Integrating genomic approaches to understand ageing

Marjolein J. Peters

Integrating Genomic Approaches to Understand Ageing

Marjolein J. Peters

ACKNOWLEDGEMENTS

The work described in this thesis was conducted at the department of Internal Medicine at Erasmus University Medical Center Rotterdam.

The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University Rotterdam; the Netherlands Organization for Scientific Research (NWO); the Netherlands Organization for Health Research and Development (ZonMW); the Research Institute for Diseases in the Elderly (RIDE); the Dutch Ministry of Education, Culture and Science; the Dutch Ministry of Health, Welfare and Sports; the European Commission (DG XII); and the Municipality of Rotterdam.

Additional funding for the work described in this thesis was provided by the European Commission (HEALTH-F2-2008-201865, GEFOS; HEALTH-F2-2008 35627, TREAT-OA); Netherlands Organisation for Scientific Research (NWO) Investments (nr. 175.010.2005.011, 911-03-012); the Netherlands Consortium for Healthy Ageing (NCHA); the Netherlands Genomics Initiative (NGI) / Netherlands Organisation for Scientific Research (NWO) project nr. 050-060-810; and Vidi grant 917103521.

The infrastructure for the CHARGE Consortium is supported in part by the National Heart, Lung, and Blood Institute grant R01HL105756.

Financial support for the publication of this thesis was kindly provided by the Erasmus University Rotterdam, the Anna Foundation (Anna Fonds), Stichting Artrose Zorg, Pfizer, and Becton Dickinson (BD).

Design and layout: Legatron Electronic Publishing, Rotterdam
Printing: Ipskamp Printing, Enschede

ISBN: 978-94-028-0061-6

"Integrating Genomic Approaches to Understand Ageing"

"Integratie van genomische onderzoeksmethodes om veroudering beter te begrijpen"

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties.

De openbare verdeding zal plaatsvinden op
woensdag 23 maart 2016 om 13.30 uur

door

Maria Josephine (Marjolein) Peters
geboren te Avereest

Erasmus University Rotterdam

Erasmus

PROMOTIECOMMISSIE

Promotor:	Prof.dr. A.G. Uitterlinden
Overige leden:	Prof.dr. J.H. Gribnau
	Prof.dr. O.H. Franco
	Prof.dr. L.H. Franke
Copromotor:	Dr. J.B.J. van Meurs

Paranimfen:	Dr. H.J.M. Kerkhof
	Dr. L. Stolk

CONTENT

Chapter 1	**General Introduction**	**7**
Chapter 2	**Transcriptomic Studies**	**27**
	2.1 The transcriptional landscape of age in human peripheral blood	29
	2.2 A meta-analysis of gene expression signatures of blood pressure and hypertension	59
	2.3 Meta-analysis of whole-blood gene expression associations with circulating lipid levels	79
	2.4 Gene transcripts associated with muscle strength: a CHARGE meta-analysis of 7,781 persons	105
Chapter 3	**Combining genetic & genomic approaches**	**127**
	3.1 Systematic identification of *trans*-eQTLs as putative drivers of known disease associations	129
	3.2 Identification of non-coding RNA target genes through *trans*-eQTL analysis	151
Chapter 4	**Genomic analysis integration for age-related musculoskeletal comorbidities**	**175**
	4.1 Genome-wide association study meta-analysis of chronic widespread pain: evidence for involvement of the 5p15.2 region	177
	4.2 Genetics of the heat pain threshold in the general population	199
	4.3 Associations between joint effusion in the knee and gene expression levels in the circulation: a meta-analysis	215
Chapter 5	**General Discussion**	**233**
Chapter 6		
	Summary	253
	Samenvatting	259
Chapter 7		
	Bibliography	265
	Authors and affiliations	277
	About the author	301
	Over de auteur	302
	Dankwoord	303

CHAPTER 1

General Introduction

Age is a major risk factor for many common diseases including cancer, cardiovascular disease, hypertension, osteoarthritis, and type 2 diabetes. The process of ageing is described as a decline in intrinsic physiological functioning over time, leading to an increased mortality rate [1]. All cells and tissues experience progressively decreased functioning over time, but it is not clear which of these changes are causal to age-related phenotypes and diseases. Although age is the most powerful risk factor for many common diseases, the underlying molecular mechanisms are still largely unknown.

Biological theories of ageing are dived in two main groups: the programmed ageing theory and the theory of cellular ageing [2]. The programmed ageing theory suggests that ageing is regulated by biological clocks operating throughout lifespan. This regulation would depend on changes in gene expression that affect systems responsible for maintenance, repair, and defense responses. The second theory of cellular ageing is based on the concept that damage, either due to environmental impacts on living organisms, normal byproducts of metabolism, or inefficient repair systems accumulates throughout the lifespan, and causes ageing. Despite the recent advances in molecular biology and genetics, the mysteries that control human lifespan are yet to be unraveled [3].

AGEING RESEARCH IN ANIMAL MODELS

The first studies into the regulation of lifespan were performed in animal models, such as yeast (*Saccharomyces cerevisiae*), round worms (*Caenorhabditis elegans*), fruit flies (*Drosophila melanogaster*), and mice (*Mus musculus*). Genetically engineered animals allow for better understanding of the molecular mechanisms of ageing. An example is a mouse model that expresses a proofreading deficient form of the mitochondrial DNA (mtDNA) polymerase [4]. The mutation resulted in randomly accumulated mtDNA mutations during the course of mitochondrial biosynthesis. The mice displayed a normal phenotype at birth and early adolescence, but subsequently acquired many features of premature ageing (such as weight loss, reduced subcutaneous fat, osteoporosis, anemia, reduced fertility, and heart enlargement) and had a reduced lifespan. These results demonstrate that the accumulation of mtDNA mutations leads to premature ageing in mice. In addition, a significant decrease in the mitochondrial energetic capacity with ageing has been identified in yeast, round worms, and fruit flies [5]. Finally, a study in human volunteers showed that the mtDNA content in muscle declined with ageing [6]. They found reduced levels of mitochondrial gene transcripts and proteins, and a declined capacity for mitochondrial energy production, resulting in lower physical function and higher insulin resistance, both more common in the elderly.

Animal models are still instrumental in studying the molecular mechanisms of disease processes: the short life span of some model species enable longitudinal studies and experimental manipulations [7]. While these studies have been crucial for the identification of some ageing regulating genes and pathways, a limitation is that the findings can be difficult to translate to human ageing. The Ageing Gene Database (GenAge) provides a publically available manually curated collection of all genes related to longevity and ageing in model organisms [8]. Today, about 300 human genes and about 2,000 genes in animals have been related to longevity and/or ageing.

AGEING RESEARCH USING CELL LINES

Next to animal studies, cell culture studies have been used to study ageing. In 1961, Hayflick *et al.* [9] discovered that human fibroblasts derived from embryonic tissues could only divide a finite number of times in culture, a phenomena called "replicative senescence". In 1990, Harley *et al.* found that telomeres (repetitive sequences at the end of your chromosomes) shorten at each passage. The telomere length contributes to the stabilization of the telomeres, and is the key in avoiding replicative senescence [10]. Recently, scientists in Japan were able to either accelerate the process of ageing within human fibroblast cell lines, or to reverse the process of ageing [11]. They targeted two genes (*GCAT* and *SHMT2*) that produce the amino acid glycine in the mitochondria, and by reprogramming the fibroblast cell lines the researchers restored age-associated respiration defects. Yet, the connection between the in vitro findings and the ageing organism remains a subject of controversy, in spite of decades of study [12].

USE OF POPULATION-BASED COHORTS

The studies described in this thesis made use of human population-based cohort studies to examine ageing and age-related comorbidities: a sample of a population is selected for longitudinal assessment of exposure-outcome relations [13]. Compared to animal- and cell line models, the results of human population-based studies are easier to translate to the human situation. The type and amount of genetic and/or genomic diversity between the samples is exploited to understand the molecular mechanisms, rather than active genetic or genomic manipulations in model systems. Population studies can be used to evaluate multiple hypotheses, particularly when data are repeatedly collected. The subjects are followed in time to record the development of risk factors for diseases.

Today, population-based epidemiological studies have identified many risk factors related to negative health outcomes. For example, the causal association between smoking and lung cancer [14] and the association between obesity and increased risk of developing post-menopausal breast cancer [15] have been identified in population-based studies. However, the results of an epidemiological study may be due to an alternative explanation, which is called confounding. Confounding occurs when both the exposure and the outcome are correlated with another risk factor. For example, one study found alcohol intake to be associated with the risk of coronary heart disease. However, smoking may have confounded the association between alcohol intake and coronary heart disease. The studies described in this thesis investigated the role of genetic and genomic factors and the interplay of such factors with environmental factors. When the biology underpinning the molecular epidemiology is used to study the effect on age-related diseases, residual confounding can be largely circumvented [16].

We used genetic studies, transcriptomic studies, and epigenetic studies to better understand the molecular mechanisms of ageing and age-related phenotypes and diseases.

GENETIC STUDIES

The human genome consists of four nucleotides: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). They form basepairs (bp) (A with T, and C with G), and the total length of the human genome is approximately 3,300,000,000 bp. The central dogma of molecular biology involves the genetic code (which is hard-wired into the DNA) to be copied to mRNA (transcription) or DNA (DNA replication). Proteins can be synthesized using the information in the mRNA as a template (Figure 1) [17]. Approximately 70% of the human genome is transcribed, of which only 2% is protein coding [18].

Although the DNA sequences of all humans are similar, no two humans are genetically identical. Today, over 88 million genetic variants have been found [19], which means that there is roughly one variant every 40 bp. Individuals are different in approximately 5.0 million sites in their genomes, so on average, 1 in 660 bp varies between two independent individuals in a population [19]. The most common DNA variants observed in a single genome are called single nucleotide polymorphisms (SNPs). A SNP is a site in the genome where one base pair of the chromosome varies, and this variant occurs in >1% of the population. If the variant is less frequent it is called a rare variant, a mutation refers to (very) rare pathological variations.

When a SNP is located in a protein-coding sequence it could result in an amino acid change, which may modify the peptide and/or the entire protein. This could alter the function, the activation, the localization, and the stability of the protein. If the SNP occurs in a region of the genome that regulates transcription of the gene, it could change the expression by affecting binding of transcription factors for example.

Hypothesis-free genome-wide association studies (GWAS) have been developed to identify SNPs associated with any trait of interest. In a GWAS, a dense set of SNPs (at least 300,000) across the genome is genotyped to survey the most common genetic variation for association with the trait of interest. The number of SNPs studied is much larger compared to a candidate gene approach, and the analysis is hypothesis free, so no prior knowledge about the biology of the trait is needed. Using haplotype information from HapMap [20,21] or the 1000 genomes project [22,23], many more SNPs that are correlated can be imputed. After such imputation, GWAS data of multiple different studies can be combined in a GWAS meta-analysis. A meta-analysis increases the chance of success to identify significantly associated SNPs, which is caused by increased power of the analysis.

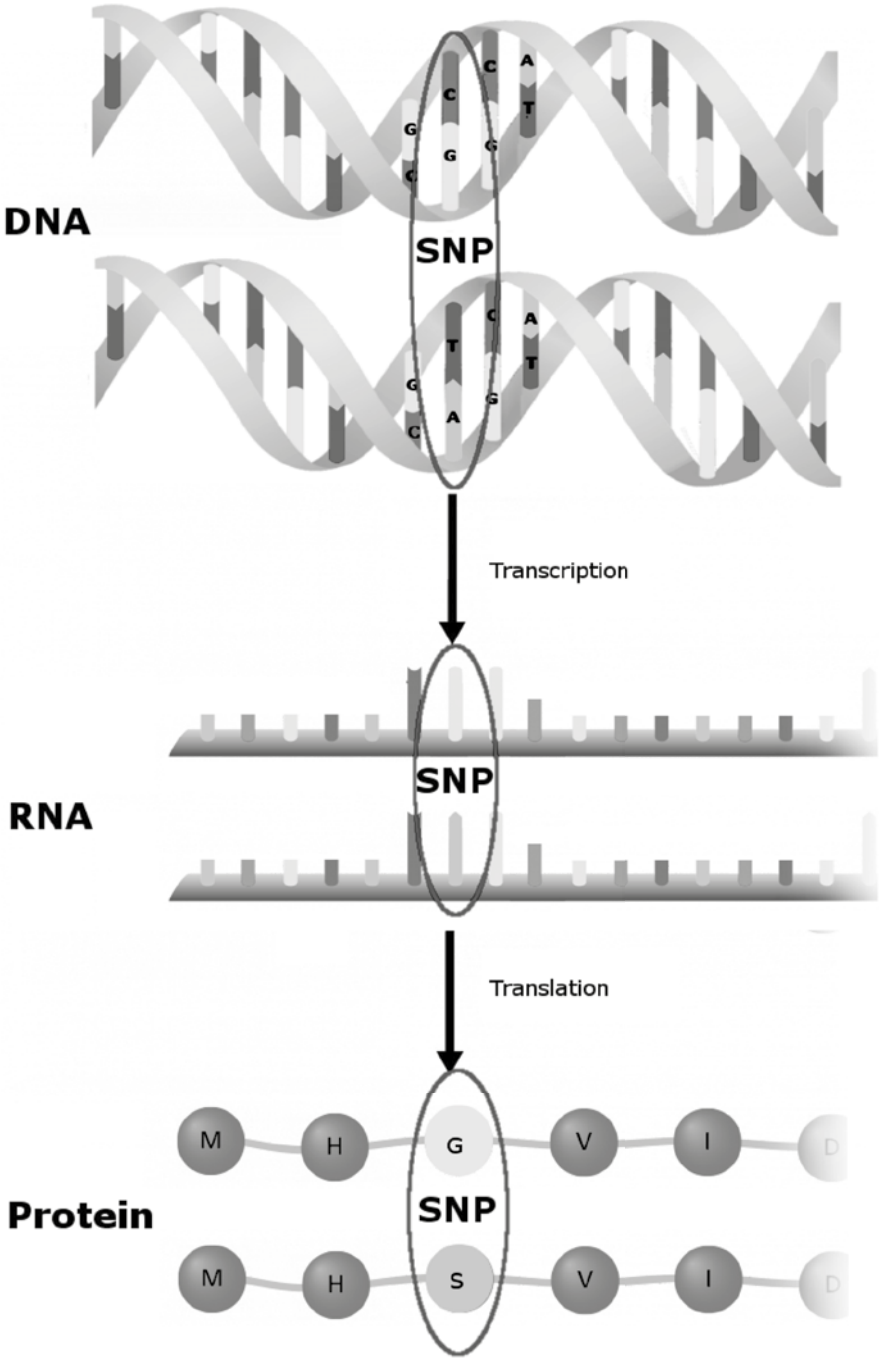


Figure 1. The central dogma of molecular biology. DNA contains the information needed to make all of our proteins, and RNA is the messenger carrying the information to the ribosomes. A SNP in the DNA could influence both the RNA and the protein structure *(image adapted from Genome Research Limited, Wellcome Trust Sanger Institute)*.

Genetic approaches in ageing research have shown that ageing and longevity are difficult to dissect. Human lifespan variation is mainly determined by environmental factors, whereas the heritability is 25-30% and expected to be polygenic [24]. Today, a number of attempts have been made to identify the genetic components of longevity by both candidate-gene studies and GWAS approaches. While thousands of SNPs have been associated with common diseases and traits [25,26], SNPs in/near *APOE*, *FOXO3*, and *5q33.3* are the only identified genetic loci consistently associated with longevity [27-35]. This may be explained by the interaction of ageing with a number of environmental factors (for example, advances in medical care, better lifestyle choices, healthier diets, etc.). Since 1985, the life expectancy at birth increased by 6 years for men and 3 years for women in the Netherlands (CBS levensverwachting 2015). Therefore, the current variation in lifespan in a population seemed to be mainly determined by non-genetic factors [36,37]. However, family studies in human centenarians and their offspring showed an important genetic component to survival to older ages beyond 100 years. And this genetic influence appears to get stronger with older age [38]. The centenarians are expected to have inherited longevity assurance mechanisms that attenuate their ageing rate and protect them from age-related diseases [39].

TRANSCRIPTOMIC STUDIES

Since ageing is characterized by many alterations at the molecular, cellular, and tissue level [40], transcriptome analyses might capture these temporal effects more robustly than genetic studies. The transcriptome is the set of all RNA molecules in one cell or a population of cells; it is highly dynamic and changes with time and in response to external environmental conditions. The transcriptome (also referred to as gene expression levels) is an important intermediate between the genetic code and the synthesized proteins (Figure 1) [17]. Next to the protein-coding genes (which are transcribed into messenger RNA (mRNA) and translated into proteins), there are many non-coding regions in the DNA which are transcribed into functional (non-coding) RNA molecules, such as transfer RNAs (tRNA), small nuclear RNAs (snRNA), ribosomal RNAs (rRNAs), micro RNAs (miRNA), and long non-coding RNAs (lncRNA). In general, non-coding RNAs function as regulators of gene expression at the transcriptional and post-transcriptional level.

The studies described in this thesis focused on mRNAs and lncRNAs. mRNAs form the largest family of RNA molecules. They are transcribed in the nucleus, transported to the cytoplasm, and translated (in the ribosomes) into polymers of amino acids, also called proteins. lncRNAs are non-protein coding RNA molecules longer than 200 nucleotides. lncRNAs are poorly conserved across species, suggesting that they may be subject to different evolutionary constraints [41]. However, a proportion of lncRNAs have been demonstrated to be biologically relevant: they can modulate the function of transcription factors by several different mechanisms, including functioning themselves as co-regulators, modifying transcription factor activity, or regulating the association and activity of co-regulators [42,43].

Gene expression levels can be measured with many different techniques, for example, with northern blotting or with reverse transcription-quantitative PCR (RT-qPCR). In this thesis, we only used genome-wide techniques (gene expression microarrays and RNA sequencing) which measure gene expression levels for a large number of RNAs within one sample.

Microarray technology to measure gene expression levels

Hybridization- based microarray approaches can be run at high throughput. Microarray data is relatively easy to analyze and the arrays are relatively inexpensive. However, the arrays have several limitations, which include: reliance upon existing knowledge of protein-coding sequences; high background levels caused by cross-hybridization [44,45], and a limited dynamic range of detection. There are differences between different microarray platforms that are available on the market. For example, Illumina arrays include only one or two probes per gene (mostly positioned on the 3' end of the gene), while Affymetrix arrays use multiple probes per exon. Therefore, combining results of different platforms can be challenging [46].

RNA sequencing (RNA-seq)

RNA sequencing uses recently developed next generation-sequencing technologies. In short, a population of RNAs is converted to a library of cDNA fragments with adaptors attached to both ends. Each molecule is then sequenced in a high-throughput manner to obtain short sequences from both ends [47]. One advantage is that it is not limited to known protein-coding sequences; it allows investigating new alternative RNA transcripts and new non-coding RNAs, and the precise location of the new transcription boundaries could be identified. In addition, RNA-seq data gives information about how exons are connected (different isoforms), and could give insight into allele specific expression. RNA-seq does not have an upper limit for quantification. Consequently, it has a large dynamic range of expression levels over which transcripts can be detected. And last but not least, genetic variants (SNPs, insertions, deletions) can be measured in the transcripts.

EPIGENETIC STUDIES

The epigenome is a series of chemical modifications that occur on the DNA or specific amino acids in histone proteins. They regulate the transcriptome by making the DNA more or less accessible for binding of proteins that regulate gene expression. There are two main types of epigenetic modifications: DNA methylation and histone modifications.

DNA methylation is the most studied and best understood epigenetic modification: a methyl group is added on the 5-position (C5) of cytosine nucleotides that are found next to a guanine nucleotide in the DNA sequence. These sites are called CpG sites. There are about 28 million CpG sites in the genome, but these are not evenly distributed: the largest part of the genome is depleted of CpG sites with less than one quarter of the expected frequency. By contrast, clusters of CpG sites (also called CpG islands) occur in promoters of housekeeping genes [48] (Figure 2). DNA methylation

is well conserved across species (plants, animals, and fungi), but it can change during life; DNA methylation is tissue specific, and it is influenced by environmental factors like exercise, food intake, smoking, and stress.

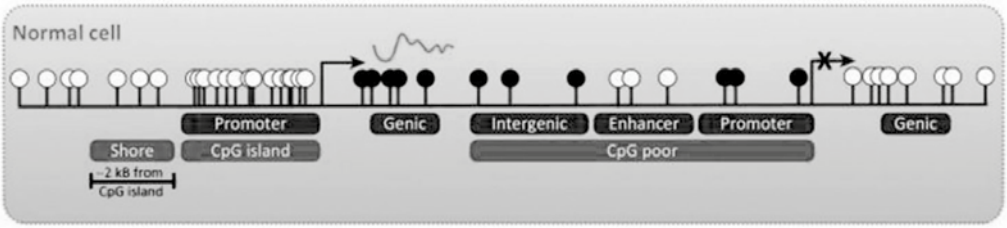


Figure 2. DNA methylation and regulation of the genome *(image adapted from Stirzaker et al.* [48]*)*. CpG islands are often associated with gene promoters and are resistant to DNA methylation. Gene expression can occur, and is highly correlated with high levels of gene body (genic) methylation. CpG-poor regions (intergenic), with the exception of enhancers, are typically methylated. Similarly, CpG-poor promoters are silenced by DNA methylation and exhibit a closed chromatin structure unless gene expression is required (tissue specific). White circles are unmethylated CpGs; black circles are methylated CpGs.

Histone modifications are catalyzed by a number of enzyme families; the best characterized modifications include acetylation and methylation of histones H3 and H4. The modifications directly alter DNA-protein interactions by changing the chromatin structure, which will alter the ability for a gene to be transcribed and expressed. Acetylation adds an acetyl group to lysine amino acids: this causes loosening of chromatin to promote gene activation [49]. Histone deacetylases (which remove the acetyl group) cause chromatin condensation or tightening and gene inactivation. Methylation can occur on lysine or arginine amino acids and can occur in mono-, di- or tri-methylation events. This mark can be associated with both gene activation and inactivation [50].

In large population-based studies, mainly DNA methylation (CpG methylation) is studied. This is mainly driven by the technological possibility to measure DNA methylation in a reasonably high throughput fashion. The most used technique to measure DNA methylation in population studies is with microarrays. Alternative techniques, such as targeted bisulfite sequencing and methylC-capture sequencing, are under development, but remain too expensive up to now.

Microarray technology to measure DNA methylation

The Infinium Human Methylation 450K BeadChips measure DNA methylation levels of 485,000 CpG sites genome wide (1.7% of all CpG sites) for 12 samples simultaneously. The primary focus of the Infinium BeadChips is on CpG islands and promoter regions: they cover 99% of the RefSeq genes (with an average of 17 CpG sites per gene) and 96% of the CpG islands. Like the gene expression

arrays, DNA methylation arrays are relatively inexpensive and easy to analyze. Therefore, they are very popular in large population based cohort studies.

INTEGRATING GENOMIC APPROACHES

To better understand the role of genetic variation and its consequences at various genomic levels during ageing and in age-related comorbidities, we analyzed integrated levels of genetic and genomic data. With integrated data analysis, we hoped to identify key genetic factors that regulate DNA methylation and gene expression levels resulting in age-related traits and diseases (Figure 3).

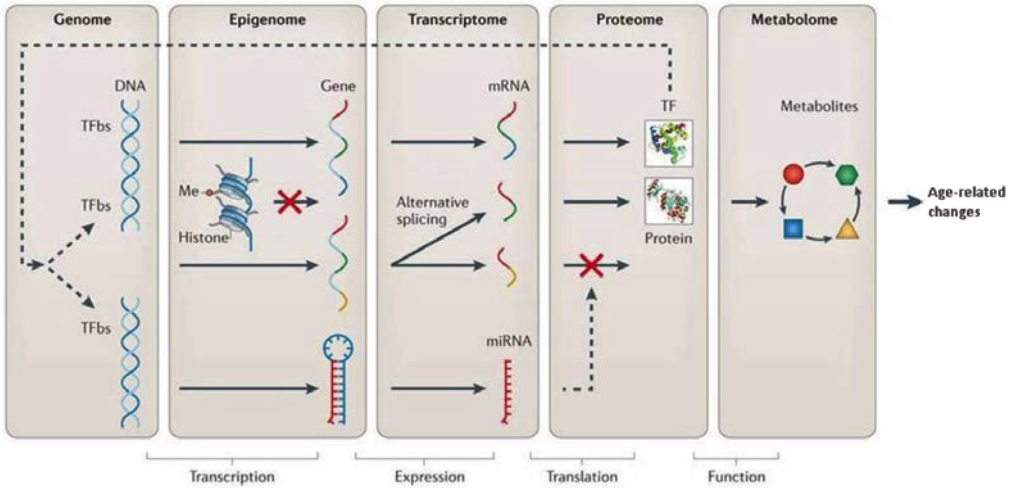


Figure 3. Biological systems multi-omics from the genome, epigenome, transcriptome, proteome and metabolome in relation to the complex phenotypes that change and occur during ageing *(image adapted from Ritchie et al.* [51]*).* Arrows indicate the flow of genetic information from the genome level to the metabolome level and, ultimately, to the phenotype of interest: ageing. The red crosses indicate inactivation of transcription or translation.

Me = methylation; TF = transcription factor; TFbs = transcription factor binding site.

The concept of "genetical genomics" was first introduced in 2001 [52]. Genetic variants were correlated with molecular quantitative traits, such as DNA methylation and gene expression levels. Instead of focusing on direct associations between SNP and disease, the molecular markers are used as endophenotypes [53]. Because the molecular markers intermediate between the genes and the disease outcome, their closer proximity to the genetic variation translates into larger effect sizes. Furthermore, given that the endophenotypes are measurable traits regardless of disease status, their use can enhance power by including both affected and unaffected study subjects.

Detection of *cis*- and *trans*-eQTLs

By combining gene expression levels and SNP variant information, it is possible to identify expression quantitative trait loci (eQTLs). eQTLs indicate what portion of the variation in gene expression is explained by the SNPs in the regulatory regions near the gene (*cis*-eQTLs or direct effects), or to SNPs that reside further away on the chromosome or even on different chromosomes (*trans*-eQTLs or indirect effects) (Figure 4).

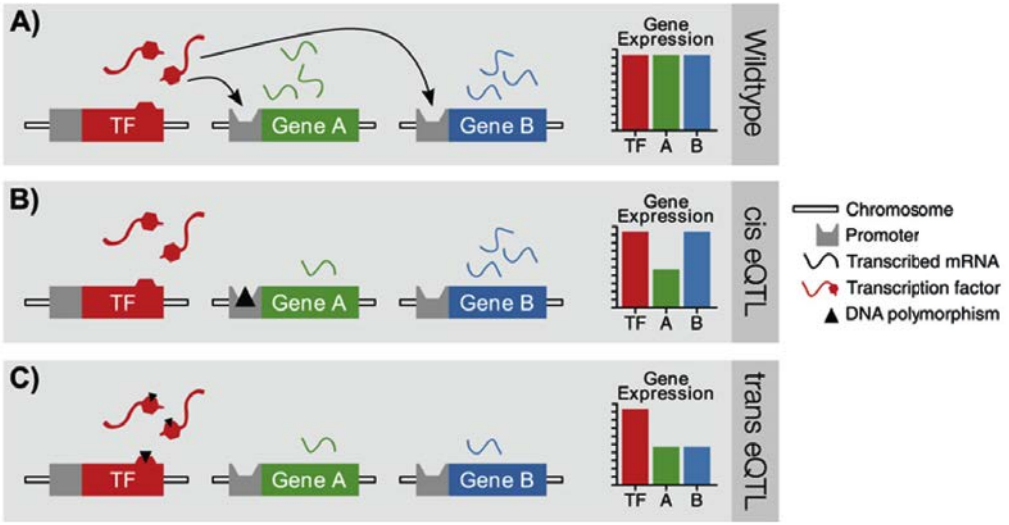


Figure 4. eQTLs can be either direct effects (*cis*-eQTLs) or distant, indirect effects (*trans*-eQTLs) *(image adapted from Wolen et al.* [54]*)*. **(A)** The left-most gene (red) codes for a transcription factor (TF') protein that activates the transcription of genes A (green) and B (blue) by binding to their respective promoters. In the normal scenario all genes are transcribed at their full potential, as indicated by the bar graph on the right. **(B)** A SNP in gene A's promoter region hinders TF binding, causing a reduction in the rate at which gene A is transcribed, while gene B is unaffected. Thus, gene A is being regulated by a *cis*-eQTL because its level of expression is associated with a nearby SNP located on the same chromosome. **(C)** A SNP in the TF gene's DNA binding region hinders binding with all downstream promoters, regardless of whether the regulated gene is located near the TF gene, like gene A, or located on an entirely different chromosome, like gene B. In fact, all genes regulated by this TF would be linked to a *trans*-eQTL at the site of this TF polymorphism.

Cis and *trans*-eQTLs are mapped by testing the correlation between SNP alleles and the variation in gene expression levels. *Cis*-eQTL SNPs are often located close to the transcription start site (TSS) of genes or within gene bodies: as the distance between the eQTL SNP and the TSS decreases, the eQTL effect size generally increases [55-57]. These SNPs may alter the transcription binding sites or other *cis*-regulatory elements that may affect transcription. In contrast to *cis*-eQTLs, the effect sizes of *trans*-eQTLs are generally small. Because of the small effect sizes, the number of reported *trans*-eQTL has remained low [58-60]. However, these *trans*-eQTL analyses are especially interesting,

because they allow us to identify downstream affected disease genes which are not implicated by GWAS studies before. Additionally, they have the ability to reveal previously unknown biological connections between two proteins, which is very important for a better understanding of the disease pathogenesis.

Combining DNA methylation and gene expression data

Apart from genetic variation influencing gene expression levels, we also investigated the possible role of the epigenome on gene expression levels. Since lifestyle and environmental factors (such as smoking [61-68], diet [69-71], and infectious disease [72]) change the epigenome, this analysis could give more insight into what genes are strongly influenced by the environment. Because epigenetic modifications are potentially reversible [73], these genes might represent interesting targets for anti-ageing therapies.

THE ROTTERDAM STUDY

All studies described in this thesis are performed within a large population based cohort study, the Rotterdam Study (RS). In the Netherlands, this study is also known as "Erasmus Rotterdam Gezondheid Onderzoek" (ERGO). The Rotterdam Study (*www.epib.nl/rotterdamstudy)* is a prospective population-based single center cohort study in the well-defined suburb of Ommoord in the city of Rotterdam, the Netherlands, studying determinants of chronic disabling diseases [74]. It consists of three sub-populations: Rotterdam Study I (RS-I) started in 1990 and consists of 7,983 persons (out of 10,215 invitees), aged 55 years and over. This cohort was extended in 1999 with 3,011 participants (out of 4,472 invitees) who had become 55 years of age or moved into the study district since the start of the study, called Rotterdam Study II (or RS-II). In 2006, a further extension of the cohort was initiated in which 3,932 subjects (out of 6,057 invitees) were included, aged 45 years and over, called Rotterdam Study III (RS-III). In total, the Rotterdam Study comprises 14,926 subjects.

All participants were examined in detail at baseline. In summary, a home interview was conducted (~2 hours) and the subjects had an extensive set of examinations (~5 hours) in a specially built research facility in the center of their district. The examinations were repeated every 3-4 years in characteristics that could change over time. An overview of baseline and follow-up visits is given in Figure 5.

The main objective of the Rotterdam Study is to investigate the prevalence and incidence of and risk factors for chronic diseases in the elderly. The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC and by the Ministry of Health, Welfare and Sport of the Netherlands, implementing the *"Wet Bevolkingsonderzoek: ERGO (Population Studies Act: Rotterdam Study)".* All participants provided written informed consent to participate in the study and to obtain information from their treating physicians.

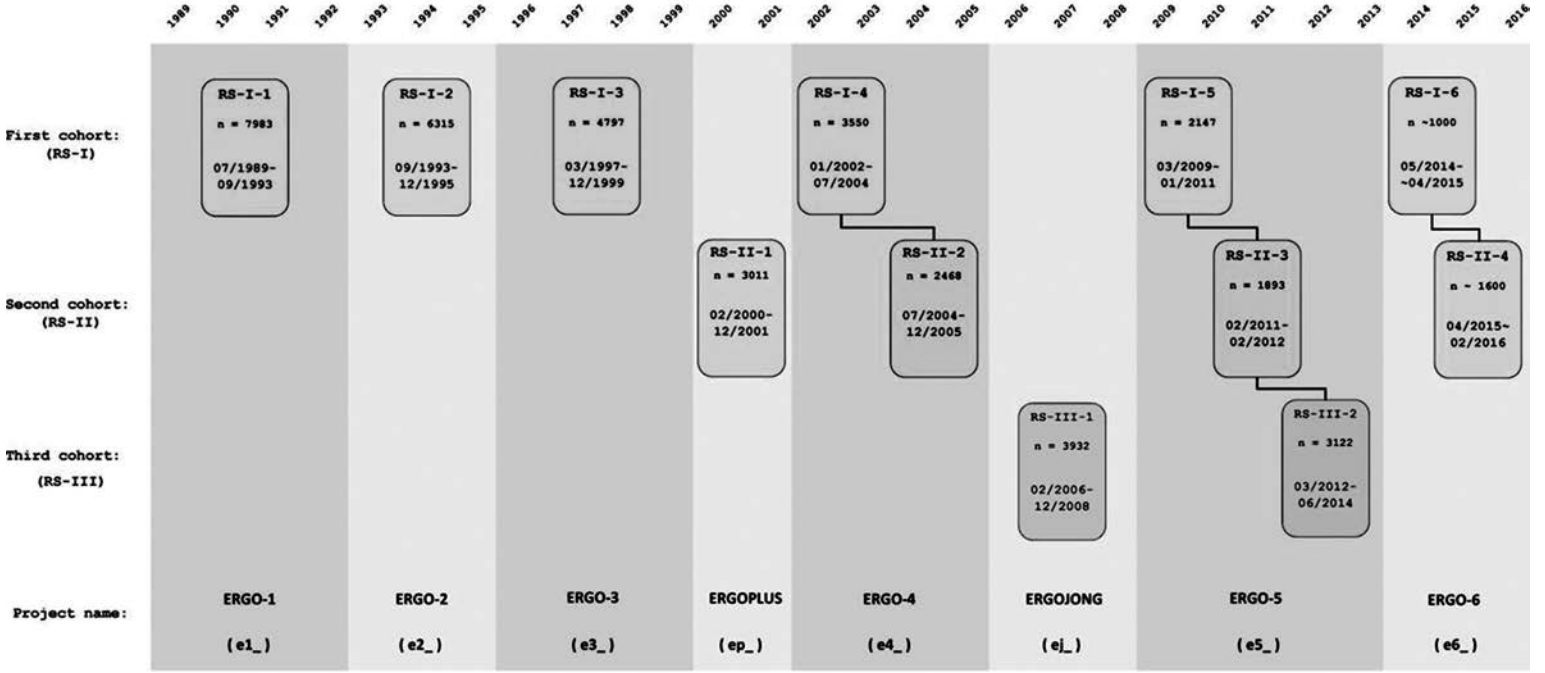


Figure 5. Overview of the Rotterdam Study (RS) sub-populations.

Measurements in blood

Both the DNA methylation and the gene expression levels have been measured in peripheral blood taken by venipuncture. For DNA methylation, the DNA was extracted from the white blood cells (stored in EDTA tubes) by standardized salting out methods. For gene expression levels, total RNA was isolated from the peripheral blood (collected in PAXgene tubes, which immediately stabilize the RNA).

PHENOTYPES STUDIED

Next to ageing and the general age-related phenotypes (for example, blood pressure, cholesterol levels, and muscle strength) [75], we focused on musculoskeletal comorbidities like osteoarthritis and chronic pain in this thesis. With ageing, musculoskeletal comorbidities have become the most frequent cause of physical activity limitations and reduced self-management behavior. Physical inactivity is a major cause of reduced quality of life, as well as many common diseases and even premature death [76].

Osteoarthritis [25] is a common degenerative joint disease affecting the whole joint characterized by pain, stiffness, and disability. The structural characteristics of the disease are articular cartilage loss, formation of new bone, increased thickness of the bone, and cyst formation.

Chronic pain causally relates to an initial local pain stimulus, such as low back pain or local pain due to OA or rheumatoid arthritis (RA). However, only a selection of patients with OA or RA will develop chronic pain, which is accompanied by central sensitization. Nerve impulses keep alerting the brain about tissue damage that no longer exists, which results in enhancement of the response.

THE CHARGE CONSORTIUM

When combining data of multiple studies in a meta-analysis approach, the power of the study can be increased, and the changes of success in discovering new loci is higher. The Cohorts for Heart and Ageing Research in Genomic Epidemiology (CHARGE) consortium was formed to facilitate genome-wide association study meta-analyses and replication opportunities among multiple large and well phenotyped longitudinal cohort studies [77]. The scientific work in the CHARGE consortium takes place in phenotype- and method-specific working groups.

In 2002, we started a Gene Expression Working Group. In this working group, we standardized the pipelines for analyzing gene expression data. Cohorts that participated in the Gene Expression Working Group are presented in Figure 6: the Framingham Heart Study (FHS, United States of America); the Grady Trauma Project (GTP, United States of America); the Heart and Vascular Health Study (HVH, United States of America); the Multi-Ethnic Study of Atherosclerosis (MESA, United States of America); the NIDDK-Phoenix Study (NIDDK/PHOENIX, United States of America); the San Antonio Family Heart Study (SAFHS, United States of America); the Genetic Epidemiology Network of Arteriopathy (GENOA, United States of America); the Atherosclerosis Risk in Communities Study (ARIC, United States of America); the Age, Gene, Environment, Susceptibility Study (AGES, Iceland); the Rotterdam Study (RS, the Netherlands); the Fehrmann *et al.* dataset (FEHRMANN, the Netherlands); the Genetics, Osteoarthritis and Progression study (GARP, the Netherlands); the KOoperative gesundheitsforschung in der Region Augsburg (KORA, Germany); the Study of Health In Pomerania (SHIP, Germany); the Invecchiare in Chianti, ageing in the Chianti area (InCHIANTI, Italy); the Estonian Gene Expression Cohort (EGCUT, Estonia); the Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome study (DILGOM, Finland); and the Brisbane Systems Genetics Study (BSGS, Australia).

AIM OF THIS THESIS

The overall objective of this thesis is to integrate different genetic and genomic approaches to better understand ageing and age-related comorbidities. We assessed the effects of genetic variants, gene expression levels, and DNA methylation levels on age-related phenotypes, and additionally we combined the different levels of -omics data (genomics, epigenomics, and transcriptomics) to identify eQTLs and potentially functional CpG-methylation sites regulating gene expression levels.

The research presented in this thesis examines genomic data from whole blood samples drawn from human population based studies. The work focuses on identifying new genetic and genomic biomarkers that are related to ageing and age-related comorbidities, rather than discovering "regenerative medicine" for anti-ageing therapies.

In *Chapter 2*, ageing and three age-related phenotypes (blood pressure and hypertension, circulating lipid levels, and muscle strength) are studied in relation with the transcriptome. In *Chapter 3*, genetic and transcriptomic data are combined to identify *cis*- and *trans*-eQTLs. In *Chapter 4*, we integrated a number of genomic analyses for different age-related musculoskeletal comorbidities (osteoarthritis and chronic pain).

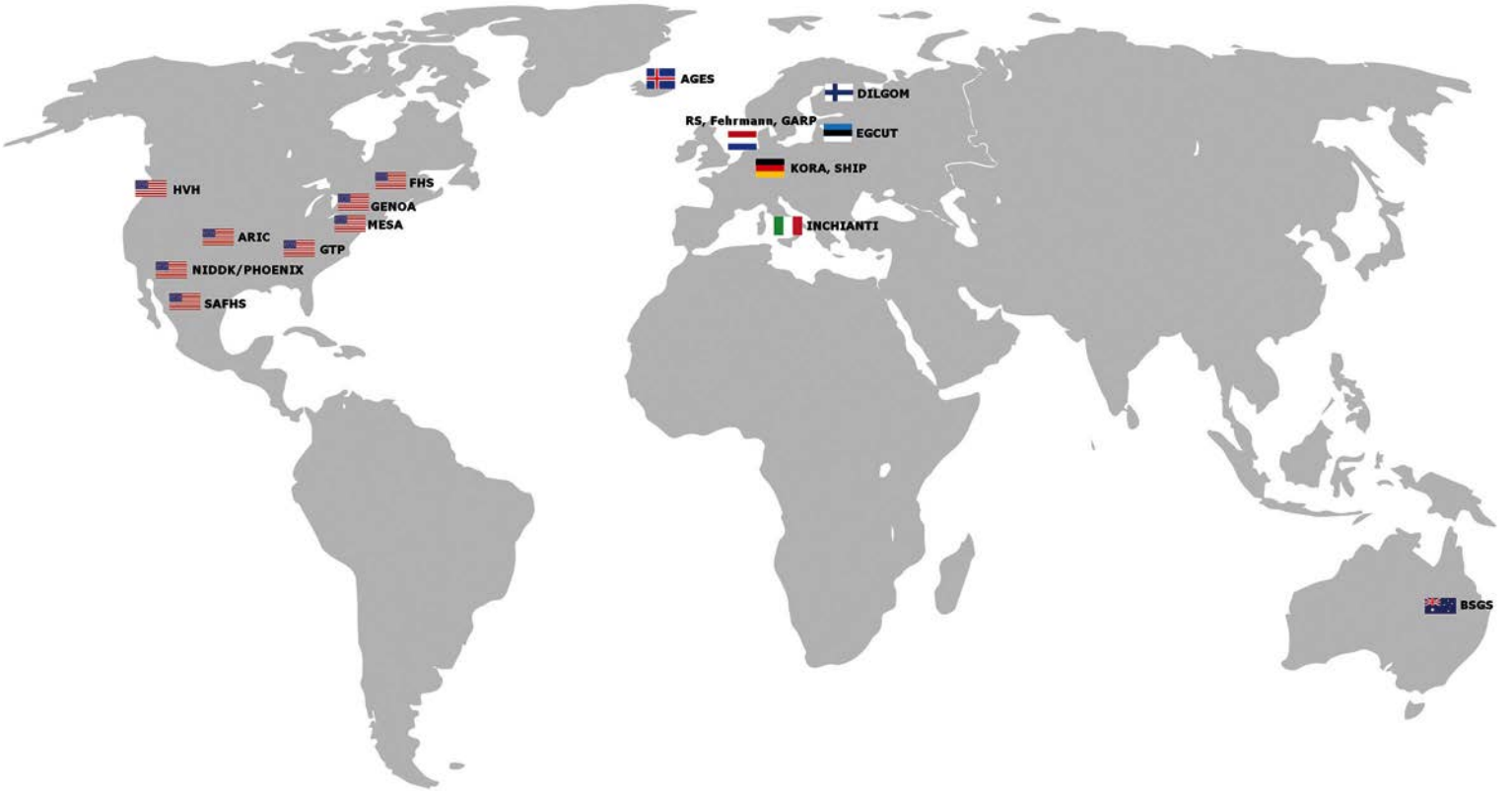


Figure 6. The CHARGE consortium: the participating cohort studies in the Gene Expression Working Group across the world.

REFERENCES

1. Flatt T (2012) A new definition of aging? Front Genet 3: 148.
2. Davidovic M, Sevo G, Svorcan P, Milosevic DP, Despotovic N, et al. (2010) Old age as a privilege of the "selfish ones". Aging and Disease 1: 139-146.
3. Jin K (2010) Modern Biological Theories of Aging. Aging and Disease 1: 72-74.
4. Bratic I, Trifunovic A (2010) Mitochondrial energy metabolism and ageing. Biochim Biophys Acta 1797: 961-967.
5. Guarente L, Kenyon C (2000) Genetic pathways that regulate ageing in model organisms. Nature 408: 255-262.
6. Short KR, Bigelow ML, Kahl J, Singh R, Coenen-Schimke J, et al. (2005) Decline in skeletal muscle mitochondrial function with aging in humans. Proc Natl Acad Sci U S A 102: 5618-5623.
7. Buffenstein R (2005) The naked mole-rat: a new long-living model for human aging research. J Gerontol A Biol Sci Med Sci 60: 1369-1377.
8. de Magalhaes JP, Budovsky A, Lehmann G, Costa J, Li Y, et al. (2009) The Human Ageing Genomic Resources: online databases and tools for biogerontologists. Aging Cell 8: 65-72.
9. Hayflick L, Moorhead PS (1961) The serial cultivation of human diploid cell strains. Exp Cell Res 25: 585-621.
10. Harley CB, Futcher AB, Greider CW (1990) Telomeres shorten during ageing of human fibroblasts. Nature 345: 458-460.
11. Hashizume O, Ohnishi S, Mito T, Shimizu A, Iashikawa K, et al. (2015) Epigenetic regulation of the nuclear-coded GCAT and SHMT2 genes confers human age-associated mitochondrial respiration defects. Sci Rep 5: 10434.
12. de Magalhaes JP (2004) From cells to ageing: a review of models and mechanisms of cellular senescence and their impact on human ageing. Exp Cell Res 300: 1-10.
13. Szklo M (1998) Population-based cohort studies. Epidemiol Rev 20: 81-90.
14. (2005) Annual smoking-attributable mortality, years of potential life lost, and productivity losses--United States, 1997-2001. MMWR Morb Mortal Wkly Rep 54: 625-628.
15. Guh DP, Zhang W, Bansback N, Amarsi Z, Birmingham CL, et al. (2009) The incidence of co-morbidities related to obesity and overweight: a systematic review and meta-analysis. BMC Public Health 9: 88.
16. Palmer L, Burton P, Smith GD (2011) An introduction to genetic epidemiology. The Policy Press - University of Bristol.
17. Crick F (1970) Central dogma of molecular biology. Nature 227: 561-563.
18. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, et al. (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 22: 1775-1789.
19. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. (2015) A global reference for human genetic variation. Nature 526: 68-74.
20. Manolio TA, Collins FS (2009) The HapMap and genome-wide association studies in diagnosis and therapy. Annu Rev Med 60: 443-456.
21. International HapMap C (2005) A haplotype map of the human genome. Nature 437: 1299-1320.
22. Patterson K (2011) 1000 genomes: a world of variation. Circ Res 108: 534-536.
23. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.
24. Deelen J (2014) Doctoral Thesis: Genetic and biomarker studies of human longevity: Leiden University.
25. Eicher JD, Landowski C, Stackhouse B, Sloan A, Chen W, et al. (2014) GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. Nucleic Acids Res.
26. Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42: D1001-1006.
27. Anselmi CV, Malovini A, Roncarati R, Novelli V, Villa F, et al. (2009) Association of the FOXO3A locus with extreme longevity in a southern Italian centenarian study. Rejuvenation Res 12: 95-104.
28. Broer L, Buchman AS, Deelen J, Evans DS, Faul JD, et al. (2014) GWAS of Longevity in CHARGE Consortium Confirms APOE and FOXO3 Candidacy. J Gerontol A Biol Sci Med Sci.

29. Nebel A, Kleindorp R, Caliebe A, Nothnagel M, Blanche H, 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.
30. Schachter F, Faure-Delanef L, Guenot F, Rouger H, Froguel P, et al. (1994) Genetic associations with human longevity at the APOE and ACE loci. Nat Genet 6: 29-32.
31. Soerensen M, Dato S, Christensen K, McGue M, Stevnsner T, et al. (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.
32. Walter S, Atzmon G, Demerath EW, Garcia ME, Kaplan RC, et al. (2011) A genome-wide association study of aging. Neurobiol Aging 32: 2109 e2115-2128.
33. Willcox BJ, Donlon TA, He Q, Chen R, Grove JS, et al. (2008) FOXO3A genotype is strongly associated with human longevity. Proc Natl Acad Sci U S A 105: 13987-13992.
34. Ganna A, Rivadeneira F, Hofman A, Uitterlinden AG, Magnusson PK, et al. (2013) Genetic determinants of mortality. Can findings from genome-wide association studies explain variation in human mortality? Hum Genet 132: 553-561.
35. Sebastiani P, Solovieff N, Dewan AT, Walsh KM, Puca A, et al. (2012) Genetic signatures of exceptional longevity in humans. PLoS One 7: e29848.
36. Skytthe A, Pedersen NL, Kaprio J, Stazi MA, Hjelmborg JV, et al. (2003) Longevity studies in GenomEUtwin. Twin Res 6: 448-454.
37. Herskind AM, McGue M, Holm NV, Sorensen TI, Harvald B, et al. (1996) The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870-1900. Hum Genet 97: 319-323.
38. Sebastiani P, Perls TT (2012) The genetics of extreme longevity: lessons from the new England centenarian study. Front Genet 3: 277.
39. Terry DF, Wilcox MA, McCormick MA, Pennington JY, Schoenhofen EA, et al. (2004) Lower all-cause, cardiovascular, and cancer mortality in centenarians' offspring. J Am Geriatr Soc 52: 2074-2076.
40. Kenyon CJ (2010) The genetics of ageing. Nature 464: 504-512.
41. Pang KC, Frith MC, Mattick JS (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet 22: 1-5.
42. Feng J, Bi C, Clark BS, Mady R, Shah P, et al. (2006) The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. Genes Dev 20: 1470-1484.
43. Panganiban G, Rubenstein JL (2002) Developmental functions of the Distal-less/Dlx homeobox genes. Development 129: 4371-4386.
44. Okoniewski MJ, Miller CJ (2006) Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics 7: 276.
45. Royce TE, Rozowsky JS, Gerstein MB (2007) Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification. Nucleic Acids Res 35: e99.
46. Xu Y, Barter MJ, Swan DC, Rankin KS, Rowan AD, et al. (2012) Identification of the pathogenic pathways in osteoarthritic hip cartilage: commonality and discord between hip and knee OA. Osteoarthritis Cartilage 20: 1029-1038.
47. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63.
48. Stirzaker C, Taberlay PC, Statham AL, Clark SJ (2014) Mining cancer methylomes: prospects and challenges. Trends Genet 30: 75-84.
49. Strahl BD, Allis CD (2000) The language of covalent histone modifications. Nature 403: 41-45.
50. Hayakawa T, Nakayama J (2011) Physiological roles of class I HDAC complex and histone demethylase. J Biomed Biotechnol 2011: 129383.
51. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D (2015) Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet 16: 85-97.
52. Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet 17: 388-391.

53. Ertekin-Taner N (2011) Gene expression endophenotypes: a novel approach for gene discovery in Alzheimer's disease. Mol Neurodegener 6: 31.
54. Wolen AR, Miles MF (2012) Identifying gene networks underlying the neurobiology of ethanol and alcoholism. Alcohol Res 34: 306-317.
55. Stranger BE, De Jager PL (2012) Coordinating GWAS results with gene expression in a systems immunologic paradigm in autoimmunity. Curr Opin Immunol 24: 544-551.
56. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, et al. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4: e1000214.
57. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, et al. (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325: 1246-1250.
58. Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, et al. (2012) Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet 44: 1084-1089.
59. Innocenti F, Cooper GM, Stanaway IB, Gamazon ER, Smith JD, et al. (2011) Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. PLoS Genet 7: e1002078.
60. Fehrmann RS, Jansen RC, Veldink JH, Westra HJ, Arends D, et al. (2011) Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet 7: e1002197.
61. Breitling LP, Yang RX, Korn B, Burwinkel B, Brenner H (2011) Tobacco-Smoking-Related Differential DNA Methylation: 27K Discovery and Replication. American Journal of Human Genetics 88: 450-457.
62. Besingi W, Johansson A (2014) Smoke-related DNA methylation changes in the etiology of human disease. Hum Mol Genet 23: 2290-2297.
63. Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, et al. (2013) Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Hum Mol Genet 22: 843-851.
64. Dogan MV, Shields B, Cutrona C, Gao L, Gibbons FX, et al. (2014) The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. BMC Genomics 15: 151.
65. Sun YV, Smith AK, Conneely KN, Chang Q, Li W, et al. (2013) Epigenomic association analysis identifies smoking-related DNA methylation sites in African Americans. Hum Genet 132: 1027-1037.
66. Harlid S, Xu Z, Panduri V, Sandler DP, Taylor JA (2014) CpG sites associated with cigarette smoking: analysis of epigenome-wide data from the Sister Study. Environ Health Perspect 122: 673-678.
67. Elliott HR, Tillin T, McArdle WL, Ho K, Duggirala A, et al. (2014) Differences in smoking associated DNA methylation patterns in South Asians and Europeans. Clin Epigenetics 6: 4.
68. Guida F, Sandanger TM, Castagne R, Campanella G, Polidoro S, et al. (2015) Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. Hum Mol Genet 24: 2349-2359.
69. Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, et al. (2008) Persistent epigenetic differences associated with prenatal exposure to famine in humans. Proc Natl Acad Sci U S A 105: 17046-17049.
70. Link A, Balaguer F, Goel A (2010) Cancer chemoprevention by dietary polyphenols: promising role for epigenetics. Biochem Pharmacol 80: 1771-1792.
71. Niculescu MD, Lupu DS (2011) Nutritional influence on epigenetics and effects on longevity. Curr Opin Clin Nutr Metab Care 14: 35-40.
72. Paschos K, Allday MJ (2010) Epigenetic reprogramming of host genes in viral and microbial pathogenesis. Trends Microbiol 18: 439-447.
73. Falahi F, van Kruchten M, Martinet N, Hospers GA, Rots MG (2014) Current and upcoming approaches to exploit the reversibility of epigenetic mutations in breast cancer. Breast Cancer Res 16: 412.
74. Hofman A, Brusselle GG, Darwish Murad S, van Duijn CM, Franco OH, et al. (2015) The Rotterdam Study: 2016 objectives and design update. Eur J Epidemiol 30: 661-708.
75. Simm A, Nass N, Bartling B, Hofmann B, Silber RE, et al. (2008) Potential biomarkers of ageing. Biol Chem 389: 257-265.

76. Martinson BC, O'Connor PJ, Pronk NP (2001) Physical inactivity and short-term all-cause mortality in adults with chronic disease. Arch Intern Med 161: 1173-1180.

77. Psaty BM, O'Donnell CJ, Gudnason V, Lunetta KL, Folsom AR, et al. (2009) Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circ Cardiovasc Genet 2: 73-80.

CHAPTER 2

Transcriptomic Studies

CHAPTER 2.1

The transcriptional landscape of age in human peripheral blood

Marjolein J. Peters*, Roby Joehanes*, Luke C. Pilling*, Claudia Schurmann*, Karen N. Conneely*, Joseph Powell*, Eva Reinmaa*, George L. Sutphin*, Alexandra Zhernakova*, Katharina Schramm*, Yana A. Wilson*, Sayuko Kobes, Taru Tukiainen, NABEC/UKBEC Consortium, Yolande F. Ramos, Harald H.H. Göring, Myriam Fornage, Yongmei Liu, Sina A. Gharib, Barbara E. Stranger, Philip L. De Jager, Abraham Aviv, Daniel Levy, Joanne M. Murabito, Peter J. Munson, Tianxiao Huan, Albert Hofman, André G. Uitterlinden, Fernando Rivadeneira, Jeroen van Rooij, Lisette Stolk, Linda Broer, Michael M.P.J. Verbiest, Mila Jhamai, Pascal Arp, Andres Metspalu, Liina Tserel, Lili Milani, Nilesh J. Samani, Pärt Peterson, Silva Kasela, Veryan Codd, Annette Peters, Cavin K. Ward-Caviness, Christian Herder, Melanie Waldenberger, Michael Roden, Paula Singmann, Sonja Zeilinger, Thomas Illig, Georg Homuth, Hans-Jörgen Grabe, Henry Völzke, Leif Steil, Thomas Kocher, Anna Murray, David Melzer, Hanieh Yaghootkar, Stefania Bandinelli, Eric K. Moses, Jack W. Kent, Joanne E. Curran, Matthew P. Johnson, Sarah Williams-Blangero, Harm-Jan Westra, Allan F. McRae, Jennifer A. Smith, Sharon L.R. Kardia, Iiris Hovatta, Markus Perola, Samuli Ripatti, Veikko Salomaa, Anjali K. Henders, Nicholas G. Martin, Alicia K. Smith, Divya Mehta, Elisabeth B. Binder, K. Maria Nylocks, Elizabeth M. Kennedy, Torsten Klengel, Jingzhong Ding, Astrid M. Suchy-Dicey, Daniel A. Enquobahrie, Jennifer Brody, Jerome I. Rotter, Yii-Der I. Chen, Jeanine Houwing-Duistermaat, Margreet Kloppenburg, P. Eline Slagboom, Quinta Helmer, Wouter den Hollander, Shannon Bean, Towfique Raj, Noman Bahkshi, Qiao Ping Wang, Lisa J. Oyston, Bruce M. Psaty, Russell P. Tracy, Grant W. Montgomery, Stephen T. Turner, John Blangero, Ingrid Meulenbelt, Kerry J. Ressler, Jian Yang*, Lude Franke*, Johannes Kettunen*, Peter M. Visscher*, Graham Greg Neely*, Ron Korstanje*, Robert L. Hanson*, Holger Prokisch*, Luigi Ferrucci*, Tonu Esko*, Alexander Teumer*, Joyce B.J. van Meurs*, Andrew D. Johnson*

** These authors contributed equally to this work*

Published: *Nat Commun 2015; **6**:8570 doi:10.1038/ncomms9570*

ABSTRACT

Disease incidences increase with age, but the molecular characteristics of ageing that lead to increased disease susceptibility remain inadequately understood. Here we perform a whole-blood gene expression meta-analysis in 14,983 individuals of European ancestry (including replication) and identify 1,497 genes that are differentially expressed with chronological age. The age-associated genes do not harbor more age-associated CpG-methylation sites than other genes, but are instead enriched for the presence of potentially functional CpG-methylation sites in enhancer and insulator regions that associate with both chronological age and gene expression levels. We further used the gene expression profiles to calculate the 'transcriptomic age' of an individual, and show that differences between transcriptomic age and chronological age are associated with biological features linked to ageing, such as blood pressure, cholesterol levels, fasting glucose, and body mass index. The transcriptomic prediction model adds biological relevance and complements existing epigenetic prediction models, and can be used by others to calculate transcriptomic age in external cohorts.

INTRODUCTION

Chronological age is a major risk factor for many common diseases including heart disease, cancer, and stroke, three of the leading causes of death. Although chronological age is the most powerful risk factor for most chronic diseases, the underlying molecular mechanisms that lead to generalized disease susceptibility are largely unknown. Genome-wide association studies (GWAS) have identified thousands of single nucleotide polymorphisms (SNPs) associated with common human diseases and traits [1,2]. Despite this success, *APOE*, *FOXO3*, and *5q33.3* are the only identified loci consistently associated with longevity [3-11]. Ageing has proven difficult to dissect in part due to its interactions with environmental influences (e.g., lifestyle choices, diet, local exposures), other genetic factors, and a large number of age-related diseases [11], making the individual factors difficult to detect.

Since studies in model organisms have shown that ageing is characterized by many alterations at the molecular, cellular, and tissue level [12], a transcriptome analysis might lend greater insight than a static genetic investigation. Therefore, the aim of this study was to exploit a large-scale population-based strategy to systematically identify genes and pathways differentially expressed as a function of chronological age. In contrast to the relatively invariable genome sequence, the transcriptome is highly dynamic and changes in response to stimuli. Previous gene expression studies in the context of ageing have primarily focused on model organisms [13-15] or have been confined to specific ageing syndromes such as Hutchinson-Gilford progeria [16]. One report identified age-related expression-modules across four separate datasets [17], while other studies examined age-associated gene expression changes in relatively small cohorts [18-22].

To our knowledge, we perform here the first large-scale meta-analysis of human age-related gene expression profiles with well powered discovery and replication stages. In addition, this is the first large-scale study testing the hypothesis that changes in gene expression with chronological age are epigenetically mediated by changes of methylation levels at specific loci. Finally, we take advantage of our large set of samples to build a transcriptomic predictor of age, and we compare our transcriptomic prediction model with the epigenetic prediction models of Horvath [23] and Hannum *et al*. [24].

We identified 1,497 genes that are differentially expressed with chronological age. These genes are enriched for the presence of potentially functional CpG-methylation sites in enhancer and insulator regions. Our transcriptomic age predictor complements the existing epigenetic prediction models, and can be used by others to calculate transcriptomic age in external cohorts.

RESULTS

1,497 genes differentially expressed with chronological age

The discovery stage included six European-ancestry studies (n=7,074 samples) with whole-blood gene expression levels for roughly half of the genes in the human genome (n=11,908 significantly expressed genes across different platforms). We identified 2,228 genes with age-associated expression in the discovery stage (P<4.2E-6) after adjusting for technical variables and confounding factors such as sex, cell counts and cigarette smoking (Supplementary Figure 1A). The replication stage included 7,909 additional whole-blood samples, in which we replicated association results for 1,497 genes (P<2.2E-5). Discovery and replication results were highly correlated (r=0.972, Supplementary Figure 1B) and complete results are shown in Supplementary Data 1. After meta-analysis of discovery and replication stages, the expression levels of 897 genes were negatively associated and 600 genes were positively correlated with chronological age. The top 50 most significantly associated genes are presented in Table 1.

Transferability of ageing-transcriptome signatures

To examine the generalizability of the results of our differential expression meta-analysis, we tested whether the 1,497 identified genes were also differentially expressed in relation to chronological age in other ancestry samples, in brain tissue, and in specific blood sub-cell-types (Supplementary Data 1). In Native Americans (n=1,457), 95% of the 1,497 genes were significantly expressed, and 71% (1,005 genes) were associated with chronological age (p<0.05). In Hispanic Americans (n=1,244), 40% of the 1,497 genes were significantly expressed, and 74% (440 genes) were associated with chronological age in the same direction (p<0.05). In African Americans (n=359), 99% of the genes were significantly expressed, and 27% (392 genes) were associated with chronological age in the same direction (p<0.05) (Supplementary Table 1).

In both types of brain tissue studies (cerebellum and frontal cortex, n=394), approximately 58% of the 1,497 genes were significantly expressed. Of these, 19% (163 genes) and 26% (229 genes) were associated with chronological age in the same direction (p<0.05) in cerebellum and frontal cortex, respectively (Supplementary Table 2, Supplementary Figure 2, and Supplementary Table 3). Among the top 50 age-associated genes, three genes were associated with chronological age in all tissues: *SERPINE2*, *LDHB*, and *BZW2* (p<0.05) (Supplementary Data 2).

Table 1. Top 50 age-associated genes. For the 50 most significant age-associated genes, the discovery P-value (and Z-score), the replication P-value (and Z-score), and the meta-analysis P-value (and sample size and Z-score) are shown. The last two columns display whether the genes were also significantly associated with age in the brain tissues cerebellum and frontal cortex.

		Discovery		Replication		Meta-analysis			Generalization	
Gene	Rank	Z-score	P-value	Z-score	P-value	N	Z-score	P-value	Cerebellum	Frontal Cortex
CD248	1	-32.48	2.32E-231	-40.13	4.07E-352	15,266	-51.46	1.62E-577	NA	NA
LRRN3	2	-29.12	2.03E-186	-33.55	7.81E-247	15,266	-44.38	3.53E-430	N	Y (-)
NELL2	3	-23.65	1.18E-123	-23.48	6.93E-122	15,266	-33.31	2.67E-243	N	Y (-)
LEF1	4	-22.18	5.57E-109	-22.46	9.38E-112	15,266	-31.56	1.22E-218	NA	NA
CCR7	5	-21.14	3.59E-99	-22.44	1.48E-111	15,266	-30.83	1.04E-208	NA	NA
ABLIM1	6	-22.32	2.34E-110	-20.73	1.71E-95	15,266	-30.41	4.41E-203	N	Y (+)
GZMH	7	18.68	7.03E-78	20.97	1.26E-97	15,266	28.07	2.39E-173	NA	NA
MYC	8	-18.96	3.36E-80	-19.51	9.94E-85	15,266	-27.20	5.96E-163	NA	NA
CD27	9	-17.65	1.07E-69	-20.68	5.13E-95	15,266	-27.15	2.76E-162	NA	NA
FAM102A	10	-19.46	2.24E-84	-18.68	7.11E-78	15,266	-26.95	5.68E-160	N	Y (+)*
SERPINE2	11	-16.08	3.71E-58	-20.95	1.91E-97	14,385	-26.34	7.66E-153	Y (-)	Y (-)**
SLC16A10	12	-20.39	2.29E-92	-16.51	3.15E-61	13,809	-26.15	1.00E-150	Y (+)	Y (-)
FCGBP	13	-15.76	5.50E-56	-20.83	2.49E-96	15,266	-25.95	1.65E-148	NA	Y (+)*
GPR56	14	17.52	9.47E-69	19.02	1.21E-80	15,266	25.86	2.03E-147	NA	NA
BACH2	15	-17.82	4.64E-71	-17.75	1.85E-70	15,266	-25.14	1.71E-139	N	NA
SYT11	16	17.23	1.72E-66	18.23	3.24E-74	15,266	25.08	8.82E-139	Y (-)	Y (-)
PDE9A	17	-17.21	2.22E-66	-18.20	5.44E-74	15,266	-25.05	1.91E-138	N	N
NG	18	-17.01	7.41E-65	-17.52	9.87E-69	15,266	-24.42	1.16E-131	NA	NA
FLNB	19	-15.78	4.26E-56	-18.61	2.87E-77	15,266	-24.36	4.94E-131	N	Y (+)**
NT5E	20	-17.45	3.29E-68	-16.59	8.23E-62	15,039	-24.06	6.98E-128	NA	NA
FGFBP2	21	17.45	3.51E-68	15.79	3.51E-56	15,266	23.47	8.43E-122	NA	NA
TGFBR3	22	15.00	7.73E-51	17.66	9.15E-70	15,266	23.13	2.41E-118	N	Y (+)*
ITM2C	23	-14.41	4.24E-47	-17.73	2.45E-70	15,266	-22.78	7.22E-115	N	N
ATF7IP2	24	-15.52	2.73E-54	-16.61	5.85E-62	15,266	-22.73	2.34E-114	NA	Y (-)*
CR2	25	-16.29	1.10E-59	-15.85	1.51E-56	15,266	-22.71	3.49E-114	NA	NA
FAIM3	26	-17.92	8.65E-72	-14.22	7.40E-46	15,266	-22.65	1.41E-113	NA	NA
PHGDH	27	-13.25	4.56E-40	-18.30	8.10E-75	15,266	-22.39	4.85E-111	N	Y (+)*
LDHB	28	-15.63	4.33E-55	-15.96	2.42E-57	15,266	-22.34	1.55E-110	Y (-)*	Y (-)**
SIRPG	29	-15.64	4.16E-55	-15.45	7.71E-54	15,266	-21.97	5.58E-107	NA	NA
FCRL6	30	13.29	2.83E-40	17.65	9.90E-70	15,266	21.95	9.70E-107	NA	NA
PDE7A	31	-15.58	9.40E-55	-15.37	2.68E-53	15,266	-21.88	4.42E-106	NA	NA
NSIP	32	-14.44	3.12E-47	-16.19	5.74E-59	15,266	-21.68	3.13E-104	N	N

Table 1. *(Continued)*

Gene	Rank	Discovery		Replication		Meta-analysis			Generalization	
		Z-score	P-value	Z-score	P-value	N	Z-score	P-value	Cerebellum	Frontal Cortex
PAICS	33	-16.00	1.26E-57	-14.34	1.29E-46	15,266	-21.42	9.39E-102	N	Y (+)**
BZW2	34	-14.93	2.19E-50	-15.18	4.55E-52	15,266	-21.29	1.42E-100	Y (-)**	Y (-)**
OXNAD1	35	-15.59	9.09E-55	-14.32	1.71E-46	15,266	-21.12	5.66E-99	NA	NA
CX3CR1	36	14.09	4.14E-45	15.66	3.04E-55	14,385	21.07	1.67E-98	NA	NA
SCML1	37	-14.00	1.58E-44	-15.69	1.92E-55	15,266	-21.01	5.02E-98	NA	NA
RPL22	38	-14.91	3.03E-50	-14.79	1.79E-49	15,266	-20.99	8.61E-98	N	Y (-)**
LDLRAP1	39	-14.57	4.19E-48	-14.82	1.15E-49	15,266	-20.78	6.69E-96	N	NA
RHOC	40	12.89	4.89E-38	15.93	3.71E-57	15,266	20.43	8.94E-93	N	Y (+)
LTB	41	-14.90	3.55E-50	-14.02	1.11E-44	15,266	-20.43	9.52E-93	NA	NA
FAM134B	42	-15.17	5.88E-52	-13.43	3.96E-41	15,266	-20.19	1.31E-90	N	N
LBH	43	-14.18	1.29E-45	-14.22	7.04E-46	15,266	-20.07	1.28E-89	NA	Y (-)**
PRSS23	44	14.07	5.76E-45	14.07	6.25E-45	15,266	19.89	5.11E-88	NA	NA
SUSD3	45	-14.26	4.01E-46	-13.91	5.30E-44	14,385	-19.87	6.90E-88	NA	NA
PIK3IP1	46	-14.93	2.02E-50	-13.13	2.16E-39	15,266	-19.81	2.58E-87	Y (+)*	Y (+)**
MFGE8	47	-12.46	1.23E-35	-15.34	4.09E-53	15,266	-19.70	2.06E-86	N	N
AGMAT	48	-13.77	4.14E-43	-14.09	4.34E-45	15,266	-19.70	2.31E-86	NA	NA
NKG7	49	14.43	3.17E-47	13.42	4.53E-41	15,266	19.67	3.67E-86	NA	NA
PPP2R2B	50	13.49	1.81E-41	14.26	4.19E-46	15,266	19.63	9.40E-86	Y (-)*	Y (-)

Y = p<0.05; Y = p<0.01; Y** = p<0.0001; N = p>=0.05; NA= not expressed; N = sample size; (-) or (+) gives the direction of the effect with age. Y = p<0.05; Y* = p<0.01; Y** = p<0.0001; N = p>=0.05; NA= not expressed; N = sample size; (-) or (+) gives the direction of the effect with age.*

Novel and known age-associated genes and pathways

To differentiate between changes caused by cell composition and other biological mechanisms, we clustered genes based on co-expression networks in GeneNetwork (see Methods) and performed pathway analysis on the clusters of co-expressed genes. Among the negatively age-correlated genes, three major clusters were identified (Figure 1, Supplementary Data 3A-M). The largest group (cluster #1, 109 genes) consisted of three sub-clusters enriched for: 1a) *RNA metabolism functions*, *ribosome biogenesis*, and *purine metabolism*, 1b) *multiple mitochondrial and metabolic pathways* including ten mitochondrial ribosomal protein (MRP) genes consistent with earlier ageing studies in mice, *Caenorhabditis elegans* [25] and *Drosophila melanogaster* [26-28], and 1c) *DNA replication, elongation and repair*, and *mismatch repair* [26]. The second cluster of negatively correlated genes (cluster #2, 57 genes) contained factors related to immunity; including *T- and B- cell signaling genes*, and *genes involved in hematopoiesis*. The third tight cluster (cluster #3) included 12 genes, of which 11 encoded cytosolic ribosomal subunits*:* 7 *RPL*-genes *(RPL8, RPL11, RPL18, RPL28, RPL30, RPL35*, and

RPL36), 3 *RPS*-genes (*RPS14, RPS16,* and *RPS29*) and *UBA52* (*ribosomal protein L40*). The other gene of the cluster (#12) was *NACA, a nascent polypeptide-associated complex alpha subunit*. The protein encoded by the *NACA* gene forms the nascent polypeptide-associated complex (NAC), which binds to nascent proteins as they emerge from the ribosome [29]. Strikingly, the mRNA abundance of many genes encoding ribosomal subunits and mitochondrial ribosomal proteins were significantly associated with chronological age: 34 ribosomal genes were significantly associated, of which 33 were negatively correlated with chronological age (Supplementary Table 4), and 10 mitochondrial ribosomal protein (MRP) genes were significantly negatively correlated with chronological age (Supplementary Table 5).

The positively age-correlated genes revealed four major clusters (Figure 1, Supplementary Data 3N-V): cluster#1 (77 genes): *innate and adaptive immunity*, cluster#2 (9 genes): *actin cytoskeleton, focal adhesion, and tight junctions*, cluster#3 (8 genes): *fatty acid metabolism and peroxisome activity*, and cluster#4 (6 genes): *lysosome metabolism and glycosaminoglycan degradation*.

For both brain tissue studies, we checked the number (and %) of overlapping age-associated genes for the different functional clusters: 24 genes (11.7% of the genes expressed in cerebellum) and 33 genes (of the genes expressed in frontal cortex) of all pathway genes (278 genes) were associated with chronological age (Supplementary Tables 6 and 7). In cerebellum, the best replicating pathway was the positively age-correlated cluster #4: *lysosome metabolism and glycosaminoglycan degradation*. In frontal cortex, the best replicating pathway was the positively age-correlated cluster #2: *actin cytoskeleton, focal adhesion, and tight junctions.*

Associations with prior ageing candidate genes

We investigated the intersection between genes significantly associated with chronological age in our study and candidate genes from previous human and animal studies (170 genes, see Supplementary Tables 8 and 9). Thirty-three of the 170 candidate genes were significantly associated with chronological age in our WB meta-analysis, including members of the mTOR/FOXO pathways (*FOXO1, VEGFB, EIF4G3, SREBF1, STAT3, and RPS6KB1*) [30], DNA repair (*ATM*) [31], and prior multi-species candidates (*LDHB, IGJ, IRF8,*and *FCGR1A*). Twenty-eight of the 33 significant age-associated genes (~85%) have the same expression directionality in our CHARGE meta-analysis as previously reported in a variety of studies in humans and other model organisms.

Premature ageing syndrome genes *ATM* (ataxia-telangiectasia), *DKC* (dyskeratosis congenita) and *WRN* (werner syndrome) all exhibited lower transcript abundance in older individuals, concordant with loss-of-function alterations in disease-related mutations. Based on the co-expression analyses, these genes clustered together with genes encoding proteins involved in DNA and RNA metabolism, DNA repair, and purine/pyrimidine metabolism. The Hutchinson-Gilford progeria gene *LMNA* showed higher mRNA levels in the elderly, consistent with earlier findings in muscle [32], and clustered with actin remodeling genes.

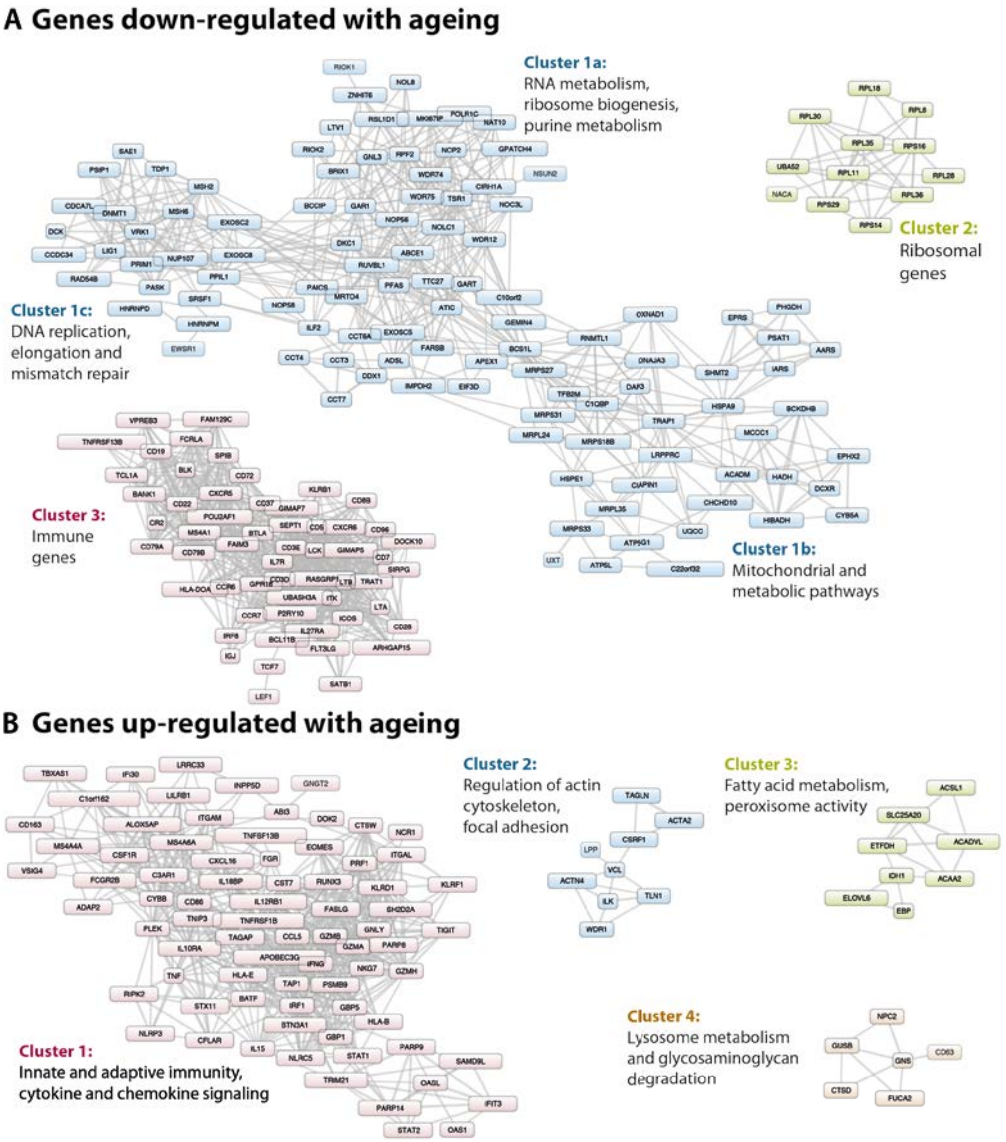


Figure 1. Pathway analysis on the clusters of co-expressed genes. We ran a co-functionality network analysis on 897 down-regulated genes with age (negative effect direction) and 600 up-regulated genes with age (positive effect direction) using GeneNetwork. With a correlation threshold of 0.7, we selected all clusters bigger than 4 genes and ran per-cluster pathway analyses using KEGG, Reactome, and GO-terms in WEBGESTALT. Benjamini & Hochberg FDR was used for multiple testing corrections. The significant threshold 0.05 after correction for multiple testing was applied. (**A**) Three clusters of down-regulated genes with age and (**B**) four clusters of genes up-regulated with age were enriched for functional pathways in KEGG, Reactome, and GO terms; the specific pathways are mentioned next to the (sub)cluster names.

Methylation association patterns for top age-associated loci

Given the possible role of the methylome in ageing, we investigated whether age-associated methylation accompanied age-associated expression for the 1,497 age-associated genes. We analyzed methylation of 135,230 CpG sites (regions of DNA where a cytosine nucleotide occurs next to a guanine nucleotide) in or near (±250kb) the age-associated genes in WB or PBMCs from seven cohorts (N=3,073). We chose CpGs in a 250kb vicinity because earlier studies have shown that methylation can regulate gene expression levels at this distance [33], and that long range enhancer activities are present and actively regulate gene expression at a wide scale [34]. We observed significant associations between methylation and chronological age for 31,331 CpG sites, and between expression and methylation for 12,280 CpG sites, based on a conservative Bonferroni threshold (P<3.7E-7) (top results for each gene in Supplementary Data 4). 1,248 of the 1,497 age-associated genes (83%) had ≥1 significant mediating CpGs and the number of significant mediating CpGs per gene ranged from 1 to 154 (Supplementary Data 4).

To test whether the age-associated genes were enriched for nearby CpG methylation sites associated with chronological age or expression, we performed a similar analysis for a set of 1,497 randomly selected genes matched for similar gene length and mean WB expression (see Methods and Supplementary Figures 3A-D). Compared to the set of random genes, age-associated genes had only mild enrichment for CpG methylation sites associated with chronological age (Figure 2A; Odds Ratio (OR) =1.04; 95% Confidence Interval (CI) =1.02-1.06; P=7.9E-5), but strong enrichment for CpG methylation sites associated with expression (Figure 2B; OR=2.68; 95%CI=2.58-2.78; P<1E-300). This pattern was consistent across all cohorts (Supplementary Figure 4) and within subsets of CpG methylation sites annotated to specific biological features (i.e., enhancer regions, promoter regions, CpG islands, etc.) (Supplementary Figure 5), and was robust to the entire range of significance thresholds (see Methods). This is consistent with a scenario where many methylation sites associate with chronological age, but only those with regulatory potential lead to altered transcript expression with chronological age.

We used Sobel tests (see Methods) for all CpG methylation sites to investigate whether the observed patterns could potentially reflect a methylation-mediated relationship between chronological age and transcript levels. In total, 1,248 of the 1,497 age-associated genes (83%) had ≥1 CpG site with a significant Sobel test after Bonferroni adjustment for the number of CpGs tested (Supplementary Data 4). These potentially mediating CpG sites were less likely to reside in CpG islands (OR=0.28; 95%CI=0.26-0.30; P<1E-300) or in promoters (OR=0.38; 95%CI=0.36-0.40; P<1E-300) and more likely to be located in enhancers (OR=2.29; 95%CI=2.17-2.41; P=2.7E-188) and insulators (OR=1.44; 95%CI=1.23-1.67; P=6.6E-6), compared to non-mediating CpGs within 250kb of age-associated genes (Supplementary Figure 6). This pattern is again consistent with the mediation of age-associated transcripts by age-associated methylation of CpG sites with specific regulatory roles.

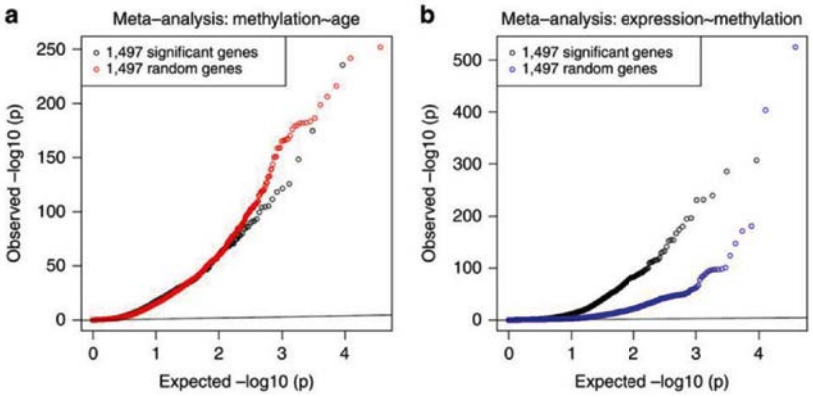


Figure 2. Age-associated genes are enriched for the presence of potentially functional methylation sites. (**A**) Quantile-Quantile (QQ) plot of the observed p-values (-log10P) for the methylation~age associations. The plot in black shows p-values from the 1,497 significant age-associated genes, whereas the plot in red shows p-values for 1,497 random genes. We do not see enrichment for the 1,497 age-associated genes. (**B**) QQ plot of the observed p-values (-log10P) for the expression~methylation associations. The plot in black shows p-values from the 1,497 significant age-associated genes, whereas the plot in blue shows p-values for 1,497 random genes. The age-associated genes are enriched for CpG methylation sites that associate with gene expression levels.

Transcriptomic age prediction as a surrogate biomarker

All 11,908 discovery genes were used to build a predictor for age using a leave-one-out-prediction meta-analysis (see Methods). For each cohort in turn, we left out that cohort as the validation sample and re-ran the discovery meta-analysis on the other cohorts to avoid overlap between the discovery and validation sample (Supplementary Data 5A). The difference between the predicted transcriptomic age and chronological age (*delta age*) may be a reflection of altered biological age (see Methods). The correlation between chronological age and transcriptomic age was significant in all cohorts (P<2E-29) (Figure 3A-H). The average absolute difference between predicted age and chronological age was 7.8 years (n=8,847 samples, Supplementary Table 10). A positive delta age, interpreted as reflecting more rapid biological ageing, was consistently associated with higher systolic and diastolic blood pressure, total cholesterol, HDL cholesterol, fasting glucose levels, and body mass index (BMI) (Table 2, Supplementary Table 11). All analyses were adjusted for chronological age, and after adjustment for BMI all phenotypes remained associated in the same direction (Table 2, Supplementary Table 12). For systolic blood pressure, the added predictive value of the transcriptomic predictor over chronological age is shown for the Rotterdam Study (Figure 4A-C). Other phenotypes showed the same pattern.

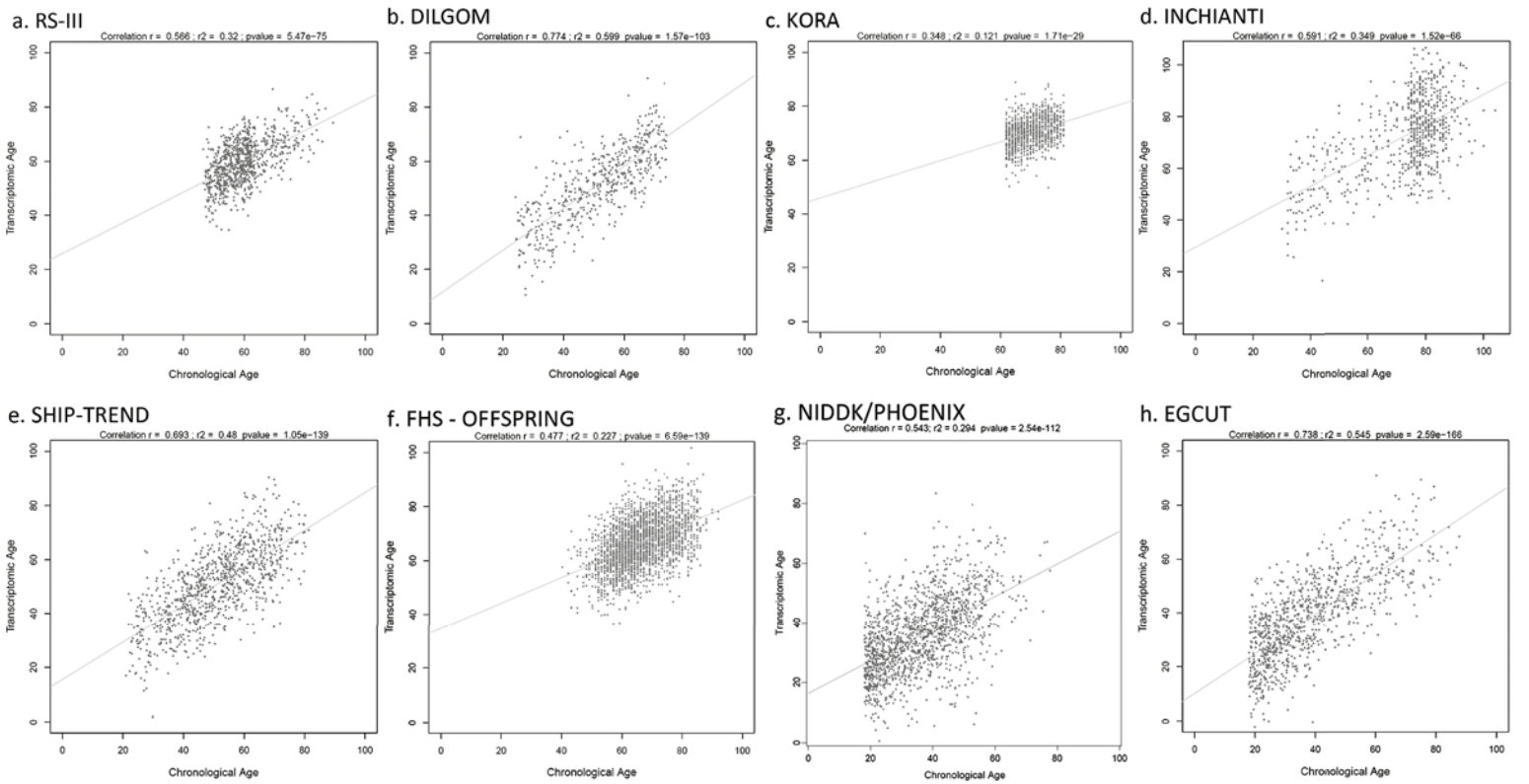


Figure 3. Transcriptomic age versus chronological age. This figure represents the correlations between chronological age (x-axis) and transcriptomic age (y-axis) in eight different cohorts: (**A**) RS-III, (**B**) DILGOM, (**C**) KORA, (**D**) InCHIANTI, (**E**) SHIP-TREND, (**F**) FHS-OFFSPRING, (**G**) NIDDK/PHOENIX, and (**H**) EGCUT. Transcriptomic age was calculated using a cohort-specific prediction formula and the measured gene expression levels of 11,908 genes. The correlation between chronological age and transcriptomic age was significant in all cohorts (P<2E-29).

Table 2. Meta-analysis of associations between transcriptomic delta age with twelve biological ageing phenotypes.

	Adjusted for Chronological Age				Adjusted for Chronological Age + BMI			
Phenotype of Interest	**Z-score**	**P-value***	**Direction**	**N**	**Z-score**	**P-value***	**Direction**	**N**
Sex: 0 = male, 1 = female	-2.7610	5.76E-03	--+--+-+	8,836	0.7500	4.53E-01	--+++--+	8,829
Systolic bloodpressure: mmHg	9.8510	6.78E-23	++++++++	8,571	9.3740	6.97E-21	++++++++	8,564
Diastolic bloodpressure: mmHg	7.7200	1.16E-14	++++++++	8,568	6.8020	1.03E-11	++++++++	8,561
Total cholesterol levels: mmol/L	5.4190	5.99E-08	+++++-++	8,688	4.6370	3.53E-06	+++++-++	8,681
HDL cholesterol levels: mmol/L	4.4630	8.07E-06	+++++-+-	8,687	5.8310	5.52E-09	+++++-++	8,680
Fasting glucose levels: mmol/L	6.9330	4.11E-12	++++++??	7,330	5.8920	3.82E-09	++++++??	7,323
Body Mass Index: kg/m^2	5.3860	7.21E-08	++++++++	8,829	NA	NA	NA	NA
Waist Hip Ratio	3.3800	7.25E-04	++??++++	4,837	1.9370	5.27E-02	++??++++	4,837
Hand grip strength: kg	-1.5120	1.31E-01	++?-????	3,651	-1.1760	2.40E-01	++?-????	3,651
Renal function	0.8740	3.82E-01	+++-+?-?	7,317	-0.4890	6.25E-01	+++-+?-?	7,310
Mini Mental State Exam Score	-1.3130	1.89E-01	--??????	1,492	-1.3810	1.67E-01	--??????	1,492
Current smoking: 0 = no, 1 = yes	5.5100	3.59E-08	+-?+++--	7,379	3.2040	1.36E-03	+-?+++--	7,379

BMI = body mass index; NA = not available.

We tested whether the transcriptomic delta age was associated with twelve biological phenotypes known to be associated with chronological age. Gene expression levels were adjusted for plate ID, RNA quality score, fasting state, sex, smoking status, and cell counts. Association results of all cohorts were meta-analyzed. After adjustment for chronological age and BMI (right columns), systolic blood pressure, diastolic blood pressure, total cholesterol levels, HDL cholesterol levels, and fasting glucose levels were significantly positively associated with delta age (p<4.17E-3). Samples predicted to be older (positive delta age) consistently had higher levels for these ageing phenotypes.

Delta age = transcriptomic age – chronological age; + Z-score = increasing phenotype with higher predicted age; - Z-score = decreasing phenotype with higher predicted age;

**if P-value<(0.05/12=4.17E-3), significance has been reached.*

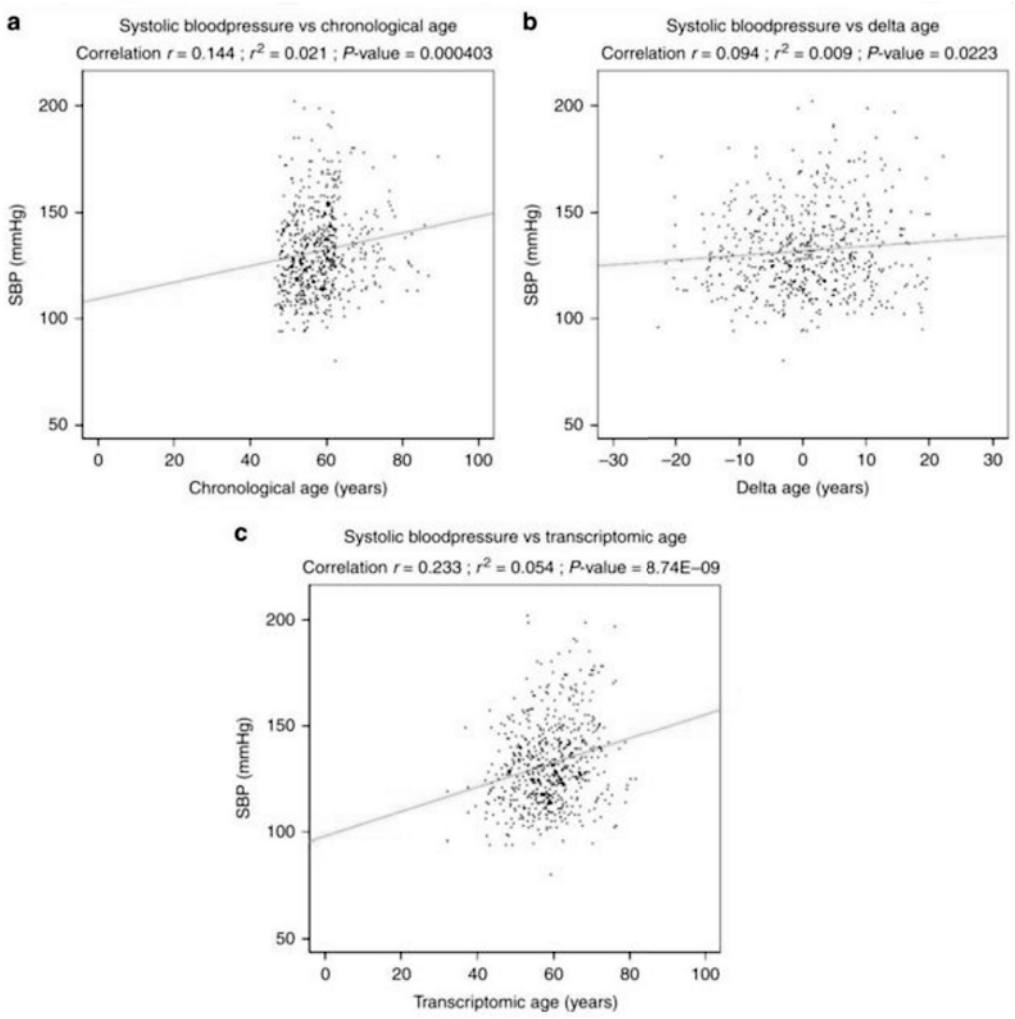


Figure 4. The added value of the transcriptomic predictor. To show the added value of the transcriptomic predictor, we choose one biological ageing phenotype systolic blood pressure (SBP), and plotted its correlation with chronological age (**A**), delta age (**B**), and the transcriptomic age (**C**) in the Rotterdam Study (n=597 samples with SBP data available). Delta age represents the difference between chronological age and transcriptomic age. SBP was plotted on the y-axis, and the age-related values were plotted on the x-axes. SBP was significantly associated with chronological age (P=4.0E-04), but SBP was even stronger associated with transcriptomic age (calculated with a cohort specific prediction formula based on gene expression levels) (P=8.7E-09), Therefore, the transcriptomic predictor adds value over chronological age alone. Other biological ageing phenotypes showed the same pattern.

We compared our transcriptomic predictor with two already published epigenetic predictors of age of Horvath [23] and Hannum *et al*. [24] in 1,396 individuals from the KORA study and the Rotterdam Study, all having gene expression levels and methylation data available. The transcriptomic predictor was less strongly correlated with chronological age than the two epigenetic predictors (Supplementary Figure 7), which can be explained by the different data types used: we used gene expression data instead of DNA methylation data.

Transcriptomic age and epigenetic age (both Hannum and Horvath) were positively correlated, with r^2 values varying between 0.10 and 0.33 (Supplementary Figure 7). Interestingly, all three age predictors were associated with different ageing phenotypes (Supplementary Tables 13 and 14), i.e., the transcriptomic predictor was significantly associated with systolic blood pressure, waist-hip-ratio, and smoking; the epigenetic Horvath predictor was associated with waist-hip-ratio only; and the epigenetic Hannum predictor was associated with fasting glucose, waist-hip-ratio, and smoking (all analyses were adjusted for chronological age, sex, and BMI). By adding two predictors into one formula (one transcriptomic predictor and one epigenetic predictor), both predictors added value (significant effect) to the phenotype associations, i.e. for waist-hip-ratio in KORA (explained variance transcriptomic predictor=0.015%, Horvath predictor=0.005%, Hannum predictor=0.006%; transcriptomic+Horvath=0.017%, and transcriptomic+Hannum=0.016%) (Supplementary Tables 15 and 16).

DISCUSSION

Age-associated changes in gene expression levels point towards altered activity in defined age-related molecular pathways that may play vital roles in the mechanisms of increased susceptibility to ageing diseases. In contrast to earlier, smaller studies [17-21] of human age-related molecular differences, we detected and replicated 1,497 age-associated genes in 14,983 individuals of European ancestry. Additionally, many of our associations were generalized across different ancestries and multiple cell and tissue types. Because we had much smaller sample sizes for both brain tissue (n=394) and the other ancestry groups (1,244 Hispanic Americans, 1,457 Native Americans, and 359 African Americans), we used a nominal p-value threshold ($P<0.05$) in these specific sub-analyses. Larger sample sizes will ultimately be needed to fully understand the transferability of the ageing-transcriptome signatures.

A potential limitation of our study is that we relied on a linear regression model to identify age-associated genes. A linear model assumes constant change over age, which may not be always correct in biological processes that stretch over several decades (adulthood). A recent study demonstrated that a quadratic regression model has a higher statistical fit to cross-sectional gene expression datasets over linear models [35]. Although we chose to apply a linear regression model in our study, we recognize that more complex models could be investigated in future studies.

Our human age-expression and pathway enrichment analysis results were consistent with known ageing mechanisms including dysregulation of transcription and translation, metabolic function, DNA damage accumulation, immune senescence, ribosome biogenesis, and mitochondrial decline. Houtkooper *et al.* [25] and others [26,27] highlighted the key role of mitochondria in ageing and longevity in model organisms. Mitochondria regulate a multitude of different metabolic and signaling pathways and also play an important role in programmed cell death [36]. The number of mitochondria decreases and their capacity to produce energy is reduced with chronological age [37-39]. Consistent with these reports, a large number of mitochondrial ribosomal proteins (*MRPL24, MRPL3, MRPL35, MRPL45, MRPS18B, MRPS26, MRPS27, MRPS31, MRPS33,* and *MRPS9*) showed lower expression at higher chronological age in our study, supporting the hypothesis that age-dependent mitochondrial dysfunction plays a causal role in human ageing.

The large immune function associated clusters (cluster #2 and cluster #1 of the negatively and positively correlated genes, respectively) reflect immune senescence. The relative abundance of immune cells in WB shifts with ageing, with naïve T-cells decreasing and highly differentiated effector and memory T-cells increasing with chronological age [28,40-44]. Consistent with immune senescence, the mRNA abundance of the chemokine receptor *CCR7* and cell differentiation antigens *CD27* and *CD28* was lower in older individuals (P=1.0E-208, P=2.8E-162, and P=5.8E-59). Notably, these results were consistent in many of the blood sub-cell-types. For example, *CCR7* was lower in older individuals across multiple cell types including CD4+ cells (P=1.0E-08), CD8+ cells (P=3.0E-15), CD14+ cells /monocytes (P=8.5E-3), and PBMCs (P=3.0E-3). This suggests that genes in the immune associated clusters reflect a biological function related to a more general ageing phenotype , at least in multiple immune cell types, and are not solely accountable to cell count differences. We also note that cell subset classification is to a greater or lesser extent artificial, reflecting our current ability to distinguish cells based on specific small sets of available markers. Accepted subpopulation of cells can often be further broken down into additional subgroups as the tools for such classification become more sophisticated. The analysis of unfractionated cell populations (such as our study) adds a layer of complexity to the interpretation, but is not necessarily less informative than the analysis of marker defined subpopulations.

Aside from the *immune clusters*, we identified and newly emphasized pathways associated with human ageing, for example, *glycosaminoglycan degradation* and *actin remodeling*. These pathways have previously been implicated in life span regulation of the model organisms *Caenorhabditis elegans* and *Drosophila melanogaster* [45-47]. Glycosaminoglycans (GAGs) influence cell migration, proliferation and differentiation, and play a role in wound healing [48,49]. Impaired degradation of GAGs in extreme lysosomal storage disorders lead to chronic, progressive effects on a variety of organs and physiologic systems [50]. Tissue repair and regeneration are known to be impaired in the elderly and inhibition of GAG degradation may be therapeutic in these contexts [51]. Our findings suggest GAG degradation as a candidate mechanism for the age-associated changes. The actin cytoskeleton is a critical structural element in eukaryotic cells that is crucial in mediating cell responses to both internal and external signals in yeast [52]. Actin dynamics have clearly been linked

to yeast replicative ageing through both reactive oxygen species mediated apoptosis and through selective sequestration of healthy mitochondria to new daughter cells during cell division [52,53]. Our pathway analysis indicates that the actin cytoskeleton may be similarly important in human ageing. While much prior effort in targeting actomyosin dynamics has been aimed at cancers, recent studies indicate that targeted modulation of these systems could also have benefits in immune-mediated pathologies [48].

In addition to these novel candidate pathways, our 1,497 age-associated genes contain genes in many pathways known to be associated with ageing. Beyond the immune-related pathways, we confirm an age-associated role for *mitochondrial function* [54], *metabolic function* [12], *ribosome biogenesis* [55], *DNA replication, elongation and repair* [56,57], *focal adhesion* [58], and *lysosome metabolism* [59], and suggest a number of new potential age-related targets within these pathways, including *TTC27* (ribosome biogenesis); *CCDC34* (ribosomal cluster); *ARHGAP15, DOCK10, FAM129C, FCRLA, GIMAP7,* and *VPREB3* (T- and B- cell signaling genes and genes involved in hematopoiesis); *GZMH, SAMD9L,* and *TAGAP* (innate and adaptive immunity). Of note, overexpression of the full-length *ARHGAP15* protein in COS-7 and HeLa cells resulted in an increase in actin stress fibers and cell contraction, relating the newly ageing emphasized actin remodeling pathway and the focal adhesion pathway in ageing to immune cell changes [60]. Thus, by using co-expression networks, we identified new genes and pathways that are likely important in human ageing, opening new avenues of inquiry for future studies.

Age-related epigenetic changes have recently been examined including a large study combining data across 7,844 non-cancer samples from 82 individual datasets to define a set of age-methylation clock genes. Only 35 of our 1,497 age-related genes were found among the genes harboring the 353 age-methylation clock CpG sites reported by Horvath [23], suggesting that our age-associated genes may not be particularly enriched for age-associated CpG methylation sites. To test this formally, we analyzed the DNA methylation sites (CpG sites only) within 250kb (upstream and downstream) of all 1,497 age-associated genes, as well as a comparison set of 1,497 randomly chosen unassociated genes. We observed that the genes exhibiting age-associated transcript levels in blood are much more likely than other genes to harbor CpG methylation sites that associate with expression levels, but are not substantially more likely to harbor methylation differences in close CpG sites associated with chronological age. These results suggest that genes showing age-related expression differences are characterized primarily by the presence of nearby CpG sites with regulatory potential, rather than by the presence of age-associated CpG methylation sites, which are abundant everywhere in the genome. A limitation of our study is that we used the Illumina Infinium HumanMethylation450K BeadChip Array for measuring methylation levels: this array queries only 1.6% of all CpGs in the genome and the CpG selection is biased towards CpG islands. In addition, we did not examine non-CpG methylated sites, which have recently been suggested to play a role in regulating gene expression as well [61]. Other techniques – whole genome bisulfite sequencing [62] and methylC-capture (MCC) sequencing [63] for example – have definite technical advantages (higher resolution

and no CpG island selection bias), but these have currently not been applied to a large number of samples.

Although the CpG selection on the methylation array is biased towards CpG islands, the CpG sites for which methylation was associated with both expression and chronological age were strongly enriched for enhancer activity. This is consistent with the concept that methylation at enhancers is more variable and may regulate gene expression in development [64] and/or in environmental responses, while promoter methylation is comparatively stable. Interestingly, the age- and expression-associated CpGs were also enriched at insulators, which function to block the communication between an enhancer and a promoter, thereby preventing inappropriate gene activation. Taken together, these results suggest that the age-associated genes reported here may be regulated by methylation of CpG sites in specific functional regions, and that studying both methylation and expression as potential joint effectors of the ageing process may significantly improve the prediction of age and identification of novel age-related genes and pathways.

Using gene expression levels as a predictive biomarker indicated that individuals having higher predicted than chronological age also have clinical features consistent with an older age, such as higher blood pressure and total cholesterol levels. Developing a strongly predictive gene expression set as a biomarker panel has clinical potential to identify subjects at risk for early biological ageing, and provide a tool for targeting susceptible individuals for early intervention. It remains to be seen whether the transcriptomic age can serve as a surrogate marker to predict age-associated decline in other tissues. Therefore, the development of a robust transcriptomic predictor for age will require independent and prospective validation across different tissues.

We observed that both the transcriptomic predictor and the epigenetic predictors were significantly associated with a number of phenotypes, but that the pattern of association differed among the predictors. Therefore, the transcriptomic age and the epigenetic age should be combined to obtain the optimal biological age prediction. A general transcriptomic prediction formula has been calculated that is freely available (Supplementary Data 5B). These results suggest that the biological mechanisms behind the transcriptomic and the epigenetic predictors are different. The exact mechanism of these differences need further examination in larger sample sizes and subgroup analysis were different diseases are studied. In addition, the predictors need to be evaluated for their prognostic value. In conclusion, gene expression levels are likely to become a valuable addition to evolving indicators of age based on epigenetic and telomeric age predictors. Ideally, a combination of transcriptomic, epigenetic, and telomeric elements could further improve and refine age prediction.

In conclusion, we have identified a compendium of genes and pathways associated with human chronological age. By leveraging transcriptional information across large, multi-ethnic cohorts, different tissue types, and genomic repositories, we captured an unprecedented overview of the complex and temporally dynamic biological pathways orchestrating the ageing process. Our list of

genes should provide a rich trove of data for future ageing studies. While the pursuit of an anti-ageing panacea in humans remains a distant goal, our work has generated new biological hypotheses and will serve as a roadmap for future studies aimed at translating findings into treatment strategies for age-related diseases.

METHODS

Study design

We performed a differential expression meta-analysis in 7,074 human peripheral blood samples from six independent cohort studies, including EGCUT (n=1,086), FHS – 2nd generation (n=2,446), INCHIANTI (n=698), KORA (n=993), ROTTERDAM STUDY (n=881), and SHIP-TREND (n=970) (Supplementary Table 17). Gene expression data for each dataset was obtained using either PAXGene (Becton Dickinson) or Tempus Tubes (Life Technologies), followed by hybridization to Illumina Whole-Genome Expression BeadChips (HT12v3 or HT12v4) or Affymetrix Human Exon 1.0 ST GeneChips.

We replicated the significantly associated transcripts in 7,909 peripheral blood samples from seven independent cohort studies, including BSGS (n=862), DILGOM (n=512), FEHRMANN (n=1,191), FHS – 3rd generation (n=3,180), GTP (n=359), HVH (n=121 on the Illumina HT12v3 platform and n=227 on the Illumina HT12v4 platform), and NIDDK/PHOENIX (n=1,457) (Supplementary Table 18). Gene expression data for these datasets was also obtained using either PAXGene (Becton Dickinson) or Tempus Tubes (Life Technologies), followed by hybridization to Illumina Whole-Genome Expression BeadChips (HT8v2, HT12v3, or HT12v4 arrays) or Affymetrix Human Exon 1.0 ST GeneChips.

We generalized the significantly replicated transcripts in 4,644 samples with other tissue types, including: CD4+ cells of EGCUT (n=302) and a Boston sample (n=213), CD8+ cells of EGCUT (n=299), CD14+ cells (or monocytes) of a Boston sample (n=213) and MESA (n=354), LCLs of GENOA (n=869), lymphocytes of SAFHS (n=1,244), PBMCs of GARP (n=134) and PMBC-MS (n=228), and brain tissue (cerebellum and frontal cortex) of NABEC-UKBEC (n=394) (Supplementary Table 19). Gene expression data of these datasets was obtained by tissue specific RNA isolation and hybridization to Illumina Whole-Genome Expression BeadChips (WG6v1, HT12v3 or HT12v4), Affymetrix Human Exon Arrays, or Affymetrix Human Gene Arrays.
The study outline is summarized in Supplementary Figure 8. The study populations, the RNA isolation methods, the amplification and labeling methods, and the array types used for each study are described in the supplementary methods. The covariates used in each study are presented in Supplementary Tables 17-19.

Phenotype

Chronological age was defined as the length of time in years between birth and blood draw, using two decimals. Detailed descriptions of the chronological age distributions, fasting status, and the

available covariates from the participating cohorts are presented in Supplementary Tables 17-19 and Supplementary Figures 9A-V.

Illumina pipeline: gene expression probes and normalization procedure

The different Illumina platforms used by the different cohorts share a large number of probes with identical 50-mer probe sequences. Therefore, we harmonized the probes across the HT12-v3 and the HT12-v4 platforms by determining the probe sequences from the different annotation files for each platform; renumbering the probes on the basis of unique probe sequences. In total, we identified 56,330 unique Illumina probes (11,453 probes measured only on the HT12-v3 platform, 7,529 probes measured only on the HT12-v4 platform, 37,348 probes measured on both platforms). Genes were declared significantly expressed in the discovery data when 1) the detection p-values calculated by GenomeStudio were <0.05 in more than 10% of all discovery samples, and 2) the probes were measured in at least two cohorts. This resulted in 23,170 transcripts considered as being significantly expressed in our Illumina discovery; these transcripts code for 15,639 well characterized unique genes. 3,484 genes have more than one Illumina probe on the HT12 platform. Illumina gene expression data was quantile-normalized to the median distribution and subsequently log2-transformed. The probe and sample means were centered to zero.

Affymetrix pipeline: gene expression probes and normalization procedure

The Affymetrix platform generated CEL files, containing both gene-based and exon-based expression levels. We used the gene-based expression levels and normalized the data using Affymetrix Power Tools: probes with RLE mean values >3.0 (range 1.34-12.71) were considered to be significantly expressed. This resulted in 16,798 well characterized unique genes in the Affymetrix discovery. Samples with all probeset RLE means >0.7 were defined as outliers and excluded from further analysis. A genetic expression SNP analysis was undertaken to locate mislabeled samples and re-identify them where possible with high confidence. After exclusions and re-identification, the RMA normalization was repeated.

Differential expression with chronological age

All Illumina studies ran a least squares linear regression model (lm) using the normalized and standardized gene expression values as dependent variables, chronological age as an explanatory variable, and with adjustments for the potential confounders: sex (factor), fasting and smoking status (both factors), plate origin (factor), RNA quality (RIN/RQS), and cell counts (# of granulocytes, lymphocytes, monocytes, erythrocytes, and platelets), so:

lm(gene expression ~ chronological age + confounders + batch effects)

The Affymetrix cohort ran a multivariate stepwise PC regression, using the normalized and standardized gene expression values as dependent variables, chronological age as an explanatory variable, and the significant technical covariates: all_probeset_mean, all_probeset_stdev, neg_control_mean, neg_control_stdev, pos_control_mean, pos_control_stdev, all_probeset_rle_mean, all_probeset_mad_residual_mean, RNA quality (RIN), and RNA processing batch. Batch was included in modeling as a random factor while all others were fixed factors.

Meta-analysis of significantly expressed genes

We ran four separate meta-analyses: one for the studies using the Illumina platforms in the discovery phase, one for the Illumina discovery studies plus the FHS Affymetrix discovery results, one for the replication sample combined, and one for the discovery samples plus replication samples for validated results in order to re-rank the final results list. For these meta-analyses, we used a sample size weighted meta-analysis based on p-values and the direction of the effects; using p-values, a Z-statistic characterizing the evidence for association was calculated. The Z-statistic summarized the magnitude and the direction of the effect. An overall Z-statistic and p-value was calculated from the weighted sum of the individual statistics. Weights were proportional to the square-root of the number of individuals examined in each sample and standardized such that the squared weights sum to 1.

We calculated the Z-scores and p-values using the Meta-Analysis Tool for genome-wide association scans (METAL) [65]. METAL is a flexible and computationally efficient command line tool that was developed for meta-analyzing GWAS studies, but can easily be adapted to gene expression studies. Because we are dealing with gene expression levels and not SNPs, we changed the SNPID column to probe IDs and gave all probes a minor allele A and a major allele G, a minor allele frequency=0.10, and a + strand. For the positions, the probe chromosomes and the midpoint position of the probes were used. Sample sizes, effect directions, and p-values were extracted from the linear model results files. We extensively tested what input parameters to use for meta-analyzing gene expression data. By using similar allele names, allele frequencies, and allele strands for all cohorts, we forced METAL to use the default meta-analysis approach. We tested an inverse variance weighted meta-analysis (using the effect size estimates and the standard errors), and found that our METAL meta-analysis results were identical to the meta-analysis results using the R package Meta.

Meta-analysis of discovery samples

To calculate which genes are significantly associated with chronological age, we ran a sample size weighted meta-analysis based on p-values and the direction of the effects of the results of the Affymetrix and the Illumina meta-analyses. Combining the 16,798 Affymetrix probes and the 15,639 Illumina probes, these platforms have 11,908 genes significantly expressed in WB samples in common.

Replication phase. Genes with a p-value <4.20E-6 (0.05 / 11,908 genes tested) were considered transcriptome-wide significantly associated with chronological age. We replicated these findings in an additional 8,009 samples (Supplementary Table 18). Replication cohorts used the same analysis plan and R-scripts as the discovery phase, however, some covariates were not available in these cohorts and ethnicities could be different than European-ancestry.

Meta-analysis of the replication cohorts. We meta-analyzed the summary statistics of the replication cohorts using METAL. Genes were considered significantly replicating if P<2.23E-5 (0.05/2,238 genes tested) and the overall Z-score was in the same direction as the overall Z-score of the discovery meta-analysis.

Meta-analysis of discovery and replication cohorts. We additionally performed a meta-analysis based on the summary statistics of all discovery and all replication cohorts and obtained two-sided P-values.

Generalization phase

To see whether our findings are specific for WB, we tried to generalize our significantly replicating transcripts in samples of other tissue types, including CD4+ cells, CD8+ cells, CD14+ cells (monocytes), LCLs, lymphocytes, PBMCs, and brain tissue (both cerebellum and frontal cortex) (Supplementary Table 19). If we had data of one tissue type of more than one cohort, we ran a meta-analysis based on the summary statistics of both cohorts. Because sample sizes of these tissue types were very small, we considered p-values <0.05 (with an identical effect direction) sufficient to document generalization of the effect.

Pathway analysis of significant genes

We used WEBGESTALT (*http://bioinfo.vanderbilt.edu/webgestalt/analysis.php*) and GeneNetwork (*http://genenetwork.nl:8080/GeneNetwork/pathway_network.html*) for pathway analysis of age-associated transcripts. First, we ran the co-functionality network analysis separately on 897 down-regulated genes and 600 up-regulated genes, using a correlation threshold of 0.7. Of 897 down-regulated genes, 192 formed cluster groups at this threshold, and of 600 up-regulated genes, 114 formed cluster groups. We next re-ran the co-expression cluster analysis on these 192 and 114 genes, using a correlation threshold of 0.65 to see if small clusters could be merged together if a lower co-expression threshold was applied. We selected clusters with 5 and more genes for pathway analysis; in total 178 and 100 down- and up-regulated genes respectively. Based on the clustering analysis, we performed per-cluster pathway analysis. Pathways were selected using KEGG, Reactome and GO-terms. In WEBGESTALT Benjamini & Hochberg FDR was used for multiple testing corrections. The significant threshold 0.05 after correction for multiple testing was applied.

Analysis of chronological age, methylation, and expression

For 3,073 blood samples with methylation data available from the Illumina 450K array, we analyzed methylation for CpG sites within 250kb of the 1,497 genes identified in the differential expression meta-analysis. For this analysis we performed a new meta-analysis of samples from seven cohorts including EGCUT (n=82), InChianti (n=485), KORA (n=735), Rotterdam Study (n=726), BSGS (n=610), GTP (n=315), and GARP (n=120); all samples were derived from whole-blood except for GARP (PBMCs). After filtering (to remove non-specific probes and probes with SNPs in the probe target as documented by Price *et al.*), 135,230 CpG sites within 250kb of the 1,497 age-associated genes were eligible for analysis [66].

Within each cohort, we fit two linear regression models where we considered as our dependent variable either standardized gene expression values for a particular gene or methylation β-values, which are measures of the proportion of DNA methylated within a sample, for a particular CpG site. In Model 1, we regressed methylation β-values on chronological age. In Model 2, we regressed

gene expression on both methylation and chronological age. In both models we adjusted for the following potential confounders as available in each cohort: sex, fasting- and smoking status (both modeled as categorical variables or factors), and cell counts (# of granulocytes, lymphocytes, monocytes, erythrocytes, and platelets). In Model 1, where methylation was the dependent variable we adjusted for chip and row on chip (both as factors). In Model 2, where the dependent variable was gene expression we adjusted for plate origin (factor) and RNA quality (RIN/RQS). For each of the age-associated genes, we fit these models separately for each CpG site within 250kb (upstream or downstream) of the gene.

To combine results from these models across cohorts, we performed a sample size weighted meta-analysis based on the t-statistics from these models. For each model, we calculated a Z-score as the weighted sum of t-statistics across the seven cohorts. As above, weights were proportional to the square-root of the number of individuals analyzed in each cohort and selected such that the squared weights sum to 1. To test for mediation of the age-expression relationship by methylation of a particular CpG site, we used the Z-scores from Model 1 and Model 2 to perform a Sobel test [67], such that our Sobel Z-score was equal to:

$$SobelZ = Z_1 Z_2 / \sqrt{Z_1^2 + Z_2^2} \quad (1)$$

where Z_1 is the meta-analysis Z-score from the association between methylation and chronological age in Model 1, and Z_2 is the meta-analysis Z-score from the association between expression and methylation, adjusted for chronological age, in Model 2. To assess overall significance for each model (Model 1, Model 2, and the Sobel test), we used a Bonferroni-adjusted α-level of .05/135,230=3.70×10^{-7} for all CpG sites tested. To assess whether sites in each gene were significant, we assessed Bonferroni significance for each gene according to the number of CpG sites tested in that gene.

To test whether the genes were enriched for CpG sites associated with chronological age in Model 1, or CpG sites associated with expression in Model 2, we performed similar analyses on a set of 1,497 random genes. We chose these genes by first selecting the 5000 least-associated genes from the original age-expression analysis. We then used the optmatch R package [68] to select a subset of 1,497 random genes that were well-matched to the 1,497 age-associated genes in terms of gene length (bp) and the log of mean expression in WB. By doing this, we obtained a set of 1,497 random genes that were similar to the 1,497 age-associated genes in distributions of gene length, mean expression, and number of CpG sites within 250kb (Supplementary Figure 3A-D). We then performed the meta-analysis for Models 1 and 2 for all eligible CpG sites (after filtering to remove sites with probes that were non-specific or harbored genetic variants) within 250kb of these genes. We used Fisher's exact test to test whether there was an increased proportion of significant ($p<\alpha$) CpG sites in each model in the age-associated genes compared to the random genes. For our main enrichment test we set $\alpha=2.37\times10^{-7}$ as in the original analysis but to ensure robustness we re-performed the

enrichment test for a wide range of α-levels, ranging from 10^{-20} to .05, and observed that results were consistent for all α-levels considered.

To identify whether the mediating CpG sites were located in functionally relevant regions, we took two main approaches. First, we intersected the CpG positions with the hg19 CpG island annotation track from UCSC Genome Browser (http://genome.ucsc.edu), to determine whether each site was located in a CpG island, CpG shore (+/- 1.5 kb from island) or CpG shelf (+/- 1.5 kb from shore). Second, we intersected the CpG positions with ENCODE's ChromHMM annotation for lymphoblastoid cell line GM12878, which uses a hidden Markov model to assign genomic features based on the combinatorial pattern of various chromatin marks [69]. The ChromHMM annotation allowed us to identify CpGs located in promoters, enhancers, and insulators. We then used Fisher's exact test to assess whether there was significant enrichment of each feature in mediating CpG sites compared to other CpG sites within the 1,497 genes.

Query of candidate age-expression associated genes and pathways

A total of 204 candidate genes were identified from a variety of sources including Mendelian ageing disorders, longevity genetics candidates [11,12,70-72], and members of key ageing pathways, mainly FOXO/mTOR, key DNA repair genes, regulators of telomere maintenance, and mitochondrial ribosomal proteins [12,25,71,73]. Additional candidates included those from past human or multi-species expression studies [74-76], and markers of naïve or differentiated immune cells [77]. Animal model gene names were translated to human homolog names. All genes and their known human alias names were searched against the discovery and replication results. Thirty-three genes were not tested due to lack of measurement or blood expression below filtered levels. Most candidate genes (n=126) were analyzed but did not meet the strict discovery thresholds to be carried forward to the replication phase (Supplementary Table 9). Of 45 genes carried into replication, 33 convincingly replicated in WB (Supplementary Table 8).

Transcriptomic age prediction as surrogate biomarker

To investigate how accurate biological age can be predicted from gene expression levels, we performed a leave-one-out prediction analysis, i.e., re-running the meta-analysis excluding each of the validation cohorts. For all models, we used the standardized residuals of the gene expression levels, which were obtained by adjusting the gene expression levels for the technical covariates (RNA quality, batch effects) and some biological covariates (sex, fasting status, smoking status, and cell counts).

To predict age, we needed to have the estimated effect sizes of the gene expression levels on chronological age (model 1: chronological age ~ gene expression). However, effect sizes from the meta-analysis were for chronological age on gene expression levels (model 2: gene expression ~ chronological age). We used an equivalent transformation to convert the effect size in model 2 to that in model 1 by the following equation:

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

where $\hat{b}$ is the effect size of the gene expression level on chronological age (model 1), based on standardized chronological age and standardized gene expression levels, so that it needs to be interpreted in standard deviation (SD) unit for both chronological age and gene expression level; z is the z-statistic for association from the meta-analysis; and n is the sample size. We then conducted an approximate ridge regression analysis based on a random effect model, which is analogue to the best linear unbiased prediction (BLUP) approach in mixed linear model analysis, to estimate the effect sizes of all 11,908 genes jointly taking correlations between probes into account. The random effect model can be written as:

$$y = X * b_R + e \quad (3)$$

where **y** is the vector of age phenotype and **X** is the matrix of gene expression level, $\mathbf{b}_R$ is a vector of effects of gene expression on age with:

$$\mathbf{b}_R \sim N(0, 1\sigma^2_b) \quad (4)$$

and **e** is a vector of residual with:

$$\mathbf{e} \sim N(0, 1\sigma^2_e) \quad (5)$$

In a ridge regression analysis, $\mathbf{b}_R$ can be estimated as

$$\hat{b}_R = (X'X + I\lambda)^{-1} X'y \quad (6)$$

where

$$\lambda = \sigma_e^2 / \sigma_b^2 \quad (7)$$

In a single probe based meta-analysis, the analysis is equivalent to:

$$\hat{b} = D^{-1}X'y \quad (8)$$

where **b** is a vector of effect sizes estimated from the meta-analysis and **D** is the diagonal matrix of **X'X**. If the gene expression level of each probe is standardized, the i-th diagonal element of D is:

$$Dii = n \quad (9)$$

with n being the sample size. We therefore have

$$X'y = D\hat{b} \quad (10)$$

so that

$$\hat{\boldsymbol{b}}\mathbf{R} = (\mathbf{X'X} + \mathbf{1}\lambda)^{-1}\,\mathbf{D}\hat{\boldsymbol{b}} = (\mathbf{R} + \mathbf{1}\lambda/\mathbf{n})^{-1}\hat{\boldsymbol{b}} \tag{11}$$

where **R** is the correlation matrix between probes. This method largely follows the method that was proposed to estimate the joint effect sizes of SNPs using summary data from GWAS and linkage disequilibrium between SNPs from a reference sample [78]. We estimated $\mathbf{b}_R$ using $\hat{\boldsymbol{b}}$ from the meta-analysis (excluding the validation cohort) and probe correlation matrix **R** from reference samples (also independent from the validation cohort).

We calibrated the parameter λ using BSGS as the validation cohort (finding a λ value that maximized prediction accuracy in BSGS) (Supplementary Figure 10) and applied it to the prediction analysis in the other validation cohorts (Supplementary Data 5A). We call this an approximate method because the correlation matrix **R** consisted of weighted averages (weighted by sample size) from up to 6 of the discovery cohorts rather than all the samples pooled together. We applied the estimates of the individual genes from the ridge regression analysis to the left-out sample (validation sample) to predict age, and calculated the correlation coefficient of chronological age and the predicted transcriptomic age (Figure 3A-H).

Since the effect sizes of the probes were estimated from the meta-analyses excluding the validation sample, the validation set is completely independent from the discovery (training) set, so that the prediction accuracy is unbiased. We created the predictor of an individual in the validation cohort as

$$Z = \sum_i \chi_{v(i)}\,\hat{b}_{R(i)} \tag{12}$$

with $\chi_{v(i)}$ being the gene expression level of i-th probe in the validation cohort. We scaled *Z* using the mean and standard deviation (SD) of chronological age from the validation cohort:

$$SZ = \mu_{age} + (Z - \mu_z) * \frac{\sigma_{age}}{\sigma_z} \tag{13}$$

where μ_{age} and σ_{age} are the mean and SD of chronological age from the validation cohort, and μ_z and σ_z are the mean and SD of the predictor *Z*. Delta age was defined as the difference between the scaled transcriptomic predicted age (*SZ*) and chronological age for each individual.

We explored whether delta age was associated with any multi-systemic biological parameter (or biomarker) of ageing, such as sex, blood pressure, cholesterol levels, glucose levels, etc. For all biomarkers used, outliers were excluded from the analysis. Associations were tested using a linear model, including the phenotype of interest as the outcome (the dependent variable) and the delta age as the independent variable; all associations were adjusted for chronological age. To overcome the effects of obesity on cardiovascular disease and other traits, we additionally adjusted for BMI in

a second model. Additionally, we tested whether the biological parameters were directly associated with chronological age (Supplementary Table 20), so:

lm(phenotype~chronological-age) (14)

Transcriptomic age prediction for external cohorts

A general transcriptomic predictor *(Z)* was generated which can be used by external researchers for future purposes. This predictor was calculated using the prediction meta-analysis of all cohorts (except BSGS on which we calibrated the λ parameter) (Supplementary Figure 10). Cohorts that have chronological age available should scale the predictor as we did for the validation cohorts (equation 13), using the mean and SD of chronological age and the mean and SD of the predictor (*Z*).

To make our predictor useful to cohorts that do not have chronological age available, we further transformed the predictor to a scaled transcriptomic predictor (in years). This scaled predictor was calculated using the mean and SD of chronological age from all discovery cohorts in the meta-analysis (equation 13). Since the individual level age data was not available, the SD of chronological age was calculated using the pooled variance method (Supplementary Table 21).

ACKNOWLEDGMENTS

The infrastructure for the CHARGE Consortium is supported in part by the National Heart, Lung, and Blood Institute grant R01HL105756. This study was funded by the European Commission (HEALTH-F2-2008-201865, GEFOS; HEALTH-F2-2008 35627, TREAT-OA), the Netherlands Organisation for Scientific Research (NWO) Investments (nr. 175.010.2005.011, 911-03-012), the Netherlands Consortium for Healthy Aging , the Netherlands Genomics Initiative (NGI) / Netherlands Organisation for Scientific Research (NWO) project nr. 050-060-810 and Vidi grant 917103521. Additional acknowledgments to specific cohorts and their support are found in Supplementary Notes 1-2.

ADDITIONAL INFORMATION

Supplementary Information accompanies this paper at:
http://www.nature.com/nature-communications/

REFERENCES

1. Eicher JD, Landowski C, Stackhouse B, Sloan A, Chen W, et al. (2014) GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. Nucleic Acids Res.
2. Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42: D1001-1006.
3. Anselmi CV, Malovini A, Roncarati R, Novelli V, Villa F, et al. (2009) Association of the FOXO3A locus with extreme longevity in a southern Italian centenarian study. Rejuvenation Res 12: 95-104.
4. Broer L, Buchman AS, Deelen J, Evans DS, Faul JD, et al. (2014) GWAS of Longevity in CHARGE Consortium Confirms APOE and FOXO3 Candidacy. The journals of gerontology Series A, Biological sciences and medical sciences.
5. Nebel A, Kleindorp R, Caliebe A, Nothnagel M, Blanche H, 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.
6. Schachter F, Faure-Delanef L, Guenot F, Rouger H, Froguel P, et al. (1994) Genetic associations with human longevity at the APOE and ACE loci. Nat Genet 6: 29-32.
7. Soerensen M, Dato S, Christensen K, McGue M, Stevnsner T, et al. (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.
8. Walter S, Atzmon G, Demerath EW, Garcia ME, Kaplan RC, et al. (2011) A genome-wide association study of aging. Neurobiol Aging 32: 2109 e2115-2128.
9. Willcox BJ, Donlon TA, He Q, Chen R, Grove JS, et al. (2008) FOXO3A genotype is strongly associated with human longevity. Proc Natl Acad Sci U S A 105: 13987-13992.
10. Ganna A, Rivadeneira F, Hofman A, Uitterlinden AG, Magnusson PK, et al. (2013) Genetic determinants of mortality. Can findings from genome-wide association studies explain variation in human mortality? Hum Genet 132: 553-561.
11. Sebastiani P, Solovieff N, Dewan AT, Walsh KM, Puca A, et al. (2012) Genetic signatures of exceptional longevity in humans. PLoS One 7: e29848.
12. Kenyon CJ (2010) The genetics of ageing. Nature 464: 504-512.
13. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, et al. (2001) The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nature Genetics 29: 389-395.
14. Jones SJM, Riddle DL, Pouzyrev AT, Velculescu VE, Hillier L, et al. (2001) Changes in gene expression associated with developmental arrest and longevity in Caenorhabditis elegans. Genome Research 11: 1346-1352.
15. Weindruch R, Kayo T, Lee CK, Prolla TA (2001) Microarray profiling of gene expression in aging and its alteration by caloric restriction in mice. Journal of Nutrition 131: 918s-923s.
16. Ly DH, Lockhart DJ, Lerner RA, Schultz PG (2000) Mitotic misregulation and human aging. Science 287: 2486-2492.
17. van den Akker EB, Passtoors WM, Jansen R, van Zwet EW, Goeman JJ, et al. (2014) Meta-analysis on blood transcriptomic studies identifies consistently coexpressed protein-protein interaction modules as robust markers of human aging. Aging Cell 13: 216-225.
18. Glass D, Vinuela A, Davies MN, Ramasamy A, Parts L, et al. (2013) Gene expression changes with age in skin, adipose tissue, blood and brain. Genome Biol 14: R75.
19. Harries LW, Hernandez D, Henley W, Wood AR, Holly AC, et al. (2011) Human aging is characterized by focused changes in gene expression and deregulation of alternative splicing. Aging Cell 10: 868-878.
20. Kent JW, Goring HHH, Charlesworth JC, Drigalenko E, Diego VP, et al. (2012) Genotype x age interaction in human transcriptional ageing. Mechanisms of Ageing and Development 133: 581-590.
21. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, et al. (2010) Genetics and Beyond - The Transcriptome of Human Monocytes and Disease Susceptibility. PLoS One 5.
22. Tan Q, Christensen K, Christiansen L, Frederiksen H, Bathum L, et al. (2005) Genetic dissection of gene expression observed in whole blood samples of elderly Danish twins. Hum Genet 117: 267-274.
23. Horvath S (2013) DNA methylation age of human tissues and cell types. Genome Biol 14: R115.

24. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, et al. (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell 49: 359-367.

25. Houtkooper RH, Mouchiroud L, Ryu D, Moullan N, Katsyuba E, et al. (2013) Mitonuclear protein imbalance as a conserved longevity mechanism. Nature 497: 451-457.

26. McCarroll SA, Murphy CT, Zou S, Pletcher SD, Chin CS, et al. (2004) Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. Nat Genet 36: 197-204.

27. Landis G, Shen J, Tower J (2012) Gene expression changes in response to aging compared to heat stress, oxidative stress and ionizing radiation in Drosophila melanogaster. Aging (Albany NY) 4: 768-789.

28. Landis GN, Abdueva D, Skvortsov D, Yang J, Rabin BE, et al. (2004) Similar gene expression patterns characterize aging and oxidative stress in Drosophila melanogaster. Proc Natl Acad Sci U S A 101: 7663-7668.

29. Lauring B, Wang S, Sakai H, Davis TA, Wiedmann B, et al. (1995) Nascent-polypeptide-associated complex: a bridge between ribosome and cytosol. Cold Spring Harb Symp Quant Biol 60: 47-56.

30. Johnson SC, Yanos ME, Kayser EB, Quintana A, Sangesland M, et al. (2013) mTOR inhibition alleviates mitochondrial disease in a mouse model of Leigh syndrome. Science 342: 1524-1528.

31. Park J, Jo YH, Cho CH, Choe W, Kang I, et al. (2013) ATM-deficient human fibroblast cells are resistant to low levels of DNA double-strand break induced apoptosis and subsequently undergo drug-induced premature senescence. Biochem Biophys Res Commun 430: 429-435.

32. Luo YB, Mitrpant C, Johnsen RD, Fabian VA, Fletcher S, et al. (2013) Investigation of age-related changes in LMNA splicing and expression of progerin in human skeletal muscles. Int J Clin Exp Pathol 6: 2778-2786.

33. Bonder MJ, Kasela S, Kals M, Tamm R, Lokk K, et al. (2014) Genetic and epigenetic regulation of gene expression in fetal and adult human livers. BMC Genomics 15: 860.

34. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. (2015) Integrative analysis of 111 reference human epigenomes. Nature 518: 317-330.

35. Gheorghe M, Snoeck M, Emmerich M, Back T, Goeman JJ, et al. (2014) Major aging-associated RNA expressions change at two distinct age-positions. BMC Genomics 15: 132.

36. Shigenaga MK, Hagen TM, Ames BN (1994) Oxidative damage and mitochondrial decay in aging. Proc Natl Acad Sci U S A 91: 10771-10778.

37. Ojaimi J, Masters CL, Opeskin K, McKelvie P, Byrne E (1999) Mitochondrial respiratory chain activity in the human brain as a function of age. Mechanisms of Ageing and Development 111: 39-47.

38. Short KR, Bigelow ML, Kahl J, Singh R, Coenen-Schimke J, et al. (2005) Decline in skeletal muscle mitochondrial function with aging in humans. Proc Natl Acad Sci U S A 102: 5618-5623.

39. Yen TC, Chen YS, King KL, Yeh SH, Wei YH (1989) Liver mitochondrial respiratory functions decline with age. Biochem Biophys Res Commun 165: 944-1003.

40. Sallusto F, Lenig D, Forster R, Lipp M, Lanzavecchia A (1999) Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. Nature 401: 708-712.

41. Lee WW, Yang ZZ, Li G, Weyand CM, Goronzy JJ (2007) Unchecked CD70 expression on T cells lowers threshold for T cell activation in rheumatoid arthritis. J Immunol 179: 2609-2615.

42. Moro-Garcia MA, Alonso-Arias R, Lopez-Larrea C (2012) Molecular mechanisms involved in the aging of the T-cell immune response. Curr Genomics 13: 589-602.

43. Pletcher SD, Macdonald SJ, Marguerie R, Certa U, Stearns SC, et al. (2002) Genome-wide transcript profiles in aging and calorically restricted Drosophila melanogaster. Curr Biol 12: 712-723.

44. Rera M, Clark RI, Walker DW (2012) Intestinal barrier dysfunction links metabolic and inflammatory markers of aging to death in Drosophila. Proc Natl Acad Sci U S A 109: 21528-21533.

45. Landis GN, Bhole D, Tower J (2003) A search for doxycycline-dependent mutations that increase Drosophila melanogaster life span identifies the VhaSFD, Sugar baby, filamin, fwd and Cctl genes. Genome Biol 4: R8.

46. Liu YL, Lu WC, Brummel TJ, Yuh CH, Lin PT, et al. (2009) Reduced expression of alpha-1,2-mannosidase I extends lifespan in Drosophila melanogaster and Caenorhabditis elegans. Aging Cell 8: 370-379.

47. Landis G, Bhole D, Lu L, Tower J (2001) High-frequency generation of conditional mutations affecting Drosophila melanogaster development and life span. Genetics 158: 1167-1176.

48. Taylor KR, Gallo RL (2006) Glycosaminoglycans and their proteoglycans: host-associated molecular patterns for initiation and modulation of inflammation. FASEB J 20: 9-22.

49. Pittman J (2007) Effect of aging on wound healing: current concepts. J Wound Ostomy Continence Nurs 34: 412-415; quiz 416-417.

50. Loegel TN, Trombley JD, Taylor RT, Danielson ND (2012) Capillary electrophoresis of heparin and other glycosaminoglycans using a polyamine running electrolyte. Anal Chim Acta 753: 90-96.

51. Didsbury A, Wang C, Verdon D, Sewell MA, McIntosh JD, et al. (2011) Rotavirus NSP4 is secreted from infected cells as an oligomeric lipoprotein and binds to glycosaminoglycans on the surface of non-infected cells. Virol J 8: 551.

52. Gourlay CW, Ayscough KR (2005) A role for actin in aging and apoptosis. Biochem Soc Trans 33: 1260-1264.

53. Higuchi R, Vevea JD, Swayne TC, Chojnowski R, Hill V, et al. (2013) Actin dynamics affect mitochondrial quality control and aging in budding yeast. Curr Biol 23: 2417-2422.

54. Bratic A, Larsson NG (2013) The role of mitochondria in aging. J Clin Invest 123: 951-957.

55. Ebersberger I, Simm S, Leisegang MS, Schmitzberger P, Mirus O, et al. (2014) The evolution of the ribosome biogenesis pathway from a yeast perspective. Nucleic Acids Res 42: 1509-1523.

56. Kenyon J, Gerson SL (2007) The role of DNA damage repair in aging of adult stem cells. Nucleic Acids Res 35: 7557-7565.

57. Petes TD, Farber RA, Tarrant GM, Holliday R (1974) Altered rate of DNA replication in ageing human fibroblast cultures. Nature 251: 434-436.

58. Wolfson M, Budovsky A, Tacutu R, Fraifeld V (2009) The signaling hubs at the crossroad of longevity and age-related disease networks. Int J Biochem Cell Biol 41: 516-520.

59. Boya P (2012) Lysosomal function and dysfunction: mechanism and disease. Antioxid Redox Signal 17: 766-774.

60. Seoh ML, Ng CH, Yong J, Lim L, Leung T (2003) ArhGAP15, a novel human RacGAP protein with GTPase binding property. FEBS Lett 539: 131-137.

61. Patil V, Ward RL, Hesson LB (2014) The evidence for functional non-CpG methylation in mammalian cells. Epigenetics 9: 823-828.

62. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462: 315-322.

63. Allum F, Shao X, Guenard F, Simon MM, Busche S, et al. (2015) Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. Nat Commun 6: 7211.

64. Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics 13: 484-492.

65. Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26: 2190-2191.

66. Price ME, Cotton AM, Lam LL, Farre P, Emberly E, et al. (2013) Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics Chromatin 6: 4.

67. Sobel ME (1982) Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. Sociological Methodology Vol. 13: pp. 290-312.

68. Hansen BB, Klopfer SO (2006) Optimal full matching and related designs via network flows. Journal of Computational and Graphical Statistics 15: 609-627.

69. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473: 43-49.

70. Barzilai N, Gabriely I, Atzmon G, Suh Y, Rothenberg D, et al. (2010) Genetic studies reveal the role of the endocrine and metabolic systems in aging. J Clin Endocrinol Metab 95: 4493-4500.

71. Kenyon C (2011) The first long-lived mutants: discovery of the insulin/IGF-1 pathway for ageing. Philosophical transactions of the Royal Society of London Series B, Biological sciences 366: 9-16.

72. Newman AB, Murabito JM (2013) The Epidemiology of Longevity and Exceptional Survival. Epidemiologic reviews.

73. Harries LW, Fellows AD, Pilling LC, Hernandez D, Singleton A, et al. (2012) Advancing age is associated with gene expression changes resembling mTOR inhibition: evidence from two human populations. Mechanisms of ageing and development 133: 556-562.

74. de Magalhaes JP, Curado J, Church GM (2009) Meta-analysis of age-related gene expression profiles identifies common signatures of aging. Bioinformatics 25: 875-881.

75. Passtoors WM, Boer JM, Goeman JJ, Akker EB, Deelen J, et al. (2012) Transcriptional profiling of human familial longevity indicates a role for ASF1A and IL7R. PLoS One 7: e27759.

76. Zahn JM, Poosala S, Owen AB, Ingram DK, Lustig A, et al. (2007) AGEMAP: a gene expression database for aging in mice. PLoS genetics 3: e201.

77. Chou JP, Ramirez CM, Wu JE, Effros RB (2013) Accelerated aging in HIV/AIDS: novel biomarkers of senescent human CD8+ T cells. PLoS One 8: e64702.

78. Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, et al. (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 44: 369-375, S361-363.

CHAPTER 2.2

A meta-analysis of gene expression signatures of blood pressure and hypertension

Tianxiao Huan*, Tõnu Esko*, Marjolein J. Peters*, Luke C. Pilling*, Katharina Schramm*, Claudia Schurmann*, Brian H. Chen, Chunyu Liu, Roby Joehanes, Andrew D. Johnson, Chen Yao, Sai-xia Ying, Paul Courchesne, Lili Milani, Nalini Raghavachari, Richard Wang, Poching Liu, Eva Reinmaa, Abbas Dehghan, Albert Hofman, André G. Uitterlinden, Dena G. Hernandez, Stefania Bandinelli, Andrew Singleton, David Melzer, Andres Metspalu, Maren Carstensen, Harald Grallert, Christian Herder, Thomas Meitinger, Annette Peters, Michael Roden, Melanie Waldenberger, Marcus Dörr, Stephan B. Felix, Tanja Zeller, International Consortium for Blood Pressure GWAS (ICBP), Ramachandran Vasan, Christopher J. O'Donnell, Peter J. Munson, Xia Yang*, Holger Prokisch*, Uwe Völker*, Joyce B.J. van Meurs*, Luigi Ferrucci*, Daniel Levy*

** These authors contributed equally to this work*

Published: *PLoS Genet 2015;* ***11:*** *doi:10.1371/journal.pgen.1005035*

ABSTRACT

Genome-wide association studies (GWAS) have uncovered numerous genetic variants (SNPs) that are associated with blood pressure (BP). Genetic variants may lead to BP changes by acting on intermediate molecular phenotypes such as coded protein sequence or gene expression, which in turn affect BP variability. Therefore, characterizing genes whose expression is associated with BP may reveal cellular processes involved in BP regulation and uncover how transcripts mediate genetic and environmental effects on BP variability. A meta-analysis of results from six studies of global gene expression profiles of BP and hypertension in whole blood was performed in 7,017 individuals who were not receiving antihypertensive drug treatment. We identified 34 genes that were differentially expressed in relation to BP (Bonferroni-corrected $p<0.05$). Among these genes, *FOS* and *PTGS2* have been previously reported to be involved in BP-related processes; the others are novel. The top BP signature genes in aggregate explain 5%-9% of inter-individual variance in BP. Of note, rs3184504 in *SH2B3*, which was also reported in GWAS to be associated with BP, was found to be a *trans* regulator of the expression of 6 of the transcripts we found to be associated with BP (*FOS*, *MYADM*, *PP1R15A*, *TAGAP*, *S100A10*, and *FGBP2*). Gene set enrichment analysis suggested that the BP-related global gene expression changes include genes involved in inflammatory response and apoptosis pathways. Our study provides new insights into molecular mechanisms underlying BP regulation, and suggests novel transcriptomic markers for the treatment and prevention of hypertension.

AUTHOR SUMMARY

The focus of blood pressure (BP) GWAS has been the identification of common DNA sequence variants associated with the phenotype; this approach provides only one dimension of molecular information about BP. While it is a critical dimension, analyzing DNA variation alone is not sufficient for achieving an understanding of the multidimensional complexity of BP physiology. The top loci identified by GWAS explain only about 1 percent of inter-individual BP variability. In this study, we performed a meta-analysis of gene expression profiles in relation to BP and hypertension in 7,017 individuals from six studies. We identified 34 differentially expressed genes for BP, and discovered that the top BP signature genes explain 5%-9% of BP variability. We further linked BP gene expression signature genes with BP GWAS results by integrating expression associated SNPs (eSNPs) and discovered that one of the top BP loci from GWAS, rs3184504 in *SH2B3*, is a *trans* regulator of expression of 6 of the top 34 BP signature genes. Our study, in conjunction with prior GWAS, provides a deeper understanding of the molecular and genetic basis of BP regulation, and identifies several potential targets and pathways for the treatment and prevention of hypertension and its sequelae.

INTRODUCTION

Systolic and diastolic blood pressure (SBP and DBP) are complex physiological traits that are affected by the interplay of multiple genetic and environmental factors. Hypertension (HTN) is a critical risk factor for stroke, renal failure, heart failure, and coronary heart disease [1]. Genome- wide association studies (GWAS) have identified numerous loci associated with BP traits [2,3]. These loci, however, only explain a small proportion of inter-individual BP variability. In aggregate the 29 loci reported by the International Consortium of Blood Pressure (ICBP) consortium GWAS account for about one percent of BP variation in the general population [3]. Most genes near BP GWAS loci are not known to be mechanistically associated with BP regulation [3]. Therefore, further studies are needed to determine whether the genes implicated in GWAS demonstrate functional relations to BP physiology and to uncover the molecular actions and interactions of genetic and environmental factors involved in BP regulation.

Alterations in gene expression may mediate the effects of genetic variants on phenotype variability. We hypothesized that characterizing gene expression signatures of BP would reveal cellular processes involved in BP regulation and uncover how transcripts mediate genetic and environmental effects on BP variability. We additionally hypothesized that by integrating gene expression profiling with genetic variants associated with altered gene expression (eSNPs or eQTLs) and with BP GWAS results, we would be able to characterize the genetic architecture of gene expression effects on BP regulation.

Several previous studies have examined the association of global gene expression with BP [4,5] or HTN [6,7]. Most of these studies, however, were based on small sample sizes and lacked replication [4,5,6,7]. To address this challenge, we conducted an association study of global gene expression levels in whole blood with BP traits (SBP, DBP, and HTN) in six independent studies. In order to avoid the possibility that the differentially expressed genes we identified reflect drug treatment effects, we excluded individuals receiving anti-hypertensive treatment. The eligible study sample included 7,017 individuals: 3,679 from the Framingham Heart Study (FHS), 972 from the Estonian Biobank (EGCUT), 604 from the Rotterdam Study (RS) [8], 597 from the InCHIANTI Study, 565 from the Cooperative Health Research in the Region of Augsburg [KORA F4] Study [9], and 600 from the Study of Health in Pomerania [SHIP-TREND] [10]. We first identified differentially expressed BP genes in the FHS (n=3,679) followed by external replication in the other five studies (n=3,338). Subsequently, we performed a meta-analysis of all 7,017 individuals from the six studies, and identified 34 differentially expressed genes associated with BP traits using a stringent statistical threshold based on Bonferroni correction for multiple testing of 7,717 unique genes. The differentially expressed genes for BP (BP signature genes) were further integrated with eQTLs and with BP GWAS results in an effort to differentiate downstream transcriptomic changes due to BP from putatively causal pathways involved in BP regulation.

RESULTS

Clinical characteristics

After excluding individuals receiving anti-hypertensive treatment, the eligible sample size was 7,017 (FHS, n=3,679; EGCUT, n=972; RS, n=604; InCHIANTI, n=597; KORA F4, n=565 and SHIP-TREND, n=600). Clinical characteristics of participants from the four studies are presented in Table 1. The mean age varied across the cohorts (FHS=51, EGCUT=36, RS=58, InCHIANTI=71, KORA F4=72 and SHIP-TREND=46 years) as did the proportion of individuals with hypertension (11% in FHS, 19% in EGCUT, 35% in RS, 45% in InCHIANTI, 26% in KORA, and 12% in SHIP).

Table 1. Clinical characteristics of the study cohorts

	FHS N=3,679	EGCUT N=972	RS N=604	InCHIANTI N=597	KORA F4 N=565	SHIP-TREND N=600
Age (years)	51 ± 12	36 ± 14	58 ± 8	71 ± 16	72 ± 5	46 ± 13
Sex, male (%)	42	49	46	46	51	43
Hypertension (%)	11	19	35	45	26	12
BMI (kg/m^2)	27.2 ± 5.3	24.8 ± 4.4	26.8 ± 4.1	27.0 ± 4.2	29.8 ± 4.6	26 ± 4.2
Systolic BP (mm Hg)	118 ± 15	122 ± 16	132 ± 20	132 ± 20	129 ± 21	120 ± 15
Diastolic BP (mm Hg)	74 ± 9	76 ± 10	82 ± 11	78 ± 10	73 ± 11	75 ± 9

Identification and replication of differentially expressed BP signature genes

At a Bonferroni corrected p<0.05, we identified 73, 31, and 8 genes that were differentially expressed in relation to SBP, DBP, and HTN, respectively in the FHS, which used an Affymetrix array for expression profiling, and 6, 1, and 1 genes in the meta-analysis of the 5 cohorts that used an Illumina array (Illumina cohorts): EGCUT, RS, InCHIANTI, KORA F4 and SHIP-TREND (Supplementary Table 1). For each differentially expressed BP gene in the FHS or in the Illumina cohorts, we attempted replication in the other group. At a replication p<0.05 (Bonferroni corrected), 13 unique genes that were identified in the FHS were replicated in the Illumina cohorts, including 10 for SBP (*CD97*, *TAGAP*, *DUSP1*, *FOS*, *MCL1*, *MYADM*, *PPP1R15A*, *SLC31A2*, *TAGLN2*, and *TIPARP*), 5 for DBP (*CD97*, *BHLHE40*, *PRF1*, *CLC*, and *MYADM*), and 2 for HTN (*GZMB* and *MYADM*) (Table 2). Each of the unique BP signature genes in the Illumina cohorts, 6 for SBP (*TAGLN2*, *BHLHE40*, *MYADM*, *SLC31A2*, *DUSP1*, and *MCL1*), 1 for DBP (*BHLHE40*) and 1 for HTN (*SLC31A2*), replicated in the FHS. All 6 Illumina cohorts BP signature genes that replicated in the FHS were among the 13 FHS BP signature genes that replicated in the Illumina cohorts. The BP signature genes identified in the FHS showed enrichment in the Illumina cohorts at pi1=0.88, 0.75, and 0.99 for SBP, DBP, and HTN respectively (pi1 value indicates the proportion of significant signals among the tested associations [11]; see details in the Methods section). Figure 1 shows that the mean gene expression levels of the top BP signature genes were consistent with the BP phenotypic changes observed in the FHS and the Illumina cohorts.

The 73 SBP signature genes in the FHS (55 of these 73 genes were measured in the Illumina cohorts) at a Bonferroni corrected p<0.05 in aggregate explained 9.4% of SBP phenotypic variance in the Illumina cohorts, and the 31 DBP signature genes from the FHS (22 of these 31 genes were measured in the Illumina cohorts) in aggregate explained 5.3% of DBP phenotypic variance in the Illumina cohorts. These results suggest that in contrast to common genetic variants identified by BP GWAS, which explain in aggregate only about 1% of inter-individual BP variation [3], changes in gene expression levels explains a considerably larger proportion of phenotypic variance in BP.

Meta-analysis of the six cohorts identifies differentially expressed BP signature genes

A meta-analysis of differential expression across all six cohorts revealed 34 differentially expressed BP genes at p<0.05 (Bonferroni corrected for 7,717 genes that were measured and passed quality control in the FHS and Illumina cohorts), including 21 for SBP, 20 for DBP, and 5 for HTN (Table 2 and Supplementary Figure 2). All of the 34 differentially expressed BP signature genes showed directional consistency in the FHS and the Illumina cohorts (Table 2). The 34 BP signature genes included all 13 genes that were cross-validated between the FHS and the Illumina cohorts. Of the 34 BP signature genes, 27 were positively correlated with BP and only 7 genes were negatively correlated. MYADM and SLC31A2 were top signature genes for SBP, DBP, and HTN. At FDR<0.2, 224 unique genes were differentially expressed in relation with BP phenotypes, including 142 genes for SBP, 137 for DBP, and 45 for HTN (details are reported in the Supplementary Text 1 and 2, and Supplementary Tables 3-5).

Functional analysis of differentially expressed BP signature genes

We used gene set enrichment analysis (GSEA) to identify the biological process and pathways associated with gene expression changes in relation to SBP, DBP, and HTN in order to better understand the biological themes within the data. As shown in Table 3, the GSEA of genes whose expression was positively associated with BP showed enrichment for antigen processing and presentation (p<0.0001), apoptotic program (p<0.0001), inflammatory response (p<0.0001), and oxidative phosphorylation (p=0.0018). The negatively associated genes showed enrichment for nucleotide metabolic process (p<0.0001), positive regulation of cellular metabolic process (p<0.0001), and positive regulation of DNA dependent transcription (p=0.0021).

Genetic effects on expression of BP signature genes

Among the 34 BP signatures genes from the meta-analysis of all 6 studies, 33 were found to have *cis*-eQTLs and 26 had *trans*-eQTLs (Figure 2A and Supplementary Table 2) based on whole blood profiling [12,13]. Of these, six master *trans*-eQTLs mapped to either five or six BP signature genes (no master *cis*-eQTL was identified). Five master *trans*-eQTLs (rs653178, rs3184504, rs10774625, rs11065987, and rs17696736) were located on chromosome 12q24 within the same linkage disequilibrium (LD) block (r2 >0.8, Figure 2B). We retrieved a peak *cis*- and *trans*-eQTL for each BP signature gene. The peak *cis*-eQTL explained 0.2-20% of the variance in the corresponding transcript levels, in contrast, the peak *trans*-eQTL accounted for very little (0.02-2%) of the corresponding transcript variance. Westra et al. also reported a similar small proportion of variance in transcript levels explained by *trans*-eQTLs [12].

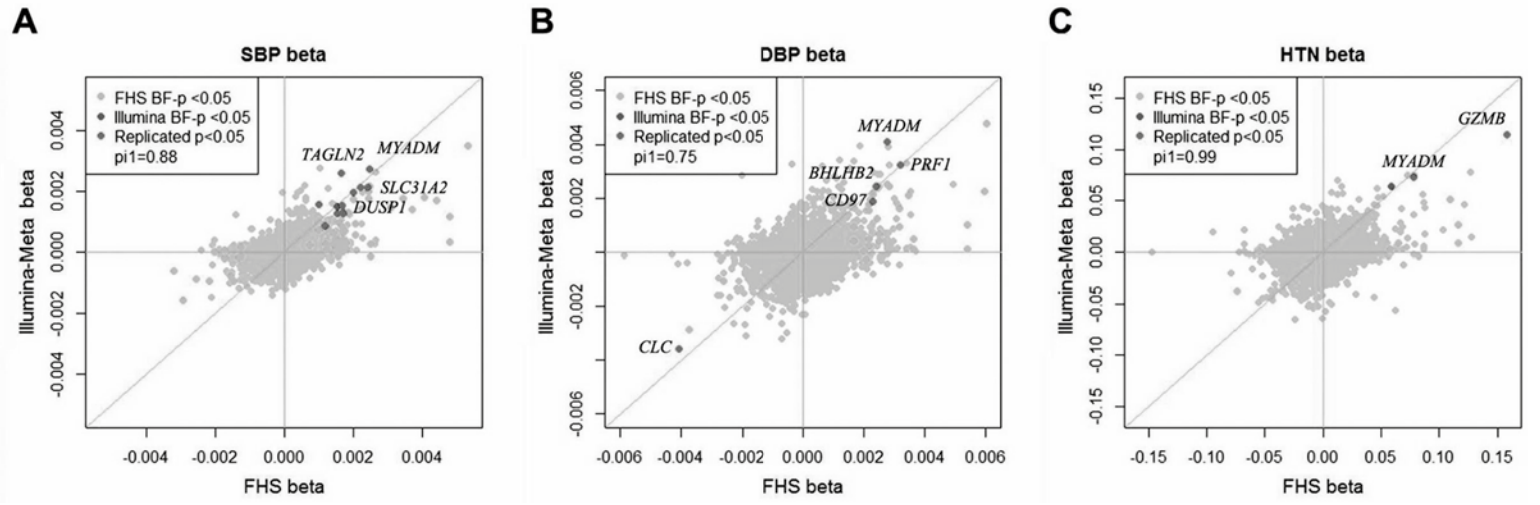


Figure 1. Effect size of differentially expressed BP genes in the FraminghamHeart Study and the Illumina cohorts. (**A**) SBP; (**B**) DBP; (**C**) HTN. The x-axis is the effect size of the differentially expressed genes in the FHS cohort and the y-axis is the effect size in the Illumina cohorts. The BP signature genes identified both in the FHS and the Illumina cohorts at p<0.05 (Bonferroni corrected) are highlighted. pi1 values indicate the proportion of significant signals among the tested associations [11] (See details in the Methods section).

Table 2. Differentially expressed genes associated with BP and hypertension at Bonferroni correction p<0.05 in meta-analysis of the six cohorts.

Gene	Chr	Gene Description	FHS Beta	FHS SE	FHS P-value	Illumina Beta	Illumina SE	Illumina P-value	Meta* Beta	Meta* SE	Meta* P-value
SBP Signature genes											
SLC31A2	9	solute carrier family 31 (coppertransporters), member2	2.4E-03	3.3E-04	1.2E-13	2.1E-03	3.3E-04	9.9E-11	2.3E-03	2.3E-04	<1E-16
MYADM	19	myeloid-associated differentiation marker	2.5E-03	3.2E-04	2.2E-14	2.7E-03	3.9E-04	2.2E-12	2.6E-03	2.5E-04	<1E-16
DUSP1	5	dual specificity phosphatase1	2.2E-03	3.9E-04	1.1E-08	2.1E-03	4.2E-04	3.7E-07	2.2E-03	2.9E-04	2.0E-14
TAGLN2	1	transgelin 2	2.0E-03	4.1E-04	1.0E-06	2.0E-03	4.0E-04	1.3E-06	2.0E-03	2.9E-04	5.8E-12
CD97	19	CD97 molecule	1.7E-03	3.2E-04	1.4E-07	1.5E-03	3.5E-04	1.6E-05	1.6E-03	2.4E-04	1.0E-11
BHLHE40	3	basic helix-loop-helix family, member e40	1.5E-03	3.4E-04	4.3E-06	1.5E-03	3.0E-04	6.4E-07	1.5E-03	2.2E-04	1.2E-11
MCL1	1	myeloid cell leukemia sequence 1 (BCL2-related)	1.0E-03	2.0E-04	7.5E-07	1.6E-03	3.2E-04	1.5E-06	1.2E-03	1.7E-04	1.4E-11
PRF1	10	perforin 1 (pore forming protein)	2.5E-03	4.1E-04	2.5E-09	1.8E-03	5.3E-04	1.0E-03	2.2E-03	3.3E-04	1.6E-11
GPR56	16	G protein-coupled receptor 56	2.0E-03	3.4E-04	3.5E-09	1.7E-03	5.8E-04	3.0E-03	1.9E-03	2.9E-04	3.9E-11
PPP1R15A	19	protein phosphatase 1, regulatory (inhibitor) subunit 15A	1.5E-03	2.6E-04	1.7E-09	1.3E-03	3.0E-04	2.8E-05	1.4E-03	2.4E-04	1.5E-08
FGFBP2	4	fibroblast growth factor binding protein 2	2.3E-03	5.0E-04	5.8E-06	2.0E-03	6.2E-04	1.5E-03	2.2E-03	3.9E-04	3.3E-08
GNLY	2	granulysin	2.6E-03	6.4E-04	3.6E-05	2.6E-03	7.2E-04	3.0E-04	2.6E-03	4.8E-04	4.0E-08
FOS	14	FBJ murine osteosarcoma viral oncogene homolog	1.7E-03	2.5E-04	1.6E-11	2.6E-03	6.3E-04	3.6E-05	2.3E-03	4.1E-04	4.8E-08
NKG7	19	natural killer cell group 7 sequence	2.3E-03	5.3E-04	1.9E-05	1.4E-03	5.5E-04	8.8E-03	1.9E-03	3.8E-04	9.4E-07
GRAMD1A	19	GRAM domain containing 1A	-6.0E-04	1.4E-04	2.1E-05	-6.7E-04	2.8E-04	1.8E-02	-6.2E-04	1.3E-04	1.1E-06
GLRX5	14	glutaredoxin 5	1.7E-03	3.9E-04	1.3E-05	1.3E-03	6.1E-04	3.5E-02	1.6E-03	3.3E-04	1.5E-06
TMEM43	3	transmembrane protein 43	7.5E-04	2.1E-04	3.0E-04	7.7E-04	2.5E-04	2.4E-03	7.6E-04	1.6E-04	2.3E-06
TIPARP	3	TCDD-inducible poly(ADP-ribose) polymerase	1.2E-03	2.3E-04	1.3E-07	8.6E-04	2.4E-04	3.3E-04	9.5E-04	2.0E-04	2.6E-06
AHNAK	11	AHNAK Nucleoprotein	9.1E-04	2.6E-04	4.1E-04	9.7E-04	3.4E-04	4.0E-03	9.3E-04	2.0E-04	5.2E-06
PIGB	15	phosphatidylinositol glycan anchor biosynthesis, class B	1.1E-03	3.1E-04	5.3E-04	6.7E-04	2.1E-04	1.9E-03	8.0E-04	1.8E-04	6.1E-06
TAGAP	6	T-cell activation RhoGTPase activating protein	1.7E-03	2.5E-04	5.7E-12	1.3E-03	3.7E-04	7.1E-04	1.4E-03	3.1E-04	6.4E-06

Table 2. *(Continued)*

Gene	Chr	Gene Description	FHS Beta	FHS SE	FHS P-value	Illumina Beta	Illumina SE	Illumina P-value	Meta* Beta	Meta* SE	Meta* P-value
DBP Signature genes											
BHLHE40	3	basic helix-loop-helix family, member e40	2.4E-03	5.1E-04	2.3E-06	2.5E-03	5.2E-04	2.8E-06	2.4E-03	3.6E-04	2.7E-11
ANXA1	9	annexin A1	3.5E-03	5.7E-04	1.2E-09	2.1E-03	7.8E-04	6.3E-03	3.0E-03	4.6E-04	6.5E-11
PRF1	10	perforin 1 (pore forming protein)	3.2E-03	6.2E-04	3.2E-07	3.2E-03	9.4E-04	5.7E-04	3.2E-03	5.2E-04	6.7E-10
KCNJ2	17	potassium inwardly-rectifying channel, subfamily J, member 2	-2.6E-03	5.6E-04	3.9E-06	-2.0E-03	5.5E-04	2.6E-04	-2.3E-03	3.9E-04	4.9E-09
CLC	19	Charcot-Leyden crystal protein	-4.1E-03	8.6E-04	2.6E-06	-3.6E-03	1.0E-03	5.7E-04	-3.9E-03	6.7E-04	5.8E-09
CD97	19	CD97 molecule	2.3E-03	4.8E-04	1.6E-06	1.9E-03	5.8E-04	1.1E-03	2.1E-03	3.7E-04	7.4E-09
IL2RB	22	interleukin2 receptor, beta	2.3E-03	4.9E-04	3.0E-06	2.2E-03	7.3E-04	2.4E-03	2.3E-03	4.1E-04	2.5E-08
S100A10	1	S100 calcium binding protein A10	3.2E-03	6.1E-04	2.4E-07	1.6E-03	6.2E-04	9.9E-03	2.4E-03	4.4E-04	4.0E-08
GPR56	16	G protein-coupled receptor 56	2.5E-03	5.2E-04	1.1E-06	2.4E-03	1.0E-03	1.7E-02	2.5E-03	4.6E-04	5.5E-08
TIPARP	3	TCDD-inducible poly(ADP-ribose) polymerase	1.3E-03	3.4E-04	1.3E-04	1.1E-03	3.1E-04	2.8E-04	1.2E-03	2.3E-04	1.4E-07
HAVCR2	5	Hepatitis A Virus Cellular Receptor 2	1.7E-03	4.6E-04	3.8E-04	1.8E-03	4.8E-04	1.8E-04	1.7E-03	3.3E-04	2.4E-07
PTGS2	1	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	-2.1E-03	4.9E-04	2.2E-05	-1.3E-03	5.1E-04	9.0E-03	-1.7E-03	3.5E-04	1.0E-06
MYADM	19	myeloid-associated differentiation marker	2.8E-03	4.9E-04	1.7E-08	4.1E-03	1.0E-03	8.6E-05	3.6E-03	7.4E-04	1.1E-06
ANTXR2	4	anthrax toxin receptor2	1.5E-03	3.3E-04	5.2E-06	8.3E-04	4.3E-04	5.5E-02	1.3E-03	2.6E-04	1.7E-06
OBFC2A	2	nucleic acid binding protein 1	-1.7E-03	3.9E-04	7.2E-06	-9.6E-04	4.6E-04	3.8E-02	-1.4E-03	3.0E-04	1.8E-06
GRAMD1A	19	GRAM domain containing 1A	-9.3E-04	2.1E-04	1.4E-05	-8.7E-04	5.0E-04	7.8E-02	-9.2E-04	2.0E-04	2.8E-06
ARHGAP15	2	Rho GTPase activating protein 15	-1.3E-03	4.1E-04	1.1E-03	-1.4E-03	4.4E-04	1.5E-03	-1.4E-03	3.0E-04	5.2E-06
FBXL5	4	F-box and leucine-rich repeat protein 5	-1.6E-03	3.7E-04	2.1E-05	-9.4E-04	4.9E-04	5.5E-02	-1.3E-03	2.9E-04	5.3E-06
SLC31A2	9	solute carrier family 31 (copper transporters), member 2	2.8E-03	4.9E-04	1.0E-08	2.4E-03	8.1E-04	2.6E-03	2.6E-03	5.6E-04	5.4E-06
VIM	10	vimentin	1.7E-03	3.8E-04	5.5E-06	7.6E-04	5.9E-04	2.0E-01	1.4E-03	3.2E-04	6.2E-06

Table 2. *(Continued)*

Gene	Chr	Gene Description	FHS Beta	FHS SE	FHS P-value	Illumina Beta	Illumina SE	Illumina P-value	Meta* Beta	Meta* SE	Meta* P-value
HTN Signature genes											
SLC31A2	9	solute carrier family 31 (copper transporters), member 2	5.9E-02	1.4E-02	1.9E-05	6.4E-02	1.4E-02	2.1E-06	6.1E-02	9.6E-03	1.8E-10
MYADM	19	myeloid-associated differentiation marker	7.8E-02	1.4E-02	1.2E-08	7.3E-02	2.1E-02	6.2E-04	7.4E-02	1.4E-02	3.0E-07
TAGAP	6	T-cell activation RhoGTPase activating protein	4.4E-02	1.1E-02	3.2E-05	3.2E-02	1.2E-02	5.3E-03	3.9E-02	7.8E-03	7.3E-07
GZMB	14	granzyme B (granzyme 2, cytotoxic T-lymphocyte-associated serine esterase 1)	1.6E-01	2.3E-02	1.1E-11	1.1E-01	3.5E-02	9.6E-04	1.3E-01	2.6E-02	1.4E-06
KCNJ2	17	potassium inwardly-rectifying channel, subfamily J, member 2	-5.2E-02	1.6E-02	8.4E-04	-4.4E-02	1.3E-02	5.5E-04	-4.7E-02	9.9E-03	1.7E-06

**Meta = meta-analysis of all six cohorts.*

Table 3. Gene set enrichment analysis for BP associated gene expression changes.

Name	Pos/Neg associated gene expression changes	Database	# of genes in pathway	NES*	P-value	FDR
DBP Signature						
Antigen processing and presentation	Positive	KEGG	37	2	<1E-04	0.01
Nature killer cell mediated cytotoxicity	Positive	KEGG	71	1.8	<1E-04	0.07
Porphyrin and chlorophyll metabolism	Positive	KEGG	15	1.7	1.0E-02	0.13
Rho protein signaling transduction	Negative	GO-BP	18	-1.8	3.9E-03	0.10
Receptor mediated endocytosis	Negative	GO-BP	16	-1.8	3.9E-03	0.17
Detection of stimulus	Negative	GO-BP	18	-1.9	9.8E-03	0.20
SBP Signature						
Nature killer cell mediated cytotoxicity	Positive	KEGG	71	1.9	1.7E-03	0.05
Apoptotic program	Positive	GO-BP	37	1.9	<1E-04	0.03
Inflammatory response	Positive	GO-BP	72	2	<1E-04	0.05
Nucleotide metabolic process	Negative	GO-BP	32	-1.9	<1E-04	0.04
Translation	Negative	GO-BP	79	-1.8	<1E-04	0.05
HTN Signature						
Antigen processing and presentation	Positive	KEGG	37	1.8	<1E-04	0.04
Oxidative phosphorylation	Positive	KEGG	52	1.8	1.8E-03	0.05
Apoptotic program	Positive	GO-BP	37	1.9	1.8E-03	0.14
Positive regulation of nucleic acid metabolic process	Negative	GO-BP	71	-1.9	<1E-04	0.08
Positive regulation of cellular metabolic process	Negative	GO-BP	105	-1.8	<1E-04	0.08
Positive regulation of transcription - DNA dependent	Negative	GO-BP	56	-1.8	2.1E-03	0.09

**NES = normalized enrichment score; GO-BP = Gene ontology- biological process; KEGG = Kyoto encyclopedia of genes and genomes.*

We then linked the *cis*- and *trans*-eQTLs of the 34 BP signature genes with BP GWAS results from the ICBP Consortium [3] and the NHGRI GWAS Catalog [14] (Figure 2 and Supplementary Table 2). We did not find any *cis*-eQTLs for the top BP signature genes that also were associated with BP in the ICBP GWAS [3]. However, the 6 master *trans*-eQTLs were all associated with BP at p<5E-8 in the ICBP GWAS [3] and were associated with multiple complex diseases or traits (Table 4). For example, rs3184504, a nonsynonymous SNP in *SH2B3* that was associated in GWAS with BP, coronary heart disease, hypothyroidism, rheumatoid arthritis, and type 1 diabetes [12], is a *trans*-eQTL for 6 of our 34 BP signature genes from the meta-analysis (*FOS*, *MYADM*, *PP1R15A*, *TAGAP*, *S100A10*, and *FGBP2*; Figure 2A-B and Table 4). These 6 genes are all highly expressed in neutrophils, and their expression levels are correlated significantly (average r2=0.04, p<1E-16). rs653178, intronic to *ATXN2* and in perfect LD with rs3184504 (r2=1), also is associated with BP and multiple other diseases in the NHGRI GWAS Catalog [14]. It also is a *trans*-eQTL for the same 6 BP signature genes (Table 4). These two SNPs are *cis*-eQTLs for expression *SH2B3* in whole blood (FDR<0.05), but not for *ATXN2* (FDR=0.4). We found that the expression of *SH2B3* is associated with expression of *MYADM*, *PP1R15A*, and *TAGAP* (at Bonferroni corrected p<0.05), but not with *FOS*, *S100A10*, or *FGBP2*. The expression of *ATXN2* was associated with expression of 5 of the 6 genes (*PP1R15A* was not associated). Supplementary Figure 3 shows the co-expression levels of the eight genes that were *cis*- or *trans*-associated with rs3184504 and rs653178 genotypes. These results suggest that there may be a pathway or gene co-regulatory mechanism underling BP regulation involving these genes that is driven by this common genetic variant (rs3184504; minor allele frequency 0.47) or its proxy SNPs.

We further checked whether the *cis*- or *trans*-eQTLs for the top 34 BP signature genes are associated with other diseases or traits in the NHGRI GWAS catalog [14]. We identified 12 *cis*-eQTLs (for 8 genes) and 6 *trans*-eQTLs (for 6 genes) that are associated with other diseases or traits in the NHGRI GWAS catalog [14] (Table 4).

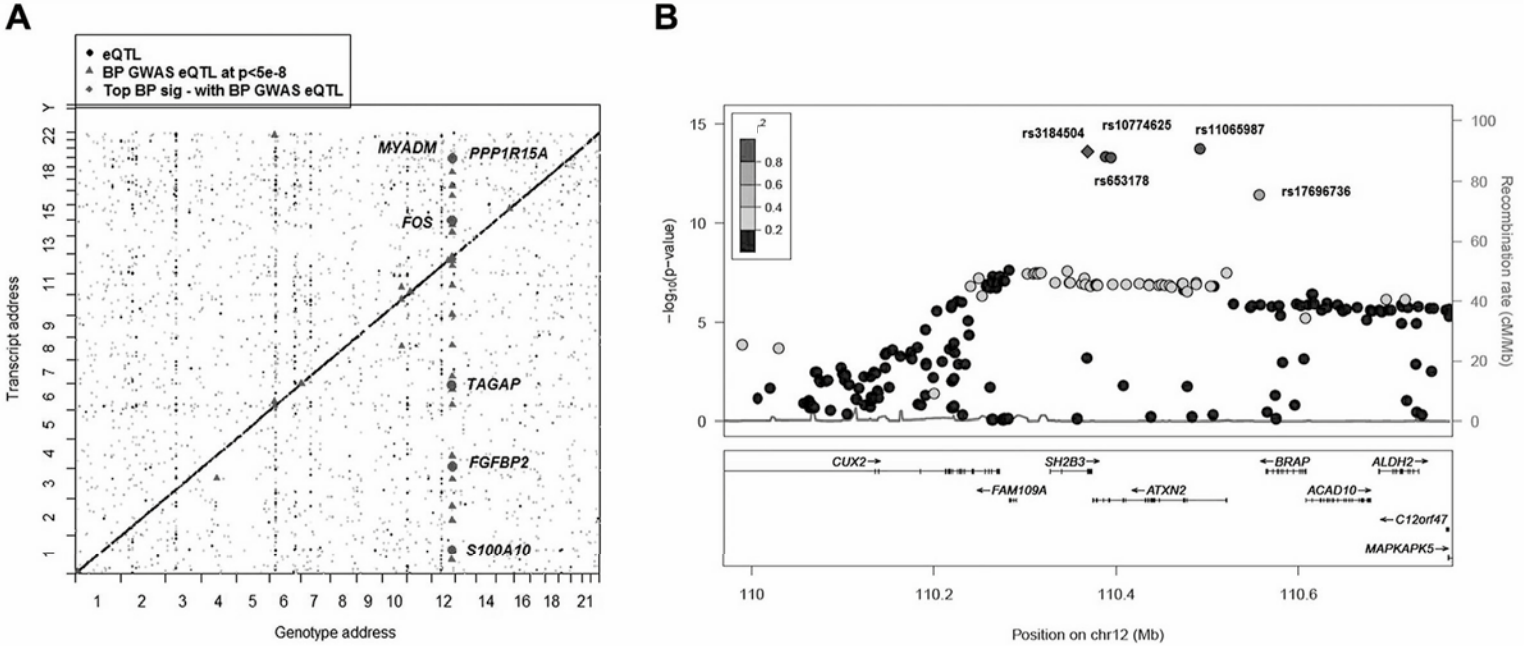


Figure 2. Global view of BP eQTLs effects on differentially expressed BP signature genes. (**A**) 2-Dimensional plot of in whole blood eQTLs vs. transcript position genome wide. eQTL-transcript pairs at FDR<0.1 are shown in black dots; those that fall along the diagonal are *cis*-eQTLs and all others are *trans*-eQTLS. eQTL-transcript pair SNPs that are associated with BP in GWAS [3] are highlighted with blue triangles. eQTL-transcript pair genes that are BP signature genes from analysis of differential gene expression in relation to BP are depicted by red circles. (**B**) Regional association plots for rs3184504 proxy QTLs that showing association with BP in ICBP GWAS [3]. -log10(p) indicated the -log10 transformed DBP association p values in ICBP GWAS [3]. Color coding indicates the strength (measured by r2) of LD of each SNP with the top SNP (rs3184504). Five master *trans*-eQTLs (also BP GWAS SNPs) for BP signature genes are labeled in the figure. This figure was drawn by LocusZoom [32].

Table 4. GWAS eQTLs for the top differentially expressed BP signature genes.

		SNP – Trait Association			SNP – Gene Association			Gene – Trait Association		
SNP	**SNP location**	**ICBP-SBP P-value**	**ICBP-DBP P-value**	**Other traits in GWAS catalogue**	**Gene**	**Chr Gene**	***Cis* or *Trans***	**SBP P-value**	**DBP P-value**	**HTN P-value**
rs3184504*	chr12	1.70E-09	2.30E-14	Coronary heart disease	MYADM	chr19	*trans*	**<1E-16&**	**1.1E-06**	**3.0E-07**
	(missense, SH2B3)			Rheumatoid arthritis	FOS	chr14	*trans*§	**4.9E-08**	3.2E-04	7.9E-05
				Type 1 diabetes	PPP1R15A	chr19	*trans*§	**1.6E-08**	1.2E-05	6.1E-04
					TAGAP	chr6	*trans*	**6.4E-06**	1.3E-04	**7.3E-07**
					S100A10	chr1	*trans*§	2.6E-04	**4.0E-08**	7.0E-05
					FGFBP2	chr4	*trans*§	**3.3E-08**	1.8E-05	5.1E-03
rs10187424	chr2 (intergenic)	-	-	Prostate cancer	GNLY	chr2	*cis*§	**4.0E-08**	2.8E-05	2.2E-04
rs411174	chr5 (intron, ITK)	-	-	Personality dimensions	HAVCR2	chr5	*cis*§	1.6E-04	**2.4E-07**	1.5E-03
rs3758354	chr9 (intergenic)	-	-	Schizophrenia, bipolar disorder and depression	ANXA1	chr9	*cis*	1.8E-03	**6.5E-11**	7.5E-03
rs1950500	chr14 (intergenic)	-	-	Height	GZMB	chr14	*cis*	7.8E-05	6.0E-05	**1.4E-06**
rs8017377	chr14 (missense, NYNRIN)	-	-	LDL cholesterol	GZMB	chr14	*cis*	7.8E-05	6.0E-05	**1.4E-06**
rs8192917	chr14 (missense, GZMB)	-	-	Vitiligo	GZMB	chr14	*cis*	7.8E-05	6.0E-05	**1.4E-06**
rs2284033	chr22 (intron, IL2RB)	-	-	Asthma	IL2RB	chr22	*cis*§	1.6E-04	**2.5E-08**	9.3E-03
rs11724635+	chr4 (intergenic)	-	-	Parkinson's disease	FBXL5	chr4	*cis*	5.9E-05	**5.3E-06**	7.0E-02
rs4333130$	chr4 (intron, ANTXR2)	-	-	Ankylosing spondylitis	ANTXR2	chr4	*cis*	2.8E-04	**1.7E-06**	4.0E-02
rs8005962	chr14 (intergenic)	-	-	Tuberculosis	GLRX5	chr14	*cis*	**1.5E-06**	1.3E-01	9.0E-02
rs7995215	chr13 (intron, GPC6)	-	-	Attention deficit hyperactivity disorder	TAGAP	chr6	*trans*	**6.4E-06**	1.3E-04	**7.3E-07**
rs12047808	chr1 (intron, C1orf125)	-	-	Multiple sclerosis (age of onset)	FOS	chr14	*trans*§	**4.9E-08**	3.2E-04	7.9E-05
rs2894207	chr6 (intergenic)	-	-	Nasopharyngeal carcinoma	AHNAK	chr11	*trans*	**5.2E-06**	6.8E-05	1.8E-03
rs3763313	chr6 (neargene 5, BTNL2)	-	-	HIV-1 control	PPP1R15A	chr19	*trans*	**1.6E-08**	1.2E-05	6.1E-04
rs9376092	chr6 (intergenic)	-	-	Beta thalassemia / hemoglobin E disease	GPR56	chr16	*trans*	**3.9E-11**	**5.5E-08**	4.9E-04

** rs653178, intronic to ATXN2 and in tight linkage disequilibrium with rs3184504 (r2=1), was also associated with BP in ICBP GWAS and all the 6 genes; + A proxy SNP rs4698412 at LD r2=1 associated with the same trait; $ A proxy SNP rs4389526 at LD r2=1 associated with the same trait; § indicated eQTL were identified from* [12]*. & highlighted p values indicated passing* transcriptome-*wide significance at Bonferroni corrected p<0*

DISCUSSION

Our meta-analysis of gene expression data from 7,017 individuals from six studies identified and characterized whole blood gene expression signatures associated with BP traits. Thirtyfour BP signature genes were identified at Bonferroni corrected p<0.05 (224 genes were identified at FDR<0.2, reported in the S1 Text). Thirteen BP signature genes replicated between the FHS and Illumina cohorts. The top BP signature genes identified in the FHS (55 genes for SBP and 22 genes for DBP) explained 5-9% of inter-individual variation in BP in the Illumina cohorts on average. Among the 34 BP signature genes (at Bonferroni corrected p<0.05), only FOS [15] and PTGS2 [16] have been previously implicated in hypertension. We did not find literature support for a direct role of the remaining signature genes in BP regulation. However, we found several genes involved in biological functions or processes that are highly related to BP, such as cardiovascular disease (*GZMB*, *ANXA1*, *TMEM43*, *FOS*, *KCNJ2*, *PTGS2*, and *MCL1*), angiogenesis (*VIM* and *TIPARP*), and ion channels (*CD97*, *ANXA1*, *S100A10*, *PRF1*, *ANTXR2*, *SLC31A2*, *TIPARP*, and *KCNJ2*). We speculate that these genes may be important for BP regulation, but further experimental validation is needed.

Seven of the 34 signature genes, including *KCNJ2*, showed negative correlation of expression with BP. *KCNJ2* is a member of the potassium inwardly-rectifying channel subfamily; it encodes the inward rectifier K+ channel Kir2.1, and is found in cardiac, skeletal muscle, and nervous tissue [17]. Most outward potassium channels are positively correlated with BP. Loss-of function mutations in ROMK (*KCNJ1*, the outward potassium channel) are associated with Bartter's syndrome, and ROMK inhibitors are used in the treatment of hypertension [18,19]. Previous studies reported that greater potassium intake is associated with lower blood pressure [20-23]. These data suggest that *KCNJ2* up-regulation may be a means of lowering BP.

By linking the BP signature genes with eQTLs and with BP GWAS results, we found several SNPs that are associated with BP in GWAS and that also are *trans* associated with several of our top BP signature genes. For example, rs3184504, a non-synonymous SNP located in exon 3 of *SH2B3*, is associated in GWAS with BP, coronary heart disease, hypothyroidism, rheumatoid arthritis, and type I diabetes [12]. rs3184504 is a common genetic variant with a minor allele frequency of approximately 0.47; the rs3184504-T allele is associated with an increment of 0.58 mm Hg in SBP and of 0.48 mm Hg in DBP [2]. rs3184504 is a *cis*-eQTL for *SH2B3*, expression of this gene was not associated with BP or hypertension in our data. However, rs3184504 also is a *trans*-eQTL for 6 of our 34 BP signature genes: *FOS*, *MYADM*, *PP1R15A*, *TAGAP*, *S100A10*, and *FGBP2*. These 6 genes are highly expressed in neutrophils [12], and are co-expressed. Prior studies have suggested an important role of neutrophils in BP regulation [24]. We speculate that these 6 BP signature genes, all driven by the same BP-associated eQTL, point to a critical and previously unrecognized mechanism involved in BP regulation. Further experimental validation is needed.

One limitation of our study is the use of whole blood derived RNA for transcriptomic profiling. GSEA showed that the top enriched biological processes for the differentially expressed BP genes

include inflammatory response. Numerous studies have shown links between inflammation and hypertension [25-27]. The top ranked genes in inflammatory response categories provide a guide for further experimental work to recognize the contributions of inflammation to alterations in BP regulation. We speculate that using similar approaches in other tissues might identify additional differentially expressed BP signature genes. In conclusion, we conducted a meta-analysis of global gene expression profiles in relation to BP and identified a number of credible gene signatures of BP and hypertension. Our integrative analysis of GWAS and gene expression in relation to BP can help to uncover the genetic and genomic architecture of BP regulation; the BP signature genes we identified may represent an early step toward improvements in the detection of susceptibility, and in the prevention and treatment of hypertension.

MATERIALS AND METHODS

Study population and ethics statement

This investigation included six studies (the Framingham Heart Study (FHS), the Estonian Biobank (EGCUT), the Rotterdam Study (RS) [8], the InCHIANTI Study, the Cooperative Health Research in the Region of Augsburg (KORA F4) Study [9], and the Study of Health in Pomerania (SHIP-TREND) [10], each of which conducted genome-wide genotyping, mRNA expression profiling, and had extensive BP phenotype data. Each of the six studies followed the recommendations of the Declaration of Helsinki. The FHS: Systems Approach to Biomarker Research (SABRe) in cardiovascular disease is approved under the Boston University Medical Center's protocol H-27984. Ethical approval of EGCUT was granted by the Research Ethics Committee of the University of Tartu (UT REC). Ethical approval of the InCHIANTI study was granted by the Instituto Nazionale Riposo e Cura Anziani institutional review board in Italy. Ethical approval of RS was granted by the medical ethics committee of the Erasmus Medical Center. The study protocol of SHIP-TREND was approved by the medical ethics committee of the University of Greifswald. KORA F4 is a population-based survey in the region of Augsburg in Southern Germany which was performed between 2006 and 2008. KORA F4 was approved by the local ethical committees. Informed consent was obtained from each study participant.

Definition of the phenotype

Hypertension (HTN) was defined as SBP ≥140 mm Hg or DBP ≥90 mm Hg. We excluded individuals receiving anti-hypertensive treatment because of the possibility that some of the differentially expressed genes we identified would reflect treatment effects. The eligible study sample included 7,017 individuals: 3,679 from FHS, 972 from EGCUT, 604 from RS, 597 from InCHIANTI, 565 from KORA F4, and 600 from SHIP-TREND.

Gene expression profiling

RNA was isolated from whole blood samples that were collected in PaxGene tubes (PreAnalytiX, Hombrechtikon, Switzerland) in FHS, RS, InCHIANTI, KORA F4 and SHIP-TREND, and in Blood RNA Tubes (Life Technologies, NY, USA) in EGCUT. Gene expression in the FHS samples used the

Affymetrix Exon Array ST 1.0. EGCUT, RS, InCHANTI, KORA F4, and SHIP-TREND used the Illumina HT12v3 (EGCUT, InCHANTI, KORA F4, and SHIPTREND) or HT12v4 (RS) array. Raw data from gene expression profiling are available online (FHS [*http://www.ncbi.nlm.nih.gov/gap*; accession number phs000007], EGCUT [GSE48348], RS [GSE33828], InCHIANTI [GSE48152], KORA F4 [E-MTAB-1708] and SHIP-TREND [GSE36382]). The details of sample collection, microarrays, and data processing and normalization in each cohort are provided in the S2 Text.

Identification and replication of differentially expressed genes associated with BP

The association of gene expression with BP was analyzed separately in each of the six studies (Equation 1). A linear mixed model was used in the FHS in order to account for family structure. Linear regression models were used in the other five studies. In each study, gene expression level, denoted by geneExp, was included as the dependent variable, and explanatory variables included blood pressure phenotypes (SBP, DBP, and HTN), and covariates included age, sex, body mass index (BMI), cell counts, and technical covariates. A separate regression model was fitted for each gene. The general formula is shown below, and the details of analyses for each study are provided in the S2 Text and S6 Table.

$$geneExp = BP + \sum_{j=1}^{m} covariates \quad \textbf{(1)}$$

The overall analysis framework is provided in Supplementary Figure 1. We first identified differentially expressed genes associated with BP (BP signature genes) in the FHS samples (Set 1) and attempted replication in the meta-analysis results from the Illumina cohorts (Set 2, see Methods, Meta-analysis). We next identified BP signature genes in the Illumina cohorts (Set 2), and then attempted replication in the FHS samples (Set 1). The significance threshold for pre-selecting BP signature genes in discovery was at Bonferroni corrected p=0.05 (in FHS, corrected for 17,318 measured genes [17,873 transcripts], and in illumina cohorts, corrected for 12,010 measured genes [14,222 transcripts] that passed quality control). Replication was established at Bonferroni corrected p=0.05, correcting for the number of pre-selected BP signatures genes in the discovery set. We computed the pi1 value to estimate the enrichment of significant p values in the replication set (the Illumina cohorts) for BP signatures identified in the discovery set (the FHS) by utilizing the R package Qvalue [11]. *Pi1* is defined as 1-*pi0*. *Pi0* value provided by the Qvalue package, represents overall probability that the null hypothesis is true. Therefore, *pi1* value represents the proportion of significant results. For genes passing Bonferroni corrected p<0.05 in the discovery set for SBP, DBP and HTN, we calculated *pi1* values for each gene set in the replication set.

Meta-analysis

We performed meta-analysis of the five Illumina cohorts (for discovery and replication purposes), and then performed meta-analysis of all six cohorts. An inverse variance weighted meta-analysis was conducted using fixed-effects or random-effects models by the metagen function in the R package Meta (*http://cran.r-project.org/web/packages/meta/index.html*). At first, we tested heterogeneity for each gene using Cochran's Q statistic. If the heterogeneity p value is significant (p<0.05), we will use

a random-effects model for the meta-analysis, otherwise use a fixed-effects model. The Benjamini-Hochberg (BH) method [28] was used to calculate FDR for differentially expressed genes in relation to BP following the meta-analysis of all six cohorts. We also used a more stringent threshold to define BP signature genes by utilizing p<6.5E-6 (Bonferroni correction for 7,717 unique genes [7,810 transcript] based on the overlap of FHS and illumina cohort interrogated gene sets).

Estimating the proportion of variance in BP attributable to BP signature genes

To estimate the proportion of variances in SBP or DBP explained by a group of differentially expressed BP signature genes (gene 1, gene 2, . . ., gene n), we used the following two models:
Full model:

$$BP = \sum_{i=1}^{n} gene\ i + \sum_{j=1}^{m} covariates \tag{2}$$

Null model:

$$BP = \sum_{j=1}^{m} covariates \tag{3}$$

The proportion of variance in BP attributable to the group of differentially expressed BP signature genes ($h^2_{BP\_sig}$) was calculated as:

$$h^2_{BP\_sig} = \max(0, \frac{\sigma^2_{G.\,null} + \sigma^2_{err.\,null} - \sigma^2_{G.\,full} + \sigma^2_{err.\,full}}{\sigma^2_{BP}}) \tag{4}$$

where σ^2_{BP} is the total phenotypic variance of SBP or DBP, $\sigma^2_{G.full}$ and $\sigma^2_{err.full}$ are the variance and error variance when modeling with the tested group of gene expression traits (gene 1, gene 2, . . ., gene n), and $\sigma^2_{G.null}$ and $\sigma^2_{err.null}$ are the variance and error variance when modeling without the tested group of gene expression traits. The proportion of the variance in BP phenotypes attributable to the FHS BP signature genes was estimated in the five Illumina cohorts, respectively, and then the average proportion values were reported. In turn, the proportion of the variance in BP phenotypes attributable to the Illumina BP signature genes was estimated in the FHS.

Identifying eQTLs and estimating the proportion of variance in gene expression attributable to single *cis*- or *trans*-eQTLs

SNPs associated with altered gene expression (i.e. eQTLs) were identified using genome-wide genotype and gene expression data in all available FHS samples (n=5,257) at FDR<0.1 (Joehanes R, submitted, 2014, and a brief summary of methods and results are provided in the S2 Text). A *cis*-eQTL was defined as an eQTL within 1 megabase (MB) flanking the gene. Other eQTLs were defined as *trans*-eQTLs. We combined the eQTL list generated in the FHS with the eQTLs generated by meta-analysis of seven other studies (n=5,300) that were also based on whole blood expression [12].

For every BP signature gene, we estimated the proportion of variance in the transcript attributable to the corresponding *cis*- or *trans*-eQTLs (h^2_{eQTL}) using the formula:

$$h^2_{eQTL} = \max(0, \frac{\sigma^2_{eQTL.\,null} + \sigma^2_{err.\,null} - \sigma^2_{eQTL.\,full} + \sigma^2_{err.\,full}}{\sigma^2_{gene}}) \quad (5)$$

where σ^2_{gene} was the total phenotypic variance of a gene expression trait; $\sigma^2_{eQTL.full}$ and $\sigma^2_{err.full}$ were the variance and the residual error, respectively, when modeling with the tested eQTL; $\sigma^2_{eQTL.null}$ and $\sigma^2_{err.null}$ were the variance and the residual error when modeling without the tested eQTL.

Functional category enrichment analysis

In order to understand the biological themes within the global gene expression changes in relation to BP, we performed gene set enrichment analysis [29] to test for enrichment of any gene ontology (GO) biology process [30] or KEGG pathways [31]. "Metric for ranking gene" parameters were configured to the beta value of the meta-analysis, in order to look at the top enriched functions for BP associated up-regulated and down-regulated gene expression changes respectively. One thousand random permutations were conducted and the significance level was set at FDR≤0.25 to allow for exploratory discovery [29].

ACKNOWLEDGMENTS

We thank the field staff in Augsburg who was involved in the conduct of the studies. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists. We thank Marjolein Peters, MSc, Ms. Mila Jhamai, Ms. Jeannette M. Vergeer-Drop, Ms. Bernadette van Ast-Copier, Mr. Marijn Verkerk and Jeroen van Rooij, BSc for their help in creating the RNA array expression database. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (*http://biowulf.nih.gov*).

ADDITIONAL INFORMATION

Supplementary Information accompanies this paper at *http://journals.plos.org/plosgenetics/*.

REFERENCES

1. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, et al. (2003) Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. Hypertension 42: 1206–1252. PMID: 14656957.
2. Levy D, Ehret GB, Rice K, Verwoert GC, Launer LJ, et al. (2009) Genome-wide association study of blood pressure and hypertension. Nat Genet 41: 677–687. doi: 10.1038/ng.384 PMID: 19430479.
3. Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, et al. (2011) Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 478: 103–109. doi: 10.1038/nature10405 PMID: 21909115.
4. Leonardson AS, Zhu J, Chen Y, Wang K, Lamb JR, et al. (2010) The effect of food intake on gene expression in human peripheral blood. Hum Mol Genet 19: 159–169. doi: 10.1093/hmg/ddp476 PMID:19837700.
5. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, et al. (2010) Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. PLoS One 5: e10693. doi: 10.1371/journal.pone.0010693 PMID: 20502693.
6. Bull TM, Coldren CD, Moore M, Sotto-Santiago SM, Pham DV, et al. (2004) Gene microarray analysis of peripheral blood cells in pulmonary arterial hypertension. Am J Respir Crit Care Med 170: 911–919. PMID: 15215156.
7. Korkor MT, Meng FB, Xing SY, Zhang MC, Guo JR, et al. (2011) Microarray analysis of differential gene expression profile in peripheral blood cells of patients with human essential hypertension. Int J Med Sci 8: 168–179. PMID: 21369372.
8. Hofman A, van Duijn CM, Franco OH, Ikram MA, Janssen HL, et al. (2011) The Rotterdam Study: 2012 objectives and design update. Eur J Epidemiol 26: 657–686. doi: 10.1007/s10654-011-9610-5 PMID:21877163.
9. Schurmann C, Heim K, Schillert A, Blankenberg S, Carstensen M, et al. (2012) Analyzing illumina gene expression microarray data from different tissues: methodological aspects of data analysis in the metaxpress consortium. PloS one 7: e50938. doi: 10.1371/journal.pone.0050938 PMID: 23236413.
10. Volzke H, Alte D, Schmidt CO, Radke D, Lorbeer R, et al. (2011) Cohort profile: the study of health in Pomerania. Int J Epidemiol 40: 294–307. doi: 10.1093/ije/dyp394 PMID: 20167617.
11. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 100: 9440–9445. PMID: 12883005.
12. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, et al. (2013) Systematic identification of trans-eQTLs as putative drivers of known disease associations. Nature genetics 45: 1238–1243. doi: 10.1038/ng.2756 PMID: 24013639.
13. Joehanes R., Huan T., C Yao, X Zhang, S Ying, et al. (2013) Genome-wide Expression Quantitative Trait Loci: Results from the NHLBI's SABRe CVD Initiative. the American Society of Human Genetics(ASHG) conference. Boston Convention Ctr. Boston, MA.
14. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106: 9362 9367. doi: 10.1073/pnas.0903103106 PMID: 19474294.
15. Rowland NE, Li BH, Fregly MJ, Smith GC (1995) Fos induced in brain of spontaneously hypertensive rats by angiotensin II and co-localization with AT-1 receptors. Brain Res 675: 127–134. PMID:7796121.
16. Beetz N, Harrison MD, Brede M, Zong X, Urbanski MJ, et al. (2009) Phosducin influences sympathetic activity and prevents stress-induced hypertension in humans and mice. J Clin Invest 119: 3597–3612. doi: 10.1172/JCI38433 PMID: 19959875.
17. Hibino H, Inanobe A, Furutani K, Murakami S, Findlay I, et al. (2010) Inwardly rectifying potassium channels: their structure, function, and physiological roles. Physiol Rev 90: 291–366. doi: 10.1152/physrev.00021.2009 PMID: 20086079.
18. Felix JP, Priest BT, Solly K, Bailey T, Brochu RM, et al. (2012) The inwardly rectifying potassium channel Kir1.1: development of functional assays to identify and characterize channel inhibitors. Assay Drug Dev Technol 10: 417–431. doi: 10.1089/adt.2012.462 PMID: 22881347.

19. Fang L, Li D, Welling PA (2010) Hypertension resistance polymorphisms in ROMK (Kir1.1) alter channel function by different mechanisms. Am J Physiol Renal Physiol 299: F1359–1364. doi: 10.1152/ajprenal.00257.2010 PMID: 20926634.

20. Cappuccio FP, MacGregor GA (1991) Does potassium supplementation lower blood pressure? A meta-analysis of published trials. J Hypertens 9: 465–473. PMID: 1649867.

21. Geleijnse JM, Kok FJ, Grobbee DE (2003) Blood pressure response to changes in sodium and potassium intake: a metaregression analysis of randomised trials. J Hum Hypertens 17: 471–480. PMID:12821954.

22. Fulgoni VL 3rd (2007) Limitations of data on fluid intake. J Am Coll Nutr 26: 588S–591S. PMID: 17921470.

23. Koliaki C, Katsilambros N (2013) Dietary sodium, potassium, and alcohol: key players in the pathophysiology, prevention, and treatment of human hypertension. Nutr Rev 71: 402–411. doi: 10.1111/nure.12036 PMID: 23731449.

24. Morton J, Coles B, Wright K, Gallimore A, Morrow JD, et al. (2008) Circulating neutrophils maintain physiological blood pressure by suppressing bacteria and IFNgamma-dependent iNOS expression in the vasculature of healthy mice. Blood 111: 5187–5194. doi: 10.1182/blood-2007-10-117283 PMID:18281503.

25. Harrison DG, Guzik TJ, Lob HE, Madhur MS, Marvar PJ, et al. (2011) Inflammation, immunity, and hypertension. Hypertension 57: 132–140. doi: 10.1161/HYPERTENSIONAHA.110.163576 PMID:21149826.

26. Harrison DG, Marvar PJ, Titze JM (2012) Vascular inflammatory cells in hypertension. Front Physiol 3:128. doi: 10.3389/fphys.2012.00128 PMID: 22586409.

27. Harrison DG, Vinh A, Lob H, Madhur MS (2010) Role of the adaptive immune system in hypertension. Curr Opin Pharmacol 10: 203–207. doi: 10.1016/j.coph.2010.01.006 PMID: 20167535.

28. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological): 289–300.

29. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–15550. PMID: 16199517.

30. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. Nature genetics 25: 25–29. PMID: 10802651.

31. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research 28: 27–30. PMID: 10592173.

32. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, et al. (2010) LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26: 2336–2337. doi: 10.1093/bioinformatics/btq419 PMID: 20634204.

CHAPTER 2.3

Meta-analysis of whole-blood gene expression associations with circulating lipid levels

Brian H. Chen*, Claudia Schurmann*, Marjolein J. Peters*, Katharina Schramm*, Luke C. Pilling*, Liliane Pfeiffer*, Roby Joehanes*, Stefania Bandinelli, Maren Carstensen-Kirberg, Paul Courchesne, L. Adrienne Cupples, Valur Emilsson, Tonu Esko, Stephan B. Felix, Lita Freeman, Harald Grallert, Dena G. Hernandez, Albert Hofman, Georg Homuth, Tianxiao Huan, Till Ittermann, Andrew D. Johnson, Sekar Kathiresan, Thomas Meitinger, Peter J. Munson, Matthias Nauck, Gina Peloso, Annette Peters, Eva Reischl, Alan T. Remaley, Michael Roden, Andrew B. Singleton, Seth G. Thacker, André G. Uitterlinden, Cornelia van Dujin, Joyce B.J. van Meurs, Melanie Waldenberger, Hanieh Yaghootkar, Saixia Ying, David Melzer*, Luigi Ferrucci*, Holger Prokisch*, Christian Herder*, Aaron Isaacs*, Daniel Levy*, Alexander Teumer* on behalf of the CHARGE Consortium Gene Expression Working Group.

** These authors contributed equally to this work*

Manuscript submitted

ABSTRACT

Genome-wide association studies (GWAS) have identified genetic loci associated with circulating lipid levels. To complement these findings, we conducted a meta-analysis of gene expression associations with triglycerides (TG), HDL cholesterol (HDL-C), LDL cholesterol (LDL-C), and total cholesterol (TC) from whole blood-derived mRNA levels using either the Affymetrix GeneChip Human Exon 1.0 ST or the Illumina HumanHT-12 Expression BeadChip arrays in 4,841 fasting individuals. Individuals taking lipid-lowering medications were excluded from analysis. Using a significance level of $p<1 \times 10^{-6}$ (corresponding to Bonferroni correction for 12,492 unique genes across four traits), expression levels for 906 genes were significantly associated with levels of at least one lipid trait (793 for TG, 489 for HDL-C, 20 for TC, and five for LDL-C). We identified a set of basophil and mast cell-related genes whose expression levels were highly associated with all four lipid traits. The two genes with the smallest p-values were *HDC* ($p_{TG}=5.3 \times 10^{-268}$, $p_{HDL\text{-}C}=5.2 \times 10^{-61}$, $p_{TC}=2.1 \times 10^{-11}$, $p_{LDL\text{-}C}=1.3 \times 10^{-21}$) and *CPA3* ($p_{TG}=3.8 \times 10^{-191}$, $p_{HDL\text{-}C}=3.6 \times 10^{-36}$, $p_{TC}=2.7 \times 10^{-21}$, $p_{LDL\text{-}C}=2.5 \times 10^{-10}$). Expression levels of these two genes explained 23.1% and 18.1% of the total variation in log-transformed TG levels, respectively. Our findings add further support for the role of basophil and mast cells in the link between dyslipidemia and atherosclerosis. Furthermore, we observed significant associations between lipid levels and the expression levels of 95 known lipid-related genes in the LIPID MAPS proteomic database and 47 genes identified by lipid. Thus, we provide evidence that gene expression data in whole blood may be helpful in identifying novel genes and pathways involved in the regulation and downstream effects of circulating lipid levels.

AUTHOR SUMMARY

Large-scale genetic studies have identified DNA sequence variants that are associated with blood lipid levels. Identifying genes that are differentially expressed in relation to lipid levels may provide additional insights into the lipid regulatory landscape and help prioritize genes for therapeutic targeting. We examined the associations between whole blood expression levels of 12,492 genes and circulating lipid levels. Using gene expression data from 4,841 individuals, we identified 906 genes whose expression was significantly associated with blood lipid levels. While our strongest associations were likely due to changes in gene expression caused by lipid levels, other associations may be involved in the regulation of lipid levels. To help identify the latter, we cross-referenced our findings with those from a large genome-wide association study (GWAS) meta-analysis of circulating lipid levels. Genes identified through the convergence of GWAS and the current gene expression study should be prioritized for further research on their potential roles in lipid regulation.

INTRODUCTION

Dyslipidemia is a highly predictive and readily modifiable risk factor for atherosclerotic cardiovascular disease (CVD) [1]. Several pharmacologic therapies have been proven to effectively reduce the risk of CVD by altering lipid levels [2-4]. The discovery of new and targeted therapeutics through genomic approaches has recently been shown to be effective for lipid lowering and CVD event prevention [5,6].

Genome-wide association studies (GWAS) have identified 157 genetic loci associated with circulating levels of total cholesterol (TC), low-density lipoprotein-cholesterol (LDL-C), high-density lipoprotein-cholesterol (HDL-C), and triglycerides (TG) [7,8]. Yet the aggregate effect of single genetic variants only explain 10-12% of the total variation in circulating lipid levels [7,8]. Additional sources of inter-individual variability in lipid levels remain to be identified [9]. Gene expression reflects the integration of multiple levels of genomic regulation; thus, studies of gene expression studies may complement GWAS by offering an alternate high throughput and unbiased approach to identifying strong genetic associations.

Identifying genes whose blood expression levels correlate with circulating lipid levels is important for two major reasons. First, these genes may be involved in upstream steps related to lipid synthesis or metabolism, or they may play a role in the downstream steps related to the pathobiology linking circulating lipids to the development of atherosclerosis [10].

The current meta-analysis of circulating lipid levels used gene expression data from whole blood-derived mRNA in five population-based cohorts totaling 4,841 individuals who were not receiving lipid lowering medication. Using an analytic strategy that integrated multiple data sources (e.g., GWAS) shown in Figure 1, we were able to recapitulate known lipid pathway genes in addition to providing potential insights into whole blood gene expression correlates of circulating lipid levels.

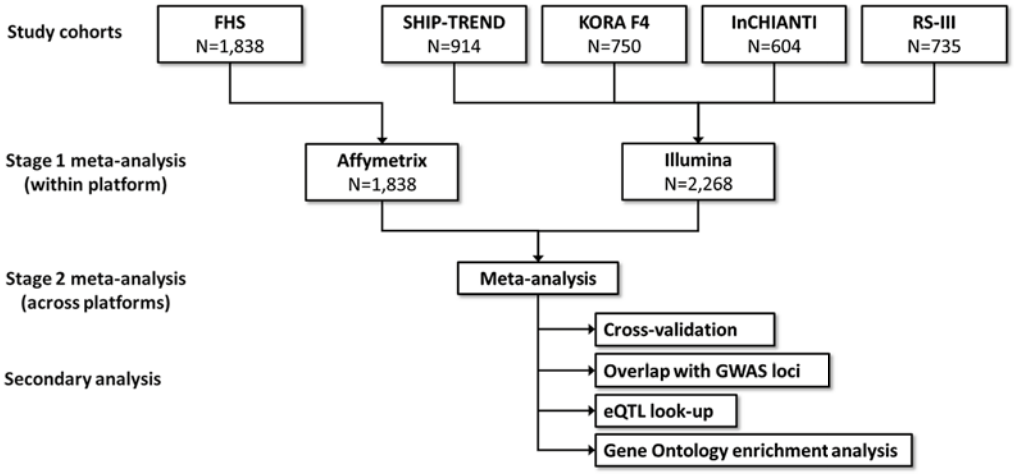


Figure 1. Overview of study design.

RESULTS

Study population characteristics

Study participants were generally middle-aged and elderly individuals from 5 observational studies with mean ages ranging from 49 to 72 years (Table 1). Individuals taking lipid-lowering medications were excluded from analysis. Lipid levels were comparable across cohorts with the exception of lower LDL-C levels in the Framingham Heart Study (FHS). Between 42 to 49% of participants in each cohort were men.

Table 1. Characteristics of study populations.

	Affymetrix cohort	Illumina cohorts			
Variables	**Framingham**	**SHIP-TREND**	**KORA F4**	**RS-III**	**InCHIANTI**
n	3,978	917	750	736	604
n (with cell counts)	1,838	914	750	735	604
Age (years), mean±SD	52 ± 13	49 ± 13	70 ± 5	59 ± 8	72 ± 16
Sex (men), n (%)	1,658 (42%)	399 (44%)	367 (49%)	325 (44%)	268 (45%)
BMI (kg/m^2), mean±SD	27.6 ± 5.6	27.2 ± 4.6	28.9 ± 4.6	27.4 ± 4.6	27.0 ± 4.3
Fasting (yes), n (%)	3,978 (100%)	917 (100%)	743 (99%)	724 (98%)	604 (100%)
Lipid levels, median (IQR)*					
Triglycerides (mmol/L)	1.0 (0.8-1.5)	1.2 (0.9-1.7)	1.3 (0.9-1.8)	1.2 (0.9-1.7)	1.2 (0.9-1.6)
Total cholesterol (mmol/L)	5.0 (4.4-5.5)	5.5 (4.8-6.3)	5.9 (5.2-6.5)	5.7 (5.0-6.3)	5.3 (4.6-6.0)
HDL cholesterol (mmol/L)	1.5 (1.2-1.9)	1.4 (1.2-1.7)	1.4 (1.2-1.7)	1.4 (1.1-1.7)	1.4 (1.2-1.7)
LDL cholesterol (mmol/L)†	2.8 (2.3-3.4)	3.4 (2.8-4.1)	3.8 (3.2-4.4)	3.6 (3.0-4.2)	3.2 (2.6-3.8)
Microarray platform	Affymetrix GeneChip Human Exon 1.0 ST	Illumina HumanHT-12 v3 BeadChip	Illumina HumanHT-12 v3 BeadChip	Illumina HumanHT-12 v4 BeadChip	Illumina HumanHT-12 v3 BeadChip

** Triglyceride conversion from mmol/L to mg/dl, multiply by 88.5. Cholesterol conversion from mmol/L to mg/dl, multiply by 38.6. † LDL cholesterol was computed using the Friedewald formula in Framingham, RS-III, and InChianti.*

Blood count and lipid correlations

Among lipid traits, LDL-C and TC were highly correlated (Spearman r=0.85). To a lesser extent HDL-C and TG also were correlated (Spearman r=-0.51). Because tissue heterogeneity may be a potential confounding factor, we assessed its potential impact by examining associations between blood cell counts and lipid levels. The strongest Spearman correlations were observed for red blood cell counts and TG (r=0.26), HDL-C (r=-0.43), and LDL-C (r=0.24) in FHS (S1 Table). Weaker correlations were observed for TG and HDL-C levels in relation to white blood cell counts (r=0.25 for TG; r=-0.18 for HDL-C), neutrophils (r=0.22 for TG; *r*=-0.17 for HDL-C), and monocytes (r=0.18 for TG; r=-0.20 for HDL-C). All aforementioned correlations had $p<1 \times 10^{-14}$. Further adjustment for age did not substantially affect these correlations (S2 Table), suggesting that the association between lipids

and blood counts were not entirely due to age-related changes in both parameters. All subsequent analyses were adjusted for age, sex, technical covariates, and blood cell counts to address potential confounding.

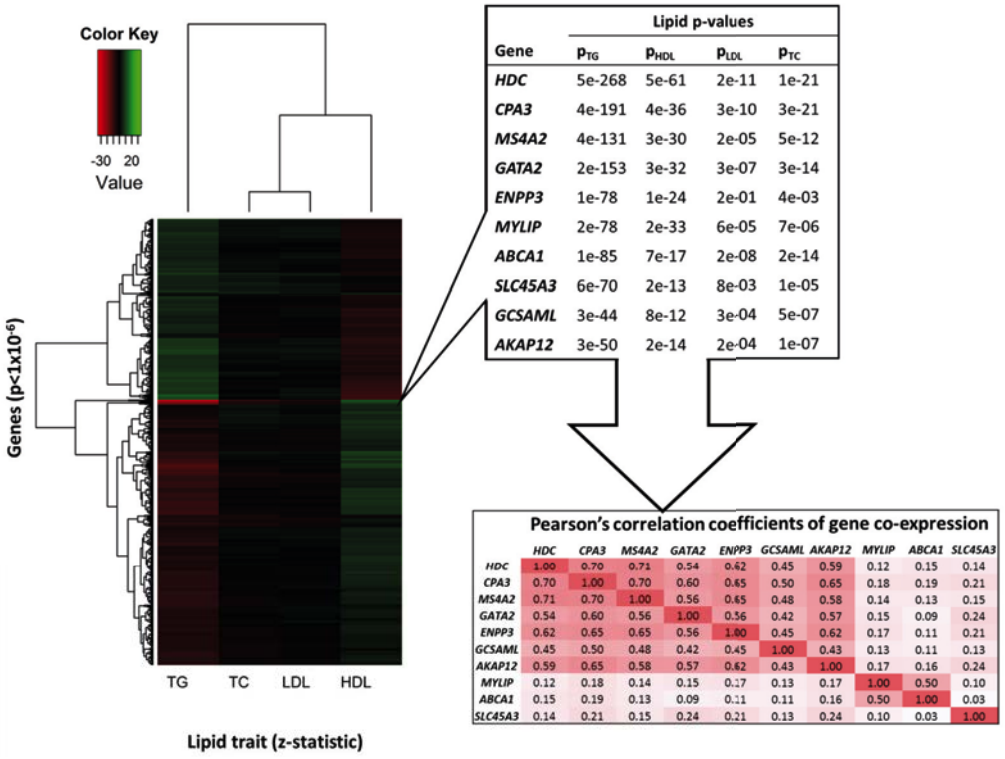
Gene expression associations with circulating lipid levels

Due to differences in microarray platforms used across cohorts, we conducted a two stage meta-analysis (Figure 1). In Stage 1, results from the cohorts using Illumina arrays were meta-analyzed. In Stage 2, results from the meta-analysis of Illumina arrays were meta-analyzed with results from the single Affymetrix cohort (FHS) using an inverse variance weighted random effects model. The Stage 2 meta-analysis of 12,492 genes common to both platforms yielded 792 genes for TG, 488 for HDL-C, 20 for TC, and five for LDL-C whose expression levels were significantly associated with lipid levels at a Bonferroni-corrected $p<1.0 \times 10^{-6}$. The ten most statistically significant expression associations for each lipid trait are presented in Table 2, and the complete results can be found in S1-S5 Files.

At a Bonferroni-corrected significance level of $p<1.0 \times 10^{-6}$, 906 genes were significantly associated with at least one lipid trait. Many of these 906 genes had similar patterns of gene expression associations with lipid traits, suggesting gene co-expression clusters (Figure 2).

In particular, one cluster of genes, with inter-correlations ranging from 0.42 to 0.71, was highly associated with TG and HDL-C. Within this cluster were two sub-clusters based on co-expression patterns. One sub-cluster included *ABCA1* and *MYLIP*, known lipid-regulating genes. The other sub-cluster included *HDC*, *CPA3*, *MS4A2*, *GATA2*, *ENPP3*, *GCSAML*, and *AKAP12*. Because several genes in the latter sub-cluster were known basophil and mast cell-specific genes, we examined the correlation between their expression levels and percent basophils in blood using data from FHS (Figure 3). As expected, moderate correlations were observed between percent basophils in blood and the expression levels of genes within this sub-cluster. The strongest of these correlations were for *CPA3* ($r=0.45$, $p=5.8 \times 10^{-93}$), *HDC* ($r=0.43$, $p=3.5 \times 10^{-84}$), and *MS4A2* ($r=0.41$, $p=5.8 \times 10^{-76}$). *ABCA1* and *MYLIP* were not correlated with percent basophils ($r=0.08$ and $r=0.05$, respectively).

In general, gene expression associations (i.e., *t*-statistics) for TC were highly correlated with those for LDL-C ($r=0.81$, $p<1 \times 10^{-16}$), and TG associations were inversely correlated with HDL-C associations ($r=-0.74$, $p<1 \times 10^{-16}$) (S1 Figure). Furthermore, results adjusting for body mass index (BMI) were highly similar to results without BMI adjustment ($r>0.94$ for all lipid traits) (S2 Figure).



Gene	Lipid p-values			
	p_{TG}	p_{HDL}	p_{LDL}	p_{TC}
HDC	5e-268	5e-61	2e-11	1e-21
CPA3	4e-191	4e-36	3e-10	3e-21
MS4A2	4e-131	3e-30	2e-05	5e-12
GATA2	2e-153	3e-32	3e-07	3e-14
ENPP3	1e-78	1e-24	2e-01	4e-03
MYLIP	2e-78	2e-33	6e-05	7e-06
ABCA1	1e-85	7e-17	2e-08	2e-14
SLC45A3	6e-70	2e-13	8e-03	1e-05
GCSAML	3e-44	8e-12	3e-04	5e-07
AKAP12	3e-50	2e-14	2e-04	1e-07

Pearson's correlation coefficients of gene co-expression

	HDC	*CPA3*	*MS4A2*	*GATA2*	*ENPP3*	*GCSAML*	*AKAP12*	*MYLIP*	*ABCA1*	*SLC45A3*
HDC	1.00	0.70	0.71	0.54	0.62	0.45	0.59	0.12	0.15	0.14
CPA3	0.70	1.00	0.70	0.60	0.65	0.50	0.65	0.18	0.19	0.21
MS4A2	0.71	0.70	1.00	0.56	0.65	0.48	0.58	0.14	0.13	0.15
GATA2	0.54	0.60	0.56	1.00	0.56	0.42	0.57	0.15	0.09	0.24
ENPP3	0.62	0.65	0.65	0.56	1.00	0.45	0.62	0.17	0.11	0.21
GCSAML	0.45	0.50	0.48	0.42	0.45	1.00	0.43	0.13	0.11	0.13
AKAP12	0.59	0.65	0.58	0.57	0.62	0.43	1.00	0.17	0.16	0.24
MYLIP	0.12	0.18	0.14	0.15	0.17	0.13	0.17	1.00	0.50	0.10
ABCA1	0.15	0.19	0.13	0.09	0.11	0.11	0.16	0.50	1.00	0.03
SLC45A3	0.14	0.21	0.15	0.24	0.21	0.13	0.24	0.10	0.03	1.00

Figure 2. Heatmap of meta-analysis Z-scores for 906 genes whose expression levels were significantly associated (after Bonferroni correction) with a circulating lipid in the meta-analysis (left inset figure). Upper inset table provides *p*-values for the associations between lipid levels and expression levels for each gene in the cluster. Lower inset table is a correlation matrix of the Pearson's correlation coefficients for the co-expression of these genes. All models adjusted for age, age^2, sex, batch effects, and blood cell counts.

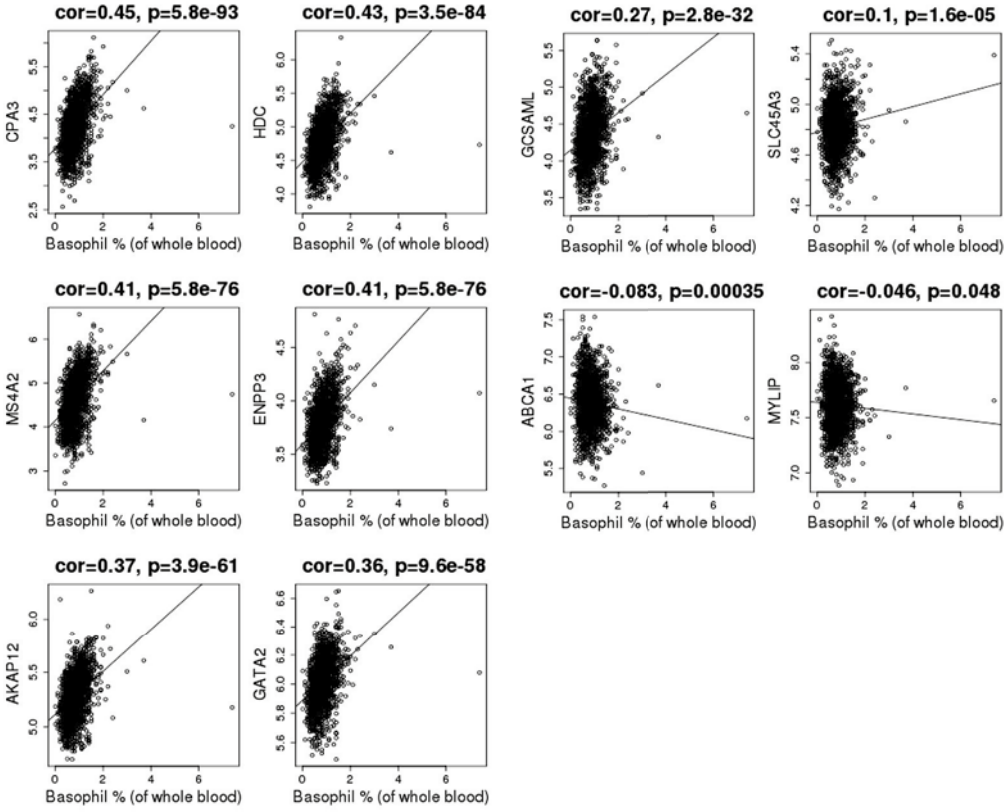


Figure 3. Scatterplots of percent basophil versus gene expression levels in whole blood. Regression line and Pearson's correlation coefficients are shown. Removal of 3 outliers for percent basophil slightly strengthened correlations for all genes except for *GCSAML*, *SLC45A3*, and *ABCA1*, whose correlations were slightly attenuated.

Cross-validation

To evaluate the reproducibility of our findings, we compared the strength and direction of gene expression associations across platforms (i.e., Stage I Affymetrix results compared to Stage I Illumina results) (Figure 4). Overall, TG and HDL-C had the strongest and most reproducible associations, whereas LDL-C and TC associations tended to be less reproducible. More importantly, gene expression levels of our top signals showed significant association with lipid levels ($p<1.0 \times 10^{-6}$) in both sets of cohorts (94 genes for TG, 40 for HDL-C, 0 for LDL-C, and 3 for TC).

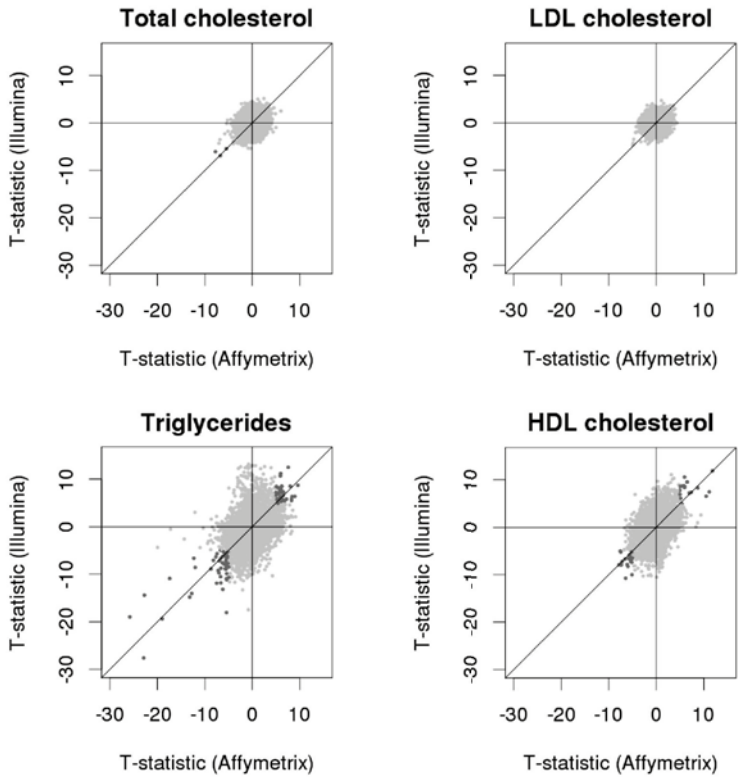


Figure 4. Plots of gene expression associations (t-statistics) with circulating lipid levels in the Affymetrix cohort versus the meta-analysis of Illumina cohorts. Genes with Bonferroni-significant associations ($p<1.0 \times 10^{-6}$) in both sets of results are highlighted in red. All models adjusted for age, age^2, sex, batch effects, and blood cell counts.

Percent of total variation explained

In aggregate, expression levels of the significant genes from our Stage 2 meta-analysis explained 39.2% of the total variation in circulating TG levels, 18.1% for HDL-C, 6.4% for TC, and 1.9% for LDL-C. Expression levels in a few individual genes explained a large percentage of total variation in log-transformed TG levels (Table 2). For example, substantial proportions of variation in TG levels were explained by *HDC* (23.1%), *SLC45A3* (15.2%), *GATA2* (17.3%), *CPA3* (18.1%), *MS4A2* (12.8%), *ABCA1* (8.1%), and *MYLIP* (6.6%). But these estimates may be inflated because the same population was used to select significant genes and to compute the estimates for percent variation explained. As a sensitivity analysis, we computed the percent of total variation explained for the Illumina cohorts using the Bonferroni-significant genes in the Affymetrix cohort (and vice versa). The mean percent of total variation for lipid levels in the Illumina cohorts explained by the significant genes (in aggregate for a given lipid trait) in the Affymetrix set was 40.7% for TG, 16.8% for HDL-C, and 3.3% for TC. No

genes were significant for LDL-C in the meta-analysis of Illumina cohorts. Conversely, the significant expression levels in the Illumina cohorts explained 43.8% for TG, 16.4% for HDL-C, 3.2% for TC, and 2.1% for LDL-C in the Affymetrix cohort. Thus, a large percentage of the total variation in blood lipid levels was explained by whole blood gene expression levels, and these findings were highly consistent across cohorts.

Gene ontology enrichment analysis

At FDR q<0.05, LDL-C was significantly enriched for apolipoprotein receptor activity, driven by *ABCA1* and *ABCA7* associations (S3 Table). TC gene expression results were highly enriched for functions related to DNA and RNA binding (S4 Table). HDL-C and TG gene expression results were enriched for protein binding (including IgE and kinase), poly(A) RNA binding, and structural constituents of ribosomes (S5 and S6 Tables).

Intersection of genetic variant and gene expression findings

Under the assumption that genes identified through both GWAS and gene expression are likely to represent expression levels that influence lipid levels (i.e., upstream effects), we cross-referenced our gene expression results with genetic loci identified by the Global Lipids Genetics Consortium (GLGC) [7,8]. We used three sources of eQTL (genetic variants associated with gene expression traits) in our analysis. The first source of eQTL was from an updated version of a previously described catalog of published eQTL in various tissues [11]. The second source was from gene expression in FHS samples (n=5,257) with 1000 Genomes-imputed genotype data. And the third source was from the Blood eQTL Browser (*http://genenetwork.nl/bloodeqtlbrowser/*) [12]. In total, we identified 5,105 *cis*-eQTL at $p<5.0 \times 10^{-8}$ or Bayes' Factor >30 that included 210 unique genes available in our meta-analysis (S7 Table). Among the *cis*-eQTL, 47 unique genes had p<0.05/210, *DPEP2* expression had the association with the lowest p-value for TG and HDL-C levels (Table 3).

Known lipid pathway proteins

Using the LIPID MAPS proteomic database (accessed on March 26, 2015), we conducted a look-up of known lipid pathway proteins in our meta-analysis results. Among the 1,205 human genes in the LIPID MAPS database, data on 699 genes were available for look-up in our results, of which 95 (13.6%) reached statistical significance at a Bonferroni corrected p<0.05/699 with a lipid trait in our meta-analysis (Table 4).

Table 2. Top 10 genes whose expression levels were associated with circulating lipid levels, based on meta-analysis P-values of Affymetrix and Illumina cohorts.

	Affymetrix cohort (n=1,838)			Illumina cohorts (n=3,003)			Meta-analysis
Gene	**$R^{2\dagger}$**	**t-statistic**	**P-value**	**$R^{2\dagger}$**	**t-statistic**	**P-value**	**P-value**
Total cholesterol							
HDC	0.02	-6.76	1.9E-11	0.02	-6.90	5.3E-12	1.3E-21
CPA3	0.03	-7.79	1.2E-14	0.01	-6.04	1.5E-09	2.7E-21
ABCA1	0.02	-7.04	2.6E-12	0.01	-4.31	1.6E-05	2.0E-14
GATA2	0.01	-5.49	4.6E-08	0.01	-5.41	6.2E-08	2.7E-14
MS4A2	0.02	-6.84	1.1E-11	0.01	-3.53	4.1E-04	5.0E-12
IFIT3	0.01	-3.98	7.1E-05	0.01	-4.41	1.0E-05	3.4E-09
CPT1A	0.02	6.01	2.2E-09	0.00	2.50	1.2E-02	2.2E-08
EPSTI1	0.01	-4.02	6.0E-05	0.00	-3.91	9.4E-05	2.4E-08
ETS1	0.00	2.45	1.5E-02	0.01	5.14	2.7E-07	2.4E-08
AKAP12	0.02	-5.81	7.5E-09	0.01	-2.30	2.1E-02	1.1E-07
LDL cholesterol							
HDC	0.01	-4.96	7.7E-07	0.01	-4.67	3.0E-06	2.1E-11
CPA3	0.01	-4.87	1.2E-06	0.01	-4.27	2.0E-05	2.5E-10
ABCA1	0.01	-4.76	2.1E-06	0.00	-3.45	5.6E-04	2.1E-08
GATA2	0.00	-3.26	1.2E-03	0.01	-3.97	7.1E-05	2.9E-07
E2F2	0.00	3.54	4.2E-04	0.00	3.61	3.1E-04	5.7E-07
GNA12	0.01	3.94	8.4E-05	0.00	3.00	2.7E-03	2.0E-06
PIM1	0.00	2.91	3.7E-03	0.00	3.59	3.3E-04	3.9E-06
STK24	0.00	-1.76	7.8E-02	0.00	-4.43	9.6E-06	4.3E-06
FTL	0.00	-2.33	2.0E-02	0.01	-3.82	1.3E-04	8.4E-06
SLC6A16	0.00	1.08	2.8E-01	0.01	4.73	2.3E-06	9.6E-06
HDL cholesterol							
HDC	0.05	11.81	4.3E-31	0.06	11.90	1.2E-32	5.2E-61
CPA3	0.05	11.16	5.1E-28	0.04	7.46	8.8E-14	3.6E-36
AHSP	0.02	-6.46	1.4E-10	0.03	-10.75	6.2E-27	1.5E-35
MYLIP	0.01	5.96	3.1E-09	0.04	10.58	3.5E-26	2.8E-33
GATA2	0.03	8.74	5.4E-18	0.05	8.28	1.2E-16	3.2E-32
TSC22D3	0.02	6.50	1.0E-10	0.03	9.54	1.4E-21	1.2E-30
MS4A2	0.04	10.49	4.9E-25	0.03	6.54	6.3E-11	2.7E-30
HBD	0.01	-5.14	3.1E-07	0.03	-10.01	1.4E-23	1.7E-28
ALAS2	0.02	-7.86	6.8E-15	0.02	-7.89	3.1E-15	4.4E-28
CEBPD	0.00	3.18	1.5E-03	0.03	11.10	1.3E-28	4.5E-27

Table 2. *(Continued)*

	Affymetrix cohort (n=1,838)			Illumina cohorts (n=3,003)			Meta-analysis
Gene	**$R^{2\dagger}$**	**t-statistic**	**P-value**	**$R^{2\dagger}$**	**t-statistic**	**P-value**	**P-value**
Triglycerides							
HDC	0.20	-22.90	3.1E-102	0.24	-27.62	7.1E-168	5.3E-268
CPA3	0.24	-25.82	1.4E-125	0.17	-18.97	3.1E-80	3.8E-191
GATA2	0.15	-18.99	1.5E-73	0.18	-19.38	1.1E-83	1.6E-153
MS4A2	0.20	-22.75	4.4E-101	0.11	-14.42	3.6E-47	4.4E-131
ABCA1	0.08	-13.22	3.6E-38	0.08	-14.83	9.4E-50	1.1E-85
ENPP3	0.13	-17.41	6.5E-63	0.06	-10.91	1.0E-27	9.7E-79
MYLIP	0.07	-12.78	7.1E-36	0.06	-14.06	6.4E-45	1.7E-78
SLC45A3	0.01	-5.44	5.9E-08	0.19	-18.05	7.9E-73	5.8E-70
AKAP12	0.16	-19.97	2.2E-80	0.01	-4.37	1.2E-05	2.8E-50
AHSP	0.02	7.54	7.4E-14	0.05	12.51	6.7E-36	1.3E-47

** Regression models adjusted for age, age^2, sex, technical covariates, and blood cell counts. † Gene expression-specific explained variance in the levels of the lipid trait (R^2)*

Table 3. Meta-analysis associations between circulating lipid levels and expression levels of genes that were *cis*-eQTL for lipid loci identified by the Global Lipids Genetics Consortium meta-analyses.*

			Total cholesterol		LDL cholesterol		HDL cholesterol		Triglycerides	
Gene	SNP†	eQTL p-value	Z	P-value	Z	P-value	Z	P-value	Z	P-value
DPEP2	rs16942887	1.21E-28	-0.49	6.24E-01	0.05	9.57E-01	**10.28**	**8.38E-25**	**-13.43**	**3.85E-41**
ZDHHC18	rs12748152	3.44E-65	-1.99	4.70E-02	0.47	6.42E-01	**5.21**	**1.84E-07**	**-10.93**	**7.92E-28**
AKT1	rs4983559	7.19E-24	-0.69	4.88E-01	-0.82	4.14E-01	**7.62**	**2.48E-14**	**-10.50**	**8.26E-26**
HDGF	rs12145743	5.32E-08	-0.20	8.38E-01	1.68	9.37E-02	**-8.46**	**2.63E-17**	**9.27**	**1.85E-20**
KLHDC8B	rs2013208	3.02E-07	-3.18	1.47E-03	-2.51	1.22E-02	**4.65**	**3.36E-06**	**-9.11**	**8.34E-20**
NTAN1	rs3198697	8.06E-14	-1.69	9.04E-02	-0.38	7.06E-01	**-9.02**	**1.89E-19**	**8.80**	**1.42E-18**
ASCC2	rs5763662	5.62E-20	1.53	1.25E-01	3.12	1.80E-03	**-7.07**	**1.56E-12**	**8.33**	**7.76E-17**
DARS	rs7570971	1.03E-27	**5.16**	**2.46E-07**	3.64	2.72E-04	-3.62	2.96E-04	**7.97**	**1.58E-15**
MARCH8	rs970548	5.58E-10	3.04	2.37E-03	3.08	2.06E-03	**-5.02**	**5.29E-07**	**7.53**	**4.88E-14**
RAF1	rs2290159	4.19E-11	-0.94	3.47E-01	-1.08	2.81E-01	**6.64**	**3.11E-11**	**-7.46**	**8.61E-14**
ERGIC3	rs2277862	4.41E-08	1.24	2.15E-01	0.90	3.66E-01	**-4.46**	**8.11E-06**	**7.38**	**1.53E-13**
MAP3K11	rs12801636	8.17E-07	-1.77	7.73E-02	-0.72	4.74E-01	2.56	1.06E-02	**-6.73**	**1.68E-11**
GPR146	rs1997243	2.41E-205	0.45	6.53E-01	1.82	6.81E-02	**-6.07**	**1.29E-09**	**6.51**	**7.35E-11**
RAC1	rs702485	9.79E-107	-1.37	1.69E-01	-1.53	1.25E-01	**4.68**	**2.87E-06**	**-6.49**	**8.37E-11**
BCKDK	rs11649653	5.04E-14	0.69	4.91E-01	1.24	2.17E-01	**5.44**	**5.30E-08**	**-6.42**	**1.40E-10**
HIST1H4C	rs1800562	5.36E-38	0.05	9.57E-01	0.26	7.97E-01	**-4.47**	**7.73E-06**	**5.94**	**2.91E-09**
ALOX5	rs970548	1.08E-19	0.40	6.87E-01	0.91	3.64E-01	**5.05**	**4.52E-07**	**-5.65**	**1.59E-08**
MACF1	rs4660293	7.56E-78	**4.35**	**1.33E-05**	2.93	3.40E-03	-2.53	1.14E-02	**5.56**	**2.64E-08**
NRBF2	rs10761731	4.41E-24	-1.31	1.90E-01	0.56	5.75E-01	1.64	1.02E-01	**-5.31**	**1.08E-07**
TOM1	rs138777	1.23E-144	0.99	3.22E-01	0.73	4.63E-01	**4.32**	**1.59E-05**	**-5.24**	**1.61E-07**
MAF1	rs11136341	1.82E-07	-0.98	3.26E-01	1.10	2.73E-01	**-6.60**	**4.24E-11**	**5.09**	**3.52E-07**

Table 3. *(Continued)*

			Total cholesterol		LDL cholesterol		HDL cholesterol		Triglycerides	
Gene	**SNP†**	**eQTL p-value**	**Z**	**P-value**	**Z**	**P-value**	**Z**	**P-value**	**Z**	**P-value**
GRINA	rs11136341	6.49E-243	-1.29	1.96E-01	1.51	1.30E-01	**-7.38**	**1.58E-13**	**5.01**	**5.46E-07**
ATP13A1	rs10401969	3.79E-128	-3.01	2.60E-03	-3.35	8.17E-04	2.64	8.29E-03	**-4.95**	**7.31E-07**
HIST1H4E	rs1800562	1.43E-36	-1.51	1.30E-01	-1.22	2.23E-01	2.40	1.63E-02	**-4.68**	**2.85E-06**
DPEP3	rs16942887	7.50E-21	0.45	6.49E-01	-0.38	7.07E-01	**6.46**	**1.03E-10**	**-4.67**	**2.97E-06**
CTSA	rs6065906	3.56E-09	-2.66	7.81E-03	-1.38	1.66E-01	1.26	2.09E-01	**-4.51**	**6.53E-06**
PGS1	rs4129767	4.72E-34	-0.90	3.70E-01	0.60	5.51E-01	0.31	7.55E-01	**-4.48**	**7.33E-06**
STAG1	rs645040	1.75E-13	1.11	2.69E-01	0.45	6.51E-01	**4.48**	**7.52E-06**	**-4.44**	**9.17E-06**
HIST1H2BF	rs1800562	7.19E-11	-0.51	6.12E-01	0.67	5.02E-01	1.13	2.60E-01	**-4.40**	**1.10E-05**
OASL	rs1169288	9.72E-08	**-4.20**	**2.67E-05**	-2.48	1.32E-02	-2.20	2.79E-02	**-4.36**	**1.29E-05**
BTN2A1	rs1800562	2.31E-05	-0.83	4.08E-01	0.68	4.94E-01	2.21	2.74E-02	**-4.30**	**1.69E-05**
TBKBP1	rs7206971	1.49E-39	-1.20	2.30E-01	-1.43	1.52E-01	2.34	1.95E-02	**-4.23**	**2.37E-05**
ARHGAP1	rs3136441	1.28E-09	2.07	3.83E-02	0.87	3.82E-01	**6.22**	**4.98E-10**	**-4.12**	**3.86E-05**
UBA7	rs2013208	7.03E-149	-2.19	2.88E-02	-2.68	7.36E-03	3.46	5.34E-04	**-4.11**	**4.00E-05**
KLRG1	rs4883201	8.27E-08	2.18	2.91E-02	1.70	8.99E-02	-1.72	8.47E-02	**3.95**	**7.70E-05**
ARL15	rs6450176	6.90E-09	-1.22	2.24E-01	-1.15	2.49E-01	2.41	1.59E-02	**-3.95**	**7.82E-05**
SF3B2	rs12801636	3.96E-05	-0.72	4.69E-01	-2.01	4.49E-02	**5.17**	**2.35E-07**	**-3.94**	**7.99E-05**
MARCH2	rs7255436	2.92E-10	-1.50	1.33E-01	-2.51	1.21E-02	**-3.70**	**2.12E-04**	**3.87**	**1.10E-04**
HIST1H2BG	rs1800562	7.19E-11	1.04	2.99E-01	1.17	2.43E-01	3.07	2.15E-03	**-3.85**	**1.17E-04**
MAPKAPK3	rs2013208	3.88E-08	-0.37	7.14E-01	-0.19	8.51E-01	**4.03**	**5.51E-05**	**-3.84**	**1.21E-04**
FAM117B	rs11694172	1.83E-07	2.83	4.61E-03	1.05	2.93E-01	-0.03	9.74E-01	**3.79**	**1.48E-04**
DGAT2	rs499974	1.45E-12	-0.50	6.19E-01	-0.10	9.19E-01	2.30	2.17E-02	**-3.76**	**1.70E-04**
GATAD2A	rs10401969	7.67E-26	-1.24	2.14E-01	-1.44	1.50E-01	**3.94**	**8.03E-05**	**-3.73**	**1.90E-04**

Table 3. *(Continued)*

			Total cholesterol		LDL cholesterol		HDL cholesterol		Triglycerides	
Gene	**SNP†**	**eQTL p-value**	**Z**	**P-value**	**Z**	**P-value**	**Z**	**P-value**	**Z**	**P-value**
APH1B	rs2652834	1.77E-05	-0.86	3.88E-01	-1.42	1.55E-01	2.27	2.32E-02	**-3.72**	**1.97E-04**
EVI5	rs7515577	8.30E-24	0.60	5.50E-01	-0.41	6.82E-01	**4.66**	**3.19E-06**	**-2.82**	**4.78E-03**
DUSP3	rs8077889	5.62E-17	-0.70	4.85E-01	0.03	9.77E-01	**-3.90**	**9.48E-05**	**2.59**	**9.59E-03**
TTC38	rs4253772	7.46E-11	0.04	9.68E-01	-0.53	5.97E-01	**3.82**	**1.33E-04**	**-2.40**	**1.63E-02**
VIM	rs10904908	3.47E-24	0.29	7.71E-01	-1.55	1.22E-01	**3.69**	**2.27E-04**	**1.10**	**2.70E-01**

** Gene expression levels that were significantly associated with at least 1 lipid trait are shown (p<0.05/210 in bold). † SNPs identified by the Global Lipids Genetics Consortium GWAS meta-analyses* [7,8].

Table 4. Meta-analysis associations between circulating lipid levels and expression levels of known lipid genes from LIPID MAPS proteome database.*

	Total Cholesterol		LDL cholesterol		HDL cholesterol		Triglycerides	
Gene	**Z**	**P-value**	**Z**	**P-value**	**Z**	**P-value**	**Z**	**P-value**
ABCA1	**-7.7**	**2.0E-14**	**-5.6**	**2.1E-08**	**8.3**	**7.3E-17**	**-19.6**	**1.1E-85**
ZDHHC18	-2.0	4.7E-02	0.5	6.4E-01	**5.2**	**1.8E-07**	**-10.9**	**7.9E-28**
ALOX5AP	-2.4	1.7E-02	-2.8	5.6E-03	**8.6**	**8.7E-18**	**-9.7**	**3.9E-22**
PISD	-2.8	5.4E-03	-1.0	3.1E-01	3.9	9.4E-05	**-9.3**	**2.0E-20**
RXRA	-1.2	2.4E-01	-1.0	3.3E-01	**6.7**	**1.5E-11**	**-8.9**	**6.1E-19**
PIK3CD	-1.2	2.4E-01	-1.5	1.4E-01	**7.7**	**1.0E-14**	**-8.8**	**9.8E-19**
GMEB2	-1.0	3.2E-01	-2.1	3.9E-02	**8.0**	**1.7E-15**	**-8.4**	**5.0E-17**
OSBP2	1.8	7.1E-02	2.4	1.6E-02	**-5.3**	**1.3E-07**	**8.4**	**5.8E-17**
SMPDL3A	-2.1	3.9E-02	-2.3	2.0E-02	**4.8**	**1.3E-06**	**-8.2**	**2.8E-16**
DPM2	-0.1	9.0E-01	2.4	1.5E-02	**-9.0**	**2.2E-19**	**8.2**	**3.0E-16**
CXCL16	-1.4	1.5E-01	-1.5	1.4E-01	**7.0**	**2.9E-12**	**-8.2**	**3.1E-16**
SORL1	-0.7	4.6E-01	-0.4	7.0E-01	**7.2**	**8.5E-13**	**-8.2**	**3.6E-16**
PTGS2	-1.3	2.0E-01	1.2	2.4E-01	2.1	3.5E-02	**-7.9**	**2.0E-15**
CARM1	0.5	6.5E-01	2.3	2.1E-02	**-7.1**	**1.0E-12**	**7.8**	**5.4E-15**
EGLN2	-2.7	6.1E-03	-2.7	6.2E-03	**4.9**	**9.3E-07**	**-7.6**	**4.0E-14**
CPT1A	**5.6**	**2.2E-08**	3.5	3.9E-04	-1.9	5.5E-02	**7.1**	**1.5E-12**
CHPT1	-1.0	3.2E-01	0.6	5.3E-01	**-6.7**	**1.6E-11**	**7.0**	**3.3E-12**
SCAP	-0.3	7.9E-01	-0.2	8.1E-01	**6.3**	**2.4E-10**	**-6.9**	**6.7E-12**
TLR4	-0.2	8.3E-01	-0.1	9.4E-01	**5.8**	**8.2E-09**	**-6.8**	**1.2E-11**
PIK3R2	1.8	6.5E-02	1.9	6.0E-02	-3.0	2.8E-03	**6.6**	**4.3E-11**
AGPAT3	0.9	3.9E-01	1.0	3.3E-01	-3.7	2.1E-04	**6.5**	**5.9E-11**
MMD	-0.7	5.0E-01	0.8	4.4E-01	1.9	6.2E-02	**-6.5**	**6.5E-11**
SERINC1	-1.0	3.1E-01	-0.3	7.4E-01	2.5	1.3E-02	**-6.5**	**9.7E-11**
BCKDK	0.7	4.9E-01	1.2	2.2E-01	**5.4**	**5.3E-08**	**-6.4**	**1.4E-10**
HPGDS	-2.7	6.6E-03	-2.3	2.0E-02	2.9	3.7E-03	**-6.4**	**1.7E-10**
ACSL5	2.3	2.4E-02	1.3	1.8E-01	-2.6	9.3E-03	**6.3**	**3.4E-10**
DBI	-1.1	2.7E-01	-0.8	4.5E-01	**-4.9**	**1.1E-06**	**6.0**	**2.0E-09**
ADIPOR1	0.5	6.4E-01	2.5	1.1E-02	**-5.8**	**7.5E-09**	**5.9**	**2.9E-09**
PRKCQ	3.6	3.3E-04	1.1	2.8E-01	0.1	8.9E-01	**5.8**	**7.1E-09**
PIGB	-0.2	8.2E-01	0.4	6.9E-01	3.5	4.9E-04	**-5.7**	**1.2E-08**
ALOX5	0.4	6.9E-01	0.9	3.6E-01	**5.0**	**4.5E-07**	**-5.7**	**1.6E-08**
PTGER2	2.5	1.2E-02	0.7	4.8E-01	-0.5	5.9E-01	**5.6**	**1.9E-08**
PREX1	0.6	5.4E-01	0.2	8.0E-01	**5.2**	**1.6E-07**	**-5.6**	**2.2E-08**
PGRMC1	-2.5	1.2E-02	-0.9	3.5E-01	0.5	6.1E-01	**-5.5**	**3.8E-08**
SDPR	-0.7	4.9E-01	0.6	5.4E-01	1.1	2.5E-01	**-5.4**	**7.9E-08**
ABHD4	-1.5	1.4E-01	-1.2	2.4E-01	3.7	2.5E-04	**-5.4**	**8.1E-08**

Table 4. *(Continued)*

	Total Cholesterol		LDL cholesterol		HDL cholesterol		Triglycerides	
Gene	**Z**	**P-value**	**Z**	**P-value**	**Z**	**P-value**	**Z**	**P-value**
ZDHHC2	0.5	6.3E-01	1.7	8.3E-02	**-5.4**	**8.3E-08**	**5.3**	**1.3E-07**
HSD17B11	-2.6	1.0E-02	-1.7	8.5E-02	2.5	1.4E-02	**-5.3**	**1.3E-07**
ST6GALNAC4	-1.1	2.5E-01	0.9	3.8E-01	**-6.2**	**4.6E-10**	**5.2**	**1.7E-07**
DAD1	-3.3	9.9E-04	-3.0	2.3E-03	2.3	1.9E-02	**-5.2**	**2.1E-07**
SLC27A2	-2.0	4.5E-02	-2.1	3.2E-02	3.4	5.9E-04	**-5.2**	**2.3E-07**
SQLE	-0.5	6.4E-01	-0.5	6.2E-01	**-4.3**	**2.0E-05**	**5.2**	**2.4E-07**
ANXA1	**4.0**	**6.5E-05**	1.0	2.9E-01	0.7	5.0E-01	**5.1**	**3.5E-07**
INPP5A	-0.5	5.9E-01	-1.0	3.3E-01	**5.6**	**2.6E-08**	**-5.1**	**4.3E-07**
PLCG2	0.5	6.0E-01	0.2	8.3E-01	**5.1**	**2.7E-07**	**-5.0**	**6.3E-07**
ACSL3	-1.9	5.9E-02	-1.9	5.7E-02	**4.1**	**5.0E-05**	**-5.0**	**6.5E-07**
DAPP1	-3.5	5.1E-04	-1.9	5.2E-02	-2.6	1.0E-02	**-4.9**	**8.8E-07**
PNPLA6	-1.6	1.2E-01	-1.6	1.2E-01	2.9	4.2E-03	**-4.9**	**1.1E-06**
PITPNM1	-1.0	3.3E-01	0.0	9.9E-01	1.1	2.9E-01	**-4.8**	**1.5E-06**
SUCLG1	1.8	7.4E-02	0.8	4.5E-01	-1.8	7.3E-02	**4.8**	**1.9E-06**
ACOX1	-0.5	6.2E-01	0.8	4.4E-01	3.9	1.1E-04	**-4.8**	**1.9E-06**
PTGDS	0.0	9.7E-01	-1.0	3.3E-01	**5.2**	**1.8E-07**	**-4.7**	**3.2E-06**
ANXA11	-1.1	2.8E-01	-1.2	2.1E-01	**4.4**	**1.0E-05**	**-4.6**	**4.3E-06**
GDE1	0.5	6.3E-01	1.7	8.1E-02	**-4.4**	**1.0E-05**	**4.6**	**4.5E-06**
ALDH9A1	2.2	3.0E-02	1.6	1.1E-01	**4.7**	**2.8E-06**	**-4.6**	**4.8E-06**
CDS2	-1.5	1.3E-01	-0.6	5.4E-01	1.9	6.3E-02	**-4.5**	**5.7E-06**
LMF2	-2.2	3.1E-02	-2.3	2.4E-02	2.6	9.4E-03	**-4.5**	**6.2E-06**
PIGX	0.0	9.9E-01	0.3	7.7E-01	3.3	9.4E-04	**-4.5**	**6.6E-06**
ACAT1	0.4	6.8E-01	-0.8	4.4E-01	-2.3	2.0E-02	**4.5**	**7.1E-06**
PIP5K1C	-0.4	6.7E-01	-0.6	5.3E-01	3.5	4.5E-04	**-4.5**	**7.2E-06**
PGS1	-0.9	3.7E-01	0.6	5.5E-01	0.3	7.5E-01	**-4.5**	**7.3E-06**
IMPA2	-0.6	5.4E-01	-0.4	7.0E-01	**4.8**	**1.8E-06**	**-4.5**	**8.5E-06**
RCHY1	-1.3	1.8E-01	-0.6	5.2E-01	2.4	1.6E-02	**-4.4**	**8.8E-06**
FDFT1	-1.2	2.4E-01	0.7	5.0E-01	**-5.2**	**2.4E-07**	**4.4**	**1.1E-05**
PICALM	-0.7	5.0E-01	0.9	3.9E-01	2.7	7.3E-03	**-4.4**	**1.1E-05**
SEC14L4	2.1	3.5E-02	2.7	6.5E-03	-3.2	1.5E-03	**4.4**	**1.3E-05**
NUDT4	**4.5**	**7.1E-06**	3.5	5.4E-04	1.5	1.4E-01	**4.4**	**1.3E-05**
GPX1	-1.9	6.2E-02	0.3	8.0E-01	**-5.5**	**3.8E-08**	**4.3**	**2.0E-05**
GK	-2.1	3.7E-02	1.9	5.7E-02	-0.8	4.1E-01	**-4.3**	**2.0E-05**
EBPL	-0.2	8.2E-01	0.4	7.1E-01	**-4.1**	**4.6E-05**	**4.2**	**2.4E-05**
TMEM55A	0.2	8.4E-01	0.0	9.7E-01	3.9	8.1E-05	**-4.2**	**3.3E-05**
SLC25A29	0.9	3.9E-01	1.1	2.6E-01	2.0	4.2E-02	**-4.1**	**3.9E-05**
PAQR6	-1.8	7.6E-02	-1.4	1.6E-01	2.1	3.7E-02	**-4.1**	**4.4E-05**

Table 4. *(Continued)*

	Total Cholesterol		LDL cholesterol		HDL cholesterol		Triglycerides	
Gene	**Z**	**P-value**	**Z**	**P-value**	**Z**	**P-value**	**Z**	**P-value**
LRP10	-0.5	6.1E-01	1.3	2.0E-01	1.3	2.1E-01	**-4.1**	**4.5E-05**
ACAA1	-3.6	3.7E-04	-2.3	2.1E-02	0.5	6.1E-01	**-4.0**	**5.4E-05**
FAAH	-0.1	9.0E-01	-0.6	5.2E-01	3.8	1.2E-04	**-4.0**	**5.8E-05**
CYB5R3	1.8	7.3E-02	2.9	3.9E-03	**-4.5**	**6.7E-06**	**3.8**	**1.5E-04**
OSBPL9	0.6	5.7E-01	-0.6	5.6E-01	**4.0**	**5.8E-05**	**-3.8**	**1.7E-04**
PLCB2	0.5	6.0E-01	-0.4	6.8E-01	**5.0**	**4.9E-07**	**-3.7**	**2.6E-04**
PTGES	0.1	9.2E-01	-0.6	5.8E-01	**4.0**	**6.0E-05**	**-3.6**	**2.8E-04**
RBM14	1.3	2.0E-01	-1.1	2.7E-01	**4.5**	**8.2E-06**	**-3.6**	**2.9E-04**
PIGN	2.4	1.6E-02	1.4	1.6E-01	**4.3**	**2.1E-05**	**-3.6**	**3.1E-04**
SREBF2	-0.2	8.7E-01	0.6	5.3E-01	**-4.3**	**2.1E-05**	**3.6**	**3.3E-04**
DNAJA1	-1.3	2.0E-01	0.0	9.8E-01	**-5.5**	**4.3E-08**	**3.5**	**5.0E-04**
PNPLA1	-0.6	5.8E-01	-1.1	2.5E-01	**4.6**	**4.2E-06**	**-3.5**	**5.2E-04**
STT3A	1.4	1.6E-01	0.2	8.6E-01	**4.7**	**2.9E-06**	**-3.0**	**2.4E-03**
AGPAT9	-0.2	8.8E-01	-1.0	3.3E-01	**4.7**	**2.5E-06**	**-2.8**	**6.0E-03**
PIP5K1B	-1.7	9.7E-02	1.8	7.1E-02	**-4.4**	**9.2E-06**	**2.6**	**9.4E-03**
HSD17B12	3.2	1.6E-03	1.0	3.2E-01	**4.9**	**8.5E-07**	**-2.4**	**1.5E-02**
PSAP	1.4	1.5E-01	-1.3	1.9E-01	**6.5**	**7.4E-11**	**-2.3**	**2.2E-02**
CYP27A1	-0.2	8.5E-01	-1.2	2.5E-01	**4.2**	**3.1E-05**	**-2.3**	**2.3E-02**
ECHDC3	0.6	5.7E-01	-0.6	5.4E-01	**4.6**	**3.9E-06**	**-2.1**	**3.5E-02**
LDLR	**-4.4**	**1.1E-05**	-3.7	2.3E-04	**-3.6**	**3.0E-04**	**2.1**	**3.8E-02**
ADIPOR2	1.3	2.1E-01	0.3	7.9E-01	**4.1**	**3.8E-05**	**-2.0**	**4.8E-02**
ALDH1A1	0.9	3.6E-01	-3.0	2.5E-03	**6.5**	**7.8E-11**	**-1.1**	**2.7E-01**

** Genes significantly associated with at least 1 lipid trait are shown (P<0.05/699 in bold).*

DISCUSSION

Meta-analyzing data on 4,841 participants from five well-characterized population-based studies, we identified 906 genes whose expression levels were associated with circulating levels of at least one lipid trait, including 793 genes for TG, 489 for HDL-C, 20 for TC, and five for LDL-C. When the Affymetrix and Illumina cohorts were analyzed separately, we identified 105 genes whose expression levels were associated with circulating lipid levels in both data sets, suggesting highly reproducible results. Furthermore, we identified 13 additional genes reported in GWAS of lipids whose expression levels were also significantly associated with lipid levels, reflecting a convergence of signals from genetic variation and gene expression.

Our strongest associations included a set of co-expressed genes related to basophils and mast cells. The strongest associations within this gene set were for *HDC* and *CPA3*, which were previously

observed in a smaller study using co-expression network analysis [13]. *HDC* codes for histidine decarboxylase, a rate-limiting enzyme catalyzing the conversion of L-histidine to histamine, a pro-inflammatory molecule secreted by mast cells, basophils, and macrophages. *HDC* expression was associated with levels of all four lipid traits. Gonen *et al.* demonstrated that incubating basophils or mast cells with low density lipoproteins (LDL) or very low density lipoproteins (VLDL) altered histamine release [14,15]. In addition, LDL and VLDL binding sites have been identified on human basophils and mast cells, providing further evidence of a biological interaction [16]. Of note, when we conducted analysis in FHS participants who were taking lipid-lowering medications (n=1,550), the association between *HDC* expression levels and LDL-C levels were not significant (p=0.97). *HDC* gene expression, however, remained strongly associated with TG, HDL-C, and TC levels. *HDC* knockout mice lose body weight in the face of a high fat diet, and develop hepatic steatosis [17]. *ApoE* and *Hdc* double knockout mice have reduced atherosclerotic areas and reduced expression of inflammatory genes despite having elevated serum cholesterol levels compared to *ApoE* single knockout mice [18]. These findings suggest that some of our observed associations may reflect downstream effects of lipid levels on gene expression, possibly through alternate or unrecognized pathways. A causal role, however, as demonstrated by knockout experiments, may also be invoked as part of a complex regulatory network. These complex relationships may partially explain the large proportion of total variation in lipid levels explained by our top associations.

CPA3 codes for carboxypeptidase A3, a metalloexopeptidase found in the secretory granules of basophils and mast cells. These granules contain enzymes that digest LDL-C particles. Specifically, carboxypeptidase A degrades apolipoprotein B and A1, the primary apolipoproteins of LDL-C and HDL-C, respectively [19-23]. In turn, this carboxypeptidase-mediated process appears to increase LDL-C uptake by macrophages and, in turn, contributing to foam cell formation [24,25]. The negative correlation between *CPA3* expression and LDL-C levels the we observed was consistent with the known biology. It remains unclear, however, whether, or how, carboxypeptidase A is involved in the regulation of other lipids.

Our findings are further supported by a large body of literature on the role of mast cells in the development of atherosclerosis. Mast cells accumulate in atherosclerotic lesions [26] and at the sites of intra-plaque hemorrhages [20,27], where they are thought to release histamine and matrix-degrading proteases that lead to fragility of microvessels that eventually leads to hemorrhage [28]. Further, mast cells may recruit leukocytes to the plaque sites, increase lipid uptake by macrophages through heparin-bound LDL particles, and proteolyse high density lipoproteins, thereby leading to foam cell formation [22,29,30]. Our study may be limited to the extent that our findings were driven by tissue-specific gene expression. Still, our findings highlight the role of basophils and mast cells as potential mediators between circulating lipid levels and atherosclerosis. We minimized confounding due to differences in blood cell abundances by adjusting for complete blood counts.

Several other limitations should be considered when interpreting our results. First, our results may not necessarily reflect gene expression in liver, the primary organ involved in the regulation

of circulating lipid levels. This may explain the lack of associations for several known lipid genes, whose expression may be tissue-specific [31-33]. Moreover, our study interrogated gene expression levels, but genetic control of lipid levels may be due, at least in part, to functional changes in genes or their protein products, as was shown to be the case with *APOE* [34]. These functional changes notwithstanding, we attempted to separate genes that may be involved in the regulation of lipid levels from those that are influenced by changes in lipids levels by examining the intersection of gene expression with GWAS results. Second, excluding participants on lipid-lowering medications may have limited our range of lipid levels, which in turn may limit the generalizability of our findings. We excluded treated individuals because of the likelihood that lipid medications alter the expression of many genes. Third, the overall weak associations with LDL-C may have been due to imprecision in the estimation of LDL-C levels through the use of the Friedewald equation in three of our cohorts. Lastly, due to the cross-sectional nature of our study, the large percentage of variation in TG and HDL-C levels explained by individual gene expression levels may be partly due to genes whose expression levels are influenced by circulating lipids rather than the other way around. One must keep in mind that this statistical measure also does not imply causality. Nonetheless, genes whose expression levels are altered by lipids are still important because they may represent mediators that link lipid levels to their associated complex diseases, as we believe to be the case for *HDC* and *CPA3* genes.

Our study identified 906 genes whose expression levels in whole blood were associated with circulating lipid levels. The expression levels of these genes, in aggregate, explained a large proportion of the total variation in the lipid phenotypes – far greater than what has been reported in GWAS. By combining genetic and gene expression associations, we identified several known lipid regulatory genes, such as *ABCA1* and *MYLIP* (i.e., upstream regulators), as well as genes that not only responded to changes in lipid levels (i.e., downstream effects) but may play a role in the development of atherosclerosis. Our study serves as a proof-of-principle that gene expression association in whole blood may be useful in identifying important genes involved in both the regulation and downstream effects of circulating lipid levels.

MATERIALS AND METHODS

Study populations

The Framingham Heart Study (FHS), a community-based, prospective study of CVD and its risk factors, began enrollment of the Offspring cohort in 1971 (n=5,124) [35,36]. The Third Generation cohort of 4,095 men and women enrolled individuals from 2002-2005 [37]. Participants underwent in-person evaluations every four to eight years. The current investigation was limited to the 2,446 FHS Offspring cohort attendees at the 8th examination cycle (2005-2008) and 3,180 Third Generation attendees at the 2nd examination cycle (2008-2011) who provided fasting plasma samples for assessment of lipid levels and whole blood for RNA isolation and who consented to have their data used for genomic research. Participants receiving lipid-lowering drug therapy at the time of RNA

collection were excluded because their expression levels may have been altered by medication use. This exclusion resulted in 3,978 FHS participants available for further analysis. Of these 3,978 participants, complete blood counts were available in 1,838 participants.

The Study of Health in Pomerania (SHIP-TREND) is a longitudinal population-based cohort study in West Pomerania, a region in the northeast of Germany, assessing the prevalence and incidence of common population-relevant diseases and their risk factors. Baseline examinations for SHIP-TREND were carried out between 2008 and 2012, comprising 4,420 participants aged 20 to 81 years. Study design and sampling methods were previously described [38]. Analyses for the present project were based on a subset of 917 fasting individuals from the SHIP-TREND study population who were not taking lipid lowering medications and had gene expression data available.

KORA F4 (Cooperative Health Research in the Region of Augsburg) is a population-based survey of the KORA project in the region of Augsburg in Southern Germany [39]. The F4 cohort consists of 3,080 individuals of German nationality. For the expression analysis we included a random subset of participants aged 61 to 82 years in F4 (2006-2008) [40,41]. After quality control and exclusion of individuals taking lipid lowering medications 750 remained for analysis.

The Rotterdam Study (RS) is a large prospective, population-based cohort study in the Ommoord district of Rotterdam, the Netherlands, investigating the prevalence, incidence, and risk factors of various chronic disabling diseases among Caucasians aged 45 years and over, as described elsewhere . The initial cohort, named Rotterdam Study I (RS-I), started in 1989, and consisted of 7,983 persons aged 55 years or over, living in the well-defined Ommoord district. In 1999, a second cohort, named Rotterdam Study II (RS-II), was started and consisted of 3,011 participants who had reached 55 years of age or moved into the study district [42]. In 2006, a further extension of the cohort was initiated in which 3,932 subjects were included, aged 45 years or over, called Rotterdam Study III (RS-III) [42].

The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC and by the Ministry of Health, Welfare and Sport of the Netherlands, implementing the Wet Bevolkingsonderzoek: ERGO (Population Studies Act: Rotterdam Study). All participants provided written informed consent to participate in the study and to obtain information from their treating physicians. Gene expression levels in whole blood were assessed in 881 subjects of RS-III, for whom serum lipid levels were also measured. Detailed methods describing the gene expression assay were described elsewhere [43].

The Invecchiare in Chianti (InCHIANTI) Study is a population-based, epidemiological study of risk factors contributing to the decline in physical functioning in late life [44]. Individuals were selected from the population registries of two small towns in Tuscany, Italy. Whole blood gene expression and fasting plasma lipid levels were available in 698 participants (two participants had missing LDL-C measures). We further excluded 90 (13%) participants who reported taking lipid-altering

medications at the time of blood draw. Thus, 606 participants remained for further analysis. Detailed methods describing the gene expression assay were described elsewhere [45].

Lipid quantification

Fasting levels of TC, HDL-C, and TG were assayed in plasma or serum samples using automated chemistry analyzers (see details in Supporting Information). Twelve individuals in Rotterdam (1.6%) and 7 individuals in KORA (1.0%) were not fasting at the time of blood draw; thus, all statistical models for those cohorts were additionally adjusted for fasting status. LDL-C was directly measured in the KORA and SHIP-TREND samples and calculated using the Friedewald equation for all other cohorts [46].

Gene expression assays

Whole blood samples were collected in PAXgene tubes (details in Supporting Materials) at the same clinic visit as the blood samples for lipid assays. RNA was extracted using the PAXgene Blood mRNA kit (Qiagen). RNA amplification was conducted using either the Ambion TotalPrep RNA Amp kit (InChianti, SHIP-TREND, KORA, and RS-III), or the NuGEN WT-Ovation Pico RNA Amplification System (FHS). FHS samples were assayed using the Affymetrix GeneChip Human Exon 1.0 ST microarray (Affymetrix, Inc., Santa Clara, California, USA). RS-III samples were assayed using the Illumina HumanHT-12 v4 (Illumina, San Diego, California, USA). KORA, SHIP-TREND, and InChianti samples were assayed using Illumina HumanHT-12 v3. Probes from the Affymetrix and Illumina arrays were mapped to RefSeq genes by matching mRNA sequences. Additional details of the mRNA processing and probe matching are described in the Supplemental Materials. The resulting Affymetrix and Illumina probe matches used for the current analysis are listed in S8 Table. For genes with multiple sets of Affymetrix and Illumina probes that matched, we selected the pair with the lowest p-value for association with each lipid in the meta-analysis.

Intersection of GWAS and gene expression findings

Three sources of *cis*-eQTL were used to link our gene expression findings to prior lipid GWAS findings. First, we used an updated version of a previously described *cis*-expression quantitative trait loci (eQTL) compilation across multiple tissues (see Supplemental Materials for details) [11]. Second, we queried *cis*-eQTL in the Blood eQTL Browser [12]. Lastly, we used the expanded FHS gene expression data (17,873 gene sets) linked to Affymetrix 500K mapping array and 50K supplemental array data. We focused on the 157 genetic loci identified to be associated with lipid levels in the Global Lipids Genetics Consortium [7,8]. After identifying *cis*-eQTL (±1 megabase upstream or downstream) with $p<5.0 \times 10^{-8}$ for these loci, we then conducted look-ups of the expression levels of those eQTL genes in relation to lipid levels in our gene expression meta-analysis.

Statistical analysis

Normalized, log-transformed gene expression levels were modeled as the dependent variable and lipid traits as the independent variable. TG levels were natural log-transformed. To account for familial relatedness among FHS participants, we used pedigree-based mixed-effect models in

the *kinship* package in R. The remaining cohorts used multivariable linear regression models. All models were adjusted for age, age^2, sex, technical covariates, and blood cell counts (see details in Supplemental Materials). Additional models further adjusted for body mass index (BMI). To account for multiple comparisons in the primary meta-analysis, we used an experiment-wide Bonferroni correction for 12,492 genes across four traits, corresponding to a nominal $p<1.0 \times 10^{-6}$. Because the arrays from Illumina cohorts were processed similarly giving effect estimates on the same scale, an inverse variance weighted random effects meta-analysis of 37,348 common probes between the HumanHT-12 v3 and v4 arrays was performed first. Meta-analysis of Affymetrix and Illumina array gene expression estimates was conducted using the sample size weighted Z-score method by combining probes that mapped to the same mRNA on both array systems (see details in Supplemental Materials). To estimate the percent of total variation in lipid levels explained by our top gene expression signals, mean partial R^2 was computed by taking the difference in model mean squared errors of the full and restricted models divided by the total variance in the lipid trait. The full model included the gene expression level(s) and other covariates (age, age^2, sex, blood counts) as the independent variables. The restricted models were identical to the full models except the gene expression level(s) was omitted from the model. The gene ontology enrichment analysis was conducted by submitting a single ranked list of genes based on meta-analysis p-values of the gene expression results for each lipid trait to the *GOrilla* web-based software (http://cbl-gorilla.cs.technion.ac.il/, accessed on March 18, 2015) [47].

ACKNOWLEDGMENTS

Rotterdam Study

We thank all study participants and staff from the Rotterdam Study, the participating general practitioners and the pharmacists. This work was funded by the European Commission (HEALTH-F2-2008-201865, GEFOS; HEALTH-F2-2008 35627, TREAT-OA 200800), the Netherlands Organisation of Scientific Research NWO Investments (numbers 175.010.2005.011, 911-03-012), the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Netherlands Genomics Initiative (NGI)/ Netherlands Consortium for Healthy Aging (NCHA) (project nr. 050-060-810), and NWO Vidi grant (#917103521). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organisation for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam.

Framingham Heart Study

The Framingham Heart Study is funded by National Institutes of Health contract N01-HC-25195. The laboratory work for this investigation was funded by the Division of Intramural Research, National Heart, Lung, and Blood Institute, National Institutes of Health. The analytical component of this project was funded by the Division of Intramural Research, National Heart, Lung, and Blood

Institute, and the Center for Information Technology, National Institutes of Health, Bethesda, MD. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster (*http://biowulf.nih.gov*) and Helix Systems (*http://helix.nih.gov*) at the National Institutes of Health, Bethesda, MD, USA.

SHIP-TREND

SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network 'Greifswald Approach to Individualized Medicine (GANI_MED)' funded by the Federal Ministry of Education and Research (grant 03IS2061A). Genome-wide data have been supported by the Federal Ministry of Education and Research (grant no. 03ZIK012) and a joint grant from Siemens Healthcare, Erlangen, Germany and the Federal State of Mecklenburg, West Pomerania. The University of Greifswald is a member of the 'Center of Knowledge Interchange' program of the Siemens AG and the Caché Campus program of the InterSystems GmbH.

KORA

This work was supported by the Ministry of Science and Research of the State of North Rhine-Westphalia (MIWF NRW) and the German Federal Ministry of Health (BMG). This study was supported in part by a grant from the German Federal Ministry of Education and Research (BMBF) to the German Center for Diabetes Research (DZD e.V.). The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements n° 261433 (BioSHaRE-EU) and n°603288 (SysVasc).

ADDITIONAL INFORMATION

Supplementary Information is available on request.

REFERENCES

1. Kannel WB, Castelli WP, Gordon T, McNamara PM (1971) Serum cholesterol, lipoproteins, and the risk of coronary heart disease. The Framingham study. Annals of internal medicine 74: 1-12.
2. Shepherd J, Cobbe SM, Ford I, Isles CG, Lorimer AR, et al. (1995) Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. West of Scotland Coronary Prevention Study Group. The New England journal of medicine 333: 1301-1307.
3. Downs JR, Clearfield M, Weis S, Whitney E, Shapiro DR, et al. (1998) Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: results of AFCAPS/TexCAPS. Air Force/Texas Coronary Atherosclerosis Prevention Study. JAMA : the journal of the American Medical Association 279: 1615-1622.
4. Sever PS, Dahlof B, Poulter NR, Wedel H, Beevers G, et al. (2003) Prevention of coronary and stroke events with atorvastatin in hypertensive patients who have average or lower-than-average cholesterol concentrations, in the Anglo-Scandinavian Cardiac Outcomes Trial--Lipid Lowering Arm (ASCOT-LLA): a multicentre randomised controlled trial. Lancet 361: 1149-1158.
5. Robinson JG, Farnier M, Krempf M, Bergeron J, Luc G, et al. (2015) Efficacy and Safety of Alirocumab in Reducing Lipids and Cardiovascular Events. N Engl J Med.
6. Sabatine MS, Giugliano RP, Wiviott SD, Raal FJ, Blom DJ, et al. (2015) Efficacy and Safety of Evolocumab in Reducing Lipids and Cardiovascular Events. N Engl J Med.
7. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466: 707-713.
8. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, et al. (2013) Discovery and refinement of loci associated with lipid levels. Nature genetics 45: 1274-1283.
9. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nature reviews Genetics 11: 446-450.
10. Yaqoob P (2003) Lipids and the immune response: from molecular mechanisms to clinical applications. Curr Opin Clin Nutr Metab Care 6: 133-150.
11. Zhang X, Gierman HJ, Levy D, Plump A, Dobrin R, et al. (2014) Synthesis of 53 tissue and cell line expression QTL datasets reveals master eQTLs. BMC Genomics 15: 532.
12. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, et al. (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature genetics 45: 1238-1243.
13. Inouye M, Silander K, Hamalainen E, Salomaa V, Harald K, et al. (2010) An immune response network associated with blood lipid levels. PLoS genetics 6: e1001113.
14. Gonen B, O'Donnell P, Post TJ, Quinn TJ, Schulman ES (1987) Very low density lipoproteins (VLDL) trigger the release of histamine from human basophils. Biochimica et biophysica acta 917: 418-424.
15. Schulman ES, Quinn TJ, Post TJ, O'Donnell P, Rodriguez A, et al. (1987) Low density lipoprotein (LDL) inhibits histamine release from human mast cells. Biochemical and biophysical research communications 148: 553-559.
16. Virgolini I, Li SR, Yang Q, Koller E, Sperr WR, et al. (1995) Characterization of LDL and VLDL binding sites on human basophils and mast cells. Arteriosclerosis, thrombosis, and vascular biology 15: 17-26.
17. Ohtsu H, Tanaka S, Terui T, Hori Y, Makabe-Kobayashi Y, et al. (2001) Mice lacking histidine decarboxylase exhibit abnormal mast cells. FEBS Lett 502: 53-56.
18. Wang KY, Tanimoto A, Guo X, Yamada S, Shimajiri S, et al. (2011) Histamine deficiency decreases atherosclerosis and inflammatory response in apolipoprotein E knockout mice independently of serum cholesterol level. Arterioscler Thromb Vasc Biol 31: 800-807.
19. Kokkonen JO, Vartiainen M, Kovanen PT (1986) Low density lipoprotein degradation by secretory granules of rat mast cells. Sequential degradation of apolipoprotein B by granule chymase and carboxypeptidase A. The Journal of biological chemistry 261: 16067-16072.
20. Pejler G, Knight SD, Henningsson F, Wernersson S (2009) Novel insights into the biological function of mast cell carboxypeptidase A. Trends in immunology 30: 401-408.

21. Paananen K, Kovanen PT (1994) Proteolysis and fusion of low density lipoprotein particles independently strengthen their binding to exocytosed mast cell granules. The Journal of biological chemistry 269: 2023-2031.
22. Kokkonen JO, Kovanen PT (1985) Low density lipoprotein degradation by rat mast cells. Demonstration of extracellular proteolysis caused by mast cell granules. The Journal of biological chemistry 260: 14756-14763.
23. Usami Y, Kobayashi Y, Kameda T, Miyazaki A, Matsuda K, et al. (2013) Identification of sites in apolipoprotein A-I susceptible to chymase and carboxypeptidase A digestion. Biosci Rep 33: 49-56.
24. Kovanen PT (1991) Mast cell granule-mediated uptake of low density lipoproteins by macrophages: a novel carrier mechanism leading to the formation of foam cells. Annals of medicine 23: 551-559.
25. Wang Y, Lindstedt KA, Kovanen PT (1995) Mast cell granule remnants carry LDL into smooth muscle cells of the synthetic phenotype and induce their conversion into foam cells. Arteriosclerosis, thrombosis, and vascular biology 15: 801-810.
26. Cairns A, Constantinides P (1954) Mast cells in human atherosclerosis. Science 120: 31-32.
27. Kaartinen M, Penttila A, Kovanen PT (1994) Accumulation of activated mast cells in the shoulder region of human coronary atheroma, the predilection site of atheromatous rupture. Circulation 90: 1669-1678.
28. Bot I, Shi GP, Kovanen PT (2015) Mast cells as effectors in atherosclerosis. Arterioscler Thromb Vasc Biol 35: 265-271.
29. Kokkonen JO, Kovanen PT (1987) Stimulation of mast cells leads to cholesterol accumulation in macrophages in vitro by a mast cell granule-mediated uptake of low density lipoprotein. Proc Natl Acad Sci U S A 84: 2287-2291.
30. Lee M, Lindstedt LK, Kovanen PT (1992) Mast cell-mediated inhibition of reverse cholesterol transport. Arterioscler Thromb 12: 1329-1335.
31. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, et al. (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 466: 714-719.
32. Jiang XC, Moulin P, Quinet E, Goldberg IJ, Yacoub LK, et al. (1991) Mammalian adipose tissue and muscle are major sources of lipid transfer protein mRNA. J Biol Chem 266: 4631-4639.
33. Abifadel M, Varret M, Rabes JP, Allard D, Ouguerram K, et al. (2003) Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nat Genet 34: 154-156.
34. Weisgraber KH, Innerarity TL, Mahley RW (1982) Abnormal lipoprotein receptor-binding activity of the human E apoprotein due to cysteine-arginine interchange at a single site. J Biol Chem 257: 2518-2521.
35. Feinleib M, Kannel WB, Garrison RJ, McNamara PM, Castelli WP (1975) The Framingham Offspring Study. Design and preliminary data. Preventive medicine 4: 518-525.
36. Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP (1979) An investigation of coronary heart disease in families. The Framingham offspring study. American journal of epidemiology 110: 281-290.
37. Splansky GL, Corey D, Yang Q, Atwood LD, Cupples LA, et al. (2007) The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. American journal of epidemiology 165: 1328-1335.
38. Volzke H, Alte D, Schmidt CO, Radke D, Lorbeer R, et al. (2011) Cohort profile: the study of health in Pomerania. International journal of epidemiology 40: 294-307.
39. Holle R, Happich M, Lowel H, Wichmann HE (2005) KORA--a research platform for population based health research. Gesundheitswesen 67 Suppl 1: S19-25.
40. Rathmann W, Strassburger K, Heier M, Holle R, Thorand B, et al. (2009) Incidence of Type 2 diabetes in the elderly German population and the effect of clinical and lifestyle risk factors: KORA S4/F4 cohort study. Diabetic medicine : a journal of the British Diabetic Association 26: 1212-1219.
41. Schramm K, Marzi C, Schurmann C, Carstensen M, Reinmaa E, et al. (2014) Mapping the genetic architecture of gene regulation in whole blood. PLoS One 9: e93844.
42. Hofman A, Darwish Murad S, van Duijn CM, Franco OH, Goedegebure A, et al. (2013) The Rotterdam Study: 2014 objectives and design update. Eur J Epidemiol 28: 889-926.
43. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, et al. (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet 45: 1238-1243.
44. Ferrucci L, Bandinelli S, Benvenuti E, Di Iorio A, Macchi C, et al. (2000) Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. Journal of the American Geriatrics Society 48: 1618-1625.

45. Harries LW, Bradley-Smith RM, Llewellyn DJ, Pilling LC, Fellows A, et al. (2012) Leukocyte CCR2 Expression Is Associated with Mini-Mental State Examination Score in Older Adults. Rejuvenation research 15: 395-404.

46. Friedewald WT, Levy RI, Fredrickson DS (1972) Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clinical chemistry 18: 499-502.

47. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10: 48.

CHAPTER 2.4

Gene transcripts associated with muscle strength: a CHARGE meta-analysis of 7,781 persons

Luke C. Pilling*, Roby Joehanes*, Tim Kacprowski*, Marjolein J. Peters*, Rick Jansen*, David Karasik, Douglas P. Kiel, Lorna W. Harries, Alexander Teumer, Joseph E. Powell, Daniel Levy, Honghuang Lin, Kathryn L. Lunetta, Peter Munson, Stefania Bandinelli, William Henley, Dena G. Hernandez, Andrew Singleton, Toshiko Tanaka, Gerard van Grootheest, Albert Hofman, André G. Uitterlinden, Reiner Biffar, Sven Gläser, Georg Homuth, Carolin Malsch, Uwe Völker, Brenda W.J.H. Penninx*, Joyce B.J. van Meurs*, Luigi Ferrucci*, Thomas Kocher*, Joanne M. Murabito*, David Melzer*

** These authors contributed equally to this work*

Published: *Physiol Genomics 2015; 00054.2015. doi: 10.1152/physiolgenomics.00054.2015*

ABSTRACT

Background: Lower muscle strength in midlife predicts disability and mortality in later life. Blood-borne factors, including growth differentiation factor 11 (*GDF11*), have been linked to muscle regeneration in animal models. We aimed to identify gene transcripts associated with muscle strength in adults.

Methods: Meta-analysis of whole blood gene expression (overall 17,534 unique genes measured by microarray) and hand-grip strength in four independent cohorts (n=7,781, ages: 20-104 years, weighted mean=56), adjusted for age, sex, height, weight, and leukocyte subtypes. Separate analyses were performed in subsets (older/younger than 60, male/female).

Results: Expression levels of 221 genes were associated with strength after adjustment for cofactors and for multiple statistical testing, including *ALAS2* (rate limiting enzyme in heme synthesis), PRF1 (perforin, a cytotoxic protein associated with inflammation), *IGF1R* and *IGF2BP2* (both insulin like growth factor related). We identified statistical enrichment for hemoglobin biosynthesis, innate immune activation and the stress response. Ten genes were only associated in younger individuals, four in males only and one in females only. For example *PIK3R2* (a negative regulator of *PI3K/AKT* growth pathway) was negatively associated with muscle strength in younger (<60 years) individuals but not older (>=60 years). We also show that 115 genes (52%) have not previously been linked to muscle in NCBI PubMed abstracts.

Conclusions: This first large-scale transcriptome study of muscle strength in human adults confirmed associations with known pathways and provides new evidence for over half of the genes identified. There may be age and sex specific gene expression signatures in blood for muscle strength.

INTRODUCTION

Muscle strength correlates with health and physical function, and poor muscle strength in midlife is a strong, independent predictor of health status decline and mortality over 25 years [1]. Sufficient muscle strength in the hands, arms and legs is needed for everyday functioning; persons with poor strength are at high risk of disability, injury from falls, and other age-related morbidities [2,3].

Hand-grip is a frequently used summary measure of strength because it correlates well with strength of other key muscles and is relatively easy to measure with high precision; Bohannon *et al.* reported strong correlations between grip and knee extension strength (Pearson R=0.77 to 0.8) in a sample aged 18 to 85 years [4], with Samson *et al.* reporting similar estimates [5]. Muscle strength (including grip strength) is a more important predictor of mortality risk than muscle mass [6,7], and grip strength (but not muscle mass) was associated with poor physical functioning in older adults [8]. The mechanisms underlying the association between lower strength and mortality are not entirely clear, but a recent large-scale multi-country follow-up study (n=142,861) reported that lower grip strength associated most strongly with cardiovascular mortality [9]. Current theories emphasize the role of denervation not compensated by adequate re-innervation, mitochondrial dysfunction, cellular senescence, inflammation, changes in microenvironment, and local skeletal changes, among other factors [10,11].

Studies of heterochronic parabiosis (connecting the blood circulations of young and old mice) found that circulating factors, in particular lower *GDF11* (growth and differentiation factor 11) in older mice, explained the lower muscle regenerative capacity in older compared to younger muscle [12-14]. It is well established that circulating factors, such as pro-inflammatory mediators and hormones (including testosterone), are strong correlates which predict the slope of decline of muscle mass and strength in ageing humans [15]. These blood borne factors may function as systemic regulators influencing muscle and may be different from gene expression patterns in muscle itself.

Previous studies of transcriptome associations with muscle strength in humans were conducted predominantly in muscle tissue, and are mostly limited by small sample size [16] or focus on candidate genes [17]. These studies can be susceptible to false negatives statistical associations due to lack of power or coverage. A transcriptome-wide study of whole blood transcript associations with grip strength conducted by the InCHIANTI Ageing Study (mean age 72 years, 71% ≥72 years old) found only one gene, *CEBPB* (CCAAT/enhancer-binding protein beta, required for macrophage-mediated muscle repair in a murine model [18]), to be associated with muscle strength in older humans after adjustment for confounders and multiple testing [19]. A follow-up study in humans found that *CEBPB* expression increased following exercise-induced muscle damage [20]. However, because of the limited sample size of both studies, important transcriptional signals may have been missed.

In the present study we sought to test associations between transcripts expressed in whole blood and hand grip strength in multiple adult human cohorts. The majority of RNA in whole blood samples is derived from immature erythrocytes and platelets (~70% from reticulocytes and ~18% from reticulated platelets); however, these are predominantly globin-related and are not actively transcribed by circulating cells [21]. The remaining approximately 12% of RNA is from circulating white blood cells of all types, driving non-globin related gene expression. We have also performed subgroup analyses by age-group and gender, to check for heterogeneity in the results. We used a robust meta-analysis framework within four independent cohorts (n=7,781 participants) from the CHARGE (Cohorts for Heart and Ageing Research in Genomic Epidemiology) consortium [22] to identify the genes whose levels of expression assessed by blood transcripts were associated with muscle strength.

METHODS

Study sample

Characteristics of the cohorts are presented in Table 1. Complete data for the planned meta-analysis were available for 7,781 participants from four cohorts; the Framingham Heart Study [23] (FHS, n=5,576, ages=24-90), the InCHIANTI study [24] (n=667, ages=30-104), the Rotterdam Study [25] (RS, n=556, ages=46-89) and the Study of Health in Pomerania [26] (SHIP, n=982, ages=20-81) (total n=7,781). The FHS study included two related generations of participants (accounted for in the statistical methodology); the FHS Generation 2 (n=2,421, ages=40-90) and Generation 3 (n=3,155, ages=24-78) cohorts were included. The Netherlands Study of Depression and Anxiety [27] (NESDA, n=1,989, ages=18-65) is also reported, but was not included in the discovery meta-analysis due to unavailable data on white cell proportions necessary for the meta-analysis protocol. Overall the cohorts were quite similar with respect to sex-distribution and sampling methods, differing only by age distribution and lower mean hand-grip strength in the RS. Detailed study design and cohort information has been previously published [28].

Four additional subset analyses were performed. The sub-samples available were 1) older participants ≥60 years (n=2,402), 2) younger participants <60 years (n=5,379), 3) male participants (n=3,557), 4) female participants (n=4,224).

Phenotype

The primary phenotype was hand-grip strength in kg (a normally distributed phenotype). In the FHS, hand grip strength was measured with a Jamar dynamometer with three trials performed in each hand, and the maximum of the six trials for each participant was used in the analysis. In InCHIANTI each participant recorded their maximum grip strength three times in each hand, and the maximum recorded value of the six trials was used. In the Rotterdam Study grip strength of the non-dominant hand was measured three times for each participant, and the maximum recorded value was used. In the SHIP cohort participants were asked to press the hand dynamometer firmly for several seconds,

once per hand (left and right), and the maximum value was used. Each participant in NESDA was measured twice with a Jamar dynamometer in their dominant hand, with the maximum recorded value used.

Table 1. Characteristics of the study cohorts.

	Meta-analysis cohorts					Additional cohort ∞
Variables	**FHS Gen 2 ***	**FHS Gen 3 ***	**InCHIANTI ~**	**RS ~**	**SHIP ~**	**NESDA**
N	2,421	3,155	667	556	982	1,989
N >= 60 years *(%)*	1,319 (54%)	48 (1%)	547 (81%)	264 (45%)	268 (27%)	158 (8%)
Sex (male), n *(%)*	1,095 (45%)	1,470 (47%)	311 (46%)	272 (46%)	435 (44%)	1,328 (67%)
Age (yrs), mean±SD	66 ± 8.9	46 ± 8.8	72 ± 15	60 ± 7.9	50 ± 14	42 ± 13
Age (yrs), min:max	40 : 90	24 : 78	30 : 104	46 : 89	20 : 81	18 : 65
WBC-counts	yes ∴	yes	yes	yes	yes	no
Hand-grip strength						
Mean±SD (Kg)	31 ± 12	38 ± 12	29 ± 12	25 ± 9.4	38 ± 12	38 ± 12
Min : max (Kg)	1 : 76	5 : 84	3 : 76	2 : 55	11 : 73	10 : 90
Microarray platform	Affymetrix GeneChip Human Exon 1.0 ST	Affymetrix GeneChip Human Exon 1.0 ST	Illumina HumanHT-12 v3 BeadChip	Illumina HumanHT-12 v4 BeadChip	Illumina HumanHT-12 v3 BeadChip	Affymetrix Human Genome U219 Array

FHS = Framingham Heart Study; RS = Rotterdam Study 3; SHIP = Study of Health in Pomerania; NESDA = Netherlands Study of Depression and Anxiety; SD = standard deviation; WBC = white blood cell
** FHS cohorts analyzed together prior to overall meta-analysis; ~ Illumina-based cohorts analyzed together prior to overall meta-analysis; ∞ cohort not included in meta-analysis due to data missing from analysis protocol; ∴ cell counts imputed in this dataset; see methods*

Peripheral gene expression data

Blood samples were drawn from participants and RNA was isolated, reverse-transcribed to cDNA, which was then amplified and hybridized to a microarray individually for each cohort; methods described in detail [28]. Briefly, the FHS used Affymetrix Human Exon 1.0 ST GeneChips, characterizing the expression of 16,798 unique genes (after exclusion of probesets with relative log expression mean values <3). The InCHIANTI and SHIP studies used the Illumina HumanHT-12 v3 Expression BeadChip Kit, and the RS used the Illumina HumanHT-12 v4 Expression BeadChip Kit, with 37,348 probes measured on both Illumina platforms (22,911 unique genes; after exclusions of probes expressed above background in <5% of participants this becomes 15,639 unique genes). These four studies in the primary analysis all used PAXgene tubes to isolate and stabilize the RNA, thereby limiting the technical variability between studies. Finally the NESDA cohort utilized the Affymetrix Human Genome U219 Array, with expression information available on 18,212 unique gene identifiers. Quantile normalization and log2 transformation was performed on the gene expression data in each cohort, and both probes and samples were z-transformed. Raw data from

gene expression profiling are available online (FHS [*NCBI dbGAP:* phs000363.v7.p8], InCHIANTI [*GEO:* GSE48152], NESDA [NCBI dbGAP: phs000486.v1.p1], RS [*GEO:* GSE33828] and SHIP-TREND [*GEO:* GSE36382]).

Systematically mapping pairs of probes to RefSeq transcripts (one Affymetrix Exon ST and one Illumina HumanHT-12 probe) found 26,746 probe-pairs corresponding to 17,534 unique RefSeq gene symbols. The assignment of a probe to one or more transcripts was performed as described previously [29]. For the Illumina arrays, the transcript sequences derived from the 48,803 probe sequences provided in the Illumina annotation file (HumanHT-12_V3_0_R3_11283641_A, version 3.0, 7/1/2010) were mapped against all available mRNA sequences provided in the UCSC genome annotation database (version 06/30/2013) using string matching. Altogether 29,818 probes were successfully mapped to one or more validated mRNAs. Probes that could neither be mapped to a unique mRNA nor to a single annotated RefSeq gene using the UCSC database were flagged accordingly in the annotation file. In total, 27,171 probes (55.7%) were unambiguously associated with a single mRNA or gene. The same method and version of the UCSC database was used for mapping the probes of the Affymetrix GeneChip Human Exon 1.0 ST microarray. For this array, probe sequences were obtained from the annotation file version HuEx-1_0-st-v2.r2 restricting the probes to the main probe types of the core dataset with unique cross hybridization type, and combining them at the level of transcript cluster. For this array system, 196,515 probes (86.0%) of 17,876 transcript clusters were unambiguously associated with a single mRNA or gene. Finally, the probes of both array systems were combined based on the same transcripts obtained from the mapping against the UCSC database.

The Human Genome Nomenclature Committee [30] list 19,060 protein-coding genes (Sept 15, 2014), less than the total "unique identifiers" mapped by the two arrays used in the overall meta-analysis; this discrepancy is due to probes on the array mapping to non-protein-coding transcripts, which we have included under the term "unique genes" or "transcripts" in this manuscript.

Statistical analysis

Using the R statistical software [31] and package "lme4" [32] each cohort performed a linear mixed effects model for each probe in their microarray data, using the probe as the outcome, muscle strength as an independent variable, and with the following covariates included as fixed effects; age, sex, height (cm), weight (kg), cell count estimates (neutrophils, monocytes, basophils and eosinophils), and fasting state (where applicable). By including these factors as covariates in the models our results are independent of inter-individual variation in, for example, lymphocyte cell counts. The following covariates were included as random effects; batch (e.g. amplification and/or hybridization), study site (in InCHIANTI), family structure (in FHS), and RNA quality (e.g. RNA Integrity Number – RIN – where available). Empirical cell counts were only available in half of the FHS cohort; the rest of the cohort was imputed using partial least square regression methods (see Online Methods for more details).

Meta-analysis

A sample-size weighted meta-analysis method was used, where an overall p-value and Z-score for each probe are calculated which together describe the significance of the effect, and the direction and magnitude, respectively; this method was chosen over the effect size/standard error method because of the multiple array technologies and technical considerations that differed between the cohorts. The analysis was done using the Meta-Analysis Tool for Genome Wide-Association Scans (METAL) [33] which took the effect size, sample size, and p-values from the individual cohort results as input (we set the "minor allele", "major allele", "minor allele frequency", and "strand" to the same fixed value for all cohorts and probes, as this package was developed for GWAS and these options are not relevant for gene expression data).

For each analysis (the primary analysis including all individuals, and the four subset analyses) the Illumina-based cohorts (InCHIANTI, RS, SHIP) were meta-analyzed together, as these technologies are very similar, then a secondary meta-analysis was performed which used the FHS results and the Illumina results as the input; these are the final meta-analysis results reported. This reduced the heterogeneity in the meta-analysis due to array differences between the cohorts.

Before interpretation of the results, probes were excluded if they were expressed in <5% of the sample or if the heterogeneity p-value calculated by METAL was <0.05. The Benjamini-Hochberg (BH) [34] false-discovery rate (FDR) correction was applied to determine the statistically significant probes for each analysis. Validation was defined as a gene with $p<0.05$ in the NESDA cohort.

Ontology enrichment and network analysis

The WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) online resource is a method for determining pathway enrichment [35]. We conducted a "Gene Ontology" analysis (database version: Nov 11, 2012) and a "Human Phenotype Ontology" analysis (database version: May 20, 2014), which uses a systematic approach to phenotype abnormalities to link them into ontologies [36]. Default analysis options were selected, including BH multiple testing adjustment, and the list of 17,534 genes included in the meta-analysis we used as the "background".

A co-expression analysis was performed in the FHS (as the study with the largest sample size) where Spearman correlations were determined between each gene significantly associated with muscle strength, and all other genes (after adjusting the data for the covariates mentioned in the Statistical Analysis methods section). Genes correlated with rho ≥0.5 were selected for visualization in Cytoscape (v3.2.1). Ontology enrichment of networks was performed in Cytoscape using the BiNGO plugin (v2.44).

A priori genes associated with muscle function

We selected sets of genes known to influence muscle function for *a priori* analysis to highlight whether the pathways in muscle tissue are also important in whole blood. Kelch proteins, including *KLHL19*, *KLHL31*, *KLHL39m*, and *KLHDC1*, are involved in skeletal muscle function and development;

canonical and non-canonical Wnt signaling play crucial roles in maintenance and development of skeletal muscle; insulin-like growth factors (IGF's) are also known to play roles in muscle growth and homeostasis, and finally TGF-β family members, including myostatin (*GDF8*) [37] and *GDF11*. Supplementation of *GDF11* in mice was reported to ameliorate the sarcopenia-like phenotype. In their 2007 study, Melov et al identified 586 unique genes expressed in muscle that were associated with endurance exercise training and differed between older and younger men [16], which we also checked for associations with muscle strength in this analysis.

Systematic literature search for genes

For each significant gene in the analysis a systematic search of literature was performed by accessing the "publications" list from GeneCards (*www.genecards.org*) [38], which has the advantage of including publications where the gene ID may have changed over time. From this list the title and abstract were downloaded from NCBI PubMed (*www.ncbi.nlm.nih.gov/pubmed*). Searches were then made within each publication for the text string "muscle", and results counted.

RESULTS

Meta-analysis: genes associated with muscle strength

Overall, 26,746 probe-pairs (corresponding probes on the Affymetrix and Illumina platforms), mapping to 17,534 unique gene identifiers, were available for the meta-analysis. Including data from all 7,781 participants, 208 unique genes (246 probe-pairs) were associated with muscle strength (FDR<0.05; Figure 1; see Supplementary Table 1 for all significant results) after correction for multiple confounders, and excluding results with significant heterogeneity (het p<0.05) between cohorts; 133 were negatively associated with grip strength, 75 were positively associated.

Of the 208 unique genes associated with muscle strength in the meta-analysis all were significant in FHS alone (nominal p<0.05), and 79 (38%) were also "independently" associated with muscle strength in the Illumina meta-analysis (nominally significant; p<0.05). Details of the statistically most significant 'top' 20 transcripts are shown in Table 2. The proportion "independently replicated" was greater for the top 30 most significant genes identified in the meta-analysis (21 of 30=70%).

Meta-analysis in subsets of the participants

The analyses in subsets of the participants identified 13 genes associated with muscle strength at transcriptome-wide significance in the meta-analysis that were not identified in the analysis of all participants together (Table 3; see Supplementary Table 2-4 for full results for each subset). We also investigated whether the 208 genes identified in the analysis of all individuals were nominally significant (p<0.05) in the subsets. 153 of the 208 genes were nominally significant in the older participants, 198 in the younger, 200 in the males and 121 in the females (Supplementary Table 1). Supplementary Table 5 includes additional information for all genes included in Tables 2 & 3.

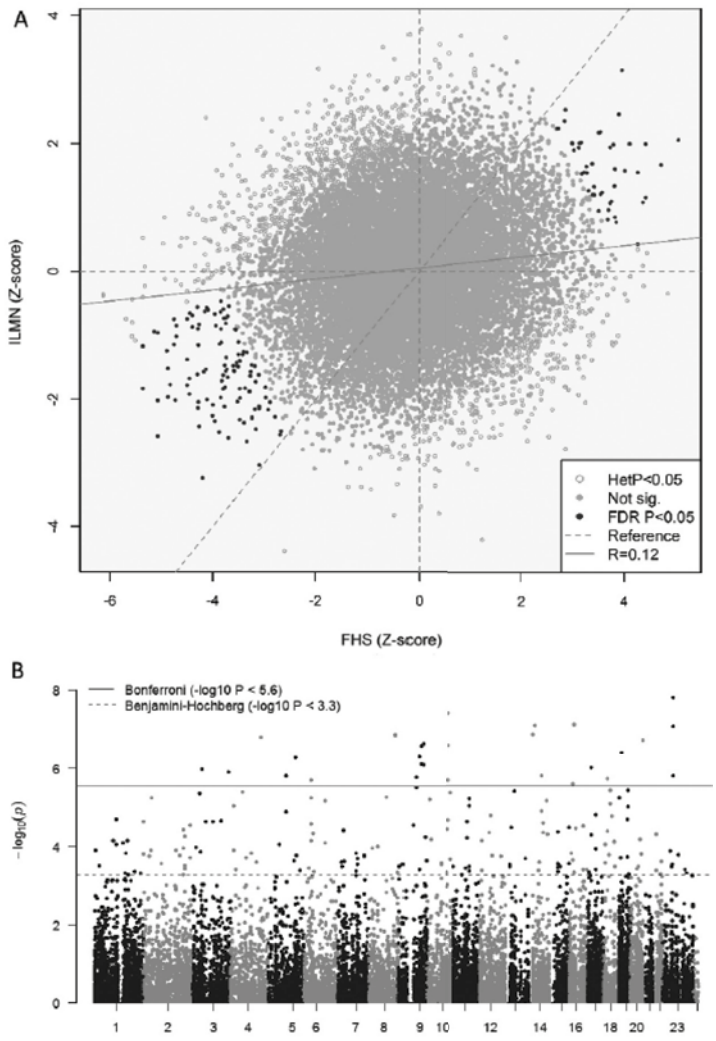


Figure 1. Gene transcripts associated with muscle strength in all participants.

(**A**) Compares the individual meta-analyses performed in the Illumina-cohorts and the FHS separately. The dark grey points represent gene transcripts significantly associated with muscle strength (FDR<0.05). The light grey points were not significant in this analysis. The unfilled grey points were excluded due to significant heterogeneity (Cochran's Q-test p<0.05 [33]). The solid grey line shows the trend across all the genes. (**B**) Shows the meta-analysis results by Manhattan plot. The dashed line indicates those probes significantly associated with grip strength after Benjamini-Hochberg correction, the solid line shows those significant after Bonferroni correction, for comparison.

Table 2. Top 20 unique genes associated with muscle strength in the meta-analysis of all participants, with robust replication in Illumina.

Meta-analysis			P-values				
Z-score	P-value	BH P-value	FHS	Illumina	Gene	Entrez-ID	Name
-5.67	1.5×10^{-8}	2.0×10^{-4}	3.9×10^{-7}	9.9×10^{-3}	*ALAS2*	212	aminolevulinate, delta-, synthase 2
5.37	7.8×10^{-8}	3.1×10^{-4}	4.3×10^{-7}	4.0×10^{-2}	*HEATR5A*	25938	HEAT repeat containing 5A
-5.28	1.3×10^{-7}	3.9×10^{-4}	2.7×10^{-5}	1.2×10^{-3}	*PNP*	4860	purine nucleoside phosphorylase
-5.17	2.4×10^{-7}	4.7×10^{-4}	1.1×10^{-6}	5.0×10^{-2}	*STOM*	2040	stomatin
5.14	2.8×10^{-7}	4.7×10^{-4}	7.9×10^{-6}	1.1×10^{-2}	*RPS6KA5*	9252	ribosomal protein S6 kinase, 90kDa, polypeptide 5
-5.09	3.7×10^{-7}	5.8×10^{-4}	5.4×10^{-5}	1.8×10^{-3}	*MBNL3*	55796	Muscleblind-Like Splicing Regulator 3
-5.07	3.9×10^{-7}	5.8×10^{-4}	2.3×10^{-6}	4.4×10^{-2}	*RAD23A*	5886	RAD23 homolog A (S. cerevisiae)
5.02	5.1×10^{-7}	6.8×10^{-4}	7.7×10^{-5}	1.7×10^{-3}	*HNRNPA0*	10949	heterogeneous nuclear ribonucleoprotein A0
-4.90	9.5×10^{-7}	1.1×10^{-3}	2.0×10^{-5}	1.5×10^{-2}	*GID4*	79018	GID Complex Subunit 4
-4.88	1.1×10^{-6}	1.1×10^{-3}	9.3×10^{-6}	3.4×10^{-2}	*TFDP1*	7027	transcription factor Dp-1
4.80	1.6×10^{-6}	1.4×10^{-3}	9.9×10^{-6}	4.7×10^{-2}	*ARRDC3*	57561	arrestin domain containing 3
-4.77	1.7×10^{-6}	1.5×10^{-3}	1.8×10^{-5}	3.3×10^{-2}	*RIOK3*	8780	RIO kinase 3
-4.71	2.5×10^{-6}	1.8×10^{-3}	1.9×10^{-5}	4.1×10^{-2}	*NTAN1*	123803	N-terminal asparagine amidase
4.66	3.1×10^{-6}	2.2×10^{-3}	8.1×10^{-4}	6.0×10^{-4}	*PDE4D*	5144	phosphodiesterase 4D, cAMP-specific
-4.62	3.8×10^{-6}	2.4×10^{-3}	1.1×10^{-4}	1.2×10^{-2}	*RGCC*	28984	Regulator Of Cell Cycle
4.61	4.1×10^{-6}	2.5×10^{-3}	9.6×10^{-5}	1.4×10^{-2}	*POLR2B*	5431	polymerase (RNA) II (DNA directed) polypeptide B
-4.59	4.4×10^{-6}	2.5×10^{-3}	6.1×10^{-5}	2.4×10^{-2}	*EIF1B*	10289	eukaryotic translation initiation factor 1B
4.57	4.8×10^{-6}	2.7×10^{-3}	5.4×10^{-4}	2.0×10^{-3}	*CIRBP*	1153	cold inducible RNA binding protein
-4.57	4.8×10^{-6}	2.7×10^{-3}	6.2×10^{-5}	2.7×10^{-2}	*ASGR2*	433	asialoglycoprotein receptor 2
-4.55	5.3×10^{-6}	2.9×10^{-3}	3.8×10^{-5}	4.5×10^{-2}	*CA1*	759	carbonic anhydrase I

BH = Benjamini-Hochberg. Ordered by meta-analysis p-value. Showing top 20 results with nominal 'replication' ($p<0.05$) in Illumina. Duplicate gene entries excluded. Full table in Supplementary Table 1.

Table 3. Thirteen genes associated with muscle strength in age or gender specific subset analyses.

Subset		Meta-analysis of subset			P-values			
Group	N	Z-score	P-value	BH P-value	FHS	Illumina	Gene	Name
Younger & Male	5379	-4.879	1.1×10^{-6}	9.5×10^{-3}	2.4×10^{-6}	1.3×10^{-1}	ASAP1	ArfGAP with SH3 domain, ankyrin repeat and PH domain 1
Younger	5379	4.495	6.9×10^{-6}	2.5×10^{-2}	7.5×10^{-5}	3.3×10^{-1}	*PKN2*	protein kinase N2
Younger	5379	-4.426	9.6×10^{-6}	2.5×10^{-2}	2.9×10^{-5}	6.1×10^{-1}	*RNF175*	ring finger protein 175
Younger	5379	-4.407	1.1×10^{-5}	2.5×10^{-2}	3.3×10^{-5}	2.4×10^{-1}	*TMED5*	transmembrane emp24 protein transport domain containing 5
Younger	5379	4.235	2.3×10^{-5}	3.5×10^{-2}	4.2×10^{-5}	4.7×10^{-1}	*ZFYVE27*	zinc finger, FYVE domain containing 27
Younger & Male	5379	-4.217	2.5×10^{-5}	3.5×10^{-2}	3.4×10^{-3}	2.5×10^{-3}	*GID8*	GID complex subunit 8
Younger	5379	-4	6.3×10^{-5}	4.2×10^{-2}	1.7×10^{-4}	1.2×10^{-2}	*PIK3R2*	phosphoinositide-3-kinase, regulatory subunit 2
Younger	5379	-3.961	7.5×10^{-5}	4.2×10^{-2}	7.3×10^{-6}	3.3×10^{-1}	*PDK4*	pyruvate dehydrogenase kinase, isozyme 4
Younger	5379	-3.961	7.5×10^{-5}	4.2×10^{-2}	2.0×10^{-6}	6.7×10^{-1}	*CTNNAL1*	catenin (cadherin-associated protein), alpha-like 1
Younger	5379	-3.937	8.3×10^{-5}	4.4×10^{-2}	7.6×10^{-5}	1.2×10^{-1}	*COQ9*	coenzyme Q9
Male	3557	4.565	5.0×10^{-6}	1.9×10^{-2}	1.2×10^{-5}	1.1×10^{-1}	*RAC1*	ras-related C3 botulinum toxin substrate 1
Male	3557	4.041	5.3×10^{-5}	4.3×10^{-2}	8.0×10^{-5}	1.9×10^{-1}	*NDUFS1*	NADH dehydrogenase (ubiquinone) Fe-S protein 1, 75kDa
Female	4224	-4.907	9.3×10^{-7}	2.5×10^{-2}	8.0×10^{-4}	1.1×10^{-4}	*DEFA4*	defensin, alpha 4, corticostatin

BH = Benjamini-Hochberg. Some genes were identified in in multiple groups; for these genes the statistics for the larger group are given; in all cases direction of association is the same. Ordered by subset, then P-value.

Table 4. Top 20 probes in the FHS analysis that did not map to a corresponding Illumina probe, ordered by P-value.

Estimate	P-value	BH P-value	Chr	Start	Gene	Name
-0.0037	1.76×10^{-6}	1.49×10^{-3}	1	144989319		
-0.0028	4.89×10^{-6}	2.88×10^{-3}	9	37800563	*DCAF10*	DDB1 and CUL4 associated factor 10
0.0031	8.57×10^{-6}	4.12×10^{-3}	1	150522766	*ADAMTSL4*	ADAMTS(a disintegrin and metalloproteinase with thrombospondin motifs)-like 4
0.0038	9.29×10^{-6}	4.15×10^{-3}	16	89980135		
0.0040	1.95×10^{-5}	5.97×10^{-3}	2	162412847	*SLC4A10*	solute carrier family 4, sodium bicarbonate transporter, member 10
0.0029	3.24×10^{-5}	7.74×10^{-3}	16	9186734		
0.0031	3.93×10^{-5}	8.51×10^{-3}	1	44440179	*ATP6V0B*	ATPase, H+ transporting, lysosomal 21kDa, V0 subunit b
-0.0027	4.10×10^{-5}	8.64×10^{-3}	3	63819562	*THOC7*	THO complex 7 homolog (Drosophila)
0.0032	5.38×10^{-5}	9.81×10^{-3}	20	3898284		
0.0019	1.49×10^{-4}	1.70×10^{-2}	20	1316212		
0.0018	1.89×10^{-4}	1.91×10^{-2}	3	128628719	*ACAD9* | *KIAA1257*	acyl-CoA dehydrogenase family, member 9 | KIAA1257
-0.0017	2.21×10^{-4}	2.08×10^{-2}	1	8021733	*PARK7*	parkinson protein 7
0.0036	3.18×10^{-4}	2.69×10^{-2}	12	57809458		
0.0010	3.20×10^{-4}	2.70×10^{-2}	1	153901987	*DENND4B*	DENN/MADD domain containing 4B
-0.0019	3.56×10^{-4}	2.87×10^{-2}	3	10157370	*BRK1*	BRICK1, SCAR/WAVE actin-nucleating complex subunit
-0.0028	3.84×10^{-4}	2.97×10^{-2}	2	95517671	*TEKT4*	tektin 4
-0.0014	3.95×10^{-4}	3.03×10^{-2}	1	247937996	*OR9H1P*	olfactory receptor, family 9, subfamily H, member 1 pseudogene
0.0018	6.01×10^{-4}	3.89×10^{-2}	9	124042152	*GSN-AS1*	GSN(gelsolin) antisense RNA 1
0.0028	6.41×10^{-4}	4.03×10^{-2}	6	20451305		
-0.0021	6.90×10^{-4}	4.19×10^{-2}	12	8024178	*NANOGP1* | *NANOG*	Nanog homeobox pseudogene 1 | Nanog homeobox

BH = Benjamini-Hochberg. Blank gene symbols were not annotated to a specific gene. Table continued in Supplementary Table 9. Ordered by P-value. Not all probes map to gene ID's.

Significant genes only available on one array

Due to differences between the array technologies and relative abundance of transcripts, not all the genes were eligible for the meta-analysis. In the analysis on all individuals, 1,123 probes (898 unique identifiers) were present on the Affymetrix Exon array that did not have a corresponding probe on the Illumina array; 21 of these probes were significantly associated with hand-grip strength after BH adjustment for multiple testing (see Table 4 for top 20 probes in the "all individuals" analysis and Supplementary Tables 6-8 for list of significant probes in each of the FHS analyses with significant results). In the Illumina array, 7,768 probes (6,119 unique gene identifiers) were available that did

not map to a gene/transcript in the Affymetrix Exon array (after excluding lowly expressed probes). None of the probes were significantly associated with muscle strength after BH multiple testing correction.

Ontology enrichment of strength-associated genes

Two analyses were performed to identify pathways using the WebGestalt web resource based on the 208 genes associated with muscle strength in the meta-analysis of all participants;

1. *Gene Ontology* analysis found that 10 biological processes were significantly enriched (FDR<0.05) (Table 5) including "hemoglobin metabolic process" and related processes, "innate immune response" (18 genes) and the stress response (55 genes); 10 molecular functions were enriched (including "protein binding genes"), and 10 cellular component pathways were enriched (including "intracellular membrane-bound organelle") (See Supplementary Table 9).
2. *Human Phenotype Ontology* analysis found 10 phenotypes significantly enriched in the genes, including "Anemia due to reduced life span of red cells", "Hemolytic Anemia", and "Abnormality of erythrocytes" (See Supplementary Table 10).

Additionally, network analysis in the FHS of all genes correlated (rho≥0.5) with the strength-associated genes (n=425) revealed four clusters with at least 10 genes, all of which had a number of significantly (FDR p<0.05) enriched pathways (Supplementary Table 11): 1) the largest cluster (n=333 of 425 genes) included "protein ubiquitination", "erythrocyte homeostasis", and "cellular metabolic process". 2) The second cluster (n=32 genes) was enriched for genes in "regulation of cell communication", "actin cytoskeleton" and "ATP metabolic process". 3) The third cluster (n=18) had two enriched ontologies only: "cell surface receptor-linked signaling pathway" and "cytolysis". 4) The final cluster (n=10) was enriched for terms including "negative regulator or immune system process" and "negative regulation of complement activation".

A priori genes associated with muscle function

Of 20 *IGF*-related genes tested in this meta-analysis two were significantly associated with muscle strength: *IGF1R* (positively associated) and *IGF2BP2* (negatively associated; meta-analysis FDR=3.2 x 10^{-2} and FDR=1.2 x 10^{-3}, respectively). Expression of myostation (*MSTN*), follistatin (*FST*) and *GDF11* were not associated with muscle strength (FDR>0.05). 40 unique Kelch genes were tested in the meta-analysis; none were associated with muscle strength in whole blood (FDR>0.05). 18 unique Wnt genes (from *WNT1* to *WNT9B*) were tested in the meta-analysis; none were associated with muscle strength in whole blood (FDR>0.05). All 10 Frizzled genes (*FZD1-10*, receptors for the Wnt pathway) were also available to test; none were associated with muscle strength. Similarly all three Dishevelled genes were available to test (*DVL1-3,* acts directly downstream of the Frizzled receptors) and none were associated with muscle strength. Of 586 genes identified by Melov *et al* that were differentially expressed in muscle tissue between old and young men following endurance training [16] four were associated with muscle strength in this analysis: *ANP32B*, *CIRBP*, *MCM7* and *MGST1*.

Most associated genes are not previously linked to muscle in the literature

For each of the 221 genes associated with muscle strength we searched in the published literature cataloged on GeneCards and NCBI Pubmed titles and abstracts, using the search term "muscle": for 115 of these 221 genes (52%) there were no mentions of muscle (as of Nov 12, 2014; Supplementary Table 12).

Few genes replicate in the NESDA cohort

NESDA was not included in the meta-analysis due to data limitations: the lack of empirically determined or reliably imputed white cell count data, the use of a different microarray technology (a predecessor to the Exon array used by FHS, much more dissimilar than the v3/v4 Illumina arrays are to one another), and a younger population than the other cohorts included in the meta-analysis (max age=65, see Table 1). As noted above, in total 221 unique genes were associated with muscle strength across all the meta-analyses performed. Of 208 genes significantly associated with muscle strength in analysis 1 (all participants) it was possible to test 144 in the NESDA cohort; 7 genes were also associated with muscle strength ($p<0.05$) in the NESDA cohort (*ACSL6, ALDH5A1, CARHSP1, FGL1, NRG1, PIGB, SIGLEC7*).

Associations with knee strength

Maximum knee and grip strength (both in Kg) were measured in 619 participants in the InCHIANTI study and were highly correlated (Pearson R=0.751), and were significantly associated after adjustment for age, sex, height and weight (coefficient=0.193, $p=8.2 \times 10^{15}$) in linear regression models with knee strength as the dependent variable.

DISCUSSION

In this discovery study we set out to determine whether specific transcript levels in blood are associated with muscle strength in multiple human cohorts including mostly middle-aged volunteers. Previous cross-sectional (and longitudinal) studies have shown that the degree and *rate* of loss of strength (and muscle mass) is greater in older participants [39]. We therefore performed stratified analyses by age and gender to determine whether transcripts or pathways associated with muscle strength in whole blood differ between these groups. 208 unique genes were associated with muscle strength in the analysis of all participants (Table 2 for 20 most robust associations). Thirteen additional unique genes were identified that were only associated when participants were separated into older/younger or male/female groups (Table 3). In total 221 unique genes were associated with muscle strength in at least one analysis, 52% of which were not previously linked to the term "muscle" in the published literature cataloged on GeneCards and PubMed (as of Nov 12, 2014).

We observe significant associations between muscle strength and expression of *IGF1R* and *IGF2BP2* (positive and negative directions of association with muscle, respectively), growth factors involved

in skeletal muscle growth [40,41]; the former is known to enhance cell survival by mediating *IGF1* signaling, and the latter modulates *IGF2* translation and has genetic variants associated with type-2 diabetes [42]. Of 586 genes that differ in expression in muscle between old and young men after endurance training [16], four were associated with grip strength in this study: *MGST1* (negative direction), an immune mediator which may protect against oxidative stress [43]; *MCM7* (positive direction), which regulates DNA replication during proliferation [44]; *CIRBP* (positive direction), which promotes inflammation in response to shock and sepsis [45]; and *ANP32B* (negative direction), a cell-cycle progression and anti-apoptosis factor [38]. These latter results suggest that most blood based gene expression associated with strength is different from that seen in muscle itself, which is not unexpected given the respective systemic regulatory versus myofibril maintenance functions involved. Further work should explore whether transcripts that alter in response to exercise show overlaps between circulating cells and muscle. Interestingly, no genes from the *Wnt* or *Kelch* pathways (both known to be important for muscle function [46,47]) were associated with strength in this analysis; nor was *GDF11*, a protein that can reverse age-related muscle dysfunction in mice [14], although as noted we observe associations between strength and expression of two IGF-related genes.

Other genes of note include *CCR6* and *PRF1* (both positively associated with muscle strength and age [48]): *CCR6* is implicated in B-cell maturation and recruitment [49]; perforin (*PRF1*) is a protein secreted by cytotoxic T-cells which creates pores in membranes to permit apoptosis-inducing granzyme into the target cell [50]. These findings are consistent with the notion that inflammation may be associated with muscle repair, maintenance and turnover, at least in part by interfering with the production and biological activity of *IGF-1* [51].

NANOG expression was measured in the FHS analysis only and is positively associated with strength in the FHS analysis (Table 4); *NANOG* can reverse ageing of some stem cells [52] and, in combination with three other genes (*OCT4*, *SOX2*, and *LIN28*, not significant in this analysis), can induce pluripotency of somatic cells [53]. This may suggest that differentiation (of whole blood cells) is inversely correlated with muscle strength, but the mechanisms are unclear.

Genes identified in subset analyses

Thirteen genes were associated with muscle strength at transcriptome-wide significance in the subset analysis only (Table 3). These were predominantly in the younger (<60 years) group and included *PIK3R2*, a negative regulator of the PI3K/AKT growth pathway; the negative expression association with strength suggests that there is increased PI3K activity (due to reduced expression of *PIK3R2*) with increasing muscle strength, in whole blood. This association is observed in the younger subset ($p=6 \times 10^{-5}$) but not in the older subset, even nominally (p=0.92), suggesting differences in growth pathway expression in blood with respect to muscle strength as individuals age. Similarly expression of *PDK4* (inhibits pyruvate dehydrogenase in mitochondria, thereby reducing the conversion of pyruvate into acetyl-CoA) is negatively associated with strength, suggesting increased pyruvate dehydrogenase activity with increased strength in younger individuals only.

PKN2, (associated with height [42] and cell-cycle progression), is positively associated with strength in the younger individuals, underlining the difference in growth pathways in whole blood between younger and older individuals. See Supplementary Table 5 for more details on the other results.

Defensin, Alpha 4, Corticostatin (*DEFA4*, negative strength association in the analysis of females only) is a cytotoxic peptide that has antimicrobial activity against Gram-negative bacteria (predominantly) [49]. In males, expression of *RAC1* (membrane-associated GTPase involved in signal transduction, including growth signals) and *NDUFS1* (member of mitochondrial complex 1, may form part of the active site) were positively associated with muscle strength; these associations may suggest that on average males have specific energy and growth-related gene expression relating to strength.

No genes were associated (transcriptome-wide) in the older participants only; although the sample size was still reasonably high (2,402 participants), variability in the strength phenotype as individuals age and development of various co-morbidities plus chronic inflammation may reduce the power to detect associations. Subset-specific gene-expression associations with strength reported here need to be replicated and added to, as we may lack statistical power in this study to detect smaller-effect associations, and the microarrays do not quantify all transcripts or isoforms present.

Enrichment analysis

WebGestalt analyses identified statistically significant enrichment for genes in the biological process "Hemoglobin Metabolic Process" and the phenotypic abnormality "Hemolytic Anemia", amongst others. Anemia is a cross-sectional correlate of muscle strength and predicts accelerated muscle strength decline with ageing [54], while hemoglobin levels are positively associated with muscle strength and density [54]. Circulating reticulocytes (erythrocyte precursors with some residual RNA present) were not adjusted for in this analysis and are likely the source of the associations with genes such as *ALAS2* (strongest meta-analysis association, negative direction – a rate-limiting step in heme biosynthesis [55]).

"Innate Immune Response" – which includes macrophages – genes were also enriched in the results. *CEBPB*, the gene implicated in the macrophage wound-healing response [18] and significantly associated with muscle strength in the 2012 study by InCHIANTI [19], did not replicate in the other cohorts. This could be due to methodological differences between the previous study and this meta-analysis, as well as differences in age distribution (81% of the InCHIANTI cohort is aged >=60 years, compared to 31% in this analysis, which includes the InCHIANTI cohort; Table 1). The implications are unclear given the mouse model evidence of plausible biological mechanism [18] and evidence in humans that exercise-induced muscle damage is associated with *CEBPB* expression changes in whole blood [20]. Further work is required in older and frail groups.

Co-expression analysis of all genes correlated with those identified in the meta-analysis to be significantly associated with strength revealed four clusters. Ontology enrichment analysis of these revealed very similar results to those identified using WebGestalt only on the genes significantly

associated with strength, emphasizing the association of immune activation and cell signaling pathways to muscle strength in whole blood, in addition to hemoglobin pathways.

LIMITATIONS

There are several potential limitations of this study including its cross-sectional design; it is not possible to determine a causal direction in this study for the associations reported, but the robustly identified markers emerging provide a sound foundation for follow-up studies to address causation. Grip strength is strongly correlated with strength in other key muscle systems (see introduction), but further work will be needed to confirm more specific gene expression associations with strength in other muscle groups. Grip strength can be influenced by non-muscle strength factors, including functional anomalies in the hands, for example caused by rheumatoid arthritis [56], and work is needed to clarify whether any of our findings reflect these alternative influences.

Another potential limitation is the mixed cell subtype composition of "whole blood": our analysis approach based on overall expression should have greater power to detect net expression changes in common immune cell types or large changes in expression of highly specific genes, but will have less power to detect smaller expression changes within less numerous cell subtypes. The cell subtype origins of the top transcripts reported here now need to be identified. Additionally, the microarray technology used across the participating cohorts was not the same. However, 21 (70%) of the top 30 meta-analysis results were independently replicated between the platforms, which suggests that the top (most strongly associated) results are very robust to cohort and array differences. Also, the current analysis has identified expressed genes statistically associated with muscle strength, but future work will be needed to identify the mechanisms underlying these associations, and whether these act on muscle directly or through indirect pathways, perhaps with effects on central command, cerebellar coordination or neural transmission. Finally, work is needed on whether the identified strength associated gene expression transcripts are predictive of subsequent changes in strength or functional decline.

CONCLUSIONS

In this first large-scale transcriptome wide study in human blood, we have identified robust associations between the expression of 221 genes and muscle strength in adults. Several known pathways were confirmed, including growth factor-related genes, the innate immune response and hemoglobin metabolism. For 115 genes this analysis appears to provide the first published link to muscle. The analysis also suggests that parts of the expression signatures may be specific to subgroups, notably with 10 genes associated with muscle strength only in younger people.

Further work is needed to establish which of the identified genes predict future changes in strength. The findings of genes via expression microarrays may help identify key changes in cell subtypes in blood contributing to strength, through studies of the cellular origins of gene expression signals. Future research should also include longitudinal data to assess whether expression of the identified genes predicts poor muscle strength or functional outcomes.

FUNDING AND ACKNOWLEDGMENTS

FHS

FHS gene expression profiling was funded through the Division of Intramural Research (Principal Investigator, Daniel Levy), National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD. Dr. Murabito is supported by NIH grant R01AG029451. Dr. Kiel is supported by NIH R01 AR41398. The Framingham Heart Study is supported by National Heart, Lung, and Blood Institute contract N01-HC-25195.

InCHIANTI

The InCHIANTI study was supported in part by the Intramural Research Program, National Institute on Ageing, NIH, Baltimore MD USA. D.M. and L.W.H. were generously supported by a Wellcome Trust Institutional Strategic Support Award (WT097835MF). W.E.H. was funded by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) for the South West Peninsula. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health in England.

The infrastructure for the NESDA study (*www.nesda.nl*) is funded through the Geestkracht program of the Netherlands Organisation for Health Research and Development (Zon-Mw, grant number 10-000-1002) and is supported by participating universities and mental health care organizations (VU University Medical Center, GGZ inGeest, Arkin, Leiden University Medical Center, GGZ Rivierduinen, University Medical Center Groningen, Lentis, GGZ Friesland, GGZ Drenthe, Scientific Institute for Quality of Healthcare (IQ healthcare), Netherlands Institute for Health Services Research (NIVEL) and Netherlands Institute of Mental Health and Addiction (Trimbos Institute).

Rotterdam Study

The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), Netherlands Organisation of Scientific Research NWO Investments (nr. 175.010.2005.011, 911-03-012), the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists. The generation and management of RNA-expression array data for the Rotterdam Study was executed and funded by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine,

Erasmus MC, the Netherlands. We thank Marjolein Peters, MSc, Ms. Mila Jhamai, Ms. Jeannette M. Vergeer-Drop, Ms. Bernadette van Ast-Copier, Mr. Marijn Verkerk and Jeroen van Rooij, BSc for their help in creating the RNA array expression database.

SHIP

SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network 'Greifswald Approach to Individualized Medicine (GANI_MED)' funded by the Federal Ministry of Education and Research (grant 03IS2061A). The University of Greifswald is a member of the 'Center of Knowledge Interchange' program of the Siemens AG and the Caché Campus program of the InterSystems GmbH.

ADDITIONAL INFORMATION

Supplementary Information accompanies this paper at *http://physiolgenomics.physiology.org/*.

REFERENCES

1. Rantanen T. Muscle strength, disability and mortality. Scand J Med Sci Sports 13: 3–8, 2003.
2. Guralnik JM, Ferrucci L, Simonsick EM, Salive ME, Wallace RB. Lower-extremity function in persons over the age of 70 years as a predictor of subsequent disability. N Engl J Med 332: 556–61, 1995.
3. Lang T, Streeper T, Cawthon P, Baldwin K, Taaffe DR, Harris TB. Sarcopenia: etiology, clinical consequences, intervention, and assessment. Osteoporos Int 21: 543–59, 2010.
4. Bohannon RW, Magasi SR, Bubela DJ, Wang YC, Gershon RC. Grip and Knee extension muscle strength reflect a common construct among adults. Muscle and Nerve 46: 555–558, 2012.
5. Samson MM, Meeuwsen IB, Crowe A, Dessens JA, Duursma SA, Verhaar HJ. Relationships between physical performance measures, age, height and body weight in healthy adults. Age Ageing 29: 235–242, 2000.
6. Cesari M, Pahor M, Lauretani F, Zamboni V, Bandinelli S, Bernabei R, et al. Skeletal muscle and mortality results from the InCHIANTI Study. J Gerontol A Biol Sci Med Sci 64: 377–84, 2009.
7. Newman AB, Kupelian V, Visser M, Simonsick EM, Goodpaster BH, Kritchevsky SB, et al. Strength, but not muscle mass, is associated with mortality in the health, aging and body composition study cohort. J Gerontol A Biol Sci Med Sci 61: 72–77, 2006.
8. Visser M, Deeg DJ, Lips P, Harris TB, Bouter LM. Skeletal muscle mass and muscle strength in relation to lower-extremity performance in older men and women. J Am Geriatr Soc 48: 381–386, 2000.
9. Leong DP, Teo KK, Rangarajan S, Lopez-Jaramillo P, Avezum A, Orlandini A, et al. Prognostic value of grip strength: findings from the Prospective Urban Rural Epidemiology (PURE) study. Lancet (2015). doi: 10.1016/S0140-6736(14)62000-6.
10. Baylis D, Ntani G, Edwards MH, Syddall HE, Bartlett DB, Dennison EM, Martin-Ruiz C, von Zglinicki T, Kuh D, Lord JM, Aihie Sayer A, Cooper C. Inflammation, Telomere Length, and Grip Strength: A 10-year Longitudinal Study. Calcif Tissue Int 95: 54–63, 2014.
11. Schaap LA, Pluijm SMF, Deeg DJH, Visser M. Inflammatory markers and loss of muscle mass (sarcopenia) and strength. Am J Med 119: 526.e9–17, 2006.
12. Carlson B, Faulkner J. Muscle transplantation between young and old rats: age of host determines recovery. Am J Physiol Physiol 256: C1262, 1989.
13. Conboy IM, Conboy MJ, Wagers AJ, Girma ER, Weissman IL, Rando TA. Rejuvenation of aged progenitor cells by exposure to a young systemic environment. Nature 433: 760–4, 2005.
14. Sinha M, Jang YC, Oh J, Khong D, Wu EY, Manohar R, et al. Restoring systemic GDF11 levels reverses age-related dysfunction in mouse skeletal muscle. Science 344: 649–52, 2014.
15. Lee CE, McArdle A, Griffiths RD. The role of hormones, cytokines and heat shock proteins during age-related muscle loss. Clin Nutr 26: 524–34, 2007.
16. Melov S, Tarnopolsky M a, Beckman K, Felkey K, Hubbard A. Resistance exercise reverses aging in human skeletal muscle. PLoS One 2: e465, 2007.
17. Patel HP, Al-Shanti N, Davies LC, Barton SJ, Grounds MD, Tellam RL, et al. Lean Mass, Muscle Strength and Gene Expression in Community Dwelling Older Men: Findings from the Hertfordshire Sarcopenia Study (HSS). Calcif. Tissue Int. (July 24, 2014). doi: 10.1007/s00223-014-9894-z.
18. Ruffell D, Mourkioti F, Gambardella A, Kirstetter P, Lopez RG, Rosenthal N, Nerlov C. A CREB-C/EBPbeta cascade induces M2 macrophage-specific gene expression and promotes muscle injury repair. Proc Natl Acad Sci U S A 106: 17475–80, 2009.
19. Harries LW, Pilling LC, Hernandez LDG, Bradley-Smith R, Henley W, Singleton AB, et al. CCAAT-enhancer-binding protein-beta expression in vivo is associated with muscle strength. Aging Cell 11: 262–268, 2012.
20. Blackwell J, Harries LW, Pilling LC, Ferrucci L, Jones A, Melzer D. Changes in CEBPB expression in circulating leukocytes following eccentric elbow-flexion exercise. J. Physiol. Sci. (November 13, 2014). doi: 10.1007/s12576-014-0350-7.
21. Affymetrix. Technical Note: Factors Affecting Blood Gene Expression [Online]. 2011. http://www.panomics.com/downloads/QG_BloodGene_TN_RevB_110719.pdf.

22. Psaty BM, O'Donnell CJ, Gudnason V, Lunetta KL, Folsom AR, Rotter JI, et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circ Cardiovasc Genet 2: 73–80, 2009.
23. Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP. An investigation of coronary heart disease in families. The Framingham offspring study. Am J Epidemiol 110: 281–90, 1979.
24. Ferrucci L, Bandinelli S, Benvenuti E, Di Iorio A, Macchi C, Harris TB, Guralnik JM. Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. J Am Geriatr Soc 48: 1618–25, 2000.
25. Hofman A, van Duijn CM, Franco OH, Ikram MA, Janssen HLA, Klaver CCW, et al. The Rotterdam Study: 2012 objectives and design update. Eur J Epidemiol 26: 657–86, 2011.
26. Völzke H, Alte D, Schmidt CO, Radke D, Lorbeer R, Friedrich N, et al. Cohort profile: the study of health in Pomerania. Int J Epidemiol 40: 294–307, 2011.
27. Penninx BWJH, Beekman ATF, Smit JH, Zitman FG, Nolen WA, Spinhoven P, et al. The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. Int J Methods Psychiatr Res 17: 121–40, 2008.
28. Huan T, Esko T, Peters MJ, Pilling LC, Schramm K, Schurmann C, et al. A Meta-analysis of Gene Expression Signatures of Blood Pressure and Hypertension. PLOS Genet 11: e1005035, 2015.
29. Schurmann C, Heim K, Schillert A, Blankenberg S, Carstensen M, Dörr M, et al. Analyzing illumina gene expression microarray data from different tissues: methodological aspects of data analysis in the metaxpress consortium. PLoS One 7: e50938, 2012.
30. HUGO Gene Nomenclature Committee. HGNC [Online]. http://www.genenames.org/.
31. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing: 2014.
32. Bates D, Mächler M, Bolker B. Fitting linear mixed-effects models using lme4. J Stat Softw … submitted: 51, 2012.
33. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26: 2190–1, 2010.
34. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B.
35. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 33: W741–8, 2005.
36. Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res 41: W77–83, 2013.
37. Lee S-J, Lee Y-S, Zimmers TA, Soleimani A, Matzuk MM, Tsuchida K, Cohn RD, Barton ER. Regulation of Muscle Mass by Follistatin and Activins. Mol Endocrinol 24: 1998–2008, 2010.
38. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. Trends Genet 13: 163, 1997.
39. Mitchell WK, Williams J, Atherton P, Larvin M, Lund J, Narici M. Sarcopenia, dynapenia, and the impact of advancing age on human skeletal muscle size and strength; a quantitative review. Front Physiol 3: 260, 2012.
40. Paoni NF, Peale F, Wang F, Errett-Baroncini C, Steinmetz H, Toy K, Bai W, Williams PM, Bunting S, Gerritsen ME, Powell-Braxton L. Time course of skeletal muscle repair and gene expression following acute hind limb ischemia in mice. Physiol Genomics 11: 263–272, 2002.
41. Sharples AP, Al-Shanti N, Hughes DC, Lewis MP, Stewart CE. The role of insulin-like-growth factor binding protein 2 (IGFBP2) and phosphatase and tensin homologue (PTEN) in the regulation of myoblast differentiation and hypertrophy. Growth Horm IGF Res 23: 53–61, 2013.
42. Hindorff L, MacArthur J, Wise A, Junkins H, Hall P, Klemm A, Manolio T. A Catalog of Published Genome-Wide Association Studies [Online]. [date unknown]. www.genome.gov/gwastudies.
43. Siritantikorn A, Johansson K, Ahlen K, Rinaldi R, Suthiphongchai T, Wilairat P, Morgenstern R. Protection of cells from oxidative stress by microsomal glutathione transferase 1. Biochem Biophys Res Commun 355: 592–6, 2007.
44. Cortez D, Glick G, Elledge SJ. Minichromosome maintenance proteins are direct targets of the ATM and ATR checkpoint kinases. Proc Natl Acad Sci U S A 101: 10078–83, 2004.

45. Qiang X, Yang W-L, Wu R, Zhou M, Jacob A, Dong W, Kuncewitch M, Ji Y, Yang H, Wang H, Fujita J, Nicastro J, Coppa GF, Tracey KJ, Wang P. Cold-inducible RNA-binding protein (CIRP) triggers inflammatory responses in hemorrhagic shock and sepsis. Nat Med 19: 1489–95, 2013.

46. Gupta VA, Beggs AH. Kelch proteins: emerging roles in skeletal muscle development and diseases. Skelet Muscle 4: 11, 2014.

47. Von Maltzahn J, Chang NC, Bentzinger CF, Rudnicki MA. Wnt signaling in myogenesis. Trends Cell Biol 22: 602–9, 2012.

48. Harries LW, Hernandez D, Henley W, Wood AR, Holly AC, Bradley-Smith RM, et al. Human aging is characterized by focused changes in gene expression and deregulation of alternative splicing. Aging Cell 10: 868–78, 2011.

49. Cesari M, Penninx BWJH, Lauretani F, Russo CR, Carter C, Bandinelli S, et al. Hemoglobin levels and skeletal muscle: results from the InCHIANTI study. J Gerontol A Biol Sci Med Sci 59: 249–54, 2004.

50. Thiery J, Keefe D, Boulant S, Boucrot E, Walch M, Martinvalet D, Goping IS, Bleackley RC, Kirchhausen T, Lieberman J. Perforin pores in the endosomal membrane trigger the release of endocytosed granzyme B into the cytosol of target cells. Nat Immunol 12: 770–7, 2011.

51. Barbieri M, Ferrucci L, Ragno E, Corsi A, Bandinelli S, Bonafè M, et al. Chronic inflammation and the effect of IGF-I on muscle strength and power in older persons. Am J Physiol Endocrinol Metab 284: E481–E487, 2003.

52. Han J, Mistriotis P, Lei P, Wang D, Liu S, Andreadis ST. Nanog reverses the effects of organismal aging on mesenchymal stem cell proliferation and myogenic differentiation potential. Stem Cells 30: 2746–59, 2012.

53. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J, Jonsdottir GA, Ruotti V, Stewart R, Slukvin II, Thomson JA. Induced pluripotent stem cell lines derived from human somatic cells. Science 318: 1917–20, 2007.

54. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 42: D756–63, 2014.

55. Khan AA, Quigley JG. Control of intracellular heme levels: heme transporters and heme oxygenases. Biochim Biophys Acta 1813: 668–82, 2011.

56. Shiratori AP, Iop R da R, Júnior NGB, Domenech SC, Gevaerd M da S. Evaluation protocols of hand grip strength in individuals with rheumatoid arthritis: A systematic review. Rev. Bras. Reumatol. 54: 140–147, 2014.

CHAPTER 3

Combining genetic & genomic approaches

CHAPTER 3.1

Systematic identification of *trans*-eQTLs as putative drivers of known disease associations

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

** These authors contributed equally to this work*

Published: *Nat Genet 2013;* **45***(10)*: *doi: 10.1038/ng.2756*

ABSTRACT

Identifying the downstream effects of disease-associated single nucleotide polymorphisms (SNPs) is challenging: the causal gene is often unknown or it is unclear how the SNP affects the causal gene, making it difficult to design experiments that reveal functional consequences. To help overcome this problem, we performed the largest expression quantitative trait locus (eQTL) meta-analysis so far reported in non-transformed peripheral blood samples of 5,311 individuals, with replication in 2,775 individuals. We identified and replicated *trans*-eQTLs for 233 SNPs (reflecting 103 independent loci) that were previously associated with complex traits at genome-wide significance. Although we did not study specific patient cohorts, we identified trait-associated SNPs that affect multiple *trans*-genes that are known to be markedly altered in patients: for example, systemic lupus erythematosus (SLE) SNP rs4917014 [1] altered C1QB and five type 1 interferon response genes, both hallmarks of SLE [2-4]. Subsequent ChIP-seq data analysis on these *trans*-genes implicated transcription factor IKZF1 as the causal gene at this locus, with DeepSAGE RNA-sequencing revealing that rs4917014 strongly alters 3' UTR levels of IKZF1. Variants associated with cholesterol metabolism and type 1 diabetes showed similar phenomena, indicating that large-scale eQTL mapping provides insight into the downstream effects of many trait-associated variants.

Genome-wide association studies (GWAS) have identified thousands of variants that are associated with complex traits and diseases. However, because most variants and their proxies are non-coding, it is generally difficult to identify the causal genes. Recently, several eQTL-mapping studies [5-8] have now shown that the majority of disease-predisposing variants actually affect gene expression levels of nearby genes (i.e. *cis*-eQTLs). A few recent studies have also identified *trans*-eQTLs [5,9-13], revealing the downstream consequences of some variants. However, the total number of reported *trans*-eQTLs is fairly low, mainly due to the severe burden of multiple testing. To improve statistical power, we performed an eQTL meta-analysis in 5,311 peripheral blood samples, from seven studies using Illumina gene expression arrays (EGCUT [14], InCHIANTI [15], Rotterdam Study [16], Fehrmann [5], HVH [17-19], SHIP-TREND [20], and DILGOM [21]) and replication analysis in another 2,775 samples. We aimed to ascertain to what extent SNPs affect genes in *cis* and *trans* and whether eQTL mapping in peripheral blood could reveal important downstream pathways that may be putative drivers of disease processes.

Our genome-wide analysis identified *cis*-eQTLs for 44% of all tested genes (6,418 genes at probe-level false discovery rate (FDR)<0.05 and 4,690 genes with more stringent Bonferroni multiple testing correction, Table 1, Supplementary Table 1, Supplementary Figures 1-3). Our *trans*-eQTL analysis focused on 4,542 SNPs that have been implicated in complex disease or traits (derived from the "Catalog of Published GWAS"). In the discovery dataset, we detected *trans*-eQTLs at FDR<0.05 for 1,513 significant *trans*-eQTLs that include 346 unique SNPs (8% of all tested SNPs, Table 1, Supplementary Table 2, Supplementary Figure 4 and 5). These SNPs affect the expression of 430 different genes (a more stringent Bonferroni correction revealed 643 significant *trans*-eQTLs, including 200 unique SNPs and 223 different genes).

We used stringent procedures for *trans*-eQTL detection (Supplementary methods), and various benchmarks to ensure reliability: for 26 *trans*-eQTL genes the eQTL SNP affected multiple probes within these genes (Supplementary Table 3), always with consistent allelic directions, suggesting that our probe filtering procedure was effective in preventing false-positive *trans*-eQTLs. We observed uniform directionality for 90% of the tested *trans*-eQTLs across all studies within our discovery meta-analysis (Supplementary Figure 5). We did not find evidence that the eQTLs were driven by differences in age or blood cell-counts between individuals (Supplementary Results and Supplementary Table 4, Supplementary Figure 6). However, we cannot exclude this possibility entirely because FACS analyses on individual cell-types had not been conducted.

To ensure reproducibility of the *trans*-eQTLs of our current meta-analysis, we performed various analyses. We replicated previously reported blood *trans*-eQTLs [5] (Supplementary Table 5, Supplementary Results and Supplementary Figure 7) and replicated *trans*-eQTLs from our discovery meta-analysis in two independent studies of peripheral blood gene expression (52% in KORA F4 [22], N=740 samples and 79% in BSGS [23], N=862 samples, FDR<0.05, Supplementary Figure 8). Irrespective of significance, 91% and 93% of all 1,513 significant *trans*-eQTL SNP-probe combinations showed consistent allelic direction in these replication cohorts as compared to the discovery analysis.

A meta-analysis of these two replication studies improved replication rates: 89% of the 1,513 *trans*-eQTLs were significantly replicated (FDR<0.05), 99.7% of which showed a consistent allelic direction. Irrespective of significance, 97% of the *trans*-eQTLs showed a consistent allelic direction in this replication meta-analysis (Supplementary Figure 8).

Table 1. Results of the *cis*- and *trans*-eQTL mapping analyses.

Summary statistics	***Cis*-eQTLs**	***Trans*-eQTLs**
Number of SNPs tested that pass QC	1,962,237	4,542 (of which 2,082 are associated with complex traits at genome-wide significance, $P<5 \times 10^{-8}$)
Number of probes tested that pass QC	29,891	34,061
Number of genes tested	14,542	16,332
Number of probes not mapping to genes	9,260	18,018
Number of statistical tests performed	11,172,453	153,134,630

Significance thresholds	***Cis*-eQTLs**		***Trans*-eQTLs**	
	Meta-analysis Z-score	**Meta-analysis P-value**	**Meta-analysis Z-score**	**Meta-analysis P-value**
FDR<0.05 significance	3.824	1.31×10^{-4}	5.022	5.12×10^{-7}
Bonferroni significance	5.867	4.5×10^{-9}	6.287	3.3×10^{-10}

***cis*-eQTL analysis**	**FDR<0.05 significance**	**Bonferroni significance**
Number of significant unique SNP-Probe pairs	664,097	395,543
Number of significant unique eQTL SNPs	397,310	266,036
Number of significant unique eQTL probes	8,228	5,738
Number of significant unique eQTL genes	6,418	4,690
Number of significant unique eQTL probes not mapping to genes	636	326

***Trans*-eQTL analysis**	**FDR<0.05 significance**	**Bonferroni significance**
Number of significant unique SNP-Probe pairs	1,513	643
Number of significant unique eQTL SNPs	346	200
Number of significant unique eQTL probes	494	240
Number of significant unique eQTL genes	430	223
Number of significant unique eQTL probes not mapping to genes	35	13

We found that some *trans*-eQTLs could also be detected in three cell-type-specific datasets (283 monocyte samples [9], 282 B-cell samples [9] and 608 HapMap lymphoblastoid cell-line (LCL) samples [24]; Supplementary Figures 9 and 10). Despite the different tissue of these three studies, we were still

able to significantly replicate 7%, 4% and 2% of the *trans*-eQTLs (FDR<0.05), respectively. As 95% of the *trans*-eQTL SNPs explained less than 3% of the total expression variance (Supplementary Figure 11), we lack statistical power to replicate most *trans*-eQTLs in these smaller replication cohorts.

We subsequently confined further analyses to 2,082 different SNPs that have been found associated with complex traits at genome-wide significant levels ('trait-associated SNPs', reported $P<5 \times 10^{-8}$, out of 4,542 unique SNPs that we tested). These 2,082 SNPs showed a significantly higher number of *trans*-eQTL effects as compared to the 2,460 tested SNPs with reported disease associations at lower significance levels ($P=8 \times 10^{-22}$, Supplementary methods and results, Supplementary Figure 12): 254 of these 2,082 SNPs show a *trans*-eQTL effect in the discovery analysis (reflecting 1,340 SNP-probe combinations, of which we significantly replicated 1,201 SNP-probe combinations, reflecting 233 different SNPs and 103 independent loci in blood). For 671 out of these 1,340 *trans*-eQTLs (50%) the trait-associated SNP was either the strongest *trans*-eQTL SNP within the locus (or in strong LD with the strongest *trans*-eQTL SNP) or unlinked to the strongest *trans*-eQTL SNP (Supplementary results and Supplementary Table 6). We observed that the 2,082 trait-associated SNPs were six times more likely to cause *trans*-eQTL effects than randomly selected SNPs (matched for distance to gene and allele frequency, $P=5.6 \times 10^{-49}$, Supplementary methods and results, Supplementary Figure 13). SNPs, associated with (auto)immune or hematological traits were twice as likely to cause *trans*-eQTLs, as compared to other trait-associated SNPs ($P=5 \times 10^{-25}$, Supplementary methods and results). We observed that trait-associated SNPs that also cause *trans*-eQTLs more often affect the expression levels of nearby transcription factors in *cis*, as compared to trait-associated SNPs that do not affect genes in *trans* (Fisher's exact P=0.032; Supplementary results), suggesting that some of the *trans*-eQTLs arise due to altered *cis* gene expression levels of nearby transcription factors.

We also examined genomic SNP properties of the *trans*-eQTLs: these SNPs (and their perfect proxies based on data from the 1000 Genomes Project [25-26]) are significantly enriched (Fisher's exact P<0.05) for mapping within miRNA binding sites (Figure 1A). They map to regions showing strong enrichment (fold-change>2.5) of histone enhancer signals in K562 (myeloid) and GM12878 (lymphoid) cell-lines (Figure 1B), when compared to six non-blood cell-lines. This myeloid and lymphoid enhancer enrichment supports the validity of our blood-derived *trans*-eQTLs. These enrichment results suggest tissue specificity, which is supported by our inability to replicate a strong *trans*-eQTL that was previously identified in adipose tissue for SNP rs4731702 [13] that is associated with both type 2 diabetes and lipid levels.

These *trans*-eQTLs can provide insight into the pathogenesis of disease. Although RNA microarray studies have revealed dysregulated pathways for many complex diseases, it is often unclear what comes first: whether the associated SNPs first cause defects in the pathways whose dysregulation ultimately leads to disease, or whether the SNPs first cause disease that then perturbs these pathways. One example is SLE, an auto-immune disease resulting in inflammation and tissue damage.

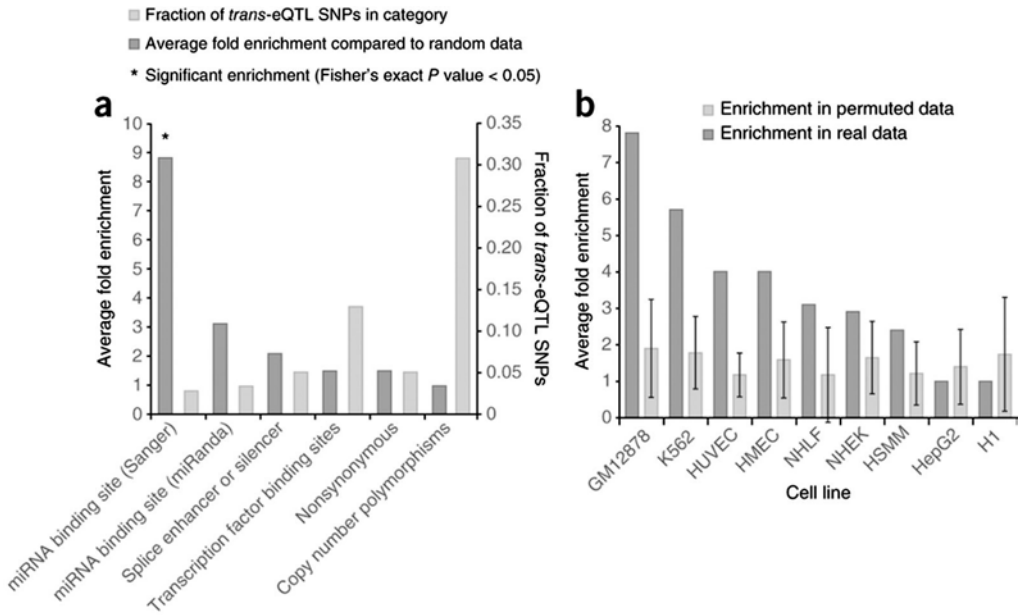


Figure 1. *Trans*-eQTL SNPs are enriched for functional elements. We investigated whether the *trans*-eQTL SNPs are enriched for certain functional elements. We used the online tools SNPInfo, SNPNexus, and HaploReg that rely upon data from, amongst others, the ENCODE project. **(A)** We observed that *trans*-eQTL SNPs are enriched for mapping within miRNA binding sites **(B)** *trans*-eQTL SNPs show strong enrichment (as annotated using HaploReg) for enhancer regions that are present in K562 (myeloid) and GM12878 (lymphoid) cell-lines (error bars represent one standard deviation).

It is known that SLE patients show markedly increased type 1 interferon (IFN-α) levels, increased expression of IFN-α response genes [4,27-28] and decreased complement *C1q* expression. We observed that four common SLE associated variants do indeed affect IFN-α response genes in *cis* (*IRF5*, *IRF7*, *TAP2* and *PSMB9*;Supplementary Table 1). However, as most SLE-associated SNPs do not map near complement or IFN-α response genes, we assessed whether these SNPs might affect complement or IFN-α response genes in *trans*. This was the case for rs4917014, for which the SLE risk allele (rs4917014*T, showing genome-wide significance in Asian populations and nominal significance in European populations [1,24]) not only increased expression of five different IFN-α response genes (*HERC5, IFI6, IFIT1, MX1* and *TNFRSF21*; Figure 2), but also decreased expression of three different probes in *CLEC10A*. In addition, we observed a nominal significant association of rs4917014*T with decreased expression of C1QB ($P=5.2 \times 10^{-6}$, FDR=0.28), a subunit of the first component of complement C1q, which has an established protective role in lupus. The complete deletion of C1q practically assures the development of SLE [29,30]. *CLEC10A* and *CLEC4C* belong to the C-type lectin family, which also includes mannose-binding lectins (MBL). While, to our knowledge, *CLEC10A* and *CLEC4C* have not been studied in the context of SLE, the role of MBL is

similar to C1q and is a risk factor for the development of autoimmunity in both humans and mice [3]. The rs4917014 *trans*-eQTLs were well replicated in the peripheral blood and monocyte replication datasets and reinforce the role of altered IFN-α mediated pathway, C-type lectin and *C1q* gene expression in SLE. In addition, people who do not have SLE, but who carry the rs497014*T risk allele already show these pathway alterations, which indicates these affected pathways are not solely a consequence of SLE, but could well precede SLE onset.

We next investigated the underlying mechanisms of the effects exerted by rs4917014. *IKZF1* is the only gene residing within the rs4917014 locus. Being a transcription factor (Ikaros family zinc finger 1), *cis*-regulatory effects of rs4917014 on *IKZF1,* that would translate in altered IKZF1 protein levels, could provide a working mechanism for the detected *trans*-eQTLs. However, since our meta-analysis initially did not show a *cis*-eQTL on the Illumina probe for *IKZF1* that is located near the 5' untranslated region (UTR) of *IKZF1*, we investigated the 3'-UTR by using DeepSAGE next-generation RNA-sequencing data of 94 peripheral blood samples. The variant rs4917014*T strongly increased the 3'-UTR expression levels of *IKZF1* (Spearman correlation=0.45, $P=6.29 \times 10^{-6}$, Zhernakova et al, submitted). We then used ChIP-seq data from the ENCODE-project [31] and observed significantly increased IKZF1 protein binding to the genomic DNA locations where the upregulated *trans*-eQTL genes map (Wilcoxon P-value=0.046), compared to IKZF1 binding to all other genic DNA. We also observed increased IKZF1 binding to the other SLE *cis*-genes outside of the *IKZF1* locus (Wilcoxon P-value=4.3×10^{-4}), thereby confirming the importance of *IKZF1* in SLE. *IKZF1* is important for other phenotypes as well: another, unlinked intronic variant within *IKZF1*, rs12718597, is associated with mean corpuscular volume (MCV) [32] and affects the 5' end of *IKZF1* in *cis*. As *IKZF1* knock-out mice show abnormal erythropoiesis [33], this suggests a causal role for *IKZF1* in MCV as well. However, although rs12718597*A increases expression of 31 *trans*-genes and decreases expression of another 19 *trans*-genes, none of the SLE *trans*-genes overlap the MCV *trans*-genes. The latter are mainly involved in hemoglobin metabolism and do not show an increased IKZF1-binding signal, Wilcoxon P=0.35. In summary, these results indicate that *IKZF1* has multiple functions and that different SNPs near *IKZF1* elicit function-specific effects.

We identified other *trans*-eQTLs showing similar phenomena: we observed that rs174546 (located in the 3'-UTR of *FADS1*, and associated with metabolic syndrome [34], LDL and total cholesterol levels [35,36]) affects *C11orf10*, *FADS1* and *FADS2* in *cis* and *LDLR* in *trans*. *LDLR* encodes the LDL receptor and contains common variants that are also associated with lipid levels [36] (Figure 3). *LDLR* gene expression levels correlated negatively ($P<3.0 \times 10^{-4}$) with total, HDL and LDL cholesterol levels in the tested cohorts (Rotterdam Study and EGCUT, Supplementary Table 7), indicating that peripheral blood is a useful tissue for gaining downstream insight into the effects of lipid SNPs.

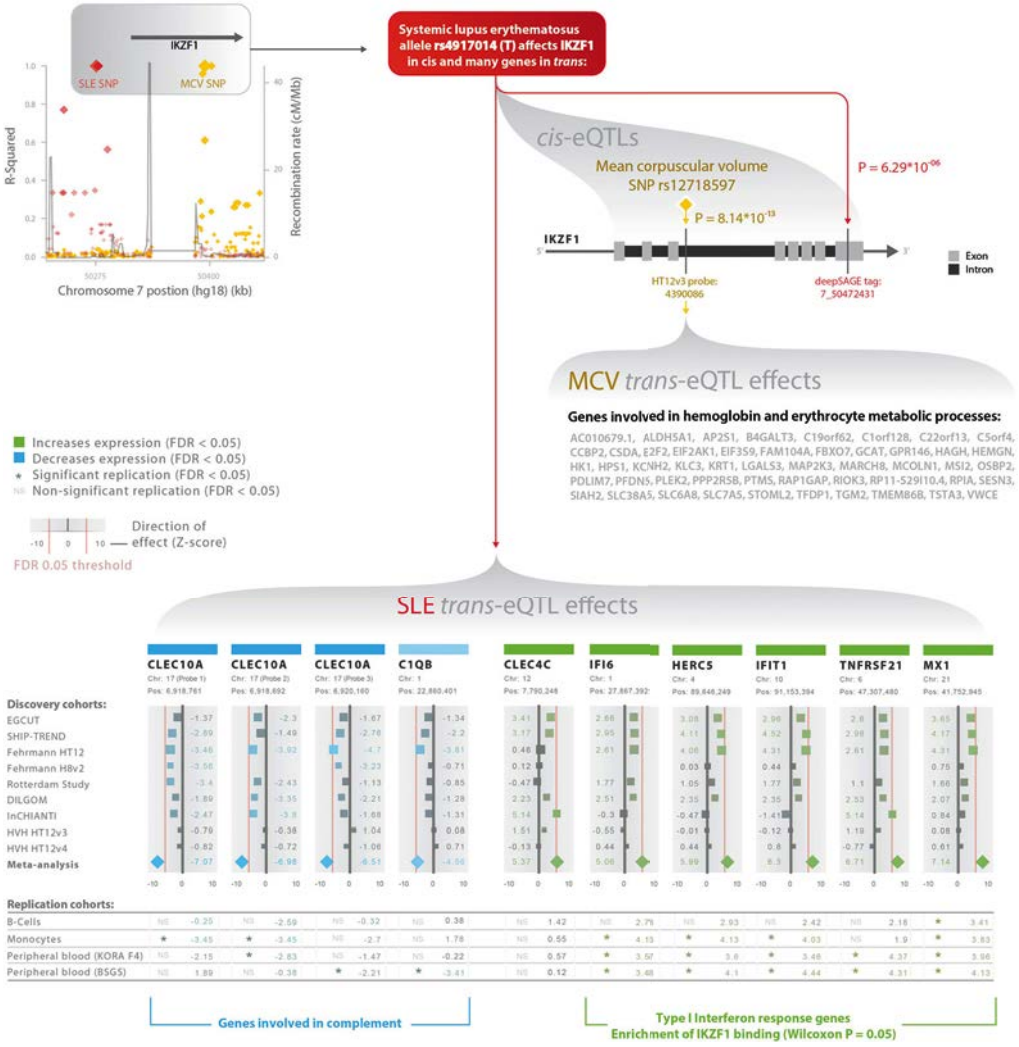


Figure 2. Independent *trans*-eQTL effects emanating from the IKZF1 locus. Systemic lupus erythematosus SNP rs4917014 and unlinked mean corpuscular volume SNP rs4917014 both affect expression of IKZF1 in *cis*. rs12718597 affects 50 *trans*-genes (mostly involved in hemoglobin metabolism) while rs4917014 affects eight different genes in *trans*: the rs4917014*T risk allele increases expression of genes involved in type I interferon response. At a somewhat lower significance threshold of FDR 0.28 rs4917014*T decreases complement C1QB expression. Both processes are hallmark features of SLE.

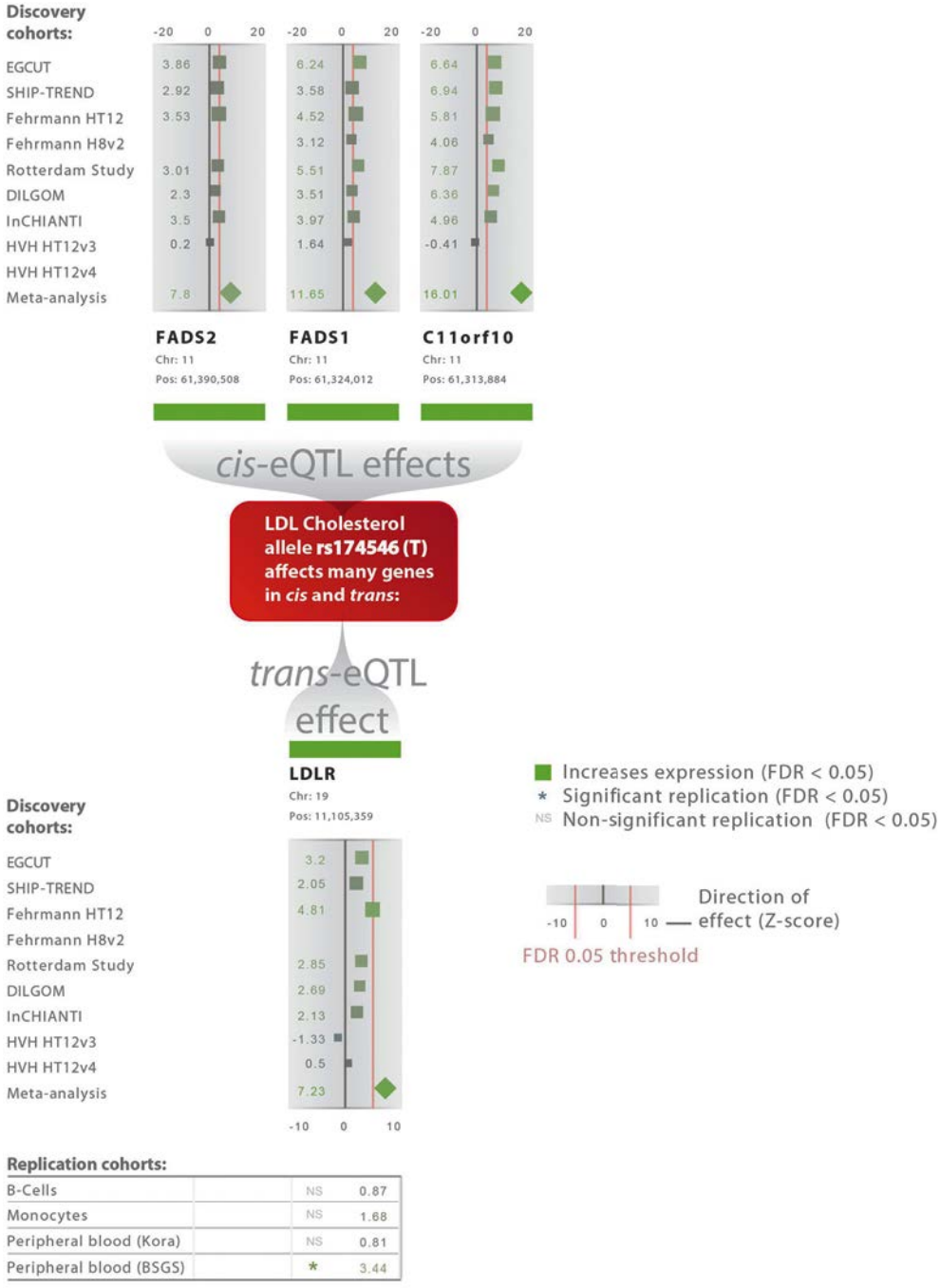
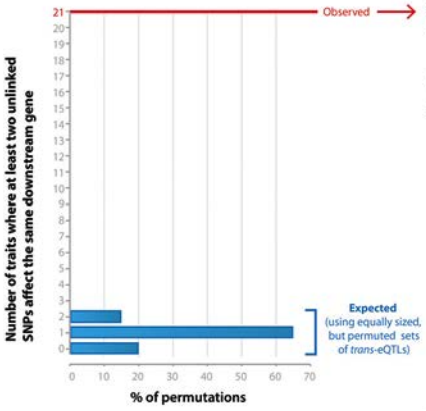


Figure 3. Cholesterol SNP rs174546 affects LDLR in *trans*. The rs174546*T allele is known to be associated with a decrease in serum LDL cholesterol and triglycerides levels. It increases the expression levels of three genes in *cis*, but also increases gene expression levels of LDLR that encodes the LDL receptor.

For 21 different complex traits, we found that at least two unlinked variants that are associated with these diseases, affected exactly the same gene in *trans*. When taking an equally sized, but permuted list of *trans*-eQTLs we would on average find only one complex trait where two unlinked SNPs affected the same gene in *trans* (Figure 4, Supplementary Table 8, Methods). Although most of these traits are hematological (e.g. mean platelet volume or serum iron levels) we also observed this convergence for blood pressure, celiac disease, multiple sclerosis, and type 1 diabetes (T1D). rs3184504 (located in an exon of *SH2B3*) and its near-perfect proxy rs653178 (located in an intronic region of *ATXN2* on chromosome 12) are associated with several auto-immune diseases including T1D [37,38], T1D auto-antibodies [37,38], celiac disease [8,39], hyperthyroidism [40], vitiligo [41], rheumatoid arthritis [39] and other complex traits such as blood pressure [42,43], chronic kidney disease [44], and eosinophil counts [45].



Trait type	Complex trait	At least two unlinked trait SNPs both affect gene(s):
Immune related traits	Type 1 diabetes	*GBP4, STAT1*
	Type 1 diabetes autoantibodies	*GBP4, STAT1*
	Celiac disease	*CXCR6, FYCO1*
	Multiple sclerosis	*CD5*
Blood pressure traits	Diastolic blood pressure	*LOC338758*
	Systolic blood pressure	*LOC338758*
Hematological traits	Hemoglobin	*ALAS2*
	Hematological Parameters	*FBXO7*
	F-cell distribution	*ESPN, PHOSPHO1, GNAS, TSPAN13, VWCE,*
	Hematocrit	*ALAS2*
	Serum markers of iron status	*ALAS2*
	Red blood cell traits	*ALAS2*
	Serum iron levels	*ALAS2*
	Glycated hemoglobin levels	*ALAS2*
	Hematology traits	*ALAS2*
	Serum hepcidin	*ALAS2*
	Beta-thalassemia	*PHOSPHO1, VWCE, TSPAN13, ESPN*
	Hematological and biochemical traits	*AL109955.37-3, RBM38, TRIM58, DPM2*
	Mean corpuscular hemoglobin	*ALAS2, C18orf10, DNAJB2, ESPN, HBM, KEL, PDZK1IP1, PIM1, PRDX5, RAP1GAP, UBXN6, VWCE, XK*
	Mean corpuscular volume	*ESPN, VWCE, PPP2R5B, PHOSPHO1, HBM, MARCH8, CCBP2, C5orf4, PLEK2, HPS1, KRT1, C1orf128, EIF2AK1, KLC3, MCOLN1, RAP1GAP, RIOX3, CSDA, LGALS3, OSBP2, PTMS, HAGH, KEL, DNAJB2, PDZK1IP1, XK, PIM1, PRDX5, C18orf10, UBXN6, ALAS2, TGM2, FBXO7, C22orf13, B4GALT3, TSTA3,*
	Mean platelet volume	*CTTN, VCL, TSPAN9, CTDSPL, PARVB, GP9, GNG11, ITGB5, CABP5, SAMD14, C19orf33, GUCY1A3, NRGN, F13A1, LIMS1, MMRN1, ESAM, PTCRA, AL353716.18, MPL, C6orf152, SH3BGRL2, RBPMS2, LY6G6F, GNB5, PRDX6, ITGA2B, ABCC3, PTGS1, MYL9, AC105339.9, HOMER2, CXCL5, AQP10*

Figure 4. For 21 complex traits, pairs of unlinked trait-associated SNPs affect the same downstream genes. We observed that for 21 different traits, there were pairs of unlinked SNPs that have previously been reported to be associated with these traits and which also affect exactly the same downstream genes in *trans*, whereas this is rarely observed when using an equally sized, but permuted list of *trans*-eQTLs.

We observed a *cis*-eQTL on *SH2B3* (FDR<0.05) and fourteen *trans*-eQTL genes (FDR<0.05, Figure 5), all highly expressed in neutrophils. Since these *trans*-eQTLs could potentially appear due to the known effect of rs3184504 on differences in cell-count proportions [45], we correlated *trans*-gene expression levels with cell counts in two cohorts (the Rotterdam Study and EGCUT) but did not observe significant correlations (Supplementary Table 6). These fourteen *trans*-eQTLs describe different biological functions: T1D disease risk allele rs3184504*T decreases expression levels of nine genes, most of which are involved in toll-like receptor signaling [46] (C12orf75, FOS, IDS, IL8, LOC338758, NALP12, PPP1R15A, S100A10 and *TAGAP*) and increases expression of five genes involved in interferon-γ response (*GBP2, GBP4, STAT1, UBE2L6* and *UPP1)*. We observed that another

T1D risk allele, rs4788084*C [37,38] on chromosome 16, increases expression of *GBP4* and *STAT1* as well (Figure 5), revealing how different T1D risk alleles converge: they both cause an increase of interferon-γ response gene expression.

In summary, our eQTL meta-analysis revealed and replicated downstream effects for 233 trait-associated SNPs. We have highlighted only a few here and shown that *trans*-eQTL mapping in blood for lipid and immune-mediated disease variants yields downstream insight which is biologically meaningful. Our results on *IKZF1* show that the two unlinked SLE and MCV variants near this gene give strikingly different yet biologically meaningful *trans*-regulatory effects. Future, larger-scale *trans*-eQTL analysis in blood will likely uncover many more of these regulatory relationships.

METHODS

Study populations

We performed a whole-genome eQTL meta-analysis of 5,311 samples from peripheral blood, divided over a total of nine datasets from seven cohorts, including EGCUT [14] (N=891), InCHIANTI [15] (N=611), Rotterdam Study [16] (N=762), Fehrmann [5] (N=1,240 on the Illumina HT12v3 platform and N=229 on the Illumina H8v2 platform), HVH [17-19] (N=43 on the Illumina HT12v3 platform and N=63 on the Illumina HT12v4 platform) SHIP-TREND [20] (N=963), and DILGOM [21] (N=509). Gene expression data for each dataset was obtained using either PAXGene (Becton Dickinson) or Tempus tubes (Life Technologies), followed by hybridization to Illumina whole-genome Expression BeadChips (HT12v3, HT12v4 or H8v2 arrays). The gene expression platforms were harmonized by matching probe sequences across the different platforms. Mappings for these sequences were obtained by mapping the sequences against the human genome build 36 (Ensembl build 54, Hg18) using BLAT, BWA and SOAPv2 sequence alignment programs. Highly stringent alignment criteria were used to ensure that probes map unequivocally to one single genomic position. Genotype data was acquired using different genotyping platforms, and harmonized by imputation, using the HapMap2 [47] Central European population as a reference. Each dataset was individually checked for sample mix-ups *using MixupMapper* [48]. For a full description of the individual datasets, results of the sample mix-up analysis, specifics on the gene expression platforms and probe mapping procedure and filtering, see Supplementary methods.

Gene expression normalization

Gene expression data was quantile-normalized to the median distribution, and subsequently $\log_2$ transformed. The probe and sample means were centered to zero. Gene expression data was then corrected for possible population structure by removal of four multi-dimensional scaling components using linear regression. We reasoned earlier that normalized gene-expression data still contains large amounts of non-genetic variation5. After population stratification correction, principal component analysis (PCA) was therefore performed on the sample correlation matrix. We performed a separate QTL analysis for each principal component (PC), to ascertain whether

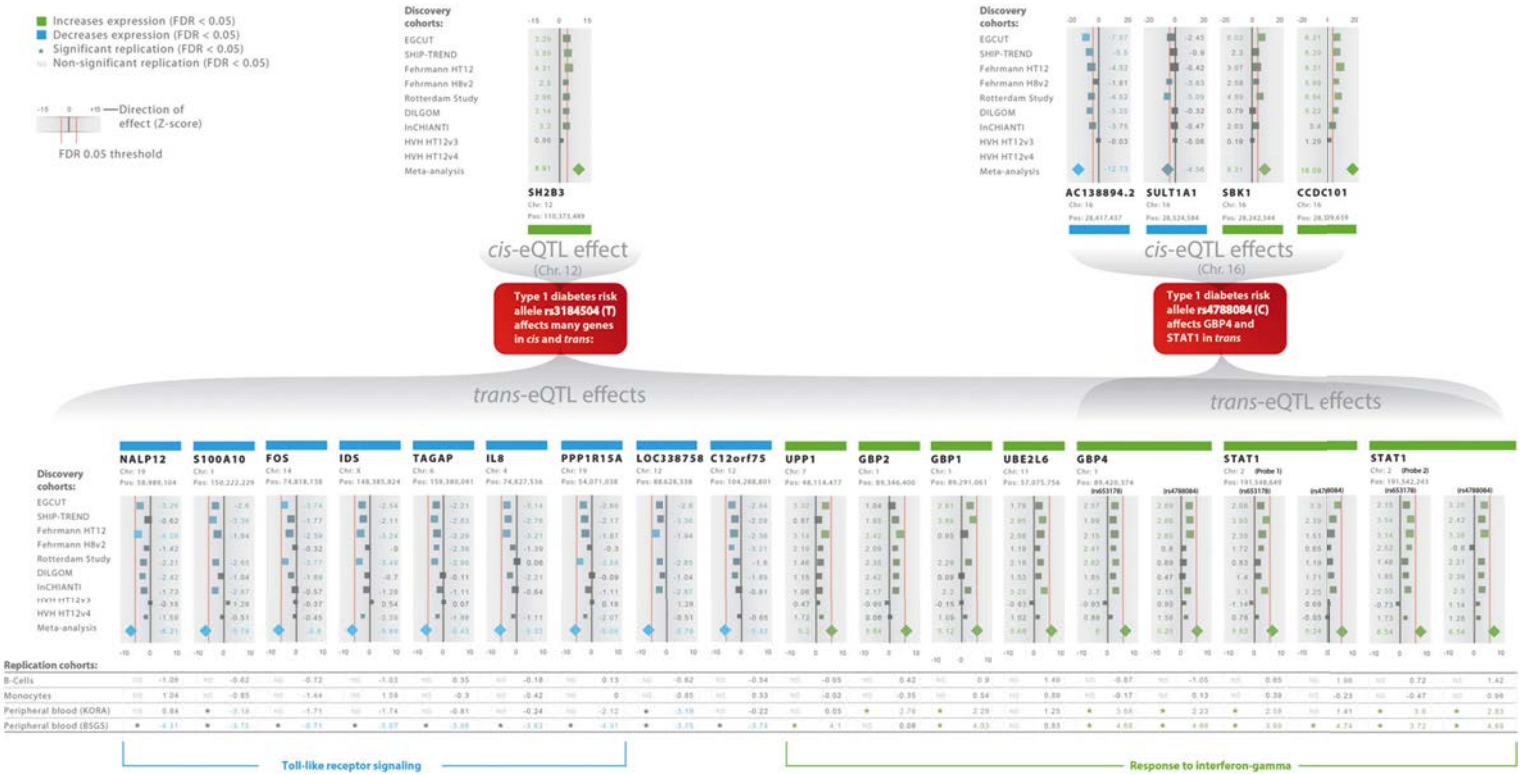


Figure 5. Two unlinked type-1 diabetes risk alleles both increase STAT1 and GBP4 expression. rs3184504*T, a risk allele for type 1 diabetes (chromosome 12), affects the expression of SH2B3 in *cis*, but also affects the expression levels of fourteen unique genes in *trans*, including two interferon-γ response genes GBP4 and STAT1. Another unlinked type-1 diabetes risk allele (rs4788084*C on chromosome 16) also increases expression levels of these two interferon-γ response genes, indicating that an elevated interferon-γ response is important in type 1 diabetes.

genetic variants could be detected that affect the PC. If we found an effect on the PC, we did not correct the expression data for these components, to ensure we would not unintentionally remove genetic effects from the expression data. Significance of these associations was established by controlling the false discovery rate (FDR), testing each association against a null-distribution created by repeating the analysis 100 times (permuting the sample labels for each iteration [49]). PCs that did not show significance at the FDR threshold of 0.0 were removed from the gene expression data by linear regression. In all but two very small datasets, the first 40 PCs were removed (excluding those components per cohort that showed a QTL effect). We observed that the removal of these 40 components revealed the highest number of eQTLs in each dataset. Although PC correction may remove some eQTL effects, we observed that the majority (95% when removing 35 PCs and 90% when removing 40 PCs) of *trans*-eQTL effects was independent of the number of PCs removed (Supplementary Figure 14).

eQTL mapping

After normalization of the data, we performed both *cis*- and *trans*-eQTL mapping. eQTLs were deemed *cis*-eQTLs, when the distance between the SNP chromosomal position and the probe midpoint was less than 250 kilobases (kb), while eQTLs with a distance greater than five megabases (mb) were defined as *trans*-eQTLs. Only SNPs with a minor allele-frequency (MAF)>0.05 and a Hardy-Weinberg equilibrium P-value>0.001 were included in the analyses. Since most cohorts had generated the gene expression data using the HT12v3 platform, we chose to only include probes that were present on this platform. We only tested SNP-probe pairs when the SNP passed quality control in at least three cohorts. Furthermore, in order to reduce issues with respect to computational time and multiple testing, we confined our *trans*-eQTL analysis to those SNPs present in the "*Catalog of Published GWAS*" (*http://www.genome.gov/gwastudies/*, accessed July 16th, 2011). We reasoned that for genes with strong *cis*-eQTL effects, the *cis*-eQTL effect may obscure the detectability of *trans*-eQTL. Therefore, we used linear regression to remove *cis*-eQTL effects prior to *trans*-eQTL mapping and observed a 12% increase in the number of detected *trans*-eQTLs (Supplementary Figure 15). For each cohort, eQTLs were mapped using a Spearman's rank correlation on the imputed genotype dosage values. We used a weighted Z-method for subsequent meta-analysis [50]. To get a realistic null-distribution, we permuted the sample identifiers labels of the expression data and repeated this analysis ten times (Supplementary Figure 16). In each permutation the sample labels were permuted. We then corrected for multiple testing by controlling the FDR at 0.05, by testing each p-value in the real data against a null-distribution created from the permuted datasets [49] (see Supplementary methods). It has been suggested that false-positive eQTL effects can arise due to polymorphisms in the probe sequences [51,52]. Therefore, we tested whether a significant *cis*-eQTL SNP was in LD (r^2>0.2) with any SNP in the *cis*-probe sequence, using the Western European subpopulations of the 1000 genomes project [25] (2011-05-21 release, 286 individuals, excluding Finnish individuals) as a reference. If we observed this to be the case the respective *cis*-eQTLs were removed. Furthermore, for each *trans*-eQTL we investigated whether portions of the probe sequence could map in the vicinity of the *trans*-eQTL SNP (which in fact would imply a *cis*-eQTL, rather than a *trans*-eQTL effect). Therefore, we tried to map the *trans*-eQTL probe sequences, using very permissive settings, within

a 5 mb window of the *trans*-eQTL SNP. SNP-probe combinations where the probe mapped with at least 15 bp within the 5 mb window, were deemed false-positive and removed from further analysis. After this filtering we recalculated the FDR for both the *cis*- and *trans*-eQTL results.

Trans-eQTL replication

Replication of the *trans*-eQTL results was carried out in five independent datasets from four cohorts, including data obtained from lymphoblastoid cell lines (HapMap3, N=604 [24]), B-cells and monocytes (Oxford [9], N=282 and N=283, respectively), and whole peripheral blood (KORA F4 [22], N=740, and BSGS [23], N=862). All the cohorts applied the same methodology as used in the discovery phase to normalize the gene expression data, check for sample mix-ups and perform *trans*-eQTL mapping, including 10 permutations in order to establish the FDR threshold at 0.05. Finally, we performed a sample-size weighted Z-score meta-analysis on the two peripheral blood replication cohorts (KORA and BSGS). Further details on these datasets can be found in the Supplementary methods.

Enhancer enrichment and functional annotation

To determine whether the significant *trans*-eQTL SNPs were enriched for functional regions on the genome, we annotated the *trans*-eQTL SNPs using SNPInfo [53], SNPNexus [54,55], and HaploReg [56], which integrate multiple data sources (such as ENCODE project data [31], Ensembl [57], and several micro-RNA databases). We limited these analyses to those *trans*-eQTL SNPs that were previously shown to be associated with complex traits at genome-wide significance levels ('trait associated SNPs', reported $P<5 \times 10^{-8}$). These SNPs were subsequently pruned (using PLINK's -clump command, using an $r^2<0.2$). We used the permuted *trans*-eQTL data to get realistic null-distributions for each of these tools: we selected equally sized sets of unlinked SNPs ($r^2<0.2$ in the Western-European subpopulations of the 1000 genomes project [25], 2011-05-21 release, 286 individuals, excluding Finnish individuals) that showed the highest significance in the permuted data, ensuring that only trait-associated SNPs are included in the null-distribution, as it is known that trait-associated SNPs in general already have different functional properties than randomly selected SNPs [58] (e.g. trait-associated SNPs typically map in closer proximity to genes than random SNPs). We also ensured that none of the SNPs in the null-distribution were affecting genes in *trans*, or were linked to those SNPs ($r^2<0.2$ in 1000 genomes). We then identified perfect proxies ($r^2=1.0$ in 1000 genomes). For SNPInfo and SNPNexus, we calculated the enrichment for each functional category using a Fisher's exact test. We determined the enhancer enrichment in nine different cell-types using HaploReg, where we averaged the enhancer enrichment over the ten permutations.

Convergence analysis

We determined which unlinked trait-associated SNPs show eQTL effects on exactly the same gene: per trait, we analyzed the SNPs that are known to be associated with this trait and assessed whether any unlinked SNP pair ($r^2<0.2$, distance between SNPs>5Mb) showed a *cis*- and/or *trans*-eQTL effect on exactly the same gene, as previously described [5]. To determine whether the number of traits for which we observed this phenomenon was higher than expected by chance, we re-ran this analysis

20 times, each time using a different set of permuted *trans*-eQTLs, equal in size to the non-permuted set of *trans*-eQTLs.

SLE IKZF1 ENCODE ChIP-seq Analysis

We used IKZF1 ChIP-seq signal data obtained from the ENCODE-project [31] (IKZF1 ChIP-seq data acquired and processed by UCSC, ENCODE March 2012 Freeze). For every human gene we determined the average signal (corrected for gene size), corrected for GC-content bias, and performed a Wilcoxon Mann-Whitney test to see whether the upregulated genes (*MX1, TNFRSF21, IFIT1/LIPA, HERC5, CLEC4C, IFI6*) showed a higher ChIP-seq signal compared to all other human genes.

Data availability

We have made a browser available for all significant *trans*-eQTL and *cis*-eQTL at *http://www.genenetwork.nl/bloodeqtlbrowser*. This browser also provides all *trans*-eQTLs that we detected at a somewhat less stringent false discovery rate of 0.5, to enable more in-depth *post-hoc* analyses.

ACKNOWLEDGMENTS

DILGOM

J.K. and S.R. were supported by funds from The European Community's Seventh Framework Programme (FP7/2007-2013) BioSHaRE, grant agreement 261433, S.R. was supported by funds from The European Community's Seventh Framework Programme (FP7/2007-2013) ENGAGE Consortium, grant agreement HEALTH-F4-2007- 201413", the Academy of Finland Center of Excellence in Complex Disease Genetics (grants 213506 and 129680), Academy of Finland (grant 251217), the Finnish foundation for Cardiovascular Research and the Sigrid Juselius Foundation. V.S. was supported by the Academy of Finland, grant number 139635 and Finnish Foundation for Cardiovascular Research. MP was partly financially supported for this work by the Finnish Academy SALVE program "Pubgensense" 129322 and by grants from the Finnish Foundation for Cardiovascular Research. The DILGOM-study was supported by the Academy of Finland, grant # 118065.

SHIP-TREND

SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Deutsche Forschungsgemeinschaft (DFG GRK840-D2), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network 'Greifswald Approach to Individualized Medicine (GANI_MED)' funded by the Federal Ministry of Education and Research (grant 03IS2061A). Genome-wide data have been supported by the Federal Ministry of Education and Research (grant no. 03ZIK012) and a joint grant from Siemens Healthcare, Erlangen, Germany and the Federal State of Mecklenburg, West Pomerania. Whole-body MR imaging was supported by a joint grant from Siemens Healthcare, Erlangen, Germany and the Federal State of Mecklenburg West Pomerania. The University of Greifswald is a member of the

'Center of Knowledge Interchange' program of the Siemens AG and the Caché Campus program of the InterSystems GmbH. The SHIP authors thank Mario Stanke for the opportunity to use his Server Cluster for the SNP imputation.

EGCUT

EGCUT received financing by FP7 grants (201413, 245536), also received targeted financing from the Estonian Government (SF0180142s08) and direct funding from the Ministries of Research and Science and Social Affairs. EGCUT studies are funded by the University of Tartu in the framework of the Center of Translational Genomics and by the European Union through the European Regional Development Fund, in the framework of the Centre of Excellence in Genomics. We thank EGCUT personnel, especially Ms. M. Hass and Mr V. Soo. EGCUT data analyses were carried out in part in the High Performance Computing Center of the University of Tartu.

HVH

HVH was supported in part by grants R01 HL085251 and R01 HL073410 from the National Heart, Lung, and Blood Institute (NHLBI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NHLBI or the National Institutes of Health, USA. Dr. Psaty serves on a DSMB for a clinical trial of a device funded by Zoll LifeCor and on the Steering Committee for the Yale Open Data Access Project funded by Medtronic. Both activities are unrelated to this work.

Rotterdam Study

We thank Pascal Arp, Mila Jhamai, Marijn Verkerk, Lizbeth Herrera, and Marjolein Peters for their help in creating the GWAS database; Karol Estrada and Maksim Struchalin for their support in creation and analysis of imputed data; Tobias A. Knoch, Anis Abuseiris, Karol Estrada, and Rob de Graaf as well as their institutions, the Erasmus GRID Office, Erasmus MC Rotterdam, The Netherlands, and especially the national German MediGRID and Services@MediGRID part of the German D-Grid, both funded by the German Bundesministerium fuer Forschung und Technology under grants #01 AK 803 A-H and # 01 IG 07015 G for access to their grid resources. The authors thank the study participants and staff from the Rotterdam Study, the participating general practitioners and the pharma*cist*s. The Rotterdam Study was funded by the European Commission (HEALTH-F2-2008-201865, GEFOS; HEALTH-F2-2008 35627, TREAT-OA 200800), Netherlands Organisation of Scientific Research NWO Investments (nos 175.010.2005.011, 911-03-012), the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Netherlands Genomics Initiative (NGI)/Netherlands Consortium for Healthy Aging (NCHA) (project nr. 050-060-810), an NWO Vidi grant (#917103521).The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organisation for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam.

Fehrmann

L.F.,H-J.W.: This study was supported by grants from the Celiac Disease Consortium (an innovative cluster approved by the Netherlands Genomics Initiative and partly funded by the Dutch Government (grant BSIK03009), Netherlands Organisation for Scientific Research (NWO-Vici grant 918.66.620, NWO-Veni grant 916.10.135 to L.F.), the Dutch Digestive Disease Foundation (MLDS WO11-30), and a Horizon Breakthrough grant from the Netherlands Genomics Initiative (grant 92519031 to L.F.). This project was supported by the Prinses Beatrix Fonds, VSB fonds, H. Kersten and M. Kersten (Kersten Foundation), The Netherlands ALS Foundation, and J.R. van Dijk and the Adessium Foundation. The research leading to these results has received funding from the European Community's Health Seventh Framework Programme (FP7/2007-2013) under grant agreement 259867. We especially thank *Cis*ca Wijmenga for helpful comments and support and Jackie Senior for critically reading the manuscript.

InCHIANTI

InCHIANTI was supported by the Wellcome Trust 083270/Z/07/Z. The InCHIANTI study was supported by contract funding from the U.S. National Institute on Aging (NIA), and the research was supported in part by the Intramural Research Program, NIA, and National Institute of Health (NIH). A.R.W. was supported by the Peninsula NIHR Clinical Research Facility. Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

KORA F4

The KORA authors acknowledge the contributions of Peter Lichtner, Gertrud Eckstein, Guido Fischer, Norman Klopp, Nicole Spada, and all members of the Helmholtz Zentrum München genotyping staff for generating the SNP data and Katja Junghans and Anne Löschner (Helmholtz Zentrum München) for generating gene expression data from both KORA and SHIP-TREND samples.The KORA research platform and the KORA Augsburg studies are financed by the Helmholtz Zentrum München, German Research Center for Environmental Health, which is funded by the BMBF and by the State of Bavaria. We thank the field staff in Augsburg who was involved in the studies. The German Diabetes Center is funded by the German Federal Ministry of Health and the Ministry of School, Science and Research of the State of North-Rhine-Westphalia. The Diabetes Cohort Study was funded by a German Research Foundation project grant to W.R. (DFG; RA 459/2-1). This study was supported in part by a grant from the BMBF to the German Center for Diabetes Research (DZD e.V.), by the DZHK (Deutsches Zentrum für Herz-Kreislauf-Forschung – German Centre for Cardiovascular Research) and by the BMBF funded Systems Biology of Metabotypes grant (SysMBo#0315494A). Additional support was given by the BMBF (National Genome Research Network NGFNplus Atherogenomics, 01GS0834) and the Leibniz Association (WGL Pakt für Forschung und Innovation). We thank Maren Carstensen, Gabi Gornitzka and Astrid Hoffmann (German Diabetes Center) for excellent technical assistance.

Oxford cell-specific

This work was supported by the Wellcome Trust (Grants 074318 [J.C.K.], 088891 [B.P.F.], and 075491/Z/04 [core facilities Wellcome Trust Centre for Human Genetics]), the European Research

Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement no. 281824 (J.C.K.) and the NIHR Oxford Biomedical Research Centre.
BSGS: This research was supported by Australian National Health and Medical Research Council (NHMRC) grants 389892, 496667, 613601, 1010374 and 1046880 and National Institutes of Health (NIH) grant GM057091. G.W.M. and P.V. are supported by the NHMRC Fellowship Scheme. J.E.P. is supported by the Australian Research Council (ARC) fellowship scheme. The funders had no role in the study design, data collection or analysis, de*cisi*on to publish, or preparation of the manuscript.

ADDITIONAL INFORMATION

Supplementary Information accompanies this paper at *http://www.nature.com/naturegenetics.*

REFERENCES

1. Han JW, et al. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat Genet. 2009;41:1234–7.
2. Bengtsson AA, et al. Activation of type I interferon system in systemic lupus erythematosus correlates with disease activity but not with antiretroviral antibodies. Lupus. 2000;9:664–71.
3. Bohlson SS, Fraser DA, Tenner AJ. Complement proteins C1q and MBL are pattern recognition molecules that signal immediate and long-term protective immune functions. Mol Immunol. 2007;44:33–43.
4. Ytterberg SR, Schnitzer TJ. Serum interferon levels in patients with systemic lupus erythematosus. Arthritis Rheum. 1982;25:401–6.
5. Fehrmann RS, et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. 2011;7:e1002197.
6. Nicolae DL, et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6:e1000888.
7. Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010;464:768–72.
8. Dubois PC, et al. Multiple common variants for celiac disease influencing immune gene expression. Nat Genet. 2010;42:295–302.
9. Fairfax BP, et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat Genet. 2012;44:502–10.
10. Innocenti F, et al. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. PLoS Genet. 2011;7:e1002078.
11. Grundberg E, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet. 2012;44:1084–9.
12. Heinig M, et al. A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. Nature. 2010;467:460–4.
13. Small KS, et al. Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nat Genet. 2011;43:561–4.
14. Metspalu A. The Estonian Genome Project. Drug Development Research. 2004;62:97–101.
15. Tanaka T, et al. Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. PLoS Genet. 2009;5:e1000338.
16. Hofman A, et al. The Rotterdam Study: 2012 objectives and design update. Eur J Epidemiol. 2011;26:657–86.
17. Heckbert SR, et al. Antihypertensive treatment with ACE inhibitors or beta-blockers and risk of incident atrial fibrillation in a general hypertensive population. Am J Hypertens. 2009;22:538–44.
18. Psaty BM, et al. The risk of myocardial infarction associated with antihypertensive drug therapies. JAMA. 1995;274:620–5.
19. Smith NL, et al. Esterified estrogens and conjugated equine estrogens and the risk of venous thrombosis. JAMA. 2004;292:1581–7.
20. Teumer A, et al. Genome-wide association study identifies four genetic loci associated with thyroid volume and goiter risk. Am J Hum Genet. 2011;88:664–73.
21. Inouye M, et al. An immune response network associated with blood lipid levels. PLoS Genet. 2010;6
22. Mehta D, et al. Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. Eur J Hum Genet. 2012
23. Powell JE, et al. The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics. PLoS One. 2012;7:e35430.
24. Wang C, et al. Genes identified in Asian SLE GWASs are also associated with SLE in Caucasian populations. Eur J Hum Genet. 2012
25. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73.
26. Patterson K. 1000 genomes: a world of variation. Circ Res. 2011;108:534–6.

27. Baechler EC, et al. Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. Proc Natl Acad Sci U S A. 2003;100:2610–5.
28. Bennett L, et al. Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. J Exp Med. 2003;197:711–23.
29. McAdam RA, Goundis D, Reid KB. A homozygous point mutation results in a stop codon in the C1q B-chain of a C1q-deficient individual. Immunogenetics. 1988;27:259–64.
30. Botto M, et al. Homozygous C1q deficiency causes glomerulonephritis associated with multiple apoptotic bodies. Nat Genet. 1998;19:56–9.
31. Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
32. Ganesh SK, et al. Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nat Genet. 2009;41:1191–8.
33. Wang JH, et al. Selective defects in the development of the fetal and adult lymphoid system in mice with an Ikaros null mutation. Immunity. 1996;5:537–49.
34. Zabaneh D, Balding DJ. A genome-wide association study of the metabolic syndrome in Indian Asian men. PLoS One. 2010;5:e11961.
35. Sabatti C, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet. 2009;41:35–46.
36. Teslovich TM, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010;466:707–13.
37. Barrett JC, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet. 2009;41:703–7.
38. Plagnol V, et al. Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. PLoS Genet. 2011;7:e1002216.
39. Zhernakova A, et al. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. PLoS Genet. 2011;7:e1002004.
40. Eriksson N, et al. Novel associations for hypothyroidism include known autoimmune risk loci. PLoS One. 2012;7:e34442.
41. Jin Y, et al. Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. Nat Genet. 2012;44:676–80.
42. Newton-Cheh C, et al. Genome-wide association study identifies eight loci associated with blood pressure. Nat Genet. 2009;41:666–76.
43. Wain LV, et al. Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. Nat Genet. 2011;43:1005–11.
44. Kottgen A, et al. New loci associated with kidney function and chronic kidney disease. Nat Genet. 2010;42:376–84.
45. Gudbjartsson DF, et al. Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. Nat Genet. 2009;41:342–7.
46. Rotival M, et al. Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. PLoS Genet. 2011;7:e1002367.
47. The International HapMap Project. Nature. 2003;426:789–96.
48. Westra HJ, et al. MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. Bioinformatics. 2011;27:2104–2111.
49. Breitling R, et al. Genetical genomics: spotlight on QTL hotspots. PLoS Genet. 2008;4:e1000232.
50. Whitlock MC. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J Evol Biol. 2005;18:1368–73.
51. Alberts R, et al. Sequence polymorphisms cause many false cis eQTLs. PLoS One. 2007;2:e622.
52. Benovoy D, Kwan T, Majewski J. Effect of polymorphisms within probe-target sequences on olignonucleotide microarray experiments. Nucleic Acids Res. 2008;36:4417–23.
53. Xu Z, Taylor JA. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. Nucleic Acids Res. 2009;37:W600–5.

54. Dayem Ullah AZ, Lemoine NR, Chelala C. SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). Nucleic Acids Res. 2012;40:W65–70.

55. Chelala C, Khan A, Lemoine NR. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. Bioinformatics. 2009;25:655–61.

56. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012;40:D930–4.

57. Flicek P, et al. Ensembl 2012. Nucleic Acids Res. 2012;40:D84–90.

58. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106:9362–7.

CHAPTER 3.2

Identification of non-coding RNA target genes through *trans*-eQTL analysis

Marjolein J. Peters, Jeroen van Rooij, Sumanta Basu, Marcus H. Stoiber, BIOS Consortium, CHARGE Consortium, Albert Hofman, André G. Uitterlinden, Leonard Lipovich, Rick Jansen, Peter A.C. 't Hoen, Lude Franke, Peter J. Bickel, James B. Brown*, and Joyce B.J. van Meurs*

** These authors contributed equally to this work*

Manuscript in preparation

ABSTRACT

Background: Non-coding RNAs are thought to be important regulators of gene expression. However, not much is known about this regulator potential. Genetic variants in or close to non-coding RNAs may influence the expression of their target genes. We aimed to identify genetic variants in non-coding RNAs that affect gene expression levels in *trans*.

Methods: We selected 38,545 SNPs in (or close to) exons of long non-coding RNAs (lncRNAs) which were likely to have functional consequences. We performed an expression quantitative trait locus (eQTL) analysis in peripheral blood samples of 652 individuals using RNA sequencing (RNA-seq) data, with replication in an additional 1,464 samples measured with RNA-seq. Significant findings were replicated in an RNA-array based *trans*-eQTL meta-analysis of 5,716 samples.

Results: We identified and replicated 2,678 *trans*-eQTLs for 1,320 lncRNAs SNPs. The majority of the identified *trans*-effects were located on the same chromosome as the analyzed SNP. When we selected only inter-chromosomal *trans*-eQTLs, we identified 195 *trans*-eQTLs: they are linked to 127 unique SNPs, and some affect multiple genes in *trans*. For example, the *cis*-eQTL SNP of lncRNA RP11-611L7.1 (rs13227497) is associated with the expression of KNS1, PI3, and ALDH1A2 in *trans*. SNPs in these genes are known to be associated with severe hand osteoarthritis, stress, and chronic pain.

Conclusion: Our *trans*-eQTL analysis for SNPs in non-coding RNAs provide new evidence for associations between SNPs in non-coding RNAs and gene expression levels on different chromosomes, indicating that *trans*-eQTLs give insight into new target genes of the non-coding RNAs.

INTRODUCTION

The first expression-quantitative trait locus (eQTL) meta-analyses in peripheral blood identified large numbers of common genetic variants regulating gene expression levels of nearby genes (*cis*-eQTLs) [1-6]. A few studies focused on *trans*-eQTLs (genetic variants influencing gene expression levels that reside further away on the chromosome or on a different chromosome), revealing the downstream consequences of genetic variants [1-3,5-9]. However, most *trans*-eQTL studies focused on SNPs that were previously found to be associated with diseases or traits.

Non-coding RNAs are thought to be involved in regulating gene expression. Dysregulation of non-coding RNAs influence tumorigenesis and neurological, cardiovascular, and developmental disease [10]. However, for most non-coding RNAs, mechanisms are unknown. In the current study, we systematically investigated the *cis*- and *trans*-eQTL effects of single nucleotide polymorphisms (SNPs) in non-coding RNAs. We focused on long non-coding RNAs (lncRNAs): they are non-protein coding RNA molecules longer than 200 nucleotides, and many have been found to function as co-regulators [11-15]. They modify transcription factor activity and regulate the activities of other co-regulators.

Genetic variants in or close to lncRNAs may affect the non-coding RNA function in different ways: the SNP may alter the binding efficiency of the lncRNA to the mRNA targets, or the SNP may change the expression of the lncRNA, which in turn may affect the expression of the target mRNA genes, or the SNP may affect the maturation efficiency of the lncRNA, which will influence the expression of the target mRNA genes.

We aimed to identify SNPs in (or close to) lncRNA having *trans*-eQTL effects. They may highlight new biological pathways and hypotheses for future studies.

METHODS

Study design

We performed the initial eQTL analysis in 652 human peripheral blood samples of the Rotterdam Study, a population-based cohort study in the district of Rotterdam, the Netherlands [16]. The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC and by the Ministry of Health, Welfare and Sport of the Netherlands , implementing the "*Wet Bevolkingsonderzoek: ERGO (Population Studies Act: Rotterdam Study)*". All participants provided written informed consent to participate in the study and to obtain information from their treating physicians.

We replicated the significantly associated eQTLs in a meta-analysis using RNA-sequencing data of 1,464 samples from peripheral blood, divided over three independent cohort studies including CODAM [17] (n=184), LL [18,19] (n=626), and LLS [20] (n=654).

For the replicating *trans*-eQTLs, we additionally checked whether the effects were similar in a meta-analysis using RNA-array data. We ran a meta-analysis including 5,716 peripheral blood samples, divided over seven independent cohort studies: DILGOM [21] (n=509), EGCUT [22] (n=891), FEHRMANN [1] (n=1,240), INCHIANTI [23] (n=611), KORA [24] (n=740), RS [25] (n=762), and SHIP-TREND [26] (n=963).

Gene expression profiling

Whole blood was collected (PAXGene Tubes – Becton Dickinson) and total RNA was isolated (PAXGene Blood RNA kits – Qiagen). To ensure a constant high quality of the RNA preparations, all RNA samples were analyzed with the Labchip GX (Calliper) according to the manufacturer's instructions. Samples exhibiting a *RQS<7* were excluded from further analyses.

For RNA-sequencing, we removed globin RNA from the total RNA (Ambions GLOBINclear kit), and generated mRNA focused libraries for RNA sequencing (Illumina's TruSeq Stranded Total RNA Library Prep Kit). We performed paired-end sequencing (2 x 50 bp) using Illumina's Hiseq2000, and pooled 10 samples per lane. We generated read sets per sample (CASAVA) and retained the reads passing the Illumina's Chastity Filter for further processing.

For the RNA-arrays, we hybridized the RNA to Illumina Whole-Genome Expression Beadchips (HT12v2 or HT12v4) using the protocol specified by the manufacturer.

RNA sequencing – preprocessing and alignment

To process the RNA-sequencing data, we used a pipeline developed by the BIOS consortium [27]. In short, the quality of the sequencing reads was checked using FastQC (v0.10.1), the adaptors were removed using *cutadapt v1.1* [28], and the low quality ends of the reads were trimmed using *sickle v1.200* [29].

The sequencing reads were mapped to the human genome v19 (hg19) using STAR *v2.3.125* [30], and we estimated expression on the exon, transcript and gene level using Ensembl *v71* annotation [31]. Post-processing steps were done with the Picard Tools software [32]. Overlapping exons were merged into meta-exons and expression was quantified for the whole meta-exon. This resulted in base counts per exon or meta-exon. Gene expression levels were calculated as the sum of expression values of all exons of each gene.

To normalize the gene expression levels, we divided the summed expression values by the gene length and the total number of reads per sample. We then used the Trimmed Mean of M-values (TMM) normalization method [33] to estimate scale factors between the samples. Finally, gene expression values were log2 transformed and standardized, and the first 25 principal components were removed, identical to the RNA array preprocessing steps [3].

Genotypes

The samples were genotyped using the Illumina 550K or the Illumina 610K arrays, and genotypes were imputed up to the 1000 Genomes Project (1000G) reference panel (phase *Iv3*) using MACH. The dataset was checked for sample mixups using *MixupMapper* [34]. *There were no mix-ups in this dataset.*

SNP selection for RNA sequencing

Of the 33,867,639 imputed variants, we selected 8,866,155 SNPs with a good imputation quality (QUAL>0.9) and a minor allele-frequency (MAF)>1%. To select the SNPs in lncRNA, we downloaded the genomic positions of all lncRNAs in the GENCODE v19 [35] database (n=23,898), and intersected the positions of the 8,866,155 good quality SNPs with the positions of the 23,898 lncRNA transcripts. 998,182 SNPs mapped in the lncRNAs, of which 51,968 SNPs mapped in the coding regions, of which 31,243 had a MAF>5% (SNPs with 0.01<MAF<0.05 were excluded). Additionally, we checked whether there are SNPs regulating the lncRNAs in *cis* using the *cis*-eQTL meta-analysis results of the BIOS consortium [27], and we identified 7,302 *cis*-eQTL SNPs. In total, we included 38,545 SNPs in or close to lncRNAs that are expressed in whole blood, which had a MAF>5%.

eQTL mapping

After normalization of the data, we performed both *cis*- and *trans*-eQTL mapping on the selected SNPs using the eQTL mapping pipeline developed by Westra *et al.* [3]. eQTLs were defined as *cis*-eQTLs, when the distance between the SNP chromosomal position and the probe midpoint was less than 250 kilobases (kb), while eQTLs with a distance greater than five megabases (mb) were defined as *trans*-eQTLs.

eQTLs were mapped using a Spearman's rank correlation on the imputed genotype dosage values. To control the false discovery rate (FDR) below 0.05, we created a null distribution by permuting sample labels of the expression data, repeating that five (for the *trans*-analysis) or ten (for the *cis*-analysis) times.

Trans-eQTL replication

All RNA-sequencing cohorts (CODAM, LLD, and LLS) applied the same methodology as used in the initial eQTL analysis to process the RNA sequencing and the genotype data. Finally, we performed a sample-size weighted Z-score meta-analysis on the three replication cohorts. Further details on these datasets can be found in the BIOS consortium manuscript [27]. We will further refer to this dataset as the BIOS-replication dataset.

Replication in RNA-array data

We performed a *trans*-eQTL meta-analysis in RNA-array cohorts (n=5,716). For the RNA-array cohorts, we selected the SNPs located in (or close to) exons of the lncRNAs that are present on the gene expression array: of the 8,866,155 SNPs with a good imputation quality (QUAL>0.9) and a minor allele-frequency (MAF)>1%, 18,986 SNPs mapped in exons of lncRNAs. Additionally, we included 749

SNPs which are known regulators of the lncRNAs [3]. In total, we included 7,950 SNPs in or close to lncRNAs that are expressed in whole blood, which had a MAF>5% (SNPs with 0.01<MAF<0.05 were excluded). All RNA-array cohorts applied the same methodology as used previously [3], including normalization and standardization, checking for sample mix-ups, principal component removal, and *trans*-eQTL mapping (including 10 permutations in order to establish the FDR threshold at 0.05). We performed a binary sample-size weighted Z-score meta-analysis on the RNA-array cohorts. We will further refer to this dataset as the CHARGE-replication dataset.

Pseudogenes

Because of ambiguous alignments in pseudogenes (which are very similar to their host gene), we filtered out *trans*-eQTLs in pseudogenes. To identify the pseudogenes, we used the GENCODE v19 [35] database, and intersected our *trans*-eQTL results with the list of 14,206 pseudogenes. In total, we excluded 1,395 *trans*-eQTLs in pseudogenes.

Data processing

All data processing was performed on the Dutch Life Science Grid. Further details about the grid can be found in the BIOS consortium manuscript [27].

RESULTS

Trans-eQTLs

In the lncRNA eQTL analysis, we tested the associations between 38,545 SNPs and 60,310 gene transcripts (Supplementary Table 1). We identified 58,327 *cis*-eQTLs and 4,556 *trans*-eQTLs effects (at FDR<0.05) (Table 1).

Table 1. The number of *cis*- and *trans*-eQTLs. Number of significant *cis*- and *trans*-eQTLs (FDR<0.05) in discovery and replication

eQTL effort	# eQTLs	# SNPs	# probes
cis-eQTLs			
Discovery (initial eQTL analysis)	58,327	19,172	13,229
BIOS-replication	**49,754 (85.3%)**	16,917	10,562
trans-eQTLs			
Discovery (initial eQTL analysis)	4,556	1,969	1,294
BIOS-replication	**4,073 (89.4%)**	1,700	1,064

BIOS-replication

In the BIOS replication, we focused on the 4,556 *trans*-eQTLs, and we replicated 89.4% (4,073 *trans*-eQTLs) with consistent allelic directions in the replication cohorts as compared to the

discovery cohort. We found 873 inter-chromosomal *trans*-eQTLs (the SNP was located in a different chromosome than the probe) and 3,200 *trans-eQTLs* with the SNP located in the same chromosome, of which 986 extended one megabase (mb) distance. After removing the *trans*-eQTLS driven by pseudogenes, we kept 2,678 *trans*-eQTLs: 195 inter-chromosomal *trans*-eQTLs, 576 *trans*-eQTLs with the SNP on the same chromosome with a distance >1mb, and 1,907 *trans*-eQTLs with the SNP on the same chromosome with a distance between 250kb and 1mb (Table 2).

Table 2. The number of replicating *trans*-eQTLs not driven by pseudogenes. Number of significant *trans*-eQTLs (FDR<0.05) splitted by *trans*-eQTL type

Type of *trans*-eQTL	**# *trans*-eQTLs**
Inter-chromosomal *trans*-eQTLs	**195**
# unique SNPs	127
# unique genes	76
Trans-eQTLs with SNP-gene distance >1mb	**576**
# unique SNPs	232
# unique genes	127
Trans-eQTLs with SNP-gene distance <1mb and >250kb	**1,907**
# unique SNPs	1,115
# unique genes	563
Total number of *trans*-eQTLs	**2,678**
# unique SNPs	1,320
# unique genes	723

Inter-chromosomal *trans*-eQTLs

We focused on the 195 significant inter-chromosomal *trans*-eQTLs which are caused by 127 unique SNPs (Supplementary Table 2): 105 SNPs are located in exons of 46 unique lncRNAs and 22 SNPs are known *cis*-eQTL SNPs for 18 unique lncRNAs [27]. Three lncRNAs (*SNHG7*, *ZNF571-AS1*, and *Z84812.4*) carry more than ten different SNPs with *trans*-eQTL effects. The lncRNA *SNHG7* is very interesting, because Boone *et al.* [36] recently showed that *SNHG7* belongs to the five most highly expressed and consistently regulated lncRNAs. It is a member of the small nucleolar host gene family, and its expression decreases by IGF1 signaling. Boone *et al.* propose that SNHG7 is a putative lncRNA oncogene that is controlled by IGF1 signaling in a feedback mechanism to prevent hyperproliferation, and that this regulation can be lost in the development or progression of breast cancer.

The 127 unique SNPs for which we found *trans*-eQTL effects were enriched for associations with immune-related diseases. In GWAS, these SNPs have been associated with diseases like celiac disease and Epstein-Barr virus immune response [37].

The 72 unique genes for which we found *trans*-eQTL effects were enriched for associations with blood- and immune related diseases: SNPs in these genes have been associated with disease like celiac disease, hematological traits, response to anti-retroviral therapy in HIV-1 infection, HIV-1 control, IgG glycosylation, immune response to smallpox, obesity related traits, and QT interval [37].

lncRNA RP11-56L13.1 affects phosphatase gene DUSP22

Two interesting *trans*-eQTL findings are the SNPs rs4247499 and rs148368513 in the first (rs4247499) and last (rs148368513) exon of the intergenic lncRNA *RP11-56L13.1*. The SNPs are in linkage (r^2=0.84 and D'=0.95), and both SNPs affect the gene expression levels of *DUSP22* (p=8.82E-24 and 6.64E-23), and the SNPs do not have any other *cis*- or *trans*-eQTL effect.

The *DUSP22* gene belongs to the tyrosine-protein phosphatase family and three SNPs located close to *DUSP22* have been linked to celiac disease, hematological traits, and response to anti-retroviral therapy in GWAS [37]. The risk alleles of the SNPs in the lncRNA *RP11-56L13.1* (rs4247499*T and rs148368513*T) increase the expression of the *DUSP22* gene. To our knowledge, the lncRNAs and the two SNPs have never been studied in the context of celiac disease, hematological traits, or response to anti-retroviral therapy.

rs13227497 and severe osteoarthritis

Another interesting results is the SNP rs13227497, which is known to be associated with the intergenic lncRNA *RP11-611L7.1* in *cis* (p=8.85E-102) [27]. In our *trans*-eQTL analysis, we observed the risk allele (rs13227497*A) to be associated with higher levels of the *KCNS1* gene (p=1.77E-14), the *PI3* gene (p=3.72E-14), and the *ALDH1A2* gene (p=4.26E-9). *KCNS1* encodes a potassium channel alpha subunit, and a missense SNP in *KCNS1* (rs734784) is known to be associated with chronic pain [38]. The *PI3* gene is a serine-type endopeptidase inhibitor and is regulated by NF-κB. NF-κB is involved in cellular responses to stimuli such as stress and free radicals. The *ALDH1A2* gene encodes retinaldehyde dehydrogenase 2 (*RALDH2*), an enzyme that catalyzes the synthesis of retinoic acid. Common variants within the *ALDH1A2* gene have been associated with severe osteoarthritis of the hand [39]. Because severe osteoarthritis goes hand In hand with chronic pain and stress, this lncRNA might affect these genes in one pathway.

lncRNA RP11-34P13.14 regulated by SNPs on 2 chromosomes

A third example is the lncRNA *RP11-34P13.14*, which is associated with five different SNPs on two different chromosomes (chr7 and chr16). Two SNPs on chromosome 7 are located in the first exon of another lncRNA *AC093627.7* (rs79615415 and rs147418006) and one SNP (rs143579933) is a *cis*-eQTL SNP of this lncRNA. The two other SNPs (rs4785780 and rs12923514) on chromosome 16 are located in an intron of the *PRDM7* gene, which is involved in transcription regulation [40]. All risk alleles are associated with higher levels of the lncRNA *RP11-34P13.14*, and only the SNPs on chromosome 16 show additional *cis*-effects. These results indicates a complex network where one lncRNA *RP11-34P13.14* is regulated by *trans*-eQTL SNPs in another lncRNA and another gene; the *PRDM7* gene.

CHARGE-replication with RNA-array data

In the CHARGE replication (based on RNA-array data), 41 of the 195 *trans*-eQTL SNP-gene combinations could be tested because of different SNP inclusion criteria (see methods). For two SNPs (4.9%), we replicated the *trans*-eQTLs effects (Table 3 and Supplementary Table 3).

Table 3. Number of *trans*-eQTLs in the CHARGE-lookup. Number of SNPs and probes tested in the RNA array meta-analysis

Summary statistics	# inter-chromosomal *trans*-eQTLs
RNA-seq	**195**
RNA-array – # of tested combinations	41
RNA-array – # of replicating *trans*-eQTLs	**2** (4.9%)

The first replicating *trans*-eQTL is the SNP rs7440274, which is located in the last exon of the lncRNA *ZNF718* on chromosome 4. The minor allele G (MAF=19.1%) is associated with lower expression levels of the *PHKB* gene (on chromosome 16) in RS (beta=-5.88, p=3.9E-9), BIOS (beta=-6.21, p=5.37E-10), and in the CHARGE meta-analysis (beta=-7.81, p=5.81E-15). The PHKB gene encodes the beta subunit of phosphorylase kinase, and earlier studies reported mutations in the PHKB gene resulting in glycogen storage disease [41].

The second replicating *trans*-eQTL is the SNP rs12939138, which is located in the last exon of the lncRNA *RP11-1094M14.11* on chromosome 17. The minor allele T (MAF=4.3%) is associated with higher expression levels of the *PTGS1* gene (on chromosome 9) in RS(beta=5.33, p=9.65E-8), BIOS (beta=5. 90, p=3.68E-9), and in the CHARGE meta-analysis (beta=7.15, p=8.58E-13). The activity of the *PTGS1* enzyme is known to be inhibited by the nonsteroidal anti-inflammatory drugs (NSAID) such as aspirin [42,43].

DISCUSSION

In this *trans*-eQTL study for SNPs thought to affect the function of non-coding RNAs, we identified and replicated 2,678 *trans*-eQTLs In (or close to) lncRNAs. We focused on the 195 inter-chromosomal *trans*-eQTLs. In the CHARGE replication based on RNA-array data, we replicated only 4.9% of the inter-chromosomal *trans*-eQTLs.

We described three inter-chromosomal *trans*-eQTL examples for which we identified potentially new associations with target genes. The SNPs in lncRNA *RP11-56L13.1* are linked to *DUSP22*, and three SNPs close to *DUSP22* are known to be associated with celiac disease, hematological traits, and response to retroviral therapy. Therefore, it might be interesting to study this lncRNA in relation

to these diseases. The SNP rs13227497 (a *cis*-eQTL SNP for lncRNA *RP11-611L7.1*) was associated with higher expression levels of *KCNS11*, *PI3*, and *ALDH1A2*, and SNPs in these genes are known to be associated with osteoarthritis related phenotypes. And five different SNPs on two different chromosomes were associated with higher expression levels of the lncRNA *RP11-34P13.14*. This might indicate a complex network where one lncRNA is regulated by different *trans*-eQTL SNPs on different chromosomes. In summary, these results indicate that the lncRNAs might be important for regulating one or more target genes which may belong to one pathway.

We replicated two of 41 (4.9%) *trans*-eQTL combinations tested in the CHARGE meta-analysis which was based on RNA-array data. The low percentage of replication could be explained by the different measurement techniques used. The RNA-arrays use 3'probes, and measure the relative amount of gene expression levels of a complete gene transcript. In contrast, with RNA-seq we measure the absolute number of reads mapping to each individual exon across the gene of interest. Therefore, the concordance between RNA-seq and RNA-arrays is not perfect. For the two replicating *trans*-eQTLs, we could say that these *trans*-eQTLs represent very robust signals: independent of the measurement technique used, the genes seem to be regulated by the lncRNA SNPs. Given the enormous size of the possible combinations for SNPs and targets (potentially detected *trans*-eQTLs), this level of replication is highly statistically significant.

To be able to replicate more *trans*-eQTLs, the RNA-seq alignment should be more similar to the RNA-array design: instead of using all exons in a gene, we could target the 3' exon only (like the RNA-array). This might improve the correlation between the gene expression levels measured with RNA-arrays and RNA-seq.

How the lncRNAs are regulating the mRNA expression is still unclear: only a small number of lncRNAs have been functionally well-characterized. Next to analyzing the effect of SNPs in lncRNAs on genome-wide expression levels, we could study the interactions between lncRNAs and proteins using RNA immuneprecipitation sequencing (RIP-seq). RIP-seq is an antibody-based technique used to map *in vivo* RNA-protein interactions. The RNA binding protein (RBP) of interest is immunoprecipitated together with its associated lncRNAs for identification of bound transcripts. With RIP-seq, we will gain more insight about the interactions between individual proteins and specific lncRNA molecules. This technique has been used to identify chromatin-modifying complexes interacting with lncRNAs like *Kcnq1ot1*, *Airn*, *Xist*, and *HOTAIR* [44].

A potential limitation of our study is that we relied on Spearman's correlation to identify eQTLs. Spearman's correlation assumes constant change over genotype (by adding one minor allele very step), which may not always be correct. We did run random forests which increased the number of significant eQTLs enormously. However, the replication rate for the random forests was worse. Although we chose to apply Spearman's correlation in our study, we recognize that more complex models should be investigated in the future.

To fully understand the functional mechanisms of lncRNAs, we could perform targeted perturbations to determine the role of the specific lncRNAs. By repressing the levels of the functional lncRNA, the dynamics behavior of the interactions can be studied. However, the reduced expression levels should have phenotypically measurable consequences.

In summary, our analysis revealed and replicated 195 inter-chromosomal *trans*-eQTL effects for 127 unique SNPs in (or close to) lncRNAs. Our complete list of *trans*-eQTLs provides new biological pathways and hypotheses for future studies. Larger *trans*-eQTL analyses will likely uncover many more of these regulatory relationships. Functional experiments would help to give more insight into the biological mechanisms underlying the *trans*-eQTL associations.

SUPPLEMENTARY TABLES

Supplementary Table 1. The number of SNPs and probes tested.

Summary statistics	*cis*-eQTLs	*trans*-eQTLs
Number of SNPs tested	38,545	38,545
Number of probes tested	60,310	60,310
Number of SNP-probe combinations tested	564,525	2,324,648,950

Supplementary Table 2. All significant inter-chromosomal *trans*-eQTLs. The *trans*-eQTLs discussed in the manuscript are indicated in light grey. The two *trans*-eQTLs significantly replicating in both the BIOS consortium (RNA-seq based data) and the CHARGE consortium (RNA-array based data) are marked in light green.

ID	SNP	GENE		lncRNA	RS			BIOS (no RS)			#cis
	CHR	CHR	SYMBOL		Effect	P	FDR	Effect	P	FDR	
rs3810491_ENSG00000236650	20	22	CTA-221G9.10	CTD-3184A7.4	-13.32	1.70E-40	0.000	-23.06	1.10E-117	0.000	4
rs77706710_ENSG00000173213	16	18	RP11-683L23.1	Z84812.4	16.99	9.77E-65	0.000	19.72	1.41E-86	0.000	4
rs7466679_ENSG00000232019	9	7	AC074183.4	SNHG7	12.30	8.53E-35	0.000	19.47	2.00E-84	0.000	4
rs77706710_ENSG00000173876	16	10	TUBB8	Z84812.4	17.10	1.56E-65	0.000	19.39	9.82E-84	0.000	4
rs77706710_ENSG00000124334	16	23	IL9R	Z84812.4	-17.35	1.98E-67	0.000	-19.19	4.50E-82	0.000	4
rs72761013_ENSG00000232019	9	7	AC074183.4	SNHG7	12.24	1.81E-34	0.000	19.10	2.62E-81	0.000	4
rs1384_ENSG00000258486,ENSG00000266422	12	14	RN7SL1,Metazoa_SRP	RP11-1143G9.4	-11.17	5.82E-29	0.000	-18.95	4.42E-80	0.000	3
rs12669559_ENSG00000017483	7	23	SLC38A5	AC020743.3	-12.55	4.03E-36	0.000	-18.38	1.98E-75	0.000	1
rs77706710_ENSG00000228463	16	1	AP006222.2	Z84812.4	16.39	2.26E-60	0.000	18.27	1.40E-74	0.000	4
rs1384_ENSG00000266037	12	14	Metazoa_SRP	RP11-1143G9.4	-8.85	8.85E-19	0.000	-17.32	3.28E-67	0.000	3
rs1384_ENSG00000265150	12	14	RN7SL2	RP11-1143G9.4	-7.91	2.60E-15	0.000	-16.89	5.29E-64	0.000	3
rs1384_ENSG00000265735	12	9	Metazoa_SRP	RP11-1143G9.4	-10.21	1.83E-24	0.000	-16.65	2.85E-62	0.000	3
rs6866_ENSG00000232019	9	7	AC074183.4	SNHG7	10.70	1.02E-26	0.000	16.39	2.17E-60	0.000	4
rs11793962_ENSG00000232019	9	7	AC074183.4	SNHG7	10.64	1.91E-26	0.000	16.39	2.30E-60	0.000	4
rs2275161_ENSG00000232019	9	7	AC074183.4	SNHG7	10.73	7.46E-27	0.000	16.39	2.37E-60	0.000	5
rs6001798_ENSG00000260986	22	15	RP11-854K16.3	BMS1P20	12.12	8.00E-34	0.000	16.37	3.15E-60	0.000	4
rs10781513_ENSG00000232019	9	7	AC074183.4	SNHG7	10.73	7.46E-27	0.000	16.36	3.51E-60	0.000	5
rs11684486_ENSG00000225655	2	9	RP11-143M1.2	RP11-395L14.4	11.08	1.57E-28	0.000	15.60	7.68E-55	0.000	4
rs2305105_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	AC016582.2	10.27	1.01E-24	0.000	15.57	1.21E-54	0.000	1
rs2105078_ENSG00000232019	9	7	AC074183.4	SNHG7	8.68	3.91E-18	0.000	15.42	1.20E-53	0.000	6
rs10870115_ENSG00000232019	9	7	AC074183.4	SNHG7	9.30	1.44E-20	0.000	15.28	1.01E-52	0.000	4
rs3739939_ENSG00000232019	9	7	AC074183.4	SNHG7	9.04	1.50E-19	0.000	15.24	2.01E-52	0.000	4
rs77706710_ENSG00000230021	16	1	RP5-857K21.4	Z84812.4	13.20	9.06E-40	0.000	15.00	6.87E-51	0.000	4
rs2305105_ENSG00000247498	19	12	RP11-392P7.6	AC016582.2	9.61	7.38E-22	0.000	14.99	8.61E-51	0.000	1
rs1384_ENSG00000263740	12	3	Metazoa_SRP	RP11-1143G9.4	-6.90	5.06E-12	0.000	-14.89	4.05E-50	0.000	3
rs7284467_ENSG00000260986	22	15	RP11-854K16.3	BMS1P20	11.49	1.42E-30	0.000	14.47	1.89E-47	0.000	4
rs2068431_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	AC016582.2	9.66	4.25E-22	0.000	14.33	1.40E-46	0.000	1
rs4803528_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	AC016582.2	9.66	4.25E-22	0.000	14.33	1.40E-46	0.000	1

Supplementary Table 2. *Continued)*

ID	SNP	GENE		lncRNA	RS			BIOS (no RS)			#cis
	CHR	CHR	SYMBOL		Effect	P	FDR	Effect	P	FDR	
rs7855405_ENSG00000232019	9	7	AC074183.4	SNHG7	9.75	1.84E-22	0.000	14.07	5.47E-45	0.000	6
rs2068431_ENSG00000247498	19	12	RP11-392P7.6	AC016582.2	8.77	1.87E-18	0.000	13.89	7.30E-44	0.000	1
rs4803528_ENSG00000247498	19	12	RP11-392P7.6	AC016582.2	8.77	1.87E-18	0.000	13.89	7.30E-44	0.000	1
rs2296763_ENSG00000236650	20	22	CTA-221G9.10	RP4-583P15.10	-9.30	1.44E-20	0.000	-13.88	8.56E-44	0.000	4
rs77706710_ENSG00000238009	16	1	RP11-34P13.7	Z84812.4	12.30	8.53E-35	0.000	13.72	7.46E-43	0.000	4
rs2275158_ENSG00000232019	9	7	AC074183.4	SNHG7	8.32	8.71E-17	0.000	13.66	1.67E-42	0.000	7
rs2074404_ENSG00000120314,ENSG00000256453	17	5	WDR55,DND1	FAM215B	8.43	3.40E-17	0.000	13.27	3.39E-40	0.000	5
rs4880136_ENSG00000232019	9	7	AC074183.4	SNHG7	-8.40	4.31E-17	0.000	13.27	3.60E-40	0.000	6
rs2275160_ENSG00000232019	9	7	AC074183.4	SNHG7	-8.40	4.31E-17	0.000	13.26	3.86E-40	0.000	7
rs4795085_ENSG00000105701	17	19	FKBP8	CTC-507E2.1	-7.58	3.43E-14	0.000	-13.13	2.08E-39	0.000	3
rs1384_ENSG00000215014	12	1	AL645728.1	RP11-1143G9.4	-7.88	3.24E-15	0.000	-13.03	7.80E-39	0.000	3
rs77706710_ENSG00000185203	16	23	WASIR1	Z84812.4	-8.54	1.31E-17	0.000	-12.77	2.36E-37	0.000	4
rs204547_ENSG00000133985	19	14	TTC9	CTC-512J12.4	-7.99	1.34E-15	0.000	-12.75	3.25E-37	0.000	0
rs77706710_ENSG00000238035	16	5	AC138035.2	Z84812.4	12.46	1.28E-35	0.000	12.57	2.97E-36	0.000	4
rs4795085_ENSG00000105701,ENSG00000268938	17	19	FKBP8,AC005387.3	CTC-507E2.1	-6.74	1.56E-11	0.000	-12.38	3.37E-35	0.000	3
rs3890901_ENSG00000105701	17	19	FKBP8	CTC-507E2.1	-6.31	2.72E-10	0.000	12.29	9.73E-35	0.000	3
rs9306349_ENSG00000260986	22	15	RP11-854K16.3	BMS1P20	-7.31	2.69E-13	0.000	-12.21	2.80E-34	0.000	5
rs77706710_ENSG00000230724,ENSG00000255229	16	11	RP11-304M2.2,RP11-304M2.3	Z84812.4	11.85	2.11E-32	0.000	12.19	3.36E-34	0.000	4
rs1128327_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	CTD-3064H18.1	6.15	7.56E-10	0.000	12.15	5.56E-34	0.000	3
rs3890901_ENSG00000105701,ENSG00000268938	17	19	FKBP8,AC005387.3	CTC-507E2.1	-5.60	2.18E-08	0.008	11.81	3.39E-32	0.000	3
rs77706710_ENSG00000250765	16	5	AC138035.1	Z84812.4	11.38	5.57E-30	0.000	11.77	5.73E-32	0.000	4
rs1119229_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	CTD-2554C21.3	-6.85	7.37E-12	0.000	-11.69	1.41E-31	0.000	5
rs7501448_ENSG00000105701	17	19	FKBP8	RP11-1094M14.10	7.42	1.19E-13	0.000	11.66	1.98E-31	0.000	3
rs10410588_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	ZNF571-AS1	-6.82	8.90E-12	0.000	-11.33	9.05E-30	0.000	6
rs13744_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	ZNF571-AS1	-6.82	8.90E-12	0.000	-11.33	9.05E-30	0.000	7
rs3829688_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	ZNF571-AS1	-6.82	8.90E-12	0.000	-11.33	9.05E-30	0.000	7
rs77706710_ENSG00000225532	16	6	XX-C2158C6.3	Z84812.4	11.20	4.19E-29	0.000	11.32	1.02E-29	0.000	4
rs7501448_ENSG00000105701,ENSG00000268938	17	19	FKBP8,AC005387.3	RP11-1094M14.10	6.85	7.39E-12	0.000	11.22	3.18E-29	0.000	3

Supplementary Table 2. *Continued)*

ID	SNP	GENE		lncRNA	RS			BIOS (no RS)			*#cis*
	CHR	CHR	SYMBOL		Effect	P	FDR	Effect	P	FDR	
rs10420891_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	CTD-2554C21.3	-7.01	2.35E-12	0.000	-11.20	3.95E-29	0.000	5
rs2075281_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	CTD-2554C21.2	-6.98	2.85E-12	0.000	-11.15	7.44E-29	0.000	5
rs10402530_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	CTD-2554C21.3	-6.98	2.85E-12	0.000	-11.13	9.04E-29	0.000	5
rs8182498_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	AC016582.2	-6.98	2.85E-12	0.000	-11.11	1.10E-28	0.000	5
rs2854275_ENSG00000211789	6	14	TRAV12-2	HLA-DQB1-AS1	6.47	9.52E-11	0.000	10.96	5.95E-28	0.000	12
rs10406379_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	CTD-3064H18.4	-6.55	5.57E-11	0.000	-10.69	1.14E-26	0.000	6
rs10412786_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	CTD-3064H18.4	-6.55	5.57E-11	0.000	-10.65	1.69E-26	0.000	6
rs62447200_ENSG00000017483	7	23	SLC38A5	AC018705.5	6.18	6.39E-10	0.000	10.63	2.14E-26	0.000	0
rs17611866_ENSG00000159882	16	19	ZNF230	TIGD7,ZNF75A,LA16c-360H6.2,LA16c-360H6.3	6.69	2.25E-11	0.000	10.49	9.71E-26	0.000	6
rs7590375_ENSG00000225655	2	9	RP11-143M1.2	AC017074.2	-7.53	5.21E-14	0.000	-10.42	2.07E-25	0.000	4
rs77706710_ENSG00000254468	16	11	RP11-304M2.6	Z84812.4	10.32	5.55E-25	0.000	10.36	3.83E-25	0.000	4
rs7599741_ENSG00000225655	2	9	RP11-143M1.2	AC017074.2	8.38	5.46E-17	0.000	10.36	3.96E-25	0.000	3
rs7599832_ENSG00000225655	2	9	RP11-143M1.2	AC017074.2	8.38	5.46E-17	0.000	10.36	3.96E-25	0.000	3
rs1119229_ENSG00000247498	19	12	RP11-392P7.6	CTD-2554C21.3	-6.61	3.88E-11	0.000	-10.31	6.07E-25	0.000	5
rs4795085_ENSG00000158856	17	8	EPB49	CTC-507E2.1	-5.94	2.82E-09	0.001	-10.31	6.32E-25	0.000	3
rs77706710_ENSG00000225532,ENSG00000230423	16	6	XX-C2158C6.3,XX-C2158C6.1	Z84812.4	8.43	3.40E-17	0.000	10.23	1.41E-24	0.000	4
rs17637907_ENSG00000105701	17	19	FKBP8	RP11-1094M14.5	6.47	9.52E-11	0.000	10.23	1.51E-24	0.000	3
rs77706710_ENSG00000225119	16	10	AL133216.1	Z84812.4	9.13	6.95E-20	0.000	10.17	2.67E-24	0.000	4
rs77706710_ENSG00000237094	16	1	RP4-669L17.10	Z84812.4	9.66	4.25E-22	0.000	10.09	6.24E-24	0.000	4
rs4247499_ENSG00000112679	16	6	DUSP22	RP11-56L13.1	6.47	9.52E-11	0.000	10.05	8.82E-24	0.000	0
rs1128327_ENSG00000247498	19	12	RP11-392P7.6	CTD-3064H18.1	6.15	7.56E-10	0.000	10.03	1.13E-23	0.000	3
rs13744_ENSG00000247498	19	12	RP11-392P7.6	ZNF571-AS1	-6.66	2.70E-11	0.000	-9.91	3.90E-23	0.000	7
rs3829688_ENSG00000247498	19	12	RP11-392P7.6	ZNF571-AS1	-6.66	2.70E-11	0.000	-9.91	3.90E-23	0.000	7
rs10420891_ENSG00000247498	19	12	RP11-392P7.6	CTD-2554C21.3	-6.42	1.35E-10	0.000	-9.90	4.30E-23	0.000	5
rs77706710_ENSG00000230724	16	11	RP11-304M2.2	Z84812.4	9.86	5.95E-23	0.000	9.90	4.37E-23	0.000	4
rs10410588_ENSG00000247498	19	12	RP11-392P7.6	ZNF571-AS1	-6.66	2.70E-11	0.000	-9.89	4.67E-23	0.000	6
rs17637907_ENSG00000105701,ENSG00000268938	17	19	FKBP8,AC005387.3	RP11-1094M14.5	5.89	3.90E-09	0.002	9.87	5.44E-23	0.000	3

Supplementary Table 2. *Continued)*

ID	SNP	GENE		lncRNA	RS			BIOS (no RS)			#cis
	CHR	CHR	SYMBOL		Effect	P	FDR	Effect	P	FDR	
rs148368513_ENSG00000112679	16	6	DUSP22	RP11-56L13.1	5.94	2.83E-09	0.001	9.85	6.64E-23	0.000	0
rs2075281_ENSG00000247498	19	12	RP11-392P7.6	CTD-2554C21.2	-6.37	1.92E-10	0.000	-9.82	8.86E-23	0.000	5
rs10402530_ENSG00000247498	19	12	RP11-392P7.6	CTD-2554C21.3	-6.37	1.92E-10	0.000	-9.81	1.05E-22	0.000	5
rs12982283_ENSG00000211750	19	7	TRBV24-1	CTD-3105H18.7	7.17	7.31E-13	0.000	9.80	1.09E-22	0.000	6
rs8182498_ENSG00000247498	19	12	RP11-392P7.6	AC016582.2	-6.37	1.92E-10	0.000	-9.79	1.26E-22	0.000	5
rs204548_ENSG00000133985	19	14	TTC9	CTC-512J12.4	-6.64	3.24E-11	0.000	-9.78	1.32E-22	0.000	0
rs5760492_ENSG00000149435	22	20	GGTLC1	AP000356.2	-6.93	4.18E-12	0.000	-9.30	1.35E-20	0.000	3
rs4381721_ENSG00000228789	19	6	HCG22	RP11-420K14.8	-5.91	3.32E-09	0.001	-9.28	1.71E-20	0.000	1
rs10406379_ENSG00000247498	19	12	RP11-392P7.6	CTD-3064H18.4	-6.39	1.61E-10	0.000	-9.24	2.40E-20	0.000	6
rs10412786_ENSG00000247498	19	12	RP11-392P7.6	CTD-3064H18.4	-6.39	1.61E-10	0.000	-9.22	2.87E-20	0.000	6
rs4522535_ENSG00000228789	19	6	HCG22	RP11-420K14.8	-6.02	1.73E-09	0.001	-9.12	7.65E-20	0.000	1
rs13388082_ENSG00000111261	2	12	MANSC1	AC009506.1	5.97	2.40E-09	0.001	-9.12	7.82E-20	0.000	6
rs6985508_ENSG00000123505	8	6	AMD1	CTD-3064M3.1	-6.39	1.61E-10	0.000	-9.11	8.18E-20	0.000	4
rs6986779_ENSG00000123505	8	6	AMD1	CTD-3064M3.1	-6.39	1.61E-10	0.000	-9.11	8.18E-20	0.000	4
rs6745815_ENSG00000111261	2	12	MANSC1	AC009961.3	5.94	2.83E-09	0.001	-9.06	1.31E-19	0.000	5
rs1048505_ENSG00000247498	19	12	RP11-392P7.6	CTD-2528L19.6	-5.44	5.36E-08	0.022	-9.04	1.54E-19	0.000	6
rs3821300_ENSG00000111261	2	12	MANSC1	AC009961.3	5.99	2.04E-09	0.001	-9.03	1.75E-19	0.000	5
rs12669559_ENSG00000198336	7	17	MYL4	AC020743.3	-6.69	2.25E-11	0.000	-8.97	2.94E-19	0.000	1
rs77706710_ENSG00000233013	16	9	FAM157B	Z84812.4	8.18	2.76E-16	0.000	8.97	3.08E-19	0.000	4
rs2306201_ENSG00000247498	19	12	RP11-392P7.6	ZNF571-AS1	-5.31	1.11E-07	0.047	-8.92	4.62E-19	0.000	7
rs73033122_ENSG00000247498	19	12	RP11-392P7.6	ZNF571-AS1	-5.31	1.11E-07	0.047	-8.90	5.40E-19	0.000	7
rs11090029_ENSG00000260986	22	15	RP11-854K16.3	BMS1P20	6.02	1.73E-09	0.001	8.87	7.30E-19	0.000	5
rs56277404_ENSG00000247498	19	12	RP11-392P7.6	ZNF571-AS1	-5.33	9.64E-08	0.041	-8.87	7.37E-19	0.000	7
rs2732485_ENSG00000134152	12	15	KATNBL1	RP11-370I10.2	5.54	2.95E-08	0.012	8.86	7.83E-19	0.000	8
rs11684486_ENSG00000237543	2	9	RP11-58A12.3	RP11-395L14.4	6.82	8.92E-12	0.000	8.63	5.93E-18	0.000	4
rs79615415_ENSG00000239906	7	1	RP11-34P13.14	AC093627.7	6.31	2.72E-10	0.000	8.61	7.43E-18	0.000	1
rs6001798_ENSG00000258410	22	15	RP11-173D3.1	BMS1P20	6.07	1.25E-09	0.001	8.60	8.31E-18	0.000	4
rs71595673_ENSG00000230880	4	1	AL583842.3	RP11-535C7.1	6.72	1.87E-11	0.000	8.59	8.47E-18	0.000	1
rs143579933_ENSG00000239906	7	1	RP11-34P13.14	AC093627.7	6.55	5.58E-11	0.000	8.43	3.60E-17	0.000	1

Supplementary Table 2. *Continued)*

ID	SNP	GENE		lncRNA	RS			BIOS (no RS)			*#cis*
	CHR	CHR	SYMBOL		Effect	P	FDR	Effect	P	FDR	
rs3807552_ENSG00000017483	7	23	SLC38A5	AC018705.5	6.50	7.98E-11	0.000	8.37	5.81E-17	0.000	1
rs3779078_ENSG00000017483	7	23	SLC38A5	AC018705.5	6.42	1.36E-10	0.000	8.35	7.05E-17	0.000	1
rs3807556_ENSG00000017483	7	23	SLC38A5	AC018705.5	6.45	1.14E-10	0.000	8.35	7.05E-17	0.000	1
rs3098393_ENSG00000116962	19	1	NID1	LINC00665	6.45	1.14E-10	0.000	8.23	1.82E-16	0.000	5
rs3111556_ENSG00000116962	19	1	NID1	LINC00665	6.45	1.14E-10	0.000	8.23	1.82E-16	0.000	5
rs1048505_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	CTD-2528L19.6	-5.94	2.82E-09	0.001	-8.22	2.01E-16	0.000	6
rs12669559_ENSG00000198892	7	1	SHISA4	AC020743.3	-5.89	3.90E-09	0.002	-8.22	2.10E-16	0.000	1
rs147418006_ENSG00000239906	7	1	RP11-34P13.14	AC093627.7	6.47	9.52E-11	0.000	8.20	2.34E-16	0.000	1
rs1384_ENSG00000239899	12	2	Metazoa_SRP	RP11-1143G9.4	-5.31	1.11E-07	0.047	-8.15	3.63E-16	0.000	3
rs12711778_ENSG00000225655	2	9	RP11-143M1.2	AC017074.1	-7.28	3.28E-13	0.000	-8.03	9.64E-16	0.000	4
rs2306201_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	ZNF571-AS1	-6.02	1.73E-09	0.001	-8.00	1.28E-15	0.000	7
rs73033122_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	ZNF571-AS1	-6.02	1.73E-09	0.001	-8.00	1.28E-15	0.000	7
rs1384_ENSG00000244642	12	8	Metazoa_SRP	RP11-1143G9.4	-5.99	2.04E-09	0.001	-7.99	1.40E-15	0.000	3
rs56277404_ENSG00000013583,ENSG00000247498	19	12	HEBP1,RP11-392P7.6	ZNF571-AS1	-6.02	1.73E-09	0.001	-7.95	1.94E-15	0.000	7
rs2854275_ENSG00000211797	6	14	TRAV17	HLA-DQB1-AS1	-7.04	1.94E-12	0.000	-7.89	3.04E-15	0.000	12
rs3111555_ENSG00000116962	19	1	NID1	LINC00665	6.39	1.61E-10	0.000	7.88	3.28E-15	0.000	5
rs3093872_ENSG00000104356	14	8	POP1	RPPH1	7.72	1.19E-14	0.000	7.78	7.31E-15	0.000	2
rs7284467_ENSG00000258410	22	15	RP11-173D3.1	BMS1P20	6.10	1.06E-09	0.001	7.72	1.15E-14	0.000	4
rs13227497_ENSG00000124134	7	20	KCNS1	RP11-611L7.1	5.94	2.83E-09	0.001	7.67	1.77E-14	0.000	4
rs1465789_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.17	-6.69	2.25E-11	0.000	7.61	2.74E-14	0.000	10
rs13227497_ENSG00000124102	7	20	PI3	RP11-611L7.1	5.41	6.23E-08	0.028	7.57	3.72E-14	0.000	4
rs7408188_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.17	-6.69	2.25E-11	0.000	7.56	4.05E-14	0.000	10
rs9926788_ENSG00000225119	16	10	AL133216.1	FAM157C	6.61	3.89E-11	0.000	7.52	5.65E-14	0.000	3
rs676387_ENSG00000140450	17	15	ARRDC4	RP11-400F19.6	-5.78	7.36E-09	0.003	-7.50	6.61E-14	0.000	5
rs4801583_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.14	-6.82	8.90E-12	0.000	-7.35	1.93E-13	0.000	10
rs2854275_ENSG00000211734	6	7	TRBV5-1	HLA-DQB1-AS1	5.39	7.20E-08	0.033	7.34	2.20E-13	0.000	12
rs6001798_ENSG00000261200	22	16	RP11-989E6.10	BMS1P20	5.44	5.37E-08	0.024	7.31	2.76E-13	0.000	4
rs7248357_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.14	-6.82	8.90E-12	0.000	-7.26	3.83E-13	0.000	10

Supplementary Table 2. *Continued)*

ID	SNP	GENE		lncRNA	RS			BIOS (no RS)			#cis
	CHR	CHR	SYMBOL		Effect	P	FDR	Effect	P	FDR	
rs72759285_ENSG00000159713,ENSG00000239194	9	16	TPPP3,U1	RP11-121A14.2	-5.65	1.60E-08	0.006	-7.22	5.38E-13	0.000	4
rs7501448_ENSG00000095303	17	9	PTGS1	RP11-1094M14.10	6.05	1.47E-09	0.001	7.20	6.18E-13	0.000	3
rs3807555_ENSG00000017483	7	23	SLC38A5	AC018705.5	5.94	2.83E-09	0.001	7.18	6.80E-13	0.000	1
rs11643147_ENSG00000225119	16	10	AL133216.1	RP11-566K11.5	5.91	3.32E-09	0.002	7.18	6.89E-13	0.000	7
rs36114385_ENSG00000225119	16	10	AL133216.1	RP11-566K11.5	5.91	3.32E-09	0.002	7.18	6.89E-13	0.000	7
rs11684486_ENSG00000224958	2	9	PGM5-AS1	RP11-395L14.4	5.99	2.04E-09	0.001	7.06	1.63E-12	0.000	4
rs8103272_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.19	-6.72	1.87E-11	0.000	-7.05	1.76E-12	0.000	10
rs4372790_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.19	-6.66	2.70E-11	0.000	-7.04	1.88E-12	0.000	10
rs4801589_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.19	-6.66	2.70E-11	0.000	-7.04	1.88E-12	0.000	10
rs2179410_ENSG00000067064	6	10	IDI1	RP1-91J24.3	5.68	1.38E-08	0.006	7.04	1.89E-12	0.000	1
rs6510152_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.19	-6.69	2.25E-11	0.000	7.01	2.36E-12	0.000	10
rs4785780_ENSG00000239906	16	1	RP11-34P13.14	RP11-356C4.5	6.13	8.93E-10	0.000	6.87	6.55E-12	0.000	5
rs3794968_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.23	5.86	4.58E-09	0.002	6.82	8.84E-12	0.000	9
rs7284467_ENSG00000261200	22	16	RP11-989E6.10	BMS1P20	5.52	3.43E-08	0.014	6.72	1.78E-11	0.000	4
rs61792256_ENSG00000102893	4	16	PHKB	ZNF718	-6.07	1.24E-09	0.001	-6.63	3.44E-11	0.000	5
rs2854275_ENSG00000211782	6	14	TRAV8-1	HLA-DQB1-AS1	5.81	6.29E-09	0.003	6.58	4.55E-11	0.000	12
rs72759301_ENSG00000159713,ENSG00000239194	9	16	TPPP3,U1	RP11-121A14.2	-5.81	6.28E-09	0.003	-6.45	1.11E-10	0.000	4
rs6001700_ENSG00000260986	22	15	RP11-854K16.3	LL22NC03-2H8.4	5.33	9.65E-08	0.043	6.41	1.50E-10	0.000	8
rs77706710_ENSG00000203496	16	10	RP11-291L22.4	Z84812.4	6.23	4.55E-10	0.000	6.36	2.01E-10	0.000	4
rs17611866_ENSG00000188474,ENSG00000214189	16	19	ZNF788,ZNF788	TIGD7,ZNF75A,LA16c-360H6.2,LA16c-360H6.3	6.53	6.67E-11	0.000	6.32	2.66E-10	0.000	6
rs7440274_ENSG00000102893	4	16	PHKB	ZNF718	-5.89	3.90E-09	0.002	-6.21	5.37E-10	0.000	4
rs148989274_ENSG00000249868	1	8	RP11-63E5.6	LINC00115	-5.44	5.36E-08	0.022	-6.09	1.15E-09	0.000	2
rs80198924_ENSG00000230587	19	2	AC093609.1	AC012309.5	5.73	1.01E-08	0.005	6.05	1.41E-09	0.000	2
rs17637907_ENSG00000095303	17	9	PTGS1	RP11-1094M14.5	5.76	8.62E-09	0.004	6.03	1.65E-09	0.000	3
rs145493205_ENSG00000249868	1	8	RP11-63E5.6	LINC00115	-5.33	9.64E-08	0.041	-5.98	2.27E-09	0.001	2
rs12939138_ENSG00000095303	17	9	PTGS1	RP11-1094M14.11	5.33	9.65E-08	0.043	5.90	3.68E-09	0.001	5
rs11084544_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.17	6.07	1.25E-09	0.001	5.89	3.90E-09	0.001	9
rs11670871_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.17	6.15	7.56E-10	0.000	5.89	3.90E-09	0.001	9
rs13227497_ENSG00000128918	7	15	ALDH1A2	RP11-611L7.1	5.39	7.20E-08	0.033	5.87	4.26E-09	0.001	4

Supplementary Table 2. *Continued)*

ID	SNP	GENE		lncRNA	RS			BIOS (no RS)			*#cis*
	CHR	CHR	SYMBOL		Effect	P	FDR	Effect	P	FDR	
rs12923514_ENSG00000239906	16	1	RP11-34P13.14	RP11-356C4.5	5.33	9.65E-08	0.043	5.85	4.91E-09	0.001	5
rs74427538_ENSG00000095303	17	9	PTGS1	CTC-507E2.1	5.36	8.34E-08	0.038	5.83	5.39E-09	0.001	5
rs77706710_ENSG00000236438	16	3	FAM157A	Z84812.4	5.91	3.32E-09	0.002	5.79	7.13E-09	0.001	4
rs12980907_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.19	6.13	8.93E-10	0.000	5.79	7.16E-09	0.001	9
rs77706710_ENSG00000231512	16	1	RP11-261C10.2	Z84812.4	6.37	1.92E-10	0.000	5.79	7.23E-09	0.001	4
rs11665688_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.19	6.13	8.93E-10	0.000	5.76	8.64E-09	0.002	9
rs11666502_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.19	6.13	8.93E-10	0.000	5.76	8.64E-09	0.002	9
rs11668821_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.19	6.13	8.93E-10	0.000	5.76	8.64E-09	0.002	9
rs12711778_ENSG00000237543	2	9	RP11-58A12.3	AC017074.1	-5.91	3.32E-09	0.001	-5.73	1.00E-08	0.002	4
rs12972898_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.19	6.21	5.39E-10	0.000	5.68	1.36E-08	0.003	9
rs4795085_ENSG00000095303	17	9	PTGS1	CTC-507E2.1	-5.60	2.18E-08	0.008	-5.64	1.66E-08	0.004	3
rs61770173_ENSG00000249868	1	8	RP11-63E5.6	FAM87B	-5.36	8.33E-08	0.036	-5.62	1.89E-08	0.004	2
rs12981649_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.16	6.18	6.39E-10	0.000	5.57	2.60E-08	0.006	9
rs185161330_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.14	6.18	6.39E-10	0.000	5.54	2.94E-08	0.006	9
rs56332086_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.14	6.18	6.39E-10	0.000	5.54	2.94E-08	0.006	9
rs6001798_ENSG00000175772	22	2	AC112229.7	BMS1P20	5.49	3.99E-08	0.017	5.54	3.03E-08	0.006	4
rs11671591_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.14	5.97	2.40E-09	0.001	5.53	3.17E-08	0.007	9
rs56069262_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.14	6.23	4.55E-10	0.000	5.45	5.15E-08	0.011	9
rs4849288_ENSG00000225655	2	9	RP11-143M1.2	AC104653.1	-5.31	1.11E-07	0.047	-5.42	5.85E-08	0.012	3
rs142862805_ENSG00000250390	20	11	RP11-338H14.1	RP4-610C12.3	5.57	2.54E-08	0.010	5.41	6.16E-08	0.013	0
rs6087197_ENSG00000250390	20	11	RP11-338H14.1	RP4-610C12.4	5.57	2.54E-08	0.010	5.40	6.50E-08	0.013	0
rs6087352_ENSG00000250390	20	11	RP11-338H14.1	RP4-610C12.4	5.57	2.54E-08	0.010	5.40	6.50E-08	0.013	0
rs77923699_ENSG00000250390	20	11	RP11-338H14.1	RP4-610C12.4	5.68	1.38E-08	0.006	5.38	7.32E-08	0.015	0
rs34188294_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.14	6.58	4.66E-11	0.000	5.38	7.57E-08	0.015	9
rs3131971_ENSG00000249868	1	8	RP11-63E5.6	FAM87B; RP11-206L10.10	-5.39	7.20E-08	0.031	-5.26	1.42E-07	0.027	3
rs55710326_ENSG00000071205	19	4	ARHGAP10	CTD-2619J13.14	6.02	1.73E-09	0.001	5.19	2.16E-07	0.037	9

Supplementary Table 3. The effect sizes (Z-scores) and p-values of the two *trans*-eQTLs replicating in the BIOS consortium (RNA-seq based data) and the CHARGE consortium (RNA-array based data).

SNP	Gene	Distance (kb)	RS			BIOS-replication			CHARGE-replication		
			N	Z-score	P-value	N	Z-score	P-value	N	Z-score	P-value
rs7440274	ENSG00000102893	NA	652	-5.89	3.90E-09	1,464	-6.21	5.37E-10	5,102	-7.81	5.81E-15
rs12939138	ENSG00000095303	NA	652	5.33	9.65E-08	1,464	5.90	3.68E-09	4,976	7.15	8.58E-13

ACKNOWLEDGEMENTS

This work was done within the framework of the Biobank-Based Integrative Omics Studies (BIOS) Consortium funded by BBMRI-NL, a research infrastructure financed by the Dutch government (NWO 184.021.007) and financially supported by the European Unions Seventh Framework Program IDEAL (FP8/2007-2011) under grant agreement No. 259679.

The infrastructure for the CHARGE Consortium is supported in part by the National Heart, Lung, and Blood Institute grant R01HL105756. This study was funded by the European Commission (HEALTH-F2-2008-201865, GEFOS; HEALTH-F2-2008 35627, TREAT-OA), Netherlands Organisation for Scientific Research (NWO) Investments (nr. 175.010.2005.011, 911-03-012), the Netherlands Consortium for Healthy Aging , the Netherlands Genomics Initiative [45] / Netherlands Organisation for Scientific Research (NWO) project nr. 050-060-810 and Vidi grant 917103521. Additional acknowledgments to specific cohorts and their support are found in the Supplementary Notes.

AFFILIATIONS

BIOS Consortium

Bastiaan T. Heijmans, Peter A.C. 't Hoen, Joyce B.J. van Meurs, Aaron Isaacs, Rick Jansen, Lude Franke, André G. Uitterlinden, Bert A. Hofman, Carla J.H. van der Kallen, Casper G. Schalkwijk, *Cis*ca Wijmenga, Coen D.A. Stehouwer, Cornelia M. van Duijn, Dasha V. Zhernakova, Diana van Heemst, Dorret I. Boomsma, Erik W. van Zwet, Ettje F. Tigchelaar, Freerk van Dijk, H. Eka D. Suchiman, Hailiang Mei, Irene Nooren, Jan Bot, Jan H. Veldink, Jenny van Dongen, Jeroen van Rooij, Joris Deelen, Jouke J. Hottenga, Leonard H. van den Berg, Maarten van Iterson, Marc Jan Bonder, Marian Beekman, Marijn Verkerk, Marleen M.J. van Greevenbroek, Martijn Vermaat, Matthijs Moed, Michael M.P.J. Verbiest, Michiel van Galen, Morris A. Swertz, Nico Lakenberg, P. Eline Slagboom, P. Mila Jhamai, Patrick Deelen, Peter van 't Hof, René Luijk, René Pool, Ruud van der Breggen, Sasha Zhernakova, Szymon M. Kielbasa, Wibowo Arindrarto

CHARGE Consortium

Anni Joensuu (DILGOM), Johannes Kettunen (DILGOM), Urmo Vosa (EGCUT), Tonu Esko (EGCUT), Harm-Jan Westra (FEHRMANN), Lude Franke (FEHRMANN), Hanieh Yaghootkar (INCHIANTI), Timothy M. Frayling (INCHIANTI), Katharina Schramm (KORA), Holger Prokisch (KORA), Tim Kacprowski (SHIP), Alexander Teumer (SHIP).

REFERENCES

1. Fehrmann RS, Jansen RC, Veldink JH, Westra HJ, Arends D, et al. (2011) Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. Plos Genetics 7: e1002197.
2. Huan T, Liu C, Joehanes R, Zhang X, Chen BH, et al. (2015) A systematic heritability analysis of the human whole blood transcriptome. Hum Genet 134: 343-358.
3. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, et al. (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet 45: 1238-1243.
4. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464: 768-772.
5. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, et al. (2010) Multiple common variants for celiac disease influencing immune gene expression. Nat Genet 42: 295-302.
6. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, et al. (2010) Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. PLoS One 5: e10693.
7. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, et al. (2012) Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat Genet 44: 502-510.
8. Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, et al. (2012) Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet 44: 1084-1089.
9. Small KS, Hedman AK, Grundberg E, Nica AC, Thorleifsson G, et al. (2011) Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nat Genet 43: 561-564.
10. Esteller M (2011) Non-coding RNAs in human disease. Nat Rev Genet 12: 861-874.
11. Feng J, Bi C, Clark BS, Mady R, Shah P, et al. (2006) The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. Genes Dev 20: 1470-1484.
12. Panganiban G, Rubenstein JL (2002) Developmental functions of the Distal-less/Dlx homeobox genes. Development 129: 4371-4386.
13. Wang Y, Zhong H, Xie X, Chen CY, Huang D, et al. (2015) Long noncoding RNA derived from CD244 signaling epigenetically controls CD8+ T-cell immune responses in tuberculosis infection. Proc Natl Acad Sci U S A 112: E3883-3892.
14. Zhang H, Zeitz MJ, Wang H, Niu B, Ge S, et al. (2014) Long noncoding RNA-mediated intrachromosomal interactions promote imprinting at the Kcnq1 locus. J Cell Biol 204: 61-75.
15. Li G, Zhang H, Wan X, Yang X, Zhu C, et al. (2014) Long noncoding RNA plays a key role in metastasis and prognosis of hepatocellular carcinoma. Biomed Res Int 2014: 780521.
16. Hofman A, Brusselle GG, Darwish Murad S, van Duijn CM, Franco OH, et al. (2015) The Rotterdam Study: 2016 objectives and design update. Eur J Epidemiol 30: 661-708.
17. Thewissen MM, Damoiseaux JG, Duijvestijn AM, van Greevenbroek MM, van der Kallen CJ, et al. (2011) Abdominal Fat Mass Is Associated With Adaptive Immune Activation: The CODAM Study. Obesity 19: 1690-1698.
18. Scholtens S, Smidt N, Swertz MA, Bakker SJ, Dotinga A, et al. (2014) Cohort Profile: LifeLines, a three-generation cohort study and biobank. Int J Epidemiol.
19. Tigchelaar EF, Zhernakova A, Dekens JA, Hermes G, Baranska A, et al. (2015) Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open 5: e006772.
20. Westendorp RG, van Heemst D, Rozing MP, Frolich M, Mooijaart SP, et al. (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.
21. Inouye M, Silander K, Hamalainen E, Salomaa V, Harald K, et al. (2010) An Immune Response Network Associated with Blood Lipid Levels. Plos Genetics 6.
22. Metspalu A (2004) The Estonian Genome Project. Drug Development Research 62: 97-101.
23. Tanaka T, Shen J, Abecasis GR, Kisialiou A, Ordovas JM, et al. (2009) Genome-Wide Association Study of Plasma Polyunsaturated Fatty Acids in the InCHIANTI Study. Plos Genetics 5.

24. Wichmann HE, Gieger C, Illig T, Grp MKS (2005) KORA-gen - Resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen 67: S26-S30.
25. Hofman A, Darwish Murad S, van Duijn CM, Franco OH, Goedegebure A, et al. (2013) The Rotterdam Study: 2014 objectives and design update. Eur J Epidemiol 28: 889-926.
26. Volzke H, Alte D, Schmidt CO, Radke D, Lorbeer R, et al. (2011) Cohort Profile: The Study of Health in Pomerania. International Journal of Epidemiology 40: 294-307.
27. Zhernakova DV, Deelen P, Vermaat M, Iterson Mv, Jansen R, et al. (2015) Unbiased identification of regulatory modifiers of genetic risk factors. Submitted
28. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal Vol. 17.
29. Joshi NA, Fass JN (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33).
30. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15-21.
31. Flicek P, Amode MR, Barrell D, Beal K, Billis K, et al. (2014) Ensembl 2014. Nucleic Acids Res 42: D749-755.
32. PicardTools (2015).
33. Aanes H, Winata C, Moen LF, Ostrup O, Mathavan S, et al. (2014) Normalization of RNA-Sequencing Data from Samples with Varying mRNA Levels. PLoS One 9.
34. Westra HJ, Jansen RC, Fehrmann RSN, Meerman GJT, Van Heel D, et al. (2011) MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. Bioinformatics 27: 2104-2111.
35. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, et al. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22: 1760-1774.
36. Boone DN, Lee AV (2015) SNHG7 Is an Insulin-like Growth Factor (IGF1) Regulated Long Non-Coding RNA Necessary for Proliferation.
37. Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42: D1001-1006.
38. Costigan M, Belfer I, Griffin RS, Dai F, Barrett LB, et al. (2010) Multiple chronic pain states are associated with a common amino acid-changing allele in KCNS1. Brain 133: 2519-2527.
39. Styrkarsdottir U, Thorleifsson G, Helgadottir HT, Bomer N, Metrustry S, et al. (2014) Severe osteoarthritis of the hand associates with common variants within the ALDH1A2 gene and with rare variants at 1p31. Nat Genet 46: 498-502.
40. Jiang GL, Huang S (2000) The yin-yang of PR-domain family genes in tumorigenesis. Histol Histopathol 15: 109-117.
41. Burwinkel B, Maichele AJ, Aagenaes O, Bakker HD, Lerner A, et al. (1997) Autosomal glycogenosis of liver and muscle due to phosphorylase kinase deficiency is caused by mutations in the phosphorylase kinase beta subunit (PHKB). Hum Mol Genet 6: 1109-1115.
42. Schror K, Rauch BH (2015) Aspirin and lipid mediators in the cardiovascular system. Prostaglandins Other Lipid Mediat.
43. Lederer B, Van Hoof F, Van den Berghe G, Hers H (1975) Glycogen phosphorylase and its converter enzymes in haemolysates of normal human subjects and of patients with type VI glycogen-storage disease. A study of phosphorylase kinase deficiency. Biochem J 147: 23-35.
44. Saxena A, Carninci P (2011) Long non-coding RNA modifies chromatin: epigenetic silencing by long non-coding RNAs. Bioessays 33: 830-839.
45. Rotimi C, Abayomi A, Abimiku A, Adabayeri VM, Adebamowo C, et al. (2014) Research capacity. Enabling the genomic revolution in Africa. Science 344: 1346-1348.

CHAPTER 4

Genomic analysis integration for age-related musculoskeletal comorbidities

CHAPTER 4.1

Genome-wide association study meta-analysis of chronic widespread pain: evidence for involvement of the 5p15.2 region

Marjolein J. Peters*, Linda Broer*, Hanneke L.D.M. Willemen*, Gudny Eiriksdottir, Lynne J. Hocking, Kate L. Holliday, Michael A. Horan, Ingrid Meulenbelt, Tuhina Neogi, Maria Popham, Carsten O. Schmidt, Anushka Soni, Ana M. Valdes, Najaf Amin, Elaine M Dennison, Niels Eijkelkamp, Tamara B. Harris, Deborah J. Hart, Albert Hofman, Frank J.P.M. Huygen, Karen A. Jameson, Gareth T. Jones, Lenore J. Launer, Hanneke J.M. Kerkhof, Marjolein de Kruijf, John McBeth, Margreet Kloppenburg, William Ollier, Ben Oostra, Antony Payton, Fernando Rivadeneira, Blair H. Smith, Albert V. Smith, Lisette Stolk, Alexander Teumer, Wendy Thomson, André G. Uitterlinden, Ke Wang, Sophie H. van Wingerden, Nigel K. Arden, Cyrus Cooper, David Felson, Vilmundur Gudnason, Gary J. Macfarlane, Neil Pendleton, P. Eline Slagboom, Tim D. Spector, Henry Völzke, Annemieke Kavelaars*, Cornelia M. van Duijn*, Frances M. K. Williams*, Joyce B. J. van Meurs*

** These authors contributed equally to this work*

Published: *Ann Rheum Dis 2013;* **72***: doi:10.1136/annrheumdis-2012-201742*

ABSTRACT

Objectives: Chronic widespread pain (CWP) is a common disorder affecting ~10% of the general population and has an estimated heritability of 48-52%. In the first large-scale genome-wide association study (GWAS) meta-analysis, we aimed to identify common genetic variants associated with CWP.

Methods: We conducted a GWAS meta-analysis in 1,308 female CWP cases and 5,791 controls of European descent, and replicated the effects of the genetic variants with suggestive evidence for association in 1,480 CWP cases and 7,989 controls. Subsequently, we studied gene expression levels of the nearest genes in two chronic inflammatory pain mouse models, and examined 92 genetic variants previously described associated with pain.

Results: The minor C-allele of rs13361160 on chromosome 5p15.2, located upstream of *CCT5* and downstream of *FAM173B*, was found to be associated with a 30% higher risk of CWP (MAF=43%; OR=1.30, 95%CI=1.19-1.42, P=1.2 x 10^{-8}). Combined with the replication, we observed a slightly attenuated OR of 1.17 (95%CI=1.10-1.24, P=4.7 x 10^{-7}) with moderate heterogeneity (I2=28.4%). However, in a sensitivity analysis that only allowed studies with joint-specific pain, the combined association was genome-wide significant (OR=1.23, 95%CI=1.14-1.32, P=3.4 x 10^{-8}, I2=0%). Expression levels of *Cct5* and *Fam173b* in mice with inflammatory pain were higher in the lumbar spinal cord, not in the lumbar dorsal root ganglions, compared to mice without pain. None of the 92 genetic variants previously described were significantly associated with pain (P>7.7x10-4).

Conclusions: We identified a common genetic variant on chromosome 5p15.2 associated with joint-specific CWP in humans. This work suggests that *CCT5* and *FAM173B* are promising targets in the regulation of pain.

INTRODUCTION

Chronic widespread pain (CWP) is a common disorder, affecting about 10% of the general population [1]. The prevalence of CWP increases with age for both men and women, but is more common in women at any age [1]. CWP represents a major underestimated health problem and is associated with substantial impairment and a reduced quality of life. It has been related to a number of physical and affective symptoms such as fatigue, psychological distress and somatic symptoms [1,2]. Chronic musculoskeletal pain is one of the most common conditions seen in rheumatology clinics and accounts for 6.2% of the total healthcare costs in the Netherlands every year [3]. Further research is needed to be able to understand the causal mechanisms and optimal treatment for CWP patients.

CWP causally relates to an initial local pain stimulus, such as an acute injury or athletic injuries or another pain state such as low back pain or local pain due to osteoarthritis (OA) or rheumatic arthritis (RA) [4-6]. However, most injured subjects do not develop chronic widespread pain, and only a part of patients with OA or RA develop CWP. We therefore hypothesize that several discrete stimuli may initiate CWP via a common final pathway that involves the generation of a central pain state through the sensitization of second order spinal neurons.

CWP is a complex trait since both environmental and genetic factors play a role in the etiology. Heritability estimates of twin studies suggest that 48%-52% of the variance in CWP occurrence is due to genetic factors implying a strong genetic component [7]. A number of studies have examined genetic variants for CWP. These candidate gene studies examined polymorphisms in genes involved in both the peripheral and the central nervous system [8]. In particular, genes involved in neurotransmission (pathway of dopamine and serotonin [9-19]), and genes important for the hypothalamic-pituitary-adrenal (HPA) axis have been considered [20]. A number of genetic variants in these candidate genes were found to be associated with CWP, individual pain sites, or experimental pain. However, no consistent significant associations have been demonstrated.

The most studied gene in relation to pain is *COMT* (catechol-O-methyltransferase); an enzyme that degrades neurotransmitters including dopamine. The variant allele of rs4680 (or V158M) results in reduced enzymatic activity due to its effect on thermostability [21], and has been associated with reduced opoid activity in response to painful stimuli resulting in increased pain sensitivity [22]. But also for *COMT*, no consistent results have been observed in genetic association studies [13,23-29].

Overall, the results have been conflicting which is likely due to the modest sample sizes used and paucity of replication. In general, candidate studies are biased by previous knowledge of the etiology of the disease under study. Since knowledge about the pathophysiology of CWP is poor, the chances of success using this approach are low. Therefore our objective was to identify genetic variants involved in CWP by means of a large-scale hypothesis-free genome-wide association study (GWAS) meta-analysis including 2,788 cases and 13,780 controls. To our knowledge, this is the first study presenting a large-scale GWAS meta-analysis of chronic pain. The prevalence of CWP is

approximately two times higher in women than in men and there is strong evidence that women tolerate less thermal and pressure pain than men [30]. Therefore only women were included in this study to reduce heterogeneity and thereby increase power.

MATERIALS AND METHODS

We performed a meta-analysis (stage 1) of GWAS data of 1,308 female Caucasian CWP cases and 5,791 female Caucasian controls, derived from five studies, and focussed our follow-up efforts on the SNPs with suggestive evidence of association ($P<1 \times 10^{-5}$) with CWP (stage 2). The study outline is summarized in Figure 1.

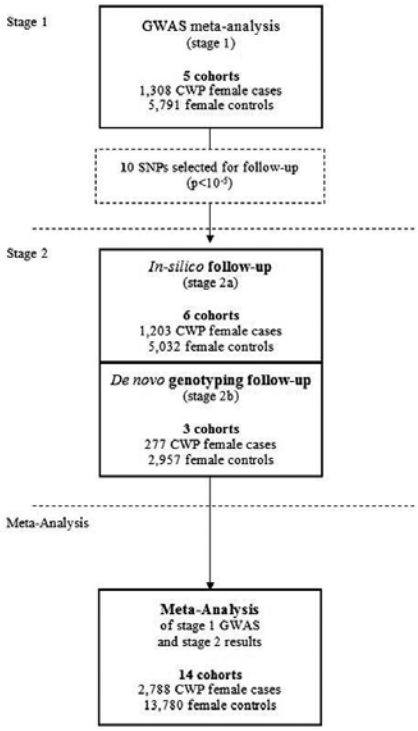


Figure 1. Study outline.

CWP = chronic widespread pain; GWAS = genome-wide association study.

Phenotype

Chronic Widespread Pain (CWP) was defined as subjects having pain in the left side of the body, in the right side of the body, above waist, below waist, and in the axial skeleton (following the *Fibromyalgia* Criteria of the American College of Rheumatology [2]). Controls were defined as subjects not having CWP. Subjects using analgesics (ATC-code: N02 [31]) were excluded from the control group. Detailed descriptions of the study specific inclusion criteria are presented in Supplementary Table 1.

Study Design Summary

We combined the summary statistics of GWAS in a meta-analysis comprising 1,308 CWP female Caucasian cases and 5,791 female Caucasian controls (stage 1). We focussed our follow-up efforts on the SNPs with suggestive evidence of association ($P<1 \times 10^{-5}$) with CWP in 1,480 CWP cases and 7,989 controls available for replication (stage 2).

Subjects

A full detailed description of all study cohorts is presented in Table 1 and in the Supplementary Methods section. For the stage 1 analysis, we included studies from the Netherlands: the Erasmus Rucphen Family study (ERF study) [32], Rotterdam Study I, II and III (RS-I, RS-II and RS-III) [33]; and the United Kingdom: TwinsUK [34,35]. All studies were approved by their institutional ethics review committees and all participants provided written informed consent. For our stage 2 analysis, we sought follow-up samples with pre-existing GWAS *in-silico* data (stage 2a) as well as *de novo* genotyping (stage 2b). The studies are from the United Kingdom: the British 1958 Birth Cohort (1958BC) [23,36-38], the Chingford Study (CHINGFORD) [39,40], the Dyne Steel DNA Bank for Aging and Cognition (DSDBAC) [41], the EPIdemiological study of FUNctional Disorders (EPIFUND) [20], and the Hertfordshire Cohort Study (HCS) [42]; from Iceland: the Age, Gene/Environment Susceptibility Study (AGES) [43]; from the United States: Framingham Osteoarthritis Study (FOA) [44]; from the Netherlands: the Genetics osteoARthritis and Progression Study (GARP) [45]; and from Germany: the *Study* of Health In Pomerania (SHIP) [46,47]. All studies were approved by the local ethics committees and all participants provided written informed consent.

Genotyping, Quality Control and Imputation

Genotyping of the stage 1 cohorts was done by Illumina Infinium HumanHap550 Beadchip (RS-I and RS-II), the Illumina Infinium HumanHap610 (RS-II, RS-III, and TwinsUK), or the Illumina Infinium HumanHap300 (ERF and TwinsUK). More details about the genotyping, Quality Control, and Imputation are shown in the Supplementary Methods section. Complete information on genotyping protocols and QC measures for all stage 1 cohorts is described in the Supplementary Material (Supplementary Table 2).Detailed descriptions of the QC and imputation procedures are provided in the Supplementary Material (Supplementary Table 3).

Genotypes of the stage 2a studies (1958BC, AGES, DSDBAC, FOA, GARP, and SHIP) were obtained from SNP arrays and imputed data. Where unavailable, proxy SNPs were selected based on high linkage disequilibrium. The stage 2b studies (CHINGFORD, EPIFUND, and HCS) performed *de novo* genotyping, using both Sequenom iPLEX and TaqMan-based assays (Supplementary Methods). Genotyping platforms, calling algorithms, quality control before imputation, imputation methods, and analysis software used were all study-specific (Supplementary Tables 4 and 5). The explicit number of follow-up SNPs genotyped in the different studies and whether the original or a proxy SNP was used is summarized in Supplementary Table 6.

GWAS-analysis in the Stage 1 Studies

CWP was analyzed as a binary trait (cases versus controls) using logistic regression under an additive model with adjustment for age and BMI (Supplementary Table 7). To adjust for population substructure, we included the 4 most important PCs as covariates in the regression analysis of RS-I, RS-II, and RS-III. These PCs were derived from an multidimensional scaling analysis of identity-by-state distances, using PLINK software [48]. Detailed descriptions of the GWAS methods are provided in Supplementary Table 8).

Stage 1 GWAS Meta-Analysis

P-values for association were combined using the Meta-Analysis Tool for genome-wide association scans (*METAL)* [49]. The genomic control method [50] as implemented in METAL was used to correct for any residual population stratification or relatedness not accounted for by the four most important PCs. A P-value<5×10^{-8} was considered genome-wide significant while a P-value<1×10^{-5} was considered suggestive [51]. Power calculations were performed using CaTS software (*www.sph.umich.edu/csg/abecasis/CaTS/*). Using Bonferroni correction (P<5×10^{-8}), power calculations showed that we had approximately 80% power to detect an odds ratio (OR) of 1.30 for SNPs with a minor allele frequency (MAF) of 0.43, given a disease prevalence of 10% for 1,308 cases and 5,791 controls in the discovery group. Using a P-value<1×10^{-5}, we had 80% power to detect an OR of 1.25.

SNP selection for replication

We aimed to select SNPs for replication (stage 2) that were enriched for signals of association with CWP. All SNPs with suggestive evidence for association in the stage 1 analyses were selected and separated into independent loci by taking the most significantly associated SNP and eliminating all SNPs that have a HapMap CEU pair-wise correlation coefficient r^2>0.8 with that SNP using the PLINK software.

Meta-analysis of stage 1 and stage 2 results

We combined the stage 1 and stage 2 association results to derive a combined meta-analysis for the suggestively associated loci. METAL was used to conduct a fixed-effects meta-analysis as in stage 1. Estimated heterogeneity variance and forest plots were generated using Comprehensive Meta-Analysis (www.meta-analysis.com).

Functional analysis of associated SNPs

To determine whether the associated SNPs have any regulatory effect on gene expression levels, we checked their effect (and the effect of the linked SNPs) on the expression levels of their neighbouring genes. We used the 1000 genomes data in the SNAP software [52,53] to identify those SNPs having a linkage disequilibrium (LD) thresholds of r^2>0.1. We searched two publicly available eQTL databases: the NCBI GTEx (Genotype-Tissue Expression) eQTL Browser (*http://www.ncbi.nlm.nih.gov/gtex/GTEX2/gtex.cgi*) and the expression Quantitative Trait Loci database (*http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/*). We used SIFT [54] to predict whether the coding non-synonymous variant causing an amino acid substitution affects protein function.

Table 1. Overview of all participating studies.

Study (stage)	Reference article	Study design	Ethnic origin	Country of origin	Medication	Age/BMI	Mean age (y)	# CWP cases	# CWP controls
Stage 1									
ERF study	[32]	Family Based Cohort	Caucasian	the Netherlands	Y	Y	46.4	149	665
RS-I	[33]	Population Based Cohort	Caucasian	the Netherlands	Y	Y	69.4	563	1,892
RS-II	[33]	Population Based Cohort	Caucasian	the Netherlands	Y	Y	67.9	110	668
RS-III	[33]	Population Based Cohort	Caucasian	the Netherlands	Y	Y	56.3	85	868
TWINSUK	[34, 35]	Twins Based Cohort	Caucasian	United Kingdom	Y	Y	51.9	401	1,698
Total # samples							**59.7**	**1,308**	**5,791**
Stage 2a									
1958BC	[23, 36-38]	Prospective Birth Cohort	Caucasian	United Kingdom	N	Y (born in 1958)	NA	315	2,206
AGES	[43]	Population Based Cohort	Caucasian	Iceland	Y	Y	76.5	173	1,204
DSDBAC	[41]	Population Based Cohort	Caucasian	United Kingdom	Y	Age only	80.1	81	219
FOA	[44]	Population Based Cohort	Caucasian	United States	Y	Y	59.3	384	814
GARP	[45]	Case Control Based	Caucasian	the Netherlands	Y	Y	58.5	67	925*
SHIP	[46, 47]	Population Based Cohort	Caucasian	Germany	Y	Y	57.6	183	589
Stage 2b									
CHINGFORD	[39, 40]	Population Based Cohort	Caucasian	United Kingdom	Y	Y	56.6	48	337
EPIFUND	[20]	Population Based Cohort	Caucasian	United Kingdom	N	Age only	49.0	139	503
HCS	[42]	Population Based Cohort	Caucasian	United Kingdom	Y	Y	66.4	90	2,117
Total # samples								**1,480**	**7,989**

**GARP consists of clinical and radiographically confirmed OA case only; therefore we used 925 randomly chosen Rotterdam Study samples as controls. BMI = body mass index; Medication Y: information about medication use available; Medication N: medication use not available; Age/BMI Y: age and BMI data are available; Age only: no BMI data are available; ERF study = Erasmus Rucpen Family study; RS = Rotterdam Study; TwinsUK = the United Kingdom Adult Twin Registry; 1958BC = 1958 Birth Cohort; AGES = Age, Gene/Environment Susceptibility study Reykjavik; DSDBAC = Dyne Steel DNA Bank for Ageing and Cognition; FOA = Framingham Osteoarthritis Study; GARP = Genetics OsteoArthritis and Progression study Leiden; SHIP = Study of Health In Pomerania; CHINGFORD = Chingford 1000 Women Study; EPIFUND = EPIdemiological study of FUNctional Disorders study; HCS = Hertfordshire Cohort Study.*

RNA expression analyses in mice

For functional follow-up, two independent mouse models of inflammatory pain were studied. The first model was based on carrageenan injections; female C57Bl/6 mice received an intraplantar injection of 20 µl λ-carrageenan (2% (w/v), Sigma-Aldrich) in saline in both hind paws [55]. The second model was based on Complete Freund's Adjuvant (CFA) injections; male C57Bl/6 mice (Harlan Laboratories) received an intraplantar injection of 20 µl CFA (Sigma-Aldrich) in saline in both hind paws [56]. Controls were injected with saline only. At day 3 (after CFA injection) or day 6 (after carrageenan injection), thermal sensitivity (heat withdrawal latency time) was measured using the Hargreaves (IITC Life Science, Woodland Hills, CA) test as described [57]. Intensity of the light beam was chosen to induce heat withdrawal latency time of approximately 8 seconds at baseline.

After measurement the mice were sacrificed and the lumbar (L2-L5) spinal cord and the dorsal root ganglions (DRG) (L2-L5) were isolated. These areas of spinal cord and DRG were selected because pain transmission from the hind paws is mediated via primary sensory neurons that have their cell bodies in the lumbar DRG, and transmit the signal to the lumbar spinal cord through sensory fibres in the dorsal roots. Total RNA was isolated and mRNA levels of *Cct5* and *Fam173b* were measured in the spinal cord and the DRG. For more details, see the Supplementary Methods section.

All experiments were performed in accordance with international guidelines and approved by the experimental animal committee of University Medical Center Utrecht (carrageenan experiment) or the United Kingdom Home Office Animals (Scientific Procedures) Act 1986 (CFA experiment). Mice used for the carrageenan experiment were bred and maintained in the animal facility of the University of Utrecht (the Netherlands).

Systematic review of genetic variants previously described

We systematically searched for associations earlier reported with pain in the HugeNavigator PhenoPedia Database [58]. We used the search term "pain" and checked all publications for genes and SNPs associated with pain at least twice. Genes and SNPs associated with drug therapy, facial pain, migraine, and postoperative pain were excluded. For all reported SNPs, we examined their association with CWP in our stage 1 meta-analysis. The significance threshold was set at $P<8 \times 10^{-4}$ using Bonferroni correction for 65 independent genetic loci. Again, power calculations were performed using CaTS software (*www.sph.umich.edu/csg/abecasis/CaTS/*). With an alpha level of 8×10^{-4}, power calculations showed that we had approximately 80% power to detect an OR of 1.22 for SNPs with an minor allele frequency of 20% or higher.

RESULTS

GWAS meta-analysis for CWP

The Manhattan plot and Quantile-Quantile (QQ) plot of the initial stage 1 meta-analysis are presented in Figure 2. In total, 2,224,068 SNPs (directly genotyped or imputed) were tested for association. The overall genomic control lambda (λ_{GC}) was 1.007, indicating no significant population stratification. We identified two SNPs which were genome-wide significant ($P<5 \times 10^{-8}$), and another 39 SNPs with suggestive evidence for association ($P<1 \times 10^{-5}$) located in ten independent genomic regions. The most significant association was observed for two imputed highly correlated SNPs ($r^2=0.97$) located upstream of the Chaperonin-Containing-TCP1-complex-5 gene (*CCT5*) and downstream of the FAMily with sequence similarity 173, member B gene (*FAM173B*) (rs13361160, $P=1.2x10^{-8}$ and rs2386592, $P=2.6 \times 10^{-8}$). For both SNPs, the minor allele (MAF=43%) was associated with a 30% higher risk for CWP (OR=1.30, 95%CI=1.19-1.42).

Meta-analysis of GWAS replication

For the ten independent SNPs with suggestive evidence, we pursued *in silico* replication data in six studies (stage 2a: 1,203 CWP cases and 5,032 controls) and performed *de novo* genotyping in subjects from three additional studies (stage 2b: 277 CWP cases and 2,957 controls) (detailed description of the studies is presented in Table 1 and Supplementary Methods). The summary results of the stage 1 and 2 meta-analysis are presented in Table 2.

After combining the results of stage 1 and stage 2, the top SNP was rs13361160 (OR=1.17, 95%CI=1.10-1.24, $P=4.7 \times 10^{-7}$, $I^2=28.4\%$). Figure 3 shows a forest plot of the association of rs13361160 with CWP across the stage 1 and stage 2 studies. The overall effect in the replication studies (stage 2 studies) was in a consistent direction but not significant (OR=1.06, 95%CI=0.98-1.16, P=0.16). In the combined analysis, moderate heterogeneity was observed ($I^2=28.4\%$). Supplementary Table 1 shows the different pain assessment methods used in the different studies to define CWP. Since four out of five stage 1 studies included joint-specific pain only (ERF, RS-I, RS-II, and RS-III), we performed a sensitivity analysis in which stage 2 cohorts using non-joint pain were excluded (1958BC, DSDBAC, EPIFUND, HCS, and SHIP). This resulted in a combined OR of 1.23 (95%CI=1.14-1.32, $P=3.4 \times 10^{-8}$, $I^2=0\%$). An overview of the results of the combined meta-analysis and the separate stage 1 and stage 2 analyses is presented in Table 3.

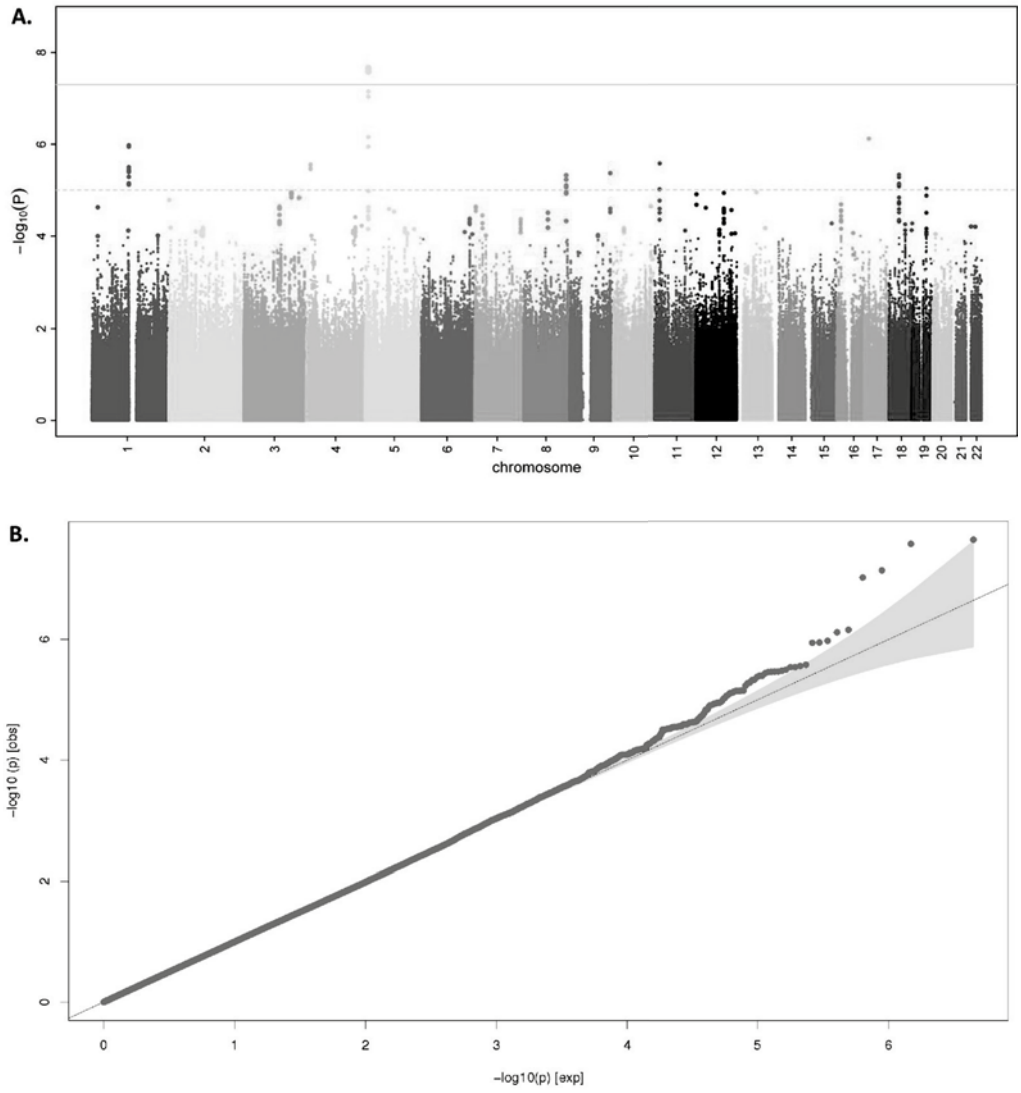


Figure 2. Genome-wide association results for chronic widespread pain (CWP) (stage 1). **(A)** Manhattan plot showing the p value of association tests for about 2 million SNPs with CWP in the stage 1 meta-analysis. SNPs are plotted on the x-axis according to their position on each chromosome. On the y-axis, the association P-values with CWP are shown (as –log 10 P-values). The grey solid horizontal line represents the P-value threshold of 5×10^{-8} (genome-wide significance). The grey dashed horizontal line represents the p value threshold of 1×10^{-5} (the level for suggestive evidence): SNPs in loci reaching 1×10^{-5} were tested for replication. **(B)** Quantile-quantile (QQ) plot of SNPs. The blue area represents the 95%CI around the test statistics. A QQ plot compares the additive model statistics to those expected under the null distribution using fixed effects for all analysed HapMAP CEU imputed SNPs passing quality control criteria.

Table 2. Association results of the 10 top hits.

SNP information					Gene information		Stage 1			Stage 2			Stage 1 & 2 (combined)		
SNP ID	CHR	MA	OA	MAF (%)	Nearest gene	Distance to gene (kb)	OR	95%CI	P	OR	95%CI	P	OR	95%CI	P
rs13361160	5	c	t	43.5	*FAM173B*	56.7	1.30	1.18-1.42	1.18×10^{-8}	1.06	0.98-1.16	0.161	1.17	1.10-1.24	4.67×10^{-7}
rs8065610	17	a	c	38.7	*PMP22*	41.9	1.26	1.15-1.38	3.86×10^{-7}	0.93	0.85-1.02	0.119	1.08	1.02-1.16	1.20×10^{-2}
rs12132674	1	a	g	29.5	*HMGCS2*	0.0	1.28	1.16-1.41	9.29×10^{-7}	1.07	0.97-1.17	0.165	1.16	1.09-1.24	1.23×10^{-5}
rs11606304	11	g	t	9.07	*TPH1*	0.0	0.55	0.43-0.70	1.47×10^{-6}	0.99	0.81-1.22	0.934	0.78	0.66-0.91	1.64×10^{-3}
rs7680363	4	a	t	6.42	*PROM1*	0.0	1.52	1.28-1.80	1.72×10^{-6}	1.04	0.86-1.25	0.688	1.27	1.12-1.44	1.52×10^{-4}
rs4837492	9	c	t	4.51	*FREQ*	56.2	1.23	1.13-1.34	2.96×10^{-6}	0.97	0.89-1.07	0.568	1.10	1.03-1.17	2.68×10^{-3}
rs524513	18	t	c	18.2	*BRUNOL4*	0.0	1.29	1.16-1.44	4.00×10^{-6}	0.96	0.85-1.09	0.537	1.13	1.04-1.23	2.75×10^{-3}
rs7835968	8	g	a	12.7	*KHDRBS3*	175.8	1.34	1.18-1.52	4.26×10^{-6}	0.98	0.86-1.12	0.787	1.16	1.06-1.27	1.43×10^{-3}
rs2249104	11	t	c	8.8	*MYOD1*	3.7	1.42	1.22-1.64	4.57×10^{-6}	0.99	0.84-1.17	0.882	1.21	1.08-1.35	8.78×10^{-4}
rs17796312	19	g	a	33.1	*FBL*	7.1	1.24	1.13-1.37	9.79×10^{-6}	1.08	0.97-1.2	0.166	1.17	1.09-1.25	2.56×10^{-5}

CHR = chromosome; MA = minor allele or effect allele (minor allele = effect allele); OA = other allele; MAF = minor allele frequency (%); OR = odds ratio; 95%CI = 95% confidence interval; P = P-value.

Table 3. Top hit association results.

Type of Analysis	Stage 1		Stage 2		Stage 1 & 2 (combined)	
SNP tested Adjustments	**OR (95%CI)**	**P**	**OR (95%CI)**	**P**	**OR (95%CI)**	**P**
rs13361160 Age, BMI, and 4 PCs	1.30 (1.19-1.42)	1.18×10^{-8}	1.06 (0.98-1.16)	0.16	1.17 (1.10-1.24)	4.67×10^{-7}
minor allele = C, other allele = T, MAF=43.5%						
Sensitivity analysis: joint pain only						
rs13361160 Age, BMI, and 4 PCs	1.30 (1.19-1.42)	1.18×10^{-8}	1.10 (0.97-1.25)	0.15	1.23 (1.14-1.32)	3.43×10^{-8}
minor allele = C, other allele = T, MAF=43.5%						

In both analyses the effect estimates of the models refer to the minor allele (=effect allele). BMI = body mass index; MAF = minor allele frequency; OR = odds ratio; 95%CI = 95% confidence interval; P = P-value.

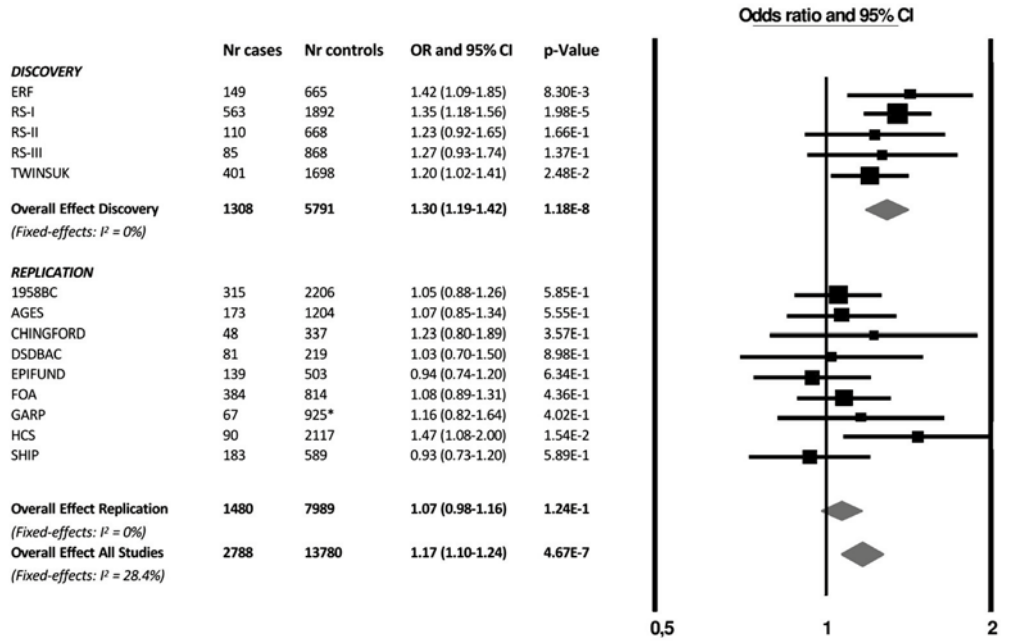


Figure 3. Forest plot of the association of rs13361160 SNP with chronic widespread pain (CWP). Study specific estimates and summary association between rs13361160 and CWP are shown.

Functional analysis of rs13361160 and rs2386592

The SNPs rs13361160 and rs2386592 (r^2=0.97) are annotated to the 5p15.2-region and located 81 kb upstream of *CCT5* and 57 kb downstream of *FAM173B* (Figure 4). We tested whether rs13361160 and rs2386592 and their linked SNPs (r^2>0.1) affected gene expression levels of *CCT5* or *FAM173B*. In total, we identified 130 SNPs in LD with our top SNPs of which two SNPs were located in the coding region: one synonymous SNP rs1042392 in the *CCT5* gene (r^2=0.16, D'=0.85) and one non-synonymous SNP rs2438652 in the *FAM173B* gene (r^2=0.17, D'=1.0) (Supplementary Table 9). The minor allele of rs2438652 causes a threonine-to-methionine substitution (T75M) which is thought to be functionally neutral. SNPs rs13361160 and rs2386592 were not recorded as influencing the expression levels of *CCT5* and *FAM173B*, however, the linked intronic SNP rs2445871 (r^2=0.14 for both) had a direct eQTL effect on *FAM173B* expression levels in liver tissue [59].

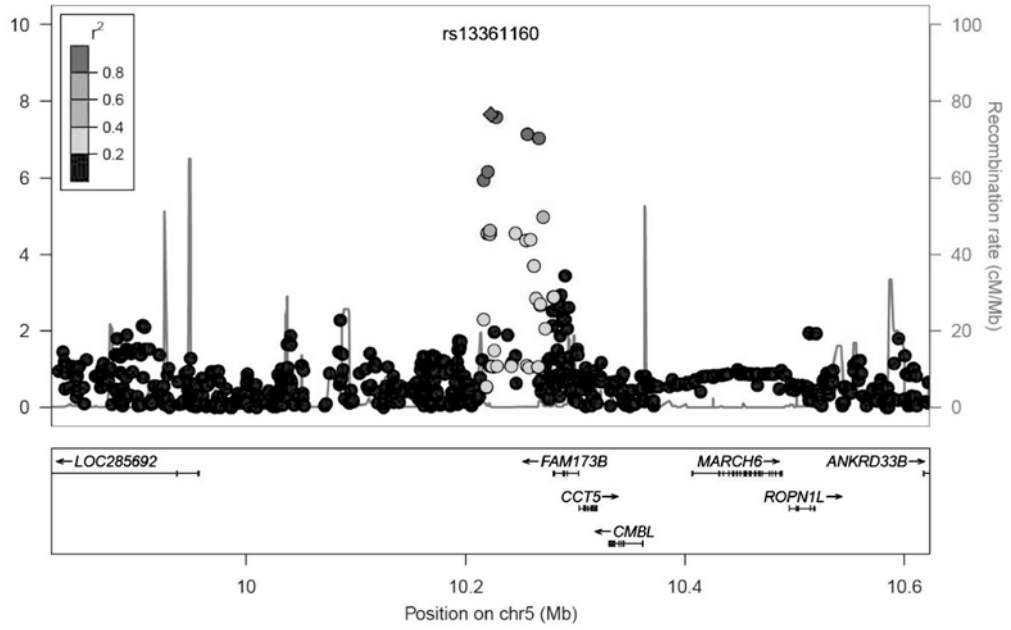


Figure 4. Regional plot of locus 5p15.2. On the x-axis, SNPs are plotted according to their position in a 400-kb window around rs13361160. On the y-axis, the association p values with chronic widespread pain are shown (as –log 10 p values). The purple diamond highlights the most significant SNP rs13361160. Blue peaks indicate recombination sites, and the SNPs surrounding the most significant SNP are color coded to identify their strength of linkage disequilibrium with the most significant SNP (pairwise r2 values of the HapMap CEU samples). Genes and the direction of transcription are shown at the bottom of the plot.

RNA expression analysis in mice

We studied gene expression levels of the two nearest genes *Cct5* and *Fam173b* in the lumbar spinal cord and the dorsal root ganglions (DRG) in two independent mouse models of chronic inflammatory pain. In both the carrageenan treated group and the Complete Freund's Adjuvant (CFA) treated group, mice had shorter heat withdrawal latency times than mice injected with saline only, confirming enhanced pain sensitivity (P<0.001) (Supplementary Figure 1).

The results from the multivariate analysis using the two genes (*Cct5* and *Fam173b* examined as dependent variables), the different treatments (saline, carrageenan, and CFA) and the different tissues (DRG and spinal cord) confirmed that there is a significant treatment effect for *Cct5* F(2,25)=3.399, p=0.0049, as well as for *Fam173b* F(2,25)=4.911, p=0.016. Moreover, both genes showed a significant tissue effect (*Cct5*: F(1,25)=13.595, p=0.001, and *Fam173b*: F(1,25)=13.522, p=0.001), as well as a significant interaction between tissue and treatment (*Cct5*: F(2,25)=6.424, p=0.006, and Fam173b: F(1,25)=4.196, p=0.027) (Figure 5). These findings indicate that in spinal cord but not in DRG both Fam173b and Cct5 expression levels were upregulated in response to two different inducers of inflammatory pain. *DRG Fam173b* and *Cct5* expression levels in CFA/carrageenan-treated mice were indistinguishable from saline-treated mice.

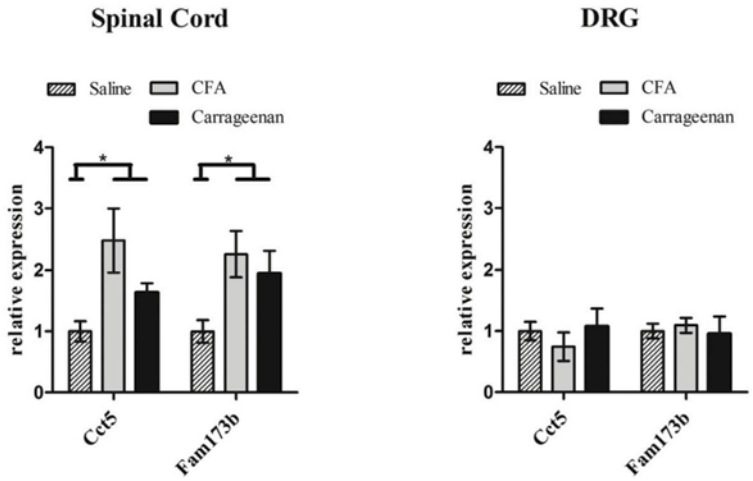


Figure 5. Quantitative PCR analysis of gene expression levels in the lumbar (L2–L5) spinal cord (A) and the dorsal root ganglions (DRG) (B) of mice after intraplantar saline (n=3), carrageenan (n=4), and Complete Freund's Adjuvant (CFA) (n=4) injection. Spinal cord and DRG were collected and analyzed for RNA levels of *Cct5* and *Fam173b*. Data were normalized for *Gapdh* and *β-actin* (housekeeping genes) expression. Data are expressed as mean±SEM, *=p<0.05.

Candidate SNPs previously associated with chronic pain.

We examined whether genetic variants previously described for association with pain were associated with CWP in our large stage 1 meta-analysis. We identified a total of 44 genes of which 136 SNPs had been reported at least twice with any pain phenotype (excluding facial pain, migraine, postoperative pain, and response to drug therapy), and we examined the association of these 136 SNPs with CWP in the GWAS stage 1 meta-analysis. Out of 136 candidate SNPs, we were able to check 92 common SNPs (MAF>5%) in 65 independent genetic loci (Supplementary Table 10). Five SNPs had a too low MAF (<=5%) and 39 SNPs were not genotyped or imputed in our meta-analysis. None of the earlier reported SNPs passed the significance threshold ($P<8 \times 10^{-4}$). Interestingly, the strongest associated SNPs are located in three genes that have been reported to be associated with pain phenotypes most frequently: *COMT* (Catechol-O-MethylTransferase), *GCH1* (GTP cyclo-hydrolase 1), and *OPRM1* (mu opioid receptor). The effects of the SNPs in *GCH1* are in the same direction as earlier reported [60-62]: individuals having the minor allele for rs10483639, rs4411417 or rs752688 have 15% less pain than those exhibiting the common alleles. The effect of the SNP rs599548 in *OPRM1* is also in the same direction as earlier reported [63]: those having the minor allele for rs599548 have 19% more pain than those exhibiting the major allele. The two *COMT* SNPs are in weak linkage disequilibrium with the well-known amino-acid changing variant rs4860, but previously have not been reported to be significantly associated with pain [23,64]. We have found a protective effect for the minor allele of rs2020917 (those having a minor allele have 15% less pain) and an adverse effect for the minor allele of rs5993883: those having the minor allele of rs5993883 have 14% more pain.

DISCUSSION

In this study, we identified a genetic variant near *CCT5* and *FAM173B* to be associated with CWP. Chronic pain coincided with higher RNA-expression of *Cct5* and *Fam173b* in the lumbar spinal cord of mouse models of inflammatory pain. This finding indicates that both genes in the 5p15.2 region are regulated in the context of inflammatory pain.

Interestingly, Bouhouche *et al.* [65] reported a human pedigree in which a *CCT5* mutation caused hereditary sensory neuropathy (OMIM=610150), a syndrome characterized by a sensory deficit in the distal portion of the lower extremities, chronic perforating ulcerations of the feet, and progressive destruction of underlying bones. Symptoms can include pain and numbness, tingling in the hands, legs or feet, and extreme sensitivity to touch. *CCT5* is a subunit of the chaperonin containing t-complex polypeptide 1 (TCP-1) which assists in protein folding and assembly in the brain [66]. *CCT5* interacts with the serine/threonine-protein phosphatase 4 catalytic subunit PP4C [67-69]. Zhang *et al*. [70] confirmed that protein phosphates like PPP4C may have a regulatory effect on the central sensitization of nociceptive transmission in the spinal cord. Interestingly, sensitization is thought to contribute to chronic inflammatory pain [71]. Since the function of the *FAM173B* gene is not yet known, it is difficult to postulate the mechanism by which this gene could influence CWP.

Further research into the genes in this locus is needed to ascertain whether either or both *CCT5* and *FAM173B* are driving the observed association.

By combining the effects across the different stage 2 studies, moderate heterogeneity was observed in the meta-analysis. This heterogeneity might be caused by different pain assessment methods used by the stage 2 cohorts. In particular, four cohorts asked the participants about joint pain specifically, while the other five also included non-joint pain. When the non-joint pain phenotype were excluded, the heterogeneity across the cohorts reduced to 0% and the overall P-value for rs13361160 now reached genome-wide significance by combining the stage 1 and stage 2 effects. This might suggest that indeed phenotype heterogeneity was introduced by including non-joint pain. In general, it is anticipated that pain is a very complex trait, with different ethiological pathways introducing phenotypic heterogeneity.

A limitation of our study is that we were not able to examine possible phenotype subgroups, such as individuals with rheumatic arthritis (RA), a chronic systemic inflammatory disorder that principally affects the synovial joints. Stratifying these groups of individuals might serve to increase power to find genetic loci. We here decided to analyse all CWP cases together, based on the hypothesis that several discrete stimuli need to initiate CWP via a common final pathway that involves the generation of a central pain state through the sensitization of second order spinal neurons. In addition, the prevalence of RA is very low (about 0.5-1%) [72], and the earlier defined GWAS hits for RA (i.e., the HLA-locus) [73] were not in our top list. So, we assume the results were not dominated by this small number of individuals with RA.

It would be helpful to dissect the phenotype of pain into quantitative sub-phenotypes, for example, by measuring pain sensitivity and pain thresholds for temperature or pressure [74], or by examining functional MRIs [75]. The use of quantitative and possibly more objective pain measurements in response to painful stimuli (rather than reported pain) will be of pivotal importance for future pain research. Because we have focused on the clinical pain definition using questionnaires and pain homunculus, we accept that we may have missed true pain susceptibility alleles. However, this study represents the largest genome-wide meta-analysis looking into the genetics of human chronic widespread pain to date. The experiments in two independent mouse models of chronic inflammatory pain showed that the expression of *Cct5* and *Fam173b* were higher in the lumbar spinal cord of mice with chronic inflammatory pain but not in de dorsal root ganglions (DRG). In the spinal cord, the expression profiles of both genes were upregulated in response to two different inducers of inflammatory pain. These findings indicate that both genes in the 5p15.2 region are co-regulated in the spinal cord during inflammation-induced pain in both independent pain models, thereby possibly contributing to the neurobiology of pain. In the lumbar DRG, containing the cell bodies of the primary sensory neurons that detect pain signals from the hind paws, *Cct5* and *Fam173b* gene expression levels did not change by inflammation. Because of these complementary results from the two independent tissues (spinal cord and DRG), we hypothesize that the 5p15.2 region is likely to play a role in spinal central pain processing and not in regulating primary sensory neuron responses.

In the study of candidate genes previously reported to be associated with a pain phenotype, we showed that none of the 92 studied variants were significantly associated with CWP in our GWAS meta-analysis. This can be explained by the fact that many of the previous reported loci were studied in relative modest sample sizes and in a large variety of pain phenotypes [76]. Power calculations show that we had approximately 80% power to detect an odds ratio (OR) as low as 1.22 for SNPs with an allele frequency of 20% or higher. So, even in this large meta-analysis power was still modest to detect small ORs and we therefore cannot exclude smaller effect sizes of the tested variants, resulting in lack of reproducibility [77]. This lack of reproducibility of SNPs in candidate genes in large GWAS meta-analyses has been shown before for other phenotypes such as BMD [78]. It is interesting to note that among the candidate SNPs, the strongest associated ones were located in the three most studied pain genes *COMT*, *GCH1*, and *OPRM1*. The directions of the effects of these SNPs were the same as reported earlier, which would support true associations.

In conclusion, our study reports a GWAS meta-analysis on CWP. We identified the genetic variant rs13361160 at the 5p15.2 locus, located 81kb upstream of the *CCT5* gene and 57kb downstream of the *FAM173B* gene, to be associated with CWP. We showed an increase in expression levels of *Cct5* and *Fam173b* in the spinal cord of inflammatory pain models of mice, and since these genes both seem to influence the central mechanism of sensitization, they may represent a novel pathway involved in pain sensation.

ACKNOWLEDGMENTS AND AFFILIATIONS

For acknowledgments, please see the supplements.

REFERENCES

1. Croft P, Schollum J, Silman A. Population study of tender point counts and pain as evidence of fibromyalgia. BMJ. 1994 Sep 17;309(6956):696-9.
2. Wolfe F, Smythe HA, Yunus MB, et al. The American College of Rheumatology 1990 Criteria for the Classification of Fibromyalgia. Report of the Multicenter Criteria Committee. Arthritis Rheum. 1990 Feb;33(2):160-72.
3. Meerding WJ, Bonneux L, Polder JJ, et al. Demographic and epidemiological determinants of healthcare costs in Netherlands: cost of illness study. BMJ. 1998 Jul 11;317(7151):111-5.
4. Buskila D, Mader R. Trauma and work-related pain syndromes: Risk factors, clinical picture, insurance and law interventions. Best Pract Res Cl Rh. 2011 Apr;25(2):199-207.
5. Leffler AS, Kosek E, Lerndal T, et al. Somatosensory perception and function of diffuse noxious inhibitory controls (DNIC) in patients suffering from rheumatoid arthritis. Eur J Pain-London. 2002;6(2):161-76.
6. Imamura M, Imamura ST, Kaziyama HHS, et al. Impact of Nervous System Hyperalgesia on Pain, Disability, and Quality of Life in Patients With Knee Osteoarthritis: A Controlled Analysis. Arthrit Rheum-Arthr. 2008 Oct 15;59(10):1424-31.
7. Kato K, Sullivan PF, Evengard B, et al. Importance of genetic influences on chronic widespread pain. Arthritis Rheum. 2006 May;54(5):1682-6.
8. Limer KL, Nicholl BI, Thomson W, et al. Exploring the genetic susceptibility of chronic widespread pain: the tender points in genetic association studies. Rheumatology (Oxford). 2008 May;47(5):572-7.
9. Nicholl BI, Holliday KL, Macfarlane GJ, et al. Association of HTR2A polymorphisms with chronic widespread pain and the extent of musculoskeletal pain: results from two population-based cohorts. Arthritis Rheum. 2011 Mar;63(3):810-8.
10. Frank B, Niesler B, Bondy B, et al. Mutational analysis of serotonin receptor genes: HTR3A and HTR3B in fibromyalgia patients. Clin Rheumatol. 2004 Aug;23(4):338-44.
11. Bondy B, Spaeth M, Offenbaecher M, et al. The T102C polymorphism of the 5-HT2A-receptor gene in fibromyalgia. Neurobiol Dis. 1999 Oct;6(5):433-9.
12. Gursoy S, Erdal E, Herken H, et al. Association of T102C polymorphism of the 5-HT2A receptor gene with psychiatric status in fibromyalgia syndrome. Rheumatol Int. 2001 Oct;21(2):58-61.
13. Tander B, Gunes S, Boke O, et al. Polymorphisms of the serotonin-2A receptor and catechol-O-methyltransferase genes: a study on fibromyalgia susceptibility. Rheumatol Int. 2008 May;28(7):685-91.
14. Su SY, Chen JJ, Lai CC, et al. The association between fibromyalgia and polymorphism of monoamine oxidase A and interleukin-4. Clin Rheumatol. 2007 Jan;26(1):12-6.
15. Offenbaecher M, Bondy B, de Jonge S, et al. Possible association of fibromyalgia with a polymorphism in the serotonin transporter gene regulatory region. Arthritis Rheum. 1999 Nov;42(11):2482-8.
16. Cohen H, Buskila D, Neumann L, et al. Confirmation of an association between fibromyalgia and serotonin transporter promoter region (5- HTTLPR) polymorphism, and relationship to anxiety-related personality traits. Arthritis Rheum. 2002 Mar;46(3):845-7.
17. Gursoy S. Absence of association of the serotonin transporter gene polymorphism with the mentally healthy subset of fibromyalgia patients. Clin Rheumatol. 2002 Jun;21(3):194-7.
18. Buskila D, Cohen H, Neumann L, et al. An association between fibromyalgia and the dopamine D4 receptor exon III repeat polymorphism and relationship to novelty seeking personality traits. Mol Psychiatry. 2004 Aug;9(8):730-1.
19. Potvin S, Larouche A, Normand E, et al. DRD3 Ser9Gly polymorphism is related to thermal pain perception and modulation in chronic widespread pain patients and healthy controls. J Pain. 2009 Sep;10(9):969-75.
20. Holliday KL, Nicholl BI, Macfarlane GJ, et al. Genetic variation in the hypothalamic-pituitary-adrenal stress axis influences susceptibility to musculoskeletal pain: results from the EPIFUND study. Ann Rheum Dis. 2010 Mar;69(3):556-60.
21. Syvanen AC, Tilgmann C, Rinne J, et al. Genetic polymorphism of catechol-O-methyltransferase (COMT): correlation of genotype with individual variation of S-COMT activity and comparison of the allele frequencies in the normal population and parkinsonian patients in Finland. Pharmacogenetics. 1997 Feb;7(1):65-71.

22. Zubieta JK, Heitzeg MM, Smith YR, et al. COMT val158met genotype affects mu-opioid neurotransmitter responses to a pain stressor. Science. 2003 Feb 21;299(5610):1240-3.
23. Hocking LJ, Smith BH, Jones GT, et al. Genetic variation in the beta2-adrenergic receptor but not catecholamine-O-methyltransferase predisposes to chronic pain: results from the 1958 British Birth Cohort Study. Pain. 2010 Apr;149(1):143-51.
24. Gursoy S, Erdal E, Herken H, et al. Significance of catechol-O-methyltransferase gene polymorphism in fibromyalgia syndrome. Rheumatol Int. 2003 May;23(3):104-7.
25. Cohen H, Neumann L, Glazer Y, et al. The relationship between a common catechol-O-methyltransferase (COMT) polymorphism val(158) met and fibromyalgia. Clin Exp Rheumatol. 2009 Sep-Oct;27(5 Suppl 56):S51-6.
26. Vargas-Alarcon G, Fragoso JM, Cruz-Robles D, et al. Catechol-O-methyltransferase gene haplotypes in Mexican and Spanish patients with fibromyalgia. Arthritis Res Ther. 2007;9(5):R110.
27. van Meurs JB, Uitterlinden AG, Stolk L, et al. A functional polymorphism in the catechol-O-methyltransferase gene is associated with osteoarthritis-related pain. Arthritis Rheum. 2009 Feb;60(2):628-9.
28. Hagen K, Pettersen E, Stovner LJ, et al. No association between chronic musculoskeletal complaints and Val158Met polymorphism in the Catechol-O-methyltransferase gene. The HUNT study. BMC Musculoskelet Disord. 2006;7:40.
29. Nicholl BI, Holliday KL, Macfarlane GJ, et al. No evidence for a role of the catechol-O-methyltransferase pain sensitivity haplotypes in chronic widespread pain. Ann Rheum Dis. 2010 Nov;69(11):2009-12.
30. Racine M, Tousignant-Laflamme Y, Kloda LA, et al. A systematic literature review of 10years of research on sex/gender and experimental pain perception - Part 1: Are there really differences between women and men? Pain. 2011 Dec 20.
31. Skrbo A, Begovic B, Skrbo S. [Classification of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes] Klasificiranje lijekova po ATC sistemu i najnovije izmjene. Med Arh. 2004;58(1 Suppl 2):138-41.
32. Aulchenko YS, Heutink P, Mackay I, et al. Linkage disequilibrium in young genetically isolated Dutch population. Eur J Hum Genet. 2004 Jul;12(7):527-34.
33. Hofman A, van Duijn CM, Franco OH, et al. The Rotterdam Study: 2012 objectives and design update. Eur J Epidemiol. 2011 Aug;26(8):657-86.
34. Spector TD, MacGregor AJ. The St. Thomas' UK Adult Twin Registry. Twin Res. 2002 Oct;5(5):440-3.
35. Spector TD, Williams FM. The UK Adult Twin Registry (TwinsUK). Twin Res Hum Genet. 2006 Dec;9(6):899-906.
36. Power C, Atherton K, Manor O. Co-occurrence of risk factors for cardiovascular disease by social class: 1958 British birth cohort. J Epidemiol Community Health. 2008 Dec;62(12):1030-5.
37. Consortium UKPsD, Wellcome Trust Case Control C, Spencer CC, et al. Dissection of the genetics of Parkinson's disease identifies an additional association 5' of SNCA and multiple associated haplotypes at 17q21. Hum Mol Genet. 2011 Jan 15;20(2):345-53.
38. Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007 Jun 7;447(7145):661-78.
39. Hart DJ, Spector TD. Cigarette smoking and risk of osteoarthritis in women in the general population: The Chingford study. Ann Rheum Dis. 1993 Feb;52(2):93-6.
40. Hart DJ, Spector TD. The relationship of obesity, fat distribution and osteoarthritis in women in the general population: the Chingford Study. J Rheumatol. 1993 Feb;20(2):331-5.
41. Rabbitt P, McInnes L, Diggle P, et al. The University of Manchester longitudinal study of cognition in normal healthy old age, 1983 through 2003. Aging, Neuropsychology and Cognition. 2004;11:P245-79.
42. Syddall HE, Aihie Sayer A, Dennison EM, et al. Cohort profile: the Hertfordshire cohort study. Int J Epidemiol. 2005 Dec;34(6):1234-42.
43. Harris TB, Launer LJ, Eiriksdottir G, et al. Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics. Am J Epidemiol. 2007 May 1;165(9):1076-87.
44. Felson DT, Zhang Y, Hannan MT, et al. The incidence and natural history of knee osteoarthritis in the elderly. The Framingham Osteoarthritis Study. Arthritis Rheum. 1995 Oct;38(10):1500-5.

45. Riyazi N, Meulenbelt I, Kroon HM, et al. Evidence for familial aggregation of hand, hip, and spine but not knee osteoarthritis in siblings with multiple joint involvement: the GARP study. Ann Rheum Dis. 2005 Mar;64(3):438-43.

46. John U, Greiner B, Hensel E, et al. Study of Health In Pomerania (SHIP): a health examination survey in an east German region: objectives and design. Soz Praventivmed. 2001;46(3):186-94.

47. Volzke H, Alte D, Schmidt CO, et al. Cohort profile: the study of health in Pomerania. Int J Epidemiol. 2011 Apr;40(2):294-307.

48. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559-75.

49. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010 Sep 1;26(17):2190-1.

50. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999 Dec;55(4):997-1004.

51. Panagiotou OA, Ioannidis JP, for the Genome-Wide Significance P. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. Int J Epidemiol. 2011 Dec 5.

52. Johnson AD, Handsaker RE, Pulit SL, et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008 Dec 15;24(24):2938-9.

53. Genomes Project C. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct 28;467(7319):1061-73.

54. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-81.

55. Willemen HL, Eijkelkamp N, Wang H, et al. Microglial/macrophage GRK2 determines duration of peripheral IL-1beta-induced hyperalgesia: contribution of spinal cord CX3CR1, p38 and IL-1 signaling. Pain. 2010 Sep;150(3):550-60.

56. Zhao J, Yuan G, Cendan CM, et al. Nociceptor-expressed ephrin-B2 regulates inflammatory and neuropathic pain. Mol Pain. 2010;6:77.

57. Hargreaves K, Dubner R, Brown F, et al. A new and sensitive method for measuring thermal nociception in cutaneous hyperalgesia. Pain. 1988 Jan;32(1):77-88.

58. Yu W, Clyne M, Khoury MJ, et al. Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. Bioinformatics. 2010 Jan 1;26(1):145-6.

59. Schadt EE, Molony C, Chudin E, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 2008 May 6;6(5):e107.

60. Tegeder I, Costigan M, Griffin RS, et al. GTP cyclohydrolase and tetrahydrobiopterin regulate pain sensitivity and persistence. Nat Med. 2006 Nov;12(11):1269-77.

61. Campbell CM, Edwards RR, Carmona C, et al. Polymorphisms in the GTP cyclohydrolase gene (GCH1) are associated with ratings of capsaicin pain. Pain. 2009 Jan;141(1-2):114-8.

62. Holliday KL, Nicholl BI, Macfarlane GJ, et al. Do genetic predictors of pain sensitivity associate with persistent widespread pain? Mol Pain. 2009;5:56.

63. Hayashida M, Nagashima M, Satoh Y, et al. Analgesic requirements after major abdominal surgery are associated with OPRM1 gene polymorphism genotype and haplotype. Pharmacogenomics. 2008 Nov;9(11):1605-16.

64. Kim H, Lee H, Rowan J, et al. Genetic polymorphisms in monoamine neurotransmitter systems show only weak association with acute post-surgical pain in humans. Mol Pain. 2006;2:24.

65. Bouhouche A, Benomar A, Bouslam N, et al. Mutation in the epsilon subunit of the cytosolic chaperonin-containing t-complex peptide-1 (Cct5) gene causes autosomal recessive mutilating sensory neuropathy with spastic paraplegia. J Med Genet. 2006 May;43(5):441-3.

66. Kubota H, Hynes G, Willison K. The chaperonin containing t-complex polypeptide 1 (TCP-1). Multisubunit machinery assisting in protein folding and assembly in the eukaryotic cytosol. Eur J Biochem. 1995 May 15;230(1):3-16.

67. Gingras AC, Caballero M, Zarske M, et al. A novel, evolutionarily conserved protein phosphatase complex involved in cisplatin sensitivity. Mol Cell Proteomics. 2005 Nov;4(11):1725-40.

68. Chen GI, Tisayakorn S, Jorgensen C, et al. PP4R4/KIAA1622 forms a novel stable cytosolic complex with phosphoprotein phosphatase 4. J Biol Chem. 2008 Oct 24;283(43):29273-84.

69. Glatter T, Wepf A, Aebersold R, et al. An integrated workflow for charting the human interaction proteome: insights into the PP2A system. Mol Syst Biol. 2009;5:237.

70. Zhang X, Ozawa Y, Lee H, et al. Histone deacetylase 3 (HDAC3) activity is regulated by interaction with protein serine/threonine phosphatase 4. Genes Dev. 2005 Apr 1;19(7):827-39.

71. Latremoliere A, Woolf CJ. Central sensitization: a generator of pain hypersensitivity by central neural plasticity. J Pain. 2009 Sep;10(9):895-926.

72. Silman AJ, Pearson JE. Epidemiology and genetics of rheumatoid arthritis. Arthritis Res. 2002;4 Suppl 3:S265-72.

73. Bax M, van Heemst J, Huizinga TW, et al. Genetics of rheumatoid arthritis: what have we learned? Immunogenetics. 2011 Aug;63(8):459-66.

74. Rolke R, Baron R, Maier C, et al. Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): standardized protocol and reference values. Pain. 2006 Aug;123(3):231-43.

75. deCharms RC, Maeda F, Glover GH, et al. Control over brain activation and pain learned by using real-time functional MRI. Proc Natl Acad Sci U S A. 2005 Dec 20;102(51):18626-31.

76. Moller AT, Jensen TS. Pain and genes: Genetic contribution to pain variability, chronic pain and analgesic responses. Eur J Pain Supplements. 2010;4:197-201.

77. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005 Aug;2(8):e124.

78. Richards JB, Kavvoura FK, Rivadeneira F, et al. Collaborative meta-analysis: associations of 150 candidate genes with osteoporosis and osteoporotic fracture. Ann Intern Med. 2009 Oct 20;151(8):528-37.

CHAPTER 4.2

Genetics of the heat pain threshold in the general population

Marjolein de Kruijf, Marjolein J. Peters, Cindy G. Boer, Carolina M. Gomez, Fernando Rivadeneira, Frank J.P.M. Huygen, André G. Uitterlinden, and Joyce B. J. van Meurs

Manuscript in preparation

ABSTRACT

Introduction: Chronic pain and pain sensitivity are complex traits with a variety of potential determinants. Although not yet fully elucidated, pain sensitivity and the risk for chronic pain are thought to be partly genetic. In our study, we attempt to further elucidate the genetic predisposition of pain sensitivity.

Methods: In a total number of 3,795 participants from the Rotterdam study (a large prospective population based cohort) heat pain thresholds (HPT) were determined. We estimated the total additive genetic influence on HPT measurements due to common genetic variation using GCTA, and we performed a genome wide association study (GWAS) to identify new loci associated with HPT in the general population. Finally, we reviewed the literature for previously reported DNA variants associated with experimental pain thresholds and tried to replicate these findings in our dataset.

Results: The overall heritability estimate of HPT was 19%. In individuals without chronic pain, this estimate was 32% compared to 9% in individuals with chronic pain. In addition, the heritability was higher in women compared to men. Our GWAS revealed one genome-wide significant signal (1:176688345:D) which is located in the twelfth intron of the *PAPPA2* gene (p= 2.48E-08). Additionally, we found six suggestive signals (P<1.0E-06). Genetic variants previously associated with pain sensitivity were not replicated in our study.

Conclusion: A significant proportion of the variability of HPT is explained by genetics. The extent to which HPT is genetically determined is higher when individuals do not experience chronic pain. Future genetic studies on pain sensitivity should take the presence of chronic pain into account since it influences the phenotype substantially. This largest genetic screen for pain sensitivity up to date provides new potential genetic loci for further research.

INTRODUCTION

Chronic pain and pain sensitivity are complex traits with a variety of potential determinants. The development of chronic pain and an increased sensitivity by sensitization of the nervous system are unintended consequences after tissue damage. In this scenario, the pain is prolonged or more severe compared to what might be expected during a normal healing process [1].

A wide variety of risk factors have been described for the development of chronic pain. One of them is an intrinsic high pain sensitivity, which can be assessed by experimental pain sensitivity measurements [2]. In theory, experimental pain sensitivity is less sensitive to bias due to disease or tissue damage, compared to more subjective pain phenotypes such as pain severity scores [3]. There are many different measurements to determine pain sensitivity, such as pain thresholds and tolerance for different stimuli. The heat pain threshold (HPT) is one of the most studied measurements for pain sensitivity: the HPT is noninvasive and can be used for measuring pain sensitivity and pain thresholds. Measurements can be done over multiple body points, and the temperature and the duration of the pain stimulus can be highly controlled [3]. Finally, there is good reproducibility between two sessions [4].

The proportion of genetic influence on pain has been under debate. In previous studies, the heritability of pain sensitivity and chronic pain has been estimated in classical twin studies. A review by Nielsen *et al.* [5] showed that the heritability estimates of specific pain phenotypes differ: for example, back and neck pain have heritability estimates ranging from 0% to 68% [6-13], osteoarthritis has heritability estimates ranging from 0% to 53% [14,15], and irritable bowel syndrome has heritability estimates ranging from 0% to 48% [16-20]. The heritability of experimental pain sensitivity has been studied scarcely, with only three previous study reports. All three reports had a twins design and used various experimental designs, such as cold pressor tests and heat pain thresholds [21-23], and relatively small sample sizes were used. Consequently, heritability estimates ranged between 0% and 60%. Other studies investigating the genetic background of pain sensitivity focused on candidate genes previously described to play a role in pain, or studied the genetic variants in modestly sized pain patient populations [24-36].

In a previous study by our group, we investigated the genetic background of chronic widespread pain [37]. Although we identified a DNA variant to be associated with CWP, we also identified significant heterogeneity among phenotype definitions among the cohorts. As for other complex traits, it would be helpful to dissect the pain phenotype into quantitative underlying endophenotypes, such as intrinsic pain sensitivity, which can be measured by experimental pain thresholds. The present study therefore focuses on experimental HPT as an endophenotype underlying the development of chronic pain.

The aim of the current study was to further elucidate the genetic predisposition of pain sensitivity, defined as the HPT. In the Rotterdam Study, a large prospective population based cohort, we

estimated the heritability of the HPT and the influence of gender and the presence of chronic pain on the heritability of the trait. We performed a genome wide association study (GWAS) to search for potential new genetic markers associated with HPT in a general population. And finally, we reviewed the literature for significantly pain sensitivity associated variants and we tried to replicate those findings in our population.

METHODS

Study population

This study is performed within the Rotterdam Study (RS), a large prospective population-based cohort study of men and women aged 45 years and over. The study design and rationale are described elsewhere in detail [38]. In summary, determinants, incidence and progression of chronic disabling diseases in the elderly are studied. The first cohort (RS-I) within the Rotterdam study started in 1990 and included 7,983 individuals ages 55 years and older. In 1999, an additional 3,011 subjects were included in Rotterdam study II (RS-II). The third cohort (RS-III) was invited in 2005, adding 3,932 individuals aged 45 years and over. All participants were examined in detail at baseline and at subsequent follow up visits, which took place approximately every six years. In summary, a home interview and extensive set of examinations at the research center was performed. For the present study, we used data from 3,795 participants for whom data on experimental pain sensitivity, data on the presence of chronic pain and genetic information were available. The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC and by the Ministry of Health, Welfare and Sport of the Netherlands, implementing the *"Wet Bevolkingsonderzoek: ERGO (Population Studies Act: Rotterdam Study)"*. All participants provided written informed consent to participate in the study and to obtain information from their treating physicians.

Genotyping

Genotyping was done using Illumina Infinium HumanHap550 Beadchips (RS-II), or the Illumina Infinium HumanHap610 Beadchips (RS-III). Details about genotyping and Quality Control have been described previously [37]. In short, a total of 2,612 subjects were genotyped in RS-II (Illumina 550 duo) and a total of 3,523 subjects in RS-III (Illumina 610 quad). Exclusion criteria were a call rate <98%, Hardy-Weinberg P-value $<10^{-6}$ and minor allele frequency <0.01%, autosomal heterozygosity, sex mismatch and outlying identity-by-state clustering estimates. A total of 2,157 for RS-II and 3,048 for RS-III passed genotyping quality control. Data was imputed with the 1000-Genomes reference panel (phase 1, version 3) using MACH version 1.0.15/1.0.16 [39]. A total number of 30,072,738 SNPs were available for association analysis.

Experimental pain sensitivity assessment: Heat pain threshold measurement

In the 3,795 participants of the Rotterdam study included in this study, quantitative sensory testing was conducted. We used a commercially available thermo-sensory analyzer, the TSA II (Medoc

Advanced Medical Systems, Durham, NC). The measurement probe had a surface of 30x30mm, and was placed on the inner site of the non-dominant forearm.

During the HPT measurement, the starting temperature of the probe was 32 degrees Celsius. Then, the probe would increase in temperature with 1.5 degrees per second until the participant ended the test or the maximum temperature of 50 degrees Celsius was reached. The participant was asked to push a large red 'quiz button' and therewith end the measurement at the moment the stimulus started to feel unpleasant or painful. After each measurement, the temperature returned to 32 degrees Celsius before the next measurement started. The HPT measurement was repeated five times in a row. For the analysis, the average temperature of the last three measurements was used.

Heritability estimation

To quantify the proportion of HPT variance explained by genetic variants, we used the restricted maximum likelihood (REML) method. This method is able to quantify heritability estimates attributable to all genetic variants and is implemented in the Genome-wide Complex Trait Analysis (GCTA) package [40]. We created one genetic relationship matrix (GRM) file for the unrelated participants in our RS-II and RS-III populations, and included all genotyped SNPs. This resulted in a GRM file of 495,775 SNPs for 3,795 samples. No pairs of individuals exceeded the GCTA standard cutoff coefficient of 0.025 for genetic relatedness. A P-value<0.05 was considered to be statistically significant in this analysis.

The mean HPT (as described before) was used as phenotype and adjustments were made for age and gender.

Additional analyses were performed in which we stratified both for gender and the presence of chronic musculoskeletal pain.

Genome Wide Association Study (GWAS) and meta-analysis

We performed two GWAS for HPT: one in RS-II and one in RS-III. We used MACH2QTL via GRIMP [41], which uses the genotype dosage values (0-2 as a continuous variables) as the predictor in a linear regression framework. HPT was used as the outcome measurement and adjustments were made for age, gender and the presence of chronic pain. In addition, the GWAS was repeated in participants without chronic pain.

Quality control was done with EasyQC [42]. The effective allele count was calculated for all SNPs by 2*minor allele frequency*R^2(correlatedness of the data)*sample size. SNPs with an effective allele count >5 were included in the meta-analysis. An effective allele count of >5 represents minor alleles appearing at least 5 times in the study population.

The summary statistics of the results of the GWAS in RS-II and RS-III were meta-analyzed using METAL (*www.sph.umich.edu/csg/abecasis/metal*) after genomic control correction to the standard

errors and p-values. METAL applies an inverse-variance methodology assuming fixed effects with Cochran's Q and I^2 metrics to quantify between-study heterogeneity.

For the GWAS, the statistical significant threshold was set on 5.0E-08. SNPs with a P-value<1.0E-06 were called suggestive signals.

Systematic review of genetic variants previously described

We systematically searched the literature for previous associations with experimental pain thresholds. We used the Human Genome Epidemiology (HuGe) Navigator Phenopedia database for this [43]. This database provides a comprehensive archive of studies assessing the associations between phenotypes and genetic variants and this database is continuously updated.

The phenopedia tool provides a list of genes previously associated with your phenotype of interest, and includes links to the articles in which these associations were published. We used the search term 'pain threshold' on 16 September 2014. All publications were manually screened for the phenotype studied and the SNPs identified. We only included the studies which investigated the association of genetic variants with measures of quantitative sensory testing. The SNPs selected for the analysis were those described to be significantly associated with the pain threshold phenotype. Additionally, an rs-id needed to be available.

For all reported SNPs, we examined their association with HPT in our GWAS meta-analysis results. The significance threshold was set at P-value<0.05.

RESULTS

Population characteristics

For this study, 1,326 individuals from the third follow up visit of RS-II and 2,469 individuals from the second follow up visit of RS-III were included in whom HPT measurements and genotype information was available. Characteristics are shown in Table 1. The participants from RS-II were significantly older and had a lower percentage of women. The prevalence of chronic pain was significantly higher in RS-II and mean HPT measures were slightly lower in RS-II compared to those in RS-III.

Table 1. Study population characteristics.

	Total	RS-II	RS-III
N=	3,795	1,326	2,469
Age, mean (SD)	65.9 (7.5)	72.6 (5.2)	62.3 (6.0)
Women, % (n)	56% (2,125)	54% (716)	57% (1,407)
Chronic pain present, % (n)	44% (1,670)	47% (623)	42% (1,037)
Heat pain threshold, mean (SD) in degrees Celsius	47.5 (3.0)	47.3 (3.2)	47.7 (2.8)

RS-II = Rotterdam Study II; RS-III = Rotterdam Study III; SD = standard deviation; n = sample size.

Heritability estimation

In the complete population, the GCTA estimate of genetic influence due to the additive effect of common SNPs was 19% (SE 0.09; P-value=0.02). Since gender is one of the major factors determining HPT, we subsequently stratified the population according to gender. We oberved the heritability estimate in women to be 35% (SE 0.20; P-value 0.04), while it was 9% in men (SE 0.28; P-value 0.38). Chronic pain is known to influence HPTs significantly through central sensitization. Therefore, we studied the heritability of HPT separately in individuals with and without chronic pain. For individuals with chronic pain, the heritability was estimated to be 8% (SE 0.20; P-value=0.35). For individuals without chronic pain, this estimate was higher and statistically significant with 32% (SE 0.17; P-value=0.03).

GWAS meta-analysis

A total of 30,072,738 markers were tested for the association with HPT in our population of in total 3,795 individuals. Genomic control inflation factors for the P-values in RS-II and RS-III were low (λ=1.01 and 1.001 respectively). The Quantile-Quantile plot indicated no substantial population stratification due to cryptic relatedness, population substructure or other biases (Figure 1). The results of the GWAS meta-analysis are summarized in a Manhattan Plot of the P-values (Figure 2).

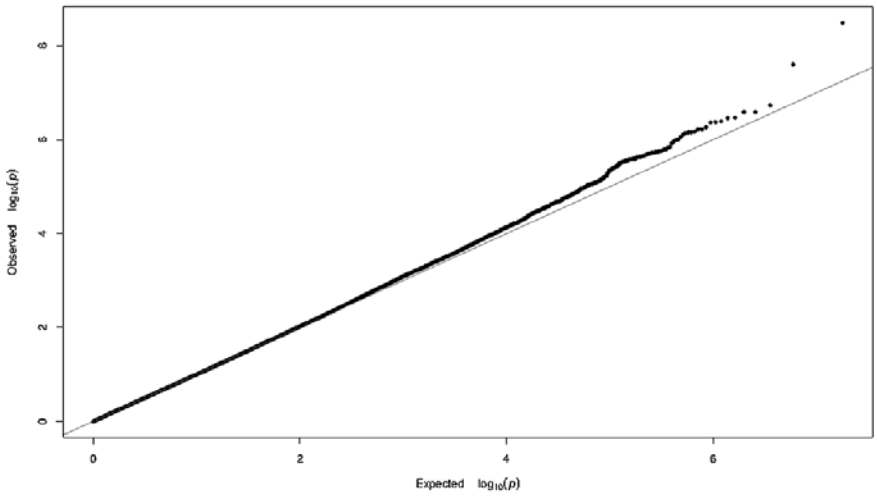


Figure 1. Quantile-quantile plot (Q-Q plot) for the GWAS meta-analysis with HPT. This plot compares additive model statistics to those expected under the null distribution using fixed-effects for all analyzed 1000G imputed SNPs passing the quality control. Analysis adjusted for the presence of chronic pain.

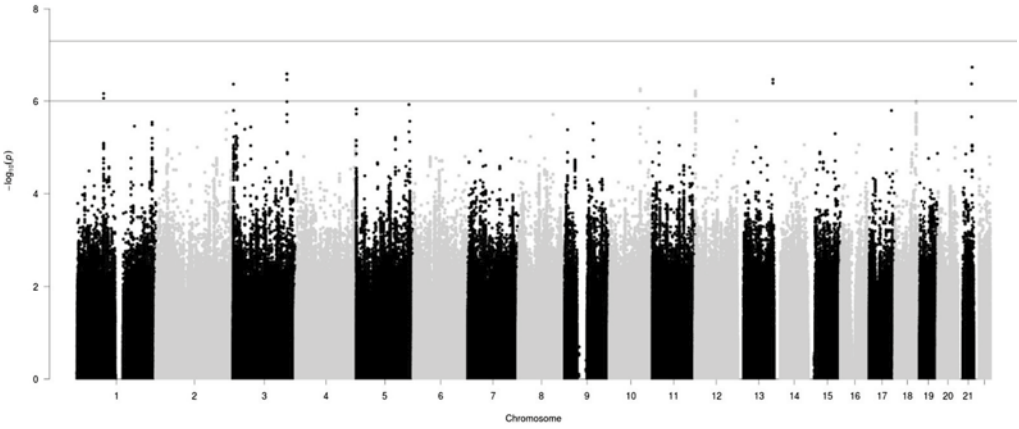


Figure 2. Manhattan plot of the P-values of the GWAS meta-analysis of HPT in RS-II and RS-III. Analysis adjusted for the presence of chronic pain. The red line represents the line for genome wide significance (P-value=5.0E-08), the blue line represents the line for suggestive signals (P-value=1.0E-06).

We identified one SNP on chromosome 1 to be genome-wide significant (P-value <5.0E-08). This SNP represents a deletion located on position 176,688,345 on chromosome 1 (P-value=2.48E-08). It is a relatively rare deletion (minor allele frequency=0.02), and it is located in the twelfth intron of the *PAPPA2* gene. The *PAPPA2* gene encodes for a protein which is thought to be a local regulator of insulin-like growth factor (*IGF*) bioavailability. *IGF* is implicated in nociceptive (pain) sensitivity of primary afferent neurons [44]. Additionally, the deletion is located 140kb downstream of the *ASTN1* gene. *ASTN1* (or astrotactin 1) is a neuronal adhesion molecule required for migration of young postmitotic neuroblasts in cortical regions of developing brain, including cerebrum, hippocampus, cerebellum, and olfactory bulb [45].

Next to the genome-wide significant hit, we found six other suggestive signals with a P-value<1.0E-06 (Table 2). Five of six top SNPs have relative low allele frequencies (MAF<0.05).

We found one locus to be suggestive in both the analyses: in the original GWAS (adjusting for chronic pain), the intronic SNP rs187924640 (MAF=0.015) was a suggestive hit (p=2.56E-07), and in the sensitivity analysis (excluding chronic pain cases) this SNP was also close to significance (P-value=8.23E-07). The *PRKC1* gene encodes for the protein kinase C iota type, which is implicated in the regulation of neuronal growth and specification [46-48].

Table 2. Genome Wide Association Analysis 'Heat Pain Threshold' adjusted for presence of chronic pain, tophits signals P-value<1.0E-06

Marker	Chr	Pos	Coded allele	Other allele	AF	Beta	StdErr	P-value (all participants)	P-value (no pain cases)	Position	Gene
1:176688345:D	1	176688345	D	I	0.02	-2.09	0.38	2.48E-08	1.11E-05	intron 12	*PAPPA2*
rs13049646	21	43651846	T	G	0.98	1.52	0.29	1.84E-07	1.24E-05	intron 2	*ABCG1*
rs187924640	3	169948817	T	G	0.02	-1.57	0.31	2.56E-07	8.23E-07	intron 1	*PRKCI*
rs512766	10	97228338	C	G	0.10	0.60	0.12	5.41E-07	2.96E-04	intron 1	*SORBS1*
rs74371079	12	1084044	A	G	0.06	-0.81	0.16	5.98E-07	2.00E-04	intron 1	*RAD52*
rs141493091	1	84375770	A	C	0.04	-1.09	0.22	6.81E-07	7.60E-04	intron 15	*TTLL7*
rs7239184	18	67129023	T	G	0.96	0.91	0.19	9.94E-07	3.52E-03	intron 1	*DOK6*

Chr = chromosome; Pos = position; Coded allele = effect allele; AF = allele frequency of coded allele; Beta = effect size of effect allele; StdErr = standard error of the effect.

Systematic review of genetic variants previously described

In the HuGe navigator, the search term 'Pain threshold' provided a total of 60 publications describing 44 different genes. After selection for pain threshold phenotypes and SNPs having an rs-id, we were left with fifteen publications. In these articles, nine SNPs in six different genes (*COMT*, *DRD3*, *OPRK*, *OPRM1*, *SLA6A4* and *HTR1A*) were previously reported to be significantly associated with pain threshold phenotypes. The selected SNPs, the direction of the effect in the previous articles and the results in our GWAS study are shown in Table 3. None of the nine SNPs were significantly associated with HPT in our GWAS meta-analysis results.

Table 3. Associations of HPT with the candidate gene SNPs.

	Coded allele	Other allele	AF	Beta	P-value (all participants)	Effect direction in literature*	References
COMT							
rs4680	A	G	0.55	-0.09	0.18	-, -, -, -, -	[25-27,31,34]
DRD3							
rs6280	T	C	0.69	0.02	0.77	-	[35]
OPRK							
rs6473799	A	G	0.77	0.09	0.27	-	[36]
rs7016778	A	T	0.88	0.15	0.17	+	[36]
rs7824175	C	G	0.92	-0.01	0.94	-	[36]
rs9479757	A	G	0.10	-0.05	0.69	+	[30]
OPRM1							
rs1799971	A	G	0.89	0.02	0.82	+,+	[24,28]
SLC6A4							
rs25531	T	C	0.93	0.025	0.89	+, +	[29,33]
HTR1A							
rs6295	C	G	0.50	0.04	0.63	+	[32]

** A negative direction (-) means a higher sensitivity for QST coinciding with a lower HPT; A positive direction (+) means a lower sensitivity for QST coinciding with a higher HPT. Coded allele = effect allele; AF = allele frequency of coded allele; Beta = effect size of effect allele.*

DISCUSSION

In this population based study, we aimed to identify the genetic background of an experimental measure of pain sensitivity, the heat pain threshold (HPT). We observed an overall heritability estimate of 19% which was dependent on gender and the presence of chronic pain. We performed a genome wide association study (GWAS) to search for potential new loci and found seven interesting new loci. In a candidate SNP approach, we were not able to replicate the earlier associated SNPs with the HPT in our study.

Although not yet fully elucidated, a significant proportion of the variability of HPT is explained by genetics. The method we used to measure heritability is different from twin and pedigree analysis. Our method uses only common DNA variants in linkage with the genotyped SNPs (on the Illumina SNP arrays) to estimate heritability, while family-based studies use all genetic variants, including rare variants [40,49]. Since there are less SNPs included in our analysis, the heritability of the trait will be underestimated. The GCTA method has been applied to other complex traits like height, and in this study a heritability of 55% for height was observed [50]. This is much lower than the heritability estimates based on twin studies, in which 89-93% of the height variance can be explained by genetics [51]. Therefore, we expect the heritability estimate of the HPT to be higher than the 19% we identified.

The advantage of GCTA is that this method is able to estimate the heritability in a large sample of unrelated individuals, which makes it more generalizable to a general population [52].

Interestingly, we found an evident difference in the heritability estimate of HPT between genders and between individuals with and without chronic pain. In women and in individuals without chronic pain, the phenotypic variance is explained genetically for one third. In men and in individuals with chronic pain, the heritability estimate was not significant. In our study sample, almost 20% of all men reached the maximum threshold of 50 degrees Celsius for the HPT. As a consequence, part of the variability of the HPT-measurement is lost, which results in lower power to measure heritability in this part of the population. Another explanation could be that the HPT in men is influenced by other, not yet identified, factors. Our results also showed that heritability of HPT is much higher in individuals without chronic pain compared to those that have pain. It is known that experimental pain sensitivity (like HPT) is influenced by the presence of chronic pain, caused by central sensitization of the nervous system [1]. We hypothesize that the presence of chronic pain overrides the subtle genetic effects observed in the general population. This may be one of the reasons why former studies were not able to find consistently influencing genes for pain sensitivity phenotypes. Therefore, the presence of chronic pain should be taken into account when performing genetic analysis on HPT and potentially for other pain sensitivity thresholds in future studies.

To the best of our knowledge, we here present results from the largest genetic study on experimental pain performed up to date. In the GWAS for the HPT adjusted for chronic pain, there was one deletion (1:176688345:D) on chromosome 1 which reached genome wide significance (p=2.48E-08). This deletion is located in the twelfth intron of the *PAPPA2* gene, of which the encoded protein is thought to be a local regulator of insulin-like growth factor (*IGF*) bioavailability. IGF is implicated to play a role in the nociceptive (pain) sensitivity of primary afferent neurons. Neurotrophy, neurogenesis and metabolic functions are shown to be influenced by *IGF* in the adult brain [53]. *In vitro*, upregulation of *IGF* showed a higher sensitivity of primary afferent neurons [54,55]. Additionally, the deletion is located 140kb downstream of the *ASTN1* gene. *ASTN1* (or astrotactin 1) is a neuronal adhesion molecule required for migration of young postmitotic neuroblasts in cortical regions of developing brain, including cerebrum, hippocampus, cerebellum, and olfactory bulb [45]. Trafficking of the

ASTN1 protein is regulated by the *ASTN2* gene [56]. Interestingly, an SNP within *ASTN2* (rs4836732) was found to be associated with the pain-related phenotype total hip replacement in women (p=6.11E-11) [57].

One suggestive hit, located within the *PRKCI* gene, was associated with HPT in both the overall analysis (including all participants) and the sensitivity analysis (without chronic pain cases). The *PRKCI* gene encodes the protein kinase C iota gene, which has been found to regulate neuronal growth in the hippocampus in embryonic rats, specification of neurons during development in cerebellar purkinje cells in zebrafish and inhibiton of spinal cord precursors, also in zebrafish [46-48]. These functional associations indicate that the *PRKCI* gene might in fact be influencing neuronal functioning.

Although very interesting, our GWAS findings need to be replicated in an independent cohort, before definite conclusions can be drawn. Since the power to detect SNPs associated with our phenotype was relatively low (n=3,795), there might be some false positive hits. Additionally, there might be interesting signals among the suggestive SNPs and replication should demonstrate the true signals. After replication of our findings, functional testing of candidate genes would help to give more insight into the biology of HPT.

In the study of candidate genes previously reported to be associated with pain sensitivity measurements, we showed that none of the SNPs was significantly associated with HPT in our GWAS meta-analysis, although our sample size was at least 10 times larger. This can be explained by the fact that many of the previous reported loci were investigated in small populations of pain patients. This could indicate that the associations found are more associated to the pain syndrome than the pain sensitivity itself. The lack of reproducibility of SNPs in candidate genes in large GWAS meta-analyses has been shown before for other phenotypes such as BMD [58].

In conclusion, our study reports a heritability estimate for HPT of 19%. We identified significant influences of gender and chronic pain on the heritability estimates of HPT. Therefore, future genetic studies on pain sensitivity should be adjusted for gender and the presence of chronic pain, or individuals with chronic pain should be excluded from the analysis. This will result in a more homogenous pain phenotype and this will increase the chances of finding new genetic loci involved. The exact genes influencing HPT remain not fully elucidated, but this study provides new potential genes for further research.

ACKNOWLEDGMENTS

We thank all study participants and staff from the Rotterdam Study, the participating general practitioners and the pharmacists.

This study was funded The Netherlands Organisation for Scientific Research (NWO) Vidi Grant 917103521. The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University (Rotterdam), Netherlands Organisation for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam.

REFERENCES

1. Woolf CJ (2011) Central sensitization: implications for the diagnosis and treatment of pain. Pain 152: S2-15.
2. Edwards RR (2005) Individual differences in endogenous pain modulation as a risk factor for chronic pain. Neurology 65: 437-443.
3. Birnie KA, Caes L, Wilson AC, Williams SE, Chambers CT (2014) A practical guide and perspectives on the use of experimental pain modalities with children and adolescents. Pain Manag 4: 97-111.
4. Meier PM, Berde CB, DiCanzio J, Zurakowski D, Sethna NF (2001) Quantitative assessment of cutaneous thermal and vibration sensation and thermal pain detection thresholds in healthy children and adolescents. Muscle Nerve 24: 1339-1345.
5. Nielsen CS, Knudsen GP, Steingrimsdottir OA (2012) Twin studies of pain. Clin Genet 82: 331-340.
6. 6Battie MC, Videman T, Levalahti E, Gill K, Kaprio J (2007) Heritability of low back pain and the role of disc degeneration. Pain 131: 272-280.
7. El-Metwally A, Mikkelsson M, Stahl M, Macfarlane GJ, Jones GT, et al. (2008) Genetic and environmental influences on non-specific low back pain in children: a twin study. Eur Spine J 17: 502-508.
8. Hartvigsen J, Nielsen J, Kyvik KO, Fejer R, Vach W, et al. (2009) Heritability of spinal pain and consequences of spinal pain: a comprehensive genetic epidemiologic analysis using a population-based sample of 15,328 twins ages 20-71 years. Arthritis Rheum 61: 1343-1351.
9. Hestbaek L, Iachine IA, Leboeuf-Yde C, Kyvik KO, Manniche C (2004) Heredity of low back pain in a young population: a classical twin study. Twin Res 7: 16-26.
10. Livshits G, Popham M, Malkin I, Sambrook PN, Macgregor AJ, et al. (2011) Lumbar disc degeneration and genetic factors are the main risk factors for low back pain in women: the UK Twin Spine Study. Ann Rheum Dis 70: 1740-1745.
11. MacGregor AJ, Andrew T, Sambrook PN, Spector TD (2004) Structural, psychological, and genetic influences on low back and neck pain: a study of adult female twins. Arthritis Rheum 51: 160-167.
12. Nyman T, Mulder M, Iliadou A, Svartengren M, Wiktorin C (2011) High heritability for concurrent low back and neck-shoulder pain: a study of twins. Spine (Phila Pa 1976) 36: E1469-1476.
13. Reichborn-Kjennerud T, Stoltenberg C, Tambs K, Roysamb E, Kringlen E, et al. (2002) Back-neck pain and symptoms of anxiety and depression: a population-based twin study. Psychol Med 32: 1009-1020.
14. Kirk KM, Bellamy N, O'Gorman LE, Kuhnert PM, Klestov A, et al. (2002) The validity and heritability of self-report osteoarthritis in an Australian older twin sample. Twin Res 5: 98-106.
15. Kujala UM, Leppavuori J, Kaprio J, Kinnunen J, Peltonen L, et al. (1999) Joint-specific twin and familial aggregation of recalled physician diagnosed osteoarthritis. Twin Res 2: 196-202.
16. Kato K, Sullivan PF, Evengard B, Pedersen NL (2009) A population-based twin study of functional somatic syndromes. Psychol Med 39: 497-505.
17. Bengtson MB, Ronning T, Vatn MH, Harris JR (2006) Irritable bowel syndrome in twins: genes and environment. Gut 55: 1754-1759.
18. Lembo A, Zaman M, Jones M, Talley NJ (2007) Influence of genetics on irritable bowel syndrome, gastro-oesophageal reflux and dyspepsia: a twin study. Aliment Pharmacol Ther 25: 1343-1350.
19. Mohammed I, Cherkas LF, Riley SA, Spector TD, Trudgill NJ (2005) Genetic influences in irritable bowel syndrome: a twin study. Am J Gastroenterol 100: 1340-1344.
20. Svedberg P, Johansson S, Wallander MA, Pedersen NL (2008) No evidence of sex differences in heritability of irritable bowel syndrome in Swedish twins. Twin Res Hum Genet 11: 197-203.
21. MacGregor AJ, Griffiths GO, Baker J, Spector TD (1997) Determinants of pressure pain threshold in adult twins: evidence that shared environmental influences predominate. Pain 73: 253-257.
22. Nielsen CS, Stubhaug A, Price DD, Vassend O, Czajkowski N, et al. (2008) Individual differences in pain sensitivity: genetic and environmental contributions. Pain 136: 21-29.
23. Norbury TA, MacGregor AJ, Urwin J, Spector TD, McMahon SB (2007) Heritability of responses to painful stimuli in women: a classical twin study. Brain 130: 3041-3049.

24. Bruehl S, Chung OY, Burns JW (2008) The mu opioid receptor A118G gene polymorphism moderates effects of trait anger-out on acute pain sensitivity. Pain 139: 406-415.
25. Diatchenko L, Nackley AG, Slade GD, Bhalang K, Belfer I, et al. (2006) Catechol-O-methyltransferase gene polymorphisms are associated with multiple pain-evoking stimuli. Pain 125: 216-224.
26. Fernandez-de-las-Penas C, Ambite-Quesada S, Rivas-Martinez I, Ortega-Santiago R, de-la-Llave-Rincon AI, et al. (2011) Genetic contribution of catechol-O-methyltransferase polymorphism (Val158Met) in children with chronic tension-type headache. Pediatr Res 70: 395-399.
27. Fernandez-de-las-Penas C, Fernandez-Lao C, Cantarero-Villanueva I, Ambite-Quesada S, Rivas-Martinez I, et al. (2012) Catechol-O-methyltransferase genotype (Val158met) modulates cancer-related fatigue and pain sensitivity in breast cancer survivors. Breast Cancer Res Treat 133: 405-412.
28. Fillingim RB, Kaplan L, Staud R, Ness TJ, Glover TL, et al. (2005) The A118G single nucleotide polymorphism of the mu-opioid receptor gene (OPRM1) is associated with pressure pain sensitivity in humans. J Pain 6: 159-167.
29. Hooten WM, Hartman WR, Black JL, 3rd, Laures HJ, Walker DL (2013) Associations between serotonin transporter gene polymorphisms and heat pain perception in adults with chronic pain. BMC Med Genet 14: 78.
30. Huang CJ, Liu HF, Su NY, Hsu YW, Yang CH, et al. (2008) Association between human opioid receptor genes polymorphisms and pressure pain sensitivity in females*. Anaesthesia 63: 1288-1295.
31. Jensen KB, Lonsdorf TB, Schalling M, Kosek E, Ingvar M (2009) Increased sensitivity to thermal pain following a single opiate dose is influenced by the COMT val(158)met polymorphism. PLoS One 4: e6016.
32. Lindstedt F, Karshikoff B, Schalling M, Olgart Hoglund C, Ingvar M, et al. (2012) Serotonin-1A receptor polymorphism (rs6295) associated with thermal pain perception. PLoS One 7: e43221.
33. Lindstedt F, Lonsdorf TB, Schalling M, Kosek E, Ingvar M (2011) Perception of thermal pain and the thermal grill illusion is associated with polymorphisms in the serotonin transporter gene. PLoS One 6: e17752.
34. Martinez-Jauand M, Sitges C, Rodriguez V, Picornell A, Ramon M, et al. (2013) Pain sensitivity in fibromyalgia is associated with catechol-O-methyltransferase (COMT) gene. Eur J Pain 17: 16-27.
35. Potvin S, Larouche A, Normand E, de Souza JB, Gaumond I, et al. (2009) DRD3 Ser9Gly polymorphism is related to thermal pain perception and modulation in chronic widespread pain patients and healthy controls. J Pain 10: 969-975.
36. Sato H, Droney J, Ross J, Olesen AE, Staahl C, et al. (2013) Gender, variation in opioid receptor genes and sensitivity to experimental pain. Mol Pain 9: 20.
37. Peters MJ, Broer L, Willemen HL, Eiriksdottir G, Hocking LJ, et al. (2013) Genome-wide association study meta-analysis of chronic widespread pain: evidence for involvement of the 5p15.2 region. Ann Rheum Dis 72: 427-436.
38. Hofman A, Brusselle GG, Darwish Murad S, van Duijn CM, Franco OH, et al. (2015) The Rotterdam Study: 2016 objectives and design update. Eur J Epidemiol 30: 661-708.
39. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34: 816-834.
40. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88: 76-82.
41. Estrada K, Abuseiris A, Grosveld FG, Uitterlinden AG, Knoch TA, et al. (2009) GRIMP: a web- and grid-based tool for high-speed analysis of large-scale genome-wide association using imputed data. Bioinformatics 25: 2750-2752.
42. Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, et al. (2014) Quality control and conduct of genome-wide association meta-analyses. Nat Protoc 9: 1192-1212.
43. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ (2008) A navigator for human genome epidemiology. Nat Genet 40: 124-125.
44. Miura M, Sasaki M, Mizukoshi K, Shibasaki M, Izumi Y, et al. (2011) Peripheral sensitization caused by insulin-like growth factor 1 contributes to pain hypersensitivity after tissue injury. Pain 152: 888-895.
45. Fink JM, Hirsch BA, Zheng C, Dietz G, Hatten ME, et al. (1997) Astrotactin (ASTN), a gene for glial-guided neuronal migration, maps to human chromosome 1q25.2. Genomics 40: 202-205.
46. Buchser WJ, Slepak TI, Gutierrez-Arenas O, Bixby JL, Lemmon VP (2010) Kinase/phosphatase overexpression reveals pathways regulating hippocampal neuron morphology. Mol Syst Biol 6: 391.

47. Roberts RK, Appel B (2009) Apical polarity protein PrkCi is necessary for maintenance of spinal cord precursors in zebrafish. Dev Dyn 238: 1638-1648.
48. Tanabe K, Kani S, Shimizu T, Bae YK, Abe T, et al. (2010) Atypical protein kinase C regulates primary dendrite specification of cerebellar Purkinje cells by localizing Golgi apparatus. J Neurosci 30: 16983-16992.
49. Yang J, Lee SH, Goddard ME, Visscher PM (2013) Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. Methods Mol Biol 1019: 215-236.
50. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, et al. (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet 47: 1114-1120.
51. Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, et al. (2003) Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res 6: 399-408.
52. Llewellyn CH, Trzaskowski M, Plomin R, Wardle J (2013) Finding the missing heritability in pediatric obesity: the contribution of genome-wide complex trait analysis. Int J Obes (Lond) 37: 1506-1509.
53. Anderson MF, Aberg MA, Nilsson M, Eriksson PS (2002) Insulin-like growth factor-I and neurogenesis in the adult mammalian brain. Brain Res Dev Brain Res 134: 115-122.
54. Lilja J, Laulund F, Forsby A (2007) Insulin and insulin-like growth factor type-I up-regulate the vanilloid receptor-1 (TRPV1) in stably TRPV1-expressing SH-SY5Y neuroblastoma cells. J Neurosci Res 85: 1413-1419.
55. Van Buren JJ, Bhat S, Rotello R, Pauza ME, Premkumar LS (2005) Sensitization and translocation of TRPV1 by insulin and IGF-I. Mol Pain 1: 17.
56. Wilson PM, Fryer RH, Fang Y, Hatten ME (2010) Astn2, a novel member of the astrotactin gene family, regulates the trafficking of ASTN1 during glial-guided neuronal migration. J Neurosci 30: 8529-8540.
57. arc OC, arc OC, Zeggini E, Panoutsopoulou K, Southam L, et al. (2012) Identification of new susceptibility loci for osteoarthritis (arcOGEN): a genome-wide association study. Lancet 380: 815-823.
58. Richards JB, Kavvoura FK, Rivadeneira F, Styrkarsdottir U, Estrada K, et al. (2009) Collaborative meta-analysis: associations of 150 candidate genes with osteoporosis and osteoporotic fracture. Ann Intern Med 151: 528-537.

CHAPTER 4.3

Associations between joint effusion in the knee and gene expression levels in the circulation: a meta-analysis

Marjolein J. Peters*, Yolande F.M. Ramos*, Wouter den Hollander, Dieuwke Schiphof, Albert Hofman, André G. Uitterlinden, Edwin H. G. Oei, P. Eline Slagboom, Margreet Kloppenburg, Johan L. Bloem, Sita M.A. Bierma-Zeinstra, Ingrid Meulenbelt*, Joyce B.J. van Meurs*

** These authors contributed equally to this work*

Manuscript submitted

ABSTRACT

Objective: To identify molecular biomarkers for early knee osteoarthritis (OA), we examined whether joint effusion in the knee associated with different gene expression levels in the circulation.

Materials and Methods: Joint effusion grades measured with magnetic resonance (MR) imaging and gene expression levels in blood were determined in women of the Rotterdam Study (N=135) and GARP (N=98). Associations were examined using linear regression analyses, adjusted for age, fasting status, RNA quality, technical batch effects, blood cell counts, and BMI. To investigate enriched pathways and protein-protein interactions, we used the DAVID and STRING webtools.

Results: In a meta-analysis, we identified 257 probes mapping to 189 unique genes in blood that were nominally significantly associated with joint effusion grades in the knee. Several compelling genes were identified such as *C1orf38* and *NFATC1*. Significantly enriched biological pathways were: response to stress, gene expression, negative regulation of intracellular signal transduction, and antigen processing and presentation of exogenous pathways.

Conclusion: Meta-analyses and subsequent enriched biological pathways resulted in interesting candidate genes associated with joint effusion that require further characterization. Associations were not transcriptome-wide significant most likely due to limited power. Additional studies are required to replicate our findings in more samples, which will greatly help to understanding the pathophysiology of OA and its relation with inflammation, and may result in biomarkers urgently needed to diagnose OA at an early stage.

INTRODUCTION

Osteoarthritis (OA) is a common, age-related, degenerative disease of the synovial joints. It is characterized by cartilage degradation, osteophyte formation, subchondral bone changes, and synovitis [1]. These characteristics can lead to joint space narrowing, pain, and loss of function, until at the end-stage of the disease total joint replacement is required. OA is a leading cause of morbidity and disability and carries high socioeconomic costs. With increasing obesity and age in the population, a massive rise in morbidity and costs attributed to OA is expected. To be able to change from symptomatic treatment at late disease state and total joint replacement towards early (secondary) prevention, it is very important to identify new osteoarthritic disease stage markers that could be measured in the early stages of OA. These markers should function as new targets or biomarkers for early disease treatment and prevention.

Radiography is routinely used to support in the diagnosis of OA. However, radiographic imaging is inadequate to detect and monitor biochemical changes within joint tissues which can occur long before symptoms are present. Magnetic resonance (MR) imaging is a non-invasive 3D imaging method with high tissue contrast that has been successfully used to visualize osteoarthritic changes [2]. Additionally to radiographic osteophyte formation and joint space loss, joint effusion can be assessed. Joint effusion is the presence of increased intra-articular fluid [3], which has been positively associated with knee pain in knee OA patients [4]. Joint effusion is known to be related with joint inflammation [5] and a recent study showed that occurrence of joint effusion is a strong predictor for development of incident radiographic OA [6].

As inflammation is increasingly considered to be an important pathway in the OA pathophysiology, efforts have been made to identify pro- and anti-inflammatory mediators (such as cytokines) which enable monitoring of the OA disease course [7-9]. With the aim to better understand the downstream consequences of inflammation in the knee, we compared gene expression levels in the blood of participants with different grades of joint effusion, as assessed by MR imaging. Ramos *et al.* already identified specific gene expression networks in blood associated with OA status [10]. Therefore, it could be advocated that blood expression profiles may reflect predisposition to OA. And because blood is a readily accessible tissue, gene expression levels associated with joint effusion may serve as molecular biomarkers for early detection of OA. We examined in two cohort studies whether joint effusion grades on MR imaging of the knee were associated with specific gene expression levels in the peripheral circulation, and subsequently performed a meta-analysis. Analysis for enrichment was performed to determine whether particular pathways were overrepresented among the genes associated with joint effusion.

MATERIALS AND METHODS

Subject selection

The Rotterdam Study (RS) is a large prospective, population-based cohort study in the district of Rotterdam, the Netherlands, investigating the prevalence, incidence, and risk factors of various chronic disabling diseases among elderly Caucasians aged 45 years and over. A detailed description of the design and rationale of the Rotterdam Study has been published elsewhere [11]. We invited the first 1,116 women aged 45-60 years visiting the research center to join a sub-study investigating early signs of knee osteoarthritis (knee OA). Participants were evaluated for the self-reported presence of rheumatoid arthritis (RA) and these cases were excluded. An additional exclusion criterion was the presence of any contra-indications for MR imaging, including weighing more than 150 kilograms. In total, 891 participants were included. For this study, we selected participants having both gene expression data and good quality knee MR imaging data available. In total, we could include 135 participants. The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC and by the Ministry of Health, Welfare and Sport of the Netherlands, implementing the *"Wet Bevolkingsonderzoek: ERGO (Population Studies Act: Rotterdam Study)"*. All participants provided written informed consent to participate in the study and to obtain information from their treating physicians.

The Genetics, Arthrosis and Progression study (GARP) consists of 191 sibling pairs (n=382) of white, Dutch ancestry. All participants (age range 40-78 years; mean age 60 years) are clinically and radiographically diagnosed with primary, symptomatic OA at multiple joint sites in the hand, or in at least two joints of the following locations: hand, spine (cervical or lumbar), knee, or hip [12]. Patients with secondary OA, such as inflammatory joint disease, major developmental diseases, bone dysplasia, major local factors or metabolic diseases as hemochromatosis were excluded. Sib pairs (n=105) with at least one subject with symptomatic hip or knee OA (but not in a radiographic end-stage) were eligible for the MR imaging sub-study [13]; in 5 out of 210 patients no MR imaging (one due to claustrophobia, one with a large knee that did not fit into the knee coil) or an MR imaging of insufficient quality (due to motion artefacts in three patients) was available. For this study, a subset of 98 women (including 28 sibs) was selected for which both gene expression data and knee MR imaging data were available. The GARP study has been approved by the Medical Ethics Committee of the Leiden University Medical Center, the Netherlands. All participants provided written informed consent to participate in the study.

Knee OA definition

In both RS and GARP, radiographs were scored to examine knee OA. Knee OA was defined as at least one definite osteophyte and definite joint space narrowing or at least two definite osteophytes (*Kellgren* and *Lawrence* (K/L) score ≥2).

MR acquisition

In the RS, all participants were scanned on a 1.5 T MRI scanner (General Electric Healthcare, Milwaukee, Wisconsin, USA) with an 8-channel cardiac coil, so that two knees could be scanned at once without repositioning the subject. The protocol consisted of a sagittal fast spin echo (FSE) proton density and T2 weighted sequence (repetition time (TR)=4,900 ms; echo time (TE)=11/90 ms, flip angle of 90-180, slice thickness 3.2 mm, field of view 15 cm^2), a sagittal FSE T2 weighted sequence with frequency selective fat suppression (TR/TE=6800/80 ms, flip angle=90-180, slice thickness=3.2 mm, field of view=15 cm^2), a sagittal spoiled gradient echo sequence with fat suppression (TR/TE=20.9/2.3 ms, flip angle=35, slice thickness=3.2 (1.6) mm, field of view=15 cm^2) and a fast-imaging employing steady-state acquisition (FIESTA) sequence (TR/TE=5.7/1.7 ms, flip angle=35, slice thickness=1.6 mm, field of view=15 cm^2). This FIESTA sequence was acquired in the sagittal plane. Total scanning time was 27 minutes for two knees per patient.

Acquisition of MR imaging in GARP was performed using a 1.5 – T MR imaging scanner (Philips Medical Systems, Best, The Netherlands) using a 4-channel transmit/receive knee coil as described elsewhere [13]. The following images were obtained: coronal proton density- and T2-weighted dual spin echo (SE) images (with TR=2,200 ms; TE=20/80 ms; 5 mm slice thickness; 0.5 mm intersection gap; 16 cm field of view; 206 x 256 acquisition matrix); sagittal proton density- and T2-weighted dual SE images (TR=2,200 ms; TE=20/80 ms; 4 mm slice thickness; 0.4 mm intersection gap; 16 cm field of view; 205 x 256 acquisition matrix); sagittal three-dimensional (3D) T1-weighted spoiled gradient echo (GE) frequency selective fat-suppressed images (TR=46 ms; TE=2,5 ms; flip angle 40°; 3.0 mm slice thickness; slice overlap 1.5 mm; no gap; 18 cm field of view; 205 x 256 acquisition matrix); and axial proton density- and T2-weighted turbo spin echo (TSE) fat-suppressed images (TR=2,500 ms; TE=7.1/40 ms; echo train length 6,2 mm slice thickness; no gap; 18 cm field of view; 205 x 256 acquisition matrix). Total acquisition time (including the initial survey sequence) was 30 min for one knee per patient. Since the original purpose of the MR imaging study in GARP was to assess progression of OA, only one knee was imaged and no images were obtained of a knee that already had a maximum K/L score of 4 [2].

Semi-quantitative joint effusion scoring

In RS, a trained reader (who was blinded for any clinical, radiographic and genetic data) scored all MR images of the knees with the semi-quantitative Knee Osteoarthritis Scoring System (KOSS), described in detail elsewhere [2]. The joint effusion grades in the tibiofemoral joint (TFJ) and the patellofemoral joint (PFJ) were scored together (grade 0-3): 0=joint effusion absent, 1=small joint effusion, 2=moderate joint effusion, and 3=massive joint effusion. The scores of the left and the right knee were summed, resulting in one grade per person ranging from 0 to 6. An experienced musculoskeletal radiologist, also blinded for any clinical, radiographic and genetic data, scored a random sample of MR images to determine the inter-observer reliability. The inter-observer reliability was moderate to good with an intra-class correlation coefficient (ICC) of 0.83.

In GARP, MR images were also scored according to KOSS [2] by three readers with 3, 15, and 25 years of experience in consensus, blinded to clinical, radiographic and genetic data, as described previously [13]. Presence of joint effusion was evaluated on T2-weigthed coronal, sagittal and axial sequences. A small, physiological sliver of synovial fluid was not recorded. A small effusion (grade 1) was present when a small amount of fluid distended one or two of the joint recesses, moderate effusion (grade 2) when more than two recesses were partially distended, and massive (grade 3) when there was full distension of all the joint recesses. As in RS, the grades were scored semi-quantitatively ranging from 0 to 3.

Because we used non-contrast-enhanced MR imaging in both GARP and RS, we could not measure synovial thickness reliably.

Gene expression levels

In RS, whole-blood was collected (PAXGene Tubes – Becton Dickinson) and total RNA was isolated (PAXGene Blood RNA kits - Qiagen). To ensure a constant high quality of the RNA preparations, all RNA samples were analyzed using the Labchip GX (Calliper) according to the manufacturer's instructions. Samples with an RNA Quality Score >7 were amplified and labelled (Ambion TotalPrep RNA), and hybridized to the Illumina HumanHT12v4 Expression Beadchips. Processing of the Rotterdam Study RNA samples was performed at the Genetic Laboratory of Internal Medicine, Erasmus University Medical Center Rotterdam, and the dataset has been deposited in the GEO database under the accession number GSE33828 [14].

For GARP, generation of gene expression levels in peripheral blood mononuclear cells (PBMCs) has been described elsewhere [10]. Gene expression data has been deposited in the GEO database under the accession number GSE48556.

Both RS and GARP samples were scanned on the Illumina iScan System (combined with an AutoLoader) using Illumina iScan image data acquisition software. Illumina GenomeStudio software (version 1.9.0) was used to generate output files for statistical analysis using R [15]. To identify transcripts that had detectable quantitative expression, we used the detection P-values reported by Illumina's GenomeStudio software. The detection P-value represents the confidence that a given transcript is expressed above the background defined by negative control probes. We called a transcript significantly expressed when the detection P-value was <0.05 in more than 50 percent of all samples. All other transcripts were excluded from analysis. Because of this stringent detection P-value cut-off, the overall false-positive rate is very small (we won't get false positive genes), whereas the false-negative rate might be higher (so we could lose some joint effusion associated genes, *i.e.,* genes that are expressed at high joint effusion grades specifically).

Statistical- and functional analysis

Raw gene expression intensities were normalized by quantile-normalization to the median distribution; gene expression levels were subsequently log2-transformed. To minimize the influence

of the overall signal levels, which may reflect RNA quantity and quality rather than a true biological difference between individuals, the probe means and sample means were centered to zero, and sample variance was linearly scaled, such that each sample had a standard deviation of one (standardization). To identify transcripts that were differentially expressed with joint effusion grades, we used four different linear regression models (lm):

- *Model 0: unadjusted: lm (probe ~ joint effusion grade)*
- *Model 1: adjusted for age + fasting status + RNA quality score (RQS) + batch + cell counts*
- *Model 2: adjusted for Model 1 + body mass index (BMI)*
- *Model 3: adjusted for Model 1 + BMI + nonsteroidal anti-inflammatory drug (NSAID) intake*

BMI was measured at the research centers (as weight in kg divided by height^2 in meters), and NSAID intake was extracted from the pharmacy records (RS) or collected via questionnaires (GARP). Because it is known that BMI is associated with markers of inflammation [16,17], and because additional adjustments for NSAID use (model 3) hardly changed the effect sizes and standard errors of the results as shown in the Supplementary Tables (S1-S2), we used model 2 for the meta-analysis and follow-up analyses. Notably, the analysis in GARP was also adjusted for sib ship in addition to age, batch, and BMI. In GARP, no adjustments were included for fasting status since blood was collected for all participants without fasting. Furthermore, gene expression levels were assessed from PBMCs and the RNA integrity number (or RQS) was at least 8.3 (36 random samples were analyzed) [10].

To be able to meta-analyze the results of both studies, we combined the 12,843 Illumina HT12v4 probes (RS) and the 12,246 Illumina HT12v3 probes (GARP) based on chromosomal position and nucleotide sequence: 9,507 probes (representing 7,408 unique genes) were similar between the two gene expression platforms and could be meta-analyzed.

We ran sample size weighted meta-analyses based on P-values and the direction of the effects. By using the P-values and the effect direction, a Z-statistic characterizing the evidence for association was calculated. The Z-statistic summarized the magnitude and the direction of the effect. An overall Z-statistic and P-value was calculated from the weighted sum of the individual statistics. Weights were proportional to the square-root of the number of individuals examined in each sample and standardized such that the squared weights sum to 1. We used the Meta-Analysis Tool for genome-wide association scans (METAL) [18] for this. METAL has been developed for meta-analyzing genetic genome-wide association studies. Because we are dealing with gene expression levels and not SNPs, we changed the SNPID column to probe IDs and assigned all probes a minor allele A and a major allele G, a minor allele frequency=0.10, and a + strand. For the positions, the probe chromosomes and the midpoint position of the probes were used. Sample sizes, effect directions, and P-values were extracted from the linear regression model results files. Probes with a meta-analysis P-value<6.75E-06 (0.05 / 7,408 genes tested) were considered transcriptome-wide significantly associated with the joint effusion grades in the knee.

Pathway analyses

Pathway analysis was done with the DAVID tool; the *Database for annotation, visualization and integrated discovery* [19]. We included all nominal significant genes (meta-analysis P-value<0.05), and checked for enrichment of any biological processes identified in the gene ontology database.

Analysis of protein interaction networks

To investigate protein interactions among the nominal significant genes, we used the *Search Tool for the Retrieval of Interacting Genes/Proteins* [20], which is available online. With the "enrichment" option, we checked for enrichment of protein-protein interactions and "*GO biological processes*".

RESULTS

Subjects

The complete characteristics of the included subjects of both RS and GARP are shown in Table 1 and Figure 1. In both RS and GARP, mean age of the subjects with and without joint effusion was not significantly different (ANOVA P-value RS=0.146, ANOVA P-value GARP=0.181). Mean BMI seemed to be higher with higher joint effusion grades, but due to small sample sizes this difference was not significant (ANOVA P-value RS=0.069, ANOVA P-value GARP=0.487).

Table 1. Subject characteristics of RS and GARP. **this can be in one or two knees.*

Grade	RS				GARP			
	#	Mean Age (±SD)	Mean BMI (±SD)	# knee OA*	#	Mean Age (±SD)	Mean BMI (±SD)	# knee OA*
Grade 0	65	54.0 (3.4)	26.6 (4.6)	3	47	60.7 (6.9)	26.5 (3.8)	21
Grade 1	30	55.1 (3.9)	28.0 (5.3)	1	46	58.9 (6.6)	25.8 (3.9)	30
Grade 2	30	54.8 (3.8)	26.9 (4.4)	1	5	58.3 (8.9)	27.2 (7.3)	4
Grade 3	6	56.2 (2.1)	29.8 (4.9)	1	0	-	-	-
Grade 4	4	52.0 (4.2)	36.8 (10.1)	1	0	-	-	-
Grade 5	0	-		-	0	-	-	-
Grade 6	0	-		-	0	-	-	-
Total	**135**	**54.5 (3.6)**	**27.4 (5.2)**	**7**	**98**	**59.7 (6.8)**	**26.1 (4.0)**	**55**

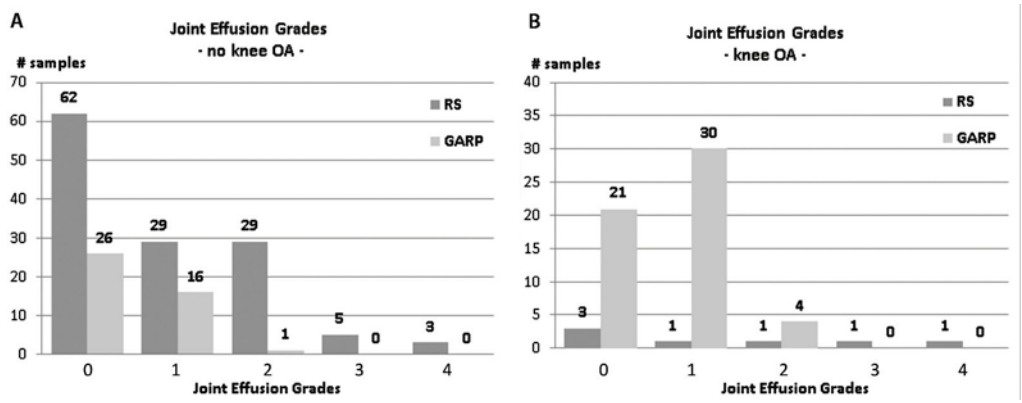


Figure 1. Joint effusion grades in subjects without knee OA (A) and samples with knee OA (B).

4

Results within the Rotterdam Study

Of the 7,408 genes tested, *CLEC4A* (C-type lectin domain family 4, member A) demonstrated the strongest association with joint effusion grades in the knee (effect size=0.407 (SE=0.120); P-value=9.57E-04). In total, 310 probes (representing 251 unique genes) were nominally significant. The top 50 results are shown in Supplementary Table 1.

Results within GARP

In GARP, the lowest P-value was found for the DNA-damage-inducible transcript 4 (*DDIT4*) gene (effect size=-1.425 (SE=0.411); P-value=5.21E-04). In total, 439 probes (representing 331 unique genes) were nominally significant (Supplementary Table 2).

Meta-analysis of the Rotterdam Study and GARP

In general, the top five genes of GARP and RS were different. To identify a common transcriptional signature for joint effusion, we performed a meta-analysis across RS and GARP. The top 20 results are shown in Table 2. All 257 nominally significant probes (representing 189 unique genes) are shown in Supplementary Table 3. The lowest P-value was found for the *C1orf38* (Chromosome 1 Open Reading Frame 38) gene, also called *THEMIS2* (Thymocyte Selection Associated Family Member 2) or *ICB-1* (Induced by Contact to Basement membrane) (Z-score=-3.356; P-value=7.90E-04). Gene expression levels of *C1orf38* were lower in samples with higher joint effusion grades in both whole blood and PBMCs (Supplementary Figure S1). Also the *DYNLL2* gene (Dynein, Light Chain, LC8-Type 2), the *NFATC1* gene (Nuclear factor of activated T-cells, cytoplasmic 1), and the *RBM4* gene (RNA Binding Motif Protein 4) were nominally associated, with respectively higher (*DYNLL2* and *NFATC1*) and lower (*RBM4*) gene expression levels correlating with advanced joint effusion grades (Supplementary Figure S2-4).

Table 2. Top 20 results of the meta-analysis (n=257)

Gene	ILMN ID	RS			GARP			META-ANALYSIS			RS position	GARP position
		Effect	SE	P-value	Effect	SE	P-value	Zscore	P-value	Dir		
C1orf38	2470240	-0.132	0.056	2.00E-02	-0.170	0.121	1.63E-01	-3.356	7.90E-04	--	56	93
GABPB1	7200431	-0.108	0.053	4.35E-02	-0.190	0.094	4.30E-02	-3.325	8.84E-04	--	185	27
TMEM97	3420541	-0.118	0.053	2.65E-02	-0.304	0.126	1.61E-02	-3.099	1.94E-03	--	95	167
DYNLL2	3400551	0.090	0.053	8.91E-02	0.101	0.103	3.27E-01	3.087	2.02E-03	++	425	33
RBM4	510132	-0.081	0.054	1.36E-01	-0.285	0.093	2.10E-03	-3.067	2.16E-03	--	739	15
PRICKLE1	1770224	0.124	0.053	2.03E-02	0.400	0.177	2.41E-02	2.993	2.76E-03	++	54	429
AP3B1	2230603	0.155	0.073	3.61E-02	0.203	0.103	4.80E-02	2.989	2.80E-03	++	133	195
TUBB2C	2070368	-0.062	0.068	3.60E-01	-0.349	0.102	5.98E-04	-2.922	3.48E-03	--	2634	2
FKBP14	6100411	0.141	0.052	7.69E-03	0.157	0.089	7.90E-02	2.915	3.56E-03	++	19	1499
GFM1\LXN	60670	0.108	0.066	1.05E-01	0.346	0.124	5.29E-03	2.899	3.74E-03	++	521	78
LYZ	4810162	-0.561	0.204	6.95E-03	-0.338	0.553	5.42E-01	-2.832	4.62E-03	--	15	2031
ARL6IP1	2690047	0.146	0.070	3.98E-02	0.189	0.112	9.07E-02	2.831	4.64E-03	++	148	359
PTPLB	6980253	0.091	0.067	1.79E-01	0.359	0.110	1.13E-03	2.809	4.96E-03	++	1098	32
MED19	3450427	-0.099	0.059	9.51E-02	-0.084	0.116	4.66E-01	-2.79	5.28E-03	--	478	117
APTX	1570138	-0.081	0.040	4.37E-02	-0.118	0.083	1.57E-01	-2.783	5.39E-03	--	174	371
NFATC1	940725	0.112	0.061	7.11E-02	0.216	0.120	7.15E-02	2.778	5.46E-03	++	319	194
RG9MTD3\SHB	1770196	0.147	0.051	5.04E-03	0.174	0.098	7.65E-02	2.755	5.87E-03	++	10	3222
POLR2J	6350333	-0.154	0.056	7.14E-03	-0.445	0.152	3.46E-03	-2.74	6.14E-03	--	20	2330
-	1070754	0.170	0.056	3.14E-03	-0.038	0.074	6.05E-01	2.737	6.20E-03	++	7	4195
IFT43	3170458	-0.130	0.043	2.99E-03	-0.152	0.081	6.15E-02	-2.734	6.26E-03	--	8	4228

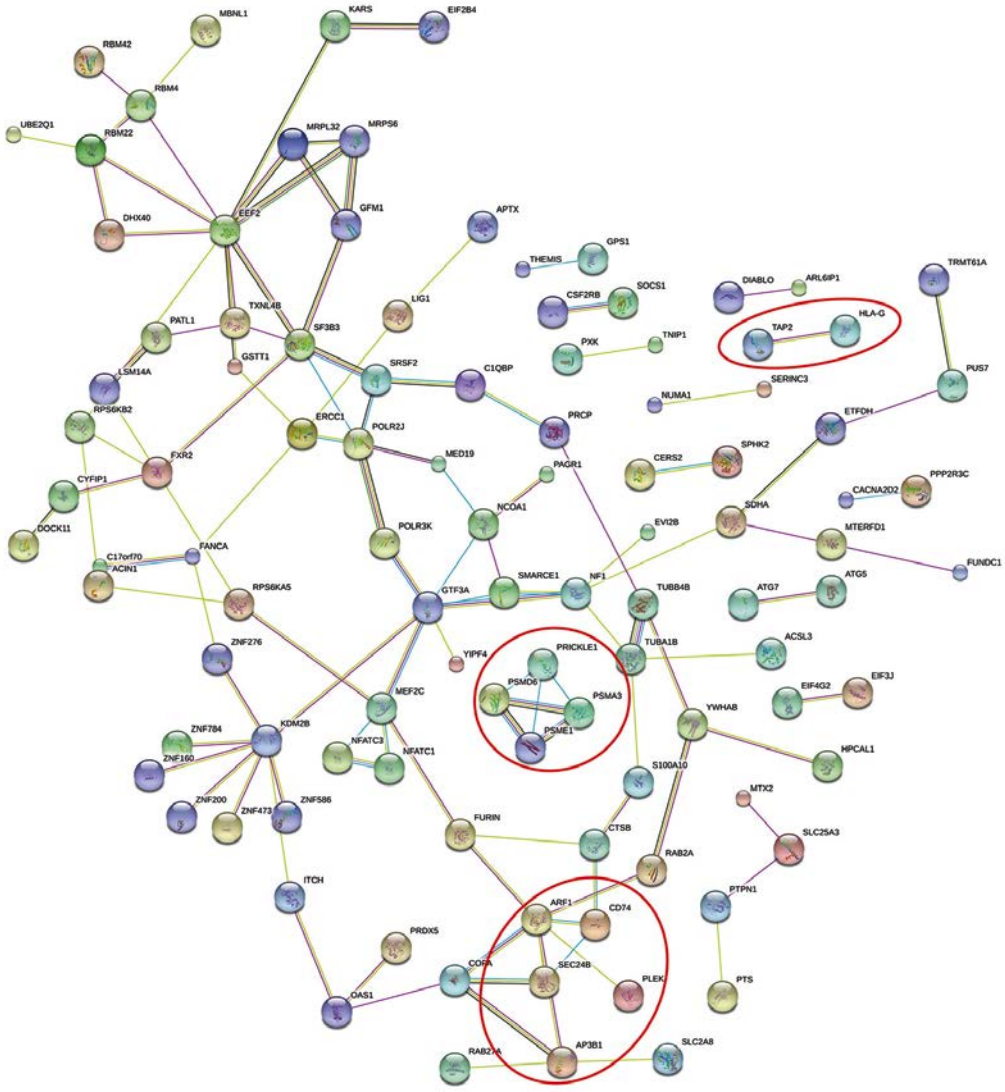


Figure 2. Protein-protein interactions determined with STRING showing interactions between the 178 nominally associated genes (P-value<0.05), marking proteins involved in antigen processing and presentation of exogenous antigens in red (GO:0019884). Disconnected proteins are hidden.

Pathway-analysis of genes nominally significant in the meta-analysis

Using the 257 nominally associated probes (P-value<0.05), 178 genes were recognized by the webtool DAVID. The most significant GO terms identified were: *intracellular protein transport* (GO:0006886: 13 of 374 genes, P-value=4.5E-04, Fold Enrichment (FE)=3.4), *response to stress* (GO:0006950: 34 of 1685 genes, P-value=1.5E-04, FE=2.0), *antigen processing and presentation of exogenous antigens*

(GO:0019884: 4 of 14 genes, P-value=3.5E-04, FE=27.8), but the three GO terms did not survive the Benjamini Hochberg multiple testing correction. Additionally, one KEGG pathway was nominally significantly enriched: *antigen processing and presentation* (hsa04612: 5 of 83 genes, P-value=0.0127, FE=5.4).

Using the webtool STRING, we did not find significantly enriched protein-protein interactions within the network of 178 genes (P-value=0.386, observed interactions=58, expected interactions=55). However, STRING confirmed two significantly enriched biological pathways identified with DAVID: *response to stress* (45 of 1685 genes, P-value=6.23E-03) and *antigen processing and presentation of exogenous antigens* (10 of 14 genes, P-value=3.44E-02). The protein-protein interactions are visualized in Figure 2. Proteins involved in the antigen processing and presentation of exogenous antigens pathway (GO:0019884) are marked red, highlighting a cluster of three proteasomes (PSMA3, PSMD6, and PSME1) important for the antigen processing pathway.

DISCUSSION

We examined whether joint effusion grades in the knee were associated with specific gene expression levels in the circulation, which could potentially serve as molecular biomarker to indicate OA in the early stage. We identified 257 nominally associated probes (P-value<0.05) mapping to 189 unique genes. *C1orf38*, *DYNLL2*, and *RBM4* were among the 5 most significant genes in the meta-analysis. Additional adjustments for BMI and NSAID intake did not notably affect the results, suggesting that the associations are consistent across all BMI ranges and in both users and non-users of NSAIDs. Subsequent pathway analyses with DAVID revealed nominal significant enrichment of genes involved in response to stress, gene expression, negative regulation of intracellular signal transduction, and antigen processing and presentation of exogenous antigens pathways. The biological pathways response to stress and antigen processing and presentation of exogenous antigens were confirmed with a second pathway analysis tool STRING.

C1orf38 is a protein-coding gene and is highly expressed in several blood cells (monocytes, dendritic cells, NK-cells, T-cells, B-cells). The gene is induced by interferon-gamma (IFN-γ), an important cytokine that orchestrates many distinct cellular processes regarding inflammation [21]. Therefore, *C1orf38* could be an interesting candidate for further research.

Cytoplasmic dynein consists of a molecular complex of several proteins including *DYNLL2*, and it is thought to play a role in movement and positioning of a wide range of organelles and complexes in the cell [22]. Notably, recent studies showed that *DYNLL2* inhibits inflammation and may also inhibit osteoclastogenesis and bone resorption via regulation of *NFκB* transcription activity [23]. This would suggest that the higher expression of *DYNLL2* in association with higher joint effusion grades is rather consequence than cause, however, this remains to be established.

RBM4 is thought to play a role in alternative splice site selection during pre-mRNA processing, and seems to be important for the regulation of the translation of pro-inflammatory genes [24].

Of note is the association of higher joint effusion grades with higher expression levels of *NFATC1* (nuclear factor of activated T cells 1). Besides its function in bone remodeling through calcium/calcineurin signaling, *NFATC1* belongs to a family of transcription factors that play a central role in inducible gene transcription during immune response [25]. Although no significant differences were found in *NFATC1* gene expression between OA-affected and unaffected tissues using microarray analyses [26-29], a slight but significant reduction was detected by RT-qPCR in OA affected cartilage [30]. In addition, Jeffries and colleagues [31] found changes in DNA methylation profiles, and it was shown that cartilage-specific ablation of *NFATC1* predisposes to development of early onset OA too [30]. Since the expression of *NFATC1* is positively associated with joint effusion it could be speculated that, in line with the increased expression of *DYNLL2*, upon occurrence of joint effusion specific pathways are activated to protect against development of OA. Consistent with this hypothesis, we observed that increased expression of *NFATC1* in association with joint effusion is much more pronounced in subjects without knee OA in GARP. Therefore, *NFATC1* might be a useful biomarker for early detection of OA. However, this should be confirmed in a longitudinal study tracking the development of the disease.

The pathway enrichment analysis results were consistent with known inflammatory disease mechanisms including response to stress and gene expression. Cellular stress and inflammation are known to reciprocally activate or inhibit each other, depending on the immune cell type and the stress-inducing signals [32]. Additionally, we identified the pathways negative regulation of intracellular signal transduction (GO:1902532) and antigen processing and presentation of exogenous antigens (GO:0019884). Hanada *et al.* [33] already highlighted a key role for the intracellular signal transduction pathways of the pro- and anti-inflammatory cytokines which activate inflammatory transcription factors such as *NF-κB*, *Smad*, and *STATs*. The antigen processing machinery can be easily linked to the inflammatory response too [34].

STRING showed the interaction between 3 proteasomes identified in the analysis (PSMA3, PSMD6, and PSME1). Proteasomes are important for degrading intracellular proteins, and recently it has been shown that mutations and polymorphisms in the proteasome are associated with several inflammatory and auto-inflammatory diseases [35]. Therefore, these genes could also be interesting targets for future studies.

Despite the identification of several compelling potential markers for early OA, major drawback of the current study was the relatively small sample size (n=233). Although gene expression data and knee MR images are available in larger datasets, the number of samples in which both measurements were determined is unfortunately limited. In addition, the data of the two cohorts (RS and GARP) was rather heterogeneous in particular due to the fact that in RS joint effusion grades were combined for two knees (sum of left and right knee), while in GARP joint effusion was determined in one randomly

selected knee. Moreover, GARP is a cohort of clinical OA cases while RS is a population-based cohort study, in which no selection was made for OA cases specifically: in RS only seven out of 135 subjects (5.2%) were diagnosed with radiographically evident knee osteoarthritis, while in GARP 55 out of 98 subjects (56.1%) had knee OA. Furthermore, in RS the analyses were adjusted for fasting status (134 of 135 subjects fasted overnight) and RNA quality scores, while in GARP non-fasting subjects were used and RNA quality scores were available in a small subset only. Finally, gene expression levels in RS were determined in whole blood, while in GARP PBMCs were used. Although a previous study showed that expression levels differ across different RNA sources (whole blood, PBMCs, and lymphoblastoid cell lines), phenotype-based differential expression analyses results were consistent in whole blood and PBMCs [36]. Taken together, it is likely that cohort heterogeneity has resulted in limited power due to which no transcriptome-wide significant probes were identified. Earlier studies confirmed the good quality and reproducibility of our gene expression arrays [10,27,37]. Another potential limitation of our study is that we did not assess recent traumatic knee injuries: traumas can increase joint effusion and dilute our associations.

In conclusion, joint effusion grades in the knee on MR imaging were nominally associated with the expression levels of 189 unique genes in blood and the identified genes were mainly involved in inflammation. Although the associations presented in this manuscript were not transcriptome-wide significant, the meta-analysis and subsequent enriched biological pathways resulted in compelling candidate genes such as *C1orf38* and *NFATC1* that could be further characterized in future research. Additional studies are needed to replicate our findings as well as to identify other genes which will greatly help in understanding the pathophysiology of OA and its relation with inflammation, and may result in biomarkers urgently needed to diagnose OA at an early stage.

ACKNOWLEDGEMENTS

This study was funded by the European Commission (HEALTH-F2-2008-201865, GEFOS; HEALTH-F2-2008 35627, TREAT-OA), Netherlands Organisation for Scientific Research (NWO) Investments (nr. 175.010.2005.011, 911-03-012), the Netherlands Consortium for Healthy Aging (NCHA), the Netherlands Genomics Initiative (NGI) / Netherlands Organization for Scientific Research (NWO) project nr. 050-060-810 and Vidi grant 917103521.

The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organisation for the Health Research and Development (ZonMw), Netherlands Organisation of Scientific Research NWO Investments (nr. 175.010.2005.011, 911-03-012), the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. We are very grateful to Dr. E. Odding, Dr. A.P. Berging, Dr.M.Reijman, Dr. S. Dahaghin, Dr. H.J.M. Kerkhof, Ms. A. van Vaalen and Mr. M. Kool for scoring the knee radiographs.

The generation and management of RNA-expression array data for the Rotterdam Study was executed and funded by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Netherlands. We thank Ms. Mila Jhamai, Ms. Jeannette M. Vergeer-Drop, Ms. Bernadette van Ast-Copier, Mr. Marijn Verkerk and Jeroen van Rooij, BSc for their help in creating the RNA array expression database.

The GARP study was supported by the Leiden University Medical Centre and the Dutch Arthritis Association. Pfizer Inc., Groton, CT, USA supported the inclusion of the GARP study. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2011) under grant agreement n° 259679 and from BBMRI-NL, a research infrastructure financed by the Dutch government: NWO 184.021.007. Expression in blood was part of the Dutch Arthritis Association grant 10-1-402. We are indebted to Dr. H.M. Kroon and Dr. N. Riyazi for scoring the radiographs and to Dr. P. Kornaat and Dr. R. Ceulemans for scoring the knee MR images.

The authors are very grateful to the study participants, the staff, the general practitioners, and the pharmacist from both participating studies.

ADDITIONAL INFORMATION

Supplementary Information is available on request.

REFERENCES

1. Dieppe PA, Lohmander LS (2005) Pathogenesis and management of pain in osteoarthritis. Lancet 365: 965-973.
2. Kornaat PR, Ceulemans RY, Kroon HM, Riyazi N, Kloppenburg M, et al. (2005) MRI assessment of knee osteoarthritis: Knee Osteoarthritis Scoring System (KOSS)--inter-observer and intra-observer reproducibility of a compartment-based scoring system. Skeletal Radiol 34: 95-102.
3. Mathison DJ, Teach SJ (2009) Approach to knee effusions. Pediatr Emerg Care 25: 773-786; quiz 787-778.
4. Lo GH, McAlindon TE, Niu J, Zhang Y, Beals C, et al. (2009) Bone marrow lesions and joint effusion are strongly and independently associated with weight-bearing pain in knee osteoarthritis: data from the osteoarthritis initiative. Osteoarthritis Cartilage 17: 1562-1569.
5. Johnson MW (2000) Acute knee effusions: a systematic approach to diagnosis. Am Fam Physician 61: 2391-2400.
6. Atukorala I, Kwoh CK, Guermazi A, Roemer FW, Boudreau RM, et al. (2014) Synovitis in knee osteoarthritis: a precursor of disease? Ann Rheum Dis.
7. Mabey T, Honsawek S (2015) Cytokines as biochemical markers for knee osteoarthritis. World J Orthop 6: 95-105.
8. Zivanovic S, Rackov LP, Zivanovic A, Jevtic M, Nikolic S, et al. (2011) Cartilage oligomeric matrix protein - inflammation biomarker in knee osteoarthritis. Bosn J Basic Med Sci 11: 27-32.
9. Attur M, Krasnokutsky S, Statnikov A, Samuels J, Li Z, et al. (2015) Low-Grade inflammation in symptomatic knee osteoarthritis: Prognostic value of inflammatory plasma lipids and peripheral blood leukocyte biomarkers. Arthritis Rheumatol.
10. Ramos YF, Bos SD, Lakenberg N, Bohringer S, den Hollander WJ, et al. (2014) Genes expressed in blood link osteoarthritis with apoptotic pathways. Ann Rheum Dis 73: 1844-1853.
11. Hofman A, Darwish Murad S, van Duijn CM, Franco OH, Goedegebure A, et al. (2013) The Rotterdam Study: 2014 objectives and design update. Eur J Epidemiol 28: 889-926.
12. Riyazi N, Meulenbelt I, Kroon HM, Ronday KH, Hellio le Graverand MP, et al. (2005) Evidence for familial aggregation of hand, hip, and spine but not knee osteoarthritis in siblings with multiple joint involvement: the GARP study. Ann Rheum Dis 64: 438-443.
13. Kornaat PR, Bloem JL, Ceulemans RY, Riyazi N, Rosendaal FR, et al. (2006) Osteoarthritis of the knee: association between clinical features and MR imaging findings. Radiology 239: 811-817.
14. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, et al. (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet 45: 1238-1243.
15. R-Development-Core-Team (2008) R: A language and environment for statistical computing.: R Foundation for Statistical Computing.
16. Kitahara CM, Trabert B, Katki HA, Chaturvedi AK, Kemp TJ, et al. (2014) Body mass index, physical activity, and serum markers of inflammation, immunity, and insulin resistance. Cancer Epidemiol Biomarkers Prev 23: 2840-2849.
17. Siervo M, Ruggiero D, Sorice R, Nutile T, Aversano M, et al. (2012) Body mass index is directly associated with biomarkers of angiogenesis and inflammation in children and adolescents. Nutrition 28: 262-266.
18. Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26: 2190-2191.
19. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4: P3.
20. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, et al. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41: D808-815.
21. Treeck O, Kindzorra I, Pauser K, Treeck L, Ortmann O (2005) Expression of icb-1 gene is interferon-gamma inducible in breast and ovarian cancer cell lines and affects the IFN gamma-response of SK-OV-3 ovarian cancer cells. Cytokine 32: 137-142.
22. Pfister KK (2015) Distinct functional roles of cytoplasmic dynein defined by the intermediate chain isoforms. Exp Cell Res.
23. Kim H, Hyeon S, Yang Y, Huh JY, Park DR, et al. (2013) Dynein light chain LC8 inhibits osteoclast differentiation and prevents bone loss in mice. J Immunol 190: 1312-1318.

24. Brudecki L, Ferguson DA, McCall CE, El Gazzar M (2013) Mitogen-activated protein kinase phosphatase 1 disrupts proinflammatory protein synthesis in endotoxin-adapted monocytes. Clin Vaccine Immunol 20: 1396-1404.

25. Takayanagi H (2005) Inflammatory bone destruction and osteoimmunology. J Periodontal Res 40: 287-293.

26. Xu Y, Barter MJ, Swan DC, Rankin KS, Rowan AD, et al. (2012) Identification of the pathogenic pathways in osteoarthritic hip cartilage: commonality and discord between hip and knee OA. Osteoarthritis Cartilage 20: 1029-1038.

27. Ramos YF, den Hollander W, Bovee JV, Bomer N, van der Breggen R, et al. (2014) Genes involved in the osteoarthritis process identified through genome wide expression analysis in articular cartilage; the RAAK study. PLoS One 9: e103056.

28. Chou CH, Wu CC, Song IW, Chuang HP, Lu LS, et al. (2013) Genome-wide expression profiles of subchondral bone in osteoarthritis. Arthritis Res Ther 15: R190.

29. Lambert C, Dubuc JE, Montell E, Verges J, Munaut C, et al. (2014) Gene expression pattern of cells from inflamed and normal areas of osteoarthritis synovial membrane. Arthritis Rheumatol 66: 960-968.

30. Greenblatt MB, Ritter SY, Wright J, Tsang K, Hu D, et al. (2013) NFATc1 and NFATc2 repress spontaneous osteoarthritis. Proc Natl Acad Sci U S A 110: 19914-19919.

31. Jeffries MA, Donica M, Baker LW, Stevenson ME, Annan AC, et al. (2014) Genome-wide DNA methylation study identifies significant epigenomic changes in osteoarthritic cartilage. Arthritis Rheumatol 66: 2804-2815.

32. Muralidharan S, Mandrekar P (2013) Cellular stress response and innate immune signaling: integrating pathways in host defense and inflammation. J Leukoc Biol 94: 1167-1184.

33. Hanada T, Yoshimura A (2002) Regulation of cytokine signaling and inflammation. Cytokine Growth Factor Rev 13: 413-421.

34. Kasajima A, Sers C, Sasano H, Johrens K, Stenzinger A, et al. (2010) Down-regulation of the antigen processing machinery is linked to a loss of inflammatory response in colorectal cancer. Hum Pathol 41: 1758-1769.

35. Gomes AV (2013) Genetics of proteasome diseases. Scientifica (Cairo) 2013: 637629.

36. Joehanes R, Johnson AD, Barb JJ, Raghavachari N, Liu P, et al. (2012) Gene expression analysis of whole blood, peripheral blood mononuclear cells, and lymphoblastoid cell lines from the Framingham Heart Study. Physiol Genomics 44: 59-75.

37. Peters MJ, Joehanes R, Pilling LC, Schurmann C, Conneely KC, et al. (2015) The transcriptional landscape of age in human peripheral blood. Nat Commun 6: 8570.

CHAPTER 5

General Discussion

The overall objective of this thesis was to discover novel genomic factors involved in ageing and age-related diseases by integrating different genomic approaches. We assessed the effects of genetic variants, gene expression levels, and DNA methylation levels on age-related phenotypes, and additionally we combined the different levels of -omics data (genomics, epigenomics, and transcriptomics). In this chapter, the challenges of population-based genomic studies are placed in a broader context. Furthermore, suggestions for future research are given.

POPULATION BASED STUDIES

When searching for novel genomic factors, there are a number of advantages of using large scale population based cohort studies. The size of the studies is very important: when studying genetics and genomics, large studies are required to identify the subtle individual effects of each of the genetic variants. A large population-based study is a good representation of the individuals living in a certain area, and both healthy and diseased subjects are included. In the Rotterdam Study, not only the number of subjects is large, but also the number of traits (phenotypes) studied in these samples. This makes it possible to study ageing in a broader context, adjust for many confounders, and study interactions between different phenotypes. Another advantage of population-based cohort studies is the longitudinal design: the period of follow-up gives the opportunity to not only study disease states in a cross-sectional fashion, but also answer research questions on incidence and progression of disease, and importantly, longitudinal data can be used to study causality. A longitudinal design also allows to introduce novel measurements in an already running cohort study with all available (historic) measurements.

The main disadvantage of large population based cohort studies is the fact that the study is not specifically designed for one's phenotype of interest. This is specifically a problem if the phenotype has a low prevalence. Examples of low prevalence diseases are Behcet's disease, Crohn's disease, Cystic Fibrosis, and Prader-Willi syndrome. Because of the low prevalence, only a handful of cases will be identified in the Rotterdam Study. With such low numbers of cases, one cannot properly study the genetics and genomics of a disease.

Another disadvantage of large population based studies is that, because of the large sample size, some measurements (e.g., RNA sequencing) remain too expensive to be applied to the complete study population and so subsets have to be selected. Therefore, selection bias can be a problem, and the power for analyses is reduced, or the measurements are done with a less expensive (and less accurate) method. In summary, robust, fast and easy-to-integrate measurements are most beneficial in large population-based studies.

Many recently started studies make use of electronic health records (ehealth records) of the general practitioners and hospitals in a certain area. Additionally, dental records, pharmacy dispensing records, education and environmental records, and social services records can be included. In

comparison with population-based studies, these studies do not have to invest in research centers and repeated measurements. However, these studies will not gain information about the health status of subjects not visiting the general practitioner and/or hospital. Therefore, population-based cohort studies are still very important.

In the United Kingdom, they started a national health resource, called the UK Biobank, which is partly funded by the Scottish and Welsh government. They recruited 500,000 people aged between 40 and 69 years in 2006-2010 from across the country to participate, with the aim to improve the prevention, diagnosis, and treatment of a wide range of serious and life-threatening illnesses. In addition to baseline measurements, participants complete detailed web-based questionnaires and confirm integrating their ehealth records [1]. Combining health measurements and ehealth records on a nation-wide level will be very valuable to better understanding of the pathogenesis of diseases. There are many more world-wide initiatives like the UK Biobank, for example, Genomics England, the Million Veteran Progam (USA), the Regeneron 100k-patient genomics study, and 23andme. 23andme is designed a bit different, because the participants pay to join this study: the company sells personal ancestry information based on your genetic variants [1], and uses the genetic information in combination with additional questionnaires for further research. Most data of these large initiatives is accessible: all researchers can apply to use the resources for health related research that is of public interest.

COMBINING DIFFERENT MEASUREMENT PLATFORMS

A main driving force in genomics has been the development of novel technology, which allows new discoveries. This also introduces challenges when combining data. Examples are the introduction of next generation sequencing technology around 2010, just after microarray technology became main stream in genomics. When combining data of multiple large population-based studies in a meta-analysis approach, similar genomic data (e.g., gene expression levels) can be measured with different platforms. For example, genome-wide gene expression levels can be measured with RNA sequencing or gene expression microarrays, and microarrays are available from several different companies, such as Illumina (San Diego, CA, USA), Affymetrix (Santa Clara, CA, USA), and Agilent Technologies (Santa Clara, CA, USA).

Combining different microarray platforms

The different microarray platforms use different probes and cover different genes. For example, the Affymetrix gene expression levels are based on multiple probes per gene, while the Illumina genes are covered by one or two probes at the 3' end of the transcript. Additionally, both array platforms use different annotation files. We decided to combine the different platforms using gene-based levels for both platforms (using average gene expression levels per gene). For non-matching genes, we examined the physical location of the probes (on the array) and intersected these to gain more overlapping genes.

Combining microarray data and RNA-seq

Additionally, different measurement techniques have different resolutions and quality. For example, RNA-seq measures read counts for each gene expressed, while microarrays use pre-designed complement sequence detection probes catching the relative amount of expressed genes, which give a level of hybridization signal. By analyzing microarray data, cross-hybridization, non-specific hybridization, and the limited detection range are important technical issues. The RNA-seq analysis does not rely on pre-designed probes, but has difficulties with alignment (for example, pseudogenes can strongly impact the alignment). These differences have impact on what expression levels can be measured, i.e., it has been shown that RNA-seq is superior in detecting low abundance transcripts (if sequenced at high sequencing depth), differentiating biologically critical isoforms, and allowing the identification of genetic variants [2].

The main advantage of combining data of different measurement platforms is that the significantly associated genes represent very robust signals. Independent of the technique used to measure the gene expression levels, the gene seems to be important for the phenotype of interest. However, the concordance between RNA-seq and microarrays is far from perfect yet. There are batch effects because of different laboratories producing the data and different production schemes of companies making the essential consumables. Such technological variation needs to be recognized and controlled for as much as possible by standardization and harmonization. In addition, the analysis pipeline for RNA-seq data is still in development. For a better concordance, we need to align the RNA-seq reads to the 3'exons only (like the Illumina array). This might improve the correlation between the gene expression levels measured with the microarrays and RNA-seq.

The Illumina Infinium Human Methylation 450K BeadChip array

In this thesis, the Illumina Infinium Human Methylation 450K Beadchip array was used for measuring DNA methylation levels in all CHARGE cohorts. The methylation arrays offer a fast and cost-effective solution for profiling a relatively large number of CpG sites in one analysis. However, the Illumina 450K array contains only 1.7% of the 28.2 million known CpG sites [3]. The definitions of the CpG locations are visualized in Figure 1. In Table 1, the locations of the CpGs on the Illumina 450K array (according to the Illumina 450K annotation file) and the locations of all known CpGs [4] are provided. For both groups of CpGs, we intersected the CpG positions with the locations of the 28,691 CpG islands (UCSC Genome Browser hg19, track "CpG islands") and their shores and shelves.

Islands were defined based on UCSC criteria: CG content >50%, a CpG observed/expected ratio >0.6, and length >200bp. Shores were defined as the 2kb up- and downstream of the CpG islands, and the shelves were defined as the 2kb up-and downstream of the shores.

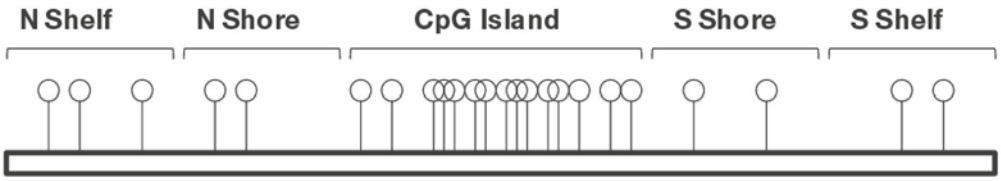


Figure 1. The definition of CpG islands and their flanking shores and shelves *(image adapted from Illumina's 450K datasheet)*.

Table 1. The number of CpGs in CpG islands, in shores, in shelves, and off island on the Illumina 450K array compared to all known CpGs.

CpG location	All known CpGs (28.2M) (%)	CpGs on 450K array (%)
CpGs in CpG islands	2,089,538 (7.4%)	150,254 (30.9%)
CpGs in shores	2,023,011 (7.2%)	112,110 (23.1%)
CpGs in shelves	1,184,491 (4.2%)	47,109 (9.7%)
CpGs off island	22,920,408 (81.2%)	176,039 (36.3%)
SUM	**28,217,448 (100%)**	**485,512 (100%)**

As presented in Table 1, the CpGs on the 450K methylation array are extremely enriched for CpGs in CpG islands and their flanking shores and shelves (96% of the CpG islands are covered) compared to all known CpGs. Illumina selected the content of the array with the guidance of a consortium of methylation experts comprising 22 members representing 19 institutions worldwide. At that point in time, it was thought that CpG islands were the most relevant areas of the genome involved in regulatory processes. Based on current knowledge, we know that CpG island methylation only accounts for a fraction of the variation in gene expression, and methylation in other domains like enhancers is hypothesized to play a far larger role than previously anticipated [5]. Table 2 shows how many of the CpGs on the 450K array are present in active promoters and/or enhancers for three different blood cell types: primary T-cells (tissue E034), primary B-cells (tissue E032), and primary monocytes (tissue E029).

Table 2. The number of CpGs in functional regions across different blood cell types. The active enhancer and promoter regions in each blood cell type were defined by the NIH Roadmap Epigenomics Consortium [6]. We intersected the locations of the CpGs with the positions of the active enhancers and promoters in three blood cell types: primary T-cells (tissue E034), primary B-cells (tissue E032), and primary monocytes (tissue E029). The number (and the percentage) of CpGs located in active enhancers and promoter regions is shown.

Blood cell type	Enhancers	Promoters	Enhancer & Promoter	Other locations
All known CpGs (28.2M)				
Primary T-cells	2,131,734 (7.6%)	1,732,227 (6.1%)	1,121 (<0.01%)	24,352,366 (86.3%)
Primary B-cells	1,981,933 (7.0%)	1,616,856 (5.7%)	1,087 (<0.01%)	24,617,572 (87.2%)
Primary monocytes	1,979,686 (7.0%)	1,585,502 (5.6%)	1,080 (<0.01%)	24,651,180 (87.4%)
450K array (485,512)				
Primary T-cells	77,348 (15.9%)	142,941 (29.4%)	67 (0.01%)	265,156 (54.6%)
Primary B-cells	74,730 (15.4%)	135,747 (28.0%)	65 (0.01%)	274,970 (56.6%)
Primary monocytes	67,405 (13.9%)	133,054 (27.4%)	74 (0.02%)	284,979 (58.7%)

As shown in Table 2, the number of CpGs in functional regions are quite similar across the three different blood cell types tested: of all known CpGs, about 7% is located in enhancer regions and about 6% is located in promoter regions. Less than 0.01% of the CpGs is located in a regions with both enhancer and promoter activity. As expected, the CpGs on the 450K array are enriched for promoter regions (approximately 28% of the CpGs on the array are located in promoter regions). In addition, the 450K array is also enriched for CpGs in functional enhancer regions: about 15% of the CpGs on the array are located in enhancers.

But of course, the array also misses many CpGs located in enhancers and promoters. Of the 2 million CpGs located in active enhancer regions, we measured only 73,000 CpGs on the 450K array (3.6%) (Table 2). These 73,000 CpGs tag 23.8% of the active enhancers in blood.

The same is true for the promoter regions: of the 1.6 million CpGs located in active promoter regions, we measured only 137,000 CpGs on the 450K array (8.3%) (Table 2). These 137,000 CpGs tag 64.3% of all active promoter regions in blood.

Therefore, all epigenetic results presented in the manuscript in chapter 2.1 on the analysis of gene expression and DNA methylation with ageing are biased towards the content of the 450K methylation array. Because of the pre-selection of the 450K CpGs, the design of the array was not hypothesis neutral and we missed many of the regulatory regions.

A new methylation array: the Illumina Infinium MethylationEPIC BeadChip

Very recently, Illumina introduced a new human methylation array, which is called the *Illumina Infinium MethylationEPIC BeadChip*. This array builds upon the 450K array (with >90% of the original

CpGs) plus an additional 350,000 CpGs in enhancer regions. These enhancer regions have been identified in the ENCODE project and the FANTOM5 project across multiple tissue types. The additional CpGs will definitely improve the coverage of the potentially functional enhancer regions. However, both methylation arrays are not able to measure allele specific methylation: the arrays measure the methylation percentage at a certain CpG site, and not which allele is methylated. Additionally, the Illumina arrays are not able to measure two CpGs close to each other. Because DNA methylation is dynamic data (this will be discussed in the next paragraph), it is not possible to impute CpGs close to each other: the correlation between CpGs may differ across different tissues and time, depending on stimuli.

Other techniques for measuring DNA methylation

To overcome the limitations of the 450K array, next-generation sequencing based modalities could be good alternatives. One option is whole-genome bisulfite sequencing (WGBS), which can quantify DNA methylation levels of all 28 million CpG sites, together with which allele is methylated [3]: single-cytosine methylation levels are measured genome-wide. However, WGBS is prohibitively expensive for most studies because this method requires resequencing the entire genome at high depth [7]. Another option is using Reduced Representation Bisulfite Sequencing (RRBS), but this method is also biased towards the CpG islands in the gene promoter and covers only a minor proportion of variable CpGs. One of the recent options is methylC-capture (MCC) sequencing. This method can be used for measuring both genetic variation and methylation levels: it targets a pre-designed selection of CpGs and SNPs in a disease-relevant tissue [8]. For example, the capture panel of Roche Nimblegen, Inc. (Madison, WI, USA) targets about five million dynamic CpGs mapping to regulatory elements in blood, and includes all SNPs linked to auto-immune diseases. The main disadvantage of MCC-sequencing is the fact that analyzing the data is still challenging: the analysis is more complex and the data storage is more demanding. Additionally, MMC-sequencing is much more expensive than running 450K arrays.

DYNAMIC DATA

As mentioned before, gene expression levels and DNA methylation are dynamic: they can vary between tissues vary in time, and vary between different disease states. In contrast to genetic variants where a specific SNP is present or not, gene expression and DNA methylation levels strongly depend on the tissue type being measured. A single sequencing experiment in whole blood will only offer information regarding that specific tissue (whole blood) at one point in time. This argues for studying immune and blood-related phenotypes when analyzing gene expression and DNA methylation levels in blood. One should be cautious in relating gene or methylation levels to other (not-blood related) phenotypes: the associations could be driven by confounders.

The eQTL studies (chapter 3.1 and 3.2) showed that gene expression levels in blood are mostly regulated by SNPs identified in GWAS studies for immune and blood-related phenotypes (for

example, celiac disease, systemic lupus erythematosus (SLE), the cholesterol metabolism, and type 1 diabetes). And this phenomenon is even stronger in the *trans*-eQTL analyses: in our blood analysis, the *trans*-eQTLs SNPs are enriched for associations with (auto)immune diseases or hematological traits. Blood itself is a heterogeneous collection of cell types, and comparisons between blood cell types showed that the number of shared eQTLs varies widely with the cell type and tissues studied. For example, a comparison of B-cells and monocytes showed that 21.8% of the detected *cis*-eQTLs were shared and only 7% of the detected *trans*-eQTLs were shared between tissues [9]. This suggests that the genetic regulation in *trans* is more cell type specific than the regulation in *cis,* and shows that the genetic regulation of gene expression is complex and differs across cell types and tissues.

MEASUREMENTS IN WHOLE BLOOD

In population-based cohorts, only easily accessible tissues can be gathered, such as blood, urine, stool, and saliva. In this thesis, we used genomic data from peripheral blood, which represents a heterogeneous pool of different cell types and subtypes with variable patterns of gene expression levels [10]. Therefore, cell composition is an important confounder, since gene expression changes with cell composition, and cell composition can be altered, for example, depending on collection time, collection conditions, and the disease state.

Complete blood counts

In the Rotterdam Study, the complete blood counts (CBCs) have been measured in the samples. The CBCs give important information about the kinds and numbers of blood cells, especially the number of white blood cells (WBC): the monocytes, the lymphocytes, and the granulocytes; the number of red blood cells (RBC): the erythrocytes; and the number of platelets: the thrombocytes (Figure 2). Normal ranges of a complete blood count for healthy males and females are given in Table 4.

Table 4. Normal ranges of complete blood counts for healthy adults. *http://www.mercynorthiowa.com/cbc-normal-ranges*

	Normal cell counts per microliter	
Type of blood cell	**Males**	**Females**
Erythrocytes (RBC)	4.3-5.8 million	3.8-5.2 million
Monocytes (WBC)	200-900	200-900
Lymphocytes (WBC)	800-4,800	800-4,800
Granulocytes (WBC)	0-30	0-30
Thrombocytes (platelets)	150,000-440,000	150,000-440,000

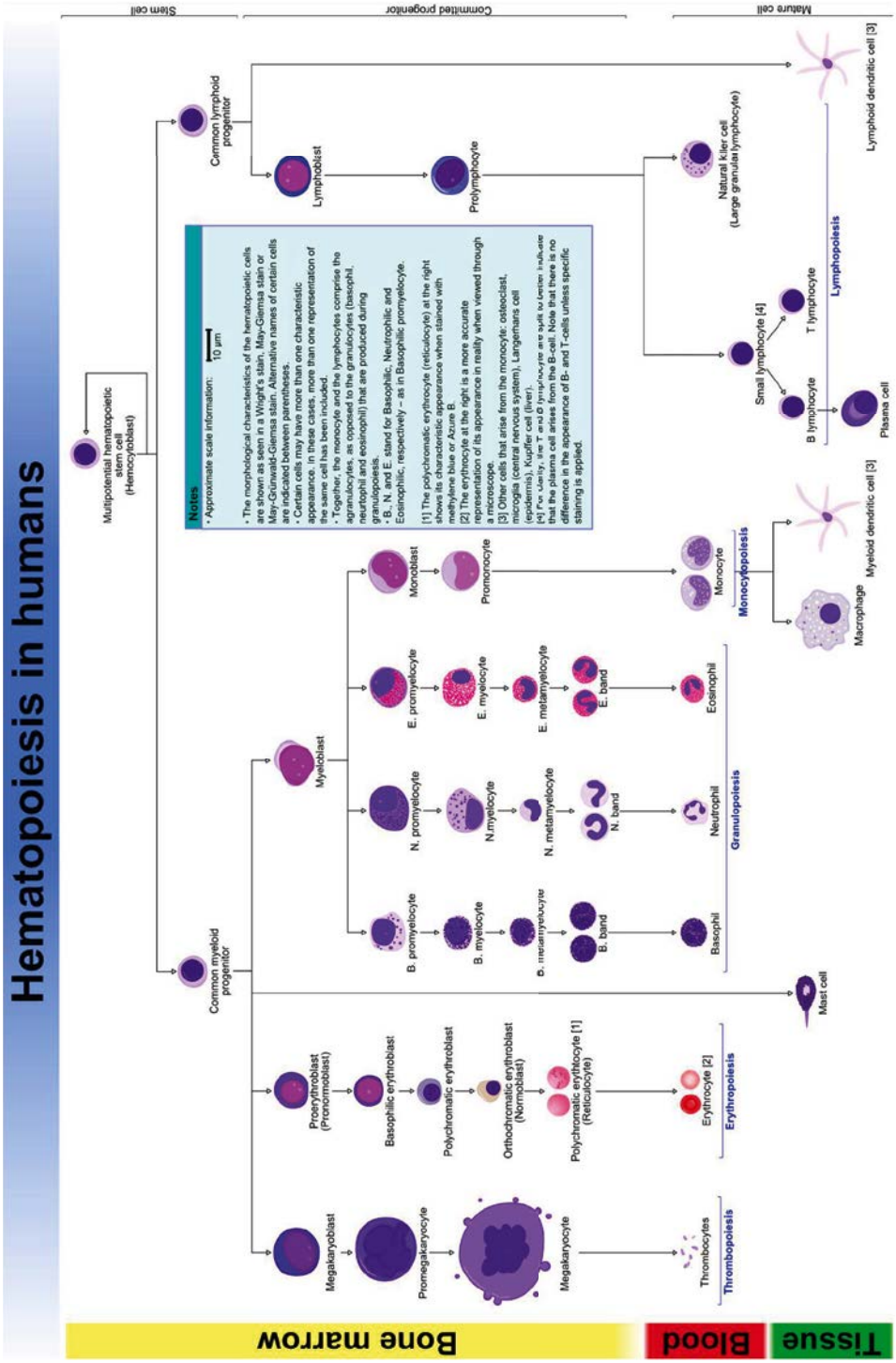


Figure 2. The development of different blood cells from hematopoietic stem cell to mature cells *(image from A. Rad via Wikimedia Commons)*

We used the CBCs to correct for the major differences in cell composition. Of course, these cell subset classifications are also artificial: they reflect the current ability to distinguish the cells based on specific small sets of available markers. It is becoming clear that each defined subpopulation of cells can be further broken down into additional subgroups, as the tools for subset classification become more sophisticated [11].

Houseman estimates

Houseman *et al.* [12] developed an algorithm to predict the blood cell type numbers based on 100 CpG methylation sites present on the Illumina 450K methylation array. The algorithm estimates the relative number of CD4+ T-cells, CD8+ T-cells, NK cells, monocytes, granulocytes, using CpG sites which are known to be differentially methylated in the different blood cell types. The algorithm was validated by predicting the blood cell type profiles of sorted white blood cell samples. The main disadvantage of this method is that the algorithm was validated in only 46 sub-cell type samples (Table 5). Additionally, in our study we identified that the correlation between the Houseman estimates and the measured CBCs is not perfect, in fact not perfect at all: the r^2 values ranged between 0.27 and 0.57 (in 726 samples of the Rotterdam Study) (Figure 3). These weak correlations are alarming results for all transcriptomic and epigenetic analyses adjusting for the Houseman estimates, since they can introduce strong biases in the reported associations.

Table 5. The number of validation samples per blood cell type, used by Houseman.

Sorted white blood cell type	# of validation samples
B-cells	6
Granulocytes	8
Monocytes	5
NK cells	11
T cells (CD4+)	8
T cells (CD8+)	5
T cells (NKT)	1
T cells (other)	5
SUM	**46**

Variations in cell composition reflecting biological differences

The possible effects of different subpopulations of cells on the phenotype should be kept in mind while interpreting the association results. Nevertheless, the differences between the cell compositions associated with disease could be very interesting too. For example, with ageing it is known that the relative abundance of immune cells in blood is shifting: the amount of naïve T-cells is decreasing and the amount of highly differentiated effector and memory T-cells is increasing with age [13-18]. This is called immunosenescence and reflects the age-associated immune deficiency, which is found in long and short-lived species.

In our transcriptome-wide association analysis with age (chapter 2.2), we adjusted for the number of CBCs. However, the subpopulations of cell (e.g., the naïve versus differentiated T-cells) were not measured in the participating studies and could therefore not be adjusted for. Interestingly, we could replicate our findings in a number of studies with cell subset specific gene expression levels. For example, CD4+ cells, CD8+ cells, CD14+ cells (monocytes), and periperheral blood mononuclear cells (PBMCs). Replicating an age-associated gene in a sub cell type suggests that these genes are not solely accountable by cell count differences, but also reflect biological functions.

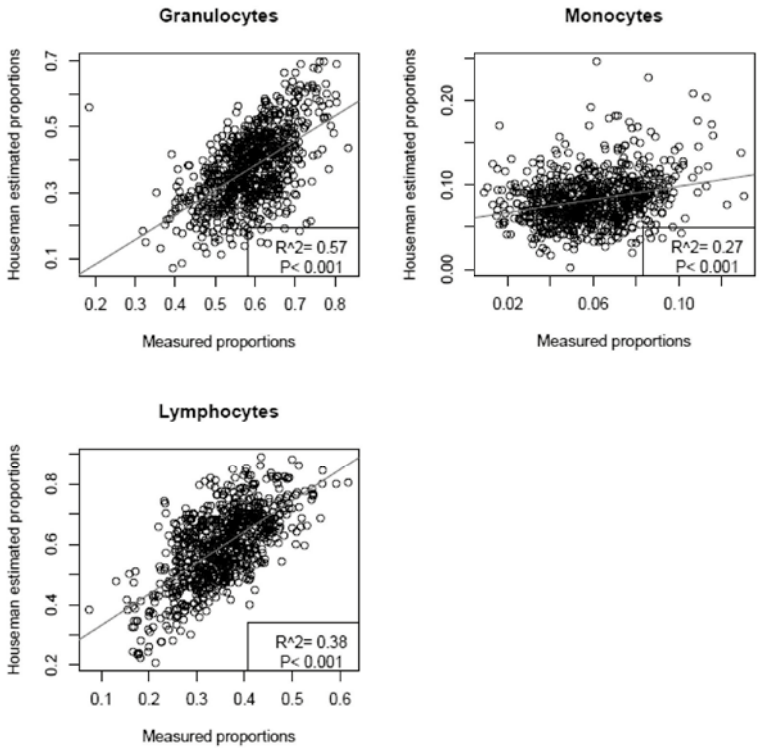


Figure 3. The correlation between the measured CBCs (x-axis) and the Houseman estimates (y-axis) for granulocytes, monocytes, and lymphocytes.

It is important to note that not only blood is a heterogeneous collection of different cell types: this is also true for other tissues, like brain cells, muscle cells, bone cells, or cartilage cells. Furthermore, these tissues are often contaminated with blood cells because of cell isolation difficulties. Therefore, the possible effects of different subpopulations of cells must be kept in mind while interpreting association results.

In this thesis, we identified 1,497 genes to be associated with ageing, of which 134 genes were known to belong to an immune function pathway (chapter 2.1). We identified 34 genes that are differentially expressed in relation to blood pressure (BP): 21 for systolic blood pressure, 20 for diastolic blood pressure, and 5 for hypertension (chapter 2.2). Next to inflammatory response, the BP genes also showed enrichment for metabolic processes and transcription regulation. For circulating lipid levels, we found 906 genes to be associated: 793 for triglycerides, 489 for HDL cholesterol, 5 for LDL cholesterol, and 20 for total cholesterol (chapter 2.3). Genes associated with triglycerides and HDL cholesterol were enriched for processes like protein binding, RNA binding, and ribosome integrity. LDL cholesterol genes were enriched for apolipoprotein receptor activity (initiating changes in cell activity), and the total cholesterol genes were highly enriched for the functions related to DNA and RNA binding. The muscle strength analysis identified 221 genes to be associated, which were enriched for hemoglobin metabolic process, hemolytic anemia, and innate immune response (chapter 2.4).

Across all four phenotypes (ageing, blood pressure, circulating lipid levels, and muscle strength), we identified 2,392 unique genes to be differentially expressed with at least one of the phenotypes. 226 genes (9.4%) were significantly associated with two or more phenotypes. One gene was identified to be associated with all four phenotypes: the *Perforin 1* (*PRF1*) gene. Its protein is important for T-cell and natural killer (NK) cell mediated destruction of cells. Gene set enrichment analysis of the 226 genes (using WebGestalt) identified enrichment for immune system diseases (24 genes, P-value=3.17E-5), inflammation (20 genes, P-value=3.17E-5), rheumatoid arthritis (14 genes, P-value=4.3E-5), and connective tissue disease (n=15, P=value=6.7E-5). Enriched molecular functions (gene ontology terms) were ribosome structure (n=13, P-value=2.33E-5), protein binding (n=119, P-value=0.026), and RNA binding (n=24, P-value=0.026). Both the enriched molecular functions and diseases emphasize the association of immune activation and cell signaling pathways in whole blood studies.

For the eQTL analyses, we checked whether the identified eQTLs were driven by differences in age or blood cell-counts between individuals (Chapter3.1). We selected 18 phenotypic measurements that were available in at least 1,500 samples (including age and CBCs), and we correlated the gene expression values with the phenotypic measurement using Spearman's rank correlation. After adjusting for the principal components (PCs) (one of the steps during the eQTL mapping), the correlations with age and CBCs were not significant anymore (p=0.31). This indicated that through removal of the PCs, we eliminated the differences in blood cell counts. However, we cannot exclude this possibility entirely because we did not conduct analyses on individual cell-types.

As discussed, analyzing genomic features in whole blood cells is more complex than interpreting cell-type specific experiments, but it is not necessarily less informative than the analysis of subpopulations. Of course, an analysis approach based on overall expression has greater power to detect expression changes in common blood cell types, and has less power to detect smaller changes in less numerous blood cell types. The cell subtype origins of the phenotype-associated

genes reported here now need to be identified by studying the associations in subtype specific cohorts.

MUSCULOSKELETAL COMORBIDITIES

Next to ageing and general age-related phenotypes, we also focused on musculoskeletal comorbidities which are the most frequent cause of physical activity limitations. We performed a GWAS for chronic widespread pain (CWP) (chapter 4.1) and a GWAS for heat pain thresholds (chapter 4.2), which is an experimental pain sensitivity marker. Many chronic pain syndromes are known to be associated with hypersensitivity to pain [19]. Additionally, we performed a transcriptomics study for gene expression levels in the circulation and joint effusion grades in the knee (chapter 4.3). Joint effusion is known to be related with joint inflammation [20], and is a strong predictor for development of incident knee OA [21].

In the GWAS meta-analysis for CWP, we found evidence for involvement of the 5p15.2 region containing two genes: *CCT5* and *FAM173B*. Although we have not conclusively identfied the causal gene in this area, there are some interesting findings which might implicate some candidates. For example, In the lumbar spinal cord of mouse models of inflammatory pain, chronic pain coincided with higher gene expression levels of both the *Cct5* gene and the *Fam173b* gene. Follow-up studies identified that the *Fam173b* reduction enhanced the resolution of inflammatory pain in this mouse model [22].

In the heat pain threshold GWAS meta-analysis, we found one deletion (1:176688345:D) to be significantly associated. The deletion is located in the twelfth intron of the *PAPPA2* gene in the 1q25.2 region. The *PAPPA2* gene is thought to be a regulator of insulin-like growth factor (*IGF*) bioavailability. IGF is important in the nociceptive pain sensitivity (processing harmful pain stimuli) of primary sensory neurons in the nerve cells. However, most interestingly we identified that having chronic pain substantially influences the heat pain threshold measurements: the genetic heritability in samples without pain (based on all genetic variants on the Illumina 550K SNP array) equals 32%, while samples having chronic pain have a heritability of only 9%. In addition, the heritability was higher in women compared to men. So, this suggests that having chronic pain and being male affects the pain sensitivity measurements with respect to the underlying genetic variants.

Both GWAS studies on pain described in this thesis were the first in its kind, and have been preceded by many small candidate gene studies reporting various "significant" associations [23]. However, in both GWAS studies, we could not replicate any of the previously reported SNPs associated with chronic pain or pain sensitivity, while our meta-analysis sample sizes were much larger than the discovery sample sizes. So, most likely the candidate gene studies were all false positives due to lack of power. Alternatively, one could speculate that very specific groups of pain patients were chosen for the candidate gene studies in which these SNPs have larger effects. Moreover, the lack of

reproducibility of SNPs in candidate genes in large GWAS meta-analyses has been shown before for many other phenotypes like BMD [24].

Last but not least, we tried to identify molecular biomarkers for early knee OA) by examining joint effusion grades in the knee and gene expression levels in the circulation. We identified 189 genes to be nominally associated with joint effusion in the knee, and several compelling genes were identified such as *C1orf38* and *NFATC1*. Significantly enriched biological pathways were: response to stress, gene expression, negative regulation of intracellular signal transduction, and antigen processing and presentation of exogenous pathways. Additional studies are needed to replicate the findings, which may result in biomarkers urgently needed to diagnose OA at an early stage.

CAUSAL DIRECTIONS

Because we studied genetics and genomics of age-related diseases in a cross-sectional design in the experiments described in this thesis, it is not possible to determine a causal direction for the associations reported. Therefore, the results of genetic and genomic studies need to be considered in light of longitudinal and functional evidence. Causality can be better understood by following the subjects in time and knowing the biological pathways involved. So, follow-up studies are needed to address causation.

For dynamic data (both gene expression and DNA methylation), a promising approach would be to infer causal networks from longitudinal data. After checking for co-regulated genes at one time point, the co-regulation can be measured at consecutive time points [25]. The continuity between the different time points allows studying the impact of major events. Because longitudinal data for both gene expression and DNA methylation data is now becoming available in the Rotterdam Study, this will hopefully provide more insight.

FUTURE DIRECTIONS

The main goal of population-based genetic and genomic studies in cohorts of elderly such as the Rotterdam Study is to identify new genes and biomarkers involved in age-related diseases, and thereby discovering new pathways to better understand the biological mechanisms of ageing.

In this thesis, we discovered SNPs and genes to be associated with ageing and age-related phenotypes. The sample sizes of the meta-analyses of GWAS data, gene expression levels, and DNA methylation levels continue to increase, thereby resulting in even more SNPs and genes associated with disease, but with increasingly smaller effect size. Like the GWAS meta-analyses done for height [26], waist-hip-ratio [27] and BMI [28] (in more than 250,000 individuals), studies including much larger sample sizes will generate even more robust associations and will reduce the noise in

the current analyses. And this will also be true for gene expression studies and DNA methylation studies: including much larger sample sizes will reduce the noise in the analyses, and will increase the number of significant findings. However, as shown in Figure 4, it is not only the sample size that determines the number of identified loci. Despite relatively large sample sizes for Alzheimer's disease and osteoarthritis (OA), only a modest number of loci were identified [29]. And the same is true for the GWAS studies for chronic widespread pain (CWP) and heat pain thresholds (HPT). In general, it is suspected that pain is an extremely heterogeneous phenotype.

For the identification of genome-wide *trans*-eQTLs, much larger sample sizes are needed too. For a genome-wide *trans*-eQTL analysis, at least 2.5 million SNPs times 60,000 gene expression probes need to be tested, resulting in 150 billion tests. The *trans*-eQTL meta-analysis described in chapter 3.1 is the largest meta-analysis done until now, including only 5,311 subjects having both genome-wide SNP data and gene expression data in blood. Therefore, the number of subjects in these analyses must increase dramatically to be able to perform the *trans*-eQTL analysis on all 2.5 million SNPs. For many other tissues, *trans*-eQTL studies have not been performed at all. This is a "big black hole", were additional research is needed to fill the enormous gap. *Trans*-eQTLs are very important to understand the downstream effect of disease variants, especially when the *trans*-eQTLs affect more genes in one pathway. The links between SNP and genes will give rise to new biological networks, which will definitely improve our understanding of complex diseases.

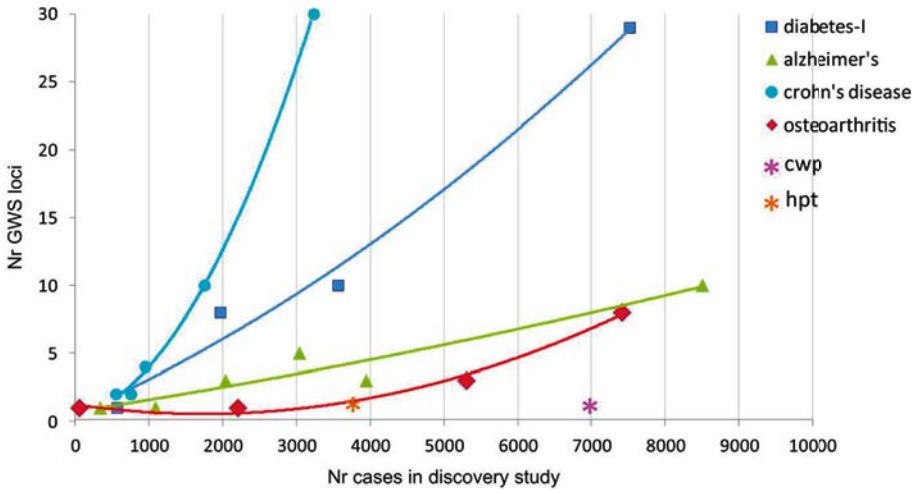


Figure 4. The power of genetic studies depends on the sample size and the phenotype (*image adapted from Van Meurs et al.* [29]). The number of GWAS hits (SNPs with a P-value<5E-08) as a function of sample size of the discovery cohort for several major complex diseases.

Because of the introduction of RNA-seq data, the current eQTL studies can now be extended with Allele Specific Expression (ASE) studies. At heterozygous positions in the genome, the percentage of reads having allele A or B can be determined. Over expression of one allele over the other (called ASE) has been observed, and can be assessed on a genome-wide scale. With the RNA-seq technology this is possible right now: because of including heterozygous genetic variants in the analysis, there are no inter-individual differences resulting in confounding factors. This is more powerful than a phenotype-based association analysis, and would be a nice addition to the current eQTL studies done.

It is anticipated that future studies will focus on more target tissues or specific cell types relevant for the disease phenotype under study. It is expected that genetic variants will have larger effects on gene expression or DNA methylation in the target tissue. For example, single cell RNA sequencing can be used to identify the gene expression profiles of one specific cell type of interest. With single cell sequencing, the gene expression levels of one specific cell are sequenced, meaning that there cannot be issues of heterogeneous cell populations. However, the selection of the single cells is still a critical step. Tissues consist of heterozygous cell population, so the selection procedure needs to be similar in all participants. Additionally, cells can be in different developmental stages (cells divide, cells grow, cells diffentiate, cells die, *etc.*), so this should be taken into account as well. Otherwise, the problem of sequencing heterozygous cell populations will be replaced by the problem of sequencing different cell populations across individuals.

It would be very interesting to get access to primary tissues, such as neuronal, bone and cartilage tissues. This will give a better insight into the molecular mechanisms underlying osteoarthritis, chronic pain, and pain sensitization. Because this will not be possible in population based cohort studies, collaborations with neurology and orthopedics departments are needed. A nice example of a project using primary tissues is the Genotype-Tissue Expression Project (GTEx): it is a pioneering project obtaining and storing a large range of organs and tissues, and testing them in the lab. These tissues were collected and stored through the National Cancer Institute's cancer Human Biobank initiative, and they created a resource that researchers can use to study how inherited changes in genes lead to common diseases [30].

Another option would be focusing on induced Pluripotent Stem (iPS) cells. iPS cells can be generated from adult cells from skin biopsies or blood cells (which can be done in population based cohort study) and upon proper stimulation they can give rise to other cell types in the body [31]. Highly differentiated somatic cells can be reprogrammed to a pluripotent stage, and these can be differentiated again. With this technique one can generate patient specific iPS cells (carrying the disease and a certain genetic background as assessed by GWAS or NGS).

Additionally, the annotation of non-coding transcripts needs to be improved. In this thesis, we focused on *trans*-eQTL effects for SNPs in long non-coding RNAs (lncRNAs), and in this analysis, we had many difficulties with the alignment of sequencing reads to the correct genes. This was caused

by all kind of transcripts that are very similar to the protein-coding genes, most of them being non-functional pseudogenes. More research is needed to improve this annotation. Next to a better annotation, it would be very interesting to measure the levels of the non-coding RNAs themselves. With the expression levels of lncRNAs, it is possible to examine whether the SNP is really affecting the non-coding RNA itself. Of course, these experiments need to be done in the target tissue.

As discussed before, it would be interesting to implement new DNA methylation measurement techniques like whole genome bisulfite sequencing (WBGS), methylC-capture (MCC) sequencing, or the *Illumina MethylationEPIC Beadchips*. These methods are not or at least less biased towards the CpG islands and gene promoters. What method to use depends on the financial circumstances and the aim of the project. When calculating costs for each method, the analysis time should be taken into account: sequencing methods generate a huge amount of data, but data storage and analysis are thus more complex. Sequence data have to be analyzed through sophisticated and computationally intensive algorithms, while array data can be easily analyzed with freely available analysis packages in R.

For the musculoskeletal comorbidities studied, less heterogeneous phenotype definitions need to be defined. The use of quantitative and possibly more objective pain measurements will be of pivotal importance to identify further genetic variants underlying chronic pain. With the heat pain thresholds GWAS, we used a quantitative measurement for pain sensitivity, but still similarly sized replication cohorts are needed that can demonstrate that the identified signals are true signals. For the joint effusion study, we experienced similar difficulties: only GARP and the Rotterdam Study have both knee MRIs and transcriptome-wide gene expression levels in blood available currently.

Our results contributed to a better understanding of the molecular mechanisms of ageing and age-related comorbidities and the inter-individual differences in these mechanisms in the general population. Our lists of genes provide a rich trove of data for future studies, and may serve as a roadmap aimed at translating findings into treatment strategies and interventions.

IS AGEING A DISEASE?

The final question remains whether ageing is programmed by biological clocks operating throughout lifespan, or whether ageing is based on accumulation of damage, or a mixture of both mechanisms? The studies in this thesis showed that transcriptome-wide gene expression levels measured in human blood cells are strongly associated with age, blood pressure, lipid levels, and muscle strength, and that DNA methylation changes in the functional regions of the DNA (enhancer regions) co-occur with these gene expression changes with age. Such enhancer regions could act as dimmer switches regulating the expression levels of nearby or distant genes, and the DNA methylation can be influenced by environmental factors (for example, changes in diet, exercise, smoking, etc.). These findings could offer very nice strategies for interventions (changing to a healthier diet, exercise more

regularly, and quit smoking), when we are sure that DNA methylation changes causally alter the gene expression levels. Earlier family studies showed that the capacity to survive to high ages is heritable: children of 90- and 100-year-olds have lower occurrences of metabolic morbidity (e.g., cardiovascular disease, diabetes, hypertension, even osteoarthritis) [32]. Additionally, their insulin sensitivity and lipid levels resemble those of the more healthy section of the population [33]. So, they are biologically younger than their spouses (age-matched unrelated individuals), and get significantly older too [34]. In summary, the theory of ageing being programmed by biological clocks feels like the preferred hypothesis: by inheriting the best genetic variants of your parents and optimally tuning the environmental factors (a healthy life style, good food patterns), one might be able to influence the pace of ageing.

REFERENCES

1. http://www.ukbiobank.ac.uk/.
2. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X (2014) Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One 9: e78644.
3. Zhang W, Spector TD, Deloukas P, Bell JT, Engelhardt BE (2015) Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. Genome Biol 16: 14.
4. Zhou X, Maricque B, Xie M, Li D, Sundaram V, et al. (2011) The Human Epigenome Browser at Washington University. Nat Methods 8: 989-990.
5. Edgar R, Tan PP, Portales-Casamar E, Pavlidis P (2014) Meta-analysis of human methylomes reveals stably methylated sequences surrounding CpG islands associated with high gene expression. Epigenetics Chromatin 7: 28.
6. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, et al. (2015) Integrative analysis of 111 reference human epigenomes. Nature 518: 317-330.
7. Stevens M, Cheng JB, Li DF, Xie MC, Hong CB, et al. (2013) Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. Genome Res 23: 1541-1553.
8. Allum F, Shao X, Guenard F, Simon MM, Busche S, et al. (2015) Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. Nat Commun 6: 7211.
9. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, et al. (2012) Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat Genet 44: 502-510.
10. Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, et al. (2003) Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci U S A 100: 1896-1901.
11. Thomas N, Heather J, Pollara G, Simpson N, Matjeka T, et al. (2012) The immune system as a biomonitor: explorations in innate and adaptive immunity. Interface Focus 3: 20120099.
12. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, et al. (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 13: 86.
13. Sallusto F, Lenig D, Forster R, Lipp M, Lanzavecchia A (1999) Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. Nature 401: 708-712.
14. Lee WW, Yang ZZ, Li G, Weyand CM, Goronzy JJ (2007) Unchecked CD70 expression on T cells lowers threshold for T cell activation in rheumatoid arthritis. J Immunol 179: 2609-2615.
15. Moro-Garcia MA, Alonso-Arias R, Lopez-Larrea C (2012) Molecular mechanisms involved in the aging of the T-cell immune response. Curr Genomics 13: 589-602.
16. Pletcher SD, Macdonald SJ, Marguerie R, Certa U, Stearns SC, et al. (2002) Genome-wide transcript profiles in aging and calorically restricted Drosophila melanogaster. Curr Biol 12: 712-723.
17. Landis GN, Abdueva D, Skvortsov D, Yang J, Rabin BE, et al. (2004) Similar gene expression patterns characterize aging and oxidative stress in Drosophila melanogaster. Proc Natl Acad Sci U S A 101: 7663-7668.
18. Rera M, Clark RI, Walker DW (2012) Intestinal barrier dysfunction links metabolic and inflammatory markers of aging to death in Drosophila. Proc Natl Acad Sci U S A 109: 21528-21533.
19. Edwards RR (2005) Individual differences in endogenous pain modulation as a risk factor for chronic pain. Neurology 65: 437-443.
20. Johnson MW (2000) Acute knee effusions: a systematic approach to diagnosis. Am Fam Physician 61: 2391-2400.
21. Atukorala I, Kwoh CK, Guermazi A, Roemer FW, Boudreau RM, et al. (2014) Synovitis in knee osteoarthritis: a precursor of disease? Ann Rheum Dis.
22. Willemen H, Kavelaars A, Cano RG, Heijnen C, Eijkelkamp N (2013) The Role Of FAM173b As a Newly Identified Regulator Of Chronic Pain. Arthritis Rheum: 2857.
23. Mogil JS (2012) Pain genetics: past, present and future. Trends Genet 28: 258-266.
24. Richards JB, Kavvoura FK, Rivadeneira F, Styrkarsdottir U, Estrada K, et al. (2009) Collaborative meta-analysis: associations of 150 candidate genes with osteoporosis and osteoporotic fracture. Ann Intern Med 151: 528-537.

25. Zhang K, Han J, Groesser T, Fontenay G, Parvin B (2012) Inference of causal networks from time-varying transcriptome data via sparse coding. PLoS One 7: e42306.
26. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, et al. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 46: 1173-1186.
27. Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, et al. (2015) New genetic loci link adipose and insulin biology to body fat distribution. Nature 518: 187-196.
28. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, et al. (2015) Genetic studies of body mass index yield new insights for obesity biology. Nature 518: 197-206.
29. van Meurs JB, Uitterlinden AG (2012) Osteoarthritis year 2012 in review: genetics and genomics. Osteoarthritis Cartilage 20: 1470-1476.
30. Consortium GT (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45: 580-585.
31. Inoue H, Nagata N, Kurokawa H, Yamanaka S (2014) iPS cells: a game changer for future medicine. EMBO J 33: 409-417.
32. Westendorp RGJ, van Heemst D, Rozing MP, Frolich M, Mooijaart SP, et al. (2009) Nonagenarian Siblings and Their Offspring Display Lower Risk of Mortality and Morbidity than Sporadic Nonagenarians: The Leiden Longevity Study. Journal of the American Geriatrics Society 57: 1634-1637.
33. Wijsman CA, Rozing MP, Streefland TCM, le Cessie S, Mooijaart SP, et al. (2011) Familial longevity is marked by enhanced insulin sensitivity. Aging Cell 10: 114-121.
34. Schoenmaker M, de Craen AJ, de Meijer PH, Beekman M, Blauw GJ, et al. (2006) Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. Eur J Hum Genet 14: 79-84.

CHAPTER 6

Summary
Samenvatting

Summary

Although age is the most powerful risk factor for many common diseases, the underlying molecular mechanisms are still largely unknown. The overall objective of this thesis was to integrate different genetic and genomic approaches to better understand ageing and age-related comorbidities. We studied the effects of genetic variants, gene expression levels, and DNA methylation levels on age-related phenotypes. Additionally, we combined the different levels of -omics data (genomics, epigenomics, and transcriptomics) to identify expression quantitative trait loci (eQTLs) and potentially functional CpG-methylation sites regulating gene expression levels.

In **Chapter 2**, the results of four transcriptome-wide gene expression meta-analyses are presented. In **Chapter 2.1**, we studied gene expression levels in whole blood of 14,983 persons, and we identified 1,497 genes that are differentially expressed with ageing. The age-associated genes were enriched for the presence of potentially functional DNA methylation sites (enhancer and insulator regions) that associated with both ageing and the gene expression levels of the genes located close by. Additionally, we used the gene expression levels to predict the "biological or transcriptomics age" of an individual, and we show that the delta age (the difference between the biological age and the chronological age) was associated with biological features like blood pressure, cholesterol levels, fasting glucose, and body mass index.
In **Chapter 2.2**, results of the transcriptome-wide association analysis for blood pressure are shown. We performed a meta-analysis in 7,017 persons that did not receive antihypertensive drugs, and identified 34 genes that were differentially expressed in relation to blood pressure. Only two genes (*FOS* and *PTGS2*) have been previously reported to be involved in blood pressure related processes. Gene set enrichment analyses suggested that the genes are involved in inflammatory response and the apoptosis pathways. The genetic variant rs3184504 in the gene *SH2B3* (which was also reported in a GWAS to be associated with blood pressure) was found to be a *trans*-regulator of six of the blood pressure associated genes (*FOS*, *MYADM*, *PP1R15A*, *TAGAP*, *S100A10*, and *FGBP2*).
In **Chapter 2.3**, we showed that gene expression levels are also associated with circulating lipid levels: triglycerides, HDL-cholesterol, LDL-cholesterol, and total cholesterol. In a meta-analysis of 4,841 fasting individuals that did not receive lipid-lowering medications, we identified 906 genes to be associated with levels of at least one lipid trait. We identified a set of basophil and mast cell-related genes whose expression levels were associated with all four lipid traits. The two genes with the smallest P-values were *HDC* and *CPA3*: expression levels of these two genes explained 23.1% and 18.1% of the total variation in log-transformed triglyceride levels, respectively. Furthermore, we identified significant associations between lipid levels and the expression levels of 95 known lipid-related genes.
In **Chapter 2.4**, we aimed to identify gene expression levels associated with muscle strength. In a meta-analysis of 7,781 persons with hand-grip strength measurements, we identified 221 genes to be associated. The associated genes were enriched for hemoglobin biosynthesis, innate immune activation, and stress response. 115 genes (52%) have not previously been linked to muscle in NCBI PubMed abstracts. Ten genes were only associated in younger individuals (<60 years), four genes

(*ASAP1*, *GID8*, *RAC1*, and *NDUFS1*) were only found in males, and one gene (*DEFA4*) was found in females only.

The relationship between genetic variants (SNPs) and gene expression levels is described in **Chapter 3**. In **Chapter 3.1**, we focused on disease-associated SNPs: we performed the largest eQTL meta-analysis in peripheral blood samples of 5,311 individuals, and replicated in another 2,775 persons. We identified and replicated *trans*-eQTLs for 233 SNPs (reflecting 103 independent loci), that were previously associated with complex traits. We identified trait-associated SNPs that affect multiple genes in *trans*: for example, the SNP rs4917014 (known to be associated with systemic lupus erythematosus or SLE) altered the expression levels of the *C1QB* gene and five type 1 interferon response genes, which are hallmarks of SLE. Variants associated with cholesterol metabolism and type 1 diabetes showed similar phenomena, indicating that large-scale eQTL mapping provides insight into the downstream effects of many trait-associated variants.
Chapter 3.2 describes a *trans*-eQTL analysis for SNPs in long non-coding RNAs (lncRNA). lncRNAs are thought to be important regulators for gene expression. In 652 persons with RNA sequencing data and replication in another 1,464 persons (with RNA sequencing data), we identified 2,678 *trans*-eQTLs for 1,320 SNPs in lncRNAs. Of these, 195 lncRNA *trans*-eQTLs showed inter-chromosomal effects (the SNP was located on a different chromosome than the gene). The 195 lncRNA *trans*-eQTLs are caused by 127 unique SNPs, and some affect multiple genes in *trans*: for example, SNP rs13227497 (a *cis*-eQTL SNP for lncRNA *RP11-611L7.1*) is associated with the expression of *KNS1*, *PI3*, and *ALDH1A2*. SNPs in these genes are known to be associated with severe hand osteoarthritis, response to stress, and chronic pain. The lncRNA might affect the three genes in one pathway, thereby influencing all three phenotypes simultaneously.

In **Chapter 4**, we integrated genetic and genomic studies for age-related musculoskeletal comorbidities. With ageing, musculoskeletal comorbidities have become the most frequent cause of physical activity limitations and reduced self-management behavior, thereby causing a reduced quality of life. **Chapter 4.1** describes the results of the first GWAS study for chronic widespread pain (CWP). CWP is a common disorder affecting about 10 percent of the general population. In the GWAS meta-analysis of 1,308 female CWP cases and 5,791 female controls, we identified two SNPs which were genome-wide significant (representing one locus) and another 39 SNPs with suggestive evidence for association (representing nine loci). After replication in another 1,480 female CWP cases and 7,989 female controls, we identified the minor C-allele of rs13361160 on chromosome 5p15.2, located upstream of *CCT5* and downstream of *FAM173B*, to be associated with a 30% higher risk of CWP. Expression levels of *Cct5* and *Fam173b* in mice with inflammatory-induced hyperalgesia were higher in the spinal cord, not in the dorsal root ganglions, compared to mice without pain. Both *CCT5* and *FAM173B* are novel genes that are associated with CWP but the underlying mechanisms involving pain sensitivity should be investigated in more detail.
In **Chapter 4.2** we describe a GWAS of the heat pain threshold (HPT). The HPT is an experimental pain sensitivity measurement: a thermo-sensory analyzer probe was placed on the inner site of the non-dominant forearm, and the participants were asked to stop the test at the moment the pain

stimulus (increase in temperature) started to feel unpleasant or painful. Our GWAS meta-analysis in 3,795 participants of the Rotterdam study revealed one genome-wide significant SNP rs192745611, which is located in the fifth exon of the *PAPPA2* gene, of which the encoded protein is thought to be a local regulator of insulin-like growth factor (*IGF*) bioavailability. *IGF* is implicated to play a role in the nociceptive (pain) sensitivity of primary afferent neurons. Neurotrophy, neurogenesis and metabolic functions are shown to be influenced by *IGF* in the adult brain, and in vitro, upregulation of *IGF* showed a higher sensitivity of primary afferent neurons.

By estimating the heritability of HPT, we found large differences between sexes and persons with or without chronic pain: the heritability estimate in females was 35%, compared to 9% in males; in persons without chronic pain the heritability estimate was 32%, compared to 8% in persons having chronic pain. Therefore, we adjusted the GWAS analysis for both sex and chronic pain status. Although very interesting, our GWAS findings need to be replicated in an independent cohort. This would give more insight into the underlying mechanisms of the HPT.

Chapter 4.3 describes a transcriptome-wide gene expression meta-analysis for joint effusion grades in the knee. Joint effusion is the presence of increased intra-articular fluid, which has been positively associated with knee pain in knee OA patients. Joint effusion is known to be related with joint inflammation and recent studies showed that the occurrence of joint effusion is a strong predictor for the development of incident knee OA. We aimed to identify molecular biomarkers for early knee OA, and included 135 females of the Rotterdam Study and 98 females of the GARP study. We identified 189 genes in blood that were nominally significantly associated with joint effusion, which were significantly enriched for response to stress, gene expression regulation, negative regulation of intracellular signal transduction, and antigen processing and presentation of exogenous pathways. The lowest P-value was found for the *C1orf38*, also called *THEMIS2* or *ICB-1*, which is highly expressed in several blood cells (monocytes, dendritic cells, NK cells, T-cells, B-cells). The gene is induced by interferon-gamma (*IFN*-γ), an important cytokine that orchestrates many distinct cellular processes regarding inflammation.

Chapter 5 provides a general discussion on the results of the studies in a broader perspective. Additionally, several recommendations for future research are presented.

Samenvatting

Ouderdom komt met gebreken. Maar eigenlijk begrijpen we nog altijd niet zo goed wat er nu precies gebeurt in het menselijk lichaam. Het doel van dit proefschrift is veroudering en ouderdomsgerelateerde ziektes beter te begrijpen door middel van integratie van genetische en genomische studies. We hebben onderzoek gedaan naar genetische variaties (SNPs), genexpressie levels en DNA methylering. Bovendien hebben we deze verschillende soorten data met elkaar gecombineerd: we hebben onderzocht welke SNPs genexpressie reguleren (dit noemt men eQTLs) en we hebben gekeken of genen die geassocieerd zijn met veroudering tegelijk verrijkt zijn met mogelijk functionele CpG methyleringssites.

In **Hoofdstuk 2** worden de resultaten van vier verschillende genexpressie studies beschreven. In **Hoofdstuk 2.1** onderzochten we de genexpressie levels in het bloed van 14.983 mensen. We hebben 1.497 genen gevonden die op de een of andere manier betrokken zijn bij het verouderingsproces. Een deel van deze genen is betrokken bij de energiehuishouding, de vetverbranding en de stevigheid en flexibiliteit van cellen. Vervolgens hebben we gekeken naar de DNA methylering (epigenetische markers) van deze genen: we zien dat de verouderingsgenen verrijkt zijn met methyleringssites in mogelijk functionele gebieden (ook wel enhancer en insulator gebieden genoemd). Deze methyleringssites zijn zowel met leeftijd als met de genexpressie levels (van de genen waarin ze liggen) geassocieerd. Tot slot hebben we de genexpressielevels gebruikt om voor ieder individu de "biologische leeftijd" te voorspellen. We laten zien dat het verschil tussen de "biologische leeftijd" en de "echte" of chronologische leeftijd geassocieerd is met bekende risicofactoren, zoals bloeddruk, cholesterol levels, bloedsuikerspiegel en de body mass index (BMI).
In **Hoofdstuk 2.2** hebben we de genexpressie levels in het bloed geassocieerd met de bloeddruk. In een meta-analyse van 7.017 mensen die geen bloeddruk verlagende medicijnen gebruikten hebben we 34 bloeddruk genen blootgelegd. Deze genen zijn geassocieerd met de bloeddruk, omdat ze met een verhoogde (of juist verlaagde) bloeddruk een ander signaal gaan afgeven. Van slechts twee genen was (wetenschappelijk) al bekend dat ze een relatie hebben met bloeddruk: dit zijn *FOS* en *PTGS2*. De totale lijst van 34 genen is belangrijk voor de afweer bij ontstekingen en voor apoptose (de afbraak van cellen). De SNP rs3184504 in het gen *SH2B3* (deze is gevonden met een genoomwijde associatie studie (GWAS) naar bloeddruk) lijkt 6 andere bloeddruk gerelateerde genen te reguleren (*FOS*, *MYADM*, *PP1R15A*, *TAGAP*, *S100A10* en *FGBP2*.
In **Hoofdstuk 2.3** worden de genen beschreven die geassocieerd zijn met circulerende lipiden: een lipiden profiel meting bestaat uit triglyceriden, HDL-cholesterol, LDL-cholesterol en totaal cholesterol. In een meta-analyse van 4.841 nuchtere deelnemers die geen lipiden-verlagende medicijnen gebruikten, hebben we 906 genen gevonden die geassocieerd zijn met tenminste één type lipide. Ook vonden we een select groepje genen dat geassocieerd was met alle vier de lipides: deze genen zijn belangrijk voor mastocyten en basofiele granulocyten. Mastocyten en basofiele granulocyten lijken sterk op elkaar: ze spelen een belangrijke rol in de immuunrespons en bij allergische reacties. De expressielevels van de meest significante genen (*HDC* en *CPA3*) verklaren

zo'n 23,1% en 18,1% van de totale variatie in trygliceriden levels. Bovendien konden we voor 96 bekende lipide genen de associatie met circulerende lipiden bevestigen.
In **Hoofdstuk 2.4** hebben we gekeken of genexpressie levels ook associëren met spierkracht. In een meta-analyse van 7.781 mensen met spierkracht-metingen in de hand, hebben we 221 genen in het bloed gevonden die geassocieerd zijn met spierkracht. Deze genen zijn belangrijk voor de aanmaak van hemoglobine, aspecifieke afweer in het immuunsysteem en de lichamelijke reactie op stress. 115 genen (52%) zijn niet eerder gevonden in relatie tot spier of spierkracht (als je zoekt in NCBI PubMed abstracts). Tien genen werden enkel gevonden in analyses specifiek voor jonge mensen (< 60 jaar), vier genen (*ASAP1*, *GID8*, *RAC1* en *NDUFS1*) werden alleen gevonden in mannen en één gen (*DEFA4*) werd enkel gevonden in vrouwen.

In **Hoofdstuk 3** werd de relatie tussen genetische varianten (SNPs) en genexpressie levels onderzocht. In **Hoofdstuk 3.1** is specifiek gekeken naar ziekte-gerelateerde SNPs: we hebben een zogenaamde "expressie quantitative trait locus" (eQTL) meta-analyse gedaan van 5.311 mensen en hebben de resultaten gerepliceerd in 2.775 mensen. We hebben *trans*-eQTLs gevonden voor 233 SNPs in 103 onafhankelijke gebieden. We hebben SNPs gevonden die meerdere genen reguleren in *trans*: bijvoorbeeld de SNP rs4917014 (een bekende SNP voor de auto-immuunziekte Lupus erythematosus of SLE): deze SNP verandert niet alleen de expressie levels van het *IKZF1* gen dichtbij (*cis*-effect), maar verandert eveneens de expressie levels van het *C1QB* gen en 5 verschillende type 1 interferon genen. Zowel het *C1QB* gen als de type 1 interferon genen zijn reeds bekend voor de ziekte SLE. Andere fenotypes (zoals het cholesterol metabolisme en type 1 diabetes) lieten vergelijkbare resultaten zien. Dit laat zien dat grote eQTL meta-analyses meer inzicht kunnen geven in de indirecte effecten van de genetische variaties.
Hoofdstuk 3.2 beschrijft een *trans*-eQTL studie specifiek voor SNPs gelegen in "long non-coding" RNAs (lncRNAs). lncRNAs heten non-coding RNAs omdat ze niet coderen voor eiwitten. Toch lijken ze belangrijk te zijn voor de regulatie van genexpressie. Daarom hebben we een *trans*-eQTL analyse gedaan in 652 mensen en hebben we de resultaten gerepliceerd in 1.464 onafhankelijke mensen. We hebben 2.678 *trans*-eQTLs gevonden voor 1.320 SNPs in lncRNAs. Slechs 195 *trans*-eQTLs waren échte *trans*-effecten, waarbij de SNP op een ander chromosoom ligt dan het gen. De *trans*-eQTLs werden veroorzaakt door 127 unieke SNPs, waarvan sommige SNPs meerdere genen reguleren. Een mooi voorbeeld is SNP rs13227497: deze SNP ligt vlakbij de lncRNA *RP11-611L7.1* en reguleert deze lncRNA in *cis*. Daarnaast reguleert de SNP ook de genexpressie levels van *KNS1*, *PI3* en *ALDH1A2* in *trans*. SNPs in deze *trans*-genen zijn al eerder geassocieerd met hand osteoarthrose (hand OA), de lichameljke reactie op stress en chronische pijn. Het lijkt er nu dus of dat de SNP rs13227497 de lncRNA reguleert en dat deze lncRNA de drie genen *KNS1*, *PI3* en *ALDH1A2* vervolgens reguleert in één netwerk. Dit moet verder onderzocht worden.

In **Hoofdstuk 4** integreren we de genetische en genomische studies met verouderingsgerelateerde musculoskeletale klachten en comorbiditeiten. Musculoskeletale aandoeningen, zoals OA en chronische gewrichtspijn, zijn de belangrijkste oorzaken van fysieke achteruitgang. Dit leidt vaak tot beperkingen in de mobiliteit, wat vervolgens weer kan leiden tot een verminderde kwaliteit van

leven. In **Hoofdstuk 4.1** beschrijven we de resultaten van de eerste GWAS studie voor algehele pijn, of in het Engels "chonic widespread pain" (CWP). Patiënten met CWP komen veel voor (ongeveer 10% van de populatie) en alleen pijnstillers nemen is voor deze mensen eigenlijk geen optie. In de GWAS meta-analyse van 1.308 vrouwen met CWP en 5.791 vrouwelijke controles hebben we twee SNPs (in één locus) gevonden die significant geassocieerd zijn met de ziekte, plus nog 39 SNPs (in negen onafhankelijke gebieden) met een p<1E-6. Na replicatie in nog 1.480 vrouwen met CWP en 7.989 controles repliceerde alleen de top SNP (rs13361160): het C-allele van deze SNP op chromosoom 5p15.2 is geassocieerd met 30% groter risico voor het hebben van CWP. De SNP ligt voor het *CCT5* gen en na het *FAM173B* gen. In muizen met chronische pijn hebben beide genen verhoogde genexpressie levels in vergelijking met muizen zonder pijn. Dit is een interessante bevinding en kan klinisch zeer interessant zijn om nieuwe aanknopingspunten te vinden om chronische pijn te bestrijden.

In **Hoofdstuk 4.2** hebben we een GWAS naar hitte pijn drempels (HPT) gedaan. De HPT is een experimentele meting van de pijngevoeligheid: er wordt een blokje op de binnenkant van de niet-dominante onderarm geplaatst wat kan variëren in temperatuur. De deelnemer wordt gevraagd de test te stoppen wanneer de pijn stimulus (de toename in temperatuur van het blokje) onaangenaam of pijnlijk begint te voelen. Onze GWAS meta-analyse van 3.795 mensen in de Rotterdam Studie geeft slechts één significante SNP: rs192745611. Deze SNP ligt in het vijfde exon van het *PAPPA2* gen. Het *PAPPA2* eiwit lijkt een belangrijke lokale regulator van het insuline-gelijkende groeifactor (*IGF*) te zijn. *IGF* speelt een belangrijk rol in de pijngevoeligheid van de sensorische zenuwcellen. In het humane brein beïnvloedt *IGF* de aanmaak van nieuwe zenuwcellen, het onderhoud van zenuwcellen en de algehele stofwisseling. En, in vitro, zorgen verhoogde *IGF* levels voor een verhoogde gevoeligheid van de zenuwcellen. We vinden grote verschillen in de schattingen voor het percentage van HPT variantie wat verklaard kan worden door genetische variaties: in vrouwen schatten we dat 35% van de HPT variantie verklaard wordt door SNPs, terwijl dit percentage in mannen slechts 9% is. Ook in mensen met chronische pijn ligt het percentage veel lager (slechts 8%) dan in mensen zonder chronische pijn (32%). De GWAS analyse is daarom geadjusteerd voor zowel geslacht als het hebben van chronische pijn. Omdat de GWAS in een relatief kleine populatie gemeten is, moeten we de resultaten eerst repliceren in een onafhankelijk cohort. Dit zal eventueel vals positieve bevindingen verwijderen en nu (net) niet significante SNPs significanter maken. Deze bevindingen kunnen meer inzicht geven in de onderliggende mechanismen van de pijngevoeligheid.

In **Hoofdstuk 4.3** hebben we gekeken of de genexpressie levels in bloed ook associëren met gewrichtseffusie scores in de knie. Gewrichteffusie is gerelateerd aan gewrichtsontsteking en recente studies hebben laten zien dat het hebben van gewrichteffusie een sterke predictor is voor het ontwikkelen van knie artrose (knie OA). Met deze studie hebben we geprobeerd moleculaire biomarkers in het bloed te vinden voor de vroege detectie van knie OA. Met 135 vrouwen van de Rotterdam Studie en 98 vrouwen van de GARP studie hebben we 189 genen gevonden waarvan de expressielevels nominaal geassocieerd zijn met de gewrichtseffusie scores in de knie. Deze genen zijn belangrijk voor de lichamelijke reactie op stress, de regulatie van genexpressie en signaaltransductie en de presentatie van antigenen. De kleinste p-waardes zijn gevonden voor het *C1orf38* gen, ook wel *THEMIS2* of *ICB-1* genoemd. Dit gen komt tot expressie in verschillende

soorten bloedcellen (monocyten, dendritische cellen, NK cellen, T-cellen, B-cellen). Het gen wordt geactiveerd door interferon-gamma (*IFN*-γ), een belangrijke cytokine die verschillende cellulaire processen bij een ontsteking reguleert.

In **Hoofdstuk 5** worden de resultaten van dit proefschrift bij elkaar gebracht en in een breder perspectief besproken. Vervolgens worden suggesties en ideëen voor vervolg onderzoek gepresenteerd.

CHAPTER 7

Bibliography
Authors and affiliations
PhD portfolio
About the author
Over de auteur
Dankwoord

Bibliography

In this thesis:

Peters MJ*, Joehanes R*, Pilling LC*, Schurmann C*, Conneely KN*, Powell J*, Reinmaa E*, Sutphin GL*, Zhernakova A*, Schramm K*, Wilson YA*, Kobes S, Tukiainen T; NABEC/UKBEC Consortium, Ramos YF, Göring HH, Fornage M, Liu Y, Gharib SA, Stranger BE, De Jager PL, Aviv A, Levy D, Murabito JM, Munson PJ, Huan T, Hofman A, Uitterlinden AG, Rivadeneira F, van Rooij J, Stolk L, Broer L, Verbiest MM, Jhamai M, Arp P, Metspalu A, Tserel L, Milani L, Samani NJ, Peterson P, Kasela S, Codd V,Peters A, Ward-Caviness CK, Herder C, Waldenberger M, Roden M, Singmann P, Zeilinger S, Illig T, Homuth G, Grabe HJ, Völzke H, Steil L, Kocher T, Murray A, Melzer D, Yaghootkar H, Bandinelli S, Moses EK, Kent JW, Curran JE, Johnson MP, Williams-Blangero S, Westra HJ, McRae AF, Smith JA, Kardia SL, Hovatta I, Perola M, Ripatti S, Salomaa V, Henders AK, Martin NG, Smith AK, Mehta D, Binder EB, Nylocks KM, Kennedy EM, Klengel T, Ding J, Suchy-Dicey AM, Enquobahrie DA, Brody J, Rotter JI, Chen YD, Houwing-Duistermaat J, Kloppenburg M, Slagboom PE, Helmer Q, den Hollander W, Bean S, Raj T, Bakhshi N, Wang QP, Oyston LJ, Psaty BM, Tracy RP, Montgomery GW, Turner ST, Blangero J, Meulenbelt I, Ressler KJ, Yang J*, Franke L*, Kettunen J*, Visscher PM*, Neely GG*, Korstanje R*, Hanson RL*, Prokisch H*, Ferrucci L*, Esko T*, Teumer A*, van Meurs JB*, Johnson AD*. The transcriptional landscape of age in human peripheral blood. *Nat Commun. 2015 Oct 22;6:8570. doi: 10.1038/ncomms9570*.

Huan T*, Esko T*, **Peters MJ***, Pilling LC*, Schramm K*, Schurmann C*, Chen BH, Liu C, Joehanes R, Johnson AD, Yao C, Ying SX, Courchesne P, Milani L, Raghavachari N, Wang R, Liu P, Reinmaa E, Dehghan A, Hofman A, Uitterlinden AG, Hernandez DG, Bandinelli S, Singleton A, Melzer D, Metspalu A, Carstensen M, Grallert H, Herder C, Meitinger T, Peters A, Roden M, Waldenberger M, Dörr M, Felix SB, Zeller T, Vasan R, O'Donnell CJ, Munson PJ, Yang X*, Prokisch H*, Völker U*, van Meurs JB*, Ferrucci L*, and Levy D*. A meta-analysis of gene expression signatures of blood pressure and hypertension. *PLoS Genet. 2015 Mar 18;11(3):e1005035, DOI: 10.1371/journal.pgen.1005035*.

Pilling LC*, Joehanes R*, Kacprowski T*, **Peters MJ***, Jansen R*, Karasik D, Kiel DP, Harries LW, Teumer A, Powell JE, Levy D, Lin H, Lunetta K, Munson PJ, Bandinelli S, Henley WE, Hernandez DG, Singleton AB, Tanaka T, van Grootheest G, Hofman A, Uitterlinden AG, Biffar R, Gläser S, Homuth G, Malsch C, Völker U, Penninx BW*, van Meurs JB*, Ferrucci L*, Kocher T*, Murabito JM*, Melzer D*. Gene transcripts associated with muscle strength: a CHARGE meta-analysis of 7,781 persons. *Physiol Genomics. 2015 Oct 20:physiolgenomics.00054.2015. doi: 10.1152/physiolgenomics.00054.2015*.

Westra HJ*, **Peters MJ***, Esko T*, Yaghootkar H*, Schurmann C*, Kettunen J*, Christiansen MW*, Fairfax BP, Schramm K, Powell JE, Zhernakova A, Zhernakova DV, Veldink JH, van den Berg LH, Karjalainen J, Withoff S, Uitterlinden AG, Hofman A, Rivadeneira F, 't Hoen PAC, Reinmaa E, Fischer K, Nelis M, Milani L, Melzer D, Ferrucci L, Singleton AB, Hernandez DG, Nalls MA, Homuth G, Nauck M, Radke D, Völker U, Perola M, Salomaa V, Brody J, Suchy-Dicey A, Gharib SA, Enquobahrie DA, Lumley T, Montgomery GW, Makino S, Prokisch H, Herder C, Roden M, Grallert H, Meitinger T, Strauch K, Li

Y, Jansen RC, Visscher PM, Knight JC, Psaty BM*, Ripatti S*, Teumer A*, Frayling TM*, Metspalu A*, van Meurs JBJ*, and Franke L*. Systematic identification of *trans*-eQTLs as putative drivers of known disease associations. *Nature Genetics 2013; 45(10): 1238-1243, DOI: 10.1038/ng.2756*.

Peters MJ*, Broer L*, Willemen HL*, Eiriksdottir G, Hocking LJ, Holliday KL, Horan MA, Meulenbelt I, Neogi T, Popham M, Schmidt CO, Soni A, Valdes AM, Amin N, Dennison EM, Eijkelkamp N, Harris TB, Hart DJ, Hofman A, Huygen FJ, Jameson KA, Jones GT, Launer LJ, Kerkhof HJ, de Kruijf M, McBeth J, Kloppenburg M, Ollier WE, Oostra B, Payton A, Rivadeneira F, Smith BH, Smith AV, Stolk L, Teumer A, Thomson W, Uitterlinden AG, Wang K, van Wingerden SH, Arden NK, Cooper C, Felson D, Gudnason V, Macfarlane GJ, Pendleton N, Slagboom PE, Spector TD, Völzke H, Kavelaars A*, van Duijn CM*, Williams FM*, and van Meurs JB*. Genome-wide association study meta-analysis of chronic widespread pain: evidence for involvement of the 5p15.2 region. *Annals of the rheumatic diseases, 2013, vol. 72, no. 3, pp. 427-436. DOI: 10.1136/annrheumdis-2012-201742*.

Other publications:

de Kruijf M, **Peters MJ**, C Jacobs L, Tiemeier H, Nijsten T, Hofman A, Uitterlinden AG, Huygen FJ, van Meurs JB. Determinants for Quantitative Sensory Testing and the Association with Chronic Musculoskeletal Pain in the General Elderly Population. *Pain practice: the official journal of World Institute of Pain, 2015, DOI: 10.1111/papr.12335*.

Westra HJ, Arends D, Esko T, **Peters MJ**, Schurmann C, Schramm K, Kettunen J, Yaghootkar H, Fairfax BP, Andiappan AK, Li Y, Fu J, Karjalainen J, Platteel M, Visschedijk M, Weersma RK, Kasela S, Milani L, Tserel L, Peterson P, Reinmaa E, Hofman A, Uitterlinden AG, Rivadeneira F, Homuth G, Petersmann A, Lorbeer R, Prokisch H, Meitinger T, Herder C, Roden M, Grallert H, Ripatti S, Perola M, Wood AR, Melzer D, Ferrucci L, Singleton AB, Hernandez DG, Knight JC, Melchiotti R, Lee B, Poidinger M, Zolezzi F, Larbi A, Wang de Y, van den Berg LH, Veldink JH, Rotzschke O, Makino S, Salomaa V, Strauch K, Völker U, van Meurs JB, Metspalu A, Wijmenga C, Jansen RC, Franke L. Cell Specific eQTL Analysis without Sorting Cells. *PLoS genetics, 2015, vol. 11, no. 5, pp. e1005223, DOI: 10.1371/journal.pgen.1005223*.

Medina-Gomez C, Felix JF, Estrada K, **Peters MJ**, Herrera L, Kruithof CJ, Duijts L, Hofman A, van Duijn CM, Uitterlinden AG, Jaddoe VW, and Rivadeneira F. Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study. *European journal of epidemiology, 2015, vol. 30, no. 4, pp. 317-330, DOI: 10.1007/s10654-015-9998-4*.

Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, Powell C, Vedantam S, Buchkovich ML, Yang J, Croteau-Chonka DC, Esko T, Fall T, Ferreira T, Gustafsson S, Kutalik Z, Luan J, Mägi R, Randall JC, Winkler TW, Wood AR, Workalemahu T, Faul JD, Smith JA, Hua Zhao J, Zhao W, Chen J, Fehrmann R, Hedman ÅK, Karjalainen J, Schmidt EM, Absher D, Amin N, Anderson D, Beekman M, Bolton JL, Bragg-Gresham JL, Buyske S, Demirkan A, Deng G, Ehret GB, Feenstra B, Feitosa MF, Fischer K, Goel A, Gong J, Jackson AU, Kanoni S, Kleber ME, Kristiansson K, Lim U, Lotay V, Mangino M, Mateo Leach

I, Medina-Gomez C, Medland SE, Nalls MA, Palmer CD, Pasko D, Pechlivanis S, **Peters MJ**, Prokopenko I, Shungin D, Stančáková A, Strawbridge RJ, Ju Sung Y, Tanaka T, Teumer A, Trompet S, van der Laan SW, van Setten J, Van Vliet-Ostaptchouk JV, Wang Z, Yengo L, Zhang W, Isaacs A, Albrecht E, Ärnlöv J, Arscott GM, Attwood AP, Bandinelli S, Barrett A, Bas IN, Bellis C, Bennett AJ, Berne C, Blagieva R, Blüher M, Böhringer S, Bonnycastle LL, Böttcher Y, Boyd HA, Bruinenberg M, Caspersen IH, Ida Chen YD, Clarke R, Daw EW, de Craen AJ, Delgado G, Dimitriou M, Doney AS, Eklund N, Estrada K, Eury E, Folkersen L, Fraser RM, Garcia ME, Geller F, Giedraitis V, Gigante B, Go AS, Golay A, Goodall AH, Gordon SD, Gorski M, Grabe HJ, Grallert H, Grammer TB, Gräßler J, Grönberg H, Groves CJ, Gusto G, Haessler J, Hall P, Haller T, Hallmans G, Hartman CA, Hassinen M, Hayward C, Heard-Costa NL, Helmer Q, Hengstenberg C, Holmen O, Hottenga JJ, James AL, Jeff JM, Johansson Å, Jolley J, Juliusdottir T, Kinnunen L, Koenig W, Koskenvuo M, Kratzer W, Laitinen J, Lamina C, Leander K, Lee NR, Lichtner P, Lind L, Lindström J, Sin Lo K, Lobbens S, Lorbeer R, Lu Y, Mach F, Magnusson PK, Mahajan A, McArdle WL, McLachlan S, Menni C, Merger S, Mihailov E, Milani L, Moayyeri A, Monda KL, Morken MA, Mulas A, Müller G, Müller-Nurasyid M, Musk AW, Nagaraja R, Nöthen MM, Nolte IM, Pilz S, Rayner NW, Renstrom F, Rettig R, Ried JS, Ripke S, Robertson NR, Rose LM, Sanna S, Scharnagl H, Scholtens S, Schumacher FR, Scott WR, Seufferlein T, Shi J, Vernon Smith A, Smolonska J, Stanton AV, Steinthorsdottir V, Stirrups K, Stringham HM, Sundström J, Swertz MA, Swift AJ, Syvänen AC, Tan ST, Tayo BO, Thorand B, Thorleifsson G, Tyrer JP, Uh HW, Vandenput L, Verhulst FC, Vermeulen SH, Verweij N, Vonk JM, Waite LL, Warren HR, Waterworth D, Weedon MN, Wilkens LR, Willenborg C, Wilsgaard T, Wojczynski MK, Wong A, Wright AF, Zhang Q, LifeLines Cohort Study, Brennan EP, Choi M, Dastani Z, Drong AW, Eriksson P, Franco-Cereceda A, Gådin JR, Gharavi AG, Goddard ME, Handsaker RE, Huang J, Karpe F, Kathiresan S, Keildson S, Kiryluk K, Kubo M, Lee JY, Liang L, Lifton RP, Ma B, McCarroll SA, McKnight AJ, Min JL, Moffatt MF, Montgomery GW, Murabito JM, Nicholson G, Nyholt DR, Okada Y, Perry JR, Dorajoo R, Reinmaa E, Salem RM, Sandholm N, Scott RA, Stolk L, Takahashi A, Tanaka T, Van't Hooft FM, Vinkhuyzen AA, Westra HJ, Zheng W, Zondervan KT, ADIPOGen Consortium, AGEN-BMI Working Group, CARDIOGRAMplusC4D Consortium, CKDGen Consortium, GLGC, ICBP, MAGIC Investigators, MuTHER Consortium, MIGen Consortium, PAGE Consortium, ReproGen Consortium, GENIE Consortium, International Endogene Consortium, Heath AC, Arveiler D, Bakker SJ, Beilby J, Bergman RN, Blangero J, Bovet P, Campbell H, Caulfield MJ, Cesana G, Chakravarti A, Chasman DI, Chines PS, Collins FS, Crawford DC, Cupples LA, Cusi D, Danesh J, de Faire U, den Ruijter HM, Dominiczak AF, Erbel R, Erdmann J, Eriksson JG, Farrall M, Felix SB, Ferrannini E, Ferrières J, Ford I, Forouhi NG, Forrester T, Franco OH, Gansevoort RT, Gejman PV, Gieger C, Gottesman O, Gudnason V, Gyllensten U, Hall AS, Harris TB, Hattersley AT, Hicks AA, Hindorff LA, Hingorani AD, Hofman A, Homuth G, Hovingh GK, Humphries SE, Hunt SC, Hyppönen E, Illig T, Jacobs KB, Jarvelin MR, Jöckel KH, Johansen B, Jousilahti P, Jukema JW, Jula AM, Kaprio J, Kastelein JJ, Keinanen-Kiukaanniemi SM, Kiemeney LA, Knekt P, Kooner JS, Kooperberg C, Kovacs P, Kraja AT, Kumari M, Kuusisto J, Lakka TA, Langenberg C, Le Marchand L, Lehtimäki T, Lyssenko V, Männistö S, Marette A, Matise TC, McKenzie CA, McKnight B, Moll FL, Morris AD, Morris AP, Murray JC, Nelis M, Ohlsson C, Oldehinkel AJ, Ong KK, Madden PA, Pasterkamp G, Peden JF, Peters A, Postma DS, Pramstaller PP, Price JF, Qi L, Raitakari OT, Rankinen T, Rao DC, Rice TK, Ridker PM, Rioux JD, Ritchie MD, Rudan I, Salomaa V, Samani NJ, Saramies J, Sarzynski MA, Schunkert H, Schwarz PE, Sever P, Shuldiner AR, Sinisalo J, Stolk RP, Strauch

K, Tönjes A, Trégouët DA, Tremblay A, Tremoli E, Virtamo J, Vohl MC, Völker U, Waeber G, Willemsen G, Witteman JC, Zillikens MC, Adair LS, Amouyel P, Asselbergs FW, Assimes TL, Bochud M, Boehm BO, Boerwinkle E, Bornstein SR, Bottinger EP, Bouchard C, Cauchi S, Chambers JC, Chanock SJ, Cooper RS, de Bakker PI, Dedoussis G, Ferrucci L, Franks PW, Froguel P, Groop LC, Haiman CA, Hamsten A, Hui J, Hunter DJ, Hveem K, Kaplan RC, Kivimaki M, Kuh D, Laakso M, Liu Y, Martin NG, März W, Melbye M, Metspalu A, Moebus S, Munroe PB, Njølstad I, Oostra BA, Palmer CN, Pedersen NL, Perola M, Pérusse L, Peters U, Power C, Quertermous T, Rauramaa R, Rivadeneira F, Saaristo TE, Saleheen D, Sattar N, Schadt EE, Schlessinger D, Slagboom PE, Snieder H, Spector TD, Thorsteinsdottir U, Stumvoll M, Tuomilehto J, Uitterlinden AG, Uusitupa M, van der Harst P, Walker M, Wallaschofski H, Wareham NJ, Watkins H, Weir DR, Wichmann HE, Wilson JF, Zanen P, Borecki IB, Deloukas P, Fox CS, Heid IM, O'Connell JR, Strachan DP, Stefansson K, van Duijn CM, Abecasis GR, Franke L, Frayling TM, McCarthy MI, Visscher PM, Scherag A, Willer CJ, Boehnke M, Mohlke KL, Lindgren CM, Beckmann JS, Barroso I, North KE, Ingelsson E, Hirschhorn JN, Loos RJ, and Speliotes EK. Genetic studies of body mass index yield new insights for obesity biology. *Nature, 2015, vol. 518, no. 7538, pp. 197-206, DOI: 10.1038/nature14177*.

Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, Mägi R, Strawbridge RJ, Pers TH, Fischer K, Justice AE, Workalemahu T, Wu JM, Buchkovich ML, Heard-Costa NL, Roman TS, Drong AW, Song C, Gustafsson S, Day FR, Esko T, Fall T, Kutalik Z, Luan J, Randall JC, Scherag A, Vedantam S, Wood AR, Chen J, Fehrmann R, Karjalainen J, Kahali B, Liu CT, Schmidt EM, Absher D, Amin N, Anderson D, Beekman M, Bragg-Gresham JL, Buyske S, Demirkan A, Ehret GB, Feitosa MF, Goel A, Jackson AU, Johnson T, Kleber ME, Kristiansson K, Mangino M, Mateo Leach I, Medina-Gomez C, Palmer CD, Pasko D, Pechlivanis S, **Peters MJ**, Prokopenko I, Stančáková A, Ju Sung Y, Tanaka T, Teumer A, Van Vliet-Ostaptchouk JV, Yengo L, Zhang W, Albrecht E, Ärnlöv J, Arscott GM, Bandinelli S, Barrett A, Bellis C, Bennett AJ, Berne C, Blüher M, Böhringer S, Bonnet F, Böttcher Y, Bruinenberg M, Carba DB, Caspersen IH, Clarke R, Daw EW, Deelen J, Deelman E, Delgado G, Doney AS, Eklund N, Erdos MR, Estrada K, Eury E, Friedrich N, Garcia ME, Giedraitis V, Gigante B, Go AS, Golay A, Grallert H, Grammer TB, Gräßler J, Grewal J, Groves CJ, Haller T, Hallmans G, Hartman CA, Hassinen M, Hayward C, Heikkilä K, Herzig KH, Helmer Q, Hillege HL, Holmen O, Hunt SC, Isaacs A, Ittermann T, James AL, Johansson I, Juliusdottir T, Kalafati IP, Kinnunen L, Koenig W, Kooner IK, Kratzer W, Lamina C, Leander K, Lee NR, Lichtner P, Lind L, Lindström J, Lobbens S, Lorentzon M, Mach F, Magnusson PK, Mahajan A, McArdle WL, Menni C, Merger S, Mihailov E, Milani L, Mills R, Moayyeri A, Monda KL, Mooijaart SP, Mühleisen TW, Mulas A, Müller G, Müller-Nurasyid M, Nagaraja R, Nalls MA, Narisu N, Glorioso N, Nolte IM, Olden M, Rayner NW, Renstrom F, Ried JS, Robertson NR, Rose LM, Sanna S, Scharnagl H, Scholtens S, Sennblad B, Seufferlein T, Sitlani CM, Vernon Smith A, Stirrups K, Stringham HM, Sundström J, Swertz MA, Swift AJ, Syvänen AC, Tayo BO, Thorand B, Thorleifsson G, Tomaschitz A, Troffa C, van Oort FV, Verweij N, Vonk JM, Waite LL, Wennauer R, Wilsgaard T, Wojczynski MK, Wong A, Zhang Q, Hua Zhao J, Brennan EP, Choi M, Eriksson P, Folkersen L, Franco-Cereceda A, Gharavi AG, Hedman ÅK, Hivert MF, Huang J, Kanoni S, Karpe F, Keildson S, Kiryluk K, Liang L, Lifton RP, Ma B, McKnight AJ, McPherson R, Metspalu A, Min JL, Moffatt MF, Montgomery GW, Murabito JM, Nicholson G, Nyholt DR, Olsson C, Perry JR, Reinmaa E, Salem RM, Sandholm N, Schadt EE, Scott RA, Stolk L, Vallejo EE, Westra HJ,

Zondervan KT, ADIPOGen Consortium, CARDIOGRAMplusC4D Consortium, CKDGen Consortium, GEFOS Consortium, GENIE Consortium, GLGC, ICBP, International Endogene Consortium, LifeLines Cohort Study, MAGIC Investigators, MuTHER Consortium, PAGE Consortium, ReproGen Consortium, Amouyel P, Arveiler D, Bakker SJ, Beilby J, Bergman RN, Blangero J, Brown MJ, Burnier M, Campbell H, Chakravarti A, Chines PS, Claudi-Boehm S, Collins FS, Crawford DC, Danesh J, de Faire U, de Geus EJ, Dörr M, Erbel R, Eriksson JG, Farrall M, Ferrannini E, Ferrières J, Forouhi NG, Forrester T, Franco OH, Gansevoort RT, Gieger C, Gudnason V, Haiman CA, Harris TB, Hattersley AT, Heliövaara M, Hicks AA, Hingorani AD, Hoffmann W, Hofman A, Homuth G, Humphries SE, Hyppönen E, Illig T, Jarvelin MR, Johansen B, Jousilahti P, Jula AM, Kaprio J, Kee F, Keinanen-Kiukaanniemi SM, Kooner JS, Kooperberg C, Kovacs P, Kraja AT, Kumari M, Kuulasmaa K, Kuusisto J, Lakka TA, Langenberg C, Le Marchand L, Lehtimäki T, Lyssenko V, Männistö S, Marette A, Matise TC, McKenzie CA, McKnight B, Musk AW, Möhlenkamp S, Morris AD, Nelis M, Ohlsson C, Oldehinkel AJ, Ong KK, Palmer LJ, Penninx BW, Peters A, Pramstaller PP, Raitakari OT, Rankinen T, Rao DC, Rice TK, Ridker PM, Ritchie MD, Rudan I, Salomaa V, Samani NJ, Saramies J, Sarzynski MA, Schwarz PE, Shuldiner AR, Staessen JA, Steinthorsdottir V, Stolk RP, Strauch K, Tönjes A, Tremblay A, Tremoli E, Vohl MC, Völker U, Vollenweider P, Wilson JF, Witteman JC, Adair LS, Bochud M, Boehm BO, Bornstein SR, Bouchard C, Cauchi S, Caulfield MJ, Chambers JC, Chasman DI, Cooper RS, Dedoussis G, Ferrucci L, Froguel P, Grabe HJ, Hamsten A, Hui J, Hveem K, Jöckel KH, Kivimaki M, Kuh D, Laakso M, Liu Y, März W, Munroe PB, Njølstad I, Oostra BA, Palmer CN, Pedersen NL, Perola M, Pérusse L, Peters U, Power C, Quertermous T, Rauramaa R, Rivadeneira F, Saaristo TE, Saleheen D, Sinisalo J, Slagboom PE, Snieder H, Spector TD, Thorsteinsdottir U, Stumvoll M, Tuomilehto J, Uitterlinden AG, Uusitupa M, van der Harst P, Veronesi G, Walker M, Wareham NJ, Watkins H, Wichmann HE, Abecasis GR, Assimes TL, Berndt SI, Boehnke M, Borecki IB, Deloukas P, Franke L, Frayling TM, Groop LC, Hunter DJ, Kaplan RC, O'Connell JR, Qi L, Schlessinger D, Strachan DP, Stefansson K, van Duijn CM, Willer CJ, Visscher PM, Yang J, Hirschhorn JN, Zillikens MC, McCarthy MI, Speliotes EK, North KE, Fox CS, Barroso I, Franks PW, Ingelsson E, Heid IM, Loos RJ, Cupples LA, Morris AP, Lindgren CM, and Mohlke KL. New genetic loci link adipose and insulin biology to body fat distribution. *Nature, 2015, vol. 518, no. 7538, pp. 187-196, DOI: 10.1038/nature14132*.

Wessel J, Chu AY, Willems SM, Wang S, Yaghootkar H, Brody JA, Dauriz M, Hivert MF, Raghavan S, Lipovich L, Hidalgo B, Fox K, Huffman JE, An P, Lu Y, Rasmussen-Torvik LJ, Grarup N, Ehm MG, Li L, Baldridge AS, Stančáková A, Abrol R, Besse C, Boland A, Bork-Jensen J, Fornage M, Freitag DF, Garcia ME, Guo X, Hara K, Isaacs A, Jakobsdottir J, Lange LA, Layton JC, Li M, Hua Zhao J, Meidtner K, Morrison AC, Nalls MA, **Peters MJ**, Sabater-Lleal M, Schurmann C, Silveira A, Smith AV, Southam L, Stoiber MH, Strawbridge RJ, Taylor KD, Varga TV, Allin KH, Amin N, Aponte JL, Aung T, Barbieri C, Bihlmeyer NA, Boehnke M, Bombieri C, Bowden DW, Burns SM, Chen Y, Chen YD, Cheng CY, Correa A, Czajkowski J, Dehghan A, Ehret GB, Eiriksdottir G, Escher SA, Farmaki AE, Frånberg M, Gambaro G, Giulianini F, Goddard WA, Goel A, Gottesman O, Grove ML, Gustafsson S, Hai Y, Hallmans G, Heo J, Hoffmann P, Ikram MK, Jensen RA, Jørgensen ME, Jørgensen T, Karaleftheri M, Khor CC, Kirkpatrick A, Kraja AT, Kuusisto J, Lange EM, Lee IT, Lee WJ, Leong A, Liao J, Liu C, Liu Y, Lindgren CM, Linneberg A, Malerba G, Mamakou V, Marouli E, Maruthur NM, Matchan A, McKean-Cowdin R, McLeod O, Metcalf GA, Mohlke KL, Muzny DM, Ntalla I, Palmer ND, Pasko D, Peter A, Rayner NW, Renström F, Rice K, Sala

CF, Sennblad B, Serafetinidis I, Smith JA, Soranzo N, Speliotes EK, Stahl EA, Stirrups K, Tentolouris N, Thanopoulou A, Torres M, Traglia M, Tsafantakis E, Javad S, Yanek LR, Zengini E, Becker DM, Bis JC, Brown JB, Adrienne Cupples L, Hansen T, Ingelsson E, Karter AJ, Lorenzo C, Mathias RA, Norris JM, Peloso GM, Sheu WH, Toniolo D, Vaidya D, Varma R, Wagenknecht LE, Boeing H, Bottinger EP, Dedoussis G, Deloukas P, Ferrannini E, Franco OH, Franks PW, Gibbs RA, Gudnason V, Hamsten A, Harris TB, Hattersley AT, Hayward C, Hofman A, Jansson JH, Langenberg C, Launer LJ, Levy D, Oostra BA, O'Donnell CJ, O'Rahilly S, Padmanabhan S, Pankow JS, Polasek O, Province MA, Rich SS, Ridker PM, Rudan I, Schulze MB, Smith BH, Uitterlinden AG, Walker M, Watkins H, Wong TY, Zeggini E, EPIC-InterAct Consortium, Laakso M, Borecki IB, Chasman DI, Pedersen O, Psaty BM, Shyong Tai E, van Duijn CM, Wareham NJ, Waterworth DM, Boerwinkle E, Linda Kao WH, Florez JC, Loos RJ, Wilson JG, Frayling TM, Siscovick DS, Dupuis J, Rotter JI, Meigs JB, Scott RA, Goodarzi MO, and EPIC-InterAct Consortium. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nature communications, 2015, vol. 6, pp. 5897, DOI: 10.1038/ncomms6897*.

Steenaard RV, Ligthart S, Stolk L, **Peters MJ**, van Meurs JB, Uitterlinden AG, Hofman A, Franco OH, and Dehghan A. Tobacco smoking is associated with methylation of genes related to coronary artery disease. *Clinical epigenetics, 2015, vol. 7, no. 1, pp. 54, DOI: 10.1186/s13148-015-0088-y*.

Ghanbari M, de Vries PS, de Looper H, **Peters MJ**, Schurmann C, Yaghootkar H, Dörr M, Frayling TM, Uitterlinden AG, Hofman A, van Meurs JBJ, Erkeland SJ, Franco OH, and Dehghan A. A Genetic Variant in the Seed Region of miR-4513 Shows Pleiotropic Effects on Lipid and Glucose Homeostasis, Blood Pressure, and Coronary Artery Disease. *Hum Mutat. 2014 Dec;35(12):1524-31, DOI: 10.1002/humu.22706*.

De Kruijf M, Kerkhof HJM, **Peters MJ**, Bierma-Zeinstra S, Hofman A, Uitterlinden AG, Huygen FJPM, and van Meurs JBJ. Finger length pattern as a biomarker for osteoarthritis and chronic joint pain: a population-based study and meta-analysis after systematic review. *Arthritis Care Res (Hoboken). 2014 Sep;66(9):1337-43, DOI: 10.1002/acr.22320*.

Loth DW, Artigas MS, Gharib SA, Wain LV, Franceschini N, Koch B, Pottinger TD, Smith AV, Duan Q, Oldmeadow C, Lee MK, Strachan DP, James AL, Huffman JE, Vitart V, Ramasamy A, Wareham NJ, Kaprio J, Wang XQ, Trochet H, Kähönen M, Flexeder C, Albrecht E, Lopez LM, de Jong K, Thyagarajan B, Alves AC, Enroth S, Enroth S, Omenaas E, Joshi PK, Fall T, Viñuela A, Launer LJ, Loehr LR, Fornage M, Li G, Wilk JB, Tang W, Manichaikul A, Lahousse L, Harris TB, North KE, Rudnicka AR, Hui J, Gu X, Lumley T, Wright AF, Hastie ND, Campbell S, Kumar R, Pin I, Scott RA, Pietiläinen KH, Surakka I, Liu Y, Holliday EG, Schulz H, Heinrich J, Davies G, Vonk JM, Wojczynski M, Pouta A, Johansson A, Wild SH, Ingelsson E, Rivadeneira F, Völzke H, Hysi PG, Eiriksdottir G, Morrison AC, Rotter JI, Gao W, Postma DS, White WB, Rich SS, Hofman A, Aspelund T, Couper D, Smith LJ, Psaty BM, Lohman K, Burchard EG, Uitterlinden AG, Garcia M, Joubert BR, McArdle WL, Musk AB, Hansel N, Heckbert SR, Zgaga L, van Meurs JB, Navarro P, Rudan I, Oh YM, Redline S, Jarvis DL, Zhao JH, Rantanen T, O'Connor GT, Ripatti S, Scott RJ, Karrasch S, Grallert H, Gaddis NC, Starr JM, Wijmenga C, Minster RL, Lederer DJ, Pekkanen

J, Gyllensten U, Campbell H, Morris AP, Gläser S, Hammond CJ, Burkart KM, Beilby J, Kritchevsky SB, Gudnason V, Hancock DB, Williams OD, Polasek O, Zemunik T, Kolcic I, Petrini MF, Wjst M, Kim WJ, Porteous DJ, Scotland G, Smith BH, Viljanen A, Heliövaara M, Attia JR, Sayers I, Hampel R, Gieger C, Deary IJ, Boezen HM, Newman A, Jarvelin MR, Wilson JF, Lind L, Stricker BH, Teumer A, Spector TD, Melén E, **Peters MJ**, Lange LA, Barr RG, Bracke KR, Verhamme FM, Sung J, Hiemstra PS, Cassano PA, Sood A, Hayward C, Dupuis J, Hall IP, Brusselle GG, Tobin MD, and London SJ. Genome-wide association analysis identifies six new loci associated with forced vital capacity. *Nature genetics, 2014, vol. 46, no. 7, pp. 669-677. DOI: 10.1038/ng.3011.*

Oei L, Hsu YH, Styrkarsdottir U, Eussen BH, de Klein A, **Peters MJ**, Halldorsson B, Liu CT, Alonso N, Kaptoge SK, Thorleifsson G, Hallmans G, Hocking LJ, Husted LB, Jameson KA, Kruk M, Lewis JR, Patel MS, Scollen S,Svensson O, Trompet S, van Schoor NM, Zhu K, Buckley BM, Cooper C, Ford I, Goltzman D, González-Macías J, Langdahl BL, Leslie WD, Lips P, Lorenc RS, Olmos JM, Pettersson-Kymmer U, Reid DM, Riancho JA,Slagboom PE, Garcia-Ibarbia C, Ingvarsson T, Johannsdottir H, Luben R, Medina-Gómez C, Arp P,Nandakumar K, Palsson ST, Sigurdsson G, van Meurs JB, Zhou Y, Hofman A, Jukema JW, Pols HA, Prince RL,Cupples LA, Marshall CR, Pinto D, Sato D, Scherer SW, Reeve J, Thorsteinsdottir U, Karasik D, Richards JB,Stefansson K, Uitterlinden AG, Ralston SH, Ioannidis JP, Kiel DP, Rivadeneira F, and Estrada K. A genome-wide copy number association study of osteoporotic fractures points to the 6p25.1 locus. *J Med Genet. 2014 Feb;51(2):122-31. DOI: 10.1136/jmedgenet-2013-102064.*

Berndt SI, Gustafsson S, Mägi R, Ganna A, Wheeler E, Feitosa MF, Justice AE, Monda KL, Croteau-Chonka DC, Day FR, Esko T, Fall T, Ferreira T, Gentilini D, Jackson AU, Luan J, Randall JC, Vedantam S, Willer CJ, Winkler TW, Wood AR, Workalemahu T, Hu YJ, Lee SH, Liang L, Lin DY, Min JL, Neale BM, Thorleifsson G, Yang J, Albrecht E, Amin N, Bragg-Gresham JL, Cadby G, den Heijer M, Eklund N, Fischer K, Goel A, Hottenga JJ, Huffman JE, Jarick I, Johansson Å, Johnson T, Kanoni S, Kleber ME, König IR, Kristiansson K, Kutalik Z, Lamina C, Lecoeur C, Li G, Mangino M, McArdle WL, Medina-Gomez C, Müller-Nurasyid M, Ngwa JS, Nolte IM, Paternoster L, Pechlivanis S, Perola M, **Peters MJ**, Preuss M, Rose LM, Shi J, Shungin D, Smith AV, Strawbridge RJ, Surakka I, Teumer A, Trip MD, Tyrer J, Van Vliet-Ostaptchouk JV, Vandenput L, Waite LL, Zhao JH, Absher D, Asselbergs FW, Atalay M, Attwood AP, Balmforth AJ, Basart H, Beilby J, Bonnycastle LL, Brambilla P, Bruinenberg M, Campbell H, Chasman DI, Chines PS, Collins FS, Connell JM, Cookson WO, de Faire U, de Vegt F, Dei M, Dimitriou M, Edkins S, Estrada K, Evans DM, Farrall M, Ferrario MM, Ferrières J, Franke L, Frau F, Gejman PV, Grallert H, Grönberg H, Gudnason V, Hall AS, Hall P, Hartikainen AL, Hayward C, Heard-Costa NL, Heath AC, Hebebrand J, Homuth G, Hu FB, Hunt SE, Hyppönen E, Iribarren C, Jacobs KB, Jansson JO, Jula A, Kähönen M, Kathiresan S, Kee F, Khaw KT, Kivimäki M, Koenig W, Kraja AT, Kumari M, Kuulasmaa K, Kuusisto J, Laitinen JH, Lakka TA, Langenberg C, Launer LJ, Lind L, Lindström J, Liu J, Liuzzi A, Lokki ML, Lorentzon M, Madden PA, Magnusson PK, Manunta P, Marek D, März W, Mateo Leach I, McKnight B, Medland SE, Mihailov E, Milani L, Montgomery GW, Mooser V, Mühleisen TW, Munroe PB, Musk AW, Narisu N, Navis G, Nicholson G, Nohr EA, Ong KK, Oostra BA, Palmer CN, Palotie A, Peden JF, Pedersen N, Peters A, Polasek O, Pouta A, Pramstaller PP, Prokopenko I, Pütter C, Radhakrishnan A, Raitakari O, Rendon A, Rivadeneira F, Rudan I, Saaristo TE, Sambrook JG, Sanders AR, Sanna S, Saramies J, Schipf

S, Schreiber S, Schunkert H, Shin SY, Signorini S, Sinisalo J, Skrobek B, Soranzo N, Stančáková A, Stark K, Stephens JC, Stirrups K, Stolk RP, Stumvoll M, Swift AJ, Theodoraki EV, Thorand B, Tregouet DA, Tremoli E, Van der Klauw MM, van Meurs JB, Vermeulen SH, Viikari J, Virtamo J, Vitart V, Waeber G, Wang Z, Widén E, Wild SH, Willemsen G, Winkelmann BR, Witteman JC, Wolffenbuttel BH, Wong A, Wright AF, Zillikens MC, Amouyel P, Boehm BO, Boerwinkle E, Boomsma DI, Caulfield MJ, Chanock SJ, Cupples LA, Cusi D, Dedoussis GV, Erdmann J, Eriksson JG, Franks PW, Froguel P, Gieger C, Gyllensten U, Hamsten A, Harris TB, Hengstenberg C, Hicks AA, Hingorani A, Hinney A, Hofman A, Hovingh KG, Hveem K, Illig T, Jarvelin MR, Jöckel KH, Keinanen-Kiukaanniemi SM, Kiemeney LA, Kuh D, Laakso M, Lehtimäki T, Levinson DF, Martin NG, Metspalu A, Morris AD, Nieminen MS, Njølstad I, Ohlsson C, Oldehinkel AJ, Ouwehand WH, Palmer LJ, Penninx B, Power C, Province MA, Psaty BM, Qi L, Rauramaa R, Ridker PM, Ripatti S, Salomaa V, Samani NJ, Snieder H, Sørensen TI, Spector TD, Stefansson K, Tönjes A, Tuomilehto J, Uitterlinden AG, Uusitupa M, van der Harst P, Vollenweider P, Wallaschofski H, Wareham NJ, Watkins H, Wichmann HE, Wilson JF, Abecasis GR, Assimes TL, Barroso I, Boehnke M, Borecki IB, Deloukas P, Fox CS, Frayling T, Groop LC, Haritunian T, Heid IM, Hunter D, Kaplan RC, Karpe F, Moffatt MF, Mohlke KL, O'Connell JR, Pawitan Y, Schadt EE, Schlessinger D, Steinthorsdottir V, Strachan DP, Thorsteinsdottir U, van Duijn CM, Visscher PM, Di Blasio AM, Hirschhorn JN, Lindgren CM, Morris AP, Meyre D, Scherag A, McCarthy MI, Speliotes EK, North KE, Loos RJ, and Ingelsson E. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nature Genetics, 2013, vol. 45, no. 5, pp. 501-512, DOI: 10.1038/ng.2606.*

Mayerle J, den Hoed CM, Schurmann C, Stolk L, Homuth G, **Peters MJ**, Capelle LG, Zimmermann K, Rivadeneira F, Gruska S, Völzke H, de Vries AC, Völker U, Teumer A, van Meurs JB, Steinmetz I, Nauck M, Ernst F, Weiss FU, Hofman A, Zenker M, Kroemer HK, Prokisch H, Uitterlinden AG, Lerch MM, and Kuipers EJ. Identification of genetic loci associated with Helicobacter pylori serologic status. *JAMA, 2013, vol. 309, no. 18, pp. 1912-1920, DOI: 10.1001/jama.2013.4350.*

Castaño Betancourt MC*, Cailotto F*, Kerkhof HJM*, Cornelis FMF, Doherty SA, Hart DJ, Hofman A, Luyten FP, Maciewicz RA, Mangino M, Metrustry S, Muir K, **Peters MJ**, Rivadeneira F, Wheeler M, Zhang W, Arden N, Spector TD, Uitterlinden AG, Doherty M, Lories RJU*, Valdes AM*, and van Meurs JBJ*. Genome-wide association and functional studies identify the DOT1L gene to be involved in cartilage thickness and hip osteoarthritis. *Proc Natl Acad Sci USA. 2012 May 22; 109(21): 8218–8223. DOI: 10.1073/pnas.1119899109*

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Lango Allen H, Lindgren CM, Luan J, Mägi R, Randall JC, Vedantam S, Winkler TW, Qi L, Workalemahu T, Heid IM, Steinthorsdottir V, Stringham HM, Weedon MN, Wheeler E, Wood AR, Ferreira T, Weyant RJ, Segrè AV, Estrada K, Liang L, Nemesh J, Park JH, Gustafsson S, Kilpeläinen TO, Yang J, Bouatia-Naji N, Esko T, Feitosa MF, Kutalik Z, Mangino M, Raychaudhuri S, Scherag A, Smith AV, Welch R, Zhao JH, Aben KK, Absher DM, Amin N, Dixon AL, Fisher E, Glazer NL, Goddard ME, Heard-Costa NL, Hoesel V, Hottenga JJ, Johansson A, Johnson T, Ketkar S, Lamina C, Li S, Moffatt MF, Myers RH, Narisu N, Perry JR, **Peters MJ**, Preuss M, Ripatti S, Rivadeneira F, Sandholt C, Scott LJ, Timpson NJ, Tyrer JP, van Wingerden S, Watanabe RM,

White CC, Wiklund F, Barlassina C, Chasman DI, Cooper MN, Jansson JO, Lawrence RW, Pellikka N, Prokopenko I, Shi J, Thiering E, Alavere H, Alibrandi MT, Almgren P, Arnold AM, Aspelund T, Atwood LD, Balkau B, Balmforth AJ, Bennett AJ, Ben-Shlomo Y, Bergman RN, Bergmann S, Biebermann H, Blakemore AI, Boes T, Bonnycastle LL, Bornstein SR, Brown MJ, Buchanan TA, Busonero F, Campbell H, Cappuccio FP, Cavalcanti-Proença C, Chen YD, Chen CM, Chines PS, Clarke R, Coin L, Connell J, Day IN, den Heijer M, Duan J, Ebrahim S, Elliott P, Elosua R, Eiriksdottir G, Erdos MR, Eriksson JG, Facheris MF, Felix SB, Fischer-Posovszky P, Folsom AR, Friedrich N, Freimer NB, Fu M, Gaget S, Gejman PV, Geus EJ, Gieger C, Gjesing AP, Goel A, Goyette P, Grallert H, Grässler J, Greenawalt DM, Groves CJ, Gudnason V, Guiducci C, Hartikainen AL, Hassanali N, Hall AS, Havulinna AS, Hayward C, Heath AC, Hengstenberg C, Hicks AA, Hinney A, Hofman A, Homuth G, Hui J, Igl W, Iribarren C, Isomaa B, Jacobs KB, Jarick I, Jewell E, John U, Jørgensen T, Jousilahti P, Jula A, Kaakinen M, Kajantie E, Kaplan LM, Kathiresan S, Kettunen J, Kinnunen L, Knowles JW, Kolcic I, König IR, Koskinen S, Kovacs P, Kuusisto J, Kraft P, Kvaløy K, Laitinen J, Lantieri O, Lanzani C, Launer LJ, Lecoeur C, Lehtimäki T, Lettre G, Liu J, Lokki ML, Lorentzon M, Luben RN, Ludwig B, MAGIC, Manunta P, Marek D, Marre M, Martin NG, McArdle WL, McCarthy A, McKnight B, Meitinger T, Melander O, Meyre D, Midthjell K, Montgomery GW, Morken MA, Morris AP, Mulic R, Ngwa JS, Nelis M, Neville MJ, Nyholt DR, O'Donnell CJ, O'Rahilly S, Ong KK, Oostra B, Paré G, Parker AN, Perola M, Pichler I, Pietiläinen KH, Platou CG, Polasek O, Pouta A, Rafelt S, Raitakari O, Rayner NW, Ridderstråle M, Rief W, Ruokonen A, Robertson NR, Rzehak P, Salomaa V, Sanders AR, Sandhu MS, Sanna S, Saramies J, Savolainen MJ, Scherag S, Schipf S, Schreiber S, Schunkert H, Silander K, Sinisalo J, Siscovick DS, Smit JH, Soranzo N, Sovio U, Stephens J, Surakka I, Swift AJ, Tammesoo ML, Tardif JC, Teder-Laving M, Teslovich TM, Thompson JR, Thomson B, Tönjes A, Tuomi T, van Meurs JB, van Ommen GJ, Vatin V, Viikari J, Visvikis-Siest S, Vitart V, Vogel CI, Voight BF, Waite LL, Wallaschofski H, Walters GB, Widen E, Wiegand S, Wild SH, Willemsen G, Witte DR, Witteman JC, Xu J, Zhang Q, Zgaga L, Ziegler A, Zitting P, Beilby JP, Farooqi IS, Hebebrand J, Huikuri HV, James AL, Kähönen M, Levinson DF, Macciardi F, Nieminen MS, Ohlsson C, Palmer LJ, Ridker PM, Stumvoll M, Beckmann JS, Boeing H, Boerwinkle E, Boomsma DI, Caulfield MJ, Chanock SJ, Collins FS, Cupples LA, Smith GD, Erdmann J, Froguel P, Grönberg H, Gyllensten U, Hall P, Hansen T, Harris TB, Hattersley AT, Hayes RB, Heinrich J, Hu FB, Hveem K, Illig T, Jarvelin MR, Kaprio J, Karpe F, Khaw KT, Kiemeney LA, Krude H, Laakso M, Lawlor DA, Metspalu A, Munroe PB, Ouwehand WH, Pedersen O, Penninx BW, Peters A, Pramstaller PP, Quertermous T, Reinehr T, Rissanen A, Rudan I, Samani NJ, Schwarz PE, Shuldiner AR, Spector TD, Tuomilehto J, Uda M, Uitterlinden A, Valle TT, Wabitsch M, Waeber G, Wareham NJ, Watkins H, Procardis Consortium, Wilson JF, Wright AF, Zillikens MC, Chatterjee N, McCarroll SA, Purcell S, Schadt EE, Visscher PM, Assimes TL, Borecki IB, Deloukas P, Fox CS, Groop LC, Haritunians T, Hunter DJ, Kaplan RC, Mohlke KL, O'Connell JR, Peltonen L, Schlessinger D, Strachan DP, van Duijn CM, Wichmann HE, Frayling TM, Thorsteinsdottir U, Abecasis GR, Barroso I, Boehnke M, Stefansson K, North KE, McCarthy MI, Hirschhorn JN, Ingelsson E, and Loos RJ. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics, 2010, vol. 42, no. 11, pp. 937-948. DOI: 10.1038/ng.686*

Heid IM, Jackson AU, Randall JC, Winkler TW, Qi L, Steinthorsdottir V, Thorleifsson G, Zillikens MC, Speliotes EK, Mägi R, Workalemahu T, White CC, Bouatia-Naji N, Harris TB, Berndt SI, Ingelsson E,

Willer CJ, Weedon MN, Luan J, Vedantam S, Esko T, Kilpeläinen TO, Kutalik Z, Li S, Monda KL, Dixon AL, Holmes CC, Kaplan LM, Liang L, Min JL, Moffatt MF, Molony C, Nicholson G, Schadt EE, Zondervan KT, Feitosa MF, Ferreira T, Lango Allen H, Weyant RJ, Wheeler E, Wood AR, MAGIC, Estrada K, Goddard ME, Lettre G, Mangino M, Nyholt DR, Purcell S, Smith AV, Visscher PM, Yang J, McCarroll SA, Nemesh J, Voight BF, Absher D, Amin N, Aspelund T, Coin L, Glazer NL, Hayward C, Heard-Costa NL, Hottenga JJ, Johansson A, Johnson T, Kaakinen M, Kapur K, Ketkar S, Knowles JW, Kraft P, Kraja AT, Lamina C, Leitzmann MF, McKnight B, Morris AP, Ong KK, Perry JR, **Peters MJ**, Polasek O, Prokopenko I, Rayner NW, Ripatti S, Rivadeneira F, Robertson NR, Sanna S, Sovio U, Surakka I, Teumer A, van Wingerden S, Vitart V, Zhao JH, Cavalcanti-Proença C, Chines PS, Fisher E, Kulzer JR, Lecoeur C, Narisu N, Sandholt C, Scott LJ, Silander K, Stark K, Tammesoo ML, Teslovich TM, Timpson NJ, Watanabe RM, Welch R, Chasman DI, Cooper MN, Jansson JO, Kettunen J, Lawrence RW, Pellikka N, Perola M, Vandenput L, Alavere H, Almgren P, Atwood LD, Bennett AJ, Biffar R, Bonnycastle LL, Bornstein SR, Buchanan TA, Campbell H, Day IN, Dei M, Dörr M, Elliott P, Erdos MR, Eriksson JG, Freimer NB, Fu M, Gaget S, Geus EJ, Gjesing AP, Grallert H, Grässler J, Groves CJ, Guiducci C, Hartikainen AL, Hassanali N, Havulinna AS, Herzig KH, Hicks AA, Hui J, Igl W, Jousilahti P, Jula A, Kajantie E, Kinnunen L, Kolcic I, Koskinen S, Kovacs P, Kroemer HK, Krzelj V, Kuusisto J, Kvaloy K, Laitinen J, Lantieri O, Lathrop GM, Lokki ML, Luben RN, Ludwig B, McArdle WL, McCarthy A, Morken MA, Nelis M, Neville MJ, Paré G, Parker AN, Peden JF, Pichler I, Pietiläinen KH, Platou CG, Pouta A, Ridderstråle M, Samani NJ, Saramies J, Sinisalo J, Smit JH, Strawbridge RJ, Stringham HM, Swift AJ, Teder-Laving M, Thomson B, Usala G, van Meurs JB, van Ommen GJ, Vatin V, Volpato CB, Wallaschofski H, Walters GB, Widen E, Wild SH, Willemsen G, Witte DR, Zgaga L, Zitting P, Beilby JP, James AL, Kähönen M, Lehtimäki T, Nieminen MS, Ohlsson C, Palmer LJ, Raitakari O, Ridker PM, Stumvoll M, Tönjes A, Viikari J, Balkau B, Ben-Shlomo Y, Bergman RN, Boeing H, Smith GD, Ebrahim S, Froguel P, Hansen T, Hengstenberg C, Hveem K, Isomaa B, Jørgensen T, Karpe F, Khaw KT, Laakso M, Lawlor DA, Marre M, Meitinger T, Metspalu A, Midthjell K, Pedersen O, Salomaa V, Schwarz PE, Tuomi T, Tuomilehto J, Valle TT, Wareham NJ, Arnold AM, Beckmann JS, Bergmann S, Boerwinkle E, Boomsma DI, Caulfield MJ, Collins FS, Eiriksdottir G, Gudnason V, Gyllensten U, Hamsten A, Hattersley AT, Hofman A, Hu FB, Illig T, Iribarren C, Jarvelin MR, Kao WH, Kaprio J, Launer LJ, Munroe PB, Oostra B, Penninx BW, Pramstaller PP, Psaty BM, Quertermous T, Rissanen A, Rudan I, Shuldiner AR, Soranzo N, Spector TD, Syvanen AC, Uda M, Uitterlinden A, Völzke H, Vollenweider P, Wilson JF, Witteman JC, Wright AF, Abecasis GR, Boehnke M, Borecki IB, Deloukas P, Frayling TM, Groop LC, Haritunians T, Hunter DJ, Kaplan RC, North KE, O'Connell JR, Peltonen L, Schlessinger D, Strachan DP, Hirschhorn JN, Assimes TL, Wichmann HE, Thorsteinsdottir U, van Duijn CM, Stefansson K, Cupples LA, Loos RJ, Barroso I, McCarthy MI, Fox CS, Mohlke KL, and Lindgren CM. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature Genetics, 2010, vol. 42, no. 11, pp. 949-960, DOI: 10.1038/ng.685*.

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segrè AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon Smith A, Mägi R, Pastinen T, Liang L, Heid IM, Luan J, Thorleifsson G, Winkler TW, Goddard ME, Sin Lo K, Palmer C, Workalemahu T, Aulchenko YS, Johansson A, Zillikens MC, Feitosa MF, Esko T, Johnson T, Ketkar S, Kraft

P, Mangino M, Prokopenko I, Absher D, Albrecht E, Ernst F, Glazer NL, Hayward C, Hottenga JJ, Jacobs KB, Knowles JW, Kutalik Z, Monda KL, Polasek O, Preuss M, Rayner NW, Robertson NR, Steinthorsdottir V, Tyrer JP, Voight BF, Wiklund F, Xu J, Zhao JH, Nyholt DR, Pellikka N, Perola M, Perry JR, Surakka I, Tammesoo ML, Altmaier EL, Amin N, Aspelund T, Bhangale T, Boucher G, Chasman DI, Chen C, Coin L, Cooper MN, Dixon AL, Gibson Q, Grundberg E, Hao K, Juhani Junttila M, Kaplan LM, Kettunen J, König IR, Kwan T, Lawrence RW, Levinson DF, Lorentzon M, McKnight B, Morris AP, Müller M, Suh Ngwa J, Purcell S, Rafelt S, Salem RM, Salvi E, Sanna S, Shi J, Sovio U, Thompson JR, Turchin MC, Vandenput L, Verlaan DJ, Vitart V, White CC, Ziegler A, Almgren P, Balmforth AJ, Campbell H, Citterio L, De Grandi A, Dominiczak A, Duan J, Elliott P, Elosua R, Eriksson JG, Freimer NB, Geus EJ, Glorioso N, Haiqing S, Hartikainen AL, Havulinna AS, Hicks AA, Hui J, Igl W, Illig T, Jula A, Kajantie E, Kilpeläinen TO, Koiranen M, Kolcic I, Koskinen S, Kovacs P, Laitinen J, Liu J, Lokki ML, Marusic A, Maschio A, Meitinger T, Mulas A, Paré G, Parker AN, Peden JF, Petersmann A, Pichler I, Pietiläinen KH, Pouta A, Ridderstråle M, Rotter JI, Sambrook JG, Sanders AR, Schmidt CO, Sinisalo J, Smit JH, Stringham HM, Bragi Walters G, Widen E, Wild SH, Willemsen G, Zagato L, Zgaga L, Zitting P, Alavere H, Farrall M, McArdle WL, Nelis M, **Peters MJ**, Ripatti S, van Meurs JB, Aben KK, Ardlie KG, Beckmann JS, Beilby JP, Bergman RN, Bergmann S, Collins FS, Cusi D, den Heijer M, Eiriksdottir G, Gejman PV, Hall AS, Hamsten A, Huikuri HV, Iribarren C, Kähönen M, Kaprio J, Kathiresan S, Kiemeney L, Kocher T, Launer LJ, Lehtimäki T, Melander O, Mosley TH Jr, Musk AW, Nieminen MS, O'Donnell CJ, Ohlsson C, Oostra B, Palmer LJ, Raitakari O, Ridker PM, Rioux JD, Rissanen A, Rivolta C, Schunkert H, Shuldiner AR, Siscovick DS, Stumvoll M, Tönjes A, Tuomilehto J, van Ommen GJ, Viikari J, Heath AC, Martin NG, Montgomery GW, Province MA, Kayser M, Arnold AM, Atwood LD, Boerwinkle E, Chanock SJ, Deloukas P, Gieger C, Grönberg H, Hall P, Hattersley AT, Hengstenberg C, Hoffman W, Lathrop GM, Salomaa V, Schreiber S, Uda M, Waterworth D, Wright AF, Assimes TL, Barroso I, Hofman A, Mohlke KL, Boomsma DI, Caulfield MJ, Cupples LA, Erdmann J, Fox CS, Gudnason V, Gyllensten U, Harris TB, Hayes RB, Jarvelin MR, Mooser V, Munroe PB, Ouwehand WH, Penninx BW, Pramstaller PP, Quertermous T, Rudan I, Samani NJ, Spector TD, Völzke H, Watkins H, Wilson JF, Groop LC, Haritunians T, Hu FB, Kaplan RC, Metspalu A, North KE, Schlessinger D, Wareham NJ, Hunter DJ, O'Connell JR, Strachan DP, Wichmann HE, Borecki IB, van Duijn CM, Schadt EE, Thorsteinsdottir U, Peltonen L, Uitterlinden AG, Visscher PM, Chatterjee N, Loos RJ, Boehnke M, McCarthy MI, Ingelsson E, Lindgren CM, Abecasis GR, Stefansson K, Frayling TM, and Hirschhorn JN. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature, 2010, vol. 467, no. 7317, pp. 832-838. DOI: 10.1038/nature09410*

Authors and affiliations

Aaron Isaacs – aaron.isaacs@gmail.com – Department of Epidemiology, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Abbas Dehghan – a.dehghan@erasmusmc.nl – Department of Epidemiology, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Abraham Aviv – avivab@njms.rutgers.edu – Center of Human Development and Aging, New Jersey Medical School, Newark, United States of America.

Alan T. Remaley – aremaley1@mail.nih.gov – Lipoprotein Metabolism Section, Cardio–Pulmonary Branch, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD, United States of America.

Albert Hofman – a.hofman@erasmusmc.nl – Department of Epidemiology, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Albert V. Smith – albert@hjarta.is – Icelandic Heart Association Research Institute, Kopavogur, Iceland; and University of Iceland, Reykjavik, Iceland.

Alexander Teumer – ateumer@uni–greifswald.de – Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany.

Alexandra Zhernakova – sashazhernakova@gmail.com – Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands.

Alicia K. Smith – aksmit3@emory.edu – Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, United States of America.

Allan F. McRae – a.mcrae@uq.edu.au – The Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia.

Ana M. Valdes – ana.valdes@kcl.ac.uk – Department of Twin Research and Genetic Epidemiology, King's College London, United Kingdom.

André G. Uitterlinden – a.g.uitterlinden@erasmusmc.nl – Department of Internal Medicine, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands; and Department of Epidemiology, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Andres Metspalu – andres.metspalu@ut.ee – Estonian Genome Center, University of Tartu, Tartu, Estonia.

Andrew B. Singleton – singleta@mail.nih.gov – Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, United States of America.

Andrew D. Johnson – johnsonad2@nhlbi.nih.gov – The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA, United States of America; and Cardiovascular Epidemiology and Human Genomics Branch, National Heart, Lung, and Blood Institute, Bethesda, MD, United States of America.

Anjali K. Henders – anjali.henders@qimr.edu.au – The Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia.

Anna Murray – A.Murray@exeter.ac.uk – Epidemiology and Public Health, University of Exeter Medical School, Exeter, United Kingdom.

Annemieke Kavelaars – A.Kavelaars@umcutrecht.nl – Laboratory of Neuroimmunology and Developmental Origins of Disease, University Medical Center Utrecht, the Netherlands.

Annette Peters – peters@helmholtz–muenchen.de – Institute of Epidemiologie II, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Neuherberg, Germany; and Research Unit of Molecular Epidemiology, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Neuherberg, Germany.

Antony Payton – tony.payton@manchester.ac.uk – Center for Integrated Genomic Medical Research, University of Manchester, Manchester, United Kingdom.

Anushka Soni – Anushka.Soni@ndorms.ox.ac.uk – NIHR Musculoskeletal Biomedical Research Unit, University of Oxford, Oxford, United Kingdom.

Astrid M. Suchy–Dicey – astridsd@gmail.com – Department of Epidemiology, University of Washington, Seattle, WA, United States of America.

Barbara E. Stranger – bstranger@medicine.bsd.uchicago.edu – Section of Genetic Medicine, Institute for Genomics and Systems Biology, University of Chicago, Chicago, Illinois, United States of America.

Ben Oostra – b.oostra@erasmusmc.nl – Department of Clinical Genetics, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Benjamin P. Fairfax – benjamin.fairfax@well.ox.ac.uk – Wellcome Trust Center for Human Genetics, Oxford, United Kingdom; and Department of Oncology, Cancer and Haematology Center, Churchill Hospital, Oxford, United Kingdom.

Blair H. Smith – blairsmith@abdn.ac.uk – Medical Research Institute, University of Dundee, Dundee, United Kingdom.

Brenda W.J.H. Penninx – b.penninx@vumc.nl – VU University Medical Center, Amsterdam, the Netherlands.

Brian H. Chen – chenbh@mail.nih.gov – The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts, United States of America; and The Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, Bethesda, Maryland, United States of America.

Bruce M. Psaty – psaty@u.washington.edu – Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology, and Health Services, University of Washington, Seattle, WA, United States of America; and Group Health Research Institute, Group Health Cooperative, Seattle, WA, United States of America.

Carolin Malsch – Carolin.Malsch@uni–wuerzburg.de – University of Greifswald, Greifswald, Germany.

Carolina Medina-Gomez – m.medinagomez@erasmusmc.nl – Department of Internal Medicine, Erasmus Medical Center, Rotterdam, the Netherlands; and The Generation R Study Group, Erasmus Medical Center, Rotterdam, the Netherlands.

Carsten O. Schmidt – carsten.schmidt@uni–greifswald.de – Institute for Community Medicine, University of Greifswald, Greifswald, Germany.

Cavin K. Ward-Caviness – cavin.wardcaviness@gmail.com – Institute of Epidemiologie II, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Neuherberg, Germany.

Chen Yao – chen.yao@nih.gov – The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts, United States of America; and The Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, Bethesda, Maryland, United States of America.

Christian Herder – christian.herder@ddz.uni–duesseldorf.de – Institute of Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany.

Christopher J. O'Donnell – odonnellc@nhlbi.nih.gov – The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts, United States of America; and The Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, Bethesda, Maryland, United States of America.

Chunyu Liu – chunyu.liu@nih.gov – The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts, United States of America; and The Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, Bethesda, Maryland, United States of America.

Cindy G. Boer – c.boer@erasmusmc.nl – Department of Internal Medicine, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Claudia Schurmann – claudia.schurmann@mssm.edu – Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany; and The Charles Bronfman Institute for Personalized Medicine, Genetics of Obesity & Related Metabolic Traits Program, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, United States of America.

Cornelia M. van Duijn – c.vanduijn@erasmusmc.nl – Department of Epidemiology, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Cyrus Cooper – cc@mrc.soton.ac.uk – MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton General Hospital, United Kingdom; and University of Oxford, Oxford, United Kingdom.

Daniel A. Enquobahrie – danenq@u.washington.edu – Department of Epidemiology, University of Washington, Seattle, WA, United States of America.

Daniel Levy – levyd@nhlbi.nih.gov – The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA, United States of America; and Population Studies Branch, National Heart, Lung, and Blood Institute, Bethesda, MD, United States of America.

Daria V. Zhernakova – d.v.zhernakova@umcg.nl – Department of Genetics, University Medical Center Groningen, Groningen, the Netherlands.

David Felson – David.Felson@manchester.ac.uk – Clinical Epidemiology Unit, Boston University School of Medicine, Boston, MA, United States of America.

David Karasik – karasik@hsl.harvard.edu – Faculty of Medicine in the Galilee, Bar–Ilan University, Safed, Israel.

David Melzer – D.Melzer@exeter.ac.uk – Epidemiology and Public Health, University of Exeter Medical School, Exeter, United Kingdom.

Deborah J. Hart – deborah.hart@kcl.ac.uk – Department of Twin Research and Genetic Epidemiology, King's College London, United Kingdom.

Dena G. Hernandez – hernand@mail.nih.gov – Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, United States of America; and Reta Lila Weston Institute and Department of Molecular Neuroscience, UCL Institute of Neurology, Queen Square, London, United Kingdom.

Dieuwke Schiphof – d.schiphof@erasmusmc.nl – Department of General Practice, Erasmus Medical Center, Rotterdam, the Netherlands.

Divya Mehta – mehta@mpipsykl.mpg.de – Max–Planck Institute of Psychiatry, Munich, Germany.

Dörte Radke – doerte.radke@uni–greifswald.de – Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany.

Douglas P. Kiel – kiel@hsl.harvard.edu – Hebrew Rehabilitation Center, Boston, United States of America; and Institute for Aging Research, Boston, United States of America.

Edwin H. G. Oei – e.oei@erasmusmc.nl – Department of Radiology, Erasmus Medical Center Rotterdam, the Netherlands.

Elaine M. Dennison – emd@mrc.soton.ac.uk – MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton General Hospital, United Kingdom; and Victoria University of Wellington, New Zealand.

Elisabeth B. Binder – ebinder@emory.edu – Max–Planck Institute of Psychiatry, Munich, Germany.

Elizabeth M. Kennedy – emkenn2@emory.edu – Department of Human Genetics, School of Medicine, Emory University, Atlanta, GA, United States of America.

Eric K. Moses – eric.moses@uwa.edu.au – Center for Genetic Epidemiology and Biostatistics, The University of Western Australia, Perth, Western Australia, Australia.

Eva Reinmaa – eva.reinmaa@ut.ee – Estonian Genome Center, University of Tartu, Tartu, Estonia.

Eva Reischl – eva.reischl@helmholtz–muenchen.de – Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany.

Fernando Rivadeneira – f.rivadeneira@erasmusmc.nl – Department of Internal Medicine, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands; and Department of Epidemiology, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Frances M. K. Williams – frances.williams@kcl.ac.uk – Department of Twin Research and Genetic Epidemiology, King's College London, United Kingdom.

Frank J.P.M. Huygen – f.huygen@erasmusmc.nl – Department of Anaesthesiology, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Gareth T. Jones – gareth.jones@abdn.ac.uk – Aberdeen Pain Research Collaboration (Epidemiology Group), University of Aberdeen, Aberdeen, United Kingdom.

Gary J. Macfarlane – g.j.macfarlane@abdn.ac.uk – Aberdeen Pain Research Collaboration (Epidemiology Group), University of Aberdeen, Aberdeen, United Kingdom.

Georg Homuth – georg.homuth@uni–greifswald.de – Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany.

George L. Sutphin – sutphin@gmail.com – Nathan Shock Center of Excellence in the Basic Biology of Aging, The Jackson Laboratory, Bar Harbor, ME, United States of America.

Gerard van Grootheest – G.Grootheest@ggzingeest.nl – Neuroscience Campus Amsterdam.

Gina Peloso – gina@broadinstitute.org – Cardiovascular Research Center, Massachusetts General Hospital, Boston, United States of America.

Graham Greg Neely – g.neely@garvan.org.au – Neuroscience Division, Garvan Institute of Medical Research, Australia and Charles Perkins Center and School of Molecular Bioscience, The University of Sydney NSW 2006 Australia.

Grant W. Montgomery – Grant.Montgomery@qimr.edu.au – Queensland Institute of Medical Research, Brisbane, Queensland, Australia.

Gudny Eiriksdottir – gudny@hjarta.is – Icelandic Heart Association Research Institute, Kopavogur, Iceland.

Hanieh Yaghootkar – H.Yaghootkar@exeter.ac.uk – Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, United Kingdom.

Hanneke J.M. Kerkhof – hanneke_kerkhof2@hotmail.com – Department of Internal Medicine, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands

Hanneke L.D.M. Willemen – H.L.D.M.Willemen@umcutrecht.nl – Laboratory of Neuroimmunology and Developmental Origins of Disease, University Medical Center Utrecht, the Netherlands.

Hans-Jörgen Grabe – grabeh@uni–greifswald.de – Department of Psychiatry and Psychotherapy, Helios Hospital Stralsund, University Medicine Greifswald, Greifswald, Germany.

Harald Grallert – harald.grallert@helmholtz–muenchen.de – Research Unit of Molecular Epidemiology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany.
Institute of Epidemiology II, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany; and German Center for Diabetes Research (DZD e.V.), Partner Munich, Munich, Germany, 28 DZHK (German Center for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany.

Harald H.H. Göring – hgoring@txbiomedgenetics.org – Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, United States of America.

Harm-Jan Westra – westra.harmjan@gmail.com – Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands.

Henry Völzke – voelzke@uni–greifswald.de – Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany.

Holger Prokisch – Prokisch@helmholtz–muenchen.de – Institute of Human Genetics, Helmholz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany; and Institute of Human Genetics, Technical University Munich, Munich, Germany.

Honghuang Lin – hhlin@bu.edu – Department of Medicine, Boston University, United States of America.

Iiris Hovatta – iiris.hovatta@helsinki.fi – Department of Biosciences, University of Helsinki, Helsinki, Finland; and Department of Mental Health and Substance Abuse Services, National Institute for Health and Welfare, Helsinki, Finland.

Ingrid Meulenbelt – I.Meulenbelt@lumc.nl – Department of Molecular Epidemiology, Leiden University Medical Center, Leiden, the Netherlands.

Jack W. Kent Jr. – jkent@txbiomedgenetics.org – Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, United States of America.

James B. Brown – benbrownofberkeley@gmail.com – Department of Statistics, University of California Berkeley, Berkeley, California, United States of America; and Department of Genome Dynamics, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America.

Jan H. Veldink – polinmz@umcutrecht.nl – Department of Neurology, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Utrecht, the Netherlands.

Jeanine Houwing-Duistermaat – j.j.houwing@lumc.nl – Department of Medical Statistics, Leiden University Medical Center, Leiden, the Netherlands.

Jennifer A. Smith – smjenn@umich.edu – Department of Epidemiology, University of Michigan, Ann Arbor, MI, United States of America.

Jennifer Brody – jeco@uw.edu – Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, United States of America.

Jeroen van Rooij – j.vanrooij@erasmusmc.nl – Department of Internal Medicine, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Jerome I. Rotter – jrotter@labiomed.org – Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Research Institute at Harbor–UCLA Medical Center, Torrance, CA, United States of America.

Jian Yang – jian.yang@uq.edu.au – The Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia; and University of Queensland Diamantina Institute, University of Queensland, Princess Alexandra Hospital, Brisbane, Queensland, Australia.

Jingzhong Ding – Jding@wakehealth.edu – Department of Internal Medicine, Wake Forest School of Medicine, Winston–Salem, North Carolina, United States of America.

Joanne E. Curran – jcurran@txbiomedgenetics.org – Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, United States of America.

Joanne M. Murabito – murabito@bu.edu – NHLBI's and Boston University's Framingham Heart Study, Framingham, MA, United States of America; and General Internal Medicine Section, Boston University, Boston, MA, United States of America.

Johan L. Bloem – J.L.Bloem@lumc.nl – Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands.

Johannes Kettunen – johannes.kettunen@helsinki.fi – Computational Medicine, Institute of Health Sciences, Faculty of Medicine, University of Oulu, Oulu, Finland; and Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland; and Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland.

John Blangero – john@txbiomedgenetics.org – Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, United States of America.

John McBeth – John.McBeth@manchester.ac.uk – Arthritis Research UK Epidemiology Unit, University of Manchester, Manchester Academic Health Science Center, Manchester, United Kingdom.

Joseph E. Powell – joseph.powell@uq.edu.au – Center for Neurogenetics and Statistical Genomics, Queensland Brain Institute, University of Queensland, St Lucia, Brisbane, Australia; and The Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia.

Joyce B.J. van Meurs – j.vanmeurs@erasmusmc.nl – Department of Internal Medicine, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Juha Karjalainen – j.karjalainen@umcg.nl – Department of Genetics, University Medical Center Groningen, Groningen, the Netherlands.

Julian C. Knight – julian@well.ox.ac.uk – Wellcome Trust Center for Human Genetics, Oxford, United Kingdom.

K. Maria Nylocks – maria.nylocks@emory.edu – Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, United States of America.

Karen A. Jameson – kaj@mrc.soton.ac.uk – MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton General Hospital, United Kingdom.

Karen N. Conneely – kconnee@emory.edu – Department of Human Genetics, School of Medicine, Emory University, Atlanta, GA, United States of America.

Kate L. Holliday – Kate.Holliday@manchester.ac.uk – Arthritis Research UK Epidemiology Unit, University of Manchester, Manchester Academic Health Science Center, Manchester, United Kingdom.

Katharina Schramm – katharina.schramm@helmholtz–muenchen.de – Institute of Human Genetics, Helmholz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany; and Institute of Human Genetics, Technical University Munich, Munich, Germany.

Kathryn L. Lunetta – klunetta@bu.edu – Boston University School of Public Health.

Ke Wang – Ke.Wang@manchester.ac.uk – Clinical Epidemiology Unit, Boston University School of Medicine, Boston, MA, United States of America.

Kerry J. Ressler – kressle@emory.edu – Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, United States of America.

Konstantin Strauch – strauch@helmholtz–muenchen.de – Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig–Maximilians–Universität, Neuherberg, Germany; and Institute of Genetic Epidemiology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany.

Krista Fischer – Krista.Fischer@ut.ee – Estonian Genome Center, University of Tartu, Tartu, Estonia

L. Adrienne Cupples – adrienne@bu.edu – The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, United States of America.

Leif Steil – steil@uni–greifswald.de – Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany.

Leonard Lipovich – llipovich@med.wayne.edu – Center for Molecular Medicine and Genetics, School of Medicine, Wayne State University, Detroit, Michigan, United States of America.

Lenore J. Launer – LaunerL@nia.nih.gov – Intramural Research Program, Laboratory of Epidemiology, Demography, and Biometry, National Institute on Aging, Bethesda, United States of America.

Leonard H. Van den Berg – polinmz@umcutrecht.nl – Department of Neurology, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Utrecht, the Netherlands.

Liina Tserel – liina.tserel@ut.ee – Molecular Pathology, Institute of Biomedicine, University of Tartu, Estonia.

Lili Milani – lili.milani@ut.ee – Estonian Genome Center, University of Tartu, Tartu, Estonia.

Liliane Pfeiffer – liliane.pfeiffer@helmholtz–muenchen.de – Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany.

Linda Broer – l.broer@erasmusmc.nl – Department of Internal Medicine, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Lisa J. Oyston – l.oyston@garvan.org.au – Neuroscience Division, Garvan Institute of Medical Research, Australia and Charles Perkins Center and School of Molecular Bioscience, The University of Sydney NSW 2006 Australia.

Lisette Stolk – stolklisette@gmail.com – Department of Internal Medicine, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Lita Freeman – litaf@mail.nih.gov – Lipoprotein Metabolism Section, Cardio–Pulmonary Branch, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, United States of America.

Lorna W. Harries – L.W.Harries@exeter.ac.uk – University of Exeter Medical School, United Kingdom.

Lude Franke – lude@ludesign.nl – Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands.

Luigi Ferrucci – FerrucciLu@grc.nia.nih.gov – Clinical Research Branch, National Institute on Aging, Baltimore, MD, United States of America.

Luke C. Pilling – L.Pilling@exeter.ac.uk – Epidemiology and Public Health, University of Exeter Medical School, Exeter, United Kingdom.

Lynne J. Hocking – l.hocking@abdn.ac.uk – Aberdeen Pain Research Collaboration (Musculoskeletal Research), University of Aberdeen, United Kingdom.

Maren Carstensen-Kirberg – maren.carstensen@ddz.uni–duesseldorf.de – Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany.

Margreet Kloppenburg – g.kloppenburg@lumc.nl – Department of Rheumatology, Leiden University Medical Center, the Netherlands; and Department of Clinical Epidemiology, Leiden University Medical Center, the Netherlands.

Mari Nelis – mari.nelis@ut.ee – Estonian Genome Center, University of Tartu, Tartu, Estonia.

Maria Popham – maria.popham@kcl.ac.uk – Department of Twin Research and Genetic Epidemiology, King's College London, United Kingdom.

Marcus Dörr – mdoerr@uni–greifswald.de – University Medicine Greifswald, Department of Internal Medicine B-Cardiology, Greifswald, Germany; and DZHK (German Center for Cardiovascular Research), partner site Greifswald, Greifswald, Germany, 32 Universitäres Herzzentrum Hamburg, Hamburg, Germany.

Marcus H. Stoiber – marcus.stoiber@gmail.com – Department of Statistics, University of California Berkeley, Berkeley, California United States of America.

Maren Carstensen – Maren.Carstensen@DDZ.uni-duesseldorf.de – Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany; and German Center for Diabetes Research (DZD e.V.), Partner Düsseldorf, Düsseldorf, Germany.

Marjolein de Kruijf – m.dekruijf@erasmusmc.nl – Department of Internal Medicine, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands; and Department of Anaesthesiology, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Mark W. Christiansen – mwchristiansen@gmail.com – Cardiovascular Health Research Unit, University of Washington, Seattle, United States of America.

Markus Perola – markus.perola@thl.fi – Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland; and Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland.

Matthew P. Johnson – mjohnson@txbiomedgenetics.org – Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, United States of America.

Matthias Nauck – matthias.nauck@uni–greifswald.de – Institute for Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Germany.

Melanie Waldenberger – waldenberger@helmholtz–muenchen.de – Research Unit of Molecular Epidemiology, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Neuherberg, Germany; and Institute of Epidemiologie II, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Neuherberg, Germany.

Michael A. Horan – michael.horan@manchester.ac.uk – Mental Health and Neurodegeneration Group, School Community Based Medicine, University of Manchester, Manchester, United Kingdom.

Michael A. Nalls – m.nalls.working@gmail.com – Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, United States of America.

Michael M.P.J. Verbiest – m.m.p.j.verbiest@erasmusmc.nl – Department of Internal Medicine, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Michael Roden – michael.roden@ddz.uni–duesseldorf.de – Institute of Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany; and Division of Endocrinology and Diabetology, University Hospital Düsseldorf, Heinrich Heine University, Düsseldorf, Germany.

Mila Jhamai – p.jhamai@erasmusmc.nl – Department of Internal Medicine, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Myriam Fornage – Myriam.Fornage@uth.tmc.edu – Division of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, University of Texas Health Sciences, Center at Houston, TX, United States of America; and Institute of Molecular Medicine, University of Texas Health Sciences Center at Houston, TX, United States of America.

Najaf Amin – n.amin@erasmusmc.nl – Department of Epidemiology, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Nalini Raghavachari – nraghavachari@mail.nih.gov – Division of Geriatrics and Clinical Gerontology National Institute on Aging, Bethesda, Maryland, United States of America.

Neil Pendleton – neil.pendleton@manchester.ac.uk – Mental Health and Neurodegeneration Group, School Community Based Medicine, University of Manchester, Manchester, United Kingdom.

Nicholas G. Martin – nick.martin@qimr.edu.au – The Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia.

Niels Eijkelkamp – N.Eijkelkamp@umcutrecht.nl – Molecular Nociception Group, Wolfson Institute for Biomedical Research, University College London, United Kingdom; and Laboratory for Translational Immunology, UMC Utrecht, Utrecht, the Netherlands.

Nigel K. Arden – nka@mrc.soton.ac.uk – MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton General Hospital, United Kingdom; and University of Oxford, Oxford, United Kingdom.

Nilesh J. Samani – njs@leicester.ac.uk – Department of Cardiovascular Sciences, University of Leicester, Leicester, United Kingdom; and National Institute for Health Research Leicester Cardiovascular Biomedical Research Unit, Glenfield Hospital, Leicester, United Kingdom.

Noman Bahkshi – n.bahkshi@garvan.org.au – Neuroscience Division, Garvan Institute of Medical Research, Australia and Charles Perkins Center and School of Molecular Bioscience, The University of Sydney NSW 2006 Australia.

P. Eline Slagboom – p.slagboom@lumc.nl – Department of Molecular Epidemiology, Leiden University Medical Center, Leiden, the Netherlands.

Pärt Peterson – part.peterson@ut.ee – Molecular Pathology, Institute of Biomedicine, University of Tartu, Estonia.

Pascal Arp – p.arp@erasmusmc.nl – Department of Internal Medicine, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Paul Courchesne – courchesnepl@nhlbi.nih.gov – The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts, United States of America; and The Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, Bethesda, Maryland, United States of America.

Paula Singmann – paula.singmann@helmholtz–muenchen.de – Research Unit of Molecular Epidemiology, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Neuherberg, Germany; and Institute of Epidemiologie II, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Neuherberg, Germany.

Peter A.C. 't Hoen – P.A.C._t_Hoen@lumc.nl – Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands.

Peter J. Bickel – bickel.peter@gmail.com – Department of Statistics, University of California Berkeley, Berkeley, California, United States of America.

Peter J. Munson – munson@mail.nih.gov – The Mathematical and Statistical Computing Laboratory, Center for Information Technology, National Institutes of Health, Bethesda, MD, United States of America.

Peter M. Visscher – peter.visscher@uq.edu.au – The Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia.

Philip L. De Jager – pdejager@rics.bwh.harvard.edu – Program in Translational NeuroPsychiatric Genomics, Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, United States of America.

Poching Liu – pcliu@nhlbi.nih.gov – Genomics Core facility Genetics & Developmental Biology Center, National Heart, Lung, and Blood Institute, Bethesda, Maryland, United States of America.

Qiao Ping Wang – q.wang@garvan.org.au – Neuroscience Division, Garvan Institute of Medical Research, Australia and Charles Perkins Center and School of Molecular Bioscience, The University of Sydney NSW 2006 Australia.

Quinta Helmer – q.helmer@xs4all.nl – Department of Medical Statistics, Leiden University Medical Center, Leiden, the Netherlands.

Ramachandran Vasan – vasan@bu.edu – The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts, United States of America.

Reiner Biffar – biffar@uni–greifswald.de – University of Greifswald, Germany.

Richard Wang – rwang50290@aol.com – Genomics Core facility Genetics & Developmental Biology Center, National Heart, Lung, and Blood Institute, Bethesda, Maryland, United States of America.

Rick Jansen – Ri.Jansen@ggzingeest.nl – Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, the Netherlands.

Ritsert C. Jansen – r.c.jansen@rug.nl – Groningen Bioinformatics Center, University of Groningen, Groningen, the Netherlands.

Robert L. Hanson – rhanson@phx.niddk.nih.gov – Phoenix Epidemiology and Clinical Research Branch, National Institute of Diabetes and Digestive and Kidney Disease, National Institutes of Health, Phoenix, Arizona, United States of America.

Roby Joehanes – robyjoehanes@hsl.harvard.edu – The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA, United States of America; and Population Studies Branch, National Heart, Lung, and Blood Institute, Bethesda, MD, United States of America.

Ron Korstanje – ron.korstanje@jax.org – Nathan Shock Center of Excellence in the Basic Biology of Aging, The Jackson Laboratory, Bar Harbor, ME, United States of America.

Russell P. Tracy – russell.tracy@uvm.edu – Department of Pathology, University of Vermont College of Medicine, Colchester, VT, United States of America.

Sai-Xia Ying – yings@mail.nih.gov – Mathematical and Statistical Computing Laboratory, Center for Information Technology, National Institutes of Health, Bethesda, Maryland, United States of America.

Samuli Ripatti – samuli.ripatti@helsinki.fi – Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland; and Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland; and Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom; and Department of Public Health, Hjelt Institute, University of Helsinki, Helsinki, Finland.

Sarah Williams-Blangero – sarah@txbiomedgenetics.org – Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, United States of America.

Sayuko Kobes – skobes@phx.niddk.nih.gov – Phoenix Epidemiology and Clinical Research Branch, National Institute of Diabetes and Digestive and Kidney Disease, National Institutes of Health, Phoenix, Arizona, United States of America.

Sebo Withoff – s.withoff@umcg.nl – Department of Genetics, University Medical Center Groningen, Groningen, the Netherlands.

Seiko Makino – seiko.makino@well.ox.ac.uk – Wellcome Trust Center for Human Genetics, Oxford, United Kingdom.

Sekar Kathiresan – sekar@broadinstitute.org – Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, United States of America.

Seth G. Thacker – thackersg@mail.nih.gov – Lipoprotein Metabolism Section, Cardio-Pulmonary Branch, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD, United States of America.

Shannon Bean – shannon.bean@jax.org – Nathan Shock Center of Excellence in the Basic Biology of Aging, The Jackson Laboratory, Bar Harbor, ME, United States of America.

Sharon L.R. Kardia – skardia@umich.edu – Department of Epidemiology, University of Michigan, Ann Arbor, MI, United States of America.

Sita M.A. Bierma-Zeinstra – s.bierma–zeinstra@erasmusmc.nl – Department of General Practice, Erasmus MC, Rotterdam, the Netherlands.

Silva Kasela – silva.kasela@ut.ee – Institute of Molecular and Cell Biology, Estonian Genome Center, University of Tartu, Tartu, Estonia.

Sina A. Gharib – sagharib@u.washington.edu – Computational Medicine Core, Center for Lung Biology, University of Washington, Seattle, WA, United States of America.

Sonja Zeilinger – sonja.zeilinger@helmholtz-muenchen.de – Research Unit of Molecular Epidemiology, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Neuherberg, Germany; and Institute of Epidemiologie II, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Neuherberg, Germany.

Sophie H. van Wingerden – sophievanwingerden@hotmail.com – Department of Epidemiology, Erasmus Medical Center Rotterdam, Rotterdam, the Netherlands.

Stefania Bandinelli – stefania.bandinelli@asf.toscana.it – Geriatric Unit, Azienda Sanitaria di Firenze, Florence, Italy.

Stephan B. Felix – felix@uni–greifswald.de – University Medicine Greifswald, Department of Internal Medicine B-Cardiology, Greifswald, Germany; and DZHK (German Center for Cardiovascular Research), partner site Greifswald, Greifswald, Germany, 32 Universitäres Herzzentrum Hamburg, Hamburg, Germany.

Stephen T. Turner – turner.stephen@mayo.edu – Division of Nephrology and Hypertension, Department of Medicine, Mayo Clinic, Rochester, MN, United States of America.

Sumanta Basu – sumbose@gmail.com – Department of Statistics, University of California Berkeley, Berkeley, California, United States of America.

Sven Gläser – sven.glaeser@uni–greifswald.de – University of Greifswald, Germany.

Tamara B. Harris – harris99@nia.nih.gov – Intramural Research Program, Laboratory of Epidemiology, Demography, and Biometry, National Institute on Aging, Bethesda, United States of America.

Tanja Zeller – t.zeller@uke.de – Universitäres Herzzentrum Hamburg, Hamburg, Germany; and DZHK (German Center for Cardiovascular Research), partner site Hamburg/Kiel/Lübeck, Hamburg, Germany.

Taru Tukiainen – taru.tukiainen@helsinki.fi – Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland; and Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland.

Thomas Illig – illig.thomas@mh-hannover.de – Hannover Unified Biobank, Hannover Medical School, Hannover, Germany.

Thomas Kocher – kocher@uni-greifswald.de – Unit of Periodontology, Department of Restorative Dentistry, Periodontology and Endodontology, University Medicine Greifswald, Greifswald, Germany.

Thomas Lumley – t.lumley@auckland.ac.nz – Department of Statistics, University of Auckland, Auckland, New Zealand.

Thomas Meitinger – meitinger@helmholtz–muenchen.de – Institute of Human Genetics, Helmholtz Zentrum München-German Research Center for Environmental Health, Neuherberg, Germany; and Institute of Human Genetics, Technische Universität München, München, Germany; and DZHK (German Center for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany.

Tianxiao Huan – tianxiao.huan@nih.gov – The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA, United States of America; and The Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, Bethesda, Maryland, United States of America.

Till Ittermann – till.ittermann@uni-greifswald.de – Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany.

Tim D. Spector – tim.spector@kcl.ac.uk – Department of Twin Research and Genetic Epidemiology, King's College London, United Kingdom.

Tim Kacprowski – tim.kacprowski@uni-greifswald.de – Interfaculty Institute of Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany.

Timothy M. Frayling – tim.frayling@pms.ac.uk – Genetics of Complex Traits, University of Exeter Medical School, Exeter, United Kingdom.

Tonu Esko – tonu.esko@ut.ee – Estonian Genome Center, University of Tartu, Tartu, Estonia; and Division of Endocrinology, Children's Hospital Boston, United States of America; and Department of Genetics, Harvard Medical School, Boston, United States of America; and Broad Insitite, Cambridge, United States of America.

Torsten Klengel – klengel@mpipsykl.mpg.de – Max–Planck Institute of Psychiatry, Munich, Germany.

Toshiko Tanaka – tanakato@mail.nih.gov – National Institute on Aging, Baltimore, United States of America.

Towfique Raj – towfique@broadinstitute.org – Division of Immunology, Department of Microbiology and Immunobiology, Harvard Medical School, Boston, Massachusetts, United States of America.

Tuhina Neogi – tneogi@bu.edu – Clinical Epidemiology Unit, Boston University School of Medicine, Boston, MA, United States of America.

Uwe Völker – voelker@uni–greifswald.de – Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany; and DZHK (German Center for Cardiovascular Research), partner site Greifswald, Greifswald, Germany.

Valur Emilsson – valur_emilsson@merck.com – Icelandic Heart Association, Iceland.

Veikko Salomaa – veikko.salomaa@thl.fi – Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland.

Veryan Codd – vc15@le.ac.uk – Department of Cardiovascular Sciences, University of Leicester, Leicester, United Kingdom; and National Institute for Health Research Leicester Cardiovascular Biomedical Research Unit, Glenfield Hospital, Leicester, United Kingdom.

Vilmundur Gudnason – v.gudnason@hjarta.is – Icelandic Heart Association Research Institute, Kopavogur, Iceland; and University of Iceland, Reykjavik, Iceland.

Wendy Thomson – wendy.thomson@manchester.ac.uk – Arthritis Research UK Epidemiology Unit, University of Manchester, Manchester Academic Health Science Center, Manchester, United Kingdom.

William Henley – W.E.Henley@exeter.ac.uk – University of Exeter Medical School, United Kingdom.

William Ollier – Bill.Ollier@manchester.ac.uk – Center for Integrated Genomic Medical Research, University of Manchester, Manchester, United Kingdom.

Wouter den Hollander – W.den_Hollander.MOLEPI@lumc.nl – Department of Molecular Epidemiology, Leiden University Medical Center, Leiden, the Netherlands.

Xia Yang – xyang123@ucla.edu – Department of Integrative Biology and Physiology, University of California, Los Angeles, Los Angeles, California, United States of America.

Yana A. Wilson – y.wilson@garvan.org.au – Neuroscience Division, Garvan Institute of Medical Research, Australia and Charles Perkins Center and School of Molecular Bioscience, The University of Sydney NSW 2006 Australia.

Yang Li – y.li01@umcg.nl – Groningen Bioinformatics Center, University of Groningen, Groningen, the Netherlands.

Yii-Der I. Chen – ichen@labiomed.org – Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Research Institute at Harbor–UCLA Medical Center, Torrance, CA, United States of America.

Yolande F. Ramos – Y.F.M.Ramos@lumc.nl – Department of Molecular Epidemiology, Leiden University Medical Center, Leiden, the Netherlands.

Yongmei Liu – yoliu@wakehealth.edu – Department of Epidemiology and Prevention, Public Health Sciences, Wake Forest School of Medicine, Winston–Salem, North Carolina, United States of America.

Members of the BIOS Consortium *(chapter 3.2)*

Bastiaan T. Heijmans, Peter A.C. 't Hoen, Joyce B.J. van Meurs, Aaron Isaacs, Rick Jansen, Lude Franke, André G. Uitterlinden, Bert A. Hofman, Carla J.H. van der Kallen, Casper G. Schalkwijk, Cisca Wijmenga, Coen D.A. Stehouwer, Cornelia M. van Duijn, Dasha V. Zhernakova, Diana van Heemst, Dorret I. Boomsma, Erik W. van Zwet, Ettje F. Tigchelaar, Freerk van Dijk, H. Eka D. Suchiman, Hailiang Mei, Irene Nooren, Jan Bot, Jan H. Veldink, Jenny van Dongen, Jeroen van Rooij, Joris Deelen, Jouke J. Hottenga, Leonard H. van den Berg, Maarten van Iterson, Marc Jan Bonder, Marian Beekman, Marijn Verkerk, Marleen M.J. van Greevenbroek, Martijn Vermaat, Matthijs Moed, Michael M.P.J. Verbiest, Michiel van Galen, Morris A. Swertz, Nico Lakenberg, P. Eline Slagboom, P. Mila Jhamai, Patrick Deelen, Peter van 't Hof, René Luijk, René Pool, Ruud van der Breggen, Sasha Zhernakova, Szymon M. Kielbasa, and Wibowo Arindrarto.

Members of the CHARGE Consortium *(chapter 3.2)*

Anni Joensuu (DILGOM), Johannes Kettunen (DILGOM), Urmo Vosa (EGCUT), Tonu Esko (EGCUT), Harm-Jan Westra (FEHRMANN), Lude Franke (FEHRMANN), Hanieh Yaghootkar (INCHIANTI), Timothy M. Frayling (INCHIANTI), Katharina Schramm (KORA), Holger Prokisch (KORA), Tim Kacprowski (SHIP), and Alexander Teumer (SHIP).

Members of the International Consortium for Blood Pressure GWAS (ICBP) *(chapter 2.2)*

Georg B. Ehret, Patricia B. Munroe, Kenneth M. Rice, Murielle Bochud, Andrew D. Johnson, Daniel I. Chasman, Albert V. Smith, Martin D. Tobin, Germaine C. Verwoert, Shih-JenHwang, Vasyl Pihur, Peter Vollenweider, Paul F. O'Reilly, Najaf Amin, Jennifer L. Bragg-Gresham, Alexander Teumer, Nicole L. Glazer, Lenore Launer, Jing Hua Zhao, Yurii Aulchenko, Simon Heath, Siim Sõber, Afshin Parsa, Jian'an Luan, Pankaj Arora, Abbas Dehghan, Feng Zhang, Gavin Lucas, Andrew A. Hicks, Anne U. Jackson, John F Peden, Toshiko Tanaka, Sarah H. Wild, Igor Rudan, Wilmar Igl, Yuri Milaneschi, Alex N. Parker,

Cristiano Fava, John C.Chambers, Ervin R. Fox,Meena Kumari,Min Jin Go, Pim van der Harst, Wen Hong Linda Kao, Marketa Sjögren, D. G. Vinay, Myriam Alexander, Yasuharu Tabara, Sue Shaw- Hawkins, Peter H. Whincup, Yongmei Liu, Gang Shi, Johanna Kuusisto, Bamidele Tayo,Mark Seielstad, Xueling Sim, Khanh-Dung Hoang Nguyen, Terho Lehtimäki, Giuseppe Matullo, Ying Wu, Tom R. Gaunt, N. Charlotte Onland-Moret, Matthew N. Cooper, Carl G.P. Platou, Elin Org, Rebecca Hardy, Santosh Dahgam, Jutta Palmen, Veronique Vitart, Peter S. Braund, Tatiana Kuznetsova, Cuno S.P.M. Uiterwaal, Adebowale Adeyemo, Walter Palmas, Harry Campbell, Barbara Ludwig, Maciej Tomaszewski, Ioanna Tzoulaki, Nicholette D. Palmer, ThorAspelund,Melissa Garcia, Yen-Pei C. Chang, Jeffrey R. O'Connell, Nanette I. Steinle, Diederick E. Grobbee, Dan E. Arking, Sharon L. Kardia, Alanna C. Morrison, Dena Hernandez, Samer Najjar, Wendy L.McArdle, David Hadley, Morris J. Brown, JohnM.Connell, Aroon D. Hingorani, Ian N.M. Day, Debbie A. Lawlor, John P. Beilby, Robert W. Lawrence, Robert Clarke, Rory Collins, Jemma C. Hopewell, Halit Ongen, Albert W. Dreisbach, Yali Li, J.H. Young, Joshua C. Bis, Mika Kähönen, Jorma Viikari, Linda S. Adair, Nanette R. Lee, Ming-Huei Chen, Matthias Olden, Cristian Pattaro, Judith A. Hoffman Bolton, Anna Köttgen, Sven Bergmann, Vincent Mooser, Nish Chaturvedi, Timothy M. Frayling, Muhammad Islam, Tazeen H. Jafar, Jeanette Erdmann, Smita R. Kulkarni, Stefan R. Bornstein, Jürgen Grässler, Leif Groop, Benjamin F. Voight, Johannes Kettunen, Philip Howard, Andrew Taylor, Simonetta Guarrera, Fulvio Ricceri, Valur Emilsson, Andrew Plump, Inês Barroso, Kay-Tee Khaw, Alan B. Weder, Steven C. Hun, Yan V. Sun, Richard N. Bergman, Francis S. Collins, Lori L. Bonnycastle, Laura J. Scott, Heather M. Stringham, Leena Peltonen, Markus Perola, Erkki Vartiainen, Stefan-Martin Brand, Jan A. Staessen, Thomas J. Wang, Paul R. Burton, Maria Soler Artigas, Yanbin Dong, Harold Snieder, Xiaoling Wang, Haidong Zhu, Kurt K. Lohman, Megan E. Rudock, Susan R. Heckbert, Nicholas L. Smith, Kerri L. Wiggins, Ayo Doumatey, Daniel Shriner, Gudrun Veldre, Margus Viigimaa, Sanjay Kinra, Dorairajan Prabhakaran, Vikal Tripathy, Carl D. Langefeld, Annika Rosengren, Dag S. Thelle, Anna Maria Corsi, Andrew Singleton, Terrence Forrester, Gina Hilton, Colin A. McKenzie, Tunde Salako, Naoharu Iwai, Yoshikuni Kita, Toshio Ogihara, Takayoshi Ohkubo, Tomonori Okamura, Hirotsugu Ueshima, Satoshi Umemura, Susana Eyheramendy, Thomas Meitinger, H.-Erich Wichmann, Yoon Shin Cho, Hyung-Lae Kim, Jong-Young Lee, James Scott, Joban S. Sehmi, Weihua Zhang, Bo Hedblad, Peter Nilsson, George Davey Smith, Andrew Wong, Narisu Narisu, Alena Stančáková, Leslie J. Raffel, Jie Yao, Sekar Kathiresan, Chris O'Donnell, Stephen M. Schwartz, M. Arfan Ikram, W.T. Longstreth Jr., Thomas H. Mosley, Sudha Seshadri, Nick R.G. Shrine, Louise V.Wain, Mario A. Morken, Amy J. Swift, Jaana Laitinen, Inga Prokopenko, Paavo Zitting, Jackie A. Cooper, Steve E. Humphries, JohnDanesh, Asif Rasheed, Anuj Goel, Anders Hamsten, HughWatkins, Stephan J.L. Bakker, Wiek H. vanGilst, Charles S. Janipalli, K. RadhaMani, Chittaranjan S. Yajnik, Albert Hofman, Francesco U.S. Mattace-Raso, Ben A. Oostra, Ayse Demirkan, Aaron Isaacs, Fernando Rivadeneira, Edward G. Lakatta, Marco Orru, Angelo Scuteri, Mika Ala-Korpela, Antti J. Kangas, Leo-Pekka Lyytikäinen, Pasi Soininen, Taru Tukiainen, Peter Würtz, Rick Twee-Hee Ong, Marcus Dörr, Heyo K. Kroemer, Uwe Völker, Henry Völzke, Pilar Galan, Serge Hercberg, Mark Lathrop, Diana Zelenika, Panos Deloukas, Massimo Mangino, Tim D. Spector, Guangju Zhai, James F. Meschia, Michael A. Nalls, Pankaj Sharma, Janos Terzic, M. J. Kranthi Kumar, Matthew Denniff, Ewa Zukowska-Szczechowska, Lynne E. Wagenknecht, F. Gerald R. Fowkes, Fadi J. Charchar, Peter E.H. Schwarz, Caroline Hayward, Xiuqing Guo, Charles Rotimi, Michiel L. Bots, Eva Brand, Nilesh J. Samani, Ozren Polasek, Philippa J. Talmud, Fredrik Nyberg,

Diana Kuh, Maris Laan, Kristian Hveem, Lyle J. Palmer, Yvonne T. van der Schouw, Juan P. Casas, Karen L. Mohlke, Paolo Vineis, Olli Raitakari, SanthiK. Ganesh, Tien Y. Wong, E Shyong Tai, Richard S. Cooper, Markku Laakso, Dabeeru C. Rao, Tamara B. Harris, Richard W. Morris, Anna F. Dominiczak, Mika Kivimaki, Michael G.Marmot, Tetsuro Miki, Danish Saleheen, Giriraj R. Chandak, Josef Coresh, Gerjan Navis, Veikko Salomaa, Bok-GheeHan, Xiaofeng Zhu, Jaspal S. Kooner, OlleMelander, Paul M. Ridker, Stefania Bandinelli, Ulf B. Gyllensten, Alan F. Wright, James F. Wilson, Luigi Ferrucci, Martin Farrall, Jaakko Tuomilehto, Peter P. Pramstaller, Roberto Elosua, Nicole Soranzo, Eric J.G. Sijbrands, David Altshuler, Ruth J.F. Loos, Alan R. Shuldiner, Christian Gieger, PierreMeneton, Andre G. Uitterlinden, Nicholas J. Wareham, Vilmundur Gudnason, Jerome I. Rotter, Rainer Rettig, Manuela Uda, David P. Strachan, Jacqueline C.M. Witteman, Anna-Liisa Hartikainen, Jacques S. Beckmann, Eric Boerwinkle, Ramachandran S. Vasan, Michael Boehnke, Martin G. Larson, Marjo-Riitta Järvelin, Bruce M. Psaty, Gonçalo R Abecasis, Aravinda Chakravarti, Paul Elliott, Cornelia M. van Duijn, Christopher Newton-Cheh, Daniel Levy, Mark J. Caulfield, Toby Johnson, the CARDIoGRAMconsortium, the CKDGen Consortium, the KidneyGen Consortium, the EchoGen consortium, and the CHARGE-HF consortium.

Members of the NABEC / UKBEC Consortium *(chapter 2.1)*

Michael A. Nalls, Dena G. Hernandez, Mark R. Cookson, Raphael J. Gibbs, John Hardy, Adaikalavan Ramasamy, Alan B. Zonderman, Allissa Dillman, Bryan Traynor, Colin Smith, Dan L. Longo, Daniah Trabzuni, Juan Troncoso, Marcel van der Brug, Michael E. Weale, Richard O'Brien, Robert Johnson, Robert Walker, Ronald H. Zielke, Sampath Arepalli, Mina Ryten, and Andrew Singleton.

PhD Portfolio

Name PhD student: Marjolein J. Peters	PhD period: September 2008 – October 2015
Erasmus MC Department: Internal Medicine	Promotor: Prof.dr. André G. Uitterlinden
Research School: NIHES & MolMed	Supervisor: Dr. Joyce B.J. van Meurs

1. PhD training

	Year	Workload (Hours/ECTS)
General academic skills		
Research Integrity	Jul 2015	0.3 ECTS
Biomedical English Writing and Communication	Dec 2011	4.0 ECTS
Course EndNote	Dec 2008	1 hour
Basis Course Systematic Literature Retrieval and Pubmed	Nov 2008	1 hour
		2 hrs & 4.0 ECTS
Research skills		
Research Management for PhD-students	Dec 2011	1.0 ECTS
Biostatistics and Research Methods (NIHES):		
– Modern Statistical Methods (EP03)	Dec 2009	4.3 ECTS
– Classical Methods for Data-analysis (CC02)	Oct 2009	5.7 ECTS
– Principles of Research in Medicine and Epidemiology (ESP01)	Sep 2009	0.7 ECTS
– Principles of Genetic Epidemiology (ESP43)	Sep 2009	0.7 ECTS
		12.4 ECTS
In-depth courses (e.g. Research school, Medical Training)		
Next Generation Sequencing Course (MGC Leiden)	Mar 2010	24 hours
Nexus Course (MolMed)	Jan 2009	16 hours
Basic Data Analysis on Gene Expression Arrays (MolMed)	Dec 2008	8 hours
Course on SNPs and Human Diseases (MolMed)	Nov 2008	40 hours
Partek Training Course (MolMed)	Oct 2008	16 hours
		104 hours
Presentations		
Identification of *cis*- and *trans*-acting variants in RNA-seq data – W*etenschapsdagen*	Jan 2015	Poster
Identification of *cis*- and *trans*-acting variants in RNA sequencing data: a "CHARGE – ENCODE" collaboration – *CHARGE meeting*	Nov 2014	Oral
Transcriptomics within the CHARGE consortium – *ENCODE meeting*	Jul 2014	Poster
Known and novel RNAs associated with smoking – *Wetenschapsdagen*	Jan 2014	Oral

	Year	Workload (Hours/ECTS)
Presentations *(Continued)*		
Known and novel RNAs associated with smoking – *ASHG*	Nov 2013	Poster
CLEC4A expression levels in blood are associated with Joint Effusion Grades – *OARSI*	Apr 2013	Oral
Identification of CLEC4A gene-expression levels in peripheral blood as a potential biomarker for knee joint effusion – *NCHA meeting*	Feb 2013	Poster
Identification of CLEC4A gene-expression levels in peripheral blood as a potential biomarker for knee joint effusion – *Wetenschapsdagen*	Jan 2013	Poster
Discovery of age-associated transcripts in a meta-analysis of 7,000 blood samples – *ASHG*	Nov 2012	Poster
The transcriptome of the Rotterdam Study: age, sex, BMI and OA asssociated genes Identified – *NCHA meeting*	Mar 2012	Poster
The transcriptome of the Rotterdam Study – *Wetenschapsdagen*	Jan 2012	Poster
The transcriptome of the Rotterdam Study: age- and sex associated genes – *ASHG*	Oct 2011	Poster
Meta-analysis of genome-wide association data implicates the 5p15.2 region to influence chronic widespread pain in women – *EFIC*	Sep 2011	Poster
Transcriptome analyses in the Rotterdam Study – *NCHA meeting*	Mar 2011	Poster
eQTL analysis in the Rotterdam Study – *CHARGE meeting*	Feb 2011	Oral
The genetics of pain – *Wetenschapsdagen*	Jan 2011	Oral
Meta-analysis of genome-wide association data implicates the 5p15.2 region to influence chronic widespread pain in women – *OARSI*	Sep 2010	Poster
Quality control for large-scale high throughput mRNA isolation and transcriptome analysis of subjects of the Rotterdam Study – *ESHG*	Jun 2010	Poster
From DNA variation to function: gene expression analyss in the Rotterdam Study – *CHARGE meeting*	Apr 2010	Oral
Large-scale high throughput mRNA isolation and transcriptome analysis of subjects of the Rotterdam Study – *Wetenschapdagen*	Jan 2010	Poster
Genome-wide CNV associations studies – *CHARGE meeting*	Oct 2009	Oral
Software methods used for genome wide analyses of copy number polymorphisms in the Rotterdam Study – *Wetenschapsdagen*	Jan 2009	Poster

	Year	Workload (Hours/ECTS)
(Inter)national conferences		
Wetenschapsdagen Antwerpen, Belgium	Jan 2015	16 hours
Wetenschapsdagen Antwerpen, Belgium	Jan 2014	16 hours
American Society of Human Genetics ASHG congress Boston, USA	Nov 2013	40 hours
Osteoarthritis Research Society International Congress Philadelphia, USA	Apr 2013	20 hours
NCHA meeting Den Haag, the Netherlands	Feb 2013	16 hours
Wetenschapsdagen Antwerpen, Belgium	Jan 2013	16 hours
American Society of Human Genetics ASHG congress San Francisco, USA	Nov 2012	40 hours
NVHG najaarscongres Papendal, the Netherlands	Sep 2012	16 hours
NCHA meeting Amersfoort, the Netherlands	Mar 2012	16 hours
Wetenschapsdagen Antwerpen, Belgium	Jan 2012	16 hours
American Society of Human Genetics ASHG congress Montreal, Canada	Oct 2011	40 hours
European Pain Federation EFIC congress Hamburg, Germany	Sep 2011	32 hours
European Human Genetics Conference Amsterdam, the Netherlands	May 2011	16 hours
NCHA meeting Amersfoort, the Netherlands	Mar 2011	16 hours
Molmed Day Rotterdam, the Netherlands	Jan 2011	8 hours
Wetenschapsdagen Antwerpen, Belgium	Jan 2011	16 hours
Osteoarthritis Research Society International Congress Brussels, Belgium	Sep 2010	20 hours
European Human Genetics Conference Gothenburg, Sweden	Jun 2010	20 hours
Advances in Genomics Symposium Gent, Belgium	Jan 2010	8 hours
European Calcified Tissue Society Vienna, Austria	May 2009	40 hours
European Human Genetics Conference Vienna, Austria	May 2009	20 hours
Genomic Variation in Health and Disease Cambridge, UK	Mar 2009	32 hours
		480 hours

	Year	Workload (Hours/ECTS)
Seminars and workshops		
CHARGE meeting Washington, USA	Nov 2014	24 hours
ENCODE consortium meeting Stanford, USA	Jul 2014	24 hours
CHARGE meeting Los Angeles, USA	Jan 2014	24 hours
CHARGE meeting Rotterdam, the Netherlands	Jun 2013	16 hours
CHARGE meeting Reykjavik, Iceland	May 2012	16 hours
CHARGE meeting Boston, USA	Feb 2011	16 hours
CHARGE meeting Houston, USA	Apr 2010	16 hours
CHARGE meeting Washington, USA	Oct 2009	16 hours
CHARGE meeting Rotterdam, the Netherlands	Apr 2009	16 hours
		168 hours
Other		
Referee activities for various international scientific journals	2008-2015	
Three months fellowship in Berkeley, USA	2014	3 months
Setting up the pain measurements in the Rotterdam Study	2012-2013	~60 hours
Interviews with participants of the Rotterdam Study	2009-2011	~250 hours
Scoring of X-rays on OA features in the Rotterdam Study	2009-2010	~100 hours
Setting up the GWAS QC pipeline of the Rotterdam Study	2008-2009	~300 hours

	Year	Workload (Hours/ECTS)
2. Teaching activities		
Lecturing		
– Epigenetics in the Rotterdam Study (oral at Genomics in Molecular Medicine Course)	Aug 2015	8 hours
– Transcriptomics in the Rotterdam Study (oral at Genomics in Molecular Medicine Course)	Aug 2015	8 hours
– RNA sequencing (oral at the NIHES course: An introduction to the analysis of the next-generation sequencing data – *GE13*)	Apr 2015	8 hours
– From DNA variation to function: gene expression analysis in consortia (oral at Genomics in Molecular Medicine Course)	Aug 2010	8 hours
		32 hours
Supervising practicals		
– NIHES Genomics in Molecular Medicine – Computer Practicals	Aug 2015	8 hours
– MOLMED SNPs and Human Diseases – Computer Practicals	Nov 2014	8 hours
– MOLMED SNPs and Human Diseases – Computer Practicals	Nov 2013	8 hours
– MOLMED SNPs and Human Diseases – Computer Practicals	Nov 2012	8 hours
– NIHES Genomics in Molecular Medicine – Computer Practicals	Aug 2012	8 hours
– MOLMED SNPs and Human Diseases – Computer Practicals	Nov 2011	8 hours
– MOLMED SNPs and Human Diseases – Computer Practicals	Nov 2010	8 hours
– NIHES Genomics in Molecular Medicine – Computer Practicals	Aug 2010	8 hours
– MOLMED SNPs and Human Diseases – Computer Practicals	Nov 2009	8 hours
– MOLMED SNPs and Human Diseases – Computer Practicals	Nov 2008	8 hours
– Hugo Course – Computer Practicals	Oct 2008	2 hours
		82 hours

	Year	Workload (Hours/ECTS)
Supervising Students		
– Joost Verlouw – Bachelor thesis – Avans Hogeschool Breda	Jul 2015	4.5 months
– Annelies Smouter – Bachelor thesis – Hogeschool Leiden	Jun 2013	9 months
– Dennis Schmitz – Bachelor thesis – Avans Hogeschool Breda	Aug 2013	9 months
– Suzanne de Maat – Master thesis – Universiteit Utrecht	Jan 2010	3 months
Other		
Organization of the Labday for department of Internal Medicine	May 2009	10 hours

About the author

Marjolein (Maria Josephine) Peters was born in Dedemsvaart (municipality Avereest) on March 30th 1984. In 2002, she completed her secondary school at the "Emmauscollege" in Rotterdam, and started her study Life Science & Technology at two universities: Delft University of Technology and Leiden University. This study focuses on fundamental and applied knowledge from the disciplines of Biology, Chemistry, Physics, Technology, Informatics, Pharmacology and Mathematics, to unlock the secrets of the human cell: the building block of life. In 2008, Marjolein achieved her Master of Science degree in Life Science & Technology at the Delft University of Technology. Her graduation thesis was on "how to prioritize genome-wide association study results by fusing genomic and functional information", performed at the department of Internal Medicine at the Erasmus Medical Center in Rotterdam. During her master, she also completed an industrial internship at DNage (part of Pharming Group NV). This internship was about ageing mice: next to wet lab experiments, she developed scripts providing research colleagues the opportunity to easily measure retina thickness and liver nuclei sizes.

In September 2008, Marjolein started her PhD research described in this thesis at the Genetic Laboratory of the department of Internal Medicine at the Erasmus Medical Center in Rotterdam, under supervision of Prof.dr. André G. Uitterlinden and Dr. Joyce B.J. van Meurs. In February 2014, she received the CHARGE Golden Tiger Award for Working Group Leadership. In June 2014, she received a Visiting Fellow Grant and visited the research groups of Dr. James B. Brown (Lawrence Berkeley National Laboratory) and Prof. Peter J. Bickel (University of California, Berkeley) for three months. These research groups are involved in the ENCODE data analysis center, and together they optimized different methods to identify *cis*- and *trans*-acting variants in RNA sequencing data of the Rotterdam Study. Additionally, Marjolein investigated the impact of these variants in non-coding RNAs on genome-wide gene expression profiles.

In January 2016, Marjolein started working at CBS (Statistics Netherlands) as a statistical researcher.

Over de auteur

Marjolein (Maria Josephine) Peters is geboren op 30 maart 1984 in Dedemsvaart (gemeente Avereest). Ze groeide op in Nieuwerkerk aan den IJssel en behaalde in 2002 haar VWO diploma aan het Emmauscollege in Rotterdam. Ze studeerde Life Science & Technology aan twee universiteiten: Technische Universiteit Delft en Universiteit Leiden. Deze studie draait om fundamenteel en toegepast onderzoek naar de levende cel: de bouwsteen van alle organismen. Dit vereist kennis en vaardigheden van vakgebieden als biologie, scheikunde, natuurkunde, techniek, informatica, farmacologie en wiskunde. In 2008 heeft Marjolein haar master diploma behaald aan de Technische Universiteit in Delft. Haar afstudeeronderzoek ging over het prioriteren van genoom-wijde associatie studie resultaten door het combineren van genomische en functionele informatie. Deze onderzoeksstage vond plaats bij de afdeling Interne Geneeskunde aan het Erasmus Medisch Centrum in Rotterdam. Ook voltooide Marjolein een bedrijfsstage bij DNage (onderdeel van Pharming Group NV). Deze stage ging over verouderende muizen: naast labexperiment ontwikkelde Marjolein scripts voor collega onderzoekers om de retina-dikte in het oog en de celkern grootte van levercellen te meten.

Op 1 september 2008 is Marjolein gestart met haar promotieonderzoek bij het Genetisch Laboratorium van de afdeling Interne Geneeskunde aan het Erasmus Medisch Centrum in Rotterdam, onder supervisie van Prof.dr. André G. Uitterlinden en Dr. Joyce B.J. van Meurs. In februari 2014 ontving Marjolein de CHARGE Golden Tiger Award voor Werkgroep Leiderschap. In juni 2014 ontving zij een CHARGE Visiting Fellow Grant, waarmee zij drie maanden naar de Verenigde Staten is geweest: ze werkte bij de onderzoeksgroepen van Dr. James B. Brown (Lawrence Berkeley National Laboratory) en Prof. Peter J. Bickel (University of California, Berkeley). Deze onderzoeksgroepen zijn onderdeel van het ENCODE data analyse centrum. Marjolein heeft hier gewerkt aan de optimalisatie van verschillende methodes om *cis*- en *trans*-eQTLs te identificeren in RNA sequencing data van de Rotterdam Studie. Daarnaast heeft Marjolein bestudeerd of *trans*-eQTL SNPs in "long non-coding RNAs" invloed hebben op genoom-wijde genexpressie levels.

Op 15 januari 2016 is Marjolein begonnen als statistisch onderzoeker bij het Centraal Bureau voor de Statistiek in Den Haag.

Dankwoord

Eén ding is zeker: promoveren doe je niet alleen! Daarom wil ik op deze plaats iedereen bedanken die de afgelopen jaren, in welke vorm dan ook, een bijdrage heeft geleverd. Ik heb geprobeerd niemand te vergeten. Mocht dat toch gebeurd zijn, dan is dat onbewust en zeker niet persoonlijk bedoeld. Hierbij dan ook voor diegene: dankjewel!

Twee belangrijke mensen die ik graag als eerste persoonlijk wil bedanken zijn mijn promotor Prof.dr. André G. Uitterlinden en mijn copromotor Dr. Joyce B.J. van Meurs:
Beste André, al in 2007 kwam ik voor mijn onderzoeksstage bij jouw genetisch lab terecht. Je vroeg me hoeveel DNA samples er in een plaat passen en ik had werkelijk geen idee! Gelukkig mocht ik toch stage komen lopen en kon ik na mijn stage blijven voor mijn promotieonderzoek. De tijd is voorbij gevlogen! Ik ben je erg dankbaar voor alle kansen die je mij gegeven hebt: ik kreeg alle ruimte om mijzelf te ontwikkelen, als wetenschapper, maar ook als persoon. Bovendien mocht ik elk jaar naar internationale congressen en kreeg ik de kans om drie maanden in Berkeley te werken: een fantastische ervaring! Samen met Joyce heb jij mij vertrouwen en kansen gegeven. Hartelijk dank voor alles.
Beste Joyce, gedurende mijn promotie waren wij een goed team. Je bent recht voor zijn raap en dat waardeer ik enorm. Door jouw kritische vragen en goede suggesties heb ik ontzettend veel geleerd. Bovendien ben ik het helemaal met je eens dat onderzoek gewoon goed moet zijn. De snelheid waarmee jij reageert op vragen en manuscripten, waar je ook bent, welke dag het ook is, dat vind ik echt bewonderenswaardig. Dankjewel voor alles! Als je ooit nog een verzoekje hebt, vraag het me gerust :)

Ook wil ik graag de mensen van de leescommissie bedanken: Prof.dr. Joost H. Gribnau, Prof.dr. Lude H. Franke, en Prof.dr. Oscar H. Franco.
Beste Joost, dankjewel voor het plaatsnemen in mijn leescommissie. Ik heb persoonlijk veel bewondering voor jouw werk. Ik denk dat het gebruik van iPS-cellen, die we vervolgens weer kunnen differentiëren tot weefselspecifieke cellen, erg belangrijk zal worden om de regulatie van genexpressie en DNA methylering beter te begrijpen.
Beste Lude, dankjewel voor alle inzichten die jij mij gegeven hebt in "big data analyse". Door mijn bezoekjes aan jouw groep heb ik veel geleerd over de methodes die jullie ontwikkeld hebben. Onze samenwerking was erg vruchtbaar en heeft een aantal mooie papers opgeleverd. Dankjewel daarvoor.
Dear Oscar, I am very happy that you joined the reading committee. Many thanks for reading this thesis, and especially for your interest in my work during CHARGE meetings and MOLEPI meetings.

Daarnaast wil ik Prof.dr. Jan H.J. Hoeijmakers, Dr. Ingrid Meulenbelt en Dr. James B. Brown bedanken voor het plaatsnemen in de commissie.
Professor Hoeijmakers, hartelijk dank voor uw bereidheid om zitting te nemen in mijn promotiecommissie en te opponeren bij de verdediging van mijn proefschrift.

Beste Ingrid, bedankt voor de fijne samenwerking. Je was altijd bereid onze bevindingen te repliceren en kritisch mee te denken over de impact van onze bevindingen. Dankjewel!
Dear Ben, thank you so much for giving me the opportunity to visit your lab in Berkeley. I really enjoyed my period in the Bay Area! I learned a lot from you and your team, both in research skills and in personal skills. It is great that you are willing to join my PhD committee. I loved your enthusiasm about my work, and I would like to finish our eQTL manuscript together!

Lieve Hanneke en Lisette, dankjulliewel dat jullie mijn paranimfen willen zijn. Ik heb ontzettend veel plezier met jullie gehad de afgelopen jaren. We kunnen (nog altijd) kletsen over onderzoek, nieuwe analyses, problemen met scripts, etc. Maar we kunnen nog veel beter kletsen over andere dingen! Ik heb veel van jullie geleerd en ik moest ook erg wennen toen jullie vertrokken van het genetisch lab voor een nieuwe uitdaging.
Hanneke, jij bent alweer in 2012 gepromoveerd en je bent vervolgens bij Pfizer gaan werken. Iedere keer ben je nieuwsgierig naar hoe het staat met mijn proefschrift en nu kan ik het echt zeggen: het is klaar! Nogmaals dank voor al je hulp en ik vind dat we de frequentie van onze etentjes wel weer een beetje kunnen opvoeren!
Lisette, jij bent nog iets langer gebleven bij het genetisch lab. Na je promotie in 2009 ben je verder gegaan als postdoc, maar vorig jaar heb ook jij de stoute schoenen aangetrokken en ben je bij MRC-Holland gaan werken. Dankjewel voor al je hulp en de vele keren dat je een hotelkamer met mij wilde delen. Het was de laatste maandjes erg stil in onze kamer...
Lieve meiden, dankjulliewel voor alle gezelligheid. Het was een mooie tijd. Ik hoop dat we nog lang vriendinnen zullen blijven!

Eline, ook jou wil ik graag persoonlijk bedanken. Je hebt zo ontzettend veel voor mij gedaan voor de afronding van mijn promotie. Het was erg fijn dat jij veel administratieve klusjes uit handen nam. Bovendien kon ik altijd bij je terecht voor een praatje. Dankjewel daarvoor!

Natuurlijk wil ik ook graag mijn collega's van het genetisch lab bedanken:
Fernando, heel erg bedankt voor je hulp bij mijn statistische problemen en voor het stellen van kritische vragen tijdens werkbesprekingen. Ik hoop dat we elkaar blijven zien bij onze loopgroep in Nesselande. Ik heb echt respect voor jouw doorzettingsvermogen en ik hoop die marathon zelf ooit nog eens te lopen... ik laat het je weten, goed?
Michael, jij kwam elke dag even om het hoekje kijken of ik tijd had voor een praatje. Altijd gezellig! Dankjewel voor je wetenschappelijke input: je hulp met betrekking tot de DNA methylering-methodes en -analyses was heel fijn! En dankjewel voor je vriendschap. Ik hoop dat we ons "rondje bruggen" en de avondjes "Rotterdam Running Crew" zullen blijven doen.
Mila en Pascal, dankjulliewel voor het sequencen van al die samples en het runnen van al die arrays. Jullie waren fijne collega's. Mila, ik heb veel van je geleerd op het lab. Je bent super attent en denkt altijd aan iedereen. Dankjewel voor alles!

Marijn, super bedankt voor je hulp bij het bouwen van onze website (*https://trap.erasmusmc.nl/*) waarmee collega-onderzoekers de biologische leeftijd van hun samples kunnen voorspellen op basis van genexpressie levels in het bloed. Ik ben er (nog altijd) heel erg trots op! Ook bedankt voor het fixen van alle computer- en serverproblemen. Dankjewel!
Marjolein en Ester, bedankt voor alle koffie en thee momentjes. Het was fijn om te kunnen sparren met jullie meiden. Soms met een lach en soms met een traan. We zijn nu alle drie bijna klaar: ik ben trots op ons! We did it!
Annemieke, dankjewel voor de gezellige tijd bij het genetisch lab. Lopen we binnenkort nog een keer een wedstrijdje samen?
Cindy, Jeroen, Annelies, Jia-Lian, Djawad en Ling, dankjewel voor jullie input tijdens werkbesprekingen. Dank voor de gezellige congressen, lunches en labuitjes. Ik wens jullie heel veel succes en plezier bij het afronden van jullie proefschrift.
Carolina, Pooja, Fatimeh, Katerina, Fjorda, Martha, Karol, and Liz, thanks for your input during my workdiscussions. Additionally, thanks for the nice conference meetings and lab days we joined together. Carolina and Pooja, good luck with finishing your PhD thesis! You are almost done too!
Carola, dank voor jouw input bij werkbesprekingen en discussies. Jouw klinische kijk geeft altijd een extra dimensie aan ons werk.
Ramazan, Robert, Linda, Stephan, Saskia, Anis, Joost (Verburg), Anke, Sarah, Manoushka, Arnoud en Iris: jullie ook bedankt voor de gezellige tijd bij het genetisch lab.
Mijn studenten: Suzanne, Annelies en Joost (Verlouw). Heel erg bedankt voor jullie bijdrage aan dit proefschrift. Ik hoop dat ik jullie iets van mijn enthousiasme voor het onderzoek heb kunnen meegeven. Joost, jij bent nu niet langer student bij ons lab. Heel veel succes met het bouwen en onderhouden van alle sequencing pipelines. En natuurlijk wil ik ook alle andere studenten van ons lab bedanken.

Verder wil ik iedereen van de vijfde verdieping bedanken voor de gezelligheid tijdens de koffies (met ....), de borrels, werkbesprekingen, de Wetenschapsdagen in Antwerpen en onze jaarlijkse labdag.

Ook wil ik graag de deelnemers van de Rotterdam Studie (ERGO) en de andere internationale cohort studies bedanken. Natuurlijk wil ik ook de mensen van het ERGO onderzoekscentrum graag bedanken voor hun werk, omdat zonder jullie dit proefschrift nooit geschreven had kunnen worden. Frank van Rooij, Jeannette Vergeer en Bernadette van Ast, jullie wil ik graag apart noemen en bedanken. Frank, voor je hulp bij het verkrijgen van alle benodigde fenotype files. Gelukkig heb je sinds de introductie van de ERGO wiki een stuk minder werk aan mij. Jeannette en Bernadette, een speciaal dankjewel voor jullie, voor het isoleren van al die RNA samples!

Ook wil ik graag mijn collega's van de epidemiologie bedanken. Abbas en Janine, Mohsen, Symen, Paul, en Daan; dankjulliewel voor de fijne samenwerking.

I also owe many thanks to all co-investigators on the projects I have been participating in. In particular, all those involved in the CHARGE gene expression working group, especially Dr. Andrew Johnson and Prof.dr. Chris O'Donnell. Thank you so much for giving me the opportunity to co-lead the gene expression working group, and more specifically our massive age project: getting our manuscript accepted in a high impact journal was very challenging. I am very proud to have our age manuscript published in Nature Communications, and it gained a lot of international media attention. Of course, I am also very proud of the other CHARGE gene expression manuscripts we published together!

Many thanks to Prof.dr. Bruce Psaty and the CHARGE research steering committee for creating a great research climate for young researchers to work in. Bruce, your support for the CHARGE fellowship gave me a unique experience in Berkeley!

Many thanks to Dr. Luke Pilling, Dr. Karen Conneely, Dr. Claudia Schurmann, Dr. Katharina Schramm, Dr. Brian Chen, Dr. Tanxiao Huan, Dr. Joseph Powell, Dr. Eva Reinmaa, Dr. Johannes Kettunen, Dr. George Sutphin, Dr. Sasha Zhernakova, Dr. Tõnu Esko, and their supervisors, for the very nice collaborations within CHARGE. The collaborative, friendly, and open nature of these projects has been very exiting: it is only through this that we can achieve the best science. Thanks for the collaborations!

Many thanks to Prof.dr. Peter J. Bickel, who welcomed me at the University of California. Peter, thanks for your interest and confidence in my work. I loved your stories about your work and family, and I learned a lot from you and Ben about statistical tools. Additionally, I would like to thank Marcus, Omid, Mu, Nathan, Taly, Sharmodeep, and Rachel. Thanks for your help during my visit. Thanks for the very useful Summer Journal Club meetings, and thanks for the nice lunches in the garden, the dinners, the birthday parties, and the barbecues.

Ook de collega's in Groningen, Leiden en Utrecht wil ik graag bedanken. Harm-Jan, dankjewel voor het fixen van alle bugs in de pijplijn. Je zult wel gek geworden zijn van al mijn vragen en problemen in de test fase. Maar we hebben uiteindelijk een mooi artikel gepubliceerd samen! Ik wens je heel veel succes met je post-doc plek in Boston. Ik hoop dat je ondertussen je draai een beetje kunt vinden daar.
Yolande, dankjewel voor onze samenwerking op het gebied van expressie en artrose. Ik hoop dat jij ook snel een nieuwe fijne plek gevonden hebt.
Bas en Peter-Bram, dankjulliewel voor de samenwerking. Ik hoop dat er nog vele mooie artikelen mogen volgen uit de BBMRI samenwerking.
Vered and Marius, thanks for our collaboration. Hopefully, our manuscript regarding quadratic regression will be accepted soon. I will keep my fingers crossed!
Hanneke en Annemieke, dankjulliewel voor de fijne samenwerking. Jullie functionele werk in muizen was een mooie toevoeging aan ons manuscript over chronische pijn.

Lieve vrienden, vriendinnen, familie en schoonfamilie, bedankt voor jullie interesse in mijn werk. En dankjulliewel voor de gezellige afspraken en gesprekken over niet-werk gerelateerde dingen!

Lieve Joost en Michèle, ik ben heel blij dat jullie er altijd voor ons zijn. Gewoon een kop koffie of een heerlijke vakantie samen, in goede en in minder goede tijden. Dankjewel voor alles! Een dikke knuffel voor jullie allebei!

Lieve pap en mam, dankjewel dat jullie er altijd voor mij zijn. We hebben aardig wat meegemaakt de afgelopen jaren en jullie deur stond (en staat) altijd wagewijd open. Ook nu zijn het weer roerige tijden en ik hoop dan ook dat we er altijd voor elkaar kunnen zijn. Jullie hebben me altijd gesteund en aangemoedigd en jullie hebben me geleerd mijzelf te zijn. Dankjulliewel voor alles! Ik hou van jullie.

Lieve Thomas, dankjewel dat jij er altijd voor mij bent. Je zorgt heel goed voor mij en je kunt me altijd aan het lachen krijgen. Bedankt voor jouw liefde en steun. Ik hou van jou!