure 6.16.

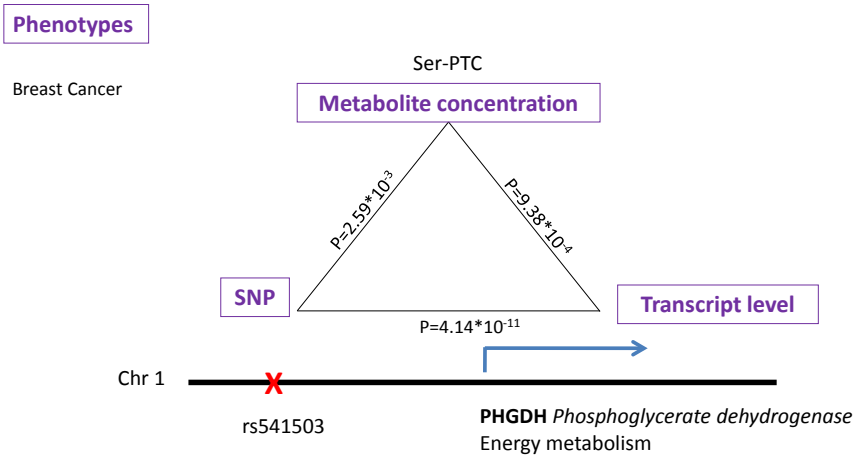

Figure 6.16.: Triangular relationship between genomics, metabolomic, and transcriptomics - *PHGDH*

2. SNP rs174547 is located intronic of *FADS1* (*Fatty Acid Desaturase 1*) and was associated with the expression level of *FADS1* (p-value $= 2.055 * 10^{-14}$) and additionally with the neighboring gene *TREM258* (*Transmembrane Protein 258*, p-value = $3.075 * 10^{-14}$). The p-value for the association between SNP and the metabolite ratio PC aa C36:3 / PC aa C36:4 (Phosphatidylcholine diacyl C36:3 / Phosphatidylcholine diacyl C36:4) was $3.58 * 10^{-69}$. The expression levels of both *FADS1* and *TREM258* were significantly

correlated with the metabolite ratio (p-value $= 1.56*10^{-2}$ and $1.68*10^{-3}$, respectively). The triangular relationship is shown in Figure 6.17.

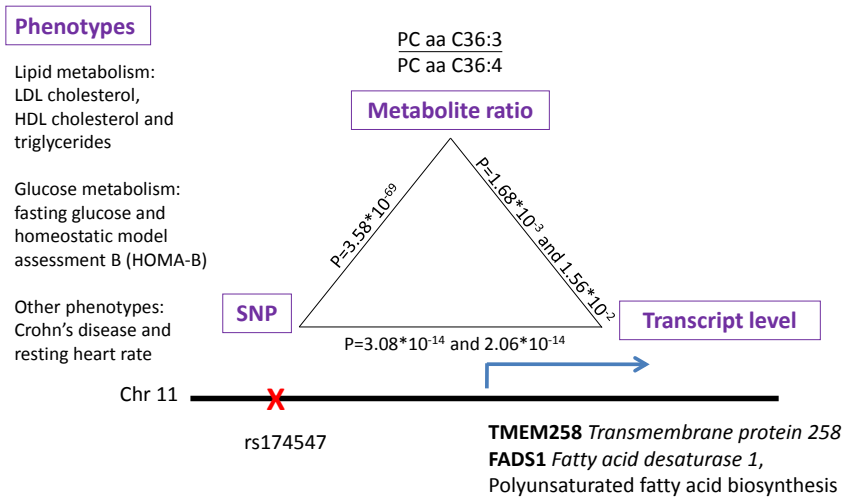

Figure 6.17.: Triangular relationship between genomics, metabolomic, and transcriptomics - *FADS1*

3. SNP rs211718 is located upstream of *ACADM* (*Acetylcoenzyme A Dehydrognase*) and was associated with the expression level of this gene (p-value $= 1.778 * 10^{-31}$) and with two different metabolites: C6(C4:1-DC) (Hexanoylcarnitine (Fumarylcarnitine), p-value $= 3.72*10^{-15}$) and the ratio C12 / C10 (Dodecanoylcarnitine / Decanoylcarnitine, p-value $= 3.87*10^{-24}$). The expression level of *ACADM* was correlated with both metabolite and the metabolite concentration (p-value $= 1.28 * 10^{-3}$ and $9.60 * 10^{-3}$, respectively). Figure 6.18 displays this triangular relationship.

Illig et al. (2010) and Suhre et al. (2011) described that all these three SNPs were known to be associated with clinical traits as cardiovascular disease, resting heart rate, Crohn's disease, and glucose as well as lipid metabolism for rs174547, medium-chain acyl-coenzyme A dehydrogenase deficiency for rs211718, and breast cancer for rs541503. In addition, *PHGDH*, *FADS1*, and *ACADM* are all enzymes encoding genes with functions in human lipid metabolism.

6.2.9. Summary of eQTLs in KORA F4

The eQTL study using data from KORA F4 was at the time of publication one of the largest studies analyzing both *cis*- and *trans*-eQTLs genome-wide in whole blood. 91% of the eQTLs could be replicated in at least one of the two replication cohorts although, there were some systematic differences in the study design between the discovery and one replication cohort.

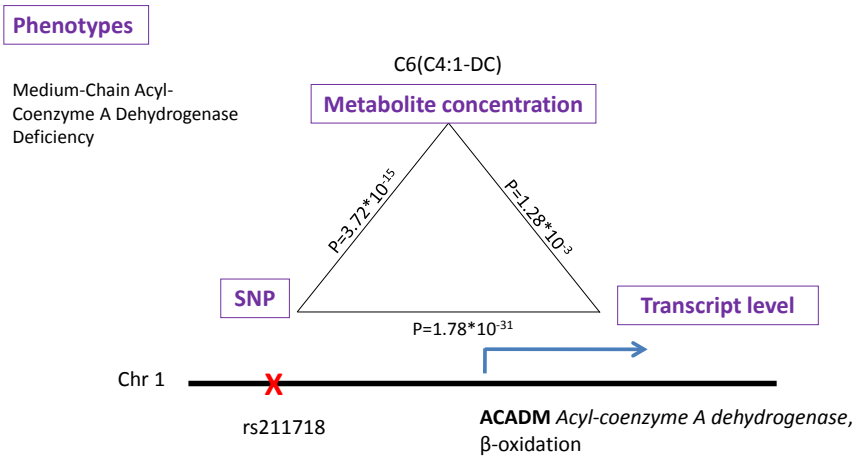

Figure 6.18.: Triangular relationship between genomics, metabolomic, and transcriptomics - *ACADM*

					Illig et al.	Schramm et al.		
					(N=1,809)	(N=717)		
SNP	Chr	Probe Id	Gene	Metabolic trait	p-value[1]	p-value[1]	p-value[2]	p-value[3]
rs541503	1	1704537	*PHGDH*	Ser-PTC	3.06^{-07}	2.59^{-03}	4.14^{-11}	9.38^{-04}
rs211718	1	1778104	*ACADM*	C6(C4:1-DC)	1.74^{-12}	3.72^{-15}	1.78^{-31}	1.28^{-03}
rs174547	11	1786759	*TMEM258*	$\frac{PCaaC36:3}{PCaaC36:4}$	3.88^{-102}	3.58^{-69}	3.08^{-14}	1.68^{-03}
rs211718	1	1778104	*ACADM*	$\frac{C12}{C10}$	5.14^{-34}	3.87^{-24}	1.78^{-31}	9.60^{-03}
rs174547	11	1670134	*FADS1*	PC aa C38:4	3.95^{-27}	1.49^{-15}	2.06^{-14}	1.28^{-02}
rs174547	11	1670134	*FADS1*	$\frac{PCaaC36:3}{PCaaC36:4}$	3.88^{-102}	3.58^{-69}	2.06^{-14}	1.56^{-02}
rs541503	1	1704537	*PHGDH*	$\frac{Thr-PTC}{Ser-PTC}$	5.62^{-11}	1.49^{-05}	4.14^{-11}	4.71^{-02}
rs272889	5	1685057	*SLC22A4*	C5	2.88^{-05}	5.47^{-05}	3.24^{-09}	7.70^{-02}
rs272889	5	1699357	*SLC22A5*	$\frac{Val-PTC}{C5}$	7.69^{-10}	4.77^{-09}	4.50^{-28}	1.15^{-01}
rs272889	5	2050911	*SLC22A4*	$\frac{Val-PTC}{C5}$	7.69^{-10}	4.77^{-09}	1.75^{-17}	1.67^{-01}
rs174547	11	1786759	*TMEM258*	PC aa C38:4	3.95^{-27}	1.49^{-15}	3.08^{-14}	3.01^{-01}
rs8396	4	1758034	*ETFDH*	SM (OH) C24:1	1.60^{-05}	2.49^{-01}	9.27^{-19}	3.30^{-01}
rs272889	5	2050911	*SLC22A4*	C5	2.88^{-05}	5.47^{-05}	1.75^{-17}	3.71^{-01}
rs964184	11	1778668	*TAGLN*	PC aa C34:2	1.30^{-05}	1.19^{-02}	1.14^{-10}	4.20^{-01}
rs272889	5	1699357	*SLC22A5*	C5	2.88^{-05}	5.47^{-05}	4.50^{-28}	4.59^{-01}
rs964184	11	1791912	*SIDT2*	$\frac{PCaaC36:2}{SMC16:0}$	2.22^{-11}	4.33^{-01}	4.63^{-09}	4.88^{-01}
rs8396	4	1758034	*ETFDH*	$\frac{C14:1-OH}{C10}$	8.22^{-15}	5.47^{-11}	9.27^{-19}	5.92^{-01}
rs964184	11	1791912	*SIDT2*	PC aa C34:2	1.30^{-05}	1.19^{-02}	4.63^{-09}	7.04^{-01}
rs272889	5	1685057	*SLC22A4*	$\frac{Val-PTC}{C5}$	7.69^{-10}	4.77^{-09}	3.24^{-09}	7.14^{-01}
rs964184	11	1778668	*TAGLN*	$\frac{PCaaC36:2}{SMC16:0}$	2.22^{-11}	4.33^{-01}	1.14^{-10}	8.53^{-01}

Table 6.6.: Results of cross-associations for congruent metQTL- and eQTL-SNPs:
For six SNPs that were both, *cis*-eQTLs and metQTLs, we calculated the association between transcriptomics, genomics, and metabolomics.
[1]indicates the p-value for the association between the SNP and the metabolic trait or the ratio of two metabolic traits
[2]indicates the p-value for the association between the SNP and the transcript level
[3]indicates the p-value for the association between the transcript level and the metabolic trait or the ratio of two metabolic traits

Together with the large overlap of eQTLs in other tissues there was evidence that whole blood might be an informative and useful tissue for the determination of regulatory effects and useful for biomarker studies.
In this study we combined data derived from multiple omics-technologies (i.e. genotypes, metabolomics, and transcriptomics) and could show that such an approach is useful to identify triangular relationships.

6.3. Replication of eQTLs in CHARGE consortium

With the KORA F4 gene expression and genotype data we also participated in a joint effort of the CHARGE consortium. In this project the KORA eQTL data were used to replicate findings of the discovery cohorts.
One CHARGE investigator Harm-Jan Westra, developed a java-based tool to calculate *cis*-eQTLs faster than using R or PLINK. The software can be downloaded at `https://github.com/molgenis/systemsgenetics/tree/master/eqtl-mapping-pipeline`. As input it requires raw expression data and imputed genotype data.

The discovery set consists of 5,311 samples from seven studies, namely

- EGCUT (n=891)
- InCHIANTI (n=611)
- Rotterdam Study (n=762)
- Fehrmann (n=1,469)
- HVH (n=106)
- SHIP-TREND (n=963)
- DILGOM (n=509)

Altogether, 1,513 *trans*-eQTLs including 346 unique SNPs having an effect on 430 different genes were detected using 4,542 SNPs from the GWAS Catalog that had already been published with different traits or diseases. 52% of these significant associations could be replicated in 740 samples of the KORA F4 study. The second replication cohort was the Brisbane Systems Genetics Study (BSGS) which also had expression data from whole blood. This study consists of 862 samples and due to the larger sample size 79% of all significant hits could be replicated. When using both studies for a meta-analysis 89% were significantly replicated and beside significance 97% showed consistent allelic direction. This rate was 91% in KORA F4 only.

6.4. Summary and discussion

In summary, three different eQTLs studies were conducted using different sample sizes. The aim of all studies was mainly to show that whole blood is a useable tissue for eQTL studies.

Our two eQTL studies could demonstrate the high robustness and reproducibility of genetically determined regulatory effects in whole blood. In both studies high replication rates could be observed and for the larger KORA F4 study there was also a high overlap with eQTLs in other tissues.
When comparing the numbers of significant *cis*- and *trans*-eQTLs in Table 6.7 it is conspicuous that the difference of significant hits from 322 samples to 890 is relatively much higher (more than ten times more hits for *cis*- and *trans*-eQTLs) than from 890 to 5,311 samples (1.4 times more hits for *cis*- and 2.4 times more hits for *trans*-eQTLs, respectively) indicating that the number of detectable eQTLs in whole blood might not be infinitely. Nevertheless, a larger eQTL study is already in progress and eQTLs are calculated in the framework of the eQTLGen Consortium. It will finally consist of more than 20,000 samples and we will join with the KORA F4 data.

Cohort	KORA F3	KORA F4	CHARGE
Samples	322	890	5,311
Number of tested SNPs (*cis*)	335,152	616,941	1,962,237
Number of tested SNP probe pairs (*cis*)	4,802,373	8,308,092	11,172,453
Bonferroni threshold (*cis*)	1.03E-8	6.02E-9	4.5E-9
Number of significant *cis*-eQTLs	363	3,449	4,690
Number of tested SNP probe pairs (*trans*)	13,878,309,168	17,867,228,301	153,134,630
Tested SNPs	335,152	616,941	4,542
Tested expression probes	41,409	28,961	34,061
Bonferroni threshold (*trans*)	3.6E-12	2.81E-12	3.3E-10
Number of significant *trans*-eQTLs	8	94	223

Table 6.7.: Summary of eQTL results from KORA F3, KORA F4, and CHARGE

In all eQTL analyses decribed above linear regression models were used to assess the association between SNPs and gene expression levels. However, linear regression requires normal distribution and in spite of extensive efforts to normalize transcriptomics data the normalization is an approximation. Beside the linear regression the Spearman's rank correlation test is often used to test the association between SNP and gene expression levels in eQTL studies (Sul et al., 2015). Its advantage is that this is a non-parametric test and it is quite robust to any deviations from the normal distribution. The software PLINK is not able to calculate this test, therefore, the analysis was repeated only for all possible *cis* SNP-probe combinations in KORA F4.
The results of the comparison between the linear model and the Spearman's rank correlation test are summarized in Table 6.8. 96% of the expression probes that were significant when analyzing all SNP-probe combinations using a linear model were also significant when using the Spearman's rank correlation test indicating comparability of both tests in our data set. For future eQTL projects it could be debated if it makes sense to prefer Spearman's rank test for a better comparability with results from other cohorts.

For the eQTL analyses in KORA F3 and F4 the multiple testing problem was solved by applying a Bonferroni correction. Although this correction is very conservative and stringent the large sample sizes of the discovery cohort (KORA F4) and of the replication cohorts allow to identify and verify a still high number of true-positive *cis*-eQTLs. In the eQTL

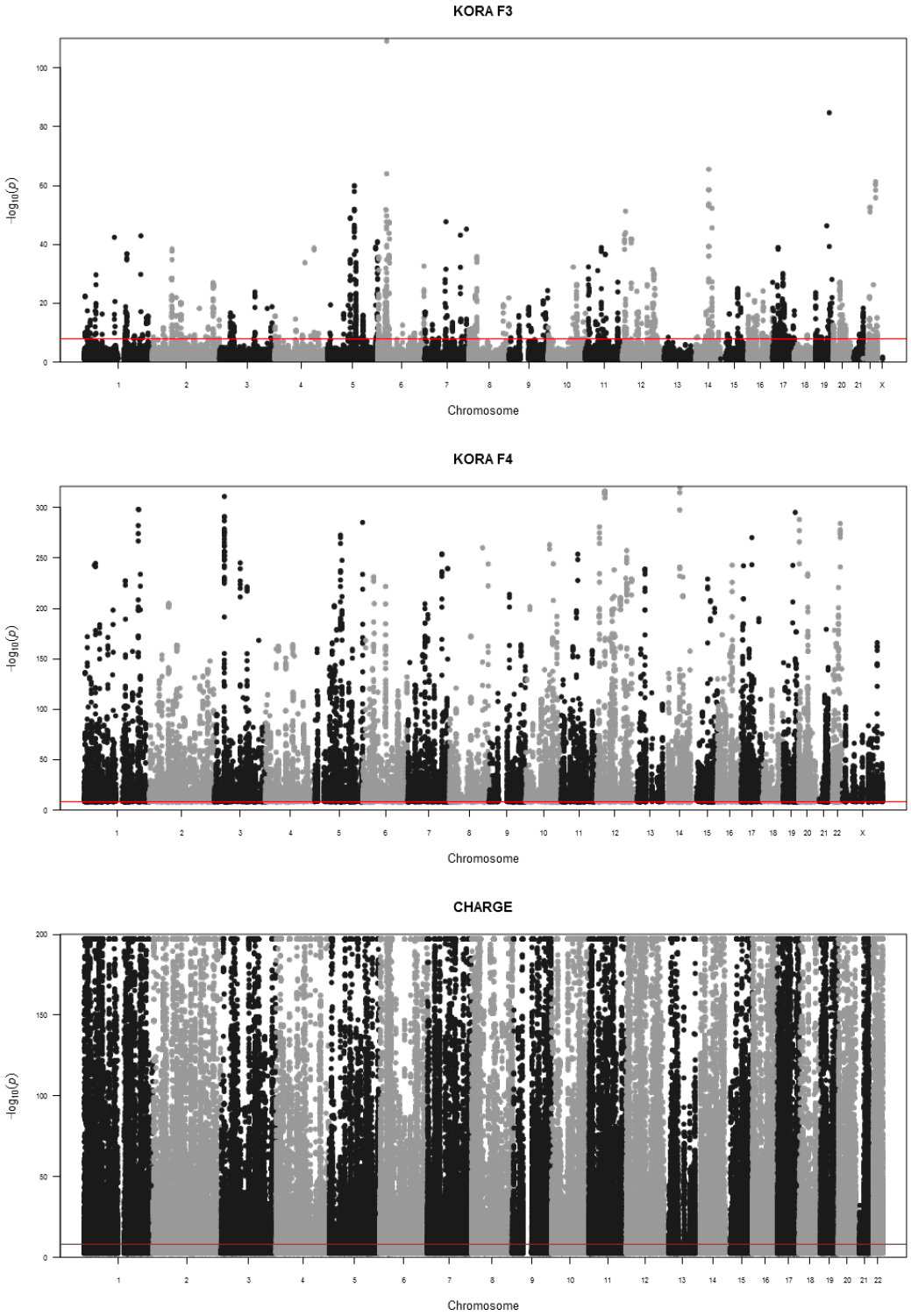

Figure 6.19.: Manhattan plots of *cis*-eQTLs in KORA F3, F4, and in CHARGE:
The genomic position of the SNPs is plotted against the -log10(p-value). The read line depicts the Bonferroni threshold of significance.

	Linear model	Spearman's rank test	Overlap
significant SNP-probe combinations	55,592	55,051	52,503
significant unique probes	4,116	4,106	3,958

Table 6.8.: Comparison of results obtained from linear regression and Spearman's rank correlation test for *cis*-hits in KORA F4:
The calculations were conducted for 8,308,092 possible SNP-probe combinations in 890 KORA F4 samples. The significance threshold was $6.02 * 10^{-09}$ (Bonferroni threshold).

project of CHARGE standard procedure for calculating p-values in eQTL studies was used: the permutation (but they also indicated the number of significant hits for the Bonferroni correction). The advantage is that it also takes into account the LD structure of the genome which is not considered when using Bonferroni correction (Sul et al., 2015). The disadvantage and the reason why it was not possible for us to use it for the *trans*-calculation is that it is very time- and computer-intensive. In the CHARGE consortium this problem was minimized by reducing the number of analyzed SNPs and by doing only ten permutations which is quite below the standard of several thousand permutations (Sul et al., 2015).
Wright et al. (2014) published the last eQTL study so far using more than 2,500 samples. This almost reached dimensions of GWAS which normally also apply a multiple testing correction that is based on Bonferroni. Normally here a significance threshold of $5 * 10^{-8}$ is used which takes the LD structure into account.

The advantage of the eQTL studies in KORA F3 and F4 was that we were able to calculate the associations between all expression levels and all available SNPs and due to the lower number of significant hits we were able to perform some additional analyses like the comparison of our two replication cohorts of KORA F4. In spite of various systematic differences in the study design of the two replication samples (fasting versus non-fasting status of participants, time of blood collection, and different laboratory tools, namely, PAXgeneTM versus TempusTM tubes) replication overlaps of both replication samples were comparable (78% and 82%). Together with a total replication overlap of 91% in at least one of the samples this demonstrates a high robustness and reproducibility of genetically determined regulatory effects in whole blood. Consequently, whole blood gene expression provides a means for the discovery of biomarkers which are of clinical relevance for the perturbation of the system in a disease status.
Finally, we could include metabolomics data to show that eQTL studies provide a hypothesis-free approach to link genetic variation with human metabolism.

7. Summary and outlook

The progress and rapid advances in the field of gene expression data during my PhD project was highly visible and is continuously ongoing.
We started with two case-control studies with small sample sizes. In one study we compared eight Parkinson patients with nine old controls. This resulted in the discovery of four differentially expressed genes (Elstner et al., 2009). The inclusion of seven young controls allowed us to describe Parkinson's disease as accelerated aging (Elstner et al., 2011).
In the second case-control study 13 NBIA patients were compared to six controls. Here, no significant differentially expressed gene could be identified, but the top 500 genes were used for a pathway analysis and showed pathways that are significantly related to mitochondrial functions (Hartig et al., 2011). Nevertheless the second project was the beginning of a large NBIA study (TIRCON project). Within the framework of this study the gene expression levels of 41 patients and 42 controls were determined. In comparison to all data that were used in this thesis no arrays were used but the RNA was sequenced. In comparison to array-based methods sequencing is less susceptible to technical artifacts, much more accurate and not limited by background noise for low signals. Despite of all the advantages so far arrays were much often used due to the lower costs. Now the RNA-sequencing is getting cheaper and therefore more lucrative also for larger sample sizes.
Therefore we started to analyze the gene expression levels obtained from RNA-sequencing of 41 PKAN patients with 42 healthy age-matched controls. Similar to the preprocessing of data from arrays the raw values have to be normalized before the analysis. Here we used the FPKM values (for details see A.2.2) that consider the number of mapped reads and the length of the genes.

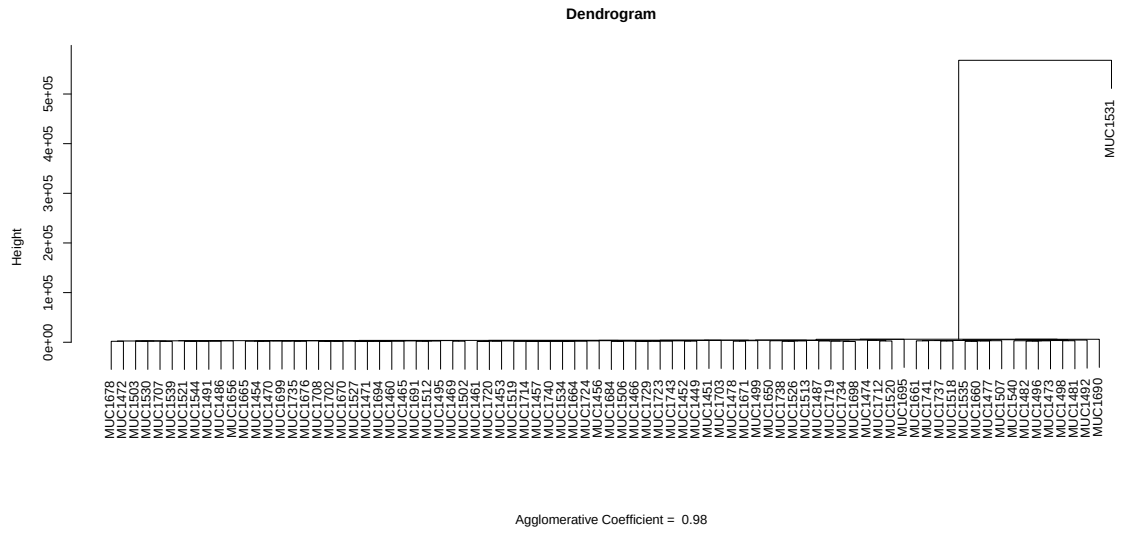

Figure 7.1.: Dendrogram of RNASeq data of 42 controls and 41 NBIA patients

In a next step an outlier was detected by clustering all samples using all expression values (see Figure 7.1). For this sample a mistake in the laboratory was already reported and a repetition is in progress. The expression levels of the remaining 40 patients and 42 controls were compared using a linear model adjusted for age and gender. Especially the age could be an important confounder due to the fact that younger patients are more likely to be affected by a more severe form of the disease.

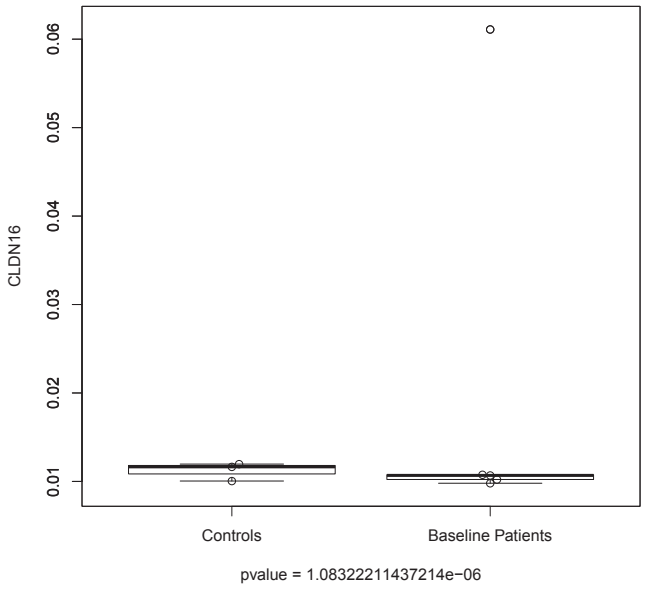

Figure 7.2.: Experiment using RNASeq data: Differentially expressed gene between controls and NBIA patients:
When comparing 40 NBIA patients and 42 controls using a linear model adjusted for age and gender, one gene was differentially expressed when applying the Benjamini-Hochberg correction.

When applying the Bonferroni correction no gene was differentially expressed between the controls and the patients. Using the Benjamini-Hochberg correction, we could identify one gene called *CLDN16* (see Figure 7.2), but the expression level of this gene was quite low

(< 1) and we cannot guarantee that the effect is not due to background noise. So far no other data set with the same patients is available and the result cannot be replicated. As the measurement of more samples is in progress, we hope that we can verify this result in a higher sample size or identify some other interesting candidate genes.

The first eQTL study using RNA sequenced data of 156 individuals was published already (Sul et al., 2015) and it should be only a question of time that these data are available for the KORA samples. Then, all analyses conducted in this thesis can be repeated in a data set with higher quality.

After the case-control studies we established one of the first larger gene expression data set from a population-based study namely KORA F3 in Munich. This data set was used to establish the first quality scores and optimize the experimental protocol. At this time-point it was one of the largest studies with 381 individuals. We could improve the quality of this data set and establish an even more homogenous data set with the samples from the KORA F4 cohort. When this data was ready for analysis, the German consortium MetaXpress was founded to harmonize the analysis of genome-wide expression data. We herein established a well-structured analysis protocol.
The KORA F3 and F4 data sets were used to analyze the effects of different phenotypes and variables on the expression levels. We started with blood pressure related phenotypes which seem to have a small impact on expression levels in whole blood. Even in the large CHARGE consortium with more than 7,000 samples only 34 genes could be identified that are associated with one of the phenotype (diastolic/systolic blood pressure or hypertension).
The impact of age on the expression levels is much higher. Here the findings in KORA F3 and F4 increased from five to 194 significantly associated genes. In the large CHARGE consortium with more than 7,000 samples we identified almost 1,500 significant genes that could be replicated in an independent cohort.
Lastly, it was analyzed if the genetic variation changes the expression levels of genes in *cis* or in *trans*. With 322 KORA F3 samples we detected 363 *cis*- and eight *trans*-associations and increased this number with 890 KORA F4 samples to 3,449 *cis*- and 94 *trans*-associations. With this sample size the eQTL study was one of the largest whole genome eQTL analysis that analyzed effects in both *cis* and *trans* in whole blood samples in European populations so far. Even in the much larger eQTL study of CHARGE with 5,311 samples not all *trans*-effects were determined as they were using only SNPs that are reported in the GWAS catalog at this time-point. The next study in CHARGE is already in progress and it is planned to include more than 20,000 samples to analyze both *cis*- and *trans*-effects genome-wide.

The next project is the genome-wide analysis of gene expression levels with methylation, which already started. For the analysis of gene expression and age we analyzed for significantly associated genes the impact of methylation on gene expression levels. Additionally, we tested if the effect of aging on the expression levels is mediated by methylation. We identified 1,248 age associated genes (83%) that had at least one mediating CpG site whereas each gene had between one and 154 mediating CpG sites.
To analyze these effects genome-wide we have to consider the high number of CpG sites (more than 450,000) and the fact that we do not know if these effects are only visible in *cis* (< 500 kb distance between gene and CpG site) or also in *trans*. To conduct these calculations in an acceptable speed and time, our collaborators from the Framingham Heart Study (FHS)

developed a GPU-based code that reduced the calculation time from several weeks to just 20 minutes. The disadvantages of this program are, that it is not very flexible and can calculate linear models only without any further covariables. We avoided this problem by adjusting the data for all relevant covariables before the calculation.
Using the Bonferroni correction we could identify in this way 3,422 *cis*- and 39,859 *trans*-associations. The replication (especially of the high number of *trans*-associations) is still ongoing. In a next step it is also possible to include genetic data to identify also SNPs that are responsible for the variation in the methylation and expression levels and to see if expression levels are a consequence of methylation or the other way round.

A. Appendix

A.1. List of abbreviations

ANOVA Analysis of variance

BMI Body mass index

BP Blood pressure

CHARGE The Cohorts for heart and aging research in genomic epidemiology

cDNA Coding deoxyribonucleic acid

cRNA Coding ribonucleic acid

DBP Diastolic blood pressure

DNA Deoxyribonucleic acid

DZHK Deutsches Zentrum für Herz-Kreislaufforschung - German Center for Cardiovascular Research

ECGUT Biobank of Estonian Genome Center of the University of Tartu

eQTL Expression quantitative locus

FDR False discovery rate

FWER Family-wise error rate

GHS Gutenberg Health Study

GWAS Genome-wide association study

HTN Hypertension

IPA Ingenuity pathway analysis software

KORA Kooperative Gesundheitsforschung in der Region Augsburg - Cooperative Health Research in the Region of Augsburg

LD Linkage disequilibrium

MAF Minor allele frequency

mRNA Messenger ribonucleic acid

NBIA Neurodegeneration with brain iron accumulation

OGTT Oral glucose tolerance test

PC Principal component

PCA Principal component analysis

PCR Polymerase chain reaction

PD Parkinson's disease

QC Quality control

qPCR Quantitative real-time polymerase chain reaction

QTL Quantitative trait locus

RIN RNA integrity number

RNA Ribonucleic acid

SBP Systolic blood pressure

SE Standard error

SNP Single nucleotide polymorphism

TSS Transcription start site

VST Variance stabilization transformation

A.2. Statistics

A.2.1. Variance stabilization transformation

The variance stabilization transformation (VST) was developed to stabilize the variance of gene expression data measured with Illumina arrays. In comparison to the log2 transformation (which is based on the average expression level of each probe), it takes the several measurements of one probe into account.
The following steps have to be performed to find a function h for Y where the variance of the transformed Y is independent from the mean (Lin et al., 2008):

1. All background probes with detection p-values higher than a predefined threshold (normally 0.01) are selected as it is assumed that these probes measure the background noise.

2. The variance of the background noise is estimated (called c_3).

3. The following formula is used to calculate c_1 and c_2:

$$\sqrt{v(u) - c_3} = c_1 u + c_2$$

where $v(u)$ is the relationship between the mean u and the variance v of the measurement Y and can be displayed in a quadratic form

$$v(u) = (s_\eta / m_\eta)^2 (u - \alpha)^2 + \sigma_\epsilon^2 = (c_1 u + c_2)^2 + c_3$$

where m_η and s_η are the mean and the standard deviation of e^η (coming from the general model $Y = \alpha + \mu e^\eta + \epsilon$), α is the offset of the general model, and σ_ϵ is the standard deviation of ϵ.

4. The transformed values are calculated using:

$$h(y) = \begin{cases} 1/c_1 \operatorname{arcsin}(c_2/\sqrt{c_3} + c_1 y/\sqrt{c_3}), & \text{when } c_3 > 0 \\ 1/c_1 \ln(c_2 + c_1 y), & \text{when } c_3 = 0 \end{cases}$$

It can be shown that the VST-transformed values are close to the log2-transformed values for high expression values.
The VST is implemented in the Bioconductor package `lumi` (Du et al., 2008).

A.2.2. RNA Sequencing and FPKM

RNA sequencing is a next-generation sequencing based technology to measure the gene expression levels. The RNA is converted to a library of cDNA fragments and each molecule is sequenced from both ends (pair-end sequencing). In our lab the Illumina HiSeq2500 was used which produces reads with a length of 100 base pairs. Afterward the reads were mapped to a reference genome (Wang et al., 2009).
To take the number of mapped reads and the length of genes in account, the expression values are indicated as FPKM values. FPKM is the abbreviation for **F**ragments **P**er **K**ilobase of transcript per **M**illion mapped reads.

A.3. Tables and Figures

genomic locus	top SNP	CHR SNP	ProbeId	Gene	CHR Gene	P-value KORA	P-value SHIP-TREND	P-value EGCUT
1	rs10736767	11	ILMN_2336609	*SYTL2*	11	1.60E-21	6.65E-15	5.29E-08
2	rs862242	5	ILMN_1653039		10	1.75E-153	7.71E-191	6.01E-262
3	rs2141180	3	ILMN_1654946	*ZSCAN18*	19	2.76E-17	7.99E-19	2.34E-18
3	rs9859077	3	ILMN_1658173	*ZNF418*	19	2.67E-12	2.39E-03	2.77E-04
3	rs2141180	3	ILMN_1724052	*ZNF814*	19	1.98E-23	4.89E-06	1.84E-16
3	rs2141180	3	ILMN_1726368	*ZNF135*	19	5.49E-14	1.16E-03	8.74E-05
4	rs6600227	16	ILMN_1658957		1	4.93E-15	1.63E-13	5.30E-09
5	rs12485738	3	ILMN_1660114	*MMRN1*	4	1.23E-14	8.20E-09	2.62E-05
5	rs12485738	3	ILMN_1668374	*ITGB5*	3	9.89E-13	8.66E-19	5.63E-06
5	rs12485738	3	ILMN_1671928	*PROS1*	3	7.82E-19	6.38E-12	1.26E-09
5	rs12485738	3	ILMN_1691264	*NAT8B*	2	1.06E-13	5.56E-13	1.49E-01
5	rs12485738	3	ILMN_1713731	*ALOX12*	17	2.61E-12	1.84E-15	1.22E-03
5	rs12485738	3	ILMN_1729453	*TSPAN9*	12	5.69E-16	2.76E-21	2.76E-12
5	rs12485738	3	ILMN_1730487	*CALD1*	7	1.26E-13	1.43E-15	1.76E-03
5	rs1344142	3	ILMN_1733324	*ITGB3*	17	6.35E-13	8.77E-16	8.61E-08
5	rs12485738	3	ILMN_1743290	*GP9*	3	1.82E-15	7.40E-18	2.10E-05
5	rs12485738	3	ILMN_1745103	*CLEC1B*	12	6.81E-13	2.01E-07	1.28E-01
5	rs1359142	6	ILMN_1767766	*PRDX2*	19	5.38E-16	4.48E-16	3.84E-17
5	rs12485738	3	ILMN_1787919	*PARVB*	22	2.77E-12	9.84E-23	2.28E-10
5	rs12485738	3	ILMN_1808590	*GUCY1A3*	4	3.58E-13	6.38E-12	3.22E-10

5	rs12485738	3	ILMN_2392189	*CTDSPL*	3	9.16E-14	1.76E-18	5.16E-10
5	rs10463053	5	ILMN_1798557	*FAM153A*	5	3.38E-21	5.34E-13	5.75E-06
6	rs10252204	7	ILMN_1660927	*FAM185A*	7	1.08E-18	1.12E-05	7.31E-02
7	rs9272346	6	ILMN_1661266		6	7.15E-116	7.00E-164	5.71E-108
7	rs502771	6	ILMN_1752592	*HLA-DRB4*	6	5.90E-45		7.07E-60
7	rs9272346	6	ILMN_1791534		6	2.09E-17	4.99E-30	2.57E-12
7	rs9272346	6	ILMN_2159694	*HLA-DRB4*	6	4.43E-29	2.55E-32	1.74E-13
7	rs9272346	6	ILMN_2165753	*HLA-A*	6	9.51E-18	3.83E-37	2.40E-06
8	rs10742583	11	ILMN_1663132	*ADCK2*	7	2.55E-14	1.18E-15	4.30E-11
8	rs10742583	11	ILMN_1688753	*PTDSS1*	8	5.79E-74	6.46E-47	7.35E-64
8	rs10742583	11	ILMN_1690963	*ASAP1*	8	5.59E-14	5.75E-16	7.16E-06
8	rs10742583	11	ILMN_1695386	*RAD51C*	17	8.91E-15	2.20E-03	4.85E-21
8	rs10742583	11	ILMN_1716445	*DTYMK*	2	1.69E-19	5.17E-13	3.92E-09
8	rs10742583	11	ILMN_1727458	*HDAC1*	1	4.30E-28	3.82E-26	8.63E-34
8	rs10742583	11	ILMN_1743049	*PWP1*	12	1.20E-36	1.65E-79	6.68E-14
8	rs10742583	11	ILMN_1752086	*TRAPPC11*	4	9.90E-16	6.92E-32	5.60E-06
8	rs10742583	11	ILMN_1754579	*SYNE2*	14	5.92E-15	4.37E-04	3.73E-19
8	rs10742583	11	ILMN_1769319	*CNBP*	3	1.01E-26	2.18E-33	1.55E-11
8	rs10742583	11	ILMN_1795428	*WDR59*	16	8.94E-23	1.79E-24	2.43E-04
8	rs10742583	11	ILMN_1795876	*GPS1*	17	4.36E-16	3.88E-21	5.88E-06
8	rs10742583	11	ILMN_1796464	*WDR37*	10	8.25E-25	1.74E-09	3.00E-17
9	rs9939012	16	ILMN_1667034	*PDPR*	16	5.82E-15	5.21E-08	7.06E-17
9	rs9939012	16	ILMN_1778013	*PDPR*	16	6.95E-15	2.70E-05	2.87E-11
10	rs225245	17	ILMN_1671686	*EPB49*	8	1.22E-12	2.96E-16	7.43E-21
10	rs225245	17	ILMN_1697529	*RNF10*	12	4.15E-16	1.87E-19	2.91E-10
11	rs12933929	16	ILMN_1671809	*DUSP22*	6	2.37E-23	8.98E-24	1.21E-17
11	rs12933929	16	ILMN_1813275	*DUSP22*	6	1.99E-25	2.70E-28	8.31E-23
11	rs12933929	16	ILMN_2159152	*TP53TG3B*	16	2.45E-15	3.60E-07	3.32E-06
12	rs3779106	7	ILMN_1673337	*PMS2CL*	7	1.50E-22	4.52E-06	3.84E-04
13	rs10417909	19	ILMN_1673991	*ATIC*	2	4.29E-21	1.28E-31	4.45E-27
14	rs6901565	6	ILMN_1676459	*ZNF785*	16	5.20E-14	1.25E-08	3.43E-04
15	rs2856553	16	ILMN_1678730	*NOMO1*	16	3.36E-21	1.49E-09	8.08E-03
15	rs2856553	16	ILMN_2126957	*NOMO1*	16	1.38E-13	6.82E-13	1.16E-04
16	rs4130579	19	ILMN_1680388		1	4.74E-14	2.61E-06	1.94E-22
17	rs12669559	7	ILMN_1684255	*MYL4*	17	2.32E-16	2.97E-13	2.51E-14
17	rs12669559	7	ILMN_1695058	*SLC38A5*	X	1.10E-16	5.61E-14	1.72E-15
17	rs12669559	7	ILMN_1787526	*C2ORF88*	2	4.72E-33	2.17E-41	2.67E-16
18	rs3745902	19	ILMN_1685843	*KIR2DL5B*	19	6.25E-24	5.69E-06	3.32E-15
18	rs3745902	19	ILMN_1793451	*KIR2DL5B*	19	3.12E-56	9.86E-30	2.45E-26
18	rs3745902	19	ILMN_2415650	*KIR2DL5B*	19	5.33E-38	6.07E-12	2.56E-20
19	rs845787	20	ILMN_1688318	*FRG1B*	20	4.56E-24	4.55E-33	5.24E-12
20	rs10861779	12	ILMN_1692962	*CTDSP2*	12	1.65E-15	3.55E-19	1.41E-06
21	rs13053817	22	ILMN_1697286	*SF3A1*	22	5.57E-13	1.89E-14	1.34E-05
22	rs12151742	2	ILMN_1652161	*PNKD*	2	1.89E-28	9.91E-08	7.22E-30
22	rs12151621	2	ILMN_1679130	*CALHM1*	10	2.24E-25	3.31E-12	4.24E-26
22	rs12151621	2	ILMN_1697317	*DYNLRB2*	16	1.12E-25	1.14E-08	8.48E-35
22	rs12151621	2	ILMN_1724158	*ZNF93*	19	7.62E-13	1.24E-04	2.30E-09
22	rs12151621	2	ILMN_1740186	*GHRHR*	7	6.93E-17	9.43E-03	1.51E-16
22	rs12151621	2	ILMN_2282282	*MLH3*	14	1.31E-38	1.50E-20	1.59E-44
23	rs5750715	22	ILMN_1697710	*IL37*	2	5.30E-27	6.02E-18	9.25E-11
24	rs592423	6	ILMN_1702447	*IGF2BP2*	3	1.20E-13	3.82E-15	4.04E-16
25	rs12189695	6	ILMN_1702866	*LINC00115*	1	9.77E-37	4.50E-09	8.02E-16
26	rs6854996	4	ILMN_1703246	*SBF1*	22	1.86E-13	7.79E-16	1.85E-04
27	rs7165535	15	ILMN_1703538	*AIF1*	6	5.58E-38	1.18E-14	9.65E-39
28	rs11247355	15	ILMN_1704376	*GLDN*	15	1.43E-19	3.61E-33	8.91E-11
29	rs11642055	16	ILMN_1709747	*EXOG*	3	1.75E-16	4.37E-13	2.51E-08
30	rs4792935	17	ILMN_1710207	*FAM178A*	10	8.44E-13	2.94E-30	4.69E-02
31	rs199448	17	ILMN_1712657	*BPTF*	17	1.45E-94	3.07E-20	3.16E-10
31	rs199448	17	ILMN_1743621	*C17ORF69*	17	1.45E-43	1.38E-16	7.21E-18
31	rs199448	17	ILMN_1772603	*LOC644172*	17	2.69E-144	3.26E-54	4.27E-07
31	rs199448	17	ILMN_1784428	*MGC57346*	17	1.32E-103	5.49E-72	1.29E-46
32	rs700415	7	ILMN_1712721	*GAST*	17	1.69E-16	3.33E-10	5.30E-17
33	rs4390943	3	ILMN_1712936	*GCFC1*	21	7.59E-36	5.29E-27	3.34E-16
34	rs615672	6	ILMN_1717261	*HLA-DRB1;HLA-DRB3*	6	7.45E-21	7.30E-06	9.68E-26
35	rs11130549	3	ILMN_1717793	*C19ORF33*	19	1.32E-12	1.74E-05	2.39E-02
36	rs1476415	21	ILMN_1732467	*OR2AG1*	11	3.05E-27	1.36E-22	5.08E-25
37	rs7005151	8	ILMN_1738793	*ZNF71*	19	2.14E-36		3.09E-21
38	rs10814410	9	ILMN_1739199	*WASH3P*	15	5.85E-22	9.84E-05	
39	rs12612045	2	ILMN_1739942	*FAM117B*	2	9.94E-15	9.09E-18	7.19E-16
40	rs11171739	12	ILMN_1740094	*BEND4*	4	1.62E-67	2.90E-54	1.05E-13
40	rs11171739	12	ILMN_1753440	*DCAF16*	4	1.55E-17	2.77E-20	1.66E-05
40	rs11171739	12	ILMN_2180866	*RPS26P11*	X	2.80E-248	3.53E-280	3.26E-17
41	rs12744267	1	ILMN_1745885	*POLR2F*	22	3.74E-13	2.04E-25	2.26E-02
42	rs791900	6	ILMN_1749070	*HLA-DPB1*	6	3.26E-21	2.89E-23	1.05E-24
43	rs7708899	5	ILMN_1754156	*FLJ45340*	7	1.14E-32	1.65E-21	1.28E-02
44	rs470119	22	ILMN_1761321	*KREMEN2*	16	1.13E-13	1.97E-05	2.33E-22
45	rs4704164	5	ILMN_1770498	*GUSBP1*	5	3.01E-19	3.17E-08	9.37E-13
45	rs4704164	5	ILMN_1780700	*GUSBP2*	6	7.79E-14	4.68E-06	1.62E-12
45	rs4704164	5	ILMN_1813191		5	6.61E-38	3.40E-41	6.97E-31
46	rs2244468	9	ILMN_1772888		17	1.23E-41	2.66E-38	9.55E-45
47	rs2217065	15	ILMN_1775590	*SCAND2*	15	9.76E-16		1.54E-05
48	rs2942219	8	ILMN_1778213	*STK33*	11	1.80E-19	2.57E-08	1.85E-01
49	rs11871616	17	ILMN_1779572	*PLEKHM1*	17	2.61E-84	4.35E-80	8.26E-21
50	rs7395116	11	ILMN_1783753	*TXNDC12*	1	2.20E-46		2.49E-16
51	rs2353678	19	ILMN_1784352	*CCM2*	7	3.52E-18	3.37E-05	1.18E-10
52	rs12936834	17	ILMN_1789407	*FAM18B2*	17	6.02E-32	3.34E-05	1.66E-19
53	rs4447245	12	ILMN_1791297		8	2.60E-28		1.88E-10
54	rs4895441	6	ILMN_1792455	*TMEM158*	3	6.10E-15	7.12E-23	1.67E-12
54	rs4895441	6	ILMN_1796678	*HBG2;HBG1*	11	7.41E-18	1.88E-38	7.22E-32
54	rs4895441	6	ILMN_2084825	*HBG2*	11	5.54E-16	6.61E-34	9.36E-28

55	rs10172646	2	ILMN_1796678	*HBG2;HBG1*	11	1.40E-19	5.28E-27	1.99E-12
55	rs10172646	2	ILMN_1806710	*ESPN*	1	2.52E-23	7.43E-19	1.03E-13
55	rs10172646	2	ILMN_2084825	*HBG2*	11	1.93E-17	1.91E-23	2.12E-12
57	rs9534145	13	ILMN_1802023	*OR7E156P*	13	8.26E-13	1.20E-01	2.20E-07
58	rs291040	16	ILMN_1803945		6	4.34E-15	2.60E-15	1.66E-10
59	rs225245	17	ILMN_1806349	*SLC6A8*	X	2.83E-14	1.12E-21	6.33E-06
60	rs6503934	17	ILMN_1810274	*HOXB2*	17	2.79E-20		1.14E-10
61	rs10000012	4	ILMN_1811443	*AVP*	20	4.92E-32	1.11E-07	2.38E-31
61	rs10000012	4	ILMN_2406439	*SPTBN4*	19	4.26E-18	9.52E-08	4.77E-26
62	rs1317548	10	ILMN_1906187	*LOC283070*	10	3.26E-25	1.71E-30	1.33E-18
63	rs642112	1	ILMN_2050023	*CCDC23*	1	8.58E-19		1.14E-20
64	rs10784774	12	ILMN_2051900	*EID2B*	19	6.00E-30	3.30E-11	6.16E-68
64	rs10784774	12	ILMN_2095653	*AFMID*	17	3.37E-30	8.42E-45	1.10E-118
64	rs10784774	12	ILMN_2130078	*CDKN2AIPNL*	5	2.74E-13	6.28E-09	4.50E-36
64	rs10784774	12	ILMN_2134381	*ITPK1-AS1*	14	2.02E-12	2.40E-15	3.93E-34
64	rs10784774	12	ILMN_2182482	*SHCBP1*	16	3.29E-68	5.99E-54	3.59E-145
64	rs10784774	12	ILMN_2334242	*CREB1*	2	5.33E-48	2.65E-59	1.61E-86
64	rs10784774	12	ILMN_2412521	*KIAA0101*	15	2.18E-35	8.33E-21	1.89E-51
65	rs4723226	7	ILMN_2089977	*FKBP9L*	7	6.28E-31	1.06E-28	4.99E-28
66	rs10203656	2	ILMN_2094416	*PLGLB2*	2	1.11E-61	6.23E-173	1.77E-63
67	rs10950029	7	ILMN_2118663	*ZNF117*	7	5.43E-24	6.49E-18	7.18E-25
68	rs3810444	19	ILMN_2134224	*ATP13A1*	19	3.48E-28	7.74E-46	1.72E-15
69	rs4665083	2	ILMN_2142752	*MANSC1*	12	2.68E-12	3.22E-06	2.53E-07
70	rs3748144	8	ILMN_2144088	*FDFT1*	8	6.45E-14	5.26E-09	1.53E-02
70	rs3763114	5	ILMN_2313926	*CDC42SE2*	5	1.73E-12	1.48E-11	6.30E-01
71	rs8058597	16	ILMN_2147105		16	6.22E-173	2.11E-79	4.64E-32
72	rs1555823	10	ILMN_2151056	*C10ORF32*	10	4.23E-21	6.28E-14	2.61E-17
73	rs11673276	19	ILMN_2175715	*KIR2DS5*	19	1.52E-59		3.84E-36
74	rs2912495	8	ILMN_2199439	*CA2*	8	1.37E-12	7.94E-16	3.03E-24
75	rs7921218	10	ILMN_2209578	*AGAP6*	10	9.43E-16	1.72E-06	1.57E-05
76	rs10116248	9	ILMN_2276758	*POFUT1*	20	1.45E-12	1.05E-22	7.64E-58
77	rs13048152	21	ILMN_2291619	*RAB3IP*	12	3.03E-34	4.16E-26	4.82E-21
78	rs11633427	15	ILMN_2320480		15	1.42E-18	3.72E-09	3.02E-06
79	rs3917989	1	ILMN_2334242	*CREB1*	2	4.51E-18	1.61E-04	1.44E-03

Table A.1.: Significant *trans*-eQTLs in KORA F4

	ILMN_1724609	ILMN_1743078	ILMN_1780382	ILMN_1786388	ILMN_1816342	ILMN_2408645
Matched_Genes	SLC2A8		LOC653566	RNF113A		SPCS2,LOC653566
Matched_CHR	9		1	X	17	
QC comment	good	no matched mRNA	good	good	good	>1 matched chr
rs28358576	5.04E-02	6.35E-01	**3.95E-51**	**1.26E-34**	6.19E-01	**3.33E-12**
rs3928306	**6.50E-28**	3.27E-05	1.98E-04	4.85E-02	1.20E-01	2.82E-01
rs2854131	7.22E-02	4.12E-01	3.01E-01	9.89E-02	**1.42E-11**	9.38E-02
rs9743	7.11E-03	9.64E-02	**1.84E-110**	3.84E-59	8.31E-01	**1.23E-22**
rs28358280	5.45E-03	9.80E-02	**5.57E-116**	**1.07E-60**	7.47E-01	**4.65E-23**
rs3915952	6.25E-01	**8.95E-11**	2.53E-03	4.52E-03	6.09E-01	6.00E-01
rs28358285	4.47E-02	1.49E-01	**5.66E-74**	**1.45E-37**	9.25E-01	**1.03E-13**
rs2853493	1.06E-02	8.06E-01	**8.09E-26**	**8.82E-17**	2.37E-04	1.40E-04
rs2853498	5.28E-02	4.82E-01	**3.64E-21**	**3.61E-14**	1.82E-03	4.60E-04
rs28359172	2.40E-02	**6.34E-14**	3.85E-03	2.66E-02	6.89E-01	5.12E-01
rs2853503	2.18E-01	4.71E-01	1.43E-01	1.17E-01	**5.56E-10**	1.75E-01
rs28359178	3.30E-02	**2.44E-14**	3.02E-03	1.70E-02	4.95E-01	7.45E-01
rs3088309	4.84E-01	**7.79E-11**	1.68E-03	2.48E-03	5.99E-01	6.95E-01

Table A.2.: Significantly associated expression probes with at least one mitochondrial SNP: Associations were calculated between all expression levels and all mitochondrial SNPs using a linear regression model adjusted for 25 Eigen-genes. The displayed SNPs were significantly associated with at least one expression probes. Significant associations with p-value $< 3.94 * 10^{-8}$ are shown in bold.

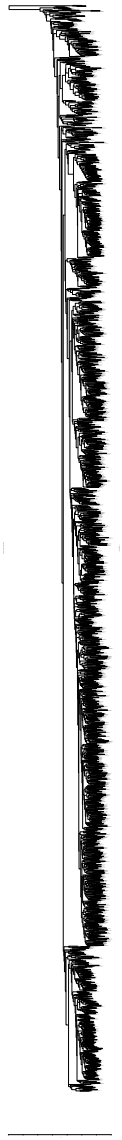

Figure A.1.: Cluster of all KORA S4, F4, and F4 OGTT samples having expression data: The expression levels of 2,509 samples were clustered using the agglomerative clustering implemented in the R package `cluster`. Three S4 samples clustered separately (at the top of the cluster) and were therefore called outliers and were removed for further analyses.

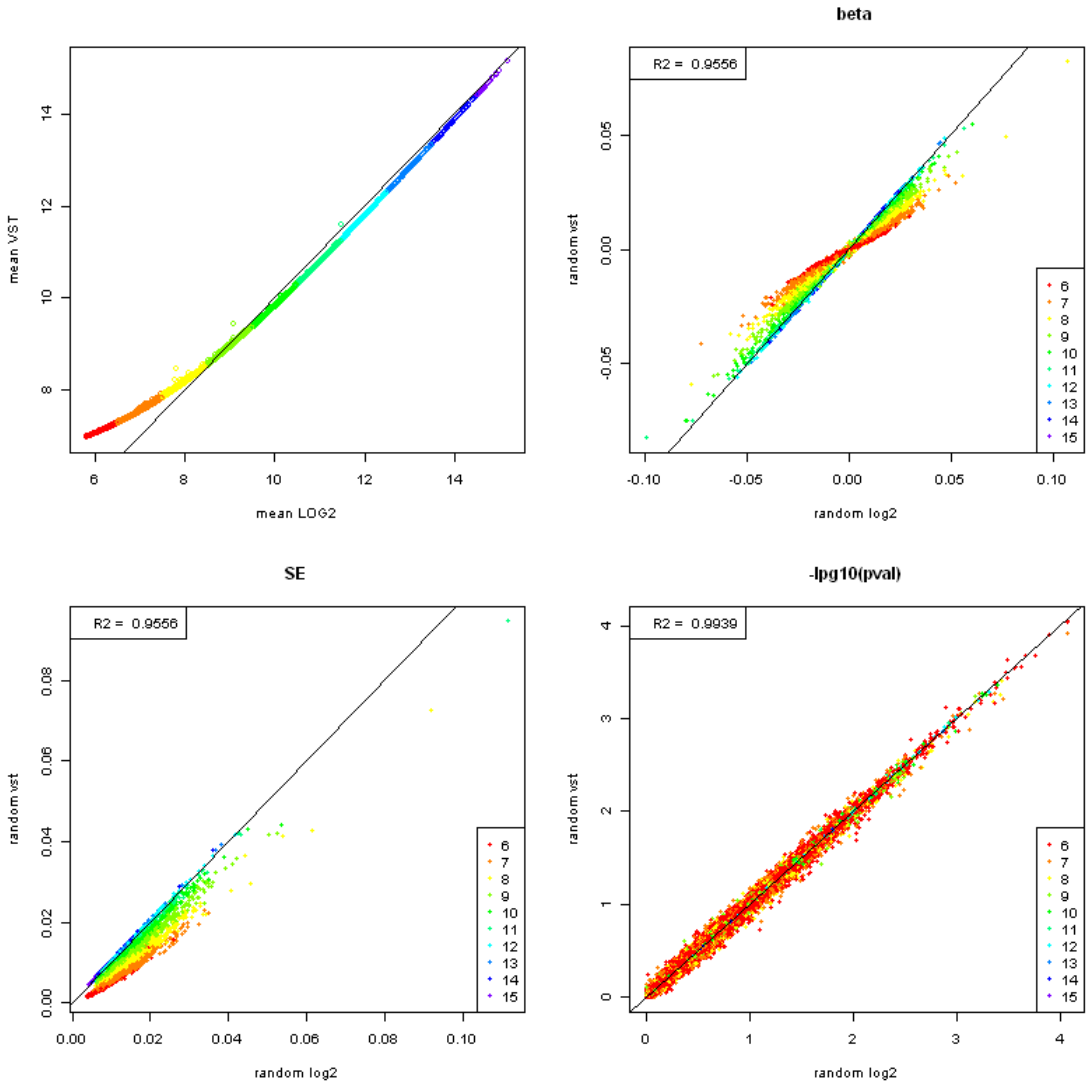

Figure A.2.: Comparison of VST and log2 transformation for a random phenotype in KORA F4:
The figures show the difference between the VST- and log2-transformed expression values in association with a random normal distributed phenotype. Each dot represents one probe and the color code is given in the legend of the plots. Differences are observed for the mean intensity values for low intensity values, for betas, and for standard errors of betas. Despite those differences the p-values are highly correlated (r^2=0.9939).

B. Declaration - Eidesstattliche Versicherung

Ich erkläre hiermit an Eides statt,
dass ich die vorliegende Dissertation mit dem Thema

Gene expression studies:
From case-control to multiple-population-based studies

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe. Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

Ort, Datum	Katharina Schramm

Bibliography

Frank W. Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nat Rev Genet*, 16(4):197–212, Apr 2015. doi: 10.1038/nrg3891. URL http://dx.doi.org/10.1038/nrg3891.

David Matthew Altshuler, Eric Steven Lander, Lauren Ambrogio, Toby Bloom, Kristian Cibulskis, Tim J Fennell, Stacey B Gabriel, David B Jaffe, Erica Shefler, Carrie L Sougnez, et al. A map of human genome variation from population scale sequencing. 2010. URL http://dash.harvard.edu/handle/1/5339502.

Nuno L Barbosa-Morais, Mark J Dunning, Shamith A Samarajiwa, Jeremy FJ Darot, Matthew E Ritchie, Andy G Lynch, and Simon Tavaré. A re-annotation pipeline for illumina beadarrays: improving the interpretation of gene expression data. *Nucleic acids research*, 38(3):e17–e17, 2010. URL http://nar.oxfordjournals.org/content/38/3/e17.short.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57:289–300, 1995.

Marina Bibikova, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M. Le, David Delano, Lu Zhang, Gary P. Schroth, Kevin L. Gunderson, Jian-Bing Fan, and Richard Shen. High density dna methylation array with single cpg site resolution. *Genomics*, 98:288–295, 2011.

E. M. Blackwood and J. T. Kadonaga. Going the distance: a current view of enhancer action. *Science*, 281(5373):60–63, Jul 1998.

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.

Marc J Bonder, Silva Kasela, Mart Kals, Riin Tamm, Kaie Lokk, Isabel Barragan, Wim A Buurman, Patrick Deelen, Jan-Willem Greve, Maxim Ivanov, et al. Genetic and epigenetic regulation of gene expression in fetal and adult human livers. *BMC genomics*, 15(1):860, 2014.

Todd M. Bull, Christopher D. Coldren, Mark Moore, Sylk M. Sotto-Santiago, David V. Pham, S Patrick Nana-Sinkam, Norbert F. Voelkel, and Mark W. Geraci. Gene microarray analysis of peripheral blood cells in pulmonary arterial hypertension. *Am J Respir Crit Care Med*, 170 (8):911–919, Oct 2004. doi: 10.1164/rccm.200312-1686OC. URL http://dx.doi.org/10.1164/rccm.200312-1686OC.

John C. Chambers, Weihua Zhang, Graham M. Lord, Pim van der Harst, Debbie A. Lawlor, Joban S. Sehmi, Daniel P. Gale, Mark N. Wass, Kourosh R. Ahmadi, Stephan J L. Bakker,

Jacqui Beckmann, Henk J G. Bilo, Murielle Bochud, Morris J. Brown, Mark J. Caulfield, John M C. Connell, H Terence Cook, Ioana Cotlarciuc, George Davey Smith, Ranil de Silva, Guohong Deng, Olivier Devuyst, Lambert D. Dikkeschei, Nada Dimkovic, Mark Dockrell, Anna Dominiczak, Shah Ebrahim, Thomas Eggermann, Martin Farrall, Luigi Ferrucci, Jurgen Floege, Nita G. Forouhi, Ron T. Gansevoort, Xijin Han, Bo Hedblad, Jaap J. Homan van der Heide, Bouke G. Hepkema, Maria Hernandez-Fuentes, Elina Hypponen, Toby Johnson, Paul E. de Jong, Nanne Kleefstra, Vasiliki Lagou, Marta Lapsley, Yun Li, Ruth J F. Loos, Jian'an Luan, Karin Luttropp, Céline Maréchal, Olle Melander, Patricia B. Munroe, Louise Nordfors, Afshin Parsa, Leena Peltonen, Brenda W. Penninx, Esperanza Perucha, Anneli Pouta, Inga Prokopenko, Paul J. Roderick, Aimo Ruokonen, Nilesh J. Samani, Serena Sanna, Martin Schalling, David Schlessinger, Georg Schlieper, Marc A J. Seelen, Alan R. Shuldiner, Marketa Sjögren, Johannes H. Smit, Harold Snieder, Nicole Soranzo, Timothy D. Spector, Peter Stenvinkel, Michael J E. Sternberg, Ramasamyiyer Swaminathan, Toshiko Tanaka, Lielith J. Ubink-Veltmaat, Manuela Uda, Peter Vollenweider, Chris Wallace, Dawn Waterworth, Klaus Zerres, Gerard Waeber, Nicholas J. Wareham, Patrick H. Maxwell, Mark I. McCarthy, Marjo-Riitta Jarvelin, Vincent Mooser, Goncalo R. Abecasis, Liz Lightstone, James Scott, Gerjan Navis, Paul Elliott, and Jaspal S. Kooner. Genetic loci influencing kidney function and chronic kidney disease. *Nat Genet*, 42(5):373–375, May 2010. doi: 10.1038/ng.566. URL `http://dx.doi.org/10.1038/ng.566`.

Lin S Chen and John D Storey. Eigen-r2 for dissecting variation in high-dimensional studies. *Bioinformatics*, 24(19):2260–2262, 2008. URL `http://bioinformatics.oxfordjournals.org/content/24/19/2260.short`.

Vivian G. Cheung, Laura K. Conlin, Teresa M. Weber, Melissa Arcaro, Kuang-Yu Jen, Michael Morley, and Richard S. Spielman. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet*, 33(3):422–425, Mar 2003. doi: 10.1038/ng1094. URL `http://dx.doi.org/10.1038/ng1094`.

Vivian G Cheung, Richard S Spielman, Kathryn G Ewens, Teresa M Weber, Michael Morley, and Joshua T Burdick. Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, 437(7063):1365–1369, Oct 2005. doi: 10.1038/nature04244. URL `http://dx.doi.org/10.1038/nature04244`.

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

William S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, Jun 2007. doi: 10.1038/nature05911. URL `http://dx.doi.org/10.1038/nature05911`.

Marilyn C. Cornelis and Frank B. Hu. Systems epidemiology: A new direction in nutrition and metabolic disease research. *Curr Nutr Rep*, 2(4), Dec 2013. doi: 10.1007/s13668-013-0052-4. URL `http://dx.doi.org/10.1007/s13668-013-0052-4`.

Zari Dastani, Marie-France Hivert, Nicholas Timpson, John R B. Perry, Xin Yuan, Robert A. Scott, Peter Henneman, Iris M. Heid, Jorge R. Kizer, Leo-Pekka Lyytikäinen, Christian

Fuchsberger, Toshiko Tanaka, Andrew P. Morris, Kerrin Small, Aaron Isaacs, Marian Beekman, Stefan Coassin, Kurt Lohman, Lu Qi, Stavroula Kanoni, James S. Pankow, Hae-Won Uh, Ying Wu, Aurelian Bidulescu, Laura J. Rasmussen-Torvik, Celia M T. Greenwood, Martin Ladouceur, Jonna Grimsby, Alisa K. Manning, Ching-Ti Liu, Jaspal Kooner, Vincent E. Mooser, Peter Vollenweider, Karen A. Kapur, John Chambers, Nicholas J. Wareham, Claudia Langenberg, Rune Frants, Ko Willems-Vandijk, Ben A. Oostra, Sara M. Willems, Claudia Lamina, Thomas W. Winkler, Bruce M. Psaty, Russell P. Tracy, Jennifer Brody, Ida Chen, Jorma Viikari, Mika Kähönen, Peter P. Pramstaller, David M. Evans, Beate St Pourcain, Naveed Sattar, Andrew R. Wood, Stefania Bandinelli, Olga D. Carlson, Josephine M. Egan, Stefan Böhringer, Diana van Heemst, Lyudmyla Kedenko, Kati Kristiansson, Marja-Liisa Nuotio, Britt-Marie Loo, Tamara Harris, Melissa Garcia, Alka Kanaya, Margot Haun, Norman Klopp, H-Erich Wichmann, Panos Deloukas, Efi Katsareli, David J. Couper, Bruce B. Duncan, Margreet Kloppenburg, Linda S. Adair, Judith B. Borja, D. I. A. G. R. A. M+ Consortium , M. A. G. I. C Consortium , G. L. G. C Investigators , MuT. H. E. R Consortium , James G. Wilson, Solomon Musani, Xiuqing Guo, Toby Johnson, Robert Semple, Tanya M. Teslovich, Matthew A. Allison, Susan Redline, Sarah G. Buxbaum, Karen L. Mohlke, Ingrid Meulenbelt, Christie M. Ballantyne, George V. Dedoussis, Frank B. Hu, Yongmei Liu, Bernhard Paulweber, Timothy D. Spector, P Eline Slagboom, Luigi Ferrucci, Antti Jula, Markus Perola, Olli Raitakari, Jose C. Florez, Veikko Salomaa, Johan G. Eriksson, Timothy M. Frayling, Andrew A. Hicks, Terho Lehtimäki, George Davey Smith, David S. Siscovick, Florian Kronenberg, Cornelia van Duijn, Ruth J F. Loos, Dawn M. Waterworth, James B. Meigs, Josee Dupuis, J Brent Richards, Benjamin F. Voight, Laura J. Scott, Valgerdur Steinthorsdottir, Christian Dina, Ryan P. Welch, Eleftheria Zeggini, Cornelia Huth, Yurii S. Aulchenko, Gudmar Thorleifsson, Laura J. McCulloch, Teresa Ferreira, Harald Grallert, Najaf Amin, Guanming Wu, Cristen J. Willer, Soumya Raychaudhuri, Steve A. McCarroll, Oliver M. Hofmann, Ayellet V. Segrè, Mandy van Hoek, Pau Navarro, Kristin Ardlie, Beverley Balkau, Rafn Benediktsson, Amanda J. Bennett, Roza Blagieva, Eric Boerwinkle, Lori L. Bonnycastle, Kristina Bengtsson Boström, Bert Bravenboer, Suzannah Bumpstead, Noël P. Burtt, Guillaume Charpentier, Peter S. Chines, Marilyn Cornelis, Gabe Crawford, Alex S F. Doney, Katherine S. Elliott, Amanda L. Elliott, Michael R. Erdos, Caroline S. Fox, Christopher S. Franklin, Martha Ganser, Christian Gieger, Niels Grarup, Todd Green, Simon Griffin, Christopher J. Groves, Candace Guiducci, Samy Hadjadj, Neelam Hassanali, Christian Herder, Bo Isomaa, Anne U. Jackson, Paul R V. Johnson, Torben Jørgensen, Wen H L. Kao, Augustine Kong, Peter Kraft, Johanna Kuusisto, Torsten Lauritzen, Man Li, Aloysius Lieverse, Cecilia M. Lindgren, Valeriya Lyssenko, Michel Marre, Thomas Meitinger, Kristian Midthjell, Mario A. Morken, Narisu Narisu, Peter Nilsson, Katharine R. Owen, Felicity Payne, Ann-Kristin Petersen, Carl Platou, Christine Proença, Inga Prokopenko, Wolfgang Rathmann, N William Rayner, Neil R. Robertson, Ghislain Rocheleau, Michael Roden, Michael J. Sampson, Richa Saxena, Beverley M. Shields, Peter Shrader, Gunnar Sigurdsson, Thomas Sparsø, Klaus Strassburger, Heather M. Stringham, Qi Sun, Amy J. Swift, Barbara Thorand, Jean Tichet, Tiinamaija Tuomi, Rob M. van Dam, Timon W. van Haeften, Thijs van Herpt, Jana V. van Vliet-Ostaptchouk, G Bragi Walters, Michael N. Weedon, Cisca Wijmenga, Jacqueline Witteman, Richard N. Bergman, Stephane Cauchi, Francis S. Collins, Anna L. Gloyn, Ulf Gyllensten, Torben Hansen, Winston A. Hide, Graham A. Hitman, Albert Hofman, David J. Hunter, Kristian Hveem, Markku Laakso, Andrew D. Morris, Colin N A. Palmer, Igor Rudan, Eric Sijbrands, Lincoln D. Stein, Jaakko Tuomilehto, Andre Uitterlinden, Mark

Walker, Richard M. Watanabe, Goncalo R. Abecasis, Bernhard O. Boehm, Harry Campbell, Mark J. Daly, Andrew T. Hattersley, Oluf Pedersen, Inês Barroso, Leif Groop, Rob Sladek, Unnur Thorsteinsdottir, James F. Wilson, Thomas Illig, Philippe Froguel, Cornelia M. van Duijn, Kari Stefansson, David Altshuler, Michael Boehnke, Mark I. McCarthy, Nicole Soranzo, Eleanor Wheeler, Nicole L. Glazer, Nabila Bouatia-Naji, Reedik Mägi, Joshua Randall, Paul Elliott, Denis Rybin, Abbas Dehghan, Jouke Jan Hottenga, Kijoung Song, Anuj Goel, Taina Lajunen, Alex Doney, Christine Cavalcanti-Proença, Meena Kumari, Nicholas J. Timpson, Carina Zabena, Erik Ingelsson, Ping An, Jeffrey O'Connell, Jian'an Luan, Amanda Elliott, Steven A. McCarroll, Rosa Maria Roccasecca, François Pattou, Praveen Sethupathy, Yavuz Ariyurek. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet*, 8(3):e1002607, 2012. doi: 10.1371/journal.pgen.1002607. URL `http://dx.doi.org/10.1371/journal.pgen.1002607`.

L M L. de Lau, P. J. Koudstaal, J C M. Witteman, A. Hofman, and M M B. Breteler. Dietary folate, vitamin b12, and vitamin b6 and the risk of parkinson disease. *Neurology*, 67(2):315–318, Jul 2006. doi: 10.1212/01.wnl.0000225050.57553.6d. URL `http://dx.doi.org/10.1212/01.wnl.0000225050.57553.6d`.

Anna L Dixon, Liming Liang, Miriam F Moffatt, Wei Chen, Simon Heath, Kenny C C Wong, Jenny Taylor, Edward Burnett, Ivo Gut, Martin Farrall, G. Mark Lathrop, Gonçalo R Abecasis, and William O C Cookson. A genome-wide association study of global gene expression. *Nat Genet*, 39(10):1202–1207, Oct 2007. doi: 10.1038/ng2109. URL `http://dx.doi.org/10.1038/ng2109`.

Timothy R Dreszer, Donna Karolchik, Ann S Zweig, Angie S Hinrichs, Brian J Raney, Robert M Kuhn, Laurence R Meyer, Mathew Wong, Cricket A Sloan, Kate R Rosenbloom, et al. The ucsc genome browser database: extensions and updates 2011. *Nucleic acids research*, 40(D1):D918–D923, 2012. URL `http://nar.oxfordjournals.org/content/40/D1/D918.short`.

Angela Döring, Christian Gieger, Divya Mehta, Henning Gohlke, Holger Prokisch, Stefan Coassin, Guido Fischer, Kathleen Henke, Norman Klopp, Florian Kronenberg, Bernhard Paulweber, Arne Pfeufer, Dieter Rosskopf, Henry Völzke, Thomas Illig, Thomas Meitinger, H-Erich Wichmann, and Christa Meisinger. SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. *Nat Genet*, 40(4):430–436, Apr 2008. doi: 10.1038/ng.107. URL `http://dx.doi.org/10.1038/ng.107`.

Pan Du, Warren A Kibbe, and Simon M Lin. lumi: a pipeline for processing illumina microarray. *Bioinformatics*, 24(13):1547–1548, Jul 2008. doi: 10.1093/bioinformatics/btn224. URL `http://dx.doi.org/10.1093/bioinformatics/btn224`.

Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren Kibbe, Lifang Hou, and Simon Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1):587, 2010.

Patrick C A. Dubois, Gosia Trynka, Lude Franke, Karen A. Hunt, Jihane Romanos, Alessandra Curtotti, Alexandra Zhernakova, Graham A R. Heap, Róza Adány, Arpo Aromaa, Maria Teresa Bardella, Leonard H. van den Berg, Nicholas A. Bockett, Emilio G. de la

Concha, Bárbara Dema, Rudolf S N. Fehrmann, Miguel Fernández-Arquero, Szilvia Fiatal, Elvira Grandone, Peter M. Green, Harry J M. Groen, Rhian Gwilliam, Roderick H J. Houwen, Sarah E. Hunt, Katri Kaukinen, Dermot Kelleher, Ilma Korponay-Szabo, Kalle Kurppa, Padraic MacMathuna, Markku Mäki, Maria Cristina Mazzilli, Owen T. McCann, M Luisa Mearin, Charles A. Mein, Muddassar M. Mirza, Vanisha Mistry, Barbara Mora, Katherine I. Morley, Chris J. Mulder, Joseph A. Murray, Concepción Núñez, Elvira Oosterom, Roel A. Ophoff, Isabel Polanco, Leena Peltonen, Mathieu Platteel, Anna Rybak, Veikko Salomaa, Joachim J. Schweizer, Maria Pia Sperandeo, Greetje J. Tack, Graham Turner, Jan H. Veldink, Wieke H M. Verbeek, Rinse K. Weersma, Victorien M. Wolters, Elena Urcelay, Bozena Cukrowska, Luigi Greco, Susan L. Neuhausen, Ross McManus, Donatella Barisani, Panos Deloukas, Jeffrey C. Barrett, Paivi Saavalainen, Cisca Wijmenga, and David A. van Heel. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet*, 42(4):295–302, Apr 2010. doi: 10.1038/ng.543. URL `http://dx.doi.org/10.1038/ng.543`.

Sandrine Dudoit, Yee Hwa Yang, Matthew J. Callow, and Terence P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12:111–139, 2002.

Georg B. Ehret, Patricia B. Munroe, Kenneth M. Rice, Murielle Bochud, Andrew D. Johnson, Daniel I. Chasman, Albert V. Smith, Martin D. Tobin, Germaine C. Verwoert, Shih-Jen Hwang, Vasyl Pihur, Peter Vollenweider, Paul F. O'Reilly, Najaf Amin, Jennifer L. Bragg-Gresham, Alexander Teumer, Nicole L. Glazer, Lenore Launer, Jing Hua Zhao, Yurii Aulchenko, Simon Heath, Siim Sõber, Afshin Parsa, Jian'an Luan, Pankaj Arora, Abbas Dehghan, Feng Zhang, Gavin Lucas, Andrew A. Hicks, Anne U. Jackson, John F. Peden, Toshiko Tanaka, Sarah H. Wild, Igor Rudan, Wilmar Igl, Yuri Milaneschi, Alex N. Parker, Cristiano Fava, John C. Chambers, Ervin R. Fox, Meena Kumari, Min Jin Go, Pim van der Harst, Wen Hong Linda Kao, Marketa Sjögren, D. G. Vinay, Myriam Alexander, Yasuharu Tabara, Sue Shaw-Hawkins, Peter H. Whincup, Yongmei Liu, Gang Shi, Johanna Kuusisto, Bamidele Tayo, Mark Seielstad, Xueling Sim, Khanh-Dung Hoang Nguyen, Terho Lehtimäki, Giuseppe Matullo, Ying Wu, Tom R. Gaunt, N Charlotte Onland-Moret, Matthew N. Cooper, Carl G P. Platou, Elin Org, Rebecca Hardy, Santosh Dahgam, Jutta Palmen, Veronique Vitart, Peter S. Braund, Tatiana Kuznetsova, Cuno S P M. Uiterwaal, Adebowale Adeyemo, Walter Palmas, Harry Campbell, Barbara Ludwig, Maciej Tomaszewski, Ioanna Tzoulaki, Nicholette D. Palmer, C. A. R. D. IoG. R. A. M consortium , C. K. D. Gen Consortium , KidneyGen Consortium , EchoGen consortium , C. H. A. R. G. E-H. F consortium , Thor Aspelund, Melissa Garcia, Yen-Pei C. Chang, Jeffrey R. O'Connell, Nanette I. Steinle, Diederick E. Grobbee, Dan E. Arking, Sharon L. Kardia, Alanna C. Morrison, Dena Hernandez, Samer Najjar, Wendy L. McArdle, David Hadley, Morris J. Brown, John M. Connell, Aroon D. Hingorani, Ian N M. Day, Debbie A. Lawlor, John P. Beilby, Robert W. Lawrence, Robert Clarke, Jemma C. Hopewell, Halit Ongen, Albert W. Dreisbach, Yali Li, J Hunter Young, Joshua C. Bis, Mika Kähönen, Jorma Viikari, Linda S. Adair, Nanette R. Lee, Ming-Huei Chen, Matthias Olden, Cristian Pattaro, Judith A Hoffman Bolton, Anna Köttgen, Sven Bergmann, Vincent Mooser, Nish Chaturvedi, Timothy M. Frayling, Muhammad Islam, Tazeen H. Jafar, Jeanette Erdmann, Smita R. Kulkarni, Stefan R. Bornstein, Jürgen Grässler, Leif Groop, Benjamin F. Voight, Johannes Kettunen, Philip Howard, Andrew Taylor, Simonetta Guarrera, Fulvio Ricceri,

Valur Emilsson, Andrew Plump, Inês Barroso, Kay-Tee Khaw, Alan B. Weder, Steven C. Hunt, Yan V. Sun, Richard N. Bergman, Francis S. Collins, Lori L. Bonnycastle, Laura J. Scott, Heather M. Stringham, Leena Peltonen, Markus Perola, Erkki Vartiainen, Stefan-Martin Brand, Jan A. Staessen, Thomas J. Wang, Paul R. Burton, Maria Soler Artigas, Yanbin Dong, Harold Snieder, Xiaoling Wang, Haidong Zhu, Kurt K. Lohman, Megan E. Rudock, Susan R. Heckbert, Nicholas L. Smith, Kerri L. Wiggins, Ayo Doumatey, Daniel Shriner, Gudrun Veldre, Margus Viigimaa, Sanjay Kinra, Dorairaj Prabhakaran, Vikal Tripathy, Carl D. Langefeld, Annika Rosengren, Dag S. Thelle, Anna Maria Corsi, Andrew Singleton, Terrence Forrester, Gina Hilton, Colin A. McKenzie, Tunde Salako, Naoharu Iwai, Yoshikuni Kita, Toshio Ogihara, Takayoshi Ohkubo, Tomonori Okamura, Hirotsugu Ueshima, Satoshi Umemura, Susana Eyheramendy, Thomas Meitinger, H-Erich Wichmann, Yoon Shin Cho, Hyung-Lae Kim, Jong-Young Lee, James Scott, Joban S. Sehmi, Weihua Zhang, Bo Hedblad, Peter Nilsson, George Davey Smith, Andrew Wong, Narisu Narisu, Alena Stančáková, Leslie J. Raffel, Jie Yao, Sekar Kathiresan, Christopher J. O'Donnell, Stephen M. Schwartz, M Arfan Ikram, WT Longstreth, Jr, Thomas H. Mosley, Sudha Seshadri, Nick R G. Shrine, Louise V. Wain, Mario A. Morken, Amy J. Swift, Jaana Laitinen, Inga Prokopenko, Paavo Zitting, Jackie A. Cooper, Steve E. Humphries, John Danesh, Asif Rasheed, Anuj Goel, Anders Hamsten, Hugh Watkins, Stephan J L. Bakker, Wiek H. van Gilst, Charles S. Janipalli, K Radha Mani, Chittaranjan S. Yajnik, Albert Hofman, Francesco U S. Mattace-Raso, Ben A. Oostra, Ayse Demirkan, Aaron Isaacs, Fernando Rivadeneira, Edward G. Lakatta, Marco Orru, Angelo Scuteri, Mika Ala-Korpela, Antti J. Kangas, Leo-Pekka Lyytikäinen, Pasi Soininen, Taru Tukiainen, Peter Würtz, Rick Twee-Hee Ong, Marcus Dörr, Heyo K. Kroemer, Uwe Völker, Henry Völzke, Pilar Galan, Serge Hercberg, Mark Lathrop, Diana Zelenika, Panos Deloukas, Massimo Mangino, Tim D. Spector, Guangju Zhai, James F. Meschia, Michael A. Nalls, Pankaj Sharma, Janos Terzic, M V Kranthi Kumar, Matthew Denniff, Ewa Zukowska-Szczechowska, Lynne E. Wagenknecht, F Gerald R. Fowkes, Fadi J. Charchar, Peter E H. Schwarz, Caroline Hayward, Xiuqing Guo, Charles Rotimi, Michiel L. Bots, Eva Brand, Nilesh J. Samani, Ozren Polasek, Philippa J. Talmud, Fredrik Nyberg, Diana Kuh, Maris Laan, Kristian Hveem, Lyle J. Palmer, Yvonne T. van der Schouw, Juan P. Casas, Karen L. Mohlke, Paolo Vineis, Olli Raitakari, . Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103–109, Oct 2011. doi: 10.1038/nature10405. URL `http://dx.doi.org/10.1038/nature10405`.

Matthias Elstner, Christopher M Morris, Katharina Heim, Peter Lichtner, Andreas Bender, Divya Mehta, Claudia Schulte, Manu Sharma, Gavin Hudson, Stefano Goldwurm, Alessandro Giovanetti, Massimo Zeviani, David J Burn, Ian G McKeith, Robert H Perry, E. Jaros, Rejko Krüger, H-Erich Wichmann, Stefan Schreiber, Harry Campbell, James F Wilson, Alan F Wright, Malcolm Dunlop, Giorgio Pistis, Daniela Toniolo, Patrick F Chinnery, Thomas Gasser, Thomas Klopstock, Thomas Meitinger, Holger Prokisch, and Douglass M Turnbull. Single-cell expression profiling of dopaminergic neurons combined with association analysis identifies pyridoxal kinase as parkinson's disease gene. *Ann Neurol*, 66(6): 792–798, Dec 2009. doi: 10.1002/ana.21780. URL `http://dx.doi.org/10.1002/ana.21780`.

Matthias Elstner, Christopher M Morris, Katharina Heim, Andreas Bender, Divya Mehta, Evelyn Jaros, Thomas Klopstock, Thomas Meitinger, Douglass M Turnbull, and Holger

Prokisch. Expression analysis of dopaminergic neurons in parkinson's disease and aging links transcriptional dysregulation of energy metabolism to cell death. *Acta Neuropathol*, 122(1):75–86, Jul 2011. doi: 10.1007/s00401-011-0828-9. URL http://dx.doi.org/10.1007/s00401-011-0828-9.

Valur Emilsson, Gudmar Thorleifsson, Bin Zhang, Amy S Leonardson, Florian Zink, Jun Zhu, Sonia Carlson, Agnar Helgason, G Bragi Walters, Steinunn Gunnarsdottir, et al. Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–428, 2008. URL http://www.nature.com/nature/journal/v452/n7186/abs/nature06758.html.

Ludwig Fahrmeir, Rita Künstler, Iris Pigeot, and Gerhard Tutz. *Statistik - Der Weg zur Datenanalyse*. Springer-Verlag, 2003.

Ludwig Fahrmeir, Thomas Kneib, and Stefan Lang. *Regression - Modelle and Methoden und Anwendungen*. Springer-Verlag, 2007.

Benjamin P Fairfax, Seiko Makino, Jayachandran Radhakrishnan, Katharine Plant, Stephen Leslie, Alexander Dilthey, Peter Ellis, Cordelia Langford, Fredrik O Vannberg, and Julian C Knight. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of hla alleles. *Nat Genet*, 44(5):502–510, May 2012. doi: 10.1038/ng.2205. URL http://dx.doi.org/10.1038/ng.2205.

Rudolf S N Fehrmann, Ritsert C Jansen, Jan H Veldink, Harm-Jan Westra, Danny Arends, Marc Jan Bonder, Jingyuan Fu, Patrick Deelen, Harry J M Groen, Asia Smolonska, Rinse K Weersma, Robert M W Hofstra, Wim A Buurman, Sander Rensen, Marcel G M Wolfs, Mathieu Platteel, Alexandra Zhernakova, Clara C Elbers, Eleanora M Festen, Gosia Trynka, Marten H Hofker, Christiaan G J Saris, Roel A Ophoff, Leonard H van den Berg, David A van Heel, Cisca Wijmenga, Gerard J Te Meerman, and Lude Franke. Trans-eqtls reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes and with a major role for the hla. *PLoS Genet*, 7(8):e1002197, Aug 2011. doi: 10.1371/journal.pgen.1002197. URL http://dx.doi.org/10.1371/journal.pgen.1002197.

Alice Gerrits, Yang Li, Bruno M. Tesson, Leonid V. Bystrykh, Ellen Weersing, Albertina Ausema, Bert Dontje, Xusheng Wang, Rainer Breitling, Ritsert C. Jansen, and Gerald de Haan. Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS Genet*, 5(10):e1000692, Oct 2009. doi: 10.1371/journal.pgen.1000692. URL http://dx.doi.org/10.1371/journal.pgen.1000692.

Christian Gieger, Ludwig Geistlinger, Elisabeth Altmaier, Martin Hrabé de Angelis, Florian Kronenberg, Thomas Meitinger, Hans-Werner Mewes, H-Erich Wichmann, Klaus M. Weinberger, Jerzy Adamski, Thomas Illig, and Karsten Suhre. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet*, 4(11):e1000282, Nov 2008. doi: 10.1371/journal.pgen.1000282. URL http://dx.doi.org/10.1371/journal.pgen.1000282.

Yoav Gilad, Scott A. Rifkin, and Jonathan K. Pritchard. Revealing the architecture of gene regulation: the promise of eqtl studies. *Trends Genet*, 24(8):408–415, Aug 2008. doi: 10.1016/j.tig.2008.06.001. URL http://dx.doi.org/10.1016/j.tig.2008.06.001.

Harald H H Göring, Joanne E Curran, Matthew P Johnson, Thomas D Dyer, Jac Charlesworth, Shelley A Cole, Jeremy B M Jowett, Lawrence J Abraham, David L Rainwater, Anthony G Comuzzie, Michael C Mahaney, Laura Almasy, Jean W MacCluer, Ahmed H Kissebah, Gregory R Collier, Eric K Moses, and John Blangero. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet*, 39(10):1208–1216, Oct 2007. doi: 10.1038/ng2119. URL `http://dx.doi.org/10.1038/ng2119`.

Daniel F. Gudbjartsson, G Bragi Walters, Gudmar Thorleifsson, Hreinn Stefansson, Bjarni V. Halldorsson, Pasha Zusmanovich, Patrick Sulem, Steinunn Thorlacius, Arnaldur Gylfason, Stacy Steinberg, Anna Helgadottir, Andres Ingason, Valgerdur Steinthorsdottir, Elinborg J. Olafsdottir, Gudridur H. Olafsdottir, Thorvaldur Jonsson, Knut Borch-Johnsen, Torben Hansen, Gitte Andersen, Torben Jorgensen, Oluf Pedersen, Katja K. Aben, J Alfred Witjes, Dorine W. Swinkels, Martin den Heijer, Barbara Franke, Andre L M. Verbeek, Diane M. Becker, Lisa R. Yanek, Lewis C. Becker, Laufey Tryggvadottir, Thorunn Rafnar, Jeffrey Gulcher, Lambertus A. Kiemeney, Augustine Kong, Unnur Thorsteinsdottir, and Kari Stefansson. Many sequence variants affecting diversity of adult human height. *Nat Genet*, 40(5):609–615, May 2008. doi: 10.1038/ng.122. URL `http://dx.doi.org/10.1038/ng.122`.

John Burdon Sanderson Haldane et al. New paths in genetics. *New Paths in Genetics.*, 1941.

Gregory Hannum, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, Rob Deconde, Menzies Chen, Indika Rajapakse, Stephen Friend, Trey Ideker, and Kang Zhang. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell*, 49(2):359–367, Jan 2013. doi: 10.1016/j.molcel.2012.10.016. URL `http://dx.doi.org/10.1016/j.molcel.2012.10.016`.

Ke Hao, Yohan Bossé, David C Nickle, Peter D Paré, Dirkje S Postma, Michel Laviolette, Andrew Sandford, Tillie L Hackett, Denise Daley, James C Hogg, W. Mark Elliott, Christian Couture, Maxime Lamontagne, Corry-Anke Brandsma, Maarten van den Berge, Gerard Koppelman, Alise S Reicin, Donald W Nicholson, Vladislav Malkov, Jonathan M Derry, Christine Suver, Jeffrey A Tsou, Amit Kulkarni, Chunsheng Zhang, Rupert Vessey, Greg J Opiteck, Sean P Curtis, Wim Timens, and Don D Sin. Lung eqtls to help reveal the molecular underpinnings of asthma. *PLoS Genet*, 8(11):e1003029, Nov 2012. doi: 10.1371/journal.pgen.1003029. URL `http://dx.doi.org/10.1371/journal.pgen.1003029`.

Alexander J. Hartemink, David K. Gifford, Tommi S. Jaakkola, and Richard A. Young. Maximum likelihood estimation of optimal scaling factors for expression array normalization. *SPIE BiOS*, 2001.

Monika B Hartig, Arcangela Iuso, Tobias Haack, Tomasz Kmiec, Elzbieta Jurkiewicz, Katharina Heim, Sigrun Roeber, Victoria Tarabin, Sabrina Dusi, Malgorzata Krajewska-Walasek, Sergiusz Jozwiak, Maja Hempel, Juliane Winkelmann, Matthias Elstner, Konrad Oexle, Thomas Klopstock, Wolfgang Mueller-Felber, Thomas Gasser, Claudia Trenkwalder, Valeria Tiranti, Hans Kretzschmar, Gerd Schmitz, Tim M Strom, Thomas Meitinger, and Holger Prokisch. Absence of an orphan mitochondrial protein and c19orf12 and causes a distinct clinical subtype of neurodegeneration with brain iron accumulation. *Am J Hum Genet*, 89

(4):543–550, Oct 2011. doi: 10.1016/j.ajhg.2011.09.007. URL http://dx.doi.org/10.1016/j.ajhg.2011.09.007.

Katharina Heim. Association analysis of ggenotype and gene expression in the kora population. Master's thesis, Ludwig-Maximilians-Universität München, 2008.

Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009. URL http://www.pnas.org/content/106/23/9362.short.

R. Holle, M.Happich, H.Löwel, and H.-E. Wichmann. KORA - a research platform for population based health research. *Gesundheitswesen*, 67:19–25, 2005.

F. C. Holstege and R. A. Young. Transcriptional regulation: contending with complexity. *Proc Natl Acad Sci U S A*, 96(1):2–4, Jan 1999.

Steve Horvath. Dna methylation age of human tissues and cell types. *Genome Biol*, 14(10):R115, 2013. doi: 10.1186/gb-2013-14-10-r115. URL http://dx.doi.org/10.1186/gb-2013-14-10-r115.

Bryan N. Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529, Jun 2009. doi: 10.1371/journal.pgen.1000529. URL http://dx.doi.org/10.1371/journal.pgen.1000529.

Tianxiao Huan, Tõnu Esko, Marjolein J. Peters, Luke C. Pilling, Katharina Schramm, Claudia Schurmann, Brian H. Chen, Chunyu Liu, Roby Joehanes, Andrew D. Johnson, Chen Yao, Sai-Xia Ying, Paul Courchesne, Lili Milani, Nalini Raghavachari, Richard Wang, Poching Liu, Eva Reinmaa, Abbas Dehghan, Albert Hofman, André G. Uitterlinden, Dena G. Hernandez, Stefania Bandinelli, Andrew Singleton, David Melzer, Andres Metspalu, Maren Carstensen, Harald Grallert, Christian Herder, Thomas Meitinger, Annette Peters, Michael Roden, Melanie Waldenberger, Marcus Dörr, Stephan B. Felix, Tanja Zeller, International Consortium for Blood Pressure G. W. A. S (I. C. B. P) , Ramachandran Vasan, Christopher J. O'Donnell, Peter J. Munson, Xia Yang, Holger Prokisch, Uwe Völker, Joyce B J. van Meurs, Luigi Ferrucci, and Daniel Levy. A meta-analysis of gene expression signatures of blood pressure and hypertension. *PLoS Genet*, 11(3):e1005035, Mar 2015. doi: 10.1371/journal.pgen.1005035. URL http://dx.doi.org/10.1371/journal.pgen.1005035.

Wolfgang Huber, Vincent J. Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S. Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, Raphael Gottardo, Florian Hahne, Kasper D. Hansen, Rafael A. Irizarry, Michael Lawrence, Michael I. Love, James MacDonald, Valerie Obenchain, Andrzej K. Ole?, Hervé Pagès, Alejandro Reyes, Paul Shannon, Gordon K. Smyth, Dan Tenenbaum, Levi Waldron, and Martin Morgan. Orchestrating high-throughput genomic analysis with bioconductor. *Nat Methods*, 12(2):115–121, Feb 2015. doi: 10.1038/nmeth.3252. URL http://dx.doi.org/10.1038/nmeth.3252.

Thomas Illig, Christian Gieger, Guangju Zhai, Werner Römisch-Margl, Rui Wang-Sattler, Cornelia Prehn, Elisabeth Altmaier, Gabi Kastenmüller, Bernet S. Kato, Hans-Werner Mewes, Thomas Meitinger, Martin Hrabé de Angelis, Florian Kronenberg, Nicole Soranzo, H-Erich Wichmann, Tim D. Spector, Jerzy Adamski, and Karsten Suhre. A genome-wide perspective of genetic variation in human metabolism. *Nat Genet*, 42(2):137–141, Feb 2010. doi: 10.1038/ng.507. URL `http://dx.doi.org/10.1038/ng.507`.

Illumina. Gene expression microarray data quality control. Technical report, Illumina and Inc., 2007.

Medea Imboden, Emmanuelle Bouzigon, Ivan Curjuric, Adaikalavan Ramasamy, Ashish Kumar, Dana B. Hancock, Jemma B. Wilk, Judith M. Vonk, Gian A. Thun, Valerie Siroux, Rachel Nadif, Florent Monier, Juan R. Gonzalez, Matthias Wjst, Joachim Heinrich, Laura R. Loehr, Nora Franceschini, Kari E. North, Janine Altmüller, Gerard H. Koppelman, Stefano Guerra, Florian Kronenberg, Mark Lathrop, Miriam F. Moffatt, George T. O'Connor, David P. Strachan, Dirkje S. Postma, Stephanie J. London, Christian Schindler, Manolis Kogevinas, Francine Kauffmann, Debbie L. Jarvis, Florence Demenais, and Nicole M. Probst-Hensch. Genome-wide association study of lung function decline in adults with and without asthma. *J Allergy Clin Immunol*, 129(5):1218–1228, May 2012. doi: 10.1016/j.jaci.2012.01.074. URL `http://dx.doi.org/10.1016/j.jaci.2012.01.074`.

Federico Innocenti, Gregory M Cooper, Ian B Stanaway, Eric R Gamazon, Joshua D Smith, Snezana Mirkov, Jacqueline Ramirez, Wanqing Liu, Yvonne S Lin, Cliona Moloney, Shelly Force Aldred, Nathan D Trinklein, Erin Schuetz, Deborah A Nickerson, Ken E Thummel, Mark J Rieder, Allan E Rettie, Mark J Ratain, Nancy J Cox, and Christopher D Brown. Identification and replication and and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet*, 7(5):e1002078, May 2011. doi: 10.1371/journal.pgen.1002078. URL `http://dx.doi.org/10.1371/journal.pgen.1002078`.

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.

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

Andrew D Johnson, Robert E Handsaker, Sara L Pulit, Marcia M Nizzari, Christopher J O'Donnell, and Paul IW de Bakker. Snap: a web-based tool for identification and annotation of proxy snps using hapmap. *Bioinformatics*, 24(24):2938–2939, 2008. URL `http://bioinformatics.oxfordjournals.org/content/24/24/2938.short`.

Peter A. Jones. Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*, 13(7):484–492, Jul 2012. doi: 10.1038/nrg3230. URL `http://dx.doi.org/10.1038/nrg3230`.

W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002. URL `http://genome.cshlp.org/content/12/6/996.short`.

Robert J. Klein, Caroline Zeiss, Emily Y. Chew, Jen-Yue Tsai, Richard S. Sackler, Chad Haynes, Alice K. Henning, John Paul SanGiovanni, Shrikant M. Mane, Susan T. Mayne, Michael B. Bracken, Frederick L. Ferris, Jurg Ott, Colin Barnstable, and Josephine Hoh. Complement factor h polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, Apr 2005. doi: 10.1126/science.1109557. URL `http://dx.doi.org/10.1126/science.1109557`.

Melvin T. Korkor, Fan Bo Meng, Shen Yang Xing, Mu Chun Zhang, Jin Rui Guo, Xiao Xue Zhu, and Ping Yang. Microarray analysis of differential gene expression profile in peripheral blood cells of patients with human essential hypertension. *Int J Med Sci*, 8(2):168–179, 2011.

Jan Krumsiek, Karsten Suhre, Thomas Illig, Jerzy Adamski, and Fabian J. Theis. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst Biol*, 5:21, 2011. doi: 10.1186/1752-0509-5-21. URL `http://dx.doi.org/10.1186/1752-0509-5-21`.

Anna Köttgen, Cristian Pattaro, Carsten A. Böger, Christian Fuchsberger, Matthias Olden, Nicole L. Glazer, Afshin Parsa, Xiaoyi Gao, Qiong Yang, Albert V. Smith, Jeffrey R. O'Connell, Man Li, Helena Schmidt, Toshiko Tanaka, Aaron Isaacs, Shamika Ketkar, Shih-Jen Hwang, Andrew D. Johnson, Abbas Dehghan, Alexander Teumer, Guillaume Paré, Elizabeth J. Atkinson, Tanja Zeller, Kurt Lohman, Marilyn C. Cornelis, Nicole M. Probst-Hensch, Florian Kronenberg, Anke Tönjes, Caroline Hayward, Thor Aspelund, Gudny Eiriksdottir, Lenore J. Launer, Tamara B. Harris, Evadnie Rampersaud, Braxton D. Mitchell, Dan E. Arking, Eric Boerwinkle, Maksim Struchalin, Margherita Cavalieri, Andrew Singleton, Francesco Giallauria, Jeffrey Metter, Ian H. de Boer, Talin Haritunians, Thomas Lumley, David Siscovick, Bruce M. Psaty, M Carola Zillikens, Ben A. Oostra, Mary Feitosa, Michael Province, Mariza de Andrade, Stephen T. Turner, Arne Schillert, Andreas Ziegler, Philipp S. Wild, Renate B. Schnabel, Sandra Wilde, Thomas F. Munzel, Tennille S. Leak, Thomas Illig, Norman Klopp, Christa Meisinger, H-Erich Wichmann, Wolfgang Koenig, Lina Zgaga, Tatijana Zemunik, Ivana Kolcic, Cosetta Minelli, Frank B. Hu, Asa Johansson, Wilmar Igl, Ghazal Zaboli, Sarah H. Wild, Alan F. Wright, Harry Campbell, David Ellinghaus, Stefan Schreiber, Yurii S. Aulchenko, Janine F. Felix, Fernando Rivadeneira, Andre G. Uitterlinden, Albert Hofman, Medea Imboden, Dorothea Nitsch, Anita Brandstätter, Barbara Kollerits, Lyudmyla Kedenko, Reedik Mägi, Michael Stumvoll, Peter Kovacs, Mladen Boban, Susan Campbell, Karlhans Endlich, Henry Völzke, Heyo K. Kroemer, Matthias Nauck, Uwe Völker, Ozren Polasek, Veronique Vitart, Sunita Badola, Alexander N. Parker, Paul M. Ridker, Sharon L R. Kardia, Stefan Blankenberg, Yongmei Liu, Gary C. Curhan, Andre Franke, Thierry Rochat, Bernhard Paulweber, Inga Prokopenko, Wei Wang, Vilmundur Gudnason, Alan R. Shuldiner, Josef Coresh, Reinhold Schmidt, Luigi Ferrucci, Michael G. Shlipak, Cornelia M. van Duijn, Ingrid Borecki, Bernhard K. Krämer, Igor Rudan, Ulf Gyllensten, James F. Wilson, Jacqueline C. Witteman, Peter P. Pramstaller, Rainer Rettig, Nick Hastie, Daniel I. Chasman, W. H. Kao, Iris M. Heid, and Caroline S. Fox. New loci associated with kidney function and chronic kidney disease. *Nat Genet*, 42(5):376–384, May 2010. doi: 10.1038/ng.568. URL `http://dx.doi.org/10.1038/ng.568`.

Amy S. Leonardson, Jun Zhu, Yanqing Chen, Kai Wang, John R. Lamb, Marc Reitman, Valur

Emilsson, and Eric E. Schadt. The effect of food intake on gene expression in human peripheral blood. *Hum Mol Genet*, 19(1):159–169, Jan 2010. doi: 10.1093/hmg/ddp476. URL http://dx.doi.org/10.1093/hmg/ddp476.

Daniel Levy, Georg B. Ehret, Kenneth Rice, Germaine C. Verwoert, Lenore J. Launer, Abbas Dehghan, Nicole L. Glazer, Alanna C. Morrison, Andrew D. Johnson, Thor Aspelund, Yurii Aulchenko, Thomas Lumley, Anna Köttgen, Ramachandran S. Vasan, Fernando Rivadeneira, Gudny Eiriksdottir, Xiuqing Guo, Dan E. Arking, Gary F. Mitchell, Francesco U S. Mattace-Raso, Albert V. Smith, Kent Taylor, Robert B. Scharpf, Shih-Jen Hwang, Eric J G. Sijbrands, Joshua Bis, Tamara B. Harris, Santhi K. Ganesh, Christopher J. O'Donnell, Albert Hofman, Jerome I. Rotter, Josef Coresh, Emelia J. Benjamin, André G. Uitterlinden, Gerardo Heiss, Caroline S. Fox, Jacqueline C M. Witteman, Eric Boerwinkle, Thomas J. Wang, Vilmundur Gudnason, Martin G. Larson, Aravinda Chakravarti, Bruce M. Psaty, and Cornelia M. van Duijn. Genome-wide association study of blood pressure and hypertension. *Nat Genet*, 41(6):677–687, Jun 2009. doi: 10.1038/ng.384. URL http://dx.doi.org/10.1038/ng.384.

Hong Li and Hongwen Deng. Systems genetics, bioinformatics and eqtl mapping. *Genetica*, 138(9-10):915–924, Oct 2010. doi: 10.1007/s10709-010-9480-x. URL http://dx.doi.org/10.1007/s10709-010-9480-x.

Choong-Chin Liew, Jun Ma, Hong-Chang Tang, Run Zheng, and Adam A. Dempsey. The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. *J Lab Clin Med*, 147(3):126–132, Mar 2006. doi: 10.1016/j.lab.2005.10.005. URL http://dx.doi.org/10.1016/j.lab.2005.10.005.

Simon M Lin, Pan Du, Wolfgang Huber, and Warren A Kibbe. Model-based variance-stabilizing transformation for illumina microarray data. *Nucleic Acids Res*, 36(2):e11, Feb 2008. doi: 10.1093/nar/gkm1075. URL http://dx.doi.org/10.1093/nar/gkm1075.

Glenn A. Maston, Sara K. Evans, and Michael R. Green. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, 7:29–59, 2006. doi: 10.1146/annurev.genom.7.080505.115623. URL http://dx.doi.org/10.1146/annurev.genom.7.080505.115623.

B. McClintock. The stability of broken ends of chromosomes in zea mays. *Genetics*, 26(2): 234–282, Mar 1941.

Divya Mehta. *Genome-wide association study to search for SNPs affecting gene expression in a general population*. PhD thesis, Technische Universität München, 2009.

Divya Mehta, Katharina Heim, Christian Herder, Maren Carstensen, Gertrud Eckstein, Claudia Schurmann, Georg Homuth, Matthias Nauck, Uwe Völker, Michael Roden, Thomas Illig, Christian Gieger, Thomas Meitinger, and Holger Prokisch. Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *Eur J Hum Genet*, Jun 2012. doi: 10.1038/ejhg.2012.106. URL http://dx.doi.org/10.1038/ejhg.2012.106.

Christa Meisinger, Holger Prokisch, Christian Gieger, Nicole Soranzo, Divya Mehta, Dieter Rosskopf, Peter Lichtner, Norman Klopp, Jonathan Stephens, Nicholas A Watkins, Panos Deloukas, Andreas Greinacher, Wolfgang Koenig, Matthias Nauck, Christian Rimmbach, Henry Völzke, Annette Peters, Thomas Illig, Willem H Ouwehand, Thomas Meitinger, H-Erich Wichmann, and Angela Döring. A genome-wide association study identifies three loci associated with mean platelet volume. *Am J Hum Genet*, 84(1):66–71, Jan 2009. doi: 10.1016/j.ajhg.2008.11.015. URL http://dx.doi.org/10.1016/j.ajhg.2008.11.015.

Andreas Menke, Monika Rex-Haffner, Torsten Klengel, Elisabeth B. Binder, and Divya Mehta. Peripheral blood gene expression: it all boils down to the rna collection tubes. *BMC Res Notes*, 5:1, 2012. doi: 10.1186/1756-0500-5-1. URL http://dx.doi.org/10.1186/1756-0500-5-1.

Jacob J. Michaelson, Salvatore Loguercio, and Andreas Beyer. Detection and interpretation of expression quantitative trait loci (eqtl). *Methods*, 48(3):265–276, Jul 2009. doi: 10.1016/j.ymeth.2009.03.004. URL http://dx.doi.org/10.1016/j.ymeth.2009.03.004.

O. J. Miller, W. Schnedl, J. Allen, and B. F. Erlanger. 5-methylcytosine localised in mammalian constitutive heterochromatin. *Nature*, 251(5476):636–637, Oct 1974.

C. Mueller, K. Schramm, C. Schurmann, S. Kwon, D. Lau A. Schillert and, A. Jagodzinski, C. Herder, G. Homuth, S. Wahl, H. Grallert, T. Illig, A. Peters, M. Dörr, X. Guo, W. Palmas, T. Meitinger, A. Teumer, M. Carstensen, P. S. Wild, H. Völzke, M. Roden, D. M. Herrington, U. Völker, A. Ziegler, Y. Liu, T. Zeller, S. Blankenberg, H. Prokisch, and S. B. Felix. Identification of blood pressure (bp) related genes by population-based transcriptome analyses and validation in a bp lowering clinical trial. In *Abstract for ASHG*, 2014.

O. Mueller, K. Hahnenberger, M. Dittmann, H. Yee, R. Dubrow, R. Nagle, and D. Ilsley. A microfluidic system for high-speed reproducible dna sizing and quantitation. *Electrophoresis*, 21(1):128–134, Jan 2000. doi: 3.0.CO;2-M. URL http://dx.doi.org/3.0.CO;2-M.

Amy Murphy, Jen-Hwa Chu, Mousheng Xu, Vincent J Carey, Ross Lazarus, Andy Liu, Stanley J Szefler, Robert Strunk, Karen DeMuth, Mario Castro, et al. Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood cd4+ lymphocytes. *Human molecular genetics*, 19(23):4745–4757, 2010. URL http://hmg.oxfordjournals.org/content/19/23/4745.short.

Steen Neldam, Margreet Lang, Russell Jones, and T. E. A. M. S. T. A-5 Investigators . Telmisartan and amlodipine single-pill combinations vs amlodipine monotherapy for superior blood pressure lowering and improved tolerability in patients with uncontrolled hypertension: results of the teamsta-5 study. *J Clin Hypertens (Greenwich)*, 13(7):459–466, Jul 2011. doi: 10.1111/j.1751-7176.2011.00468.x. URL http://dx.doi.org/10.1111/j.1751-7176.2011.00468.x.

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 J F. 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*, 41(6):666–676, Jun 2009. doi: 10.1038/ng.361. URL `http://dx.doi.org/10.1038/ng.361`.

Yuliya S. Nikolova and Ahmad R. Hariri. Can we observe epigenetic effects on human brain function? *Trends Cogn Sci*, 19(7):366–373, Jul 2015. doi: 10.1016/j.tics.2015.05.003. URL `http://dx.doi.org/10.1016/j.tics.2015.05.003`.

C. M. O'Connor and J. U. Adams. *Essentials of Cell Biology*. Cambridge, MA: NPG Education, 2010.

Len A. Pennacchio, Wendy Bickmore, Ann Dean, Marcelo A. Nobrega, and Gill Bejerano. Enhancers: five essential questions. *Nat Rev Genet*, 14(4):288–295, Apr 2013. doi: 10.1038/nrg3458. URL `http://dx.doi.org/10.1038/nrg3458`.

Marjolein J. Peters, Roby Joehanes, Luke C. Pilling, Claudia Schurmann, Karen N. Conneely, Joseph Powell, Eva Reinmaa, George L. Sutphin, Alexandra Zhernakova, Katharina Schramm, Yana A. Wilson, Sayuko Kobes, Taru Tukiainen, N. A. B. E. C/U. K. B. E. C Consortium , Yolande F. Ramos, Harald H H. Göring, Myriam Fornage, Yongmei Liu, Sina A. Gharib, Barbara E. Stranger, Philip L. De Jager, Abraham Aviv, Daniel Levy, Joanne M. Murabito, Peter J. Munson, Tianxiao Huan, Albert Hofman, André G. Uitterlinden, Fernando Rivadeneira, Jeroen van Rooij, Lisette Stolk, Linda Broer, Michael M P J. Verbiest,

Mila Jhamai, Pascal Arp, Andres Metspalu, Liina Tserel, Lili Milani, Nilesh J. Samani, Pärt Peterson, Silva Kasela, Veryan Codd, Annette Peters, Cavin K. Ward-Caviness, Christian Herder, Melanie Waldenberger, Michael Roden, Paula Singmann, Sonja Zeilinger, Thomas Illig, Georg Homuth, Hans-Jörgen Grabe, Henry Völzke, Leif Steil, Thomas Kocher, Anna Murray, David Melzer, Hanieh Yaghootkar, Stefania Bandinelli, Eric K. Moses, Jack W. Kent, Joanne E. Curran, Matthew P. Johnson, Sarah Williams-Blangero, Harm-Jan Westra, Allan F. McRae, Jennifer A. Smith, Sharon L R. Kardia, Iiris Hovatta, Markus Perola, Samuli Ripatti, Veikko Salomaa, Anjali K. Henders, Nicholas G. Martin, Alicia K. Smith, Divya Mehta, Elisabeth B. Binder, K Maria Nylocks, Elizabeth M. Kennedy, Torsten Klengel, Jingzhong Ding, Astrid M. Suchy-Dicey, Daniel A. Enquobahrie, Jennifer Brody, Jerome I. Rotter, Yii-Der I. Chen, Jeanine Houwing-Duistermaat, Margreet Kloppenburg, P Eline Slagboom, Quinta Helmer, Wouter den Hollander, Shannon Bean, Towfique Raj, Noman Bakhshi, Qiao Ping Wang, Lisa J. Oyston, Bruce M. Psaty, Russell P. Tracy, Grant W. Montgomery, Stephen T. Turner, John Blangero, Ingrid Meulenbelt, Kerry J. Ressler, Jian Yang, Lude Franke, Johannes Kettunen, Peter M. Visscher, G Gregory Neely, Ron Korstanje, Robert L. Hanson, Holger Prokisch, Luigi Ferrucci, Tonu Esko, Alexander Teumer, Joyce B J. van Meurs, and Andrew D. Johnson. The transcriptional landscape of age in human peripheral blood. *Nat Commun*, 6:8570, 2015. doi: 10.1038/ncomms9570. URL `http://dx.doi.org/10.1038/ncomms9570`.

Enrico Petretto, Jonathan Mangion, Nicholas J. Dickens, Stuart A. Cook, Mande K. Kumaran, Han Lu, Judith Fischer, Henrike Maatz, Vladimir Kren, Michal Pravenec, Norbert Hubner, and Timothy J. Aitman. Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet*, 2(10):e172, Oct 2006. doi: 10.1371/journal.pgen.0020172. URL `http://dx.doi.org/10.1371/journal.pgen.0020172`.

Anna Portela and Manel Esteller. Epigenetic modifications and human disease. *Nat Biotechnol*, 28(10):1057–1068, Oct 2010. doi: 10.1038/nbt.1685. URL `http://dx.doi.org/10.1038/nbt.1685`.

Jörg Rahnenführer. *Unsupervised learning methods for the analysis of microarray data*. Computational Biology and Applied Algorithmics and Max Planck Institute for Informatics, Saarbrücken and Germany, March 2004.

Adaikalavan Ramasamy, Adrian Mondry, Chris C. Holmes, and Douglas G. Altman. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med*, 5 (9):e184, Sep 2008. doi: 10.1371/journal.pmed.0050184. URL `http://dx.doi.org/10.1371/journal.pmed.0050184`.

Bruce Richardson. Impact of aging on dna methylation. *Ageing Res Rev*, 2(3):245–261, Jul 2003.

Mariana G. Rosca, Bernard Tandler, and Charles L. Hoppel. Mitochondria in cardiac hypertrophy and heart failure. *J Mol Cell Cardiol*, 55:31–41, Feb 2013. doi: 10.1016/j.yjmcc.2012.09.002. URL `http://dx.doi.org/10.1016/j.yjmcc.2012.09.002`.

Sherri Rose and Mark J van der Laan. Why match? investigating matched case-control study designs with causal effect estimation. *Int J Biostat*, 5(1):Article 1, 2009. doi: 10.2202/1557-4679.1127. URL `http://dx.doi.org/10.2202/1557-4679.1127`.

Daimei Sasayama, Hiroaki Hori, Seiji Nakamura, Ryo Miyata, Toshiya Teraishi, Kotaro Hattori, Miho Ota, Noriko Yamamoto, Teruhiko Higuchi, Naoji Amano, and Hiroshi Kunugi. Identification of single nucleotide polymorphisms regulating peripheral blood mrna expression with genome-wide significance: An eqtl study in the japanese population. *PLoS One*, 8(1):e54967, 2013. doi: 10.1371/journal.pone.0054967. URL http://dx.doi.org/10.1371/journal.pone.0054967.

Eric E Schadt, Cliona Molony, Eugene Chudin, Ke Hao, Xia Yang, Pek Y Lum, Andrew Kasarskis, Bin Zhang, Susanna Wang, Christine Suver, Jun Zhu, Joshua Millstein, Solveig Sieberts, John Lamb, Debraj GuhaThakurta, Jonathan Derry, John D Storey, Iliana Avila-Campillo, Mark J Kruger, Jason M Johnson, Carol A Rohl, Atila van Nas, Margarete Mehrabian, Thomas A Drake, Aldons J Lusis, Ryan C Smith, F. Peter Guengerich, Stephen C Strom, Erin Schuetz, Thomas H Rushmore, and Roger Ulrich. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol*, 6(5):e107, May 2008. doi: 10.1371/journal.pbio.0060107. URL http://dx.doi.org/10.1371/journal.pbio.0060107.

Katharina Schramm, Carola Marzi, Claudia Schurmann, Maren Carstensen, Eva Reinmaa, Reiner Biffar, Gertrud Eckstein, Christian Gieger, Hans-Jörgen Grabe, Georg Homuth, Gabriele Kastenmüller, Reedik Mägi, Andres Metspalu, Evelin Mihailov, Annette Peters, Astrid Petersmann, Michael Roden, Konstantin Strauch, Karsten Suhre, Alexander Teumer, Uwe Völker, Henry Völzke, Rui Wang-Sattler, Melanie Waldenberger, Thomas Meitinger, Thomas Illig, Christian Herder, Harald Grallert, and Holger Prokisch. Mapping the genetic architecture of gene regulation in whole blood. *PLoS One*, 9(4):e93844, 2014. doi: 10.1371/journal.pone.0093844. URL http://dx.doi.org/10.1371/journal.pone.0093844.

Andreas Schroeder, Odilo Mueller, Susanne Stocker, Ruediger Salowsky, Michael Leiber, Marcus Gassmann, Samar Lightfoot, Wolfram Menzel, Martin Granzow, and Thomas Ragg. The rin: an rna integrity number for assigning integrity values to rna measurements. *BMC Mol Biol*, 7:3, 2006. doi: 10.1186/1471-2199-7-3. URL http://dx.doi.org/10.1186/1471-2199-7-3.

Claudia Schurmann, Katharina Heim, Arne Schillert, Stefan Blankenberg, Maren Carstensen, Marcus Dörr, Karlhans Endlich, Stephan B. Felix, Christian Gieger, Harald Grallert, Christian Herder, Wolfgang Hoffmann, Georg Homuth, Thomas Illig, Jochen Kruppa, Thomas Meitinger, Christian Müller, Matthias Nauck, Annette Peters, Rainer Rettig, Michael Roden, Konstantin Strauch, Uwe Völker, Henry Völzke, Simone Wahl, Henri Wallaschofski, Philipp S. Wild, Tanja Zeller, Alexander Teumer, Holger Prokisch, and Andreas Ziegler. Analyzing illumina gene expression microarray data from different tissues: methodological aspects of data analysis in the metaxpress consortium. *PLoS One*, 7(12):e50938, 2012. doi: 10.1371/journal.pone.0050938. URL http://dx.doi.org/10.1371/journal.pone.0050938.

Michael E Sobel. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology*, 13(1982):290–312, 1982.

Nicole Soranzo, Tim D. Spector, Massimo Mangino, Brigitte Kühnel, Augusto Rendon, Alexander Teumer, Christina Willenborg, Benjamin Wright, Li Chen, Mingyao Li, Perttu

Salo, Benjamin F. Voight, Philippa Burns, Roman A. Laskowski, Yali Xue, Stephan Menzel, David Altshuler, John R. Bradley, Suzannah Bumpstead, Mary-Susan Burnett, Joseph Devaney, Angela Döring, Roberto Elosua, Stephen E. Epstein, Wendy Erber, Mario Falchi, Stephen F. Garner, Mohammed J R. Ghori, Alison H. Goodall, Rhian Gwilliam, Hakon H. Hakonarson, Alistair S. Hall, Naomi Hammond, Christian Hengstenberg, Thomas Illig, Inke R. König, Christopher W. Knouff, Ruth McPherson, Olle Melander, Vincent Mooser, Matthias Nauck, Markku S. Nieminen, Christopher J. O'Donnell, Leena Peltonen, Simon C. Potter, Holger Prokisch, Daniel J. Rader, Catherine M. Rice, Robert Roberts, Veikko Salomaa, Jennifer Sambrook, Stefan Schreiber, Heribert Schunkert, Stephen M. Schwartz, Jovana Serbanovic-Canic, Juha Sinisalo, David S. Siscovick, Klaus Stark, Ida Surakka, Jonathan Stephens, John R. Thompson, Uwe Völker, Henry Völzke, Nicholas A. Watkins, George A. Wells, H-Erich Wichmann, David A. Van Heel, Chris Tyler-Smith, Swee Lay Thein, Sekar Kathiresan, Markus Perola, Muredach P. Reilly, Alexandre F R. Stewart, Jeanette Erdmann, Nilesh J. Samani, Christa Meisinger, Andreas Greinacher, Panos Deloukas, Willem H. Ouwehand, and Christian Gieger. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the haemgen consortium. *Nat Genet*, 41(11):1182–1190, Nov 2009. doi: 10.1038/ng.467. URL `http://dx.doi.org/10.1038/ng.467`.

Stranger, Barbara E, Forrest, Matthew S, Clark, Andrew G, Minichiello, Mark J, Deutsch, Samuel, Lyle, Robert, Hunt, Sarah, Kahl, Brenda, Antonarakis, Stylianos E, Tavaré, Simon, Deloukas, Panagiotis, Dermitzakis, and Emmanouil T. Genome-wide associations of gene expression variation in humans. *PLoS Genetics*, 1:695–704, 2005.

Barbara E Stranger, Matthew S Forrest, Mark Dunning, Catherine E Ingle, Claude Beazley, Natalie Thorne, Richard Redon, Christine P Bird, Anna de Grassi, Charles Lee, Chris Tyler-Smith, Nigel Carter, Stephen W Scherer, Simon Tavaré, Panagiotis Deloukas, Matthew E Hurles, and Emmanouil T Dermitzakis. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853, Feb 2007a. doi: 10.1126/science.1136678. URL `http://dx.doi.org/10.1126/science.1136678`.

Barbara E Stranger, Alexandra C Nica, Matthew S Forrest, Antigone Dimas, Christine P Bird, Claude Beazley, Catherine E Ingle, Mark Dunning, Paul Flicek, Daphne Koller, Stephen Montgomery, Simon Tavaré, Panos Deloukas, and Emmanouil T Dermitzakis. Population genomics of human gene expression. *Nat Genet*, 39(10):1217–1224, Oct 2007b. doi: 10.1038/ng2142. URL `http://dx.doi.org/10.1038/ng2142`.

K. Struhl. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, 98(1):1–4, Jul 1999. doi: 10.1016/S0092-8674(00)80599-1. URL `http://dx.doi.org/10.1016/S0092-8674(00)80599-1`.

Karsten Suhre and Christian Gieger. Genetic variation in metabolic phenotypes: study designs and applications. *Nat Rev Genet*, 13(11):759–769, Nov 2012. doi: 10.1038/nrg3314. URL `http://dx.doi.org/10.1038/nrg3314`.

Karsten Suhre, So-Youn Shin, Ann-Kristin Petersen, Robert P. Mohney, David Meredith, Brigitte Wägele, Elisabeth Altmaier, C. A. R. D. IoG. R. A. M. , Panos Deloukas, Jeanette Erdmann, Elin Grundberg, Christopher J. Hammond, Martin Hrabé de Angelis, Gabi Kastenmüller, Anna Köttgen, Florian Kronenberg, Massimo Mangino, Christa Meisinger,

Thomas Meitinger, Hans-Werner Mewes, Michael V. Milburn, Cornelia Prehn, Johannes Raffler, Janina S. Ried, Werner Römisch-Margl, Nilesh J. Samani, Kerrin S. Small, H-Erich Wichmann, Guangju Zhai, Thomas Illig, Tim D. Spector, Jerzy Adamski, Nicole Soranzo, and Christian Gieger. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, 477(7362):54–60, Sep 2011. doi: 10.1038/nature10354. URL http://dx.doi.org/10.1038/nature10354.

Jae Hoon Sul, Towfique Raj, Simone de Jong, Paul I W. de Bakker, Soumya Raychaudhuri, Roel A. Ophoff, Barbara E. Stranger, Eleazar Eskin, and Buhm Han. Accurate and fast multiple-testing correction in eqtl studies. *Am J Hum Genet*, 96(6):857–868, Jun 2015. doi: 10.1016/j.ajhg.2015.04.012. URL http://dx.doi.org/10.1016/j.ajhg.2015.04.012.

David Sulzer. Multiple hit hypotheses for dopamine neuron loss in parkinson's disease. *Trends Neurosci*, 30(5):244–250, May 2007. doi: 10.1016/j.tins.2007.03.009. URL http://dx.doi.org/10.1016/j.tins.2007.03.009.

TIRCON. Website. URL http://tircon.eu/.

Trygve Tollefsbol. *Handbook of epigenetics: the new molecular and medical genetics*. Academic Press, London, 2010.

Stephen T. Turner. qqman: an r package for visualizing gwas results using q-q and manhattan plots. biorXiv DOI: 10.1101/005165, 2014. URL biorXivDOI:10.1101/005165.

Stephen T. Turner, Gary L. Schwartz, Arlene B. Chapman, Amber L. Beitelshees, John G. Gums, Rhonda M. Cooper-DeHoff, Eric Boerwinkle, Julie A. Johnson, and Kent R. Bailey. Plasma renin activity predicts blood pressure responses to beta-blocker and thiazide diuretic as monotherapy and add-on therapy for hypertension. *Am J Hypertens*, 23(9):1014–1022, Sep 2010. doi: 10.1038/ajh.2010.98. URL http://dx.doi.org/10.1038/ajh.2010.98.

Anja Victor. Bewertung genetischer Forschungsergebnisse - Methoden und vermeidbare Fehler. *G+G Wissenschaft*, 7:14–22, 2007.

Peter M. Visscher, Matthew A. Brown, Mark I. McCarthy, and Jian Yang. Five years of gwas discovery. *Am J Hum Genet*, 90(1):7–24, Jan 2012. doi: 10.1016/j.ajhg.2011.11.029. URL http://dx.doi.org/10.1016/j.ajhg.2011.11.029.

E. Volkin and L. Astrachan. Intracellular distribution of labeled ribonucleic acid after phage infection of escherichia coli. *Virology*, 2(4):433–437, Aug 1956.

Henry Völzke, Dietrich Alte, Carsten Oliver Schmidt, Dörte Radke, Roberto Lorbeer, Nele Friedrich, Nicole Aumann, Katharina Lau, Michael Piontek, Gabriele Born, et al. Cohort profile: the study of health in pomerania. *International journal of epidemiology*, page dyp394, 2010.

Louise V. Wain, Germaine C. Verwoert, Paul F. O'Reilly, Gang Shi, Toby Johnson, Andrew D. Johnson, Murielle Bochud, Kenneth M. Rice, Peter Henneman, Albert V. Smith, Georg B. Ehret, Najaf Amin, Martin G. Larson, Vincent Mooser, David Hadley, Marcus Dörr, Joshua C. Bis, Thor Aspelund, Tõnu Esko, A Cecile J W. Janssens, Jing Hua Zhao,

Simon Heath, Maris Laan, Jingyuan Fu, Giorgio Pistis, Jian'an Luan, Pankaj Arora, Gavin Lucas, Nicola Pirastu, Irene Pichler, Anne U. Jackson, Rebecca J. Webster, Feng Zhang, John F. Peden, Helena Schmidt, Toshiko Tanaka, Harry Campbell, Wilmar Igl, Yuri Milaneschi, Jouke-Jan Hottenga, Veronique Vitart, Daniel I. Chasman, Stella Trompet, Jennifer L. Bragg-Gresham, Behrooz Z. Alizadeh, John C. Chambers, Xiuqing Guo, Terho Lehtimäki, Brigitte Kühnel, Lorna M. Lopez, Ozren Polašek, Mladen Boban, Christopher P. Nelson, Alanna C. Morrison, Vasyl Pihur, Santhi K. Ganesh, Albert Hofman, Suman Kundu, Francesco U S. Mattace-Raso, Fernando Rivadeneira, Eric J G. Sijbrands, Andre G. Uitterlinden, Shih-Jen Hwang, Ramachandran S. Vasan, Thomas J. Wang, Sven Bergmann, Peter Vollenweider, Gérard Waeber, Jaana Laitinen, Anneli Pouta, Paavo Zitting, Wendy L. McArdle, Heyo K. Kroemer, Uwe Völker, Henry Völzke, Nicole L. Glazer, Kent D. Taylor, Tamara B. Harris, Helene Alavere, Toomas Haller, Aime Keis, Mari-Liis Tammesoo, Yurii Aulchenko, Inês Barroso, Kay-Tee Khaw, Pilar Galan, Serge Hercberg, Mark Lathrop, Susana Eyheramendy, Elin Org, Siim Sõber, Xiaowen Lu, Ilja M. Nolte, Brenda W. Penninx, Tanguy Corre, Corrado Masciullo, Cinzia Sala, Leif Groop, Benjamin F. Voight, Olle Melander, Christopher J. O'Donnell, Veikko Salomaa, Adamo Pio d'Adamo, Antonella Fabretto, Flavio Faletra, Sheila Ulivi, Fabiola M. Del Greco, Maurizio Facheris, Francis S. Collins, Richard N. Bergman, John P. Beilby, Joseph Hung, A William Musk, Massimo Mangino, So-Youn Shin, Nicole Soranzo, Hugh Watkins, Anuj Goel, Anders Hamsten, Pierre Gider, Marisa Loitfelder, Marion Zeginigg, Dena Hernandez, Samer S. Najjar, Pau Navarro, Sarah H. Wild, Anna Maria Corsi, Andrew Singleton, Eco J C. de Geus, Gonneke Willemsen, Alex N. Parker, Lynda M. Rose, Brendan Buckley, David Stott, Marco Orru, Manuela Uda, LifeLines Cohort Study , Melanie M. van der Klauw, Weihua Zhang, Xinzhong Li, James Scott, Yii-Der Ida Chen, Gregory L. Burke, Mika Kähönen, Jorma Viikari, Angela Döring, Thomas Meitinger, Gail Davies, John M. Starr, Valur Emilsson, Andrew Plump, Jan H. Lindeman, Peter A C 't Hoen, Inke R. König, EchoGen consortium , Janine F. Felix, Robert Clarke, Jemma C. Hopewell, Halit Ongen, Monique Breteler, Stéphanie Debette, Anita L. Destefano, Myriam Fornage, AortaGen Consortium , Gary F. Mitchell, C. H. A. R. G. E Consortium Heart Failure Working Group , Nicholas L. Smith, KidneyGen consortium , Hilma Holm, Kari Stefansson, Gudmar Thorleifsson, Unnur Thorsteinsdottir, C. K. D. Gen consortium , Cardiogenics consortium , CardioGram , Nilesh J. Samani, Michael Preuss, Igor Rudan, Caroline Hayward, Ian J. Deary, H-Erich Wichmann, Olli T. Raitakari, Walter Palmas, Jaspal S. Kooner, Ronald P. Stolk, J Wouter Jukema, Alan F. Wright, Dorret I. Boomsma, Stefania Bandinelli, Ulf B. Gyllensten, James F. Wilson, Luigi Ferrucci, Reinhold Schmidt, Martin Farrall, Tim D. Spector, Lyle J. Palmer, Jaakko Tuomilehto, Arne Pfeufer, Paolo Gasparini, David Siscovick, David Altshuler, Ruth J F. Loos, Daniela Toniolo, Harold Snieder, Christian Gieger, Pierre Meneton, Nicholas J. Wareham, Ben A. Oostra, Andres Metspalu, Lenore Launer, Rainer Rettig, David P. Strachan, Jacques S. Beckmann, Jacqueline C M. Witteman, Jeanette Erdmann, Ko Willems van Dijk, Eric Boerwinkle, Michael Boehnke, Paul M. Ridker, Marjo-Riitta Jarvelin, Aravinda Chakravarti, Goncalo R. Abecasis, Vilmundur Gudnason, Christopher Newton-Cheh, Daniel Levy, Patricia B. Munroe, Bruce M. Psaty, Mark J. Caulfield, Dabeeru C. Rao, Martin D. Tobin, Paul Elliott, and Cornelia M. van Duijn. Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. *Nat Genet*, 43(10):1005–1011, Oct 2011. doi: 10.1038/ng.922. URL `http://dx.doi.org/10.1038/ng.922`.

Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

J. D. WATSON and F. H. CRICK. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, Apr 1953.

Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1): D1001–D1006, 2014.

Harm-Jan Westra, Marjolein J. Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W. Christiansen, Benjamin P. Fairfax, Katharina Schramm, Joseph E. Powell, Alexandra Zhernakova, Daria V. Zhernakova, Jan H. Veldink, Leonard H. Van den Berg, Juha Karjalainen, Sebo Withoff, André G. Uitterlinden, Albert Hofman, Fernando Rivadeneira, Peter A C. 't Hoen, Eva Reinmaa, Krista Fischer, Mari Nelis, Lili Milani, David Melzer, Luigi Ferrucci, Andrew B. Singleton, Dena G. Hernandez, Michael A. Nalls, Georg Homuth, Matthias Nauck, Dörte Radke, Uwe Völker, Markus Perola, Veikko Salomaa, Jennifer Brody, Astrid Suchy-Dicey, Sina A. Gharib, Daniel A. Enquobahrie, Thomas Lumley, Grant W. Montgomery, Seiko Makino, Holger Prokisch, Christian Herder, Michael Roden, Harald Grallert, Thomas Meitinger, Konstantin Strauch, Yang Li, Ritsert C. Jansen, Peter M. Visscher, Julian C. Knight, Bruce M. Psaty, Samuli Ripatti, Alexander Teumer, Timothy M. Frayling, Andres Metspalu, Joyce B J. van Meurs, and Lude Franke. Systematic identification of trans eqtls as putative drivers of known disease associations. *Nat Genet*, 45(10):1238–1243, Oct 2013. doi: 10.1038/ng.2756. URL `http://dx.doi.org/10.1038/ng.2756`.

Janis E Wigginton, David J Cutler, and Goncalo R Abecasis. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet*, 76(5):887–893, May 2005. doi: 10.1086/429864. URL `http://dx.doi.org/10.1086/429864`.

Philipp S Wild, Christoph R Sinning, Alexander Roth, Sandra Wilde, Renate B Schnabel, Edith Lubos, Tanja Zeller, Till Keller, Karl J Lackner, Maria Blettner, Ramachandran S Vasan, Thomas Münzel, and Stefan Blankenberg. Distribution and categorization of left ventricular measurements in the general population: results from the population-based gutenberg heart study. *Circ Cardiovasc Imaging*, 3(5):604–613, Sep 2010. doi: 10.1161/CIRCIMAGING.109.911933. URL `http://dx.doi.org/10.1161/CIRCIMAGING.109.911933`.

Wrba, Dolznig, and Mannhalter. *Genetik verstehen*. UTB, 2007.

Fred A. Wright, Patrick F. Sullivan, Andrew I. Brooks, Fei Zou, Wei Sun, Kai Xia, Vered Madar, Rick Jansen, Wonil Chung, Yi-Hui Zhou, Abdel Abdellaoui, Sandra Batista, Casey Butler, Guanhua Chen, Ting-Huei Chen, David D'Ambrosio, Paul Gallins, Min Jin Ha, Jouke Jan Hottenga, Shunping Huang, Mathijs Kattenberg, Jaspreet Kochar, Christel M. Middeldorp, Ani Qu, Andrey Shabalin, Jay Tischfield, Laura Todd, Jung-Ying Tzeng, Gerard van Grootheest, Jacqueline M. Vink, Qi Wang, Wei Wang, Weibo Wang, Gonneke Willemsen, Johannes H. Smit, Eco J. de Geus, Zhaoyu Yin, Brenda W J H. Penninx, and Dorret I. Boomsma. Heritability and genomics of gene expression in peripheral blood.

Nat Genet, 46(5):430–437, May 2014. doi: 10.1038/ng.2951. URL http://dx.doi.org/10.1038/ng.2951.

Qinghua Xu, Shujuan Ni, Fei Wu, Fang Liu, Xun Ye, Bruno Mougin, Xia Meng, and Xiang Du. Investigation of variation in gene expression profiling of human blood by extended principle component analysis. *PLoS One*, 6(10):e26905, 2011. doi: 10.1371/journal.pone.0026905. URL http://dx.doi.org/10.1371/journal.pone.0026905.

Tanja Zeller, Philipp Wild, Silke Szymczak, Maxime Rotival, Arne Schillert, Raphaele Castagne, Seraya Maouche, Marine Germain, Karl Lackner, Heidi Rossmann, Medea Eleftheriadis, Christoph R Sinning, Renate B Schnabel, Edith Lubos, Detlev Mennerich, Werner Rust, Claire Perret, Carole Proust, Viviane Nicaud, Joseph Loscalzo, Norbert Hübner, David Tregouet, Thomas Münzel, Andreas Ziegler, Laurence Tiret, Stefan Blankenberg, and François Cambien. Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PLoS One*, 5(5):e10693, 2010. doi: 10.1371/journal.pone.0010693. URL http://dx.doi.org/10.1371/journal.pone.0010693.

Andreas Ziegler, Inke R. Konig, and Friedrich Pahlke. *A statistical approach to genetic epidemiology: concepts and applications*. Wiley-Blackwell, 2010.