A systematic analysis of the genetic influence on the transcriptome in human longevity

Dissertation

in fulfillment of the requirements for the degree “Dr. rer. nat.”

of the Faculty of Mathematics and Natural Sciences

at the Christian-Albrechts-University of Kiel

submitted by

Geetha Venkatesh

Kiel, 2014

First examiner:	Prof. Dr. Thomas Roeder
Second examiner:	Prof. Dr. Almut Nebel
Date of oral examination:	10.11.2014
Approved for print on:	10.11.2014
Approved by:	Prof. Dr. Wolfgang J. Duschl, Dean

“I am neither especially clever nor especially gifted. I am only very, very curious.”

(Albert Einstein)

To my mom and husband

Table of contents

Abbreviations

Abbreviation	Full form
ACE	A: additive genetics, C: common environment, E: unique environment
AIC	Akaike Information Criterion
BWA	Burrows-Wheeler Aligner
cDNA	complementary DNA
DESA	department of economic and social affairs
DNA	deoxyribonucleic acid
dsDNA	double stranded DNA
DZ	dizygotic
EDTA	ethylenediaminetetraacetic acid
eQTL	expression quantitative trait loci
FDR	false discovery rate
FPKM	Fragments Per Kilobase of exon per Million fragments mapped
G×A	genotype-age
GEMMA	Genome-wide Efficient Mixed Model Association
GO	gene ontology
GWAS	genome-wide association scan
IGF-1	insulin-like growth factor-1
IIS	insulin/ IGF-1 signalling (see IGF-1)
LLI	long-lived individuals
Mb	megabase
miRNA	microRNA
ml	milliliter
mRNA	messenger RNA
mtDNA	mitochondrial DNA
MZ	monozygotic
PCR	polymerase chain reaction
qPCR	quantitative PCR
qRT-PCR	quantitative reverse transcription PCR (see PCR)
RNA	ribonucleic acid
ROS	reactive oxygen species
SAM	sequence alignment/map format
SBS	sequencing-by-synthesis
SDS	Sequence Detection Systems
SEM	structural equation modeling
SNP	single nucleotide polymorphism

Abbreviation	Full form
ssDNA	single stranded DNA
TOR	target of rapamycin
UTR	untranslated region
UV	ultraviolet

List of tables

List of figures

1 Introduction

Over the last 200 years average life expectancy of humans has increased significantly. Especially since the beginning of the 20th century, it has increased from around 50 years to around 80-85 years in developed countries. Much of this dramatic rise in life expectancy can be ascribed to socio-economic developments as well as advancements in medical sciences which helped to conquer early and mid-life mortality during the first half of the 20th century. The further improvement in life expectancy was driven by a reduction in late-life mortality in the past 50 years (Christensen et al. 2006; Kirkwood 2008) (Figure 1-1).

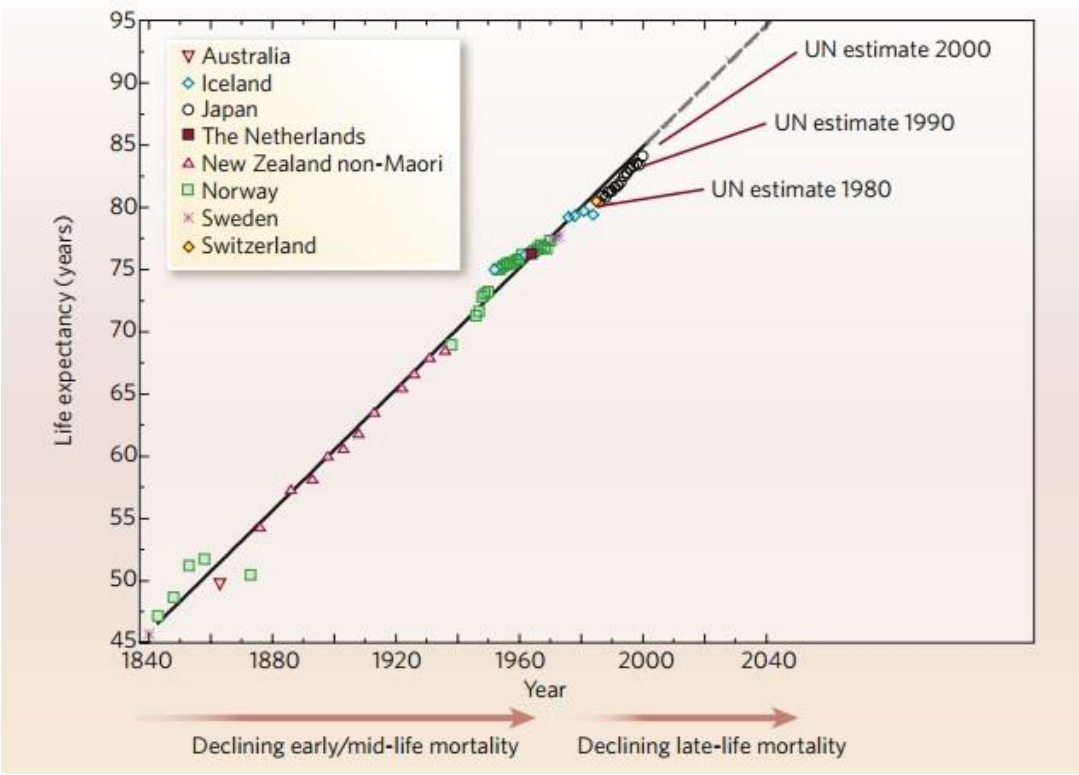

Figure 1-1 Steady increase in life expectancy over the last 200 years. The graph depicts the life expectancy in the then longest-living country. During the 19th and early 20th century, the increase in life expectancy was due to decline in early and mid-life mortality, which was followed by decline in late-life mortality in the latter half of the 20th century (Source: Kirkwood 2008).

Due to the unprecedented increase in life expectancy during the past century, the population aged 80 and over is growing faster than any younger section of the older population. In 2000, the 80+ segment of the population were around 69 million, and by 2050 people over 80 are projected to reach almost 379 million worldwide (i.e., 4.1% of the total human population) (Population division, DESA, United Nations). Since chronological age is the major risk factor for many diseases such as cardiovascular disease, type 2 diabetes, Alzheimer's and cancer, this change in the age structure of the population will have major impacts on society, economy, health care and the social system (Christensen et al. 2009). Consequently, a primary focus should be to decipher what contributes to healthy ageing. One key element of this approach is to understand the underlying molecular mechanisms as presented in this study, which will help to extend the health span of the elderly and reduce the risk of disabilities and diseases.

1.1 The phenotype of longevity

Although age is the major risk factor for many age-related diseases, there is a subset of elderly who do not follow the disease paradigm. They live to the 95th percentile of their respective birth cohort and are classified as long-lived individuals (LLI) (Gudmundsson et al. 2000). A majority of such long-lived people have been healthy and independent for most of their lives and experience a relatively rapid terminal decline (Hitt et al. 1999). They seem to attain exceptional ages not because they age more slowly but because they postpone ageing and reach advanced age in a better state of health (Vaupel 2010). A study analyzing the health of supercentenarians (110-119 years), semisupercentenarians (105-109 years), centenarians (100-104 years), nonagenarians (87-99 years) and younger controls (47-86 years) observed that the older the age group, the later the onset of diseases such as cancer, cardiovascular disease, dementia, and stroke as well as of cognitive and functional impairment (Andersen et al. 2012). The years spent with age-related diseases decrease with increasing age, which is a

phenomenon called compression of morbidity (Andersen et al. 2012) (Figure 1-2). Furthermore, it was shown that the LLI fall into three morbidity profiles – survivors (onset of at least one age-related disease before age 80), delayers (onset of at least one age-related disease between 80 and 99 years of age) and escapers (onset of at least one age-related disease after age 100) – and that 87% of male and 83% of female subjects delayed or escaped age-related diseases (Evert et al. 2003). Additionally, the frequency of survivors decreases and the frequency of escapers increases with age (Andersen et al. 2012) (Figure 1-3). Moreover, it has been observed that women live longer than men and make up a larger proportion of the older population. Although far fewer men reach the centenarian mark, they are usually healthier than their female counterparts (Franceschi et al. 2000b). Taken together, LLI represent a good model to study healthy longevity and to investigate key factors associated with it (Hitt et al. 1999; Franceschi and Bonafè 2003; Perls 2006; Engberg et al. 2009).

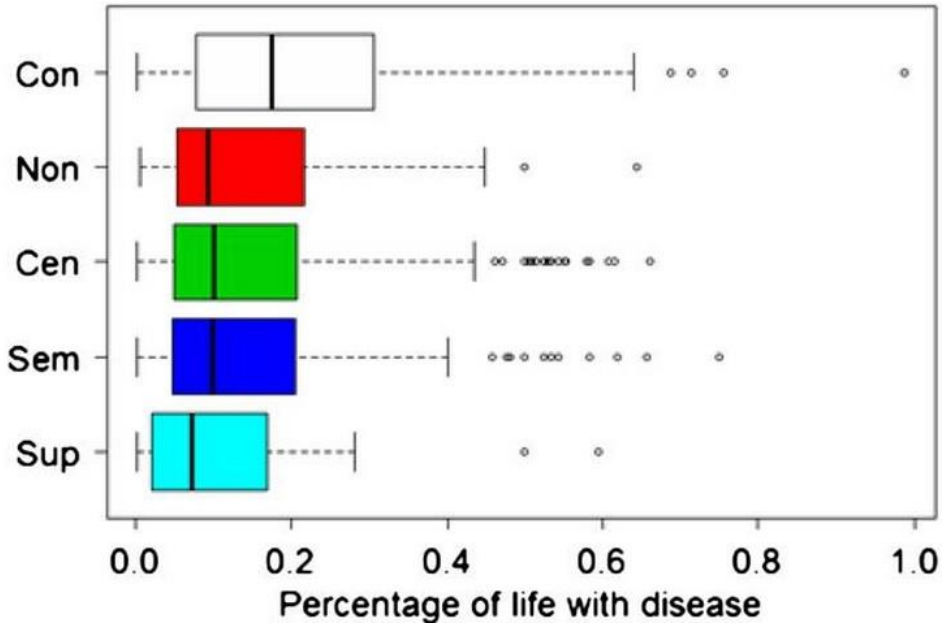

Figure 1-2 Compression of morbidity.The boxplot depicts the percentage of life spent with disease in different age groups; Con=controls, Non=nonagenarians, Cen=centenarians, Sem=semisupercentenarians, Sup=supercentenarians (Source: Andersen et al. 2012).

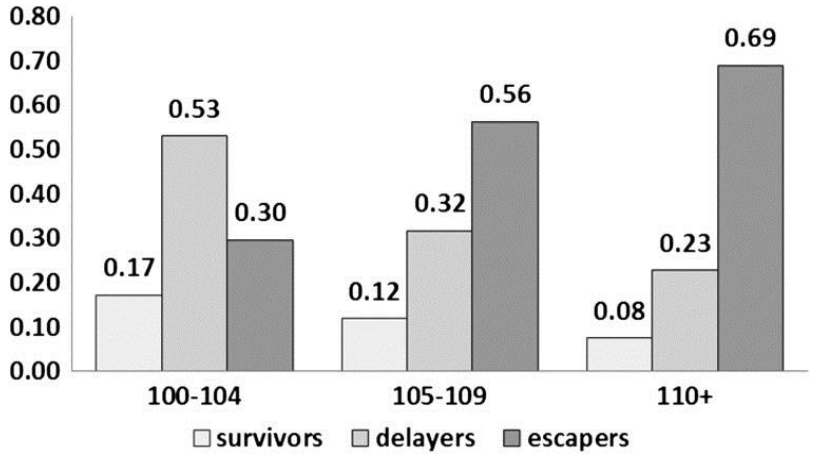

Figure 1-3 Frequency of survivors, delayers and escapers among three age groups namely, centenarians (100-104 years), semisupercentenarians (105-109 years) and supercentenarians (110+ years) (Source: Andersen et al. 2012).

1.2 Genetic epidemiology of human longevity

Family-based studies have provided useful information on the influence of heritable factors, indicating that longevity is a complex phenotype partly influenced by genetic determinants. In 1934, a positive correlation was found between the lifespans of LLI (90 years or older) and the lifespans of their parents and grandparents (Pearl 1934). In 1974, another study reported that children of long-lived parents tended to have long lifespan as well (Abbott et al. 1974). Furthermore, an increased relative probability of survival amongst siblings of nonagenarians (Westendorp et al. 2009) and of centenarians has been reported (Perls et al. 1998, 2002b; Willcox et al. 2006). In Iceland, first degree relatives of those living to the 95th percentile were twice as likely to also live to the 95th percentile as compared to a control population (Gudmundsson et al. 2000). In addition, male and female siblings of centenarians were 17 times and 8 times more likely to become centenarians themselves compared with the controls, respectively (Perls et al. 2002b). Besides an increased recurrence probability for surviving to extreme ages, the offspring of LLI have a lower prevalence of age-related diseases than age-matched controls (Terry et al. 2004; Westendorp et al. 2009). Additionally, they show beneficial

or youthful physiological characteristics for lipid and lipoprotein particle profiles, glucose metabolism and insulin sensitivity (Rozing et al. 2010; Slagboom et al. 2011; Deelen et al. 2013). All of these observations strongly argue that genetic factors play a significant role in attaining extreme ages. However, family-based studies cannot separate genetic effects from environmental effects based solely on phenotypic resemblance among family members. Twin studies, on the other hand, can distinguish between the two (Vogler 2001).

Longevity studies conducted on twins have consistently found that approximately 20-30% of the variation in adult lifespan is accounted for by variation in genetic factors (McGue et al. 1993; Herskind et al. 1996; Ljungquist et al. 1998). Combined analyses using more than 20,000 Scandinavian twins showed that the genetic influences on lifespan are minimal prior to age 60 and only increase thereafter (vB Hjelmborg et al. 2006), indicating that genetic variations, which have no impact in early life, are beneficial for older individuals. As the genetic influence on lifespan increases, it reaches approximately 40% in those individuals who survive beyond 85 years (Murabito et al. 2012). These findings encouraged the ongoing search for genes that affect longevity in humans at advanced ages (vB Hjelmborg et al. 2006).

1.3 Genetics of longevity

1.3.1 Insights from model organisms

Model organisms have significant similarity to humans with respect to genes, anatomy, and physiology (Danielle Simmons 2007). This similarity provided the opportunity to study longevity in model organisms such as *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and mice. Given their small size, short life span and the possibility to control environmental conditions and genotype, these model organisms are advantageous for studying genetic variants that are associated with

longevity (Christensen et al. 2006). Several studies in model organisms have provided significant insights into the type of genes that can influence lifespan. The first lifespan-extending mutation was found in the *C. elegans age-1* gene (Friedman and Johnson 1988). The mean lifespan of the *age-1* mutant was observed to be approximately 65% longer than that of the wild type (Friedman and Johnson 1988). Another significant lifespan-extending mutation was identified in the *C. elegans daf-2* gene, and it was shown that *daf-2* mutants live more than twice as long as the wild type (Kenyon et al. 1993). Both of these genes have mammalian homologues: *age-1* is homologous to the mammalian gene that encodes phosphatidylinositol-3-OH kinase catalytic subunits (PIK3C), which are positioned downstream of the insulin receptor (Morris et al. 1996), and *daf-2* is homologous to the genes encoding the insulin receptor and insulin-like growth factor-1 (IGF-1) receptor (Kimura et al. 1997). These two genes act as key players in the insulin/ IGF-1 signalling (IIS) pathway (Figure 1-4), which was the first pathway shown to influence lifespan in *C. elegans*. The major target of this pathway is *daf-16* that encodes the transcription factor DAF-16 (Forkhead box O (FOXO)), a protein that regulates the expression of other genes that mediate metabolic processes, stress resistance and antimicrobial response (Murphy et al. 2003). Reduced activity of DAF-2 signalling not only causes an increase in lifespan, but also induces metabolic changes (Kimura et al. 1997). Similar to *C. elegans*, reduced activity of the IIS pathway extends lifespan also in other model organisms such as *Drosophila* (Tatar et al. 2001) and mice (Holzenberger et al. 2003). Taken together, this indicates that IIS is an evolutionarily conserved regulator of longevity.

Another important pathway associated with lifespan extension is the target of rapamycin (TOR) pathway. TOR is a major amino acid and nutrient-sensing protein kinase that regulates growth and metabolism (Kapahi et al. 2010). Similar to the IIS pathway, reduced activity of the TOR pathway was shown to increase lifespan in *C. elegans* (Vellai et al. 2003), *Drosophila* (Kapahi et al. 2004), yeast (Kaeberlein et al. 2005) and mice (Harrison et al. 2009).

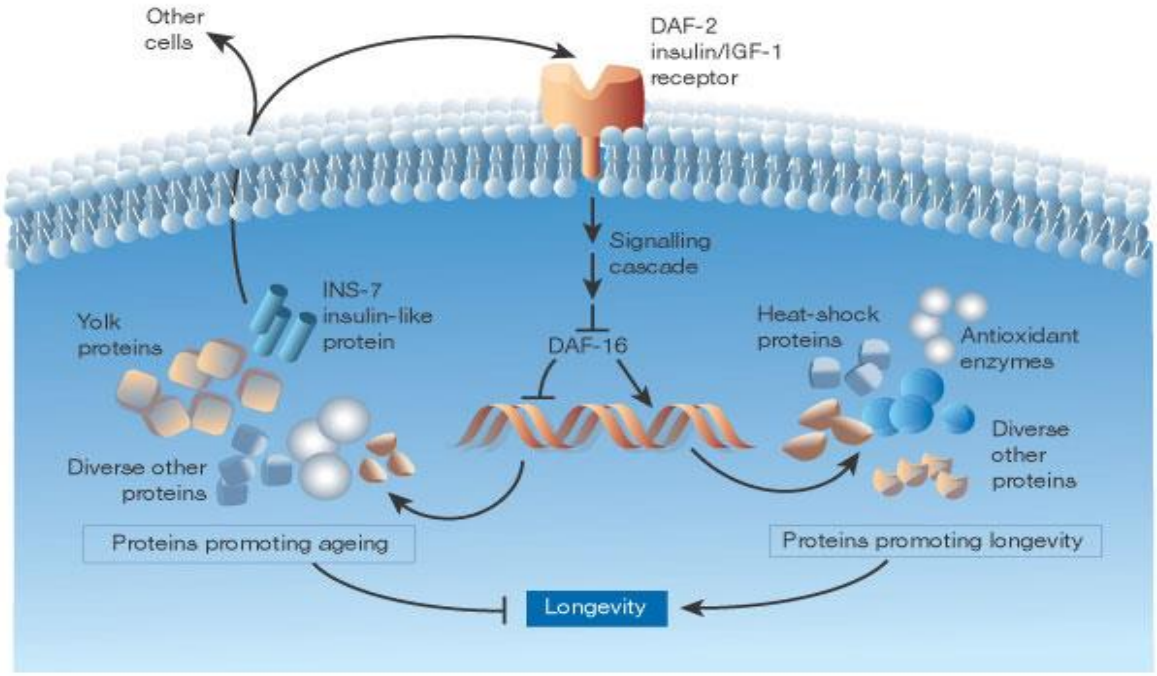

Figure 1-4 The IIS pathway found in *C. elegans*.The major components of the IIS pathway namely, daf-2, daf-16 and ins-7 have been found to regulate lifespan in *C. elegans*. Reduced activity of DAF-2 signalling, which induces lifespan extension in C. elegans, requires the activity of the transcription factor DAF-16. DAF-16 is known to inhibit expression of the DAF-2 agonist *ins-7*, and the reduced levels of INS-7 reduced DAF-2 activity which leads to lifespan extension (Source: Gems and McElwee 2003).

Apart from IIS and TOR pathway genes, a number of other longevity genes have been identified in model organisms. The pathways and associated genes contributing to longevity are available in databases such as the NetAge (Tacutu et al. 2010) and GenAge (de Magalhaes and Toussaint 2004).

Many genes that influence lifespan in model organisms have corresponding human homologues (Kuningas et al. 2008). The insights from model organisms have provided the motivation to elucidate whether similar longevity genes and mechanisms exist in humans.

1.3.2 Findings in humans

Family-based studies and studies on model organisms support the notion that human longevity is also partially influenced by genetic factors. Although substantial progress has already been made in model organisms to unravel the genetic basis of longevity, thus far, variations in only two genes namely, apolipoprotein E (*APOE*) and forkhead box O3A gene (*FOXO3A*) have been found to be consistently associated with longevity in humans. Variants in *APOE* (Schächter et al. 1994; Zhang et al. 1998; Blanché et al. 2001; Christensen et al. 2006) and *FOXO3A* (Willcox et al. 2008; Flachsbart et al. 2009; Li et al. 2009) were identified through candidate gene association studies. Genome-wide association scans (GWAS) of LLI versus younger controls had only identified *APOE* with genome-wide significance (Nebel et al. 2011; Deelen et al. 2011, 2014). A recent meta-GWAS has revealed a novel locus on chromosome 5q33.3, a region of unknown function, to confer survival beyond 90 years of age (Deelen et al. 2014).

The genetic variants in the *APOE* gene were the first to be consistently associated with human longevity (Schächter et al. 1994; Zhang et al. 1998; Blanché et al. 2001; Christensen et al. 2006). *APOE,* which plays an important role in regulating lipoproteins (Smith 2002; Christensen et al. 2006), has three common alleles namely, Ɛ2, Ɛ3 and Ɛ4, which code for the isoforms APOE2, APOE3 and APOE4 respectively (Schächter et al. 1994). These three isoforms of *APOE* interact differently with specific lipoprotein receptors, thus altering the levels of circulating cholesterols. *APOE* allele distributions vary significantly between LLI and young controls in different populations (Gerdes et al. 2000) (Figure 1-5). The Ɛ4 allele of *APOE*, which has repeatedly been associated with increased risk of cardiovascular disease (Kumar et al. 2012) and Alzheimer's disease (Liu et al. 2013), is significantly less frequent in LLI than in younger adults (Schächter et al. 1994; Garatachea et al. 2014). Therefore, it has been proposed that the *APOE* Ɛ4 is a mortality factor rather than a longevity conferring allele (Gerdes et al. 2000). This led to the hypothesis that individuals who attain extreme old age are likely to lack many of the disease-predisposing variations (Perls et al. 2002a; Perls and Terry 2003). However,

more recently, it has been shown that LLI carry the same number of known disease alleles as the general population (Beekman et al. 2010; Sebastiani and Perls 2012). Hence, it has been suggested that LLI may have genetic variations that confer protection against age-related disease risks (Bergman et al. 2007). Although many biologically plausible genes with positive associations with longevity have been described, further confirmation of their effects is pending (Christensen et al. 2006). One exception is *FOXO3A*, which is the second gene with variants that are repeatedly associated with longevity.

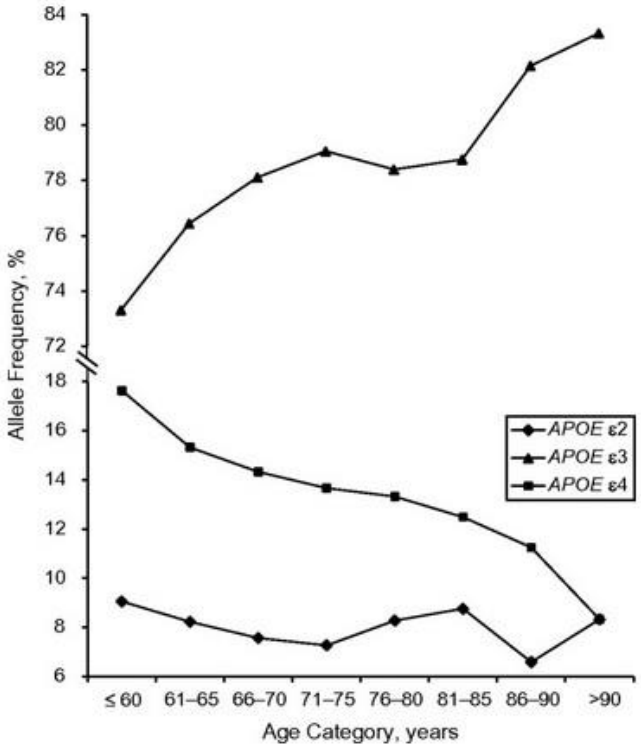

Figure 1-5 Frequencies of the three common *APOE* alleles, Ɛ2, Ɛ3 and Ɛ4 with respect to age (Source: McKay et al. 2011).

FOXO3A, a gene that encodes an evolutionarily conserved key transcription factor of the IIS pathway, is one of the homologues of *daf-16* in *C. elegans.* It is known that the daf-16

transcription factor significantly influences metabolism and lifespan in model organisms (Ogg et al. 1997; White 2003; Giannakou et al. 2004). These aspects suggested *FOXO3A* as a promising candidate for genetic studies of human longevity. In 2008, Willcox *et al.* (2008) first described a statistically significant association of variants in the *FOXO3A* gene with longevity in long-lived American men of Japanese ancestry. Later, several case-control studies confirmed the association in German (Flachsbart et al. 2009), Italian (Anselmi et al. 2009), Chinese (Li et al. 2009), American (European descent) (Pawlikowska et al. 2009) and Danish (Soerensen et al. 2010) populations. A recent meta-analysis has further validated the strong association between variants in the *FOXO3A* gene and longevity (Bao et al. 2014). In all these studies, the genetic variants associated with longevity are intronic and are not linked to known coding variants. So far, neither the functional variant in *FOXO3A* is known, nor are the mechanisms through which human longevity is influenced. A recent study proposed that coding variants of the *FOXO3A* may not be key players and intronic variants may be more important (Donlon et al. 2012).

Although the estimated genetic contribution to longevity is around 20-30%, variations in only two genes (*APOE* and *FOXO3A*) have been replicated in multiple populations. The effect sizes of these longevity-associated variants are small with odds ratios for survival to age 100 being 1.26 and 1.45 for *FOXO3A* and *APOE*, respectively (Wheeler and Kim 2011). In other words, these variants account for only a small portion of the estimated genetic contribution to human longevity. A large part of the reported heritability therefore remains to be explained.

Given the complexity of the interacting mechanisms and the involvement of possibly numerous genes with small effects (Christensen et al. 2006), it has been difficult to identify additional variants that influence longevity, and to replicate these results in independent samples. The genetic approaches in longevity research have not been as successful in identifying the genes underlying the essential mechanisms as in other complex phenotypes, for instance, inflammatory bowel disease (Franke et al. 2010).

Therefore, new avenues are needed to identify additional longevity-associated genes and the underlying molecular mechanisms. One approach has been to perform transcriptome studies in the field of human longevity.

1.4 Previous transcriptome studies in human longevity

The transcriptome of an individual reflects the influence of genetic variation as well as the response to the environment (Deelen et al. 2013). Investigating the transcriptomes of LLI could provide insights into the biological processes associated with longevity. The advent of the whole genome microarrays provided the opportunity to study whole human transcriptomes in a high-throughput approach. Since then, several explorative studies have been performed to identify the transcriptomic profiles marking human longevity. Transcriptional profiles of various tissues, including brain (Hong et al. 2008; Lu et al. 2004), eye (Segev et al. 2005), kidney (Rodwell et al. 2004), skin (Lener et al. 2006), lymphocytes (Hong et al. 2008), skeletal muscles (Zahn et al. 2006) and blood (Harries et al. 2011; Passtoors et al. 2012; Marttila et al. 2013) have been studied to investigate whether there are changes in gene expression levels with age. The expression of some genes was found not only to denote the chronological age but also reflect the function of the source organs and thus their biological age (Slagboom et al. 2011; Passtoors et al. 2012). The genes that were differentially expressed with chronological age in different tissues have been associated with biological terms like energy metabolism, immune response, cell cycle/cell growth, signal transduction, mRNA processing, and mitochondrial electron transport chain (Rodwell et al. 2004; Zahn et al. 2006; Hong et al. 2008; Harries et al. 2011; Passtoors et al. 2012; Marttila et al. 2013).

Most of the transcriptome studies performed in longevity research use cross-sectional designs where LLI are compared to young adults or groups of increasing age are examined. As pointed out by a hallmark review in the research field of longevity, the main problem in cross-sectional designs is a lack of appropriate control groups, as

cohort-specific characteristics might confound comparisons between LLI and younger controls (Christensen et al. 2006). Controls used in these studies originate from other birth cohorts than the LLI and thus the differences in two samples might confound the observations. Moreover, with cross-sectional designs, it is not possible to distinguish which changes in gene expression are causal and which are consequential to longevity (Passtoors et al. 2008). To address these constraints, a few studies have been performed using a family-based design, in which LLI are compared to their middle-aged offspring, and the middle-aged offspring are further compared to similar-aged controls from the general population (Passtoors et al. 2012). Another approach is to use longitudinal study designs, where individuals from a specific birth cohort are followed over time. Although longitudinal designs are less prone to biases that are associated with the selection of controls, it is difficult to perform such studies due to logistical challenges (Christensen et al. 2006). As of now, prospective designs have been used in large epidemiological studies but not in transcriptomics.

Besides study design problems, limited sample sizes and limited availability of relevant tissue are the major problems in longevity research. Only few large transcriptome studies have been performed to elucidate molecular patterns associated with longevity (Harries et al. 2011; Passtoors et al. 2012; Marttila et al. 2013). The majority of transcriptome studies were performed using whole blood, as blood is one of the easiest accessible biomaterials. Since blood is in contact with all tissues, it may also reflect in part the physical health of the whole body (Passtoors et al. 2012). In addition, whole blood shares significant similarities with other tissues on a transcriptome level (Sullivan et al. 2006). Therefore, studying gene expression in blood is both biologically meaningful and practical in the field of human longevity. However, while performing transcriptome studies using whole blood, different blood cell populations must be considered. The differences in blood cell counts between LLI and younger controls may confound observed changes in gene expression (Deelen et al. 2013). Thus, differential gene

expression observed in whole blood transcriptome studies should be adjusted for blood cell subsets.

1.5 RNA sequencing in longevity research

Many technologies have been developed over the years to quantify gene expression, including Northern blotting, quantitative reverse transcription polymerase chain reaction (qRT-PCR), hybridization- and sequencing-based approaches. Among these, microarrays and the recent RNA-sequencing (RNA-seq) are capable of analyzing thousands of genes simultaneously.

Microarrays contain thousands of different oligonucleotides - short single stranded DNA molecules, which are immobilized on to the solid support (glass slide, silicon chips or nylon membrane). The principle behind microarray technology is hybridization between DNA probes on the array and labeled RNA molecules derived from the sample of interest. The amount of mRNA hybridized to each probe indicates the corresponding gene expression, which can be estimated from the hybridization signal (Tarca et al. 2006).

Although microarrays are the most popular approach for gene expression studies and are relatively inexpensive, they pose several limitations. For instance, the accuracy of expression measurements suffer from background noise, especially for transcripts present in low abundance. Furthermore, probes differ considerably in their hybridization properties. Thus, hybridization results from a single sample may not provide a reliable measure of the relative expression of different transcripts. Finally, usage of microarray technology limits the interrogation to only those genes for which probes are designed (Marioni et al. 2008).

In contrast to microarray technology, RNA-seq involves direct sequencing of transcripts using high-throughput sequencing technologies. The expression of each transcript is measured by the number of sequenced fragments that align to the transcript, which is expected to correlate with its abundance level (Rapaport et al. 2013). RNA-seq has been used to measure gene expression levels and has the potential to overcome the limitations posed by microarrays. Unlike microarrays, RNA-seq is not limited to investigating transcripts that correspond to existing genomic sequence but can also be applied to non-model organisms whose genomic sequences are yet to be determined (Wang et al. 2009). Besides, RNA-seq permits examining novel transcripts (Hutchins et al. 2012), allele-specific expression (Degner et al. 2009), and alternative splicing (Pan et al. 2008). Furthermore, RNA-seq has very low background noise because DNA sequences can be unambiguously mapped to unique regions of the genome. Additionally, RNA-seq does not have an upper limit for quantification and thus has a large dynamic range of expression levels. Finally, RNA-seq has high technical and biological reproducibility (Wang et al. 2009). Therefore, RNA-seq is becoming a powerful technology for transcriptomic studies.

Since RNA-seq can provide more accurate and reliable genome-wide gene expression data than any alternative method, it has been used to study gene expression in several diseases, including Alzheimer's (Twine et al. 2011), schizophrenia (Fillman et al. 2012; Hwang et al. 2013; Sainz et al. 2013), renal cell carcinoma (Xiong et al. 2014), lung cancer (Cheng et al. 2012) and immune-mediated diseases (Zhang et al. 2012). But RNA-seq based transcriptome profiling has not been performed in longevity research thus far. This study provides the first comprehensive insights into the transcriptome of LLI using RNA-seq.

1.6 Objectives of the study

The overall aim of this investigation is to gain knowledge about the essential mechanisms that let some people survive to an exceptional age in good health. The information obtained from studies as this one represent a starting point for approaches aiming to extend the health span of the elderly and thus may ease the pressure on healthcare systems. The specific objectives of this thesis are:

- To identify the molecular patterns associated with human longevity by comparing the transcriptomes of LLI with those of young controls.
- To investigate the heritability of transcriptional activity by using monozygotic and dizygotic twins.
- To examine the influence of natural genetic variability on gene expression by performing expression quantitative trait loci (eQTL) analysis, which provides an opportunity 1) to determine the primary (genetic) basis of the regulation of gene expression and 2) to gain insights into otherwise unknown genotype-phenotype correlations.
- To investigate the genotype-age interaction effects on the genes that are differentially expressed between LLI and young controls.

To achieve the above mentioned objectives, a large number of probands were recruited: 55 LLI and 73 young controls from Germany along with 48 long-lived twins, 48 middle-aged twins, 10 unrelated LLI and 10 unrelated middle-aged individuals from Denmark. The transcriptomes from whole blood of all the individuals were sequenced with a state-of-the-art technology, RNA-seq. The obtained data was subsequently analyzed using a pipeline, which was developed during the course of this study, to accomplish the specific goals.

2 Materials and Methods

The work contained in this thesis was performed according to the workflow depicted in Figure 2-1. German samples (see Table 2-1) used in this study were recruited by the PopGen Biobank in Kiel and Danish samples were recruited by our Danish collaborators. RNA and DNA extraction, RNA-sequencing and genotyping were performed on the platforms at the Institute of Clinical Molecular Biology, Kiel. Each step is described in detail in the following sections.

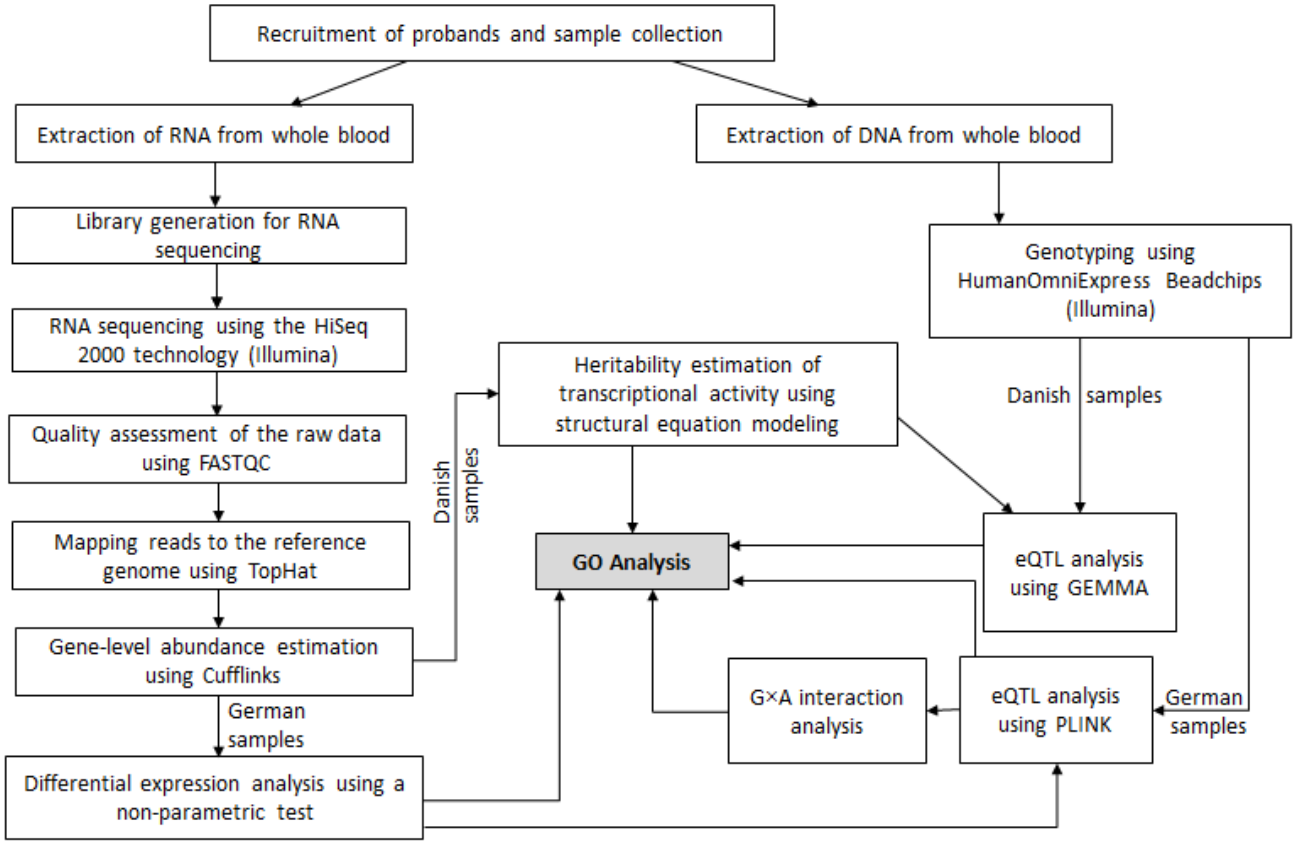

Figure 2-1 Analysis pipeline/workflow used in the present study

2.1 Study participants and sample collection

An overview of the samples used in this study is provided in Table 2-1. All the German subjects were recruited between October 2009 and November 2011 from different regions of Germany. The recruitment was carried out as previously described (Nebel et al. 2005). In brief, the contact with the participants was established by obtaining their names and addresses from local registry offices. At first, a letter explaining the study and seeking permission to send a questionnaire and a blood sampling kit was sent out to all potential participants. Subjects were advised to fill out the questionnaire and take a blood sample (20-ml) with the help of their general practitioner. A complete summary of the socioeconomic condition, quality of life, and health status of the subjects was recorded from the questionnaires. Primary criteria for inclusion in the study population were ability to perform daily activities on their own, good health and absence of overt severe diseases. All probands signed a written informed consent to participate in the study. Approval for the study was received from the Ethics Committee of the Christian-Albrechts-University, and also from local data protection authorities.

All the Danish samples were recruited by our collaborators at the University of Southern Denmark, Odense and sent to us. The approval to use these samples in this study was received from the Ethics Committee in Odense, Denmark.

Table 2-1 Overview of the study participants

Study groups (Country)	Age range in years	No. of females / No. of males (total number)	Total (N)
Young individuals (Germany)	20-55	45/28 (73)	244
Long-lived individuals (Germany)	91-104	40/15 (55)	
Middle-aged twins (Denmark)	58-60	32/16 (48)	
Long-lived twins (Denmark)	83-92	32/16 (48)	
Middle-aged unrelated individuals (Denmark)	58-60	4/6 (10)	
Long-lived unrelated individuals (Denmark)	83-92	8/2 (10)	

2.2 Extraction of RNA and DNA from blood samples

Total RNA including miRNA was extracted from whole blood samples collected in PAXgene Blood RNA tubes (QIAGEN). These tubes contain a reagent that lyses blood cells and immediately stabilizes intracellular RNA in order to prevent RNA degradation. RNA molecules longer than 18 nucleotides were purified from the stabilized blood samples using the PAXgene Blood miRNA Kit (QIAGEN) following the manufacturer's instructions. Briefly, lysed blood cells were first centrifuged, washed and subsequently subjected to proteinase K treatment. The cell lysate was then homogenized using PAXgene shredder spin column. Isopropanol was added to optimize binding conditions, and the lysate was subjected to PAXgene RNA spin column. Total RNA bound to the PAXgene silica-membrane was treated with DNase to remove genomic DNA contamination. RNA was washed and eluted in buffer.

DNA was extracted from EDTA whole blood samples using the Invisorb Blood Giga Kit following the manufacturer's protocol. In brief, erythrocytes were first lysed and

removed. In the subsequent step, leukocytes were isolated from interfering hemoglobin and were subjected to second lysis step. Afterwards, DNA was separated from protein fragments using proteinase K and precipitated by adding 96% ethanol to the lysate. The precipitated DNA was washed with 70% ethanol, dried and resuspended in low salt buffer for further downstream applications.

2.3 Library generation for RNA sequencing

Paired-end libraries were prepared using the Illumina TruSeq RNA Sample Preparation Kit (low-throughput protocol) following the manufacturer's protocol. Briefly, the messenger RNA (mRNA) was purified from the total RNA using polyA selection and then chemically fragmented. Following the fragmentation, the first cDNA strand was synthesized using reverse transcriptase and random primers. Next, the second cDNA strand was synthesized to create double-stranded cDNA. The cDNA fragments were then subjected to end repair process, addition of a single 'A' nucleotide to the 3' end, and ligation of the adapters. The ligated products were subsequently purified and amplified with PCR to produce a final double-stranded cDNA library. The libraries were quality-checked using the Agilent Bioanalyzer 2100. The Illumina TruSeq RNA Sample Preparation Kit employed here does not preserve the strandedness of the RNA transcripts.

2.4 RNA sequencing

The paired-end libraries were sequenced using the HiSeq 2000 technology (Illumina). Before sequencing, the prepared libraries were first subjected to bridge amplification in order to generate clonal clusters. Single-molecule DNA fragments were amplified in a flow cell whose surface consists of covalently attached oligonucleotides which are

complementary to the adapters ligated to the library fragments. To obtain clonal clusters, DNA fragments were first hybridized to the oligonucleotides on the flow cell and then amplified by bridge amplification (Figure 2-2). During this process, the hybridized DNA fragments arch over and hybridize to an adjacent oligonucleotide on the flow cell. In the next step, DNA polymerase was used to copy the DNA fragments from the hybridized oligonucleotides, thus forming double stranded DNA (dsDNA) bridges. These dsDNA bridges were subsequently denatured to form two single stranded DNA (ssDNA) strands that again arch over and hybridize to adjacent oligonucleotides. This process was repeated multiple times, which results in millions of individual, dense clonal clusters. This cluster generation step was carried out on the Illumina cBOT instrument.

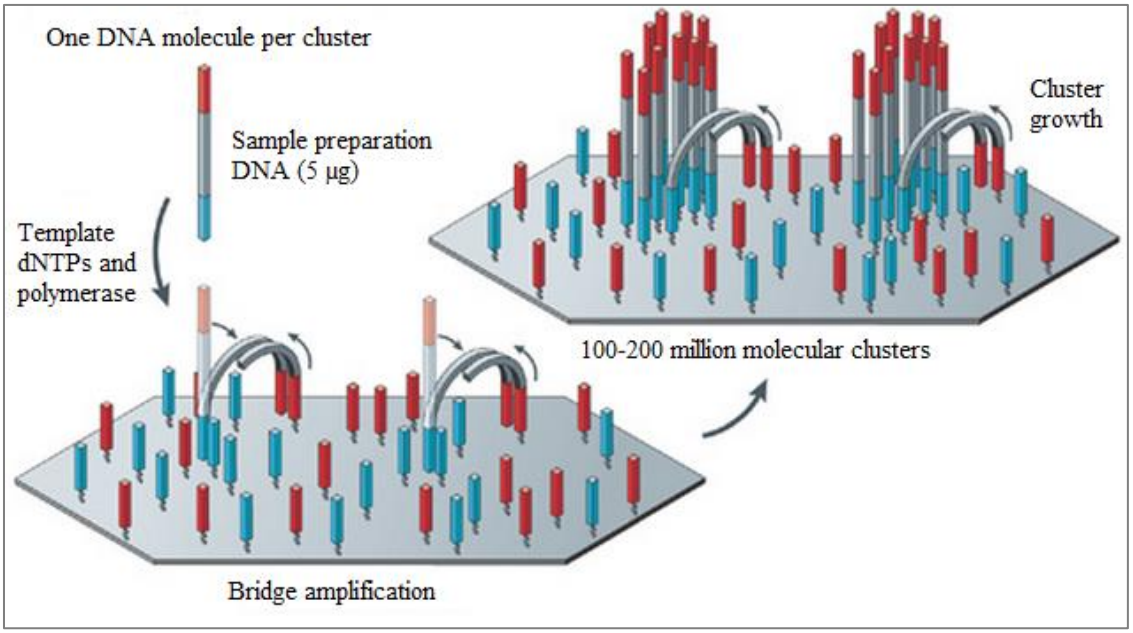

Figure 2-2 Bridge amplification to obtain clonal clusters. It is composed of two basic steps: initial priming and extending of the single-stranded, single-molecule template, and bridge amplification of the immobilized template with immediately adjacent primers to form clusters (Source: Metzker 2010).

The generated clusters were sequenced using Illumina's well-established "sequencing-by-synthesis" (SBS) chemistry. SBS chemistry employs four reversible terminator nucleotides (Figure 2-3), each labeled with a different fluorescent dye and a chemically

inactivated 3'-OH group. Sequencing was initialized by hybridizing primers to the templates on the flow cell. Next, four fluorescently labeled nucleotides were added that compete for incorporation into the growing DNA strands. After incorporation of a single nucleotide, the free nucleotides were washed off and the fluorescence from the incorporated nucleotide was measured by laser excitation and recorded in an imaging step. Subsequently, the fluorescent dye was cleaved and the procedure was repeated for the adjacent base. Since these were paired-end libraries, sequencing was carried out from both ends of the adapter ligated fragments.

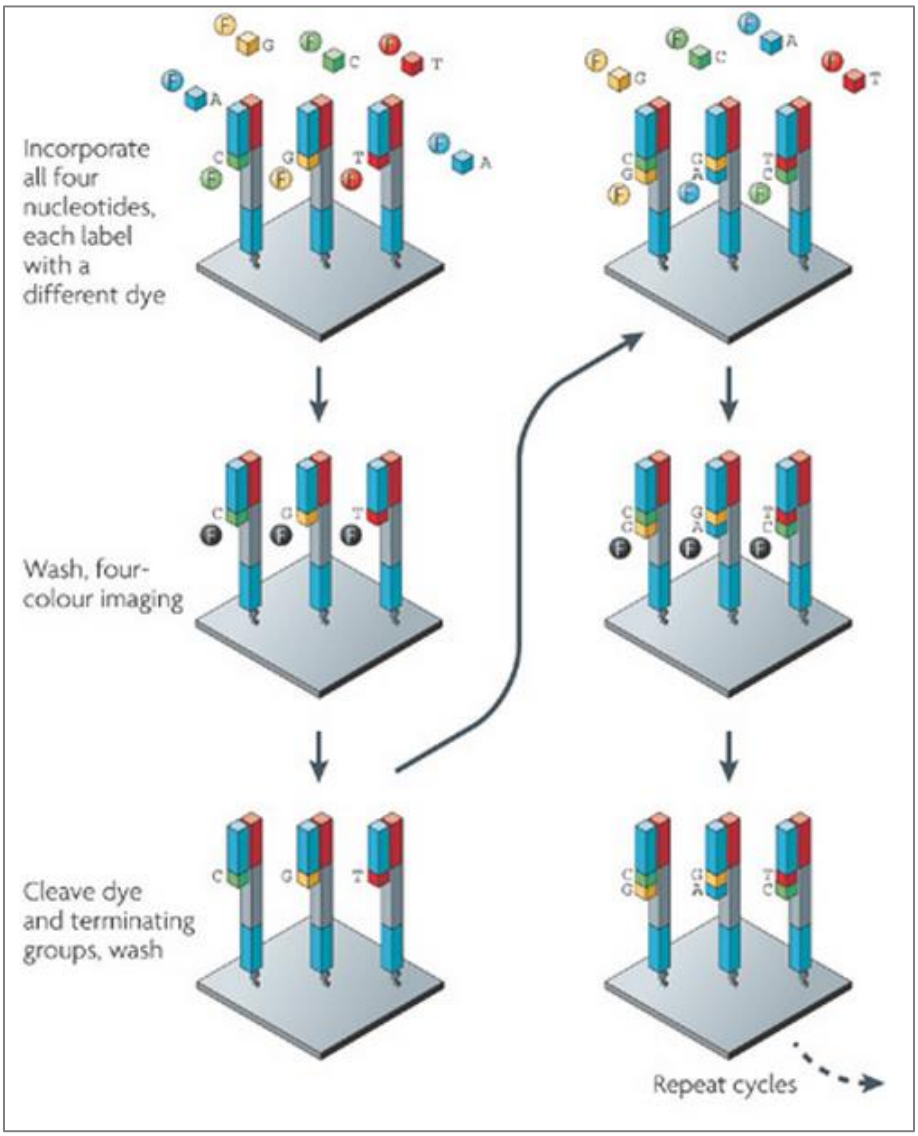

Figure 2-3 Four-color reversible termination method used by Illumina (Source: Metzker 2010).

2.5 Primary and secondary analysis of the sequence data

Primary analysis (i.e., image analysis, base calling and quality scoring) was performed on-instrument (Illumina, HiSeq 2000). During the base calling phase, the calls of the mixed clusters cannot be determined accurately. To address this issue, Illumina uses a built-in chastity filter to identify clusters with a low signal to noise ratio. Chastity for a given base call is defined as follows:

$$\text{chastity} = \frac{\text{highest intensity}}{\text{highest intensity} + \text{second highest}} \qquad (1)$$

By default the chastity threshold is 0.6. Clusters pass the filter if no more than one base call in the first 25 cycles has a chastity of below 0.6. The reads that failed the Illumina chastity filter were removed using the Illumina CASAVA-1.8 FASTQ filter v0.1. After chastity filtering, the quality of the reads was further assessed using FASTQC v0.10.0 (Andrews 2010).

The raw reads were then mapped against the human reference (hg19) using TopHat v2.0.4 (Trapnell et al. 2009), a fast splice junction aligner. The most recent UCSC genome build (hg19) and a pre-built index were downloaded from the TopHat homepage (http://tophat.cbcb.umd.edu/) and used as a reference. TopHat aligned reads to the reference genome in two phases. In the first phase, TopHat used Bowtie2 v2.0.0 (Langmead and Salzberg 2012) to align all reads to the reference genome allowing no more than 2 mismatches. Reads that were not mapped in the first phase were set aside. TopHat then generated a database of putative exons and potential splice junctions from the aligned reads. The unmapped reads from the first phase were then aligned to putative exons in this database while a seed and extend alignment method was employed to align reads to possible splice sites. The alignments were conducted using TopHat's default parameters except for the below mentioned:

--mate-inner-dist (r) and --mate-std-dev: The mean inner distance between pairs and the standard deviation for the distribution on inner distances between pairs were

determined for each sample by aligning a subset of reads (around 5 million) against the human reference (hg19) using BWA v0.5.9 (Burrows-Wheeler Aligner) (Li and Durbin 2009). The insert size and standard deviation values were calculated from the obtained SAM files (Sequence Alignment/Map, http://samtools.sourceforge.net) using Picard's CollectInsertSizeMetrics (http://picard.sourceforge.net/). The combined read length was then subtracted from the calculated insert size to obtain the mate inner distance. These values were passed to TopHat for alignment.

2.6 Gene assembly and quantification

Aligned reads were processed using Cufflinks v2.0.2 (Trapnell et al. 2010). Cufflinks constructed a directed acyclic read overlap graph and assembled transcripts by finding a minimal path cover through minimum cost maximum matching. Additionally, the gene-level expression was calculated and reported in terms of fragments per kilobase of exon per million fragments mapped (FPKM). The supplied reference annotation (hg19) was used by Cufflinks as the basis for the transcript assembly and gene-level abundance estimation. Cufflinks was instructed to run a bias detection and correction algorithm (Roberts et al. 2011) in order to improve accuracy of abundance estimates by providing a multifasta reference file (hg19) via a command option --frag-bias-correct. Additionally, other command line options mentioned in Table 2-2 were enabled.

Table 2-2 List of command line options used in cufflinks while assembling and quantifying genes.

Command line options	Function
--upper-quartile-norm	Normalizes FPKM values by the upper quartile of the number of reads mapped to individual loci instead of the total number of sequenced reads. This can improve the robustness for less abundant genes while performing differential expression analysis
--multi-read-correct	Instructs cufflinks to do an initial transcript abundance estimation procedure to more accurately weight reads mapping to multiple positions in the genome
--compatible-hits-norm	Considers only those reads that are compatible with reference transcript as the number of mapped reads while estimating FPKM values

2.7 Detection of differentially expressed genes

The obtained FPKM estimates for genes were processed using custom Perl and R scripts to identify the differentially expressed genes. Firstly, FPKM estimates from all the German samples were constructed into a matrix where each row represented a gene and each column represented a sample. Between-sample normalization was carried out using the quantile normalization method in order to make genes comparable between samples, and FPKM estimates were then log2 transformed. Only those genes expressed (i.e., log2 FPKM > 0) in more than 5% of the samples were considered and differential analysis was conducted using a non-parametric Wilcoxon rank sum test. To address the issue of multiple testing, p-values were adjusted using Benjamini and Hochberg's method (Benjamini and Hochberg 1995). A p-value cutoff of 0.001 was applied for selecting the differentially expressed genes.

2.8 Filtering blood cell-specific genes

It is evident that hematopoietic changes occur with advanced age (Pang et al. 2011; Kuranda et al. 2011). Since whole blood samples were used in this study, differential expression of few genes may be attributed to age-related hematopoietic changes. To address this issue, the differentially expressed genes that highly correlated ($r \geq 0.8$) with blood cells (Palmer et al. 2006) were filtered and omitted from further analysis.

2.9 Heritability estimation

The gene expression data from the Danish twins was subjected to between-sample normalization, and log2 transformed as described for the German samples. The genes which were expressed (i.e., log2 FPKM > 0) in more than 95% of the samples were considered for heritability estimation. The heritability of gene expression (transcriptional activity) was estimated using standard structural equation modeling (SEM) (Neale et al. 1992), which is a statistical technique extensively used in quantitative genetics to estimate genetic and environmental influences on phenotypes of interest. Under classical biometric analyses of twin data, the total phenotypic variance in quantitative traits can be partitioned into two components, namely, variance that is associated with genetic effects, and variance that is associated with environmental effects. This can be represented as below:

$$V_P = V_G + V_E \qquad (2)$$

Where:

V_P = total phenotypic variance in a trait

V_G = genetic variance

V_E = environmental variance

The genetic variance can be further decomposed into additive and dominance components and the environmental variance can be decomposed into shared and non-shared environmental components. Thus, the total phenotypic variance can be written as:

$$V_P = V_A + V_D + V_E + V_C \quad (3)$$

Where:

V_A = additive genetic variance

V_D = dominance genetic variance (non-additive genetic effects that represent interactions between alleles at a locus)

V_E = unique environmental variance (environmental exposures unique to each twin)

V_C = common environmental variance (environmental exposures shared by both twins)

These different variance components can be estimated using twin data, since monozygotic (MZ) twins share all additive (A) and dominant (D) genetic effects whereas dizygotic (DZ) pairs share 50% of additive (A) and 25 % of dominant (D) genetic effects (Rijsdijk and Sham 2002). Also, both MZ and DZ pairs have equal shared environmental effects.

The heritability of transcriptional activity was estimated by considering only the additive genetic effects. Thus, the ACE model was fitted using biometric modeling (SEM), and was allowed to compete with more parsimonious models, namely, AE, CE and E. The best model was chosen based on the Akaike Information Criterion (AIC) (Akaike 1974) for non-nested models and the likelihood ratio test for nested models. The final heritability estimate was obtained from the best model. The R statistical package called “Mets” (v0.1-13) (http://CRAN.R-project.org/package=mets) was used for the biometric modeling.

2.10 Genotyping

DNA samples from all the German individuals (73 young and 55 LLI) were subjected for genotyping. In case of the Danish samples, only one individual from each MZ twin pair was selected as MZ twins are genetically identical. Therefore, 22 MZ, 50 DZ twins and 20 unrelated individuals from Denmark were used. In total, DNA samples from 220 individuals were genotyped using HumanOmniExpress BeadChips (Illumina). The arrays monitor 730,525 single nucleotide polymorphisms (SNPs) across the genome. The infinium HD (High-Density) Assay was used for genotyping following the manufacturer's instructions. In brief, the samples were subjected to whole genome amplification followed by fragmentation. In subsequent steps, the unlabeled DNA fragments were hybridized to the BeadChip and extended with single labeled nucleotides followed by signal detection using a scanner. The resulting genotype data was subjected to basic quality control using the GenomeStudio software (Illumina).

Additionally, two *APOE* variants namely, rs429358 and rs7412, that collectively determine the *APOE* genotype (Lindahl-Jacobsen et al. 2013), and three *FOXO3A* variants (rs3800231, rs2764264 and rs4945816) (Flachsbart et al. 2009) that are known to be associated with longevity were genotyped using predesigned Taqman genotyping assays (Life Technologies). These variants were genotyped in all the above mentioned samples, except in 4 unrelated individuals from Denmark due to a lack of DNA. The genotyping was performed according to the manufacturer's instructions. Briefly, the reaction mixtures consisting of DNA, PCR reaction mix components, and the TaqMan probes were prepared and the reaction was carried out on the 7900HT Real-time PCR system (Life Technologies). After completion of the run, genotypes were determined by measuring the fluorescence with the Sequence Detection Systems (SDS, Applied Biosystems) 2.0 software.

2.11 Expression quantitative trait loci analysis – investigating *cis*-regulation of gene expression

Expression quantitative trait loci (eQTL) analysis is a statistical method that correlates gene expression with genotype data. eQTL analysis provides an opportunity to investigate the influence of genetic variation on mRNA expression levels. Genetic variants can influence gene expression in many ways, for example by altering the promoter or coding sequences of genes, their splice junctions, or other gene regulatory elements. A gene can be regulated by two types of eQTLs, namely *cis* and *trans*. Variants that affect genes that are in their close proximity are termed *cis* regulators and variants that affect genes elsewhere on the genome are referred to as *trans* regulators.

The gene expression and genotype data from the German and Danish samples were correlated separately. Since the samples were anonymized before genotyping, both gene expression and genotype data were passed on to a trustee (PopGen biobank) in order to match them. The obtained matched data contained no true sample identity. As the German samples were from unrelated individuals and the Danish samples were from twins, two different approaches were employed to perform the eQTL analysis. The details are described below.

In case of the German samples, the *cis*-eQTL analysis was carried out for the genes that were differentially expressed between young and long-lived individuals. The analysis was performed with a 1-Mb window, which means only those variants that were located within a 1-Mb sequence both upstream and downstream from the starting and end points of the gene were included in the analysis. Some variants were used more than once because of overlapping intervals. The selected variants within the 1-Mb window were analyzed for association with gene expression using PLINK v1.07 (Purcell et al. 2007). PLINK is a well-established command-line tool, which focuses on analysis of genotype/phenotype data. It provides features that facilitate inclusion of multiple covariates when testing for quantitative trait and genetic variant association. *Cis*-eQTL

analysis was performed on each differentially expressed gene using a linear model (option --linear), in which age and gender were provided as covariates. Only those variants with minor allele frequency above 1% and missingness below 5% (variants having missing values in less than 5% of the individuals) were used. The obtained *cis*-eQTLs were adjusted for multiple testing using Benjamini and Hochberg's method (Benjamini and Hochberg 1995). The associations with a p-value ≤ 0.05 were considered as potential *cis*-eQTLs.

In case of the Danish twins, the *cis*-eQTL analysis was performed on those genes which had heritable transcriptional activity. As described for the German samples, a 1-Mb window was used. The analysis was carried out using Genome-wide Efficient Mixed Model Association (GEMMA) (Zhou and Stephens 2012). As a first step of the analysis, four input files containing genotypes, phenotypes (gene expression), covariates and relatedness matrix were prepared as per GEMMA's requirements. The gene expression and genotype data was prepared in the PLINK binary ped format. Since MZ twins are genetically identical, only one of each twin pair was genotyped and the same data was used for the other twin. The standardized relatedness matrix file was constructed in an n x n format, where each row and each column corresponds to individuals in the same order as in the genotype file. The value in i^{th} row and j^{th} column indicates the relatedness between i^{th} and j^{th} individuals. Also, a covariate file containing age and gender information was provided. GEMMA then performed the association test between gene expression and genotype data by fitting a univariate linear mixed model. Only those variants that had missingness below 5% and minor allele frequency above 1% were analyzed. The *cis*-eQTLs for each gene were corrected for multiple testing using Benjamini and Hochberg's method and a p-value cutoff of 0.05 was applied.

2.12 Investigation of genotype-age interaction effects on gene expression

The eQTL analysis identifies the variants that influence the expression of a gene, but it cannot identify the changes in the influence of variants on gene expression due to their interaction with age. Thus, the contribution of genotype-age (G×A) interaction to the variance observed in differentially expressed genes was investigated. For this analysis, only the differentially expressed genes with at least one significant *cis*-eQTL were considered. The G×A interaction for each gene and the most significant *cis*-eQTL were investigated by employing a linear mixed model. Two models namely, a full model with a G×A interaction term and a reduced model without an interaction term, were fitted along with age, gender (fixed effects) and sequencing batch number (random effects), using a R package called lme4. The best fit model was selected based on AIC. If the best fit model was a full model with a G×A interaction term, then the gene was considered to have a G×A interaction effect on its expression.

2.13 Functional analysis: Gene ontology

Gene ontology (GO) is widely used to describe biological features of genes and their products. It provides a controlled vocabulary to categorize the functions of genes and gene products in terms of their biological processes, molecular functions and cellular components. GO analysis requires two input gene lists namely, a reference list that comprises all the genes considered in an experiment and a query list which contains the genes of interest. The analysis was carried out on different query lists as mentioned below:

a) Differentially expressed genes between long-lived and young individuals

b) Differentially expressed genes that were associated with *cis*-eQTLs

c) Differentially expressed genes that were not associated with *cis*-eQTLs

d) Differentially expressed genes that were associated with *cis*-eQTLs and also exhibited G×A interaction

e) Genes that showed heritable transcriptional activity

f) Genes that showed heritable transcriptional activity and *cis*-eQTL association

Significant enrichment or depletion of GO terms was identified for each query list using the two-sided Fisher's exact test, followed by Benjamini and Hochberg's correction for multiple testing.

2.14 Biological validation using the Danish samples

In order to validate the results from the differential expression analysis and the *cis*-eQTL analysis (from German samples) Danish samples were used. Since the German dataset consisted of unrelated individuals and the Danish dataset consisted of twins, one individual was selected from each twin pair in order to create an equivalent validation dataset. Additionally, 10 unrelated cases (83-92 years) and 10 unrelated controls (58-60 years) from Denmark were used in the validation dataset. In total, 68 samples (48 individuals selected from 48 twin pairs, of which 24 were cases and 24 were controls, and additional 10 cases and 10 controls) were used in the validation dataset. The transcriptome data from these 68 samples was analyzed as described previously (section 2.7) to identify and validate the differentially expressed genes detected in the German dataset.

Furthermore, the genotype and the expression data from the 68 Danish samples were used in order to validate the results from the *cis*-eQTL analysis. The analysis was carried out using PLINK as described in section 2.11.

2.15 Technical validation using real-time qPCR

An independent laboratory technique, quantitative PCR (qPCR) was used to validate the results from the transcriptome sequencing. The German samples (Table 2-1) which were used for the transcriptome sequencing were employed in this analysis. However, five samples (three young individuals and two LLI) were not included in the analysis due to lack of RNA. In order to select candidate genes for validation, the differentially expressed genes were first ranked based on their effect size (fold change), significance in the analysis between young and long-lived individuals, and regulation of their expression by *cis*-eQTLs. Based on the ranks, 80 candidate genes (see Table 3-5) were selected for validation using RT-qPCR. Additionally, three groups of controls were included, namely:

a) four endogenous control genes (see Table 2-3)

b) five genes which are known or presumed to have a genetic association with longevity, but were not differentially expressed between young and LLI (see Table 2-3)

c) five genes with no or low variation in expression levels among the samples (see Table 3-4)

Real-time PCR (RT-PCR) was performed according to the manufacturer's guidelines (Life Technologies) using a 7900HT Real-time PCR system (Applied Biosystems, California, USA). In brief, total RNA was reverse-transcribed to cDNA using the High Capacity RNA-to-cDNA kit and the quantity of the cDNA was measured by determining the UV absorbance (A_{260}/A_{280}) method. The reaction mixtures consisting of cDNA samples and PCR reaction mix components were prepared and then loaded on microfluidic cards, containing the selected TaqMan Gene Expression assays. The reactions were carried out on the 7900HT Real-time PCR system (Applied Biosystems, California, USA). Upon completion of the run, the data was preprocessed by the Sequence Detection System

(SDS) 2.0 software. Gene expression levels were measured using the delta-delta CT method (Livak and Schmittgen 2001). The three above mentioned groups of control genes were investigated for variation in the expression levels across the samples and the control group with minimum variation was selected as the control. In this case, genes from group (b) were selected as controls and the median of their expression levels was used to normalize the expression levels of the candidate genes. Differences between young and long-lived individuals were determined using the Wilcoxon rank sum test and the p-values were corrected for multiple testing using Benjamini and Hochberg's method (Benjamini and Hochberg 1995).

Table 2-3 First two groups of controls and their TaqMan assays used for validation.

Gene symbol	Gene name	TaqMan assay ID
Endogenous controls		
18s rRNA	-	Hs99999901_s1
PPIB	peptidylprolyl isomerase B (cyclophilin B)	Hs00168719_m1
TRAP1	TNF receptor-associated protein 1	Hs00212474_m1
PGK1	phosphoglycerate kinase 1	Hs00943178_g1
Known or presumed to have a genetic association with longevity, but not differentially expressed		
APOE	apolipoprotein E	Hs00171168_m1
ACE	angiotensin I converting enzyme	Hs00174179_m1
SIRT3	sirtuin 3	Hs00953477_m1
LMNA	lamin A/C	Hs00153462_m1
FOXO3	forkhead box O3	Hs00818121_m1

3 Results

3.1 General mapping statistics

In the presented study, transcriptome libraries from blood samples were sequenced in 244 individuals using Illumina's Truseq technology. For each individual, an average of 48 million (SD=14M) sequences was generated. When Illumina's chastity filter was applied to the raw sequence data in order to filter the sequences based on the signal purity, on an average approximately 90% (SD=4.9) of the sequences passed the filter. Out of these, an average of 85% (SD=4) of the sequences mapped to the human reference (hg19 build).

3.2 Genome-wide transcriptome profiling

3.2.1 Differential expression between long-lived and young individuals

Transcriptomes from young (n=73, age=20–55 years) and long-lived German individuals (n=55, age=91–104 years) were analyzed to identify differentially expressed genes. Gene expression levels were determined with the Cufflinks software (see 2.6), which reported gene abundance as FPKM. The annotation file (hg19) used in this context contained 23,368 genes. Out of these, 19,991 genes were expressed (i.e., log2 FPKM > 0) in at least 5% of the samples and were considered for the differential expression analysis. To ensure high statistical data reliability, stringent cutoff criteria (Benjamini – Hochberg corrected p-value ≤ 0.001; false discovery rate (FDR) ≤ 1%) were applied. Appling this filter, a total of 6,217 genes were found to be differentially expressed between long-lived and young individuals. Of these genes, 3,525 were downregulated and 2,689 genes were upregulated in LLI, revealing that the differential gene expression pattern was dominated by downregulated transcriptional activities in LLI. In addition to up- and downregulation of differentially expressed genes, for three genes an on/off effect could be observed, which cannot be quantified as a fold change: for a given gene, if the

median expression value of the young individuals was zero while the median expression value of the LLI was not zero, the corresponding gene was considered switched on in LLI (LOC100132735). The opposite effect was observed as well (*SLC7A3*, *OR10G2*). Among the differentially expressed genes, 178 showed significant gender effect (Benjamini – Hochberg corrected p-value ≤ 0.05; FDR ≤ 5%). A list of these genes is provided in the supplementary material (Table 8-1). The top 50 upregulated and downregulated genes in LLI (after filtering blood cell-specific genes (see 3.2.2)) are listed in Table 3-1 and Table 3-2, respectively. The remaining differentially expressed genes are provided as an electronic attachment.

Table 3-1 The top 50 genes which were upregulated in long-lived individuals.

Gene symbol	Gene name	p-value*	Fold change
DNAJC1	DnaJ (Hsp40) homolog, subfamily C, member 1	3.12E-13	1.593
LOC100271722	hypothetical protein LOC100271722	9.38E-12	1.817
PHLDA3	pleckstrin homology-like domain, family A, member 3	1.36E-11	3.143
ANKRD22	ankyrin repeat domain 22	1.40E-11	2.913
ODF3B	outer dense fiber of sperm tails 3B	1.59E-11	2.102
FCGR1A	Fc fragment of IgG, high affinity Ia, receptor (CD64)	1.67E-11	3.207
MS4A4A	membrane-spanning 4-domains, subfamily A, member 4A	3.08E-11	2.491
SAP30	Sin3A-associated protein, 30kDa	6.73E-11	1.740
POLB	polymerase (DNA directed), beta	2.03E-10	1.707
PSMB8	proteasome (prosome, macropain) subunit, beta type, 8	2.27E-10	1.237
VAMP5	vesicle-associated membrane protein 5	2.27E-10	2.163
PAQR4	progestin and adipoQ receptor family member IV	2.81E-10	1.717
ZBP1	Z-DNA binding protein 1	3.12E-10	1.480
FCGR1C	Fc fragment of IgG, high affinity Ic, receptor (CD64), pseudogene	3.66E-10	3.575
TRIP4	thyroid hormone receptor interactor 4	4.45E-10	1.441
SLC27A3	solute carrier family 27 (fatty acid transporter), member 3	4.50E-10	1.304
HOXB7	homeobox B7	4.70E-10	3.503
DES	desmin	4.71E-10	2.899
PDGFRB	platelet-derived growth factor receptor, beta polypeptide	8.29E-10	2.629
MANEAL	mannosidase, endo-alpha-like	8.60E-10	2.473
RIPK3	receptor-interacting serine-threonine kinase 3	8.93E-10	1.448
BATF2	basic leucine zipper transcription factor, ATF-like 2	9.56E-10	2.925
FAM20A	family with sequence similarity 20, member A	9.66E-10	2.419
LGALS1	lectin, galactoside-binding, soluble, 1	1.01E-09	1.830
SQRDL	sulfide quinone reductase-like (yeast)	1.22E-09	1.261
BLVRA	biliverdin reductase A	1.39E-09	1.673

Gene symbol	Gene name	p-value*	Fold change
PARP3	poly (ADP-ribose) polymerase family, member 3	1.41E-09	1.585
SMARCD3	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 3	1.49E-09	1.778
DDO	D-aspartate oxidase	1.55E-09	2.292
SCO2	SCO2 cytochrome c oxidase assembly protein	1.79E-09	2.223
CDKN2B-AS1	CDKN2B antisense RNA 1	1.81E-09	2.206
TP53I3	tumor protein p53 inducible protein 3	1.87E-09	2.080
CD63	CD63 molecule	1.93E-09	1.234
IFI35	interferon-induced protein 35	1.93E-09	1.693
IL18	interleukin 18 (interferon-gamma-inducing factor)	2.34E-09	1.733
PSME2	proteasome (prosome, macropain) activator subunit 2 (PA28 beta)	2.34E-09	1.427
IGFBP7	insulin-like growth factor binding protein 7	2.99E-09	1.536
FAM110B	family with sequence similarity 110, member B	3.06E-09	1.573
RAB24	RAB24, member RAS oncogene family	3.14E-09	1.392
HSD3B7	hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 7	3.18E-09	2.273
CLIC1	chloride intracellular channel 1	3.85E-09	1.355
FBXO6	F-box protein 6	4.89E-09	1.511
APOL4	apolipoprotein L, 4	5.28E-09	2.772
ETV7	ets variant 7	5.50E-09	3.566
C5orf56	chromosome 5 open reading frame 56	6.15E-09	1.439
C11orf75	chromosome 11 open reading frame 75	6.23E-09	1.512
RASGEF1A	RasGEF domain family, member 1A	6.38E-09	1.591
TMEM62	transmembrane protein 62	6.48E-09	1.506
DDB2	damage-specific DNA binding protein 2, 48kDa	6.56E-09	1.534
C16orf7	chromosome 16 open reading frame 7	6.80E-09	1.686

*p-value is corrected for multiple testing

Table 3-2 The top 50 genes which were downregulated in long-lived individuals.

Gene symbol	Gene name	p-value*	Fold change
LRRN3	leucine rich repeat neuronal 3	9.03E-17	-9.112
CD248	CD248 molecule, endosialin	1.57E-15	-5.469
NELL2	NEL-like 2 (chicken)	1.82E-14	-2.332
NOG	noggin	4.46E-14	-3.949
ABLIM1	actin binding LIM protein 1	6.13E-14	-2.320
NT5E	5'-nucleotidase, ecto (CD73)	8.94E-14	-2.841
CAMK4	calcium/calmodulin-dependent protein kinase IV	1.19E-13	-2.060
FAM102A	family with sequence similarity 102, member A	1.19E-13	-2.000
RCAN3	RCAN family member 3	2.30E-13	-2.099
RASGRF2	Ras protein-specific guanine nucleotide-releasing factor 2	3.42E-13	-2.431

Gene symbol	Gene name	p-value*	Fold change
SERPINE2	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 2	3.42E-13	-2.790
NPAS2	neuronal PAS domain protein 2	3.70E-13	-4.513
SPTBN1	spectrin, beta, non-erythrocytic 1	3.70E-13	-2.269
N4BP3	NEDD4 binding protein 3	5.59E-13	-2.631
SFRP5	secreted frizzled-related protein 5	5.59E-13	-12.998
KLHL3	kelch-like family member 3	5.85E-13	-1.788
CHMP7	charged multivesicular body protein 7	7.04E-13	-1.591
GPRASP1	G protein-coupled receptor associated sorting protein 1	7.26E-13	-2.340
CACHD1	cache domain containing 1	9.12E-13	-5.505
NIPAL3	NIPA-like domain containing 3	1.48E-12	-1.501
ZNF549	zinc finger protein 549	1.82E-12	-1.490
AQP3	aquaporin 3 (Gill blood group)	2.12E-12	-2.174
EDAR	ectodysplasin A receptor	3.08E-12	-3.476
ZFYVE9	zinc finger, FYVE domain containing 9	3.08E-12	-1.733
CCR7	chemokine (C-C motif) receptor 7	3.40E-12	-2.111
NR3C2	nuclear receptor subfamily 3, group C, member 2	5.01E-12	-1.967
BACH2	BTB and CNC homology 1, basic leucine zipper transcription factor 2	5.18E-12	-1.989
FAM171A1	family with sequence similarity 171, member A1	5.47E-12	-1.736
PRKCA	protein kinase C, alpha	5.47E-12	-1.838
KIF5C	kinesin family member 5C	8.13E-12	-1.884
SPON1	spondin 1, extracellular matrix protein	8.77E-12	-2.601
CCR6	chemokine (C-C motif) receptor 6	9.67E-12	-2.082
ZNF551	zinc finger protein 551	9.79E-12	-1.932
MAN1C1	mannosidase, alpha, class 1C, member 1	1.03E-11	-1.870
ITPKB	inositol 1,4,5-trisphosphate 3-kinase B	1.06E-11	-1.496
SLC16A10	solute carrier family 16, member 10 (aromatic amino acid transporter)	1.07E-11	-3.204
SLC4A10	solute carrier family 4, sodium bicarbonate transporter-like, member 10	1.07E-11	-7.548
PLEKHG4	pleckstrin homology domain containing, family G (with RhoGef domain) member 4	1.17E-11	-1.627
PPFIBP1	PTPRF interacting protein, binding protein 1 (liprin beta 1)	1.17E-11	-2.289
EPHA4	EPH receptor A4	1.33E-11	-2.431
MYC	V-myc myelocytomatosis viral oncogene homolog (avian)	1.59E-11	-2.138
PLCG1	phospholipase C, gamma 1	1.67E-11	-1.646
ZNF304	zinc finger protein 304	2.13E-11	-1.579
AMIGO1	adhesion molecule with Ig-like domain 1	2.75E-11	-1.731
CDR2	cerebellar degeneration-related protein 2, 62kDa	3.08E-11	-1.534
DNMT3A	DNA (cytosine-5-)-methyltransferase 3 alpha	3.08E-11	-1.371
PIK3IP1	phosphoinositide-3-kinase interacting protein 1	3.08E-11	-1.801
C2orf89	chromosome 2 open reading frame 89	3.14E-11	-1.774
TMEM63A	transmembrane protein 63A	3.45E-11	-1.523
LDLRAP1	low density lipoprotein receptor adaptor protein 1	3.52E-11	-1.548

*p-value is corrected for multiple testing

3.2.2 Blood cell-specific genes

Since whole blood samples were used in this study, differential expression of few genes may be attributed to age-related hematopoietic changes. To address this issue, the differentially expressed genes that highly correlated with blood cells ($r \geq 0.8$) were omitted from the further analyses. A study conducted by Palmer et al. (2006) had identified gene expression signatures for the major constituent cells of blood, including B-cells, CD4+ T-cells, CD8+ T-cells, and granulocytes. This observation was used as a basis for categorizing blood cell-specific genes in the present study. However, as no validated erythrocyte-specific mRNA transcripts have been described previously, it was not possible to filter genes associated with this cell type from the analysis. The only study investigating erythrocyte-specific genes worked with erythrocyte cell populations of 99.99% purity (Kabanova et al. 2009). Taking into account the extremely low RNA content of these anucleate cells, it is likely that a very small proportion of other blood cells can influence the results dramatically. Consequently, the resulting data cannot be used to precisely identify erythrocyte-specific genes.

Out of the 6,217 genes that were significantly differentially expressed between long-lived and young individuals, 139 were blood cell-specific and thus were omitted from the further analyses. Of these genes, 112 were downregulated and 27 were upregulated in LLI. It was observed that the B-cell and T-cell-specific genes together constituted for more than 75% of the identified blood cell-specific genes (Figure 3-1). A list of these genes is provided in the supplementary material (Table 8-2). After filtering the 139 blood cell-specific genes, the remaining 6,078 differentially expressed genes were considered for the further analyses.

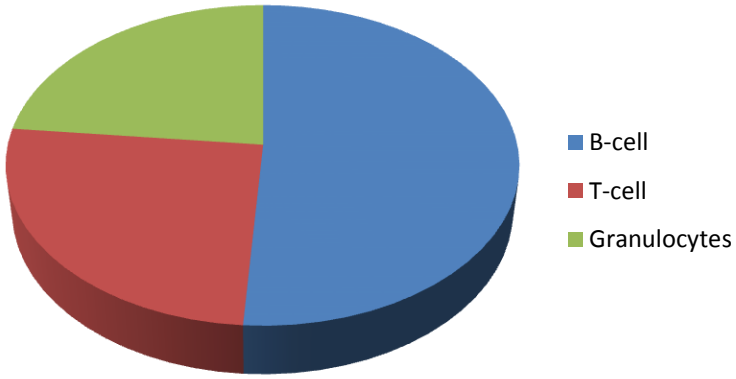

Figure 3-1 Distribution of significant blood cell specific genes (B-cell, T-cell, granulocytes)

3.2.3 Functional grouping of the differentially expressed genes

The 6,078 genes that were differentially expressed between long-lived and young individuals were associated with biological processes in order to place them into a biological context. GO analysis (see 2.13) reported that 44 and 75 biological processes were significantly represented with respect to the genes upregulated and downregulated in LLI, respectively. Among the 75 biological processes associated with downregulated genes, 59 were significantly enriched, which means many more genes than expected by chance were associated with these processes. On the other hand, 16 biological processes were significantly depleted – a smaller number of genes than expected by chance were associated with these processes. Similarly, out of 44 biological processes associated with upregulated genes, 16 were significantly enriched and 28 were depleted. The top 20 biological processes represented by up- and downregulated genes are depicted in Figure 3-2. Interestingly, nine of the top 20 processes were metabolism-related and all of them contained more downregulated genes than expected by chance, indicating the depletion of these processes in LLI. All the significantly represented biological processes are presented in the supplementary material (Table 8-3).

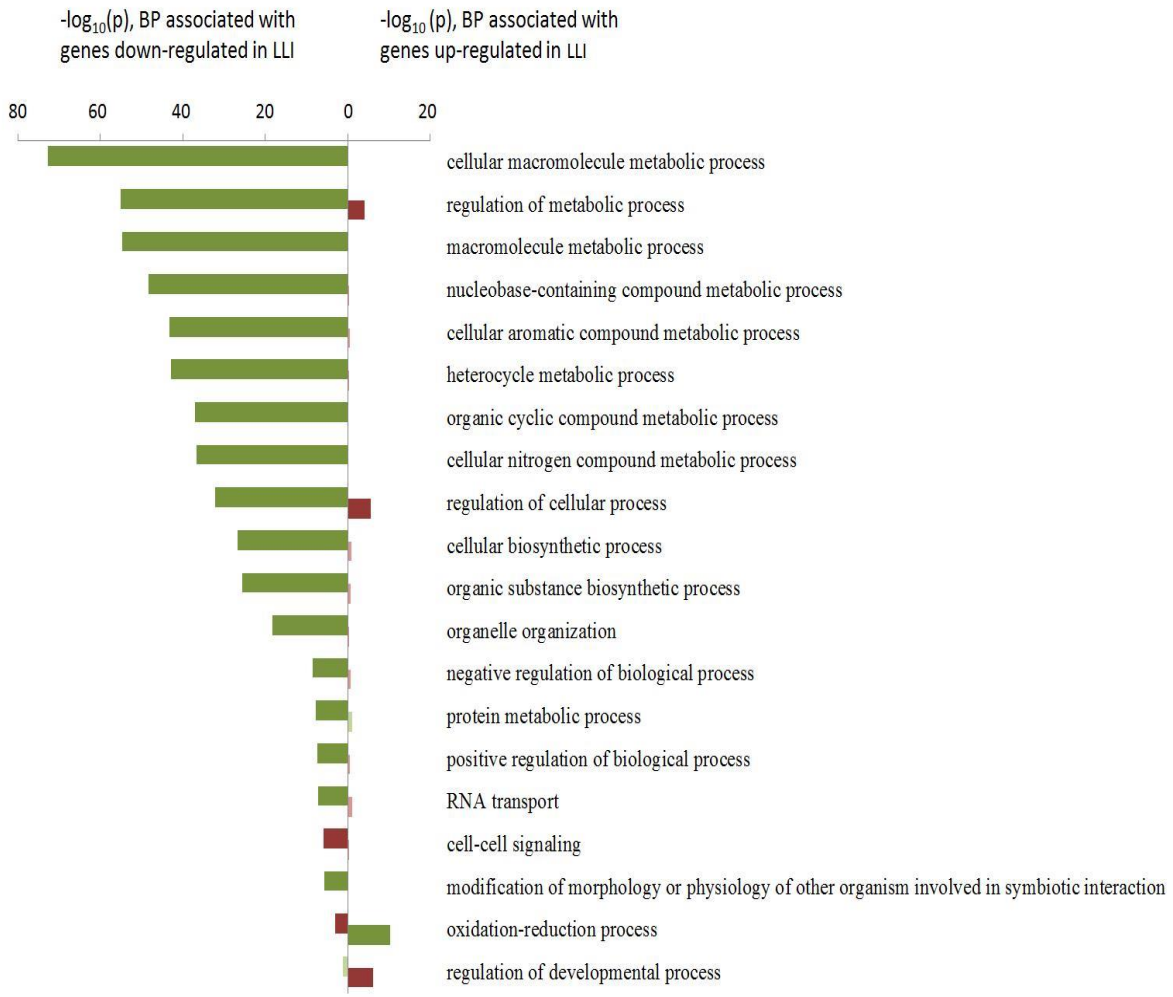

Figure 3-2 The top 20 biological processes (BP) with respect to up- and downregulated genes in case of long-lived individuals (LLI) vs. young. Green = significantly enriched processes, red = significantly depleted processes, light green = enriched processes but not significant and light red= depleted processes but not significant. Interestingly, 9 out of the top 20 were metabolism-related processes.

3.2.4 Heritable transcriptional activity

Transcriptomes from the Danish twins were used to estimate the heritability of gene expression. The expression levels were determined as described in section 2.6. Out of 23,368 genes that were described in the annotation file (hg19), 15,640 were expressed

(i.e., log2 FPKM > 0) in more than 95% of the blood samples and were considered for the analysis. The heritability of gene expression (transcriptional activity) was estimated by employing standard structural equation modeling (biometric modeling) (Neale et al. 1992). In this study, the narrow-sense heritability (h^2), which is the additive genetic portion of the phenotypic variance, was assessed for each gene. For this purpose, four different models (ACE, AE, CE and E) were fitted for each gene and heritability estimates were obtained from the best-fitting model. The best model was chosen based on the AIC (Akaike 1974) for non-nested models (AE and CE are non-nested models) and the likelihood ratio test for nested models (ACE, AE, CE and E are nested models). Of the 15,640 genes, the expression of 3,029 genes was identified to be heritable with values ranging from 0.30 to 0.99. It was observed that the heritability estimates for around 55.5% (1,683 of 3,029) of the genes were between 0.4 and 0.6 (Figure 3-4). Interestingly, the variation in expression for about 99.9% of the genes with heritable transcriptional activity was better explained by the AE model and for two genes (*NCAPG2* and *ARL17A*) it was better explained by the ACE model. Furthermore, of the 15,640 genes, the variation in expression for around 73.5% of the genes was explained by unique or non-shared environmental factors, indicating that a large proportion of the genes are influenced by non-genetic factors rather than by genetic factors (Figure 3-3). The top 50 genes with heritable transcriptional activity are provided in the supplementary material (Table 8-4) and the remaining genes are provided as an electronic attachment.

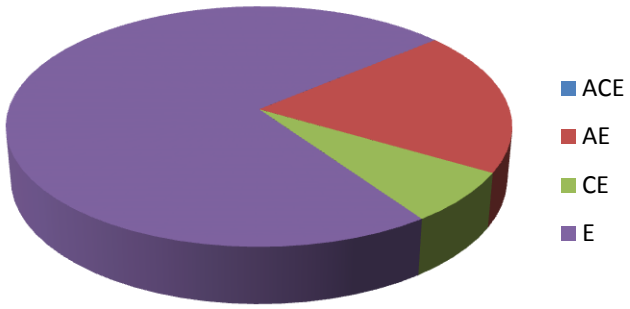

Figure 3-3 Distribution of genes with respect to different models, namely, ACE, AE, CE and E. (A: additive genetic factors, C: common or shared environmental factors, E: unique or non-shared environmental factors).

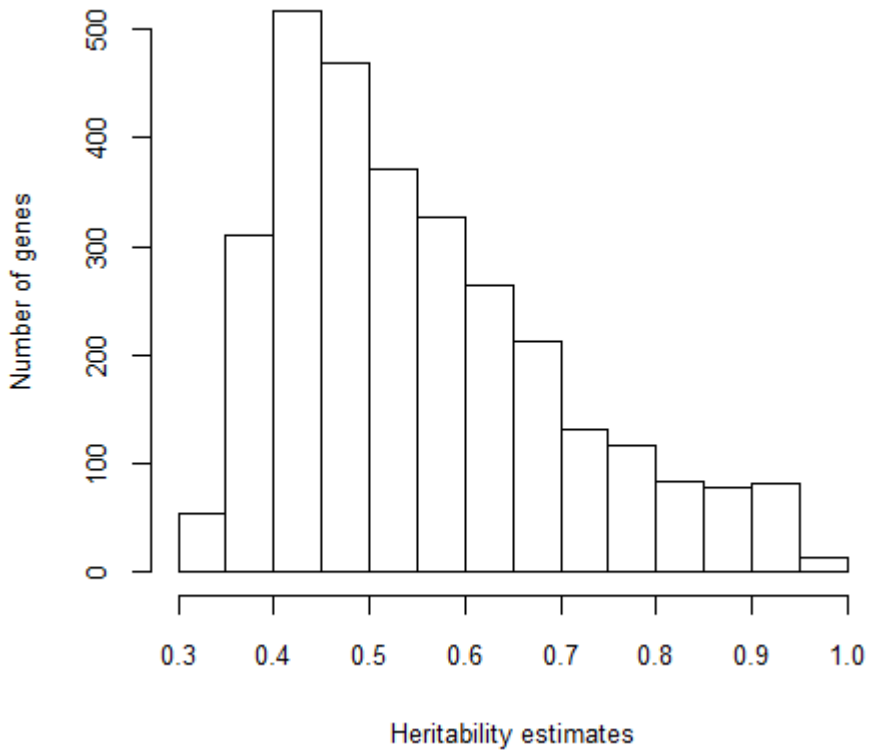

Figure 3-4 Frequency distribution of heritability estimates derived from gene expression.

3.2.5 Functional grouping of the genes with heritable transcriptional activity

The 3,029 genes with heritable transcriptional activity (see 3.2.4) associated with 693 biological processes, of which 16 were significantly depleted, and 11 were significantly enriched.

Significantly depleted GO terms included metabolic processes, cell cycle, organelle organization and RNA transport. On the other hand, significantly enriched GO terms included defense response, cell-cell signaling, cell adhesion, cell communication, regulation of hormone levels, reactive nitrogen species metabolic process and extracellular structure organization.

Interestingly, among the 27 significantly represented biological processes, nine were metabolism-related (Figure 3-5) and they contained a smaller number of heritable genes than expected by chance, illustrating that the genetic contribution to these processes plays only a minor role.

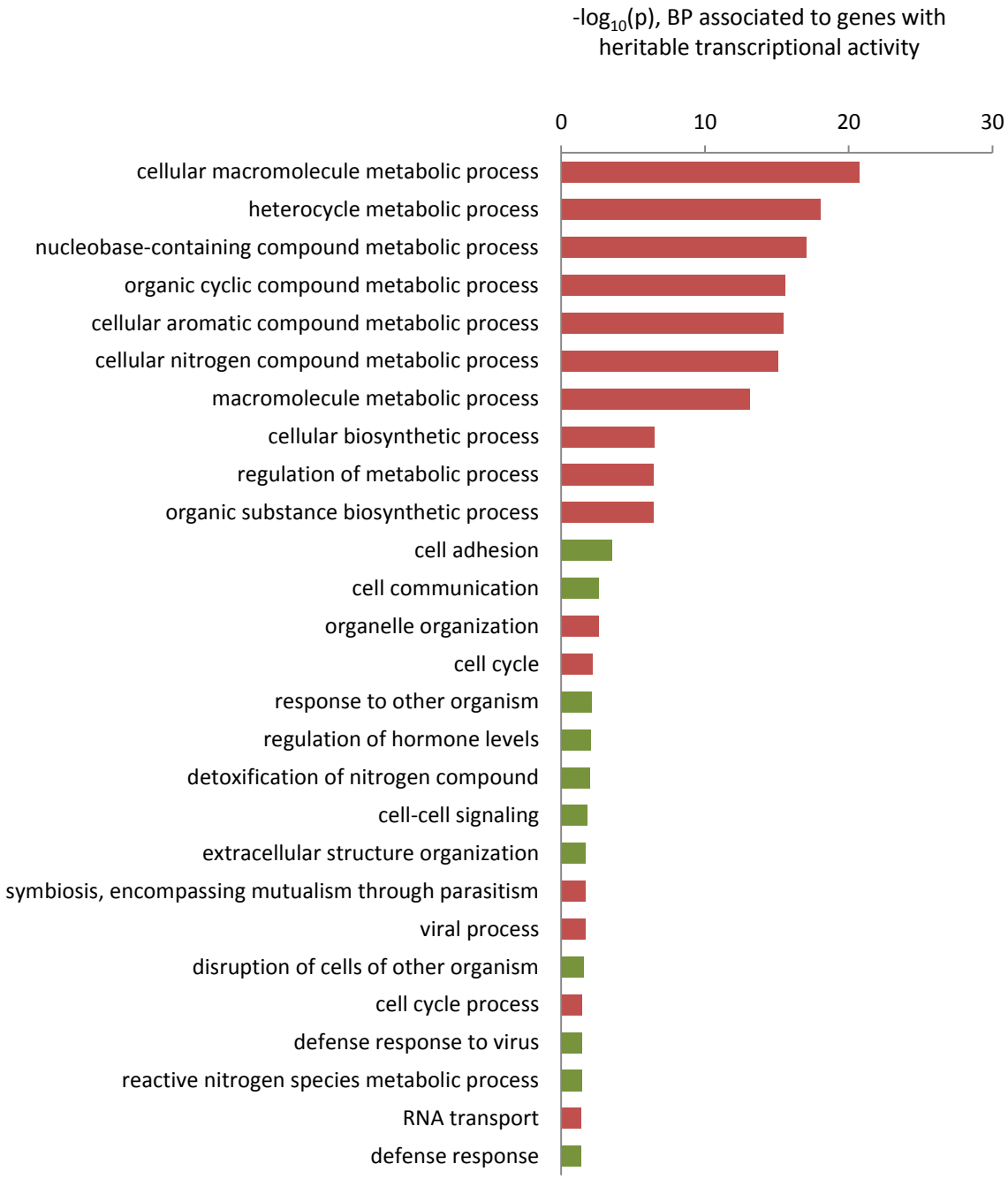

Figure 3-5 Biological processes (BP) associated with the 3,029 genes with heritable transcriptional activity; green= significantly enriched biological processes and red=significantly depleted processes.

3.3 Genotype data

In total, 220 samples (73 young and 55 long-lived individuals from Germany and 50 DZ, 22 MZ and 20 unrelated individuals from Denmark) were genotyped using HumanOmniExpress BeadChips (Illumina). Of the 220 samples, six samples (all from Denmark, two long-lived DZ and four middle-aged DZ) had a sample call rate below 90% and thus were removed from the analysis. Furthermore, SNPs with a minor allele frequency below 1%, call rate below 95% and Hardy-Weinberg equilibrium testing p-value <=0.0001 were excluded from the analysis. After quality control, 636,904 SNPs were used for the eQTL analysis for the German samples (73 young and 55 LLI). In the case of Danish samples, 621,985 SNPs passed the quality criteria and were employed for the eQTL analysis. Since MZ twins are genetically identical, only one individual from each pair was genotyped and the same data was used for the other twin. Thus, the eQTL analysis was performed with 44 MZ (24 middle-aged and 20 LLI) and 44 DZ twins (24 middle-aged and 20 LLI). The data from 20 unrelated individuals from Denmark was used for validating the *cis*-eQTL results obtained from the German samples (see 3.5.2).

Additionally, five variants (two *APOE* variants: rs429358 and rs7412, which collectively define the *APOE* genotypes and three *FOXO3A* variants: rs3800231, rs2764264 and rs4945816) were genotyped using TaqMan genotyping assays (Life Technologies). These variants were genotyped in all of the above mentioned samples, but for four Danish unrelated individuals due to the lack of DNA. All of the five variants had a call rate of >97% and a minor allele frequency >1%.

3.4 Regulation of gene expression by genetic factors

To identify eQTLs that have putative *cis*-regulatory effects on the genes that were differentially expressed (see 3.2.1) between long-lived and young individuals, all the expression traits were analyzed for association with SNPs that were located in a 1-Mb

window (both upstream and downstream from the starting and end points of the gene). Similarly, all the expression traits with heritable transcriptional activity (see 3.2.4) were analyzed for association with SNPs that were within a 1-Mb window to identify putative *cis*-eQTLs. Additionally, G×A interaction effects were investigated for the differentially expressed genes that exhibited significant association with *cis*-variants. From now on, the differentially expressed genes with significant *cis*-effects are called eQTL genes and the differentially expressed genes without significant *cis*-effects are called non-eQTL genes.

3.4.1 *Cis*-regulation of differentially expressed genes

Cis-eQTL analysis focusing on SNPs located within a 1-Mb window around each gene identified 8,757 statistically significant *cis*-eQTLs associated with 1,200 genes that were differentially expressed between long-lived and young individuals (downregulated in LLI=703 genes and upregulated in LLI=497 genes). Of the identified *cis*-eQTLs, 10.7% were located either in an intron or in the untranslated regions (UTRs) of the genes of interest. The remaining 89.3% were located either upstream or downstream of the genes. Six of the most statistically significant *cis*-eQTLs identified in the 3' and 5' UTRs are depicted in Figure 3-8. The *cis*-eQTLs with the best p-values were strikingly concentrated in the close proximity (0.2 Mb) of the genes (Figure 3-6). It was observed that about 75.6% (908 genes of 1,200) of the expression traits had more than one statistically significant *cis*-eQTL (Figure 3-7). Furthermore, the *cis*-eQTLs that influenced the same expression trait were in high LD ($r^2 \geq 0.5$). All the 8,757 statistically significant *cis*-eQTLs are provided as an electronic attachment.

In order to ensure the quality of the identified *cis*-eQTLs, around 10% (870 of 8,757) of eQTL-hit-SNPs were selected randomly and their genotype cluster plots were subjected to visual inspection. Out of the 870 eQTL-hit-SNPs, only 14 SNPs (1.6%) had either missing genotypes or overlapping clusters of different genotypes and thus did not pass

the visual inspection, which is a statistically negligible proportion and emphasizes the good quality of the generated genotype data.

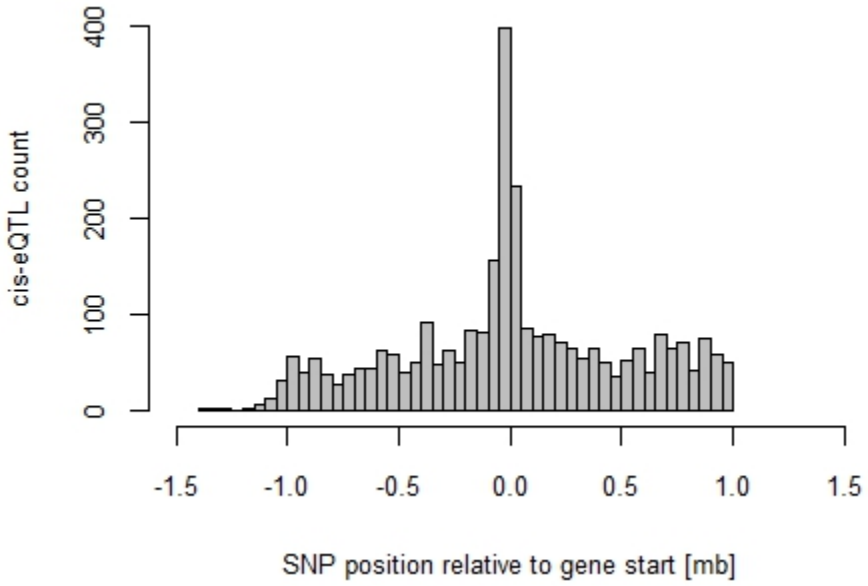

Figure 3-6 Spatial distribution of the best *cis*-eQTLs with best p-values relative to the gene start (in the German samples).

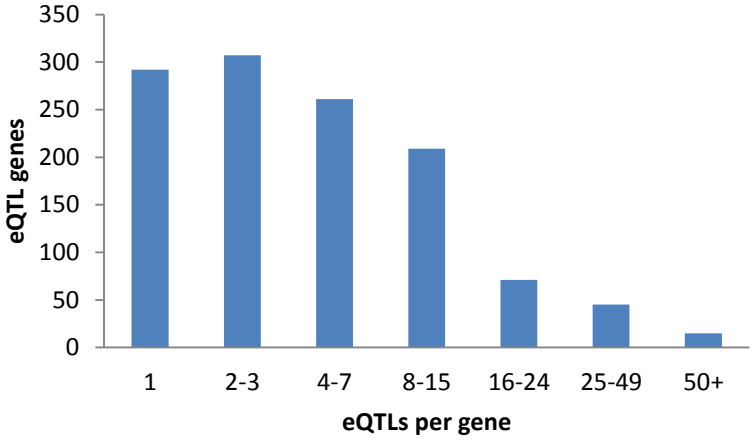

Figure 3-7 Histogram showing the number of *cis*-eQTLs per gene in the German sample.

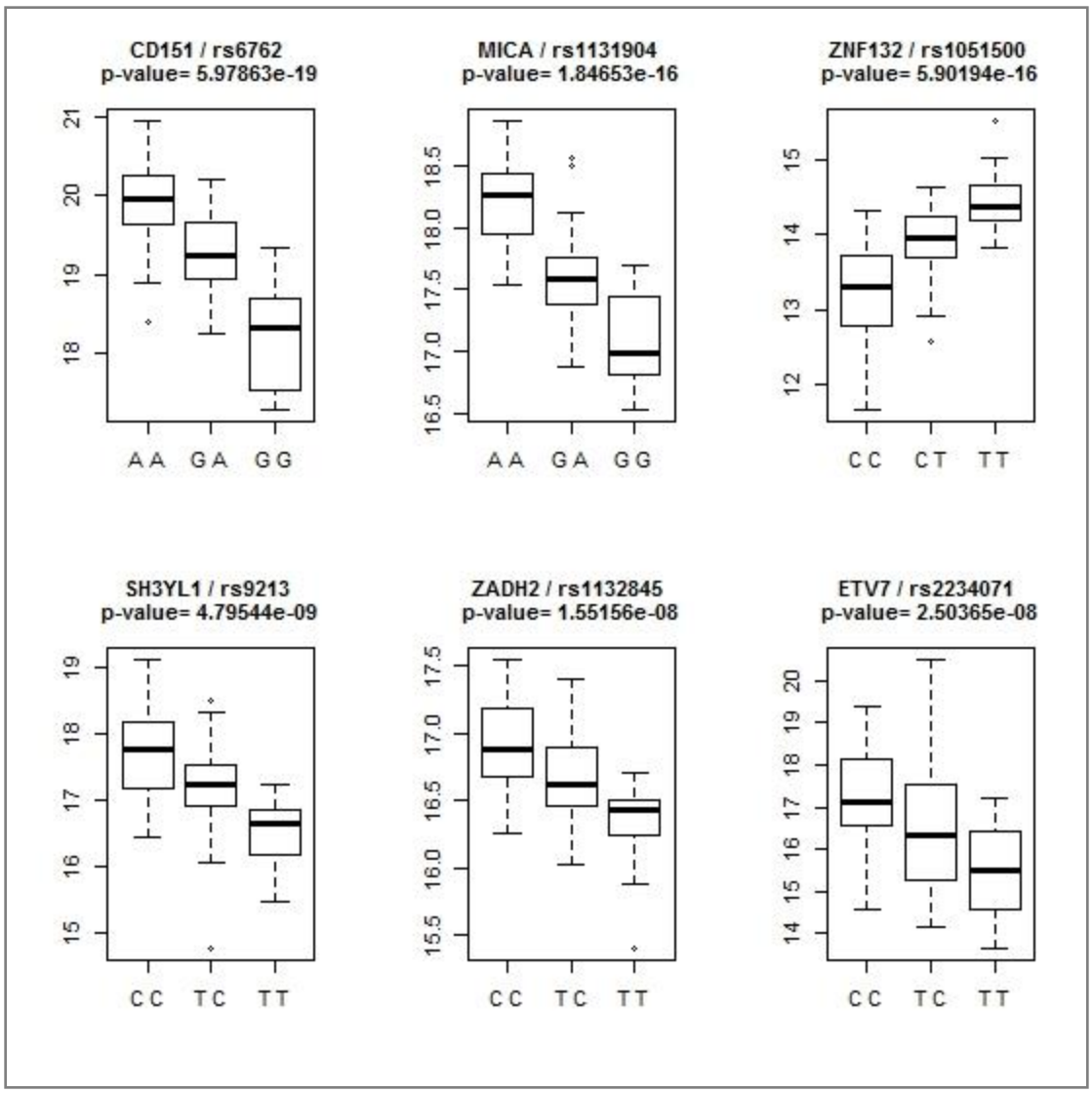

Figure 3-8 Box plot depiction of the 6 most significant *cis*-eQTLs identified in the 3' and 5' UTR regions. Five are 3' UTR variants and one (ETV7) is a 5' UTR variant; p-values are corrected for multiple testing; x-axis represents the genotypes and y-axis represents the log2 FPKM of the gene.

3.4.2 Genotype-age interaction effects on differentially expressed genes

The eQTL analysis identifies the genetic variants that influence the expression of a gene, but it cannot identify a change in the variants' influence on gene expression due to an interaction with age. Therefore, the contribution of the G×A interaction to the variation observed in differentially expressed genes was investigated. Of the 6,078 differentially

expressed genes, the expression of 1,200 was significantly influenced by *cis*-eQTLs (see 3.4.1). Among these 1,200, 908 genes exhibited significant association with more than one *cis*-eQTL. Thus, for each gene, only the *cis*-eQTLs with the best p-values were considered for this analysis. Of the 8,757 *cis*-eQTLs identified, 2,986 demonstrated the most significant *cis* associations (i.e., the best p-values). G×A interaction analysis focusing on these 2,986 *cis*-eQTLs identified 1,164 G×A interactions associated with 599 genes. All the identified G×A interactions are provided as an electronic attachment.

3.4.3 Contribution of *cis*-regulation and G×A interaction to biological processes

GO analysis of the differentially expressed genes identified 107 significant biological processes (see 3.2.3). These processes (GO terms) can be broadly classified into three categories namely, metabolic processes, immune and defense response, and cell and tissue regeneration. The investigation of *cis*-regulation and G×A interaction effects on the differentially expressed genes helped to elucidate the contribution of these factors to the biological processes. It was observed that about 20% of the genes involved in each of the identified processes were regulated by *cis*-variants, of which 50% exhibited G×A interaction effects. All the significant biological processes identified, and the contribution of *cis*-regulation and G×A interaction effects to each of these processes are presented in Table 3-3.

Table 3-3 Contribution of *cis*-regulation and genotype-age interaction effects to biological processes. Columns 2-5 categorize the processes into metabolic, immune and defense response, cell and tissue regeneration and others; %GeneticContribution indicates the percentage of genetic contribution to each of the processes; GA interaction = genotype-age interaction.

GO term name	metabolism	Immune and defense	cell and tissue regeneration	other	No. of genes in depleted processes	No. of genes in enriched processes	non-eQTL-contribution to depletion	eQTL(withoutGA)-contribution to depletion	GA-Interaction contribution to depletion	non-eQTL-contribution to enrichment	eQTL(withoutGA)-contribution to enrichment	GA-Interaction contribution to enrichment	%GeneticContribution
cellular macromolecule metabolic process	1	0	0	0	1497	0	1239	123	135	0	0	0	17
regulation of metabolic process	1	0	0	0	1887	0	1554	165	168	0	0	0	18
macromolecule metabolic process	1	0	0	0	1555	0	1280	128	147	0	0	0	18
nucleobase-containing compound metabolic process	1	0	0	0	1085	0	890	96	99	0	0	0	18
cellular aromatic compound metabolic process	1	0	0	0	1105	0	902	101	102	0	0	0	18
heterocycle metabolic process	1	0	0	0	1101	0	900	99	102	0	0	0	18
organic cyclic compound metabolic process	1	0	0	0	1125	0	914	105	106	0	0	0	19
cellular nitrogen compound metabolic process	1	0	0	0	1118	0	915	100	103	0	0	0	18
regulation of cellular process	0	0	0	1	2728	0	2242	241	245	0	0	0	18
cellular biosynthetic process	0	0	0	1	918	0	744	84	90	0	0	0	19
organic substance biosynthetic process	0	0	0	1	938	0	756	86	96	0	0	0	19
organelle organization	0	0	1	0	509	0	427	49	33	0	0	0	16
oxidation-reduction process	0	0	0	1	0	185	0	0	0	158	14	13	15
negative regulation of biological process	0	0	0	1	743	0	616	66	61	0	0	0	17
protein metabolic process	1	0	0	0	705	0	586	56	63	0	0	0	17
positive regulation of biological process	0	0	0	1	806	0	651	80	75	0	0	0	19
RNA transport	0	0	0	1	55	0	48	5	2	0	0	0	13

regulation of developmental process	0	0	1	0	139	0	106	15	18	0	0	0	24
cell-cell signaling	0	0	0	1	0	70	0	0	0	54	6	10	23
modification of morphology or physiology of other organism involved in symbiotic interaction	0	0	0	1	115	0	97	11	7	0	0	0	16
generation of precursor metabolites and energy	1	0	0	0	0	138	0	0	0	121	10	7	12
cell communication	0	0	0	1	0	99	0	0	0	79	7	13	20
modification of morphology or physiology of other organism	0	0	0	1	115	0	97	11	7	0	0	0	16
small molecule metabolic process	1	0	0	0	0	745	0	0	0	600	77	68	19
embryonic morphogenesis	0	0	1	0	19	0	14	2	3	0	0	0	26
immune response-activating signal transduction	0	1	0	0	91	0	81	5	5	0	0	0	11
interaction with host	0	0	0	1	115	0	98	10	7	0	0	0	15
embryo development	0	0	1	0	95	0	80	9	6	0	0	0	16
regulation of signaling	0	0	0	1	686	0	561	61	64	0	0	0	18
intracellular transport	0	0	0	1	267	0	223	25	19	0	0	0	16
cellular response to endogenous stimulus	0	0	0	1	249	0	201	25	23	0	0	0	19
pattern specification process	0	0	0	1	21	0	18	0	3	0	0	0	14
phosphorus metabolic process	1	0	0	0	352	0	295	33	24	0	0	0	16
cell cycle	0	0	1	0	162	120	132	17	13	101	11	8	17
organ morphogenesis	0	0	1	0	23	0	17	4	2	0	0	0	26
cell-cell adhesion	0	0	0	1	19	0	15	2	2	0	0	0	21
cell adhesion	0	0	0	1	64	0	48	10	6	0	0	0	25
cell migration	0	0	0	1	47	0	31	9	7	0	0	0	34
macromolecule methylation	0	0	0	1	44	0	35	4	5	0	0	0	20
regulation of response to stimulus	0	0	0	1	845	0	685	77	83	0	0	0	19
protein transport	0	0	0	1	256	0	212	27	17	0	0	0	17
ion transport	0	0	0	1	0	120	0	0	0	91	15	14	24
system process	0	0	0	1	0	129	0	0	0	93	23	13	28
organic acid metabolic process	1	0	0	0	0	125	0	0	0	92	16	17	26
single-organism catabolic process	0	0	0	1	0	75	0	0	0	60	8	7	20
cell cycle process	0	0	1	0	236	0	197	23	16	0	0	0	17
cellular localization	0	0	1	0	88	0	70	10	8	0	0	0	20
cellular response to stress	0	1	0	0	259	201	217	22	20	163	15	23	17
cellular response to chemical stimulus	0	0	0	1	364	0	295	38	31	0	0	0	19
Golgi vesicle transport	0	0	0	1	57	0	50	5	2	0	0	0	12

regulation of cellular component organization	0	0	1	0	312	0	256	29	27	0	0	0	18
gene silencing	0	0	0	1	26	0	19	4	3	0	0	0	27
viral process	0	1	0	0	156	0	130	14	12	0	0	0	17
symbiosis, encompassing mutualism through parasitism	0	1	0	0	156	0	130	14	12	0	0	0	17
tube morphogenesis	0	0	1	0	1	0	1	0	0	0	0	0	0
regulation of body fluid levels	0	0	0	1	47	0	37	5	5	0	0	0	21
cellular developmental process	0	0	1	0	211	0	152	25	34	0	0	0	28
tissue morphogenesis	0	0	1	0	22	0	19	2	1	0	0	0	14
antigen processing and presentation of exogenous antigen	0	1	0	0	0	44	0	0	0	38	5	1	14
mitochondrial transport	0	0	0	1	0	31	0	0	0	28	1	2	10
regulation of cellular component biogenesis	0	0	1	0	100	0	79	12	9	0	0	0	21
coagulation	0	1	0	0	39	0	30	5	4	0	0	0	23
organonitrogen compound metabolic process	1	0	0	0	0	199	0	0	0	156	22	21	22
single-organism biosynthetic process	0	0	0	1	0	47	0	0	0	38	3	6	19
establishment or maintenance of cell polarity	0	0	1	0	33	0	26	3	4	0	0	0	21
hematopoietic or lymphoid organ development	0	1	0	0	60	0	49	6	5	0	0	0	18
maintenance of location in cell	0	0	1	0	34	0	28	4	2	0	0	0	18
regulation of protein stability	0	0	0	1	38	0	37	0	1	0	0	0	3
antigen processing and presentation of peptide antigen	0	1	0	0	0	45	0	0	0	38	6	1	16
system development	0	0	0	1	56	0	39	7	10	0	0	0	30
regulation of RNA stability	0	0	0	1	19	0	18	0	1	0	0	0	5
stem cell maintenance	0	0	1	0	33	0	29	1	3	0	0	0	12
cellular response to abiotic stimulus	0	0	0	1	55	0	45	5	5	0	0	0	18
digestion	0	0	0	1	0	1	0	0	0	0	1	0	100
cellular macromolecule localization	0	0	1	0	66	0	52	8	6	0	0	0	21
cell differentiation	0	0	1	0	147	0	108	17	22	0	0	0	27
morphogenesis of a branching structure	0	0	1	0	7	0	6	1	0	0	0	0	14
xenobiotic metabolic process	1	0	0	0	0	13	0	0	0	9	2	2	31
positive regulation of molecular function	0	0	0	1	253	0	200	24	29	0	0	0	21
glutathione derivative metabolic process	1	0	0	0	0	11	0	0	0	7	4	0	36
regionalization	0	0	1	0	13	0	12	0	1	0	0	0	8

cellular component movement	0	0	1	0	123	0	94	13	16	0	0	0	24
regulation of blood pressure	0	0	0	1	0	8	0	0	0	6	1	1	25
response to other organism	0	1	0	0	0	74	0	0	0	56	11	7	24
organ development	0	0	1	0	102	0	75	16	11	0	0	0	26
hydrogen transport	0	0	0	1	0	18	0	0	0	18	0	0	0
protein localization	0	0	1	0	82	0	63	12	7	0	0	0	23
immune response-regulating cell surface receptor signaling pathway involved in phagocytosis	0	1	0	0	25	0	21	3	1	0	0	0	16
nuclear import	0	0	0	1	27	0	22	3	2	0	0	0	19
regulation of membrane potential	0	0	0	1	0	19	0	0	0	14	0	5	26
response to fibroblast growth factor stimulus	0	0	0	1	52	0	43	5	4	0	0	0	17
activation of innate immune response	0	1	0	0	44	0	41	1	2	0	0	0	7
synapse organization	0	0	0	1	4	0	2	1	1	0	0	0	50
extracellular structure organization	0	0	0	1	0	35	0	0	0	26	6	3	26
defense response to virus	0	1	0	0	0	32	0	0	0	26	4	2	19
signal transduction	0	0	0	1	448	0	362	36	50	0	0	0	19
cell death	0	0	0	1	192	0	165	18	9	0	0	0	14
regulation of immune system process	0	1	0	0	234	0	204	15	15	0	0	0	13
cellular response to stimulus	0	0	0	1	903	0	739	86	78	0	0	0	18
endosomal transport	0	0	0	1	42	0	36	4	2	0	0	0	14
response to virus	0	1	0	0	0	46	0	0	0	38	5	3	17
cerebral cortex development	0	0	1	0	14	0	8	3	3	0	0	0	43
viral entry into host cell via membrane fusion with the plasma membrane	0	1	0	0	0	3	0	0	0	2	0	1	33
tube development	0	0	1	0	2	0	2	0	0	0	0	0	0
cellular catabolic process	0	0	0	1	0	220	0	0	0	178	23	19	19
regulation of localization	0	0	0	1	308	0	253	31	24	0	0	0	18
regulation of multicellular organismal process	0	0	0	1	228	0	174	29	25	0	0	0	24

3.4.4 *Cis*-regulation of genes with heritable transcriptional activity

Cis-eQTL analysis was performed for all the 3,029 genes with heritable transcriptional activity. Of these genes, 1,248 exhibited a statistically significant association with 18,066 *cis*-eQTLs (see electronic attachment). Around 85% of the *cis*-regulated genes exhibited association with multiple *cis*-eQTLs (Figure 3-9). Furthermore, the *cis*-eQTLs with the most significant associations were concentrated in the close proximity (0.2 Mb) of the genes (Figure 3-10). The investigation of the proportion of *cis*-eQTLs in different heritability categories (h^2=0.3-0.39, 0.4-0.49, 0.5-0.59 and ≥0.6) showed that the genes with higher heritability estimates ($h^2 \geq 0.6$) associated with a larger proportion of *cis*-eQTLs. About 60.4% of the identified *cis*-eQTLs associated with the genes that had heritability estimates ≥0.6 (Figure 3-11).

The quality of these *cis*-eQTLs was examined by visually inspecting the genotype cluster plots of 10% (1,806 of 18,066) of eQTL-hit-SNPs. Out of the 1,806 eQTL-hit-SNPs, the cluster plots of only 40 SNPs (2.2%) did not pass the visual inspection i.e., had either missing genotypes or overlapping clusters of different genotypes.

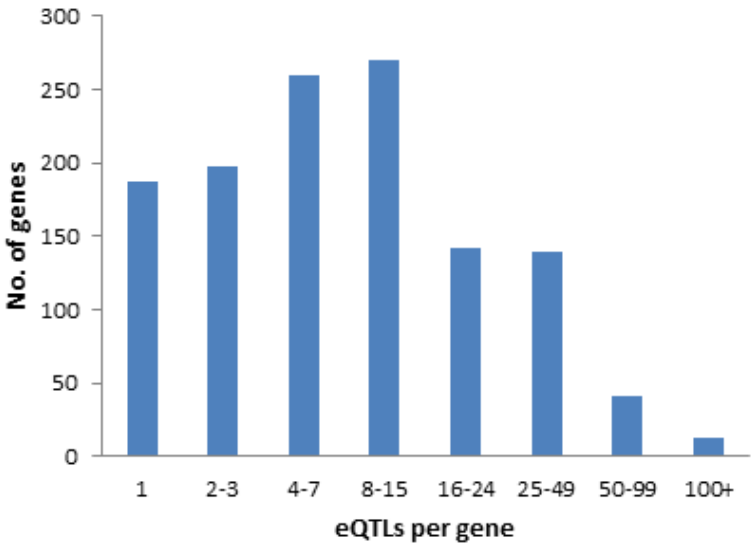

Figure 3-9 Histogram showing the number of eQTLs per gene in the Danish sample.

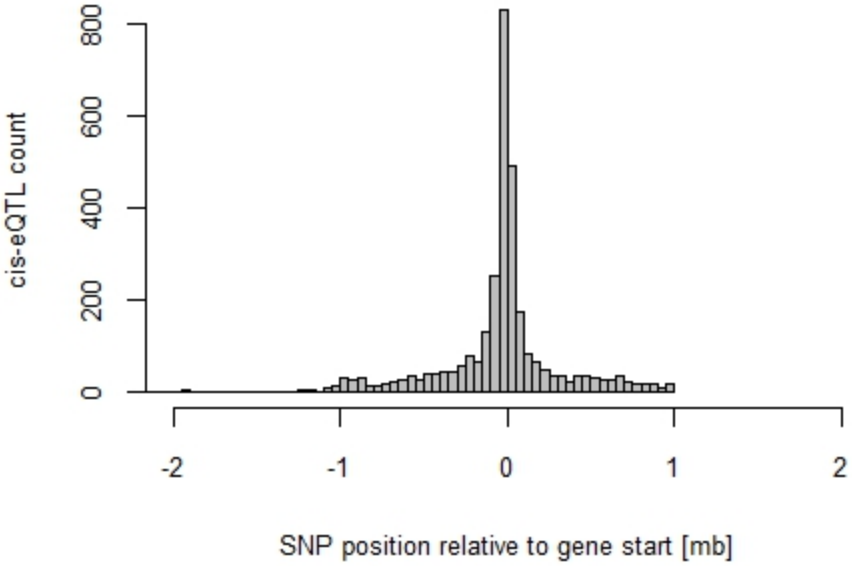

Figure 3-10 Spatial distribution of the best *cis*-eQTLs relative to the gene start position (in the Danish sample).

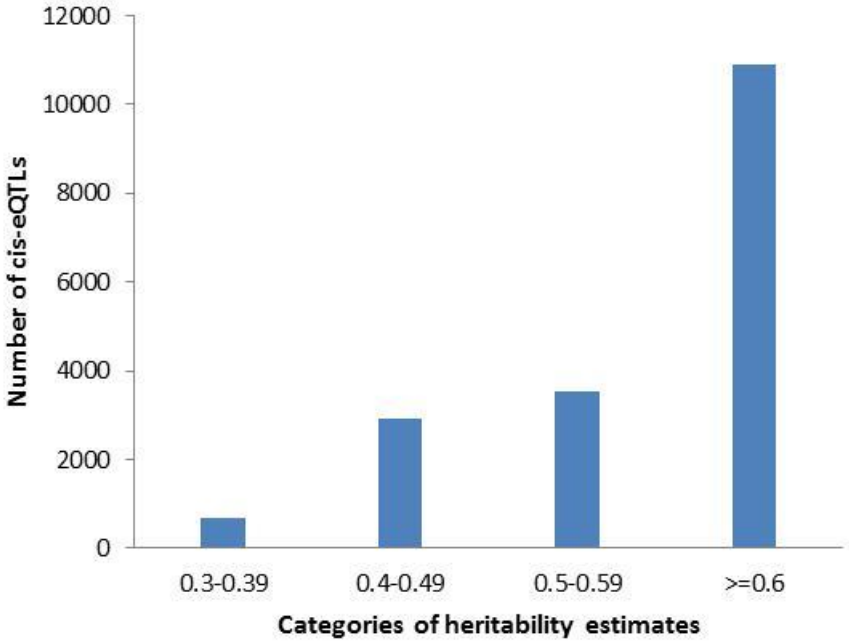

Figure 3-11 Proportion of *cis*-eQTLs in different heritability categories.

3.4.5 *Cis*-regulatory effects of known age-associated variants

The *APOE* genotypes that are determined from rs429358 and rs7412 as well as the *FOXO3A* variants namely, rs3800231, rs2764264 and rs4945816 have been repeatedly associated with longevity (Flachsbart et al. 2009; Lindahl-Jacobsen et al. 2013). Hence, the effect of these variants on their respective mRNA levels was investigated in the German as well as in the Danish sample. The eQTL analysis revealed that there is no significant *cis* association between these variants and the corresponding mRNA levels.

3.5 Biological validation using the Danish samples

In order to validate the findings from the differential expression analysis and the *cis*-eQTL analysis (from the German sample), the Danish samples were used. Since the Danish dataset consisted of twins, the validation set was created by selecting one individual from each twin pair (see 2.14). Additionally, 10 unrelated cases (83–92 years) and 10 unrelated controls (58–60 years) from Denmark were included in the validation set. Therefore, in total, 34 cases and 34 controls were employed for this analysis.

3.5.1 Differential expression analysis

Transcriptomes from the 68 Danish samples (cases=34 [83–92 years], controls=34 [58–60 years]) were used to investigate the 6,078 genes that were found to be differentially expressed in the German sample. Out of the 6,078 genes, 5,183 were regulated in the same direction as in the German dataset (i.e., exhibited the same trend), while 1,661 showed a Benjamini – Hochberg corrected p-value ≤ 0.05 and a FDR ≤ 5% (see electronic attachment). However, when the same cutoff criteria (Benjamini – Hochberg corrected p-value ≤ 0.001 and a FDR ≤ 0.1%) as in the German study were applied to the Danish

validation set, only 317 genes of the 6,078 were identified as differentially expressed and these genes showed the same trend as in the German dataset.

Since the validation dataset was smaller than the German sample, the genes with p-value ≤ 0.05 were considered for the subsequent gene ontology analysis. Interestingly, among the top 25 significantly represented biological processes, nine were metabolism-related and contained more downregulated genes than expected by chance as observed in the German sample. All the significantly represented processes are presented in the supplementary material (Table 8-5).

3.5.2 Expression quantitative trait loci analysis

Transcriptome and genotype data from the 68 Danish samples were used to validate the results obtained from the *cis*-eQTL analysis that was performed with the German dataset. Of the 68 samples, genotype data from four samples (two controls and two cases) had a sample call rate below 90% and were excluded from the analysis. Therefore, 32 cases (83–92 years) and 32 controls (58–60 years) were used for this analysis. Additionally, SNPs with a minor allele frequency ≤ 1%, call rate ≤ 95% and Hardy-Weinberg equilibrium testing p-value <=0.0001 were omitted. In total, 618,263 SNPs passed the quality criteria and were employed for the eQTL analysis.

Of the 1,200 genes which exhibited a significant association with *cis*-variants in the German sample, 374 genes exhibited significant *cis* association in the Danish dataset as well. *Cis*-eQTL analysis detected 3,332 statistically significant *cis*-eQTLs associated with these 374 genes. However, out of the 3,332 *cis*-eQTL signals, 2,017 were unique to this dataset. As observed in the German sample, the *cis*-eQTLs with the best p-values were concentrated in the close proximity (0.2 Mb) of the genes (Figure 3-12). All the *cis*-eQTLs identified in the Danish dataset are provided as an electronic attachment.

As previously described for the German sample and the Danish twins, the genotype cluster plots of 10% (333 of 3,332) of eQTL-hit-SNPs were subjected to visual inspection. Among the 333 eQTL-hit-SNPs, only 8 (2.4%) failed the visual inspection.

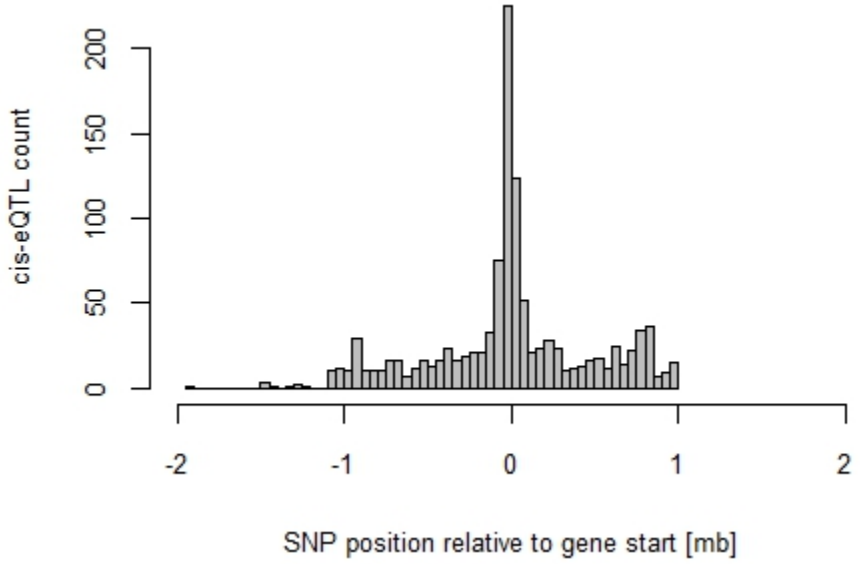

Figure 3-12 Spatial distribution of the best *cis*-eQTLs relative to the gene start position (in the Danish validation set).

3.6 Technical replication of selected genes by real-time qPCR

To replicate the results from the transcriptome sequencing, the expression levels of selected genes were measured by RT-qPCR in the German dataset. Since five samples (three young and two LLI) lacked RNA, the remaining 70 young individuals [20–55 years] and 53 LLI [91–102 years] were used for this analysis. Eighty genes (Table 3-5) were selected based on their ranks with respect to effect size (fold change), significance in the analysis between long-lived and young individuals, and regulation of their expression by *cis*-eQTLs. Additionally three control groups were selected, including (a) four endogenous control genes (Table 2-3); (b) five genes with no or low variation in expression levels among the samples (Table 3-4); and (c) five genes which are known or

presumed to have a genetic association with longevity, but were not differentially expressed between long-lived and young individuals (Table 2-3). Of these three control groups, three genes (*RER1*, *E2F4* and *BFAR*) from group (b) showed minimum variation in the expression levels across the samples and, therefore, the median of their expression levels was used to normalize the expression of the 80 candidate genes. Out of the 80, the signals of six genes turned out to be undetermined and therefore were not included in the analysis. From the remaining 74 genes, 50 were significantly differentially expressed between long-lived and young individuals (Benjamini – Hochberg corrected p-value ≤ 0.05; FDR ≤ 5%) and showed the same trend as in the transcriptome sequencing data (see supplementary material Table 8-6). Furthermore, permutation analysis revealed that, by chance, only 0.042 of the 74 genes could be expected to be differentially expressed. However, finding 50 genes as differentially expressed is more than expected by chance (Fisher's exact test p-value= 7.57×10^{-17}). The average correlation coefficient (between RNA-sequencing and RT-qPCR data) for the top 10 differentially expressed genes was 0.81, indicating that the transcriptome sequencing results were highly correlated with the RT-qPCR values. Figure 3-13 shows the correlation plot for one of the genes (ETV7).

Table 3-4 Five genes with no or low variation in expression levels in the German sample.

Gene symbol	Gene name	TaqMan assay ID
RER1	RER1 retention in endoplasmic reticulum 1 homolog (S. cerevisiae)	Hs00199824_m1
SENP5	SUMO1/sentrin specific peptidase 5	Hs00381410_m1
PRPF3	pre-mRNA processing factor 3	Hs00757030_m1
E2F4	E2F transcription factor 4, p107/p130-binding	Hs00608098_m1
BFAR	bifunctional apoptosis regulator	Hs00275423_m1

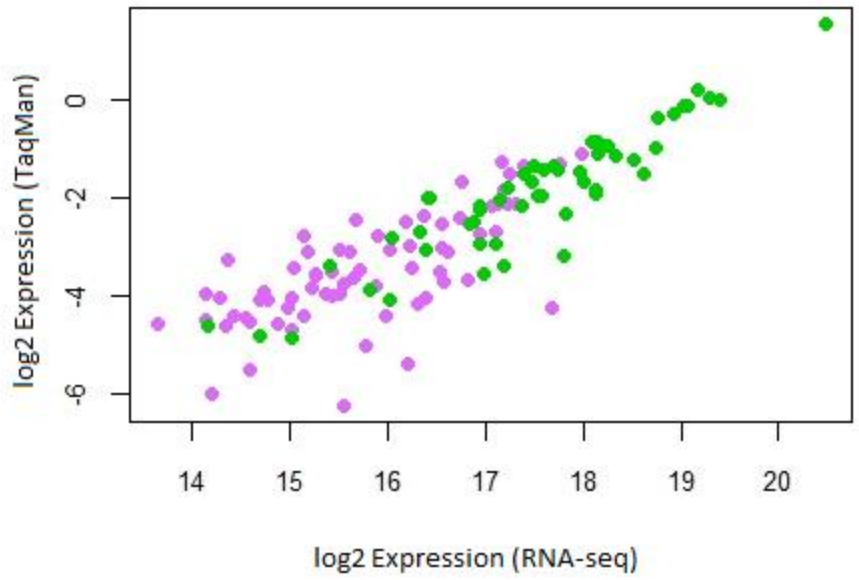

Figure 3-13 Correlation of expression levels between RNA-seq and TaqMan for the gene ETV7. Pink=young individuals, green=long-lived individuals.

Table 3-5 List of 80 candidate genes and their TaqMan assays used for validation.

Gene symbol	Gene name	TaqMan assay ID
APOL4	apolipoprotein L, 4	Hs00540930_m1
ETV7	ets variant 7	Hs00903229_m1
C1QC	complement component 1, q subcomponent, C chain	Hs00757779_m1
C8G	complement component 8, gamma polypeptide	Hs00167188_m1
HP	haptoglobin	Hs00605928_g1
NGFR	nerve growth factor receptor	Hs00609977_m1
SMTNL1	smoothelin-like 1	Hs00418171_m1
MMP23B	matrix metallopeptidase 23B	Hs04187882_g1
IFNG	interferon, gamma	Hs00989291_m1
SEPT4	septin 4	Hs00910208_g1
SATB2	SATB homeobox 2	Hs00392652_m1
FZD4	frizzled family receptor 4	Hs00201853_m1
ASGR2	asialoglycoprotein receptor 2	Hs00154160_m1
RNASE3	ribonuclease, RNase A family, 3	Hs01923184_s1
MDK	midkine (neurite growth-promoting factor 2)	Hs00171064_m1
DDB2	damage-specific DNA binding protein 2, 48kDa	Hs03044953_m1
BLVRA	biliverdin reductase A	Hs00167599_m1
TCN1	transcobalamin I (vitamin B12 binding protein, R binder family)	Hs01055542_m1

Gene symbol	Gene name	TaqMan assay ID
SPHK1	sphingosine kinase 1	Hs00184211_m1
LMCD1	LIM and cysteine-rich domains 1	Hs00205871_m1
LAP3	leucine aminopeptidase 3	Hs00429769_m1
C1QA	complement component 1, q subcomponent, A chain	Hs00706358_s1
IGFBP2	insulin-like growth factor binding protein 2, 36kDa	Hs01040719_m1
CLEC3B	C-type lectin domain family 3, member B	Hs00162844_m1
SARNP	SAP domain containing ribonucleoprotein	Hs00793492_m1
PRTN3	proteinase 3	Hs01597752_m1
NAALADL1	N-acetylated alpha-linked acidic dipeptidase-like 1	Hs00193839_m1
AMH	anti-Mullerian hormone	Hs00174915_m1
CTSL	cathepsin L	Hs00377632_m1
ZSCAN9	zinc finger and SCAN domain containing 9	Hs00196838_m1
PRSS23	protease, serine, 23	Hs00359912_m1
NT5C3B	5'-nucleotidase, cytosolic IIIB	Hs00369454_m1
MARCO	macrophage receptor with collagenous structure	Hs00198935_m1
CCL23	chemokine (C-C motif) ligand 23	Hs00270756_m1
CXCL10	chemokine (C-X-C motif) ligand 10	Hs01124251_g1
TBKBP1	TBK1 binding protein 1	Hs00207037_m1
CTSK	cathepsin K	Hs00166156_m1
CTSG	cathepsin G	Hs00175195_m1
LXN	latexin	Hs00220138_m1
ELF3	E74-like factor 3 (ets domain transcription factor, epithelial-specific)	Hs00963881_m1
FEZ1	fasciculation and elongation protein zeta 1 (zygin I)	Hs00192714_m1
SFRP5	secreted frizzled-related protein 5	Hs00169366_m1
BATF2	basic leucine zipper transcription factor, ATF-like 2	Hs00293780_m1
ETS1	v-ets avian erythroblastosis virus E26 oncogene homolog 1	Hs00428293_m1
SATB1	SATB homeobox 1	Hs00161515_m1
SPTBN1	spectrin, beta, non-erythrocytic 1	Hs00162271_m1
GZMH	granzyme H (cathepsin G-like 2, protein h-CCPX)	Hs00277212_m1
PLAG1	pleiomorphic adenoma gene 1	Hs00231236_m1
ZNF229	zinc finger protein 229	Hs00970997_m1
HIVEP2	human immunodeficiency virus type I enhancer binding protein 2	Hs00198801_m1
AXIN2	axin 2	Hs00610344_m1
LRRN3	leucine rich repeat neuronal 3	Hs01087516_m1
ZNF154	zinc finger protein 154	Hs01068902_m1
ZNF135	zinc finger protein 135	Hs00987427_g1
TSPYL5	TSPY-like 5	Hs00603217_s1
PRKCA	protein kinase C, alpha	Hs00925193_m1
MAN1C1	mannosidase, alpha, class 1C, member 1	Hs00220595_m1
PDGFRB	platelet-derived growth factor receptor, beta polypeptide	Hs01019589_m1

Gene symbol	Gene name	TaqMan assay ID
ZNF551	zinc finger protein 551	Hs00292939_m1
ANK3	ankyrin 3, node of Ranvier (ankyrin G)	Hs00253210_m1
SERPINB10	serpin peptidase inhibitor, clade B (ovalbumin), member 10	Hs00192370_m1
RAPGEF3	Rap guanine nucleotide exchange factor (GEF) 3	Hs00183449_m1
RNASE4	ribonuclease, RNase A family, 4	Hs00377763_m1
DYNLT1	dynein, light chain, Tctex-type 1	Hs00831821_s1
MAPK8IP1	mitogen-activated protein kinase 8 interacting protein 1	Hs00271363_m1
ELANE	elastase, neutrophil expressed	Hs00975994_g1
CYB5R2	cytochrome b5 reductase 2	Hs00212055_m1
CAPG	capping protein (actin filament), gelsolin-like	Hs00156249_m1
GSTM4	glutathione S-transferase mu 4	Hs00426432_m1
NUDT18	nudix (nucleoside diphosphate linked moiety X)-type motif 18	Hs00368715_m1
CCL4	chemokine (C-C motif) ligand 4	Hs00237011_m1
NRG1	neuregulin 1	Hs00247620_m1
PLA2G16	phospholipase A2, group XVI	Hs00912734_m1
ATF3	activating transcription factor 3	Hs00231069_m1
NR1H3	nuclear receptor subfamily 1, group H, member 3	Hs00172885_m1
ARG1	arginase 1	Hs00968979_m1
PPP1R3G	protein phosphatase 1, regulatory subunit 3G	Hs03805942_s1
FAM109B	family with sequence similarity 109, member B	Hs03025388_s1
MDP1	magnesium-dependent phosphatase 1	Hs04194678_s1
HOXB7	homeobox B7	Hs00270131_m1

4 Discussion

The objective of this study was to identify the transcriptome patterns that are associated with the complex phenotype of human longevity and investigate the effects of genetic variation on those patterns. Comparing the RNA-sequencing-based transcriptome profiles (from blood) of LLI with those of young controls revealed thousands of genes that were differentially expressed. This data was subsequently integrated with genotype information from the same individuals, resulting in the identification of several genes that are regulated by *cis*-variants. About 50% of the *cis*-regulated genes exhibited a G×A interaction, illustrating that the genetic influence on mRNA expression changes with age. Additionally, a heritability analysis employing monozygotic (MZ) and dizygotic (DZ) twins identified thousands of genes with heritable transcriptional activity. To place the differentially expressed genes, *cis*-regulated genes and genes with heritable transcriptional activity in a biological context, a GO analysis was performed. It highlighted three broad functional categories, namely metabolic processes, immune and defense responses and cell and tissue regeneration.

4.1 Adaptation of the eQTL study design for longevity research

Human longevity is a complex phenotype that is influenced by both genetic and environmental factors, including gene-environment interactions (Tan et al. 2002). Moreover, in contrast to most complex diseases, human longevity is a very heterogeneous phenotype. Genetic approaches in longevity research have thus far provided very limited insights into the determinants and mechanisms underlying longevity. Therefore, researchers have focused on alternative methods to identify longevity-associated genes and the biological processes functionally linked to these genes. One such approach has been to perform transcriptome studies (Rodwell et al. 2004; Zahn et al. 2006; Hong et al. 2008; Harries et al. 2011; Passtoors et al. 2012;

Marttila et al. 2013). A majority of transcriptome studies in longevity research use cross-sectional designs. Using these designs, it is not possible to distinguish between changes in gene expression that cause or are consequences of longevity.

Natural genetic variation affects mRNA expression levels and thereby impacts molecular and physiological phenotypes (Ackermann et al. 2013). Since DNA variants do not change over time for a given individual, integrating genotype information with gene expression data can help to explain the age-associated variation in gene expression. Therefore, in the present study, an eQTL analysis was employed to investigate the association between genetic variants and mRNA expression levels. Considering the above-mentioned limitations of a cross-sectional study design, this approach provides an opportunity to determine the primary (genetic) basis of gene regulation and to gain insight into otherwise unknown genotype-phenotype correlations.

4.2 Differentially expressed genes in blood between long-lived individuals and young controls

In this thesis, RNA-sequencing-based transcriptome profiling was employed to investigate differentially expressed genes between long-lived and young individuals. Contrary to previous transcriptomic studies using blood samples, a large number of genes were found to be differentially expressed between the two groups. Out of 19,991 genes that were expressed in more than 5% of the German sample, 6,078 (~30%) were significantly differentially expressed after filtering for blood cell-specific genes. Among these differentially expressed genes, 44% were upregulated and 56% were downregulated in LLI. This observed imbalance between up- and downregulated genes is consistent with previous longevity studies (Tan et al. 2008; Harries et al. 2011; Passtoors et al. 2012). Additionally, the expression levels of 178 genes exhibited a statistically significant gender effect. Interestingly, of the 178 genes, 17 were X-linked, which was significantly more than expected by chance (p-value = 0.00038). Moreover,

these X-linked genes showed the most significant gender effect, in agreement with previous findings (Tan et al. 2008).

Previous blood transcriptome studies in longevity research have identified very few genes (around 2–7%) associated with age (Harries et al. 2011; Passtoors et al. 2012). All previous transcriptome studies in the field of longevity have employed microarray technology. Here, using an RNA-sequencing-based approach yielded a substantially larger proportion of differentially expressed genes (~30%). This large proportion can be attributed to the ability of RNA-seq to quantify highly abundant as well as rare transcripts (i.e., high dynamic range) and to the large number of individuals included in this study. The power to detect significantly differentially expressed genes increases 1) with the accuracy and dynamic range of a measuring method and 2) with the sample size (here, 128 individuals). Identifying a higher number of differentially expressed genes using RNA-seq in comparison to microarrays is in line with the literature (Marioni et al. 2008; Su et al. 2011).

The most statistically significant differentially expressed gene was leucine-rich repeat neuronal 3 (*LRRN3)*, which showed a 9.1-fold decreased expression in LLI at a significance level of $p=9.03 \times 10^{-17}$. *LRRN3* is involved in the activation of MAPK activity and endocytosis and, interestingly, it has been observed as the top candidate in several other longevity studies (Hong et al. 2008; Harries et al. 2011; Passtoors et al. 2012). Similarly, other genes such as endosialin (*CD248)*, NEL-like 2 (*NELL2*), actin binding LIM protein 1 (*ABLIM1*), myobrevin *(VAMP5*) and chemokine (C-C motif) receptor 7 (*CCR7*) identified as top candidates in this study were also reported by other researchers (Harries et al. 2011; Passtoors et al. 2012). This concordance may serve as proof-of-principle and confirms the validity of the approach employed in the present study. This might also indicate that, in contrast to genetic association findings, observations on a functional level have a higher validation rate and thus are more commonly shared between different study populations.

In the present study, the mechanistic target of rapamycin (*MTOR)* gene, which encodes a serine/threonine kinase, was downregulated in LLI compared to young individuals. The downregulation of *MTOR* is associated with lifespan extension in model organisms and its increased activity in humans has been associated with a spectrum of cancers (Guertin and Sabatini 2005). Additionally, the v-myc avian myelocytomatosis viral oncogene homolog *(MYC)*, which encodes the transcription factor MYC, was downregulated in LLI. Approximately 50% of human cancers exhibit overexpression of MYC, which is negatively correlated with patient survival (Tansey 2014). The genes *MTOR* and *MYC* were reported to be downregulated in LLI in previous studies as well (Passtoors et al. 2012). Taken together, the observed downregulation of *MTOR* might indicate a lower cancer risk, yet this lower mRNA expression was not associated with particular genetic variants in the present study.

Many other factors are known to influence human longevity, including alterations to nutrient-sensing pathways such as TOR or IIS. A recent study investigated the expression of 40 genes that belong to the well-described core of the mTOR pathway and reported that seven were differentially expressed between LLI and control individuals (Passtoors et al. 2013). Five of these seven genes were found to be differentially expressed (Benjamini–Hochberg corrected p-value ≤ 0.001; FDR ≤ 0.1%) in the present study as well. This concordance further supports the validity of the approach chosen here. Finally, the identification of a large number of differentially expressed genes emphasizes the notion that longevity is a complex phenotype influenced by many genes and biological processes.

4.3 Age-associated gene regulation influenced by *cis*-variants

The eQTL analysis was used in this study to determine the genetic basis of age-associated gene regulation. Of all the 6,078 differentially expressed genes, 19.74% (1,200 of 6,078) had at least one significant *cis*-eQTL. In line with previous eQTL studies,

genetic variation with strong impact on gene expression were predominantly found in close proximity to transcription start sites (Innocenti et al. 2011; Stranger et al. 2012; Bryois et al. 2014). Furthermore, many expression traits were associated with multiple *cis*-eQTLs and *cis*-eQTLs associated with the same expression trait were in high LD ($r^2 \geq 0.5$) as previously reported (Kabakchiev and Silverberg 2013). Taking into account that genes under genetic control were influenced by seven variants on average in the German sample, these findings further support the hypothesis of a complex underlying molecular pattern, where different variants may lead to the same phenotype.

Although eQTL analyses are a common approach to identify variants that influence gene expression, they do not reveal if the regulation of gene expression changes with age. Therefore, investigating G×A interactions can provide further insight into the effect of age on the influence of genetic variation on mRNA levels. More than 400 G×A interactions have previously been reported in human skin and adipose tissues (Glass et al. 2013) and more than 600 have been reported in blood (Kent et al. 2012). In the present study, when focusing only on the most significant *cis*-eQTLs for each gene, it was observed that 9.9% of the differentially expressed genes (368 that were downregulated and 231 that were upregulated in LLI) exhibited a G×A interaction. This is in agreement with a previous longevity study performed in blood which reported that 13.94% of the differentially expressed genes showed G×A interactions (Kent et al. 2012). Interestingly, the most highly differentially expressed gene, *LRRN3* did not exhibit a G×A interaction in this study, but it was found to have a significant G×A interaction by Kent et al. (2012). In addition, the *UBQLNL* gene, whose expression and *cis* genotype are associated with a longitudinal cancer risk and display a G×A interaction (Kent et al. 2012), was not differentially expressed in the present study. These discrepancies may be due to the sample-specific differences, especially the different age ranges of individuals included in the two studies.

Overall, age-associated genes were partially influenced by genetic factors and the observation of G×A interaction effects on gene expression indicates that the impact of genetic variants on mRNA expression changes with age. However, the cross-sectional study design used here does not allow to address if the observed G×A interaction effects on mRNA expression are a life-long phenomenon or does it occur only at advanced age.

4.4 Influence of known *APOE* and *FOXO3A* variants on gene expression in blood

Several genetic studies have been performed to identify the determinants that influence the complex longevity phenotype. Although many putative loci associated with longevity have been reported, variations in only two genes have been consistently replicated thus far. These two genes are *APOE* (Schächter et al. 1994; Zhang et al. 1998; Blanché et al. 2001) and *FOXO3A* (Willcox et al. 2008; Flachsbart et al. 2009; Li et al. 2009). *APOE* is known to play an important role in regulating lipoproteins (Smith 2002; Christensen et al. 2006) whereas *FOXO3A* encodes the evolutionarily conserved transcription factor forkhead box protein O 3A (FOXO3A), a homolog of daf-16 in *C. elegans*. It is well described that the transcription factor encoded by *daf-16* functions as a key regulator in the IIS pathway and influences metabolism and lifespan in model organisms (Ogg et al. 1997; Giannakou et al. 2004).

Although variations in *APOE* and *FOXO3A* have been consistently associated with longevity, the influence of these variants on mRNA levels remains to be elucidated. In this study, two *APOE* (rs429358 and rs7412) and three *FOXO3A* (rs3800231, rs2764264 and rs4945816) genotypes that are associated with longevity were investigated to evaluate their effects on *APOE* and *FOXO3A* expression, respectively. The results indicated that these genotypes have no significant effect on their respective gene expression in blood. It is not known whether these genotypes influence the gene expression in other tissues. The major sites for *APOE* expression are liver and brain

(Elshourbagy et al. 1985), whereas *FOXO3A* is expressed in the major insulin responsive tissues such as liver, adipocytes and skeletal muscle (Banasik et al. 2011). The genotype-dependent expression in the relevant tissues remains to be examined. Considering that the current study analyzed *cis*-associations only, it is likely that the *APOE* and *FOXO3A* variants may influence transcripts located in *trans*. Identifying such *trans*-interactions, however, is a challenging task, as the power to detect such effects requires substantially larger sample numbers.

4.5 Large proportion of age-associated genes are regulated independent of *cis*-variants

In this study, thousands of genes were identified as differentially expressed between long-lived and young individuals. Of the differentially expressed genes, only 19.74% were identified to be significantly regulated by *cis*-variants and only 9.9% were identified to exhibit G×A interaction. In contrast, more than 80% of the differentially expressed genes did not exhibit significant association with *cis*-variants. It is likely that the expression of these genes is influenced by *trans*-variants or possibly by environmental factors, which were not investigated in this thesis. The impact of non-genetic factors is supported by the heritability analysis performed in this study that has revealed that the variation in gene expression for a large proportion of genes can be explained by environmental rather than by genetic factors. The individual contribution of genetic and non-genetic factors remains unclear unless *trans*-interactions have been investigated in detail. Finally, assessing the nature and contribution of non-genetic factors to human longevity may be one of the most challenging tasks in this research field, as it is next to impossible to perform longitudinal studies with human subjects in controlled environments.

4.6 Transcriptional activity as a heritable trait

Phenotypic variation among individuals, such as differences in gene expression, is a natural phenomenon that occurs in normal conditions or can be caused by disease states. Several studies indicate that a considerable proportion of variation in gene expression can be attributed to genetic factors, and therefore, is heritable (Schadt et al. 2003; Cheung et al. 2003; Tan et al. 2005). In addition to genome-wide significant eQTLs, which are considered heritable, other polygenic variation may exist that fails to reach statistical significance by normal genotype-expression association analysis. The classical twin design, comparing resemblance in MZ and DZ twin pairs, provides distinct advantages in the efficiency of heritability estimation (van Dongen et al. 2012; Wright et al. 2014). In the present study, the gene expression profiles of 96 twins (44 MZ and 52 DZ) were examined employing structural equation modeling to quantify the heritability of gene expression in blood. Analysis of 15,640 genes expressed in at least 95% of the samples facilitated to estimate the extent to which genetic factors contribute to the variation in gene expression.

The analysis demonstrated that there was a significant genetic component underlying the regulation of gene expression in blood. The expression levels of 19.36% (3,029 of 15,640) of genes were under significant genetic control. This is in contrast to a recent publication that reported a proportion of heritable genes of 4.2% (777 of 18,392) in blood (Wright et al. 2014). However, a study performed with human lymphocytes has reported a much higher proportion (85%) of transcripts under genetic control (Göring et al. 2007). This variability in the proportion of genes exhibiting heritable expression demonstrates that gene expression levels should be considered with caution when used as traits because an expression profile is a snapshot of the cellular state and is tissue-specific. The extent of genetic control (estimated as narrow-sense heritability) on individual genes in the current study ranged from $h^2 = 0.30$ to $h^2 = 0.99$. Though 19.36% of the genes were under significant genetic regulatory control, only 6.2% of genes had

considerably high expression heritability estimates ($h^2 \geq 0.6$). Since the analysis was performed with corrections for age and gender, it indicates that non-genetic factors influence the expression of the majority of genes.

The 3,029 genes identified as being under significant genetic control were further subjected to eQTL analysis to identify *cis*-eQTLs that have significant influence on expression. It was observed that 1,248 genes had at least one significant eQTL after correcting for multiple testing (Benjamini and Hochberg's method, FDR ≤ 5%). When investigating the proportion of *cis*-eQTLs in different heritability categories, it was revealed that the genes with high heritability estimates ($h^2 \geq 0.6$) were likely to have a larger proportion of *cis*-eQTLS than genes with low heritability estimates. Furthermore, the genes with at least one significant *cis*-eQTL had significantly higher heritability (P-value = 2.6×10^{-15}), which is in concordance with previous studies (Wright et al. 2014). Taken together, the heritability of gene expression and the proportion of *cis*-eQTLs associated with heritable genes illustrate the role of genetic control in biological processes.

4.7 Prominent depletion of metabolic processes in long-lived individuals

The relationship between metabolism and longevity has been a topic of interest for more than a century. In 1908, the German Physiologist Max Rubner discovered a relationship between longevity, metabolic rate and body size. He described that larger animals live longer and have slower metabolic rates than smaller animals. Thus, he proposed that longevity is inversely proportional to metabolic rate (Rubner 1908). Subsequently, the American biologist Raymond Pearl expanded Rubner's work and proposed the "rate-of-living" hypothesis, which suggests a direct link between metabolic output and longevity in eukaryotes (Pearl 1928). Three decades later, Denham Harman formulated the widely accepted oxidative damage model (Harman

1956). This model proposes that reactive oxygen species (ROS) are produced during normal metabolism, and they subsequently damage important biological molecules (DNA, proteins and lipids), and eventually cause senescence. It has been suggested that an increased metabolic rate promotes increased production of ROS and thus accelerates senescence (Beckman and Ames 1998). In this context, caloric restriction is claimed to extend lifespan by reducing metabolic rate, and therefore ROS production rate (Braeckman et al. 2006). However, in 2004, the mathematician and biologist Lloyd Demetrius suggested that the most important factor involved in lifespan is not metabolic rate or oxidative stress, but metabolic stability, which is the capacity of an organism to maintain steady state concentration of metabolites. He proposed that caloric restriction prolongs lifespan by increasing metabolic stability (Demetrius 2004).

Based on these theories, several studies have been performed in model organisms as well as in humans to elucidate the link between metabolism and longevity. Interestingly, an inverse relationship between metabolism and longevity has not been observed in populations of *Drosophila melanogaster* with different lifespans (van Voorhies et al. 2003; Khazaeli et al. 2005). In *C. elegans*, though reduced metabolic rates have been observed in several long-lived mutants (daf-2 and age-1 mutants) (Van Voorhies and Ward 1999; Van Raamsdonk et al. 2010; Lee et al. 2012;), there have been contradictory results as well (Braeckman et al. 2002). In humans, it has been reported that a high metabolic rate is a risk factor for natural mortality (Ruggiero et al. 2008; Jumpertz et al. 2011). Furthermore, individuals (40–96 years) who are fully functional and free of major medical conditions have significantly lower resting metabolic rates than those with disease and functional impairments (Schrack et al. 2014). As mitochondria are semiautonomous organelles producing most of the ROS and metabolic rate is a mitochondrial-associated parameter (Lehmann et al. 2008), researchers have investigated the relative amount of mitochondrial DNA (mtDNA) in LLI. A study by Van Leeuwen et al. (2014) demonstrated that LLI exhibit lower mtDNA content than their offspring. Furthermore, the offspring of LLI also have reduced mtDNA content compared

to their partners. These results may indicate that I) metabolic rate as an inherited trait impacts longevity or II) reduced mtDNA serves as an indicator of a passive lifestyle (e.g., reduced physical activity and moderate food intake). However, these findings still need to be replicated. Taken together, the results of these studies do not clarify whether a reduced metabolic rate predisposes individuals to longevity or is a consequence of advanced age.

In this study, though metabolic rates were not measured, a prominent depletion of metabolic processes in LLI was observed on the transcriptome level. The GO analysis of differentially expressed genes revealed that this depletion in several metabolism-related processes was the result of the downregulation of a large number of metabolism-associated genes. This number was considerably higher than expected by chance, implying a significant depletion of metabolic processes in LLI.

As described in section 1.3.1, there is substantial evidence from model organisms that the reduced activity of the IIS pathway not only causes an increase in lifespan but also induces metabolic changes (Kimura et al. 1997). In this study, genes involved in the IIS pathway such as insulin-like growth factor 1 receptor (*IGF1R*), insulin-like growth factor 2 receptor (*IGF2R*), phosphatidylinositol-3-phosphate 5-kinase (*PIP5K3/PIKFYVE*), phosphoinositide 3-kinase regulatory subunit 4 (*PIK3R4*), protein kinase C, alpha (*PRKCA*), protein kinase A catalytic subunit beta (*PRKACB*) and interaction protein for cytohesin exchange factors 1 (*IPCEF1*) were downregulated in LLI, in concordance with a previous study (Abdul Rahman et al. 2013). The key transcription factor of the IIS pathway, FOXO3A, which is one of the homologues of *daf-16* in *C. elegans*, was not differentially regulated between LLI and young controls. However, another homologue of *daf-16*, forkhead box O1 (*FOXO1*), which plays a significant role in regulating whole body energy metabolism (Gross et al. 2008), was downregulated in LLI. These observations further support the link between longevity and low metabolic rate (Rubner 1908) and indicate that depletion of metabolism may be associated with longevity. A GO

analysis based on transcriptome data generally only shows an association between a phenotype and a molecular profile. In contrast, the eQTL approach used here revealed that the depletion of metabolic processes is at least a partially causative factor underlying the longevity phenotype. Consequently, this study presents molecular data supporting a potentially causative link between metabolism and longevity in humans on the transcriptome level.

4.8 Remodeling of immune and defense responses in long-lived individuals

A popular concept among researchers is that advanced age is associated with deterioration of the immune system. However, Franceschi et al. (1995) proposed that most immune parameters undergo a continuous remodeling with age, where some parameters increase, some decrease and others remain unchanged, rather than a unidirectional deterioration. In this study, among the immune related genes that were differentially expressed, several were upregulated and several others were downregulated in LLI. The genes, such as chemokine (C-C motif) receptor 7 (*CCR7*), chemokine (C-C motif) receptor 6 (*CCR6*), CD27 molecule (*CD27*), lymphotoxin-beta (*LTB*) and endosialin (*CD248*) were downregulated in LLI, which is in agreement with other studies (Hong et al. 2008; Vo et al. 2010; Harries et al. 2011; Marttila et al. 2013). Several genes involved in antigen processing and presentation were upregulated in the elderly. These genes include cathepsin H (*CTSH*), cathepsin L1 (*CTSL1),* major histocompatibility complex, class II, DR alpha (*HLA-DRA*) and major histocompatibility complex, class II, DP beta 1 (*HLA-DPB1*). One of the major characteristics associated with advanced age is chronic low-grade inflammation (inflamm-ageing) (Franceschi et al. 2000a), which arises from continual antigenic stress. Inflammatory genes, such as interleukin-6 signal transducer (*IL6ST*) and CD40 ligand (*CD40LG*) were downregulated in LLI as reported previously (Abdul Rahman et al. 2013). In contrast to a previous study

(Vo et al. 2010), tumor necrosis factor superfamily member 1A (*TNFRSF1A*) was downregulated in LLI. Additionally, anti-inflammatory cytokines like interleukin-10 (*IL10*) and interleukin-18 (*IL18*) were upregulated. Overall, the changes observed in LLI on the transcriptome level indicate that immune-related processes are not altered in single direction over time, but appear to be continuously remodeled as proposed by Franceschi. This remodeling starts at birth and ensures an on-going adaptation of the immune system to the dynamically changing requirements of different ages.

According to the network theory, the ageing process is indirectly controlled by a variety of defense and anti-stress responses (Franceschi et al. 2000a). There are various stressors including physical (heat, UV and gamma radiation), chemical (reducing sugars, products of metabolism such as oxygen-free radicals) and biological (viruses, bacteria) agents. In the present study, several differentially expressed genes were observed to be associated with the GO terms such as defense response to virus and bacteria (GO:0051607, GO:0042742), cellular response to stress (GO:0033554) and response to other organisms (GO:0051707). In this context, some heat shock proteins such as heat shock 70 kDA protein 8 (*HSPA8*), heat shock 70 kDA protein 9 (*HSPA9*) and heat shock 90 kDa protein 1, and beta (*HSP90AB1*) were found to be downregulated in LLI. Heat shock 70 kDA protein 7 (*HSPA7*) and the genes encoding the members of the peroxiredoxin family of antioxidant enzymes, *PRDX1* and *PRDX5* were upregulated in the elderly. These findings are supported by previous studies demonstrating that overexpression of *PRDX5* in *Drosophila* increases resistance to oxidative stress and extends their lifespan by up to 30% under normal conditions (Radyuk et al. 2009), and *PRDX1* has been reported to play an important role in defense against oxidants in ageing mice (Neumann et al. 2003).

One of the prominent events associated with age is thymic involution, the progressive shrinking of the thymus (Lynch et al. 2009). As a consequence, production of naïve T-cells declines with age, while the proportion of memory T-cells increases (Lynch et al. 2009). It has been observed that older individuals are more susceptible to infectious diseases than younger population (Bender 2003). For instance, influenza is a common

infection in the elderly and is associated with increased morbidity (Gavazzi and Krause 2002). Vaccination can prevent infection and protect the vulnerable elderly population from such infectious diseases. However, the age-associated changes in the immune system as mentioned above affect the efficacy of vaccines in the elderly (Aspinall et al. 2007). Most vaccines are less efficient in aged individuals because of changes in the immune system (Weinberger and Grubeck-Loebenstein 2012). Although direct functional inferences cannot be drawn from the results of the current study, the observed alterations in genes related to immune processes might play a role in the reduced efficacy of vaccines.

4.9 Depletion of cell and tissue regeneration in long-lived individuals

With age, there is a gradual decline in the homeostatic and regenerative capacity of all tissues and organs (Kirkwood 2005; Liu and Rando 2011). The age-related changes in cellular components caused by the accumulation of random damage in somatic cells, oxidative damage and other mechanisms lead to physiological changes and to impaired responses of tissues to injury (Conboy and Rando 2005). The reduced responses to injury are exemplified by poor wound healing of skin in older individuals (Gosain and DiPietro 2004). The age-associated decreases in tissue regenerative responses can be attributed to a decreased number of adult stem cells and decline in their ability to engage in tissue repair (Pekovic and Hutchison 2008).

Repair mechanisms play a crucial role in cell regeneration and the maintenance of genomic stability, and are consequently considered central to longevity (Kenyon and Gerson 2007). Although several repair-related processes such as DNA repair (GO:0006281), mismatch repair (GO:0006298), nucleotide-excision repair (GO:0006289) and base-excision repair (GO:0006284) are defined GO terms, they were not significantly represented by the differentially expressed genes. Therefore, the results of

the present study do not allow drawing conclusions about the involvement of repair in longevity.

In contrast, many more downregulated genes than expected by chance were associated with several cellular processes such as cell cycle, organelle organization, cellular localization, stem cell maintenance, regulation of cellular component organization and maintenance of cell polarity. On the other hand, far fewer upregulated genes than expected by chance were associated with processes such as regulation of developmental process, tissue and organ morphogenesis, cell differentiation and cellular component movement. Together, these findings imply that these processes are significantly depleted in LLI, further supporting the hypothesis of an age-associated impaired regenerative capacity.

4.10 Longevity-associated molecular patterns are driven by genetic and environmental factors

Twin studies have suggested that the longevity phenotype is influenced by both genetic and environmental factors. In this study, a similar observation was made on the transcriptome level, where the variation in gene expression was found to be influenced by both of these factors. Of the differentially expressed genes, 19.74% were under the influence of *cis*-variants, while the remaining genes were potentially controlled by *trans*-variants and/or non-genetic components, like environmental factors. The heritability analysis using a classical twin design further confirmed that the expression variation in approximately 20% of the genes could be explained by genetics, while for 80% of the genes the expression variation is explained by environmental influences.

The functional categorization of the differentially expressed genes led to the identification of three broad categories: metabolic processes, immune and defense response and cell and tissue regeneration. The investigation of the effects of *cis*-

regulation and G×A interactions on gene expression provided an opportunity to determine the contribution of these factors to the biological processes associated with longevity. Around 20% of the genes involved in each of the processes were regulated by *cis*-interactions. Consequently, the majority of genes and processes may be driven by *trans*-interactions and/or environmental factors. Figure 4-1 depicts the biological processes significantly represented and the contribution of genetic factors to each of them. Furthermore, the heritability analysis illustrated that for most genes under genetic control, only part of their variation at the transcript level is explained by genetics. This emphasizes that genetic contribution to longevity might still be substantially influenced by non-genetic factors, thus increasing the heterogeneity in the results and lowering the power of such studies. At the same time, this observation encourages further studies to focus on functional aspects of the molecular phenotype of longevity.

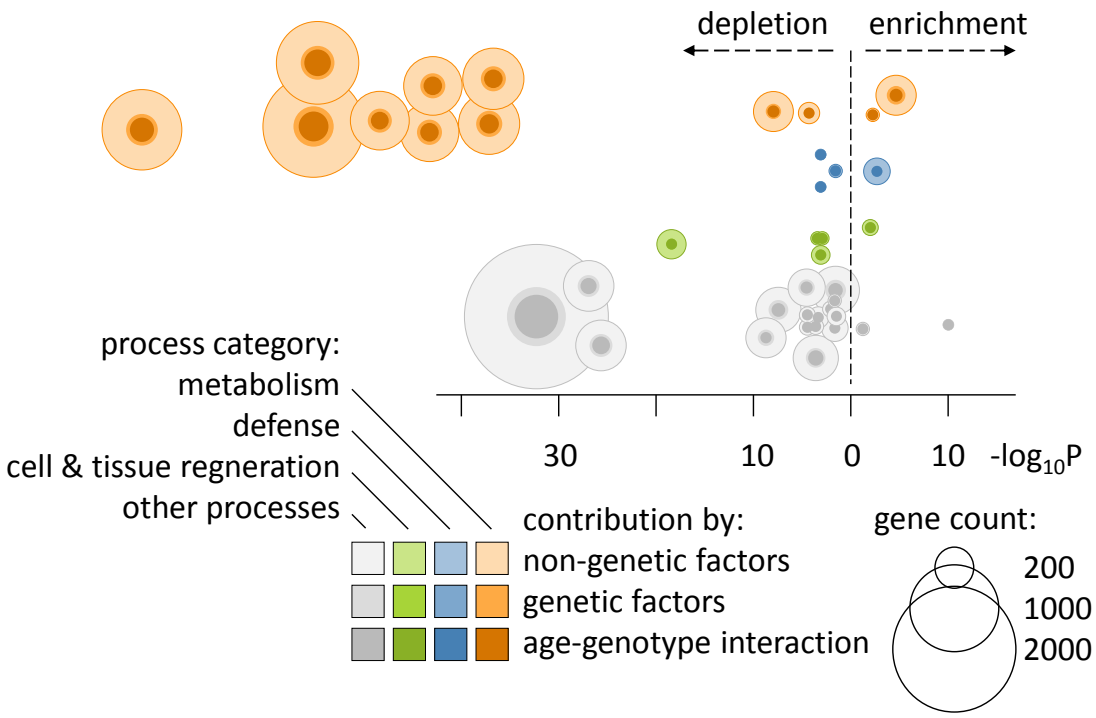

Figure 4-1 The contribution of genetic and non-genetic factors to the biological processes associated with the differentially expressed genes. Circles illustrate different biological processes, color coded by process category. The number of genes involved in each process is reflected by the area of the circle.

4.11 A large number of variants contribute to the same longevity-associated biological processes

In the present study, the eQTL analysis identified 8,757 significant *cis*-eQTLs associated with 1,200 genes, indicating that genes were regulated by multiple variants; on average, each gene was regulated by seven variants (in the German sample). These observations suggest that a large number of heterogeneous genetic components influence longevity-associated biological processes, which finally may lead to the same phenotype, longevity (Figure 4-2). Consequently, the same molecular phenotype (e.g., the depletion of metabolic processes) can result from multiple different genetic variants. The scenario created by this hypothesis could explain how different population-specific architectures

lead to a shared longevity phenotype without the same underlying genetic factors. These findings, therefore, partially explain why the replication of longevity associations is a challenging task.

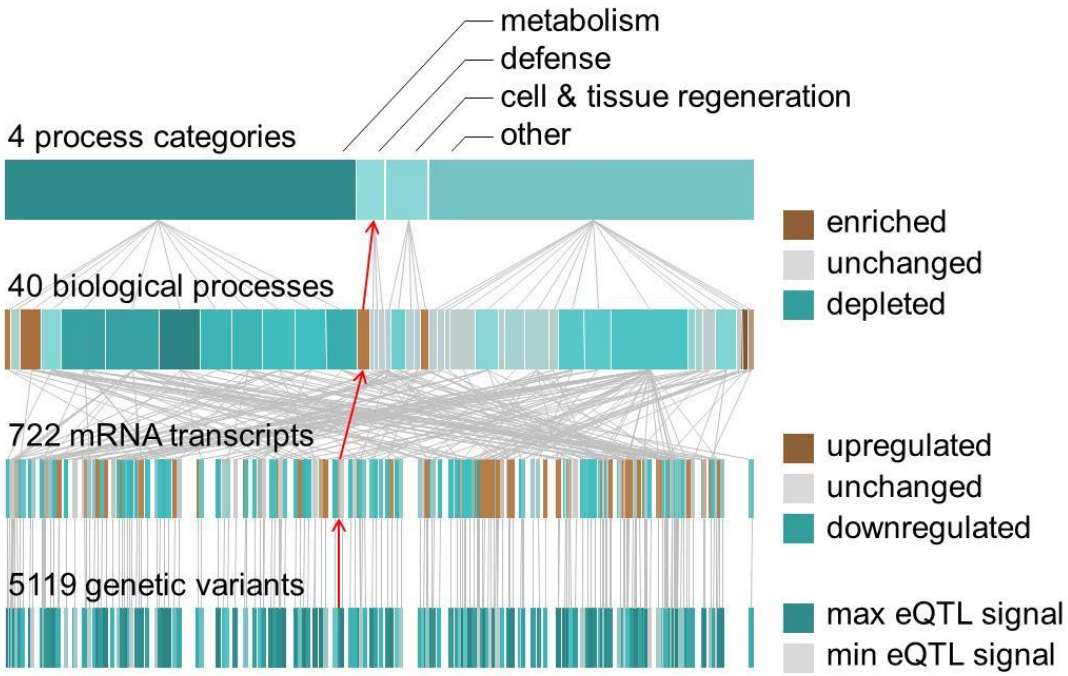

Figure 4-2 Hierarchical organization of eQTL effects on longevity

4.12 Transcriptome patterns associated with normal ageing vs. longevity

In the present study, thousands of genes were differentially expressed between the two age groups, long-lived and young individuals. The changes observed in the mRNA expression levels may be due to the normal ageing processes or may reflect the healthy ageing as the individuals investigated were exceptionally old (91–104 years). However, the cross-sectional approach employed here does not distinguish between changes that are the result of normal ageing and those that contribute to healthy ageing and

predispose individuals to the longevity phenotype. A longitudinal study, in which individuals from a given birth cohort are followed over time, may help discriminate between these changes. Additionally, since the controls used in cross-sectional designs originate from birth cohorts separate from those of the LLI, cohort-specific differences might confound the observations. However, using a longitudinal approach, individuals from a specific birth cohort are followed over time, controlling for cohort-specific differences. The findings from longitudinal studies however are very specific to the investigated cohort. While the cross-sectional design has a limited ability to demonstrate causality, results obtained from longitudinal studies are not broadly applicable across birth cohorts, and therefore, caution is necessary when making inferences regarding the causal effects on longevity.

4.13 Conclusion and outlook

In conclusion, RNA-sequencing-based genome-wide expression and eQTL analysis documented the depletion of metabolic processes, which is potentially causative for longevity in humans. Moreover, only a small proportion of this depletion of metabolism was explained by genetic contributions, further supporting the influence of non-genetic factors like environment. In addition, the remodeling of immune and defense response and depletion of cell and tissue regeneration were observed in LLI. Finally, this study provides a model to explain why longevity associations are difficult to replicate in alternative populations: in a hierarchical manner, a large number of different genetic variants can contribute to the expression of several genes, which in turn influences only a small set of biological processes. Consequently, many different genetic variants can lead to the same molecular outcome.

Although the age-associated genes were controlled by both genetic and non-genetic factors, the individual contribution of these factors remains unclear and requires further

investigation. It is evident from the literature that gene expression is regulated by epigenetic processes (Gibney and Nolan 2010), microRNA (miRNA) (Cannell et al. 2008) as well as natural genetic variation (Williams et al. 2007). Among the epigenetic modifications, DNA methylation is the best known and has a crucial role in the regulation of gene expression (Aguilera et al. 2010). Additionally, there is a growing body of evidence highlighting the role of DNA methylation and miRNAs in healthy ageing and longevity (ElSharawy et al. 2012; Heyn et al. 2012). Further investigation of the DNA methylation status and miRNAs in the samples used here could provide an even more detailed understanding of the molecular patterns associated with longevity.

5 Summary (English)

The demographic change as a result of continuously increasing life expectancy will be a major challenge for society, economy and health care. Besides other approaches addressing these challenges, understanding the molecular mechanisms of human longevity can contribute to extend the health span of the elderly and reduce the risk of disabilities and diseases. As human longevity is influenced by genetic and environmental factors, this study aims to monitor the functional consequences of both, employing an expression quantitative trait loci (eQTL) approach based on RNA-sequencing and genotype data. Blood samples from 244 individuals (age range, 20–104 years; 148 unrelated individuals and 96 twins) were analyzed in order to 1) investigate the transcriptome patterns associated with human longevity; 2) assess the influence of natural genetic variation on the identified transcriptome patterns; 3) investigate the heritability of transcriptional activity. Approximately 20% of the 6,078 genes that were differentially expressed between long-lived individuals (LLI) and young controls were influenced by *cis*-variants, and 50% of those *cis*-regulated genes exhibited genotype-age interaction effects. Consequently, gene regulation is influenced by both genetic and non-genetic factors. This observation was also supported by the subsequent heritability analysis carried out in twins. Interestingly, functional categorization of the differentially expressed genes revealed that metabolism-associated processes were significantly depleted in LLI, supporting the hypothesis that reduced metabolic rates contribute to human longevity. The eQTL approach employed here showed that the depletion of metabolic processes is potentially causative for longevity, yet the genetic contribution of about 20% indicates that non-genetic factors play a key role. Furthermore, the findings illustrate that different genetic variations can lead to the same molecular outcome, explaining the difficulties in validating genetic association findings in longevity research. Taken together, the presented results can contribute to the understanding of longevity, which is a starting point when addressing the challenging tasks resulting from the massive demographic changes in our society.

6 Summary (German)

Die kontinuierlich steigende Lebenserwartung führt zu einem demographischen Wandel, der Gesellschaft, Wirtschaft und Gesundheitssystem vor enorme Herausforderungen stellt. Um langfristig gesundes Altern zu ermöglichen, sowie um Krankheitsrisiken in älteren Bevölkerungsgruppen zu verringern, ist ein Verständnis der molekularen Mechanismen der Langlebigkeit dringend erforderlich. Da Langlebigkeit bei Menschen sowohl durch genetische Faktoren als auch durch Umweltfaktoren beeinflußt wird, war es Ziel der vorliegenden Studie, die funktionellen Konsequenzen beider Einflüsse abzubilden. Hierzu wurde ein sogenannter eQTL-Ansatz gewählt, der Sequenzierungsbasierte Trankriptomdaten mit Genotypisierungdaten verbindet. Blutproben von 244 Individuen (Altersspanne, 20–104 Jahre; 148 unverwandte Individuen und 96 Zwillinge) wurden eingesetzt um 1) die altersassoziierten Transkriptommuster zu untersuchen; 2) um den Einfluß genetischer Variation auf die identifizierten Transkriptommuster zu ermitteln und 3) um die Vererbbarkeit von Transkriptommustern darzustellen. Etwa 20% der 6,078 altersabhängig regulierten Gene unterlagen einer genetischen Kontrolle durch *cis*-Varianten, von denen wiederum etwa 50% einen Interaktionseffekt zwischen Alter und Genotyp zeigten. Daraus konnte abgeleitet werden, daß regulierte Gene sowohl durch genetische als auch durch nicht genetische Faktoren beeinflußt werden. Diese Beobachtung wurde zusätzlich durch eine nachfolgende Heritabilitätsanalyse bestätigt. Das zentrale Ergebnis der vorliegenden Arbeit ergab sich durch die funktionelle Kategorisierung der regulierten Gene: Mit Metabolismus verknüpfte biologische Prozesse zeigten in langlebigen Individuen eine signifikante Depletion. Dies ist im Einklang mit der Vermutung, daß reduzierte metabolische Raten zur Langlebigkeit beitragen. Der hier gewählte eQTL-Ansatz ermöglicht zwar, eine potentielle Kausalität dieser Depletion für Langlebigkeit zu folgern, jedoch deutet der genetische Beitrag von 20% auf einen überwiegenden Anteil nicht-genetischer Einflüsse hin. Der Befund, daß verschiedene genetische Varianten zum selben molekularen Ergebnis führen können, erklärt darüber hinaus die anhaltenden

Probleme, genetische Assoziationen in der Altersforschung zu validieren. Abschließend können die hier vorgelegten Ergebnisse zum Verständniss der Langlebigkeit beitragen, und stellen somit eine wichtige Voraussetzung dar, um sich den Herausforderungen in Folge des massiven demographischen Wandels zu stellen.

7 References

Abbott MH, Murphy EA, Bolling DR, Abbey H. 1974. The familial component in longevity. A study of offspring of nonagenarians. II. Preliminary analysis of the completed study. *Johns Hopkins Med J* **134**: 1–16.

Abdul Rahman A, Abdul Karim N, Abdul Hamid NA, Harun R, Ngah WZW. 2013. Senescence-related changes in gene expression of peripheral blood mononuclear cells from octo/nonagenarians compared to their offspring. *Oxid Med Cell Longev* **2013**: 189129.

Ackermann M, Sikora-Wohlfeld W, Beyer A, Wells CA. 2013. Impact of natural genetic variation on gene expression dynamics. *PLoS Genet* **9**: e1003514.

Aguilera O, Fernández AF, Muñoz A, Fraga MF. 2010. Epigenetics and environment: a complex relationship. *J Appl Physiol* **109**: 243–251.

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Automat Contr* **19**.

Andersen SL, Sebastiani P, Dworkis DA, Feldman L, Perls TT. 2012. Health span approximates life span among many supercentenarians: Compression of morbidity at the approximate limit of life span. *Journals Gerontol - Ser A Biol Sci Med Sci* **67 A**: 395–405.

Andrews S. 2010. FastQC: A quality control tool for high throughput sequence data. *babraham Bioinforma* 1.

Anselmi CV, Malovini A, Roncarati R, Novelli V, Villa F, Condorelli G, Bellazzi R, Puca AA. 2009. Association of the FOXO3A locus with extreme longevity in a southern Italian centenarian study. *Rejuvenation Res* **12**: 95–104.

Aspinall R, Del Giudice G, Effros RB, Grubeck-Loebenstein B, Sambhara S. 2007. Challenges for vaccination in the elderly. *Immun Ageing* **4**: 9.

Banasik K, Ribel-Madsen R, Gjesing AP, Wegner L, Andersson A, Poulsen P, Borglykke A, Witte DR, Pedersen O, Hansen T, et al. 2011. The FOXO3A rs2802292 G-allele associates with improved peripheral and hepatic insulin sensitivity and increased skeletal muscle-FOXO3A mRNA expression in twins. *J Clin Endocrinol Metab* **96**: E119–24.

Bao J-M, Song X-L, Hong Y-Q, Zhu H-L, Li C, Zhang T, Chen W, Zhao S-C, Chen Q. 2014. Association between FOXO3A gene polymorphisms and human longevity: a meta-analysis. *Asian J Androl* **16**: 446–52.

Beckman KB, Ames BN. 1998. The free radical theory of aging matures. *Physiol Rev* **78**: 547–581.

Beekman M, Nederstigt C, Suchiman HED, Kremer D, van der Breggen R, Lakenberg N, Alemayehu WG, de Craen AJM, Westendorp RGJ, Boomsma DI, et al. 2010. Genome-wide association study (GWAS)-identified disease risk alleles do not compromise human longevity. *Proc Natl Acad Sci U S A* **107**: 18046–18049.

Bender BS. 2003. Infectious disease risk in the elderly. *Immunol Allergy Clin North Am* **23**: 57–64.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B* **57**: 289 – 300.

Bergman A, Atzmon G, Ye K, MacCarthy T, Barzilai N. 2007. Buffering mechanisms in aging: A systems approach toward uncovering the genetic component of aging. *PLoS Comput Biol* **3**: 1648–1656.

Blanché H, Cabanne L, Sahbatou M, Thomas G. 2001. A study of French centenarians: are ACE and APOE associated with longevity? *C R Acad Sci III* **324**: 129–135.

Braeckman BP, Demetrius L, Vanfleteren JR. 2006. The dietary restriction effect in C. elegans and humans: Is the worm a one-millimeter human? *Biogerontology* **7**: 127–133.

Braeckman BP, Houthoofd K, Vanfleteren JR. 2002. Assessing metabolic activity in aging Caenorhabditis elegans: concepts and controversies. *Aging Cell* **1**: 82–88; discussion 102–103.

Bryois J, Buil A, Evans DM, Kemp JP, Montgomery SB, Conrad DF, Ho KM, Ring S, Hurles M, Deloukas P, et al. 2014. Cis and Trans Effects of Human Genomic Variants on Gene Expression ed. C.D. Brown. *PLoS Genet* **10**: e1004461.

Cannell IG, Kong YW, Bushell M. 2008. How do microRNAs regulate gene expression? *Biochem Soc Trans* **36**: 1224–1231.

Cheng P, Cheng Y, Li Y, Zhao Z, Gao H, Li D, Li H, Zhang T. 2012. Comparison of the gene expression profiles between smokers with and without lung cancer using RNA-Seq. *Asian Pac J Cancer Prev* **13**: 3605–9.

Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen K-Y, Morley M, Spielman RS. 2003. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet* **33**: 422–425.

Christensen K, Doblhammer G, Rau R, Vaupel JW. 2009. Ageing populations: the challenges ahead. *Lancet* **374**: 1196–1208.

Christensen K, Johnson TE, Vaupel JW. 2006. The quest for genetic determinants of human longevity: challenges and insights. *Nat Rev Genet* **7**: 436–448.

Conboy IM, Rando TA. 2005. Aging , Stem Cells and Tissue Regeneration ND ES SC CE. 407–410.

Danielle Simmons. 2007. the use of animal models in studying genetic disease: trangenesis and induced mutation. *Nat Cell Biol* **9**: 993–9.

de Magalhaes JP, Toussaint O. 2004. GenAge: a genomic and proteomic network map of human ageing. FEBS letters **571**: 243-247.

Deelen J, Beekman M, Capri M, Franceschi C, Slagboom PE. 2013. Identifying the genomic determinants of aging and longevity in human population studies: Progress and challenges. *BioEssays* **35**: 386–396.

Deelen J, Beekman M, Uh H-W, Broer L, Ayers KL, Tan Q, Kamatani Y, Bennet AM, Tamm R, Trompet S, et al. 2014. Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Hum Mol Genet*.

Deelen J, Beekman M, Uh HW, Helmer Q, Kuningas M, Christiansen L, Kremer D, van der Breggen R, Suchiman HED, Lakenberg N, et al. 2011. Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited. *Aging Cell* **10**: 686–698.

Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**: 3207–3212.

Demetrius L. 2004. Caloric restriction, metabolic rate, and entropy. *J Gerontol A Biol Sci Med Sci* **59**: B902–B915.

Donlon TA, Curb JD, He Q, Grove JS, Masaki KH, Rodriguez B, Elliott A, Willcox DC, Willcox BJ. 2012. FOXO3 gene variants and human aging: Coding variants may not be key players. *Journals Gerontol - Ser A Biol Sci Med Sci* **67**: 1132–1139.

ElSharawy A, Keller A, Flachsbart F, Wendschlag A, Jacobs G, Kefer N, Brefort T, Leidinger P, Backes C, Meese E, et al. 2012. Genome-wide miRNA signatures of human longevity. *Aging Cell* **11**: 607–16.

Elshourbagy N a, Liao WS, Mahley RW, Taylor JM. 1985. Apolipoprotein E mRNA is abundant in the brain and adrenals, as well as in the liver, and is present in other peripheral tissues of rats and marmosets. *Proc Natl Acad Sci U S A* **82**: 203–7.

Engberg H, Oksuzyan A, Jeune B, Vaupel JW, Christensen K. 2009. Centenarians - A useful model for healthy aging? A 29-year follow-up of hospitalizations among 40000 Danes born in 1905. *Aging Cell* **8**: 270–276.

Evert J, Lawler E, Bogan H, Perls T. 2003. Morbidity profiles of centenarians: survivors, delayers, and escapers. *J Gerontol A Biol Sci Med Sci* **58**: 232–237.

Fillman SG, Cloonan N, Catts VS, Miller LC, Wong J, McCrossin T, Cairns M, Weickert CS. 2012. Increased inflammatory markers identified in the dorsolateral prefrontal cortex of individuals with schizophrenia. *Mol Psychiatry*.

Flachsbart F, Caliebe A, Kleindorp R, Blanché H, von Eller-Eberstein H, Nikolaus S, Schreiber S, Nebel A. 2009. Association of FOXO3A variation with human longevity confirmed in German centenarians. *Proc Natl Acad Sci U S A* **106**: 2700–5.

Franceschi C, Bonafè M. 2003. Centenarians as a model for healthy aging. *Biochem Soc Trans* **31**: 457–461.

Franceschi C, Bonafè M, Valensin S, Olivieri F, De Luca M, Ottaviani E, De Benedictis G. 2000a. Inflamm-aging. An evolutionary perspective on immunosenescence. *Ann N Y Acad Sci* **908**: 244–254.

Franceschi C, Monti D, Sansoni P, Cossarizza A. 1995. The immunology of exceptional individuals: the lesson of centenarians. *Immunol Today* **16**: 12–16.

Franceschi C, Motta L, Valensin S, Rapisarda R, Franzone A, Berardelli M, Motta M, Monti D, Bonafè M, Ferrucci L, et al. 2000b. Do men and women follow different trajectories to reach extreme longevity? Italian Multicenter Study on Centenarians (IMUSCE). *Aging (Milano)* **12**: 77–84.

Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, et al. 2010. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* **42**: 1118–1125.

Friedman DB, Johnson TE. 1988. A mutation in the age-1 gene in Caenorhabditis elegans lengthens life and reduces hermaphrodite fertility. *Genetics* **118**: 75–86.

Garatachea N, Emanuele E, Calero M, Fuku N, Arai Y, Abe Y, Murakami H, Miyachi M, Yvert T, Verde Z, et al. 2014. ApoE gene and exceptional longevity: Insights from three independent cohorts. *Exp Gerontol* **53**: 16–23.

Gavazzi G, Krause KH. 2002. Ageing and infection. *Lancet Infect Dis* **2**: 659–666.

Gems D, McElwee JJ. 2003. Ageing: Microarraying mortality. *Nature* **424**: 259–261.

Gerdes LU, Jeune B, Ranberg K a, Nybo H, Vaupel JW. 2000. Estimation of apolipoprotein E genotype-specific relative mortality risks from the distribution of genotypes in centenarians and middle-aged men: apolipoprotein E gene is a “frailty gene,” not a “longevity gene”. *Genet Epidemiol* **19**: 202–10.

Giannakou ME, Goss M, Jünger MA, Hafen E, Leevers SJ, Partridge L. 2004. Long-lived Drosophila with overexpressed dFOXO in adult fat body. *Science* **305**: 361.

Gibney ER, Nolan CM. 2010. Epigenetics and gene expression. *Heredity (Edinb)* **105**: 4–13.

Glass D, Viñuela A, Davies MN, Ramasamy A, Parts L, Knowles D, Brown A a, Hedman AK, Small KS, Buil A, et al. 2013. Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome Biol* **14**: R75.

Göring HHH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JBM, Abraham LJ, Rainwater DL, Comuzzie AG, et al. 2007. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* **39**: 1208–1216.

Gosain A, DiPietro LA. 2004. Aging and Wound Healing. *World J Surg* **28**: 321–326.

Gross DN, van den Heuvel a PJ, Birnbaum MJ. 2008. The role of FoxO in the regulation of metabolism. *Oncogene* **27**: 2320–36.

Gudmundsson H, Gudbjartsson DF, Frigge M, Gulcher JR, Stefánsson K. 2000. Inheritance of human longevity in Iceland. *Eur J Hum Genet* **8**: 743–749.

Guertin DA, Sabatini DM. 2005. An expanding role for mTOR in cancer. *Trends Mol Med* **11**: 353–61.

HARMAN D. 1956. Aging: a theory based on free radical and radiation chemistry. *J Gerontol* **11**: 298–300.

Harries LW, Hernandez D, Henley W, Wood AR, Holly AC, Bradley-Smith RM, Yaghootkar H, Dutta A, Murray A, Frayling TM, et al. 2011. Human aging is characterized by focused changes in gene expression and deregulation of alternative splicing. *Aging Cell* **10**: 868–878.

Harrison DE, Strong R, Sharp ZD, Nelson JF, Astle CM, Flurkey K, Nadon NL, Wilkinson JE, Frenkel K, Carter CS, et al. 2009. Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature* **460**: 392–395.

Herskind AM, McGue M, Holm N V, Sørensen TI, Harvald B, Vaupel JW. 1996. The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870-1900. *Hum Genet* **97**: 319–323.

Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, Diez J, Sanchez-Mut J V., Setien F, Carmona FJ, et al. 2012. Distinct DNA methylomes of newborns and centenarians. *Proc Natl Acad Sci* **109**: 10522–10527.

Hitt R, Young-Xu Y, Silver M, Perls T. 1999. Centenarians: the older you get, the healthier you have been. *Lancet* **354**: 652.

Holzenberger M, Dupont J, Ducos B, Leneuve P, Géloën A, Even PC, Cervera P, Le Bouc Y. 2003. IGF-1 receptor regulates lifespan and resistance to oxidative stress in mice. *Nature* **421**: 182–187.

Hong M-G, Myers AJ, Magnusson PKE, Prince JA. 2008. Transcriptome-wide assessment of human brain and lymphocyte senescence. *PLoS One* **3**: e3024.

Hutchins AP, Poulain S, Fujii H, Miranda-Saavedra D. 2012. Discovery and characterization of new transcripts from RNA-seq data in mouse CD4+ T cells. *Genomics* **100**: 303–313.

Hwang Y, Kim J, Shin JY, Kim JI, Seo JS, Webster MJ, Lee D, Kim S. 2013. Gene expression profiling by mRNA sequencing reveals increased expression of immune/inflammation-related genes in the hippocampus of individuals with schizophrenia. *Transl Psychiatry* **3**: e321.

Innocenti F, Cooper GM, Stanaway IB, Gamazon ER, Smith JD, Mirkov S, Ramirez J, Liu W, Lin YS, Moloney C, et al. 2011. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet* **7**.

Jumpertz R, Hanson RL, Sievers ML, Bennett PH, Nelson RG, Krakoff J. 2011. Higher energy expenditure in humans predicts natural mortality. *J Clin Endocrinol Metab* **96**: E972–E976.

Kabakchiev B, Silverberg MS. 2013. Expression quantitative trait loci analysis identifies associations between genotype and gene expression in human intestine. *Gastroenterology* **144**.

Kabanova S, Kleinbongard P, Volkmer J, Andrée B, Kelm M, Jax TW. 2009. Gene expression analysis of human red blood cells. *Int J Med Sci* **6**: 156–159.

Kaeberlein M, Powers RW, Steffen KK, Westman EA, Hu D, Dang N, Kerr EO, Kirkland KT, Fields S, Kennedy BK. 2005. Regulation of yeast replicative life span by TOR and Sch9 in response to nutrients. *Science* **310**: 1193–1196.

Kapahi P, Chen D, Rogers AN, Katewa SD, Li PWL, Thomas EL, Kockel L. 2010. With TOR, less is more: A key role for the conserved nutrient-sensing TOR pathway in aging. *Cell Metab* **11**: 453–465.

Kapahi P, Zid BM, Harper T, Koslover D, Sapin V. 2004. Regulation of Lifespan in Drosophila by Modulation of Genes in the TOR Signaling Pathway. *Curr Biol* **14**: 885–890.

Kent JW, Göring HHH, Charlesworth JC, Drigalenko E, Diego VP, Curran JE, Johnson MP, Dyer TD, Cole SA, Jowett JBM, et al. 2012. Genotype×age interaction in human transcriptional ageing. *Mech Ageing Dev* **133**: 581–590.

Kenyon C, Chang J, Gensch E, Rudner A, Tabtiang R. 1993. A C. elegans mutant that lives twice as long as wild type. *Nature* **366**: 461–464.

Kenyon J, Gerson SL. 2007. The role of DNA damage repair in aging of adult stem cells. *Nucleic Acids Res* **35**: 7557–65.

Khazaeli AA, Van Voorhies W, Curtsinger JW. 2005. Longevity and metabolism in Drosophila melanogaster: genetic correlations between life span and age-specific metabolic rate in populations artificially selected for long life. *Genetics* **169**: 231–242.

Kimura KD, Tissenbaum HA, Liu Y, Ruvkun G. 1997. daf-2, an insulin receptor-like gene that regulates longevity and diapause in Caenorhabditis elegans. *Science* **277**: 942–946.

Kirkwood TBL. 2008. A systematic look at an old problem. *Nature* **451**: 644–647.

Kirkwood TBL. 2005. Understanding the odd science of aging. *Cell* **120**: 437–47.

Kumar NT, Liestøl K, Løberg EM, Reims HM, Brorson S-H, Mæhlen J. 2012. The apolipoprotein E polymorphism and cardiovascular diseases—an autopsy study. *Cardiovasc Pathol*.

Kuningas M, Mooijaart SP, van Heemst D, Zwaan BJ, Slagboom PE, Westendorp RGJ. 2008. Genes encoding longevity: from model organisms to humans. *Aging Cell* **7**: 270–80.

Kuranda K, Vargaftig J, de la Rochere P, Dosquet C, Charron D, Bardin F, Tonnelle C, Bonnet D, Goodhardt M. 2011. Age-related changes in human hematopoietic stem/progenitor cells. *Aging Cell* **10**: 542–546.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Lee I, Hendrix A, Kim J, Yoshimoto J, You YJ. 2012. Metabolic Rate Regulates L1 Longevity in C. elegans. *PLoS One* **7**.

Lehmann G, Segal E, Muradian KK, Fraifeld VE. 2008. Do mitochondrial DNA and metabolic rate complement each other in determination of the mammalian maximum longevity? *Rejuvenation Res* **11**: 409–417.

Lener T, Moll PR, Rinnerthaler M, Bauer J, Aberger F, Richter K. 2006. Expression profiling of aging in the human skin. *Exp Gerontol* **41**: 387–397.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li Y, Wang WJ, Cao H, Lu J, Wu C, Hu FY, Guo J, Zhao L, Yang F, Zhang YX, et al. 2009. Genetic association of FOXO1A and FOXO3A with longevity trait in Han Chinese populations. *Hum Mol Genet* **18**: 4897–4904.

Lindahl-Jacobsen R, Tan Q, Mengel-From J, Christensen K, Nebel A, Christiansen L. 2013. Effects of the APOE ε2 allele on mortality and cognitive function in the oldest old. *Journals Gerontol - Ser A Biol Sci Med Sci* **68**: 389–394.

Liu C-C, Liu C-C, Kanekiyo T, Xu H, Bu G. 2013. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nat Rev Neurol* **9**: 106–18.

Liu L, Rando TA. 2011. Manifestations and mechanisms of stem cell aging. *J Cell Biol* **193**: 257–266.

Livak KJ, Schmittgen TD. 2001. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods* **25**: 402–408.

Ljungquist B, Berg S, Lanke J, McClearn GE, Pedersen NL. 1998. The effect of genetic factors for longevity: a comparison of identical and fraternal twins in the Swedish Twin Registry. *J Gerontol A Biol Sci Med Sci* **53**: M441–M446.

Lu T, Pan Y, Kao S-Y, Li C, Kohane I, Chan J, Yankner BA. 2004. Gene regulation and DNA damage in the ageing human brain. *Nature* **429**: 883–891.

Lynch HE, Goldberg GL, Chidgey A, Van den Brink MRM, Boyd R, Sempowski GD. 2009. Thymic involution and immune reconstitution. *Trends Immunol* **30**: 366–373.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**: 1509–1517.

Marttila S, Jylhävä J, Nevalainen T, Nykter M, Jylhä M, Hervonen A, Tserel L, Peterson P, Hurme M. 2013. Transcriptional Analysis Reveals Gender-Specific Changes in the Aging of the Human Immune System. *PLoS One* **8**.

McGue M, Vaupel JW, Holm N, Harvald B. 1993. Longevity is moderately heritable in a sample of Danish twins born 1870-1880. *J Gerontol* **48**: B237–B244.

McKay GJ, Silvestri G, Chakravarthy U, Dasari S, Fritsche LG, Weber BH, Keilhauer CN, Klein ML, Francis PJ, Klaver CC, et al. 2011. Variations in apolipoprotein e frequency with age in a pooled analysis of a large group of older people. *Am J Epidemiol* **173**: 1357–1364.

Metzker ML. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* **11**: 31–46.

Morris JZ, Tissenbaum HA, Ruvkun G. 1996. A phosphatidylinositol-3-OH kinase family member regulating longevity and diapause in Caenorhabditis elegans. *Nature* **382**: 536–539.

Murabito JM, Yuan R, Lunetta KL. 2012. The search for longevity and healthy aging genes: Insights from epidemiological studies and samples of long-lived individuals. *Journals Gerontol - Ser A Biol Sci Med Sci* **67 A**: 470–479.

Murphy CT, McCarroll SA, Bargmann CI, Fraser A, Kamath RS, Ahringer J, Li H, Kenyon C. 2003. Genes that act downstream of DAF-16 to influence the lifespan of Caenorhabditis elegans. *Nature* **424**: 277–283.

Neale M., Cardon L. (Eds.). 1992. Methodology for genetic studies of twins and families **No. 67**. Springer.

Nebel A, Croucher PJP, Stiegeler R, Nikolaus S, Krawczak M, Schreiber S. 2005. No association between microsomal triglyceride transfer protein (MTP) haplotype and longevity in humans. *Proc Natl Acad Sci U S A* **102**: 7906–7909.

Nebel A, Kleindorp R, Caliebe A, Nothnagel M, Blanch?? H, Junge O, Wittig M, Ellinghaus D, Flachsbart F, Wichmann HE, et al. 2011. A genome-wide association study confirms APOE as the major gene influencing survival in long-lived individuals. *Mech Ageing Dev* **132**: 324–330.

Neumann CA, Krause DS, Carman C V, Das S, Dubey DP, Abraham JL, Bronson RT, Fujiwara Y, Orkin SH, Van Etten RA. 2003. Essential role for the peroxiredoxin Prdx1 in erythrocyte antioxidant defence and tumour suppression. *Nature* **424**: 561–565.

Ogg S, Paradis S, Gottlieb S, Patterson GI, Lee L, Tissenbaum HA, Ruvkun G. 1997. The Fork head transcription factor DAF-16 transduces insulin-like metabolic and longevity signals in C. elegans. *Nature* **389**: 994–999.

Palmer C, Diehn M, Alizadeh A, Brown P. 2006. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics* **7**: 115.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**: 1413–1415.

Pang WW, Price E a, Sahoo D, Beerman I, Maloney WJ, Rossi DJ, Schrier SL, Weissman IL. 2011. Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. *Proc Natl Acad Sci U S A* **108**: 20012–7.

Passtoors WM, Beekman M, Deelen J, van der Breggen R, Maier AB, Guigas B, Derhovanessian E, van Heemst D, de Craen AJM, Gunn DA, et al. 2013. Gene expression analysis of mTOR pathway: association with human longevity. *Aging Cell* **12**: 24–31.

Passtoors WM, Beekman M, Gunn D, Boer JM, Heijmans BT, Westendorp RGJ, Zwaan BJ, Slagboom PE. 2008. Genomic studies in ageing research: The need to integrate genetic and gene expression approaches. In *Journal of Internal Medicine*, Vol. 263 of, pp. 153–166.

Passtoors WM, Boer JM, Goeman JJ, van den Akker EB, Deelen J, Zwaan BJ, Scarborough A, van der Breggen R, Vossen RHAM, Houwing-Duistermaat JJ, et al. 2012.

Transcriptional profiling of human familial longevity indicates a role for ASF1A and IL7R. *PLoS One* **7**.

Pawlikowska L, Hu D, Huntsman S, Sung A, Chu C, Chen J, Joyner AH, Schork NJ, Hsueh WC, Reiner AP, et al. 2009. Association of common genetic variation in the insulin/IGF1 signaling pathway with human longevity. *Aging Cell* **8**: 460–472.

Pearl R. 1928. The rate of living: being an account of some experimental studies on the biology of life duration. New York: Knopf.

Pearl R, Pearl RW. 1934. The Ancestry of the Long-lived. Johns Hopkins University Press: Baltimore

Pekovic V, Hutchison CJ. 2008. Adult stem cell maintenance and tissue regeneration in the ageing context: the role for A-type lamins as intrinsic modulators of ageing in adult stem cells and their niches. *J Anat* **213**: 5–25.

Perls T, Kunkel LM, Puca AA. 2002a. The genetics of exceptional human longevity. *J Am Geriatr Soc* **50**: 359–368.

Perls T, Terry D. 2003. Genetics of exceptional longevity. In *Experimental Gerontology*, Vol. 38 of, pp. 725–730.

Perls TT. 2006. The different paths to 100. *Am J Clin Nutr* **83**: 484S–487S.

Perls TT, Bubrick E, Wager CG, Vijg J, Kruglyak L. 1998. Siblings of centenarians live longer. *Lancet* **351**: 1560.

Perls TT, Wilmoth J, Levenson R, Drinkwater M, Cohen M, Bogan H, Joyce E, Brewster S, Kunkel L, Puca A. 2002b. Life-long sustained mortality advantage of siblings of centenarians. *Proc Natl Acad Sci U S A* **99**: 8442–8447.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**: 559–575.

Qihua Tan, Jinghua Zhao, Shuxia Li, Lene Christiansen TAK, Christensen K. 2008. Differential and correlation analyses of microarray gene expression data in the CEPH Utah families. *Genomics* **92**: 94–100.

Radyuk SN, Michalak K, Klichko VI, Benes J, Rebrin I, Sohal RS, Orr WC. 2009. Peroxiredoxin 5 confers protection against oxidative stress and apoptosis and also promotes longevity in Drosophila. *Biochem J* **419**: 437–445.

Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D. 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* **14**: R95.

Rijsdijk F V, Sham PC. 2002. Analytic approaches to twin data using structural equation models. *Brief Bioinform* **3**: 119–133.

Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* **12**: R22.

Rodwell GEJ, Sonu R, Zahn JM, Lund J, Wilhelmy J, Wang L, Xiao W, Mindrinos M, Crane E, Segal E, et al. 2004. A transcriptional profile of aging in the human kidney. *PLoS Biol* **2**.

Rozing MP, Westendorp RGJ, De Craen AJM, Fr??lich M, De Goeij MCM, Heijmans BT, Beekman M, Wijsman CA, Mooijaart SP, Blauw GJ, et al. 2010. Favorable glucose tolerance and lower prevalence of metabolic syndrome in offspring without diabetes mellitus of nonagenarian siblings: The Leiden longevity study. *J Am Geriatr Soc* **58**: 564–569.

Rubner M. 1908. Das Problem der Lebensdauer und seine Beziehungen zu Wachstum und Ernährung. Oldenbourg.

Ruggiero C, Metter EJ, Melenovsky V, Cherubini A, Najjar SS, Ble A, Senin U, Longo DL, Ferrucci L. 2008. High basal metabolic rate is a risk factor for mortality: the Baltimore Longitudinal Study of Aging. *J Gerontol A Biol Sci Med Sci* **63**: 698–706.

Sainz J, Mata I, Barrera J, Perez-Iglesias R, Varela I, Arranz MJ, Rodriguez MC, Crespo-Facorro B. 2013. Inflammatory and immune response genes have significantly altered expression in schizophrenia. *Mol Psychiatry* **18**: 1056–7.

Schächter F, Faure-Delanef L, Guénot F, Rouger H, Froguel P, Lesueur-Ginot L, Cohen D. 1994. Genetic associations with human longevity at the APOE and ACE loci. *Nat Genet* **6**: 29–32.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, et al. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297–302.

Schrack JA, Knuth ND, Simonsick EM, Ferrucci L. 2014. “IDEAL” Aging is associated with lower resting metabolic rate: The baltimore longitudinal study of aging. *Journal of the American Geriatrics Society*.

Sebastiani P, Perls TT. 2012. The genetics of extreme longevity: lessons from the new England centenarian study. *Front Genet* **3**: 277.

Segev F, Mor O, Segev A, Belkin M, Assia EI. 2005. Downregulation of gene expression in the ageing lens: a possible contributory factor in senile cataract. *Eye (Lond)* **19**: 80–85.

Slagboom PE, Beekman M, Passtoors WM, Deelen J, Vaarhorst AAM, Boer JM, van den Akker EB, van Heemst D, de Craen AJM, Maier AB, et al. 2011. Genomics of human longevity. *Philos Trans R Soc Lond B Biol Sci* **366**: 35–42.

Smith JD. 2002. Apolipoproteins and aging: Emerging mechanisms. *Ageing Res Rev* **1**: 345–365.

Soerensen M, Dato S, Christensen K, McGue M, Stevnsner T, Bohr VA, Christiansen L. 2010. Replication of an association of variation in the FOXO3A gene with human longevity using both case-control and longitudinal data. *Aging Cell* **9**: 1010–1017.

Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, Sekowska M, Smith GD, Evans D, Gutierrez-Arcelus M, et al. 2012. Patterns of Cis regulatory variation in diverse human populations. *PLoS Genet* **8**.

Su Z, Li Z, Chen T, Li QZ, Fang H, Ding D, Ge W, Ning B, Hong H, Perkins RG, et al. 2011. Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys. *Chem Res Toxicol* **24**: 1486–1493.

Sullivan PF, Fan C, Perou CM. 2006. Evaluating the comparability of gene expression in blood and brain. *Am J Med Genet - Neuropsychiatr Genet* **141 B**: 261–268.

Tacutu R, Budovsky A, Fraifeld VE. 2010. The NetAge database: A compendium of networks for longevity, age-related diseases and associated processes. *Biogerontology* **11**: 513–522.

Tan Q, Benedictis G De, Ukraintseva S V, Franceschi C, Vaupel JW, Yashin AI. 2002. ARTICLE A centenarian-only approach for assessing gene ± gene interaction in human longevity. *Eur J Hum Genet* 119–124.

Tan Q, Christensen K, Christiansen L, Frederiksen H, Bathum L, Dahlgaard J, Kruse TA. 2005. Genetic dissection of gene expression observed in whole blood samples of elderly Danish twins. *Hum Genet* **117**: 267–274.

Tan Q, Zhao J, Li S, Christiansen L, Kruse TA, Christensen K. 2008. Differential and correlation analyses of microarray gene expression data in the CEPH Utah families. *Genomics* **92**: 94–100.

Tansey WP. 2014. Mammalian MYC Proteins and Cancer. *New J Sci* **2014**: 1–27.

Tarca AL, Romero R, Draghici S. 2006. Analysis of microarray experiments of gene expression profiling. *Am J Obstet Gynecol* **195**: 373–388.

Tatar M, Kopelman A, Epstein D, Tu MP, Yin CM, Garofalo RS. 2001. A mutant Drosophila insulin receptor homolog that extends life-span and impairs neuroendocrine function. *Science* **292**: 107–110.

Terry DF, Wilcox MA, McCormick MA, Pennington JY, Schoenhofen EA, Andersen SL, Perls TT. 2004. Lower all-cause, cardiovascular, and cancer mortality in centenarians' offspring. *J Am Geriatr Soc* **52**: 2074–2076.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–5.

Twine NA, Janitz K, Wilkins MR, Janitz M. 2011. Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. *PLoS One* **6**.

Van Dongen J, Slagboom PE, Draisma HHM, Martin NG, Boomsma DI. 2012. The continuing value of twin studies in the omics era. *Nat Rev Genet* **13**: 640–653.

Van Leeuwen N, Beekman M, Deelen J, Van Den Akker EB, De Craen AJM, Slagboom PE, 'T Hart LM. 2014. Low mitochondrial DNA content associates with familial longevity: The Leiden Longevity Study. *Age (Omaha)* **36**: 1463–1470.

Van Raamsdonk JM, Meng Y, Camp D, Yang W, Jia X, Bénard C, Hekimi S. 2010. Decreased energy metabolism extends life span in Caenorhabditis elegans without reducing oxidative damage. *Genetics* **185**: 559–571.

Van Voorhies WA, Khazaeli AA, Curtsinger JW. 2003. Selected contribution: long-lived Drosophila melanogaster lines exhibit normal metabolic rates. *J Appl Physiol* **95**: 2605–2613; discussion 2604.

Van Voorhies WA, Ward S. 1999. Genetic and environmental conditions that increase longevity in Caenorhabditis elegans decrease metabolic rate. *Proc Natl Acad Sci U S A* **96**: 11399–11403.

Vaupel JW. 2010. Biodemography of human ageing. *Nature* **464**: 536–542.

vB Hjelmborg J, Iachine I, Skytthe A, Vaupel JW, McGue M, Koskenvuo M, Kaprio J, Pedersen NL, Christensen K. 2006. Genetic influence on human lifespan and longevity. *Hum Genet* **119**: 312–321.

Vellai T, Takacs-Vellai K, Zhang Y, Kovacs AL, Orosz L, Müller F. 2003. Genetics: influence of TOR kinase on lifespan in C. elegans. *Nature* **426**: 620.

Vo TKD, Godard P, de Saint-Hubert M, Morrhaye G, Bauwens E, Debacq-Chainiaux F, Glupczynski Y, Swine C, Geenen V, Martens HJ, et al. 2010. Transcriptomic biomarkers of human ageing in peripheral blood mononuclear cell total RNA. *Exp Gerontol* **45**: 188–194.

Vogler GP. 2000. The value of sibling and other ‘‘relational’’data for biodemography and genetic epidemiology. Cells and surveys: Should biological measures be included in social science research, 110-132.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**: 57–63.

Weinberger B, Grubeck-Loebenstein B. 2012. Vaccines for the elderly. *Clin Microbiol Infect* **18**: 100–108.

Westendorp RGJ, Van Heemst D, Rozing MP, Fr??lich M, Mooijaart SP, Blauw GJ, Beekman M, Heijmans BT, De Craen AJM, Slagboom PE. 2009. Nonagenarian siblings and their offspring display lower risk of mortality and morbidity than sporadic nonagenarians: The Leiden longevity study. *J Am Geriatr Soc* **57**: 1634–1637.

Wheeler HE, Kim SK. 2011. Genetics and genomics of human ageing. *Philos Trans R Soc Lond B Biol Sci* **366**: 43–50.

White MF. 2003. Insulin signaling in health and disease. *Science* **302**: 1710–1711.

Willcox BJ, Donlon TA, He Q, Chen R, Grove JS, Yano K, Masaki KH, Willcox DC, Rodriguez B, Curb JD. 2008. FOXO3A genotype is strongly associated with human longevity. *Proc Natl Acad Sci U S A* **105**: 13987–13992.

Willcox BJ, Willcox DC, He Q, Curb JD, Suzuki M. 2006. Siblings of Okinawan centenarians share lifelong mortality advantages. *J Gerontol A Biol Sci Med Sci* **61**: 345–354.

Williams RBH, Chan EKF, Cowley MJ, Little PFR. 2007. The influence of genetic variation on gene expression. *Genome Res* **17**: 1707–1716.

Wright F a, Sullivan PF, Brooks AI, Zou F, Sun W, Xia K, Madar V, Jansen R, Chung W, Zhou Y-H, et al. 2014. Heritability and genomics of gene expression in peripheral blood. *Nat Genet* **46**: 430–437.

Xiong Z, Yu H, Ding Y, Feng C, Wei H, Tao S, Huang D, Zheng SL, Sun J, Xu J, et al. 2014. RNA Sequencing Reveals Upregulation of RUNX1-RUNX1T1 Gene Signatures in Clear Cell Renal Cell Carcinoma. *Biomed Res Int* **2014**: 450621.

Zahn JM, Sonu R, Vogel H, Crane E, Mazan-Mamczarz K, Rabkin R, Davis RW, Becker KG, Owen AB, Kim SK. 2006. Transcriptional profiling of aging in human muscle reveals a common aging signature. *PLoS Genet* **2**: 1058–1069.

Zhang JG, Ma YX, Wang CF, Lu PF, Zhen SB, Gu NF, Feng GY, He L. 1998. Apolipoprotein E and longevity among Han Chinese population. *Mech Ageing Dev* **104**: 159–167.

Zhang W, Ferguson J, Ng SM, Hui K, Goh G, Lin A, Esplugues E, Flavell RA, Abraham C, Zhao H, et al. 2012. Effector CD4+ T cell expression signatures and immune-mediated disease associated genes. *PLoS One* **7**.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**: 821–824.

8 Supplementary material

8.1 Genes with significant gender effect

Table 8-1 Genes that exhibited significant gender effect

Gene symbol	Gene loci	Gender Coefficient[ɬ]	p-value	Adjusted p-value
HDHD1	chrX:6966960-7066231	0.709	5.77E-25	1.75E-21
RPS4X	chrX:71492452-71497141	0.545	2.12E-13	4.3E-10
TXLNG	chrX:16804554-16862642	0.487	2.03E-12	3.09E-09
EIF1AX	chrX:20142635-20159966	0.506	1.87E-11	2.27E-08
ZFX	chrX:24167761-24234372	0.498	3.09E-11	3.13E-08
EIF2S3	chrX:24073064-24096927	0.454	2.36E-10	1.95E-07
ZRSR2	chrX:15808573-15841382	0.461	2.78E-10	1.95E-07
CA5B	chrX:15756411-15805748	0.484	2.88E-10	1.95E-07
KDM6A	chrX:44732422-44971845	0.476	5.81E-10	3.54E-07
KDM5C	chrX:53220502-53254604	0.432	4.17E-08	2.3E-05
VSIG1	chrX:107288199-107322414	0.391	4.55E-08	2.3E-05
SEPT6	chrX:118749687-118827333	0.408	1.10E-07	5.16E-05
PLD6	chr17:17104308-17109646	0.413	1.98E-07	8.6E-05
AZU1	chr19:827830-832017	-0.410	2.44E-07	9.89E-05
OXNAD1	chr3:16306666-16347594	0.341	3.23E-07	0.000123
LTB	chr6:31548335-31550202	0.385	3.67E-07	0.000131
SYT11	chr1:155829259-155854990	-0.386	6.63E-07	0.000224
CDK16	chrX:47077527-47089394	0.362	1.01E-06	0.000309
LOC439949	chr10:6622386-6627323	0.382	1.02E-06	0.000309
OPLAH	chr8:145106166-145115584	0.380	1.32E-06	0.000382
GUSB	chr7:65425672-65447301	-0.385	1.44E-06	0.000398
SCML4	chr6:108023363-108145521	0.380	1.53E-06	0.000405
TBX21	chr17:45810609-45823485	-0.384	1.65E-06	0.000417
PTGDS	chr9:139871955-139876194	-0.353	2.53E-06	0.000616
RNF157	chr17:74138533-74236390	0.371	3.32E-06	0.000777
OSBPL5	chr11:3108345-3186582	-0.349	3.90E-06	0.000878
ZSCAN18	chr19:58595208-58629793	0.335	4.38E-06	0.000952
LILRB1	chr19:55128628-55149004	-0.380	4.64E-06	0.000973
PIK3R3	chr1:46505811-46598708	-0.347	4.88E-06	0.000989
PDGFRB	chr5:149493401-149535422	-0.313	6.29E-06	0.001233
DEFA4	chr8:6793344-6795786	-0.367	6.91E-06	0.001311
REEP4	chr8:21995532-21999448	-0.335	7.11E-06	0.001311

Gene symbol	Gene loci	Gender Coefficient[†]	p-value	Adjusted p-value
LGR6	chr1:202163117-202288889	-0.324	7.79E-06	0.001389
PTPN18	chr2:131113579-131132982	-0.356	8.09E-06	0.001389
CD40LG	chrX:135730335-135742549	0.314	8.27E-06	0.001389
HAPLN3	chr15:89420518-89438770	0.367	8.45E-06	0.001389
MTERFD2	chr2:242026508-242041747	0.350	8.76E-06	0.001401
TMEM63A	chr1:226033232-226070420	0.284	9.10E-06	0.001419
SUSD1	chr9:114803060-114937556	-0.347	1.07E-05	0.001628
GPR56	chr16:57653909-57698944	-0.361	1.23E-05	0.001827
C13orf15	chr13:42031541-42045013	0.332	1.88E-05	0.002723
TTC38	chr22:46663860-46689905	-0.316	2.52E-05	0.003513
ABI3	chr17:47287588-47300587	-0.315	2.54E-05	0.003513
HAVCR2	chr5:156512842-156536248	-0.326	2.81E-05	0.003804
ZFP28	chr19:57050316-57068170	0.306	2.98E-05	0.003938
PIK3IP1	chr22:31677578-31688520	0.261	3.15E-05	0.00408
CD63	chr12:56119229-56122910	-0.310	3.44E-05	0.004285
CUX1	chr7:101459183-101927250	-0.347	3.45E-05	0.004285
LRRC16B	chr14:24521205-24538937	-0.316	3.57E-05	0.004328
FANCI	chr15:89787193-89860362	-0.318	3.63E-05	0.004328
CMKLR1	chr12:108681820-108733094	-0.337	3.77E-05	0.004404
MYO1G	chr7:45002259-45018704	-0.328	4.67E-05	0.005295
GAL3ST4	chr7:99756864-99766373	0.323	4.70E-05	0.005295
GOLM1	chr9:88641057-88715116	-0.323	4.92E-05	0.005444
WDR19	chr4:39184023-39287430	0.317	5.29E-05	0.00574
CCR7	chr17:38710021-38721736	0.256	5.80E-05	0.006189
ADPRH	chr3:119298522-119308792	-0.325	6.20E-05	0.006426
LINC00246B	chrX:72161567-72163589	0.298	6.23E-05	0.006426
SLC22A23	chr6:3269206-3456793	0.305	7.06E-05	0.007123
EPHX2	chr8:27348518-27402439	0.289	7.18E-05	0.007123
ELANE	chr19:852290-856246	-0.333	7.26E-05	0.007123
VMAC	chr19:5904851-5910263	0.304	7.55E-05	0.007292
ICOS	chr2:204801470-204826298	0.303	8.43E-05	0.008007
LRRN3	chr7:110731061-110765509	0.186	8.71E-05	0.00815
TMEM204	chr16:1578741-1605581	0.311	8.87E-05	0.008155
CTSG	chr14:25042723-25045466	-0.322	8.98E-05	0.008155
C2orf89	chr2:85048795-85108252	0.258	0.000109	0.009763
RRBP1	chr20:17594322-17662928	-0.326	0.000117	0.010315
ASCL2	chr11:2289727-2292182	-0.279	0.000121	0.010487
FBXL16	chr16:742499-755825	0.325	0.000124	0.010645
GPR183	chr13:99946788-99959749	0.312	0.000126	0.010668
CAMK4	chr5:110559946-110820748	0.227	0.000133	0.01104

Gene symbol	Gene loci	Gender Coefficient[t]	p-value	Adjusted p-value
CD274	chr9:5450502-5470567	0.304	0.000142	0.011632
NKG7	chr19:51874873-51875960	-0.299	0.000153	0.012391
ZNF101	chr19:19779662-19791138	0.292	0.000166	0.013316
PLXNB1	chr3:48445260-48471460	-0.299	0.000169	0.013341
TIMD4	chr5:156346292-156390266	0.306	0.000175	0.013648
ZNF862	chr7:149535508-149564568	0.312	0.000186	0.014349
LOC100130231	chr8:126934766-126963441	0.291	0.000206	0.015688
SFTPD	chr10:81697495-81708861	-0.295	0.00021	0.015737
N6AMT1	chr21:30244512-30257693	0.275	0.000225	0.016658
ALDH3B1	chr11:67776047-67796743	-0.282	0.000231	0.016902
FAM179A	chr2:29204163-29275096	-0.273	0.000237	0.017188
BACH2	chr6:90636246-91006627	0.213	0.000274	0.019574
KPNA2	chr17:66031847-66042970	-0.280	0.000295	0.020695
PODN	chr1:53527723-53551174	-0.278	0.000299	0.020695
NT5E	chr6:86159301-86205509	0.218	0.000299	0.020695
CD244	chr1:160799949-160832692	-0.268	0.000309	0.021101
SLC27A3	chr1:153747767-153752633	-0.241	0.000327	0.021984
CASP4	chr11:104813593-104839325	0.271	0.000333	0.021984
RPS6KA4	chr11:64126624-64139687	-0.284	0.000335	0.021984
PLOD3	chr7:100849257-100861011	-0.297	0.000336	0.021984
ZNF607	chr19:38187263-38210691	0.280	0.000356	0.022842
STXBP1	chr9:130374485-130454995	0.292	0.000357	0.022842
SERINC5	chr5:79407049-79551898	0.230	0.000365	0.023153
LOC100505696	chr3:15295690-15306005	0.287	0.000373	0.02327
RCAN3	chr1:24828840-24863510	0.197	0.000375	0.02327
CHN1	chr2:175664041-175870107	0.280	0.000384	0.023586
ANXA2	chr15:60639349-60690185	-0.261	0.000419	0.025439
FAM102A	chr9:130702860-130742812	0.208	0.000422	0.025439
FLJ39653	chr4:16228285-16259810	0.295	0.000432	0.0256
GLB1L	chr2:220101502-220110131	-0.273	0.000434	0.0256
AK5	chr1:77747661-78025654	0.267	0.000442	0.025861
MLC1	chr22:50497819-50524358	-0.281	0.000447	0.02588
N4BP3	chr5:177540555-177553107	0.218	0.000451	0.02588
FUCA2	chr6:143815948-143833020	-0.259	0.000456	0.025901
FIG4	chr6:110012423-110146634	-0.285	0.000466	0.026224
AQP3	chr9:33441151-33447631	0.216	0.000486	0.026962
MAST4	chr5:65892175-66465423	0.272	0.000488	0.026962
PDE4A	chr19:10527448-10580307	-0.261	0.0005	0.027192
KIAA0930	chr22:45588122-45636650	-0.279	0.000501	0.027192
MRPS25	chr3:15090018-15106816	0.286	0.000516	0.027791

Gene symbol	Gene loci	Gender Coefficient[t]	p-value	Adjusted p-value
RCC2	chr1:17733250-17766250	-0.276	0.000526	0.02784
SH3TC1	chr4:8201059-8242830	-0.268	0.000526	0.02784
GCET2	chr3:111839687-111852152	0.276	0.000573	0.030054
TBC1D4	chr13:75858808-76056250	0.242	0.000594	0.03086
STMN3	chr20:62271060-62284780	0.286	0.000635	0.032727
CTIF	chr18:46065426-46389586	-0.284	0.000641	0.032743
TCN2	chr22:31003069-31023047	-0.255	0.000646	0.032765
PRTN3	chr19:840984-848175	-0.280	0.000666	0.033487
IGSF9B	chr11:133785184-133826880	0.268	0.000677	0.033741
INADL	chr1:62208148-62629591	0.228	0.000694	0.034317
SOCS1	chr16:11348273-11350039	0.270	0.000701	0.034382
CA5BP1	chrX:15693038-15721474	0.269	0.000715	0.034773
C1orf21	chr1:184356149-184598155	-0.259	0.000738	0.035601
KIAA0748	chr12:55343864-55378456	0.239	0.000752	0.036036
RASSF4	chr10:45455218-45490172	-0.248	0.000783	0.037194
KIR2DL3	chr19:55249973-55264504	-0.273	0.000796	0.037395
OAF	chr11:120081746-120100650	-0.271	0.000801	0.037395
ZNF548	chr19:57901217-57913919	0.242	0.000805	0.037395
GBP2	chr1:89573309-89591799	0.273	0.000823	0.037905
SLAMF7	chr1:160709076-160724601	-0.278	0.000834	0.037905
AXIN2	chr17:63524682-63557740	0.220	0.000835	0.037905
ITPRIPL1	chr2:96991061-96994091	-0.265	0.00089	0.040091
C1orf216	chr1:36179476-36184790	-0.238	0.000899	0.040095
FXYD6	chr11:117707690-117748201	-0.265	0.000903	0.040095
PIM2	chrX:48770458-48776413	0.270	0.000913	0.040101
ZNF385A	chr12:54762919-54785083	-0.272	0.000916	0.040101
CD96	chr3:111260925-111371206	0.232	0.000938	0.040745
GZMB	chr14:25100160-25103432	-0.256	0.000951	0.040765
OTOF	chr2:26680070-26781566	-0.283	0.000952	0.040765
BIRC3	chr11:102188180-102210135	0.260	0.000967	0.041132
PTPN9	chr15:75759461-75871625	-0.266	0.000982	0.041463
CD33	chr19:51728334-51743274	-0.259	0.001	0.041948
PTPRK	chr6:128289923-128841870	0.251	0.001009	0.041948
PLEKHB1	chr11:73357222-73373864	0.271	0.001014	0.041948
MPRIP	chr17:16946073-17095962	0.225	0.001024	0.041948
DHRS12	chr13:52342130-52378293	0.243	0.001029	0.041948
NOG	chr17:54671059-54672951	0.187	0.001041	0.041948
KANK1	chr9:504702-746103	0.268	0.001044	0.041948
FAM69A	chr1:93298285-93427079	0.250	0.001059	0.041948
CARD9	chr9:139258407-139268133	-0.259	0.001065	0.041948

Gene symbol	Gene loci	Gender Coefficient[ŧ]	p-value	Adjusted p-value
AMIGO1	chr1:110049445-110052336	0.218	0.001073	0.041948
FAM113B	chr12:47610051-47630443	0.258	0.001074	0.041948
CLDND2	chr19:51870351-51872257	-0.248	0.001076	0.041948
PAK6	chr15:40509628-40569688	-0.264	0.001112	0.043059
BCL2	chr18:60790578-60986613	0.227	0.001125	0.043316
DGKE	chr17:54911459-54946036	0.252	0.001144	0.043587
FCGR1B	chr1:120926127-120935944	0.224	0.001149	0.043587
ZIK1	chr19:58095627-58103758	0.241	0.001159	0.043587
SLC4A2	chr7:150755298-150773614	-0.249	0.001161	0.043587
ANXA4	chr2:69969126-70053596	-0.272	0.001198	0.043934
KCTD11	chr17:7255207-7258262	-0.246	0.001198	0.043934
APLP2	chr11:129939715-130014706	-0.264	0.001203	0.043934
CDR2	chr16:22357256-22385938	0.216	0.001209	0.043934
FAM134B	chr5:16473146-16617167	0.220	0.001214	0.043934
C2orf18	chr2:26987141-27004099	-0.256	0.001214	0.043934
PLOD1	chr1:11994723-12035599	-0.265	0.001241	0.044663
ZNF470	chr19:57078889-57094262	0.230	0.001271	0.04547
PRSS23	chr11:86511490-86522273	-0.263	0.001287	0.045765
EMP3	chr19:48828628-48833810	-0.263	0.001301	0.046005
RAB13	chr1:153954127-153958806	-0.232	0.001319	0.04636
DPP3	chr11:66247483-66277130	-0.260	0.00135	0.046958
IL18BP	chr11:71709957-71713965	-0.257	0.001351	0.046958
ANKRD40	chr17:48770550-48785270	-0.240	0.001369	0.047323
SLC1A4	chr2:65215578-65251000	-0.256	0.001382	0.047494
NELL2	chr12:44902057-45307711	0.178	0.001443	0.04926
IL6ST	chr5:55230924-55290821	0.231	0.00145	0.04926

[ŧ]Positive gender coefficient means upregulated in females and negative coefficient means downregulated in females.

8.2 All the genes that highly correlated with blood cell populations

Table 8-2 The genes that were highly correlated with blood cell populations

Gene symbol	Loci	Cell type	p-value[*]	Fold change
TCF7	chr5:133450401-133483920	T	2.43E-13	-2.39856663
LEF1	chr4:108968700-109090112	T	5.59E-13	-2.57074344
BCL11B	chr14:99635624-99737822	T	2.78E-12	-1.76062008

Gene symbol	Loci	Cell type	p-value*	Fold change
CR2	chr1:207627644-207663240	B	3.08E-12	-3.76381326
ITGA6	chr2:173292313-173371181	T	5.14E-12	-2.37591824
CD28	chr2:204571197-204603636	T	3.14E-11	-1.96158576
KYNU	chr2:143635194-143799885	B	3.60E-11	1.785216714
CDKN2A	chr9:21967750-21994490	B	5.09E-11	2.793669958
LEPROTL1	chr8:29952921-29995222	T	1.59E-10	-1.56190513
DPP4	chr2:162848754-162931052	T	2.75E-10	-1.65053464
IL7R	chr5:35856990-35876923	T	2.12E-09	-4.56587447
ITK	chr5:156607906-156682109	T	3.67E-09	-2.23731557
RNASE2	chr14:21423629-21424594	GRANS	5.73E-09	3.037626956
AAK1	chr2:69685126-69870977	T	6.48E-09	-1.3342541
SUN2	chr22:39130718-39152024	B	9.44E-09	-1.35764176
ATG4A	chrX:107334898-107397901	B	2.48E-08	1.337542386
PDK1	chr2:173420778-173463862	GRANS	3.40E-08	-1.55828365
PMEPA1	chr20:56223447-56286592	B	3.94E-08	-1.82848518
RAB30	chr11:82692477-82782884	B	6.24E-08	-1.71492263
TXK	chr4:48068409-48136273	T	7.47E-08	-1.76161535
PLXDC1	chr17:37219555-37307902	T	8.96E-08	-1.68919167
RHOH	chr4:40198526-40246281	B	1.09E-07	-1.41751388
OSBPL10	chr3:31702316-32023342	B	1.37E-07	-1.86839612
STRBP	chr9:125883907-126030855	B	1.40E-07	-1.52026108
CD3G	chr11:118215058-118224497	T	1.45E-07	-2.17305748
AKTIP	chr16:53525191-53537170	T	1.67E-07	-1.30988158
TNFRSF10A	chr8:23048969-23082680	GRANS	1.81E-07	-1.61562624
KRIT1	chr7:91828282-91875414	GRANS	2.21E-07	-1.70020532
INPP4B	chr4:142949181-143767604	T	2.65E-07	-1.53426759
DENND4B	chr1:153901976-153919154	B	2.88E-07	-1.27486711
SLFN5	chr17:33570085-33594761	T	3.58E-07	-2.35187419
INPP4A	chr2:99061320-99207496	T	5.44E-07	-1.60455817
AFF3	chr2:100163715-100759037	B	5.74E-07	-2.00251351
TNFRSF25	chr1:6521213-6526255	T	7.17E-07	-1.41812169
BANK1	chr4:102711763-102995969	B	1.01E-06	-1.68607565
CSF2RB	chr22:37309674-37336479	GRANS	1.38E-06	-2.25989869
BCL7A	chr12:122459860-122499950	B	1.91E-06	-1.57385447
CHERP	chr19:16628699-16653263	B	1.93E-06	-1.30829367
STAT6	chr12:57489186-57505196	B	2.02E-06	-1.26965447
SEL1L3	chr4:25749048-25864610	B	2.02E-06	-1.52646456
S100P	chr4:6695565-6698897	GRANS	2.49E-06	2.268856333
SOAT1	chr1:179262848-179327814	GRANS	2.68E-06	-1.36178056
TCF4	chr18:52889561-53303188	B	2.85E-06	-1.72305208

Gene symbol	Loci	Cell type	p-value*	Fold change
FCRL1	chr1:157764193-157789940	B	3.14E-06	-2.66749882
ANXA1	chr9:75766780-75785307	T	3.21E-06	1.703869068
ARHGAP10	chr4:148653452-148993927	B	3.69E-06	1.446942521
PCSK5	chr9:78505559-78977255	T	4.10E-06	-1.59743834
DENND5B	chr12:31535156-31743952	B	4.37E-06	-1.92996297
MLLT3	chr9:20344967-20622514	T	4.49E-06	-1.27728012
WIPF1	chr2:175424301-175547627	B	4.89E-06	-1.55767299
FYB	chr5:39105353-39270759	T	4.93E-06	-2.13059611
CXCR2	chr2:218990012-219001976	GRANS	4.93E-06	-1.74039383
C11orf24	chr11:68028802-68039469	B	5.20E-06	1.304781394
CHD7	chr8:61591323-61780586	B	5.25E-06	-1.60233256
APBA2	chr15:29213839-29410516	T	6.10E-06	-1.40568735
AIM2	chr1:159032274-159046647	B	8.57E-06	1.664838234
RALGPS2	chr1:178694299-178889237	B	8.57E-06	-1.83463858
CCND2	chr12:4382901-4414522	T	9.63E-06	-1.26485679
SEC14L1	chr17:75084724-75213181	GRANS	9.99E-06	-1.65019851
HUWE1	chrX:53559062-53713673	B	1.05E-05	-1.31201267
RALGAPB	chr20:37101485-37207504	B	1.09E-05	-1.51038157
XYLT1	chr16:17196180-17564738	B	1.34E-05	-1.45000951
RCSD1	chr1:167599473-167675486	B	1.46E-05	-1.54096829
CXCR5	chr11:118754474-118766980	B	1.51E-05	-1.57968065
PSEN2	chr1:227058272-227083804	B	1.58E-05	1.324624056
MAL	chr2:95691478-95719735	T	1.62E-05	-1.48756006
FBLN5	chr14:92335754-92414046	T	1.96E-05	-1.5188418
LGALS3	chr14:55595934-55612148	GRANS	2.49E-05	1.263370708
CECR6	chr22:17597188-17602257	GRANS	2.70E-05	1.532266751
TMED8	chr14:77808113-77843396	B	3.36E-05	-2.4736054
WDR11	chr10:122610686-122669038	B	3.45E-05	-1.33398908
SNX29	chr16:12070601-12668146	B	3.55E-05	-1.55610738
IRF8	chr16:85932773-85956211	B	3.58E-05	-1.3155995
ATP6V1B2	chr8:20054703-20079207	GRANS	4.29E-05	-1.2955278
SWAP70	chr11:9685627-9774507	B	5.21E-05	-1.48165763
SNX10	chr7:26331514-26413949	B	5.26E-05	-1.80573425
MS4A1	chr11:60223281-60238225	B	5.26E-05	-2.08502361
LDHB	chr12:21788274-21810789	GRANS	5.56E-05	-1.31628978
HLA-DMA	chr6:32916390-32920899	B	5.76E-05	1.21468926
ZNF207	chr17:30677156-30697468	B	6.43E-05	-1.22581395
SHFM1	chr7:96318078-96339203	T	7.07E-05	1.330616373
KIAA0226L	chr13:46916136-46961635	B	8.07E-05	-1.43925697
LY86	chr6:6588933-6655216	B	9.37E-05	1.420867778

Gene symbol	Loci	Cell type	p-value*	Fold change
IL4R	chr16:27325250-27376099	B	9.37E-05	-1.4286932
LYPLA2	chr1:24117645-24122029	GRANS	9.64E-05	1.248433939
KLHL14	chr18:30252633-30352974	B	9.81E-05	-2.60789518
MME	chr3:154797435-154901518	GRANS	0.000102	-1.48782137
TSPAN3	chr15:77336359-77363570	B	0.000103	-1.30257191
VPS28	chr8:145648999-145653927	B	0.000111	1.208069301
MXD1	chr2:70142172-70170076	GRANS	0.000112	-1.57999332
CD6	chr11:60739112-60787848	T	0.000116	-1.1990339
TRIO	chr5:14143828-14509458	B	0.000125	-1.58201107
FCRL2	chr1:157715522-157746922	B	0.000125	-1.83220983
SPEG	chr2:220299699-220358354	T	0.000138	-1.6892117
KDM4B	chr19:4969123-5153608	B	0.00015	-1.30929965
RTN3	chr11:63448921-63527363	GRANS	0.000153	-1.26492486
CDCA7L	chr7:21940516-21985542	B	0.000156	-1.29572908
MGAM	chr7:141695678-141806547	GRANS	0.000156	-1.82350858
SRGAP2	chr1:206516199-206637783	B	0.000164	1.2179099
BRD4	chr19:15348300-15391262	B	0.000182	-1.26693715
CD200	chr3:112051915-112081658	B	0.000183	-1.78723988
MIR600HG	chr9:125871772-125877756	B	0.000186	-1.3383663
TMC6	chr17:76108998-76128488	T	0.000186	-1.17608314
TRERF1	chr6:42192668-42419783	T	0.000196	-1.33627049
GLUL	chr1:182350838-182361341	GRANS	0.000207	-1.31327274
CREB5	chr7:28338939-28865511	GRANS	0.000217	-1.44919846
DHRS9	chr2:169921298-169952677	GRANS	0.000221	1.265155403
TNFRSF10C	chr8:22960326-22974950	GRANS	0.000256	-1.24709389
ADAM19	chr5:156904311-157002783	B	0.00026	-1.41520947
PRICKLE1	chr12:42852139-42983572	B	0.00027	-1.61295203
PIM1	chr6:37137921-37143204	GRANS	0.000272	-1.15709034
ODC1	chr2:10580507-10588453	B	0.000279	-1.24208157
IL1R2	chr2:102608305-102644884	GRANS	0.000307	1.418261391
DNAJC10	chr2:183580998-183643255	B	0.000343	-1.53821367
KIAA0125	chr14:106383837-106398502	B	0.000355	-1.54837624
MOB3B	chr9:27325206-27529850	B	0.000355	-1.28591649
MRPL49	chr11:64889654-64894841	B	0.00037	-1.10489414
EBF1	chr5:158122922-158526788	B	0.000384	-1.61732941
TRAT1	chr3:108541630-108573714	T	0.000393	-1.61682195
BLNK	chr10:97951454-98031333	B	0.000418	-1.3335382
NAMPT	chr7:105888731-105925638	GRANS	0.000436	-1.49973432
ARL5B	chr10:18948312-18966940	GRANS	0.000451	-2.17621957
CAMP	chr3:48264836-48266981	GRANS	0.000455	2.013874324

Gene symbol	Loci	Cell type	p-value*	Fold change
HLA-DMB	chr6:32902405-32908847	B	0.000462	1.167226789
KLHL12	chr1:202860229-202896371	GRANS	0.000462	-1.10967421
E2F5	chr8:86089618-86126753	B	0.000522	-1.5173253
BCL6	chr3:187439164-187463513	GRANS	0.000553	-1.66406785
CCL5	chr17:34198495-34207377	T	0.000614	1.611345773
AQP9	chr15:58430407-58478110	GRANS	0.000628	-1.2500323
RTKN2	chr10:63952952-64028466	T	0.000643	-1.80099565
SP140	chr2:231090444-231177930	B	0.000654	1.24826587
FAM129C	chr19:17634109-17664648	B	0.000699	-1.61811313
LAT	chr16:28996146-29002104	T	0.000717	-1.21717842
CITED2	chr6:139693396-139695785	GRANS	0.000816	-1.24346728
POU2AF1	chr11:111222980-111250157	B	0.000893	-1.32788055
FCGR2B	chr1:161632904-161648444	B	0.000945	1.275381058
RXRA	chr9:137218315-137332431	GRANS	0.000945	-1.159029

* p-value is with respect to differential expression and is corrected for multiple testing; negative fold change represents downregulation and positive fold change represents upregulation of the gene in LLI; B=B-cell, T=T-cell, GRANS=granulocyte

8.3 GO terms with respect to up- and downregulated genes in the German sample

Table 8-3 All the significantly represented biological processes with respect to the genes upregulated and downregulated in LLI (in the German sample).

Biological processes	enriched/depleted	p-value*
With respect to the upregulated genes		
oxidation-reduction process	+	6.58584E-11
regulation of developmental process	-	7.07123E-07
generation of precursor metabolites and energy	+	3.10797E-06
regulation of cellular process	-	3.10797E-06
small molecule metabolic process	+	1.5728E-05
embryonic morphogenesis	-	1.5728E-05
regulation of signaling	-	4.08626E-05
pattern specification process	-	7.01662E-05
regulation of metabolic process	-	8.71359E-05
organ morphogenesis	-	0.000220242
cell-cell adhesion	-	0.000233777
cell adhesion	-	0.000235619

Biological processes	enriched/depleted	p-value*
cell migration	-	0.000235619
cellular response to stress	+	0.001547862
regulation of body fluid levels	-	0.001547862
tube morphogenesis	-	0.001547862
cellular developmental process	-	0.001559711
tissue morphogenesis	-	0.001559711
antigen processing and presentation of exogenous antigen	+	0.001953199
mitochondrial transport	+	0.002082563
coagulation	-	0.003328394
cell cycle	+	0.006790666
antigen processing and presentation of peptide antigen	+	0.006790666
system development	-	0.006790666
cell differentiation	-	0.009982644
morphogenesis of a branching structure	-	0.010505961
regionalization	-	0.011770173
glutathione derivative metabolic process	+	0.011770173
cellular component movement	-	0.011770173
response to other organism	+	0.013280966
hydrogen transport	+	0.013280966
organ development	-	0.013280966
synapse organization	-	0.01793919
single-organism catabolic process	+	0.018980645
hematopoietic or lymphoid organ development	-	0.025171996
defense response to virus	+	0.025171996
regulation of response to stimulus	-	0.027368965
signal transduction	-	0.030708144
cellular response to endogenous stimulus	-	0.036340253
response to virus	+	0.039529868
tube development	-	0.04358095
viral entry into host cell via membrane fusion with the plasma membrane	+	0.04358095
cellular catabolic process	+	0.04358095
regulation of multicellular organismal process	-	0.049479985
With respect to the downregulated genes		
cellular macromolecule metabolic process	+	1.68181E-73
regulation of metabolic process	+	7.38584E-56
macromolecule metabolic process	+	2.0363E-55
nucleobase-containing compound metabolic process	+	5.34186E-49
cellular aromatic compound metabolic process	+	7.022E-44

Biological processes	enriched/depleted	p-value*
heterocycle metabolic process	+	1.30312E-43
organic cyclic compound metabolic process	+	8.25218E-38
cellular nitrogen compound metabolic process	+	2.4924E-37
regulation of cellular process	+	5.83663E-33
cellular biosynthetic process	+	1.58186E-27
organic substance biosynthetic process	+	2.62531E-26
organelle organization	+	4.74181E-19
negative regulation of biological process	+	2.72502E-09
protein metabolic process	+	1.6398E-08
positive regulation of biological process	+	5.03564E-08
RNA transport	+	7.07176E-08
cell-cell signaling	-	1.14189E-06
modification of morphology or physiology of other organism involved in symbiotic interaction	+	1.99594E-06
cell communication	-	3.32407E-06
modification of morphology or physiology of other organism	+	8.34453E-06
immune response-activating signal transduction	+	1.71139E-05
interaction with host	+	2.69314E-05
embryo development	+	3.83194E-05
intracellular transport	+	4.71315E-05
cellular response to endogenous stimulus	+	5.03087E-05
phosphorus metabolic process	+	7.199E-05
cell cycle	+	0.000207797
macromolecule methylation	+	0.000342191
regulation of response to stimulus	+	0.000355204
protein transport	+	0.000380449
ion transport	-	0.000380449
system process	-	0.000423557
single-organism catabolic process	-	0.000455124
organic acid metabolic process	-	0.000455124
regulation of signaling	+	0.000515664
cell cycle process	+	0.000547866
cellular localization	+	0.00060626
cellular response to stress	+	0.000641649
cellular response to chemical stimulus	+	0.000681059
oxidation-reduction process	-	0.000832884
Golgi vesicle transport	+	0.000958635
regulation of cellular component organization	+	0.001028958
gene silencing	+	0.001212995

Biological processes	enriched/depleted	p-value*
viral process	+	0.001214434
symbiosis, encompassing mutualism through parasitism	+	0.001214434
generation of precursor metabolites and energy	-	0.002098202
regulation of cellular component biogenesis	+	0.002415636
organonitrogen compound metabolic process	-	0.003771562
single-organism biosynthetic process	-	0.004152248
establishment or maintenance of cell polarity	+	0.004802956
hematopoietic or lymphoid organ development	+	0.005683962
regulation of protein stability	+	0.006389161
maintenance of location in cell	+	0.006389161
regulation of RNA stability	+	0.007549147
stem cell maintenance	+	0.008089919
cellular response to abiotic stimulus	+	0.008513448
digestion	-	0.008858864
cellular macromolecule localization	+	0.00887381
xenobiotic metabolic process	-	0.011159448
positive regulation of molecular function	+	0.011159448
small molecule metabolic process	-	0.012258654
regulation of blood pressure	-	0.013099716
protein localization	+	0.013474158
nuclear import	+	0.013741009
immune response-regulating cell surface receptor signaling pathway involved in phagocytosis	+	0.013741009
regulation of membrane potential	-	0.013741009
response to fibroblast growth factor stimulus	+	0.015266984
activation of innate immune response	+	0.015346946
extracellular structure organization	-	0.018782087
cell death	+	0.031538537
regulation of immune system process	+	0.036331984
cellular response to stimulus	+	0.036331984
endosomal transport	+	0.038372379
cerebral cortex development	+	0.042836664
regulation of localization	+	0.045866007

*p-value is corrected for multiple testing; plus sign represent enriched processes and minus sign represent depleted processes.

8.4 The top 50 genes with heritable transcriptional activity

Table 8-4: The top 50 genes with heritable transcriptional activity; the best fit model for these genes was AE (A=additive genetic effect, E=unique environmental effect).

Gene symbol	Gene name	Best fit model	Heritability estimate
MDGA1	MAM domain containing glycosylphosphatidylinositol anchor 1	AE	0.989
FAM110C	family with sequence similarity 110, member C	AE	0.980
HLA-DRB5	major histocompatibility complex, class II, DR beta 5	AE	0.977
HLA-DQB1	major histocompatibility complex, class II, DQ beta 1	AE	0.976
FAM118A	family with sequence similarity 118, member A	AE	0.971
PPP1R17	protein phosphatase 1, regulatory subunit 17	AE	0.968
HLA-DQA2	major histocompatibility complex, class II, DQ alpha 2	AE	0.968
LOC338817	uncharacterized LOC338817	AE	0.964
LOC253039	uncharacterized LOC253039	AE	0.960
ZNF630	zinc finger protein 630	AE	0.958
GAS1	growth arrest-specific 1	AE	0.956
ERAP2	endoplasmic reticulum aminopeptidase 2	AE	0.951
BEGAIN	brain-enriched guanylate kinase-associated	AE	0.951
LRRC37A2	leucine rich repeat containing 37, member A2	AE	0.950
SNORA70	small nucleolar RNA, H/ACA box 70	AE	0.948
LILRA3	leukocyte immunoglobulin-like receptor, subfamily A (without TM domain), member 3	AE	0.948
NNAT	neuronatin	AE	0.948
CHRNA7	cholinergic receptor, nicotinic, alpha 7 (neuronal)	AE	0.947
PAWR	PRKC, apoptosis, WT1, regulator	AE	0.945
ADAMTS8	ADAM metallopeptidase with thrombospondin type 1 motif, 8	AE	0.944
GSTM1	glutathione S-transferase mu 1	AE	0.944
MYCBPAP	MYCBP associated protein	AE	0.944
LOC100288778	WAS protein family homolog 1 pseudogene	AE	0.943
LOC254099	uncharacterized LOC254099	AE	0.942
CNTNAP2	contactin associated protein-like 2	AE	0.942
SPATA20	spermatogenesis associated 20	AE	0.941
CYP2B7P1	cytochrome P450, family 2, subfamily B, polypeptide 7 pseudogene 1	AE	0.940
LOC650623	BEN domain containing 3 pseudogene	AE	0.939
USP32P1	ubiquitin specific peptidase 32 pseudogene 1	AE	0.939
RGS17	regulator of G-protein signaling 17	AE	0.938
HCG4	HLA complex group 4 (non-protein coding)	AE	0.936
TCL1B	T-cell leukemia/lymphoma 1B	AE	0.936
BTN3A2	butyrophilin, subfamily 3, member A2	AE	0.935

Gene symbol	Gene name	Best fit model	Heritability estimate
LOC654433	uncharacterized LOC654433	AE	0.934
NIPAL4	NIPA-like domain containing 4	AE	0.934
MTUS2	microtubule associated tumor suppressor candidate 2	AE	0.934
TREML4	triggering receptor expressed on myeloid cells-like 4	AE	0.934
APOBEC3B	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3B	AE	0.933
DCLK2	doublecortin-like kinase 2	AE	0.932
C17orf97	chromosome 17 open reading frame 97	AE	0.931
TEKT4P2	tektin 4 pseudogene 2	AE	0.931
KANSL1-AS1	KANSL1 antisense RNA 1	AE	0.931
CCDC163P	coiled-coil domain containing 163, pseudogene	AE	0.930
SIGLEC14	sialic acid binding Ig-like lectin 14	AE	0.930
LOC100134868	uncharacterized LOC100134868	AE	0.929
PGM5	phosphoglucomutase 5	AE	0.927
LOC283089	hypothetical LOC283089	AE	0.927
DPYSL4	dihydropyrimidinase-like 4	AE	0.927
COL9A3	collagen, type IX, alpha 3	AE	0.926
LOC339666	uncharacterized LOC339666	AE	0.925

8.5 GO terms identified in the validation dataset (the Danish sample)

Table 8-5 All the significantly represented biological processes with respect to the genes upregulated and downregulated in LLI (in the validation dataset).

Biological processes	enriched/depleted	p-value*
With respect to the upregulated genes		
generation of precursor metabolites and energy	+	0.000113
cellular response to stress	+	0.000113
response to other organism	+	0.000113
oxidation-reduction process	+	0.000113
antigen processing and presentation of exogenous antigen	+	0.000523
response to virus	+	0.000691
viral process	+	0.000691
symbiosis, encompassing mutualism through parasitism	+	0.000691
antigen processing and presentation of peptide antigen	+	0.00082
cell cycle process	+	0.001366

Biological processes	enriched/depleted	p-value*
regulation of multi-organism process	+	0.001694
regulation of body fluid levels	-	0.007592
defense response to virus	+	0.008569
hydrogen transport	+	0.009093
cell cycle	+	0.034471
coagulation	-	0.037905
embryonic morphogenesis	-	0.037905
chromosome segregation	+	0.038531
cytokine production	+	0.041176
organ development	-	0.041176
regulation of developmental process	-	0.041415
defense response	+	0.041426
system development	-	0.041426
cellular catabolic process	+	0.041426
With respect to the downregulated genes		
cellular macromolecule metabolic process	+	5.64E-20
regulation of metabolic process	+	3.45E-17
macromolecule metabolic process	+	1.85E-16
nucleobase-containing compound metabolic process	+	2.6E-13
heterocycle metabolic process	+	6.46E-12
cellular aromatic compound metabolic process	+	6.46E-12
organic cyclic compound metabolic process	+	5.77E-10
regulation of cellular process	+	1.94E-09
cellular nitrogen compound metabolic process	+	2.43E-09
positive regulation of biological process	+	1.23E-08
organelle organization	+	4.09E-08
negative regulation of biological process	+	2.12E-06
RNA transport	+	6.36E-06
regulation of developmental process	+	7.13E-05
viral process	+	8.59E-05
symbiosis, encompassing mutualism through parasitism	+	8.59E-05
modification of morphology or physiology of other organism involved in symbiotic interaction	+	9.08E-05
modification of morphology or physiology of other organism	+	0.000191
cellular response to chemical stimulus	+	0.000349
regulation of cellular component organization	+	0.00045
cellular biosynthetic process	+	0.000535
protein metabolic process	+	0.000999

Biological processes	enriched/depleted	p-value*
interaction with host	+	0.001251
regulation of response to stimulus	+	0.001272
organic substance biosynthetic process	+	0.001303
thymic T cell selection	+	0.001566
macromolecule methylation	+	0.001688
embryo development	+	0.002603
response to organic substance	+	0.003583
cellular response to endogenous stimulus	+	0.004464
nuclear import	+	0.005275
positive regulation of molecular function	+	0.00533
response to fibroblast growth factor stimulus	+	0.007193
cell cycle process	+	0.007488
gene silencing	+	0.00777
regulation of RNA stability	+	0.008396
regulation of immune system process	+	0.010293
somatic stem cell maintenance	+	0.01146
ribonucleoprotein complex biogenesis	+	0.012142
response to topologically incorrect protein	+	0.012992
carbohydrate transport	+	0.014469
regulation of signaling	+	0.014469
stem cell maintenance	+	0.019812
cellular component disassembly	+	0.03276
single-organism biosynthetic process	-	0.033952
nitrogen compound transport	+	0.034536
single-organism catabolic process	-	0.036281
cell cycle	+	0.043801
regulation of multicellular organismal process	+	0.044714
protein export from nucleus	+	0.046254
ribonucleoprotein complex localization	+	0.048642

*p-value is corrected for multiple testing; plus sign represent enriched processes and minus sign represent depleted processes.

8.6 RT-qPCR results

Table 8-6 RT-qPCR results: The candidate genes that showed differential expression.

Gene symbol	TaqMan assay ID	p-value*	Fold change
SATB1	Hs00161515_m1	4.73E-13	-2.42993
PRKCA	Hs00925193_m1	4.73E-13	-2.64929
MAN1C1	Hs00220595_m1	1.10E-12	-2.79214
ETS1	Hs00428293_m1	3.71E-12	-4.05408
HIVEP2	Hs00198801_m1	4.39E-10	-3.69304
SPTBN1	Hs00162271_m1	6.35E-10	-2.32152
GZMH	Hs00277212_m1	3.69E-09	4.008934
ETV7	Hs00903229_m1	3.16E-08	3.099065
ZNF551	Hs00292939_m1	1.07E-07	-2.3261
DYNLT1	Hs00831821_s1	1.89E-07	1.468047
RNASE3	Hs01923184_s1	3.05E-06	2.371746
DDB2	Hs03044953_m1	3.05E-06	1.319008
CAPG	Hs00156249_m1	5.59E-06	1.419846
AXIN2	Hs00610344_m1	1.22E-05	-3.13352
BLVRA	Hs00167599_m1	1.87E-05	1.438033
LRRN3	Hs01087516_m1	1.87E-05	ł
C1QA	Hs00706358_s1	2.09E-05	2.038951
TSPYL5	Hs00603217_s1	3.53E-05	-2.28605
MDK	Hs00171064_m1	5.16E-05	2.266511
TCN1	Hs01055542_m1	5.58E-05	1.568374
HOXB7	Hs00270131_m1	5.99E-05	ł
CCL4	Hs00237011_m1	6.53E-05	1.657879
CCL23	Hs00270756_m1	6.57E-05	3.97179
ZNF154	Hs01068902_m1	1.05E-04	-2.30519
ZNF135	Hs00987427_g1	2.08E-04	-3.38452
IFNG	Hs00989291_m1	0.000293	2.376178
IGFBP2	Hs01040719_m1	4.46E-04	4.344896
MARCO	Hs00198935_m1	0.000682	2.547829
ARG1	Hs00968979_m1	0.000854	1.641064
NT5C3B	Hs00369454_m1	0.001062	1.555315
MMP23B	Hs04187882_g1	0.00115	14.95619
SEPT4	Hs00910208_g1	0.001276	1.308085
CYB5R2	Hs00212055_m1	0.001652	1.281561
LAP3	Hs00429769_m1	0.001889	1.391026
SPHK1	Hs00184211_m1	0.002469	1.486209
RAPGEF3	Hs00183449_m1	0.002517	ł

Gene symbol	TaqMan assay ID	p-value*	Fold change
NUDT18	Hs00368715_m1	0.002907	2.217249
PLA2G16	Hs00912734_m1	0.003144	1.331811
PDGFRB	Hs01019589_m1	0.004556	1.923654
CTSL	Hs00377632_m1	0.005949	1.319879
HP	Hs00605928_g1	0.011211	1.467194
CTSK	Hs00166156_m1	0.011289	1.180602
NR1H3	Hs00172885_m1	0.01165	1.283037
CXCL10	Hs01124251_g1	0.011791	1.39991
APOL4	Hs00540930_m1	0.013613	1.072431
ELANE	Hs00975994_g1	0.017904	1.698913
PRTN3	Hs01597752_m1	0.018185	1.640084
CTSG	Hs00175195_m1	0.029258	2.182746
SERPINB10	Hs00192370_m1	0.034501	1.692342
ELF3	Hs00963881_m1	0.045846	1.469627

*p-value is corrected for multiple testing ;ł the median expression level for these genes was zero either in LLI or in young individuals. Therefore, fold change could not be quantified for these genes.

9 Curriculum vitae

Personal data

Name	Geetha Venkatesh
Address	Harmsstr. 62
	D-24114 Kiel
	Germany
Date of birth	01.11.1984
Place of birth	Nagari, Andhra Pradesh, India
Citizenship	Indian
Marital status	Married

Education

1990 – 1994	The primary school, Bangalore, India
1994 – 1997	The higher primary school, Bangalore, India
1997 – 2000	The higher secondary school, Bangalore, India
2000 – 2002	Pre-university course, Bangalore, India

Study

2002 – 2006	Bachelor of Engineering in Biotechnology at Acharya Institute of Technology, Bangalore, India
2007 – 2010	Master of Science in Life Science Informatics at Bonn University, Bonn, Germany
2010 – Today	PhD work at the Institute of Clinical Molecular Biology, Christian-Albrechts University of Kiel, Germany

10 Declaration

Declaration

Apart from the advice of my supervisors, Dr. Robert Häsler and Prof. Dr. Almut Nebel, this thesis is completely the result of my own work. No part of it has been submitted to any other board for another qualification.

Erklärung

Hiermit erkläre ich, daß diese Dissertation, abgesehen von der Beratung durch meine akademischen Lehrer Dr. Robert Häsler und Prof. Dr. Almut Nebel, nach Inhalt und Form meine eigene Arbeit ist. Sie hat weder im Ganzen noch zum Teil an anderer Stelle im Rahmen eines Promotionsverfahrens vorgelegen.

Kiel, ……. …….

(Geetha Venkatesh)

11 Acknowledgement

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible without the support and guidance that I received from many people. It gives me a great pleasure in expressing my gratitude to all those people who have supported me.

To begin with, my greatest appreciation and gratitude goes to my adviser, **Prof. Almut Nebel** for her trust and support. Her advice on both research as well as on my career have been priceless. She made it possible for me to have many wonderful discussions, which often provided fuel for the next round I needed to complete this long journey. The help and guidance given by her from time to time shall carry me a long way in the journey of life on which I am about to embark.

I would like to say a very big thank you to my supervisor, **Dr. Robert Haesler**, who is always calm, relaxed and cared so much about my work. He was very patient and was always providing clear explanations when I was lost, constantly driving me with energy when I was tired, and always, always giving me his time, in spite of anything else that was going on. He has a wonderful knack for knowing when to help and when to send me away to figure it out on my own.

I would like to express my sincere thanks to **Prof. Philip Rosenstiel** for providing me valuable inputs when I embarked on this journey. I also thank **Prof. Qihua Tan** for generously sharing his time and knowledge in our collaborative work on twins. I would like to thank our collaborators, **Prof. Lene Christiansen** and **Prof. Kaare Christensen** (University of Southern Denmark, Odense) for all the valuable discussions.

I would like to thank the **PopGen Biobank** for the recruitment of probands and sample collection. Many thanks to **Dorina Oelsner** for extracting RNA from the samples and also for her expert help with real-time gene expression studies. I would also like to thank **Ilona Urbach** for preparing the samples for genotyping, **Tanja Wesse** and **Genotyping**

lab for carrying out genotyping. Also, I wish to thank **Melanie Friskovec**, **Dr. Markus Schilhabel** and all the members of sequencing lab for transcriptome sequencing.

I wish to thank all my colleagues at the **Institute of Clinical Molecular Biology** who were very supportive during the course of my PhD thesis. A special thanks to **Dr. Friederike Flachsbart** for all the fruitful discussions and continuous support throughout this long journey. Many thanks to **Dr. Matthias Barann** who introduced me to the RNA-sequencing analysis. Also, thank you to **Frauke Degenhardt** for helping me with proof reading of this thesis.

The biggest thank you goes to **all my friends** without whom I would have never made it. Thank you for all the moral and emotional support, thoughts and ideas, joy and laughter.

I wish to thank the authors, developers and maintainers of the open source and tools used in this work. I would like to appreciate all the researchers whose works I have used, initially in understanding my field of research and later for updates.

Last but not least, I would like to thank **my family** for their love and encouragement. I would like to express my deepest gratitude to **my mom and husband** who stood by me and shared with me both the great and the difficult moments of life. Thank you so much for your love and care.

This work was funded by the INTERREG 4A programme Syddanmark-Schlewig-K.E.R.N. financed by EU funds from the European Regional Development Fund.