# Multi-omics strategies for detecting gene-environment interactions

Patrick Deelen

# Multi-omics strategies for detecting gene-environment interactions

**Patrick Deelen**

Patrick Deelen

Multi-omics strategies for detecting gene-environment interactions

First printing, 2019

Printed by: Ipskamp Printing

Cover design by: Eline Deelen

Printing of this thesis was financially supported by: University of Groningen, University Medical Center Groningen, the Groningen University Institute for Drug Exploration (GUIDE) and the Graduate School for Medical Sciences (GSMS), Groningen.



www.patrickdeelen.eu

ISBN: 978-94-034-1600-7 (printed version)
ISBN: 978-94-034-1599-4 (electronic version)

university of groningen

# Multi-omics strategies for detecting gene-environment interactions

## PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. E. Sterken
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Monday 6 May 2019 at 12:45 hours

by


## Patrick Deelen

born on 30 May 1986
Rotterdam, The Netherlands

## Supervisors

Prof. M.A. Swertz
Prof. L.H. Franke
Prof. C. Wijmenga

## Assessment committee

Prof. M.J.H. Kas
Prof. L. Wessels
Prof. A. van Gool

## Paranymphs

Marc Jan Bonder
Pieter B.T. Neerincx

*To Marieke & John*

## Propositions

**1.** 1+1≥2; integration of multiple datasets enables analyses not possible in individual datasets. (this thesis)

**2.** Allelic imbalance of RNA-seq data can reveal regulatory effects of rare pathogenic variants. (this thesis)

**3.** Changes in DNA methylation can reveal the downstream effects of genetic risk factors. (this thesis)

**4.** The environmental component of complex disease development can be mediated through altered gene regulation. (this thesis)

**5.** Regulatory effects of disease-associated variants are not driven by random co-localization. (this thesis)

**6.** Large-scale population transcriptomics can be used to aid the interpretation of diagnostic genome sequencing. (this thesis)

**7.** In the near future, genetic profiling will be requested via a general practitioner and will become part of standard newborn screening.

**8.** High-density molecular profiling, such as transcriptomics, metabolomics, and microbiomics, will become standards tools of medical specialists allowing personalized medicine.

**9.** BBMRI-NL and Lifelines show that large-scale infrastructure projects and biobanking efforts are essential to develop the methods needed for personalized medicine.

**10.** For personalized medicine to be successful, we need to collaborate.

**11.** All countries should have their own biobanks to reflect their genetic diversity and environmental factors.

**12.** Nothing in biology makes sense except in the light of evolution. (Theodosius Dobzhansky)

## Contents

1

# General introduction and outline of the thesis

The DNA of human cells consists of around six billion nucleotide bases of which, on average, 15 million bases differ between individuals. These genetic differences between individuals result in differences in every trait one can think of, including the risk of developing both Mendelian and more complex diseases (Figure 1) [1]. These differences are the result of genetic mutations that occurred in earlier generations and been passed on to offspring ever since. This genetic diversity is an essential part of evolution as genetic variants may have positive, neutral and negative effects on one's overall fitness (i.e. one's ability thrive and reproduce). Selection on fitness has resulted in the emergence of complex species, and it facilitates adaptation to specific circumstances and environments. Adaptive evolution allowed early Europeans to digest cow milk at adult ages [2] and Andeans to live in the low-oxygen

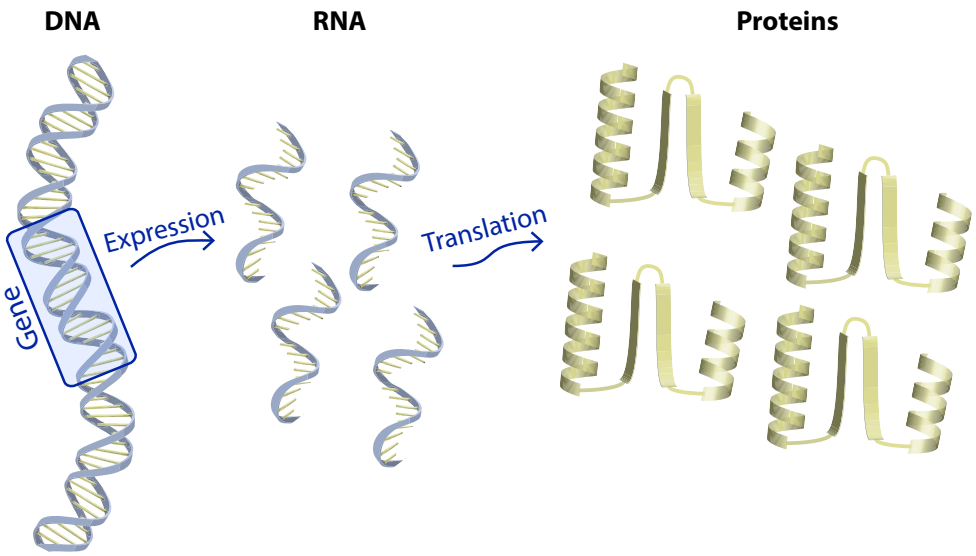



***Figure 1: Genes, expression and genetic variation.*** *DNA can be thought of as a library containing all the information needed to grow and maintain the human body. Almost all cells in our body contain a full copy of this library, which contains the instructions for making the proteins needed for a cell to function. A piece of our DNA that contains instructions to make a protein is called a gene, and it acts like a book in the genetic library. Much like a real library - which cannot be used as a workshop - DNA cannot be used to directly make proteins out of a gene. A cell first needs to make a copy of the gene in the form of RNA in a process called gene expression, and the RNA thus created can be then used as template for the production of proteins. Upon conception, you inherit half a copy of the genetic library from each your parents. However, this copying process is not perfect, and small errors are introduced that lead to genetic variation. Sometimes variants arise within a gene that lead to the protein it produces being different. Or a variant occurs in the region of the DNA that regulates how many RNA copies of a gene a cell should make, and changing the number of RNA copies can affect how much of a protein is produced within a cell. Both these kinds of variants can have an effect on a phenotype or disease risk.*

environment found at high altitudes [3]. However, since the processes that cause mutations are mostly random, they are, by definition, agnostic to the outcome. This means that introduced variants can also have deleterious effects that result in disease.

Many diseases are, at least in part, driven by harmful genetic variants. In genetics, we basically divide diseases in two classes: we discriminate between monogenic or Mendelian diseases, which are caused by variants in a single gene, and complex diseases caused by variants in multiple genes. Since monogenic diseases typically follow the inheritance patterns described by Mendel [4], they are referred to as Mendelian diseases. The inheritance of complex diseases is not as straightforward. One does not inherit a complex disease directly but instead inherits a disease predisposition caused by many different genetic variants [5,6]. Environmental influences such as diet and lifestyle combined with genetic predisposition eventually can cause a complex disease to manifest. Knowledge about the variants that underlie a disease can help us to diagnose individuals, understand the biological mechanisms of a disease, and could eventually aid new drug development.



In the past decade, there have been marked advances in our ability to identify genetic factors that drive both Mendelian and complex diseases. High-throughput DNA sequencing technologies have boosted our ability to identify the genes underlying Mendelian diseases [7]. Over 3,000 genes have now been implicated in Mendelian diseases, and for roughly 25% of the cases with a suspected Mendelian disease the disease-causing variants have been identified [8]. Our understanding of the genetics of complex diseases has been especially driven by the success of genome-wide association studies (GWAS). Since the first GWAS in 2005, more than 40,000 genetic risk factors have been identified for hundreds of complex traits and diseases (Figure 2, Box 1).

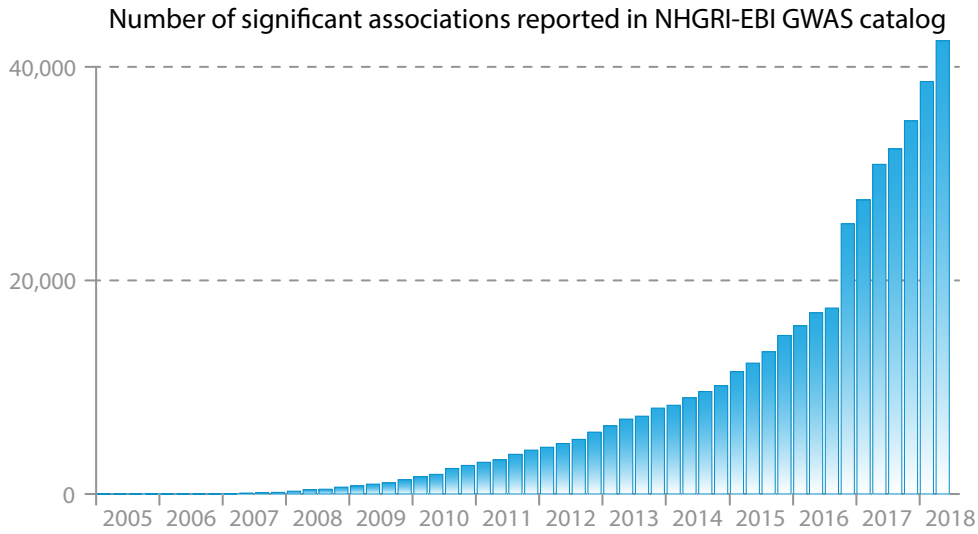



***Figure 2: Yearly growth in significant associations of genetic variants to common phenotypes and diseases (2005-2018).*** *The number of significant (p-value ≤ 5 x $10^{-8}$) GWAS associations reported in the NHGRI-EBI GWAS catalog [9].*

***Box 1: Genome Wide Association Studies (GWAS).*** *Using GWAS it is possible to identify relationships between genetic variation and phenotypic traits or diseases. This is done for many different variants around the genome. The trait studied can be quantitative, like height, or binary, like being healthy vs being affected by a specific disease. For each variant, a GWAS tests if there is a relation to the trait under investigation. A GWAS does not identify the causative variants that are actually responsible for the trait or the increased disease risk, what it does is point to the regions (loci) in the genome that harbor the causative variants. Because the effect of individual variants is typically very small, GWAS studies require a large number of samples to reach sufficient statistical power to establish reliable associations.*

## Measuring genetic variation

There are many techniques that can be used to detect genetic variation, the most comprehensive being DNA sequencing (DNA-seq) technologies. The first sequenced human genome was completed in 2003, and it is estimated that the cost of this project exceeded half a billion dollars [10]. Even with rapid cost reductions since, genome DNA-seq for one individual still costs approximately a thousand US dollars. Given that a GWAS typically require thousands to hundreds of thousands of samples, performing DNA-seq on all these samples is not currently financially feasible.

A more economical alternative to DNA-seq is to make use of genotyping chips. These microarray-chips can be used to determine the genotypes of, typically, between 200,000 to 2,000,000 known polymorphisms, and they are commonly used for GWAS [11]. This is, however, only a small fraction of the tens of millions of variants that are currently known [12,13]. Despite the limited number of variants typed by these chips, microarray-chips have led to the identification of many disease- and trait-associated loci using GWAS. However, it is easy to miss associations if a region on the genome is poorly covered by the used chip and the ability to fine-map the association to a smaller genomic region or putative causative variant using chips is limited.

These limitations can be partially overcome by performing genotype imputation to fill in the gaps in the genotyping chip data using a more densely typed reference panel (Figure 3) [14]. Conceptually, these imputation algorithms work by taking the measured genotypes

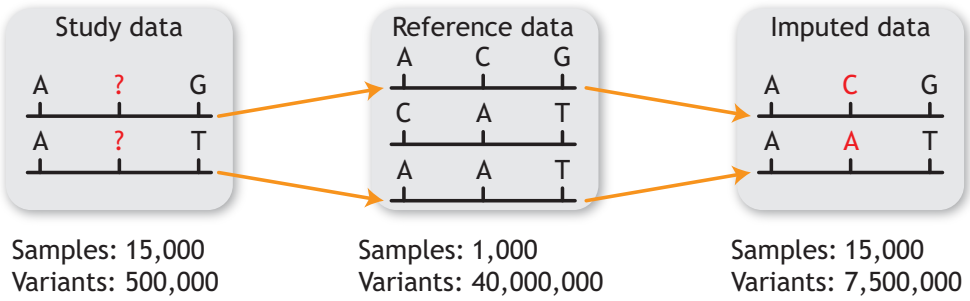



***Figure 3: Simplified explanation of genotype imputation.*** *Conceptually speaking, genotype imputation matches the haplotypes of study subjects to reference haplotypes in order to fill in un-typed variants. In practice, this is performed using complex statistical models to obtain probabilities for the imputed genotypes. The numbers in the figure are illustrative of the gain in high quality genotypes that can be obtained using imputation.*



to identify matching haplotypes in the reference dataset. These matching reference haplotypes are then used to make inferences about the genotypes that have not been directly measured in the study data. While this process does not yield perfect genotypes at an individual level, and is therefore not suited for rare disease diagnostics, it can be used to study complex traits using GWAS. A GWAS can use these imprecise genotypes for detection of associated loci and to zoom in more closely on the region containing the causative variant(s) [15]. Imputation does, however, require careful preparation of the study data and selection of a suitable reference dataset.

## Challenges in the interpretation of genetic variants

Identification of the variants involved is only the first step in understanding a disease and in the eventual development of new treatments. To start with, we need to understand the downstream molecular effects of the identified genetic variants. Even for a variant within a gene, it is difficult to predict what the results will be: the protein coded by this gene might be damaged or completely abolished (loss-of-function) or it may have acquired a new function (gain-of-function) (Figure 4A). Yet genes only make up a small part of our genome. Interestingly, most genetic risk factors associated to complex traits and diseases are found to be located within non-coding regions of the genome. These variants can have a regulatory role – affecting, for instance, the level of expression of a gene – which results in increased or decreased protein production (Figure 4B). Furthermore, since the internal biochemical processes in a cell are highly complex, the downstream effects on the overall working of a cell often remain unclear.

Another complication is that, depending on the environment, it can be favorable or detrimental to carry a specific DNA variant (i.e. allele). For instance, the variant that causes sickle cell anemia occurs more often in Africa because carrying one copy of this variant provides protection against malaria, but having two copies leads to lethal sickle cell disease [16,17], and other examples are known of alleles for which there is evidence for both beneficial and harmful effects [18–21].

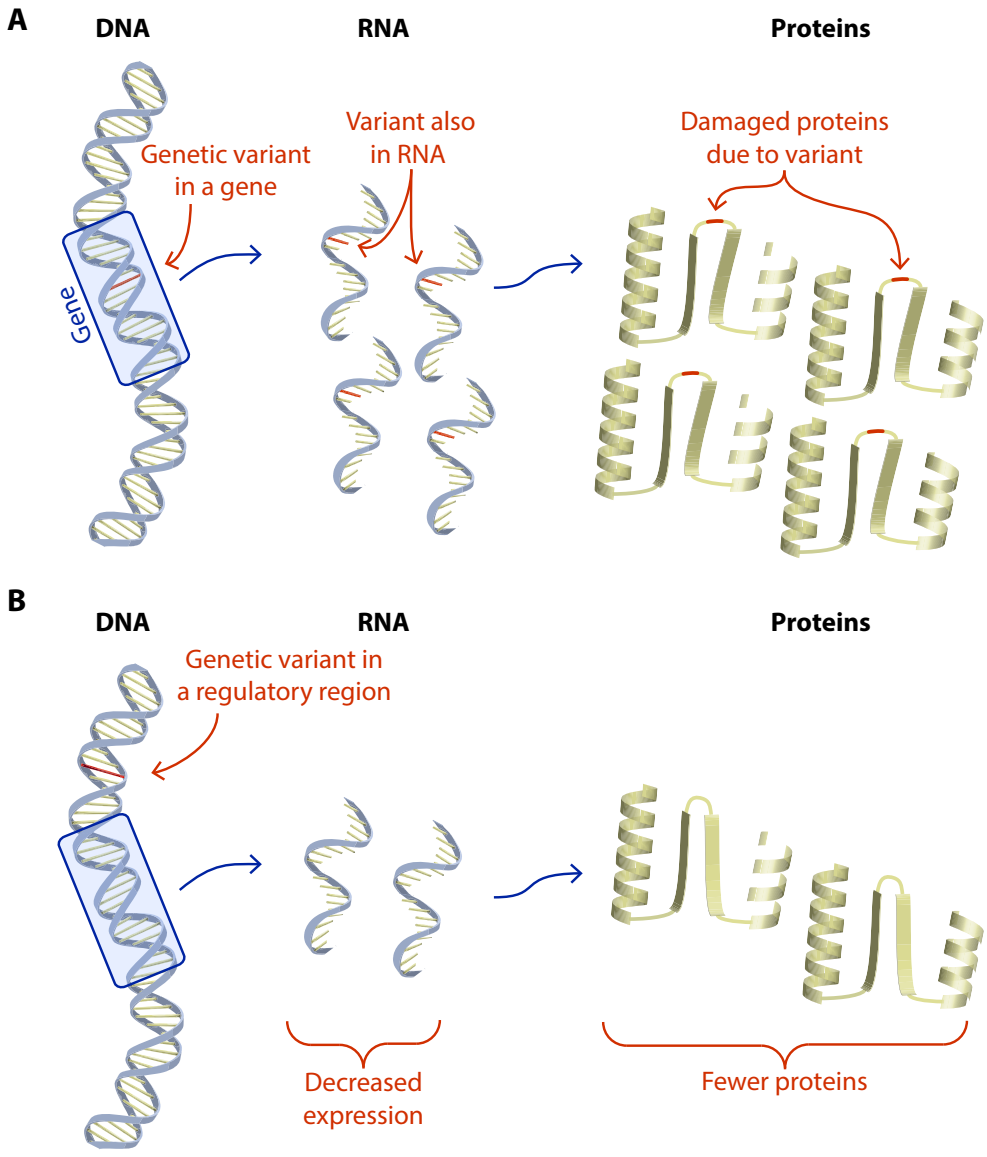



***Figure 4: Effects of genetic variants on protein function and protein production.*** *A) A genetic variant within a gene will also be present in the RNA, resulting in an altered protein. Most Mendelian diseases that have been described are the result of variants within genes. B) Variants outside of genes can still have a regulatory effect on a gene. In the example shown here, decreased expression results in fewer RNA molecules, which in turn result in a reduced number of proteins. The majority of currently identified risk factors for complex diseases have regulatory effects.*

In the case of complex diseases, there is an additional complication. The genetic variants identified as risk factors through GWAS are not necessarily the causative variants responsible for the increased disease risk. This is because neighboring variants in the genome are often inherited together and are therefore strongly correlated. This correlation, which is known as linkage disequilibrium, makes it difficult to pinpoint association signals to specific causative variants. It is thus challenging to link risk factors to genes. Initially the gene closest to the most significantly associated variant was considered the likely causative gene. However, clear examples are now known where this is not the case, even if the most significant variant maps within a gene [22]. Given that the majority of the variants underlying complex diseases are not changing the genes themselves but are instead affecting regulation of genes [23], and that these regulatory effects can act over large distances (sometimes even hundreds of kilobases), alternative strategies are needed to establish the causal gene (or genes) for a given genomic locus.



One strategy to prioritize potential causative genes is by assessing how the genetic risk factors affect the expression levels of nearby or distal genes, the so-called expression quantitative trait loci (eQTLs, Box 2). This has become a common procedure when interpreting genetic association studies and has helped to generate new hypotheses about how individual variants contribute to disease development via altered expression levels [24].

***Box 2: Expression quantitative trait loci (eQTLs).*** *The abundance of gene expression varies from person to person and is therefore considered a quantitative trait. Using the same principles as used for GWAS, it is possible to identify variants that influence the expression abundance. If a variant is correlated to gene expression, it is called an eQTL. We discriminate between* cis*-eQTLs and* trans*-eQTLs. Formally,* cis*-eQTLs only affect genes on the same physical chromosome harboring the causative allele, while* trans*-eQTLs also affect expression of genes on different chromosomes. In practice, we usually call eQTLs* cis *if the variant and the gene are nearby (for instance within a 1 megabase window) and* trans *if variants are further away from a gene. There are also many other types of molecular QTLs such as cytokine QTLs (cQTLs) and methylation QTLs (meQTLs).*

However, these studies also have limitations. For instance, a genetic variant can affect the expression of multiple genes: a phenomenon called pleiotropy. Due to pleiotropy, it is possible for a disease-associated variant to alter the expression levels of genes that are not necessarily related to disease etiology, which makes it difficult to establish which gene is involved in disease development (Figure 5). Here it is important to keep in mind that there might also be multiple causal genes and that it is also possible that none of the genes with altered expression are involved in the disease development.

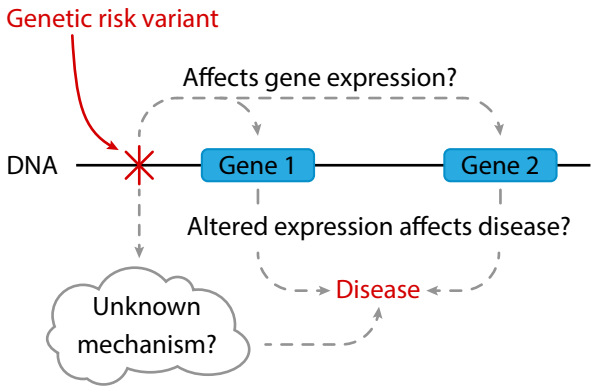



***Figure 5: Complexity of interpretation of genetic risk variants.*** *There are many mechanisms by which a genetic risk factor can alter disease susceptibility. One possible route is that the variant regulates the expression of a nearby gene, and that this modified gene expression in turn affects disease development. It is also possible that multiple genes are regulated by the risk variant, and altered expression of one or multiple of these genes affects the disease. Alternatively, it is also possible that none of the nearby genes are relevant to disease development. It might be that the expression of the genes is altered by the variant, but that this does not have an effect on disease development. In these cases, some other biological process is affected by the variant that is important for disease development.*

## The importance of environmental exposures on complex disease development

Complex diseases are not only driven by genetics: environmental exposures are also important risk factors. A clear example of an environmental effect is the gluten trigger in Celiac disease [25]. Carriers of specific variants in the HLA region who eat gluten are at high risk of developing an autoinflammatory reaction, resulting in Celiac disease. Similar gene-environment (GxE) interactions have also been identified in other complex diseases, for instance the effect of viruses on the development of multiple sclerosis and type 1 diabetes [26,27]. We also know that this happens on a smaller scale, some genetic variants only have an effect on expression levels when triggered by external stimulations [28]. With the recent availability of multi-omics datasets – such as Lifelines Deep [29] or 500FG [30] – that contain information on

multiple molecular levels and various phenotypes for large numbers of individuals, it is now becoming feasible to investigate how environmental triggers affect disease development and molecular phenotypes.

## This thesis

The overall aim of my thesis is to get a better understanding of the molecular consequences of genetic variants and how environmental differences shape these downstream effects. In part 1 of this thesis, I describe work related to genetic reference panels and specialized software to create the infrastructure needed to facilitate genetic research. In part 2, I focus on biological research questions aimed at identifying the downstream effects of genetic variation and the environmental exposures that alter these downstream effects. In part 3, I discuss the work presented in this thesis.



### Part 1 - Infrastructure to enrich and jointly analyze genetic data

In Chapter 2, we show the added value of using the, Genome of the Netherlands (GoNL) data that has been generated by BBMRI-NL, as a reference panel to improve the power of genetic association studies. We also show that the GoNL data can aid studies with non-Dutch European samples.

In Chapter 3, we present Genotype Harmonizer, a software package that allows for easy and more accurate pre-processing of genotype data prior to genotype imputation. Genotype Harmonizer also allows easy integration of different studies in a meta-analysis setting.

### Part 2 - Population based analysis of genetic risk factors

In Chapter 4, we show that it is possible to reliably call genotypes based on publicly available RNA-seq samples, which enables the investigation of tissue-specific effects of variants affecting gene expression. Additionally, we use this data to ascertain the expression effects of rare variants known to cause rare Mendelian diseases.

In Chapter 5, we show how the genetic regulation of cytokines is modulated by pathogenic stimulations. This provides new insights into how environmental factors modulate immunological responses.

In Chapter 6, we gain insights into the regulatory effects of genetic risk factors by investigating the downstream effects on DNA-methylation. Here we used data from several Dutch biobanks integrated by BBMRI-NL. We show that risk variants near transcription factors (genes with a regulatory function) affect the DNA-methylation of regions in the genome under the control of these transcription factors.

In Chapter 7, we present a method that allows us to detect cell-type- or stimuli-dependent regulatory effects using whole blood data by again using data integrated by BBMRI-NL. This also enables us to use these context-dependent effects to reconstruct regulatory networks. We find a significant enrichment of co-localization with disease-associated variants, and we can now show that a subset of these are context-dependent.

In Chapter 8, using predicted disease-gene associations, we identify candidate genes that are likely involved in Mendelian diseases. These predictions are made by again using publicly available RNA-seq data. We show how this can be used to speed up and increase the diagnostic yield of clinical sequencing.

### Part 3 - Discussion

In Chapter 9, I discuss the work presented in this thesis and some possibilities how improved insights into genetic variation can be useful in medical practice.

## References


1. Levy, S. *et al.* The Diploid Genome Sequence of an Individual Human. *PLoS Biol.* **5,** e254 (2007).
2. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74,** 1111–20 (2004).
3. Moore, L. G. *et al.* Maternal adaptation to high-altitude pregnancy: an experiment of nature--a review. *Placenta* **25 Suppl A,** S60-71 (2004).
4. Mendel, J. G. Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines Brünn* **4,** 3–47 (1866).
5. Fisher, R. A. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Trans. R. Soc. Edinburgh* **52,** 399–433 (1919).
6. Badano, J. L. & Katsanis, N. Beyond Mendel: an evolving view of human genetic disease transmission. *Nat. Rev. Genet.* **3,** 779–789 (2002).
7. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* **97,** 199–215 (2015).
8. Jamuar, S. S. & Tan, E.-C. Clinical application of next-generation sequencing for Mendelian diseases. *Hum. Genomics* **9,** 10 (2015).
9. EBI GWAS Catalog. Available at: https://www.ebi.ac.uk/gwas/. (Accessed: 9th August 2018)
10. National Human Genome Research Institute. The Cost of Sequencing a Human Genome. Available at: https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/. (Accessed: 21st August 2017)
11. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101,** 5–22 (2017).
12. Gibbs, R. A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
13. Nouhravesh, N. *et al.* Analyses of more than 60,000 exomes questions the role of numerous genes previously associated with dilated cardiomyopathy. *Mol. Genet. genomic Med.* **4,** 617–623 (2016).
14. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10,** 387–406 (2009).
15. de Bakker, P. I. W. *et al.* Efficiency and power in genetic association studies. *Nat. Genet.* **37,** 1217–1223 (2005).
16. Allison, A. C. Protection afforded by sickle-cell trait against subtertian malareal infection. *Br. Med. J.* **1,** 290–4 (1954).

17. Kwiatkowski, D. P. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am. J. Hum. Genet.* **77,** 171–92 (2005).
18. Reich, D. *et al.* Reduced Neutrophil Count in People of African Descent Is Due To a Regulatory Variant in the Duffy Antigen Receptor for Chemokines Gene. *PLoS Genet.* **5,** e1000360 (2009).
19. Capellini, T. D. *et al.* Ancient selection for derived alleles at a GDF5 enhancer influencing human growth and osteoarthritis risk. *Nat. Genet.* **49,** 1202–1210 (2017).
20. Genovese, G. *et al.* Association of Trypanolytic ApoL1 Variants with Kidney Disease in African Americans. *Science (80-. ).* **329,** (2010).
21. Byars, S. G. *et al.* Genetic loci associated with coronary artery disease harbor evidence of selection and antagonistic pleiotropy. *PLOS Genet.* **13,** e1006328 (2017).
22. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507,** 371–5 (2014).
23. Chatterjee, S. & Ahituv, N. Gene Regulatory Elements, Major Drivers of Human Disease. *Annu. Rev. Genomics Hum. Genet.* **18,** annurev-genom-091416-035537 (2017).
24. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).
25. Green, P. H. R. & Cellier, C. Celiac Disease. *N. Engl. J. Med.* **357,** 1731–1743 (2007).
26. Kakalacheva, K., Münz, C. & Lünemann, J. D. Viral triggers of multiple sclerosis. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1812,** 132–140 (2011).
27. Filippi, C. M. & von Herrath, M. G. Viral trigger for type 1 diabetes: pros and cons. *Diabetes* **57,** 2863–71 (2008).
28. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343,** 1246949 (2014).
29. Tigchelaar, E. F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5,** (2015).
30. Netea, M. G. *et al.* Understanding human immune function using the resources from the Human Functional Genomics Project. *Nat. Med.* **22,** 831–833 (2016).

# 2

**European Journal of Human Genetics, 2014**

Patrick Deelen[1,2], Androniki Menelaou[3], Elisabeth M. van Leeuwen[4], Alexandros Kanterakis[1,2], Freerk van Dijk[1,2], Carolina Medina-Gomez[5,6,7], Laurent C. Francioli[3], Jouke Jan Hottenga[8], Lennart C. Karssen[4], Karol Estrada[5,6,9,10], Eskil Kreiner-Møller[5,6,11], Fernando Rivadeneira[5,6,7], Jessica van Setten[3], Javier Gutierrez-Achury[1], Harm-Jan Westra[1], Lude Franke[1], David van Enckevort[2,12], Martijn Dijkstra[1,2], Heorhiy Byelas[1,2], Cornelia M. van Duijn[6], Genome of the Netherlands Consortium, Paul I. W. de Bakker[3,13,14,15], Cisca Wijmenga[1], Morris A. Swertz[1,2]

# Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'

## Abstract

Although genome-wide association studies (GWAS) have identified many common variants associated with complex traits, low-frequency and rare variants have not been interrogated in a comprehensive manner. Imputation from dense reference panels, such as the 1000 Genomes Project (1000G) or HapMap, enable testing of ungenotyped variants for association. Here we present the results of imputation using a large, new population-specific panel: the Genome of The Netherlands (GoNL). We benchmarked the performance of the 1000G and GoNL reference sets by comparing imputation genotypes to 'true' genotypes typed on ImmunoChip in three European populations (Dutch, British and Italian). GoNL showed significant improvement in the imputation quality for rare variants (MAF 0.05%–0.5%) compared to 1000G. In Dutch samples, the mean observed Pearson correlation, $r^2$, increased from 0.61 to 0.71. We also saw improved imputation accuracy for other European populations (in the British samples, $r^2$ improved from 0.58 to 0.65, and in the Italians from 0.43 to 0.47). A combined reference set comprising 1000G and GoNL improved the imputation of rare variants even further. The Italian samples benefitted the most from this combined reference (the mean $r^2$ increased from 0.47 to 0.50). We conclude that the creation of a large population-specific reference is advantageous for imputing rare variants and that a combined reference panel across multiple populations yields the best imputation results.



1 Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
2 Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
3 Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands
4 Genetic Epidemiology Unit, Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands
5 Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands
6 Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands
7 Netherlands Genomics Initiative (NGI)-sponsored Netherlands Consortium for Healthy Aging (NCHA)
8 Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands
9 Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA
10 Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA
11 COPSAC; Copenhagen Prospective Studies on Asthma in Childhood; Faculty of Health Sciences, University of Copenhagen
12 NBIC BioAssist, Netherlands Bioinformatics Center, Nijmegen, The Netherlands
13 Department of Epidemiology, University Medical Center Utrecht, Utrecht, The Netherlands
14 Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA
15 Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA
Corresponding author: Morris A. Swertz

## Introduction

Although genome-wide association studies (GWAS) have been very effective in identifying loci associated to diseases or traits [1], it has proved difficult to fine-map the association signals to causal variants [2,3]. To overcome these limitations, there has been increasing interest in the interrogation of less frequent variants, especially given the enrichment of deleterious alleles at low frequencies [4–7]. There are specialized chips that can assess a larger number of rare variants, like the ImmunoChip [8] or Metabochip [9], although they do not provide uniform genome-wide coverage. Hence, most investigators will use statistical imputation from SNP arrays in GWAS using dense reference panels.

Imputation using a densely typed reference set can be performed to infer untyped variants that can be used to improve the power of a GWAS [10] and there are numerous examples where imputation has effectively enriched the results in GWAS [11,12]. While most large studies have so far been based on meta-analysis of HapMap-based imputations across cohorts, the primary limitation is that HapMap is essentially restricted to common variation (MAF > 5%). Thanks to the sequencing of larger samples, such as 1000G, more complete reference panels are now being assembled, setting off a new wave of meta-analyses.



The power of detecting an association in a GWAS is determined by its sample size and effective genome-wide coverage of the included variants, among other things [13,14]. The effective coverage depends directly on the number and quality of the imputed genotypes [15]. In turn, the quality of the reference panel will depend largely on the number of samples, the quality of the haplotypes, and the number of variants included [16].

The Genome of The Netherlands (GoNL) has the potential to provide a good imputation reference panel. GoNL is a population-based sequencing project, in which 769 Dutch samples were sequenced at on average 14X coverage [17]. In particular, the fact that GoNL sequenced trios (231) or quartets (19) has enabled improved haplotype phasing by using one of the children [18]. The GoNL imputation reference set contains 998 unrelated haplotypes. In this paper we report a quantitative analysis to assess the quality of imputed genotypes from using both GoNL and 1000G in Dutch and other European populations.

We adopted a 'gold standard' approach using samples genotyped on two distinct platforms, HumanHap550 and ImmunoChip. Hap550 is a commonly used genotyping chip designed to tag as many haplotypes as possible using common variants. ImmunoChip, however, is a fine-mapping chip: it contains a large number of low-frequency and rare variants for a limited number of loci (primarily selected based on loci identified in immune-related traits). Starting from the Hap550-genotyped SNPs, we were able to impute a large number of variants present on ImmunoChip. We then compared these imputed genotypes to the measured ('gold standard') genotypes on ImmunoChip to quantify the imputation performance. We have such a dataset for three European populations: the Dutch, British, and Italians. For each population we used 745 samples genotyped on both the platforms. These three populations allowed us to ascertain population-specific differences in the imputation quality of SNPs.

# Material & Methods

## Genome of the Netherlands

The Genome of The Netherlands (GoNL) is a project in which 769 individuals from different Dutch provinces were sequenced at on average 14X coverage [17]. All samples are part of either one of the 231 trios or one of the 19 quartets. The phasing was performed using the trio information [18], and for the quartets, one of the children was used to enhance the phasing. Due to sequence failures of two parents, from different trios, these samples were excluded from the imputation reference sets. Instead, from these two trios, we used the haplotype of the child that was not present in the other parent. This resulted in an imputation reference set containing 998 unrelated haplotypes. We used GoNL release 4 for all our analyses (see www.nlgenome.nl). The current GoNL release 5 also contains over 1 million indels but did not change the SNPs.

## Benchmarking samples

Samples from a celiac disease patient cohort were selected, since they had been genotyped on both the Hap550 and ImmunoChip [19]. The 745 Dutch and the 745 British samples were all cases, while the 745 Italian samples comprised 371 cases and 374 controls. The clustering for the genotype calling of the ImmunoChip data was performed manually in the past, to ensure proper genotyping results.

The Hap550 (516,426 SNPs) data was filtered on MAF > 1% and HWE p-value > $1 \times 10^{-4}$ for each population separately. The ImmunoChip (113,991 SNPs) data was filtered on MAF > 0.05% and HWE p-value of $1 \times 10^{-4}$. Both datasets are filtered on variants present in both the 1000G reference set as in the GoNL reference set. After QC the Dutch, British and Italian Hap550 data contain 509,888, 509,984 and 510,225 SNPs. The ImmunoChip data contains in the same order 107,383, 107,212 and 107,611 SNPs.

## Combining 1000G and GoNL data

The reference set combining data from 1000G and GoNL was created using the IMPUTE2 option: "--merge_ref_panels". This merged reference set was written to a file and subsequently used for the benchmarking. Since our benchmarking data is filtered for variants present in both reference sets, we did not assess the imputations of variants that are unique to either reference set.

## Pre-phasing

The 745 samples for each population were pre-phased using SHAPEIT2 [20]. This was done per chromosome using the default settings.

## Imputation

The imputations were performed using IMPUTE2 2.3.0 [16]. The different populations were imputed separately and in chunks of 5 Mb. For the comparison using an equal number of identical European haplotypes, we performed an imputation using all 379 European 1000G samples and a random selection of 379 GoNL samples. The random selection of GoNL samples was performed stratified on the Dutch provinces. These samples were selected using the IMPUTE2 option: "--exclude_samples_h".

We used MOLGENIS compute [21] to implement the imputation pipeline, run the 8,835 imputation chunks in parallel on a PBS compute cluster, and to keep track of the 15 imputations (five for each population). All pipelines are available as open source via http://www.molgenis.org/wiki/ComputeStart.

### Gold standard method

As stated above, we used samples genotyped on two distinct platforms. We imputed the Hap550 genotypes from these samples and compared the imputed genotypes to the SNPs previously only present in the ImmunoChip data. We used the ImmunoChip data as our 'gold standard'. The concordance between imputed genotypes and ImmunoChip genotypes was determined by calculating the Pearson correlation $r^2$ between the imputed dosage and ImmunoChip observed genotypes. The mean concordances were calculated for three MAF bins: rare (≥ 0.05% and < 0.5%), low-frequency (≥ 0.5% and < 5%) and common (> 5%) SNPs. The MAF used to stratify the SNPs into the bins was calculated separately for each population. The results were plotted using R 2.14.2 [22]. The significance of the differences between the reference sets was calculated using the Wilcoxon signed-rank test implementation in R.



### Principal component analysis

The principal component analysis (PCA) was performed using the EIGENSOFT 4.2 package [23]. The components were calculated using the European 1000G, GoNL, and the 3 GWAS datasets that we used for benchmarking. Before the components were calculated, all datasets were filtered to only include variants with a MAF > 5%. A joint dataset, featuring variants present in all 5 datasets, was created. This dataset was again filtered for MAF > 5%, the merged data was also filtered on HWE > $1x\ 10^{-4}$ and a call rate of 95%. This dataset was pruned using PLINK 1.07 [24] with the "--indep-pairwise" option, windows: 1000, step: 5, $r^2$ threshold: 0.2. The first component explained 0.33% of the variation and the second 0.10%. All subsequent components described less than 0.06%.

## Results

We stratified our analysis into three groups: common variants (MAF ≥ 5%), low-frequency variants (MAF 0.5%–5%), and rare variants (MAF 0.05%–0.5%). We focused mainly on the rare variants, since these are more difficult to impute and most can be gained in terms of imputation quality when using a better reference set. We observed a large increase in the imputation quality of rare variants when using GoNL as the reference compared to 1000G (Figure 1, Table 1). The mean observed Pearson correlation ($r^2$)

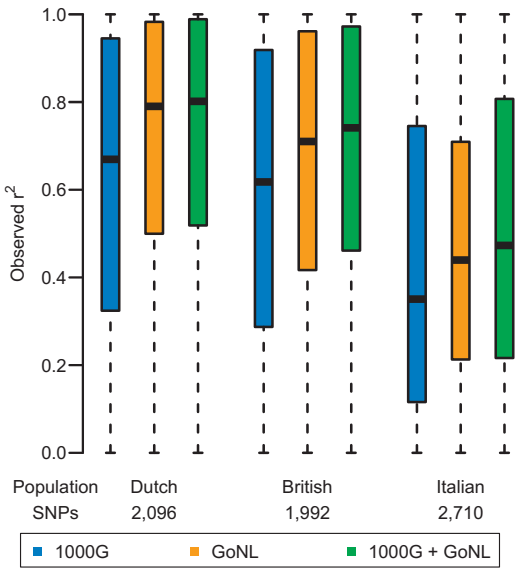



***Figure 1: Comparison of imputation quality of rare variants using the 1000G data, GoNL, and the combined reference panel.***

showed a significant increase from 0.61 to 0.71 for Dutch samples (Wilcoxon p-value = $7.16 \times 10^{-60}$). The British and Italian imputations also showed a significant improvement when imputing rare variants, from 0.58 to 0.65 ($p = 3.70 \times 10^{-35}$) and from 0.43 to 0.47 ($p = 2.64 \times 10^{-13}$), respectively. GoNL also significantly outperformed the 1000G reference set in the imputation of variants with higher MAFs (Figures/Appendices S1, S2, S3).

| Reference set | Dutch | British | Italian |
|---|---|---|---|
| 1000G | 0.61 | 0.58 | 0.43 |
| GoNL | 0.71 | 0.65 | 0.47 |
| 1000G + GoNL | 0.72 | 0.67 | 0.50 |

***Table 1: Mean observed r2 of rare variants.*** *Differences in the mean imputation quality between the reference sets was significant for each population ($p < 0.001$).*

Using a combined reference set composed of the 1000G and GoNL samples, we could improve the imputation further. The imputation of rare variants using the combined reference in Dutch and British samples showed a small increase in quality compared to GoNL-only imputation, respectively 0.02 ($p = 1.16 \times 10^{-3}$) and 0.02 ($p = 2.70 \times 10^{-5}$). The Italians benefitted most from the combined reference with an increase of 0.04 ($p = 3.62 \times 10^{-30}$) compared to a GoNL-only reference, resulting in a mean concordance for rare variants of 0.5. The differences in imputation quality when using the combined reference set for more frequent alleles were either very small or not significant (Figure S1, Tables S2 & S3).

A striking trend in these results is that the imputation quality of rare variants in Italians samples is lower than that in Dutch and British samples. The Dutch and Italian samples were genotyped at the same center and have similar call rates, and there were no indications that the genotyping quality of the Italian samples was lower. However, a principal component analysis (PCA) revealed that the Italian samples were not as well represented by either 1000G or GoNL compared to the Dutch and British GWAS samples used for benchmarking (Figure 2).

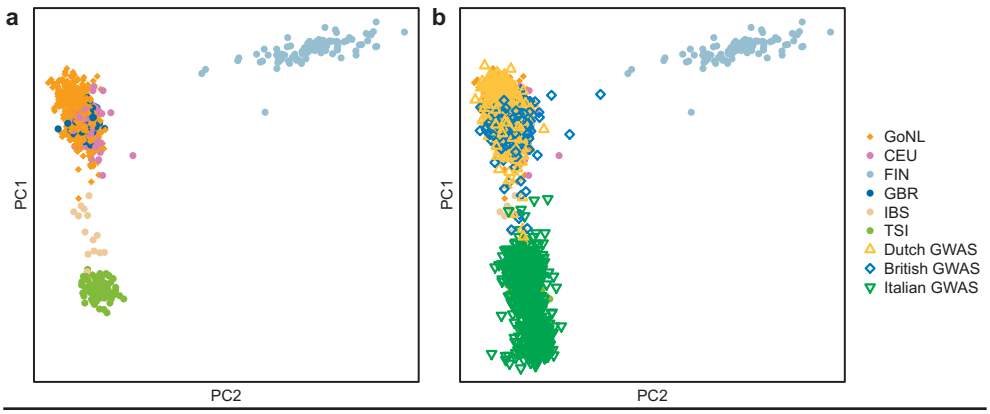



***Figure 2: Clustering of reference and study samples.*** *PC1 and PC2 reveal 3 main clusters: Tuscans from Italy (TSI), Finnish (FIN), and a Western European cluster with the CEU (Utah Residents with Northern and Western European ancestry), the GBR (British) and the GoNL samples (Panel a). Panel b shows that most of our GWAS samples clustered in a similar way to the corresponding 1000G/GoNL samples.*

We assessed whether the better performance of GoNL compared to 1000G was due to the larger number of European haplotypes in the reference set (998 vs. 578 in 1000G). We did this by performing an imputation using solely the 379 European samples in 1000G and a random subset of 379 GoNL samples. We found that the GoNL subset also significantly outperformed the European 1000G subset (Table 2).

| Reference set | Dutch | British | Italian |
|---|---|---|---|
| 1000G European | 0.59 | 0.57 | 0.40 |
| GoNL random subset 379 samples | 0.68 | 0.64 | 0.45 |

***Table 2: Mean observed $r^2$ of rare variants for reference sets of equal sample size from 1000G and GoNL (all of European descent).*** *Differences in the mean imputation quality between the reference sets was significant for each population ($p < 0.001$).*

Our experimental design also allowed us to assess the calibration of the posterior probabilities of the genotypes as they are output by IMPUTE2. We observed that the posterior probabilities were, in general, well calibrated, although we did observe a few deviations for low-frequency and rare variants (Figure 3A). To ascertain if these deviations in posterior probabilities affect the predicted imputation quality, the IMPUTE2 info metric, we plotted the predicted quality against the observed $r^2$. This showed a strong correlation between the predicted and observed quality for common variants and low-frequency variants (correlation of 0.97 and 0.91, respectively; Figure 3B & 3C). However, the info metric is not as accurate for rare variants, and the correlation with observed $r^2$ dropped to 0.70 (Figure 3D). We also observed some discrepancies where a near perfect imputation was predicted while in fact there was poor imputation, and vice versa when assessing rare variants.

## Discussion

We have shown that the new GoNL reference set provides higher downstream imputation accuracy than the 1000G reference set, not only for Dutch samples, but also for other European populations studied in this paper. Aside from the increase in imputation quality of rare variants in Dutch samples from 0.61 (1000G) to 0.71 (GoNL), we also observed an increase in imputation quality in British (0.58 to 0.65) and Italian (0.43 to 0.47) samples. We show that GoNL yielded better imputed genotypes for at least these European populations. A combined reference set, of 1000G and GoNL, increased the mean imputation quality of rare variants even further to 0.72, 0.67 and 0.50 for the Dutch, British and Italians, respectively.

By selecting an identical number of European haplotypes from 1000G and from GoNL, we showed a strong added value for GoNL in all the tested populations, confirming that the trio design of GoNL and the resulted accurate haplotypes aid the downstream imputation quality. We also observed a population-specific added value of GoNL when imputing Dutch samples. The added value (i.e. mean increase in imputation quality) was largest when comparing GoNL to 1000G in imputing the Dutch samples. Of course, it was already known that a better matched reference set will result in better imputed genotypes [25], however, the results from this paper were based on low-frequency variants and we show that there is also an inter-European effect of reference sets.

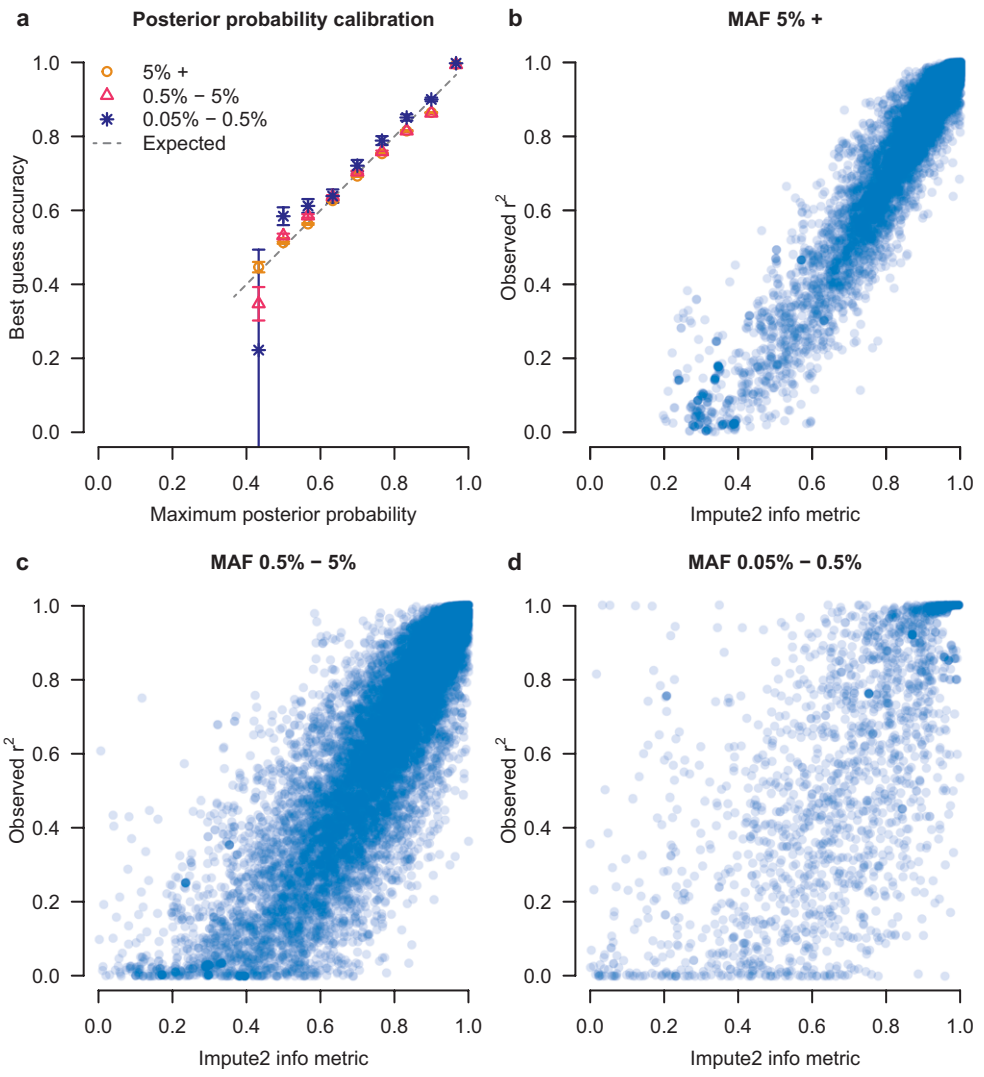



***Figure 3: Calibration of posterior probabilities.*** *The posterior probabilities were, in general, well calibrated, although there were a few deviations from the expected accuracy (panel a). For common and low-frequency variants (panels b & c), we observed a strong correlation (r2 0.97 and 0.91, respectively) between the IMPUTE2 info metric and the observed r2. However, for the rare variants (panel d), the relation between predicted and observed quality was less profound. We also observed a correlation of 0.70 and several large deviations from the diagonal.*

It is important to note that we only assessed variants present on the ImmunoChip. Although these variants were not randomly selected, we have no reason to assume that the imputation quality will be positively biased or that they do not represent low-frequency variants in general. The ImmunoChip was made to fine map loci previously associated to autoimmune diseases using a large number of low-frequency and rare variants.

We were encouraged to observe that the posterior probabilities were, in general, well calibrated with respect to the gold standard genotypes. We observed no adverse effects on the accuracy of the IMPUTE2 info metrics, although for rare variants we did observe a few instances with large deviations between the predicted and observed quality. This is in line with previous observations [26]. This observed inaccuracy also emphasizes the importance of validating associations from imputed genotypes.



It was shown earlier that a larger and more diverse reference set can improve the imputation of low-frequency variants [27]. We observed that a combination of 1000G and GoNL showed limited added value for the imputation of rare variants in the Dutch and British samples. It was, however, interesting to observe that the imputation of the Italian samples was improved more by this combined reference panel, leading us to speculate that populations that are poorly represented in the reference panel benefit more from a large and diverse reference set. Despite the limited added value for the Dutch and British datasets, such a large reference set may still be of interest for consortia aiming to impute cohorts of both European and non-European origin. All these cohorts can be imputed using the same combined reference set and then use IMPUTE2 to automatically select the best matching haplotypes [25]. We should note that we were only able to assess variants present in both reference sets, since there are very few variants on the ImmunoChip that are unique to either GoNL or 1000G. Nonetheless, our results show that population-specific reference sets and cosmopolitan panels, such as 1000G, can augment each other. This even holds true for the imputation of samples with ancestry other than those present in the population-specific reference sets, which provides further motivation for international efforts towards large and integrated reference sets.

## Acknowledgments

This study was made possible by rainbow grant 2 from BBMRI-NL to MS, a research infrastructure financed by, the Netherlands Organization for Scientific Research (NWO project 184.021.007). We thank the Target project (http://www.rug.nl/target) for providing the compute infrastructure, and the BigGrid/eBioGrid project (http://www.ebiogrid.nl) for sponsoring the pipeline implementation. We thank Jackie Senior for careful reading and editing the manuscript.

This study made use of data generated by the Genome of the Netherlands project, which is funded by the Netherlands Organization for Scientific Research (grant no. 184021007). The data was made available as a Rainbow Project of BBMRI-NL. Samples were contributed by LifeLines (http://lifelines.nl/lifelines-research/general), the Leiden Longevity Study (http://www.healthy-ageing.nl; http://www.langleven.net), the Netherlands Twin Registry (NTR: http://www.tweelingenregister.org), the Rotterdam studies, (http://www.

erasmus-epidemiology.nl/rotterdamstudy), and the Genetic Research in Isolated Populations program (http://www.epib.nl/research/geneticepi/research.html#gip). The sequencing was carried out in collaboration with the Beijing Institute for Genomics (BGI).

## Author contributions

PD, AM, MS, PB, CW: Writing of main manuscript. All: Discussion of experimental design in Genome of The Netherlands imputation working group. EL, AK, LK, CM, JH, FD: Revision of manuscript. PD, FD, MD, HB, LF, HW, AK, EK, CM: Implementation of analysis.

## Additional material

The following supplements are available with the on-line version of this paper.

Figure S1: Graphical comparison of the mean imputation quality using 1000G, GoNL, and the merged reference set comprising 1000G and GoNL.
Table S2: Mean of imputation quality when using the different reference sets.
Table S3: Wilcoxon signed-rank test p-values for all pairwise comparisons of reference sets.

## References


1. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 9362–7 (2009).
2. Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44,** 1294–301 (2012).
3. Shea, J. *et al.* Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nat. Genet.* **43,** 801–5 (2011).
4. Kryukov, G. V *et al.* Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80,** 727–39 (2007).
5. Cirulli, E. T. *et al.* Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11,** 415–25 (2010).
6. Lee, S. *et al.* Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13,** 762–75 (2012).
7. Huyghe, J. R. J. *et al.* Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat. Genet.* **45,** 197–201 (2013).
8. Cortes, A. *et al.* Promise and pitfalls of the Immunochip. *Arthritis Res. \& Ther.* **13,** 101 (2011).
9. Keating, B. J. *et al.* Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PLoS One* **3,** e3583 (2008).
10. Hao, K. *et al.* Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. *BMC Genet.* **10,** 27 (2009).

11. Holm, H. *et al.* A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat. Genet.* **43,** 316–20 (2011).
12. Li, Y. *et al.* Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10,** 387–406 (2009).
13. de Bakker, P. I. W. *et al.* Efficiency and power in genetic association studies. *Nat. Genet.* **37,** 1217–1223 (2005).
14. Flannick, J. *et al.* Efficiency and power as a function of sequence coverage, SNP array density, and imputation. *PLoS Comput. Biol.* **8,** e1002604 (2012).
15. Zheng, J. *et al.* A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genet. Epidemiol.* **35,** 102–10 (2011).
16. Howie, B. *et al.* A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).
17. Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur. J. Hum. Genet.* (2013). doi:10.1038/ejhg.2013.118
18. Menelaou, A. *et al.* Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics* **29,** 84–91 (2013).
19. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* **43,** 1193–201 (2011).
20. Delaneau, O. *et al.* Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10,** 5–6 (2013).
21. Byelas, H. *et al.* Scaling Bio-Analyses from Computational Clusters to Grids. in *IWSG* (ed. Kiss, T.) **993,** (CEUR-WS.org, 2013).
22. R Core Team. *R: A language and environment for statistical computing*. *Vienna, Austria: R Foundation for Statistical Computing* **1,** (R Foundation for Statistical Computing, 2008).
23. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38,** 904–909 (2006).
24. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
25. Howie, B. *et al.* Genotype imputation with thousands of genomes. *G3 genes - genomes - Genet.* **1,** 457–70 (2011).
26. Li, L. *et al.* Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. *PLoS One* **6,** e24945 (2011).
27. Jostins, L. *et al.* Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur. J. Hum. Genet.* **19,** 662–6 (2011).


**BMC Research Notes, 2014**

Patrick Deelen[1,2,*], Marc Jan Bonder[2,*], K. Joeri van der Velde[1,2], Harm-Jan Westra[2], Erwin Winder[1,2], Dennis Hendriksen[1,2], Lude Franke[2] and Morris A. Swertz[1,2]

# Genotype Harmonizer: automatic strand alignment and format conversion for genotype data integration

## Abstract

### Background

To gain statistical power or to allow fine mapping, researchers typically want to pool data before meta-analyses or genotype imputation. However, the necessary harmonization of genetic datasets is currently error-prone because of many different file formats and lack of clarity about which genomic strand is used as reference.

### Results

Genotype Harmonizer (GH) is a command-line tool to harmonize genetic datasets by automatically solving issues concerning genomic strand and file format. GH solves the unknown strand issue by aligning ambiguous A/T and G/C SNPs to a specified reference, using linkage disequilibrium patterns without prior knowledge of the used strands. GH supports many common GWAS/NGS genotype formats including PLINK, binary PLINK, VCF, SHAPEIT2 & Oxford GEN. GH is implemented in Java and a large part of the functionality can also be used as Java 'Genotype-IO' API. All software is open source under license LGPLv3 and available from www.molgenis.org/systemsgenetics.

### Conclusions

GH can be used to harmonize genetic datasets across different file formats and can be easily integrated as a step in routine meta-analysis and imputation pipelines.



1 University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, the Netherlands
2 University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands
* Equal contributions
Corresponding authors: Patrick Deelen & Morris A. Swertz

## Background

Genome-wide association studies (GWAS) increasingly require the integration of multiple genetic data sets to reach sufficient resolution and statistical power, either by imputing missing genotypes or by pooling datasets for a meta-analysis. However, there are two major challenges to be resolved: 1) the large number of different file formats used by the genetics community, and 2) the ambiguous A/T and G/C single nucleotide polymorphisms (SNPs) for which the strand is not obvious. For many statistical analyses, such as meta-analyses of GWAS [1] and genotype imputation [2], it is vital that the datasets to be used are aligned to the same genomic strand.

Genotype data can be coded on either the forward genomic strand or the reverse genomic strand (e.g. a SNP coded T/G on the forward strand would be coded A/C on the reverse strand). The strand used to store the genotypes is not always the same within a dataset (i.e. the same strand may not be used for all variants) or between the different datasets to be aligned (i.e. the same strand may not be used for a variant present in both datasets); these differences can be intentional [3] or accidental. To complicate matters, most of the common file formats do not define the strand used. For some types of SNPs, it is fairly straightforward to detect and correct the strand differences. For example, a T/G SNP is non-ambiguous as its complement on the other strand is A/C. However, G/C and T/A variants are ambiguous or cryptic as their complementary alleles are C/G and A/T, respectively. This ambiguity means it is more difficult to detect and resolve strand issues for these SNPs.

Of course, it is possible to simply exclude all ambiguous variants, however, modern genotyping chips often contain many A/T and G/C SNPs; the ImmunoChip has 25,740 such SNPs (1.7% of all SNPs), the ExomeChip 244,771 (11.9%) and the Omni5-quad 144.578 (3.4%). Simply excluding these variants will limit the power of a GWAS meta-analysis where the A/T or G/C variant is the causal variant or is in higher LD to the causal variant. In the case of imputation it has also been shown that more input genotypes yield imputed genotypes of higher quality [4], so if it is possible to include the A/T and G/C variants, this is more desirable. In the cases where the strand of the genotypes is known, there are many solutions to easily correct the strands of one dataset or to simply state explicitly the strand used, for example as is possible in IMPUTE2 [5] or METAL [6]. In practice, however, this information is not always available or trustworthy.

One solution to the problem of unknown strands is to compare the minor allele between two datasets. However, use of the minor allele is not ideal as it can differ between datasets and populations, especially for common variants. PLINK [7] employs a more powerful approach to detect strand inconsistencies between cases and controls. However, this method requires many manual steps, re-coding of phenotypes before and after the actual alignment, manual alignment of the non-ambiguous SNPs and merging the data into one dataset, and finally a script needs to be written to parse the alignment results from PLINK to determine the actual alignment. When using PLINK, it is not possible to align genotypes with posterior probabilities.

## Implementation

Here, we present Genotype Harmonizer (GH): a new command-line tool to automate genotype data harmonization. GH can read commonly used file formats (PLINK, binary PLINK, VCF, SHAPEIT2 & Oxford GEN) and align a study dataset to a specified reference without any prior knowledge of the strand used. After alignment, GH writes data back to a chosen format (PLINK, binary PLINK, SHAPEIT2 or Oxford GEN). All handling of the genotype data and loading genotypes from the different formats is implemented in our Genotype IO library, which also allows integration of the harmonization tools into other software. GH consists of 25,000 lines of code with a high unit test coverage of over 60% at conditional level and continuous build testing. GH is written in Java and has been tested under Linux, Windows, and OS-X. All source code is available at www.github.com/molgenis/systemsgenetics.

GH implements a fully automated method that assigns the strand of ambiguous SNPs by selecting nearby non-ambiguous SNPs that are in linkage disequilibrium (LD) in both the study data and the reference data. GH correlates the estimated haplotype frequencies between the study data and the reference data. If GH finds more negative correlations than positive ones in haplotype frequencies, the ambiguous SNP is swapped to the other strand. When GH is unable to align a SNP (e.g. because of a lack of surrounding SNPs), this ambiguous SNP is excluded from the set. It is possible to prevent exclusion of variants that could not be aligned using LD, GH can optionally perform alignment using the minor allele for variants that have a minor allele frequency below a specified value.

## Results

### Usage in an imputation workflow

We advise applying GH to pre-phased data before imputation. When pre-phasing using SHAPEIT2 [8] and imputing using IMPUTE2, GH can read the SHAPEIT2 output directly and can write aligned results in the same format for direct use by IMPUTE2 (Figure 1a). Performing the alignment after the pre-phasing step ensures that pre-phasing does not need to be repeated when imputing using a different reference set or a newer version of a reference set. GH can also update the variant identifiers of the study data to match the reference set identifiers using the --update-id option. An example command is:

```
GenotypeHarmonizer.sh --input shapeit2Output --ref refInVcf
--output targetPath --update-id
```

### Usage to harmonize GWAS data

GH can also be used in merging or meta-analysis of different GWAS datasets (Figure 1b). One of the datasets can be used as a reference and the other datasets can be aligned to it, or all the cohorts can be aligned to a public reference set. It is possible to include all the variants present in the study data that are not in the reference set using the --keep option. After alignment the datasets can be investigated using a meta-analysis or can be merged into a single dataset. An example command is:

```
GenotypeHarmonizer.sh --input dataset1 --ref dataset2 --output
dataset1Aligned --update-id --keep
```

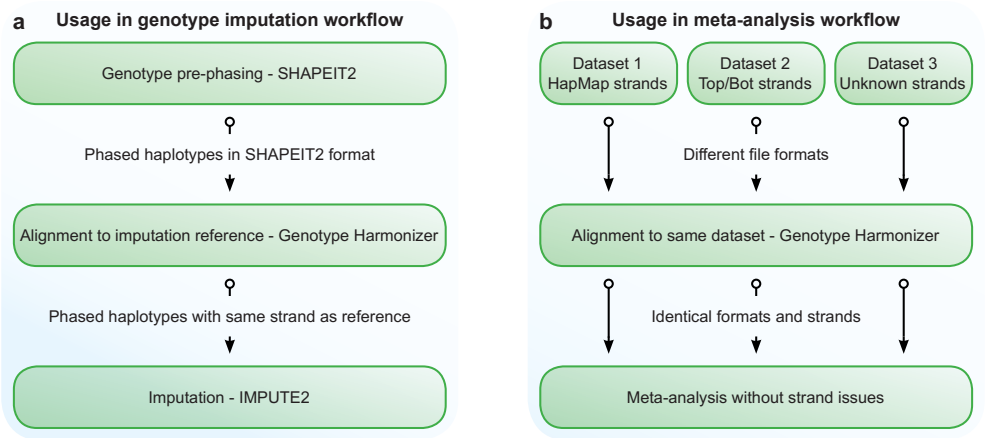



***Figure 1: Usage of Genotype Harmonizer.*** a) GH can be applied after the pre-phasing of the genotypes, preventing the need to redo the phasing for each new version of a haplotype reference set. b) GH can be used to align and reformat genotype datasets allowing easy merging or meta-analysing of data. By aligning all datasets to a public reference, the genotype data can be kept private by consortia members.

## Performance

GH requires 6:35 minutes to align a GWAS dataset consisting of 168,408 SNPs and 25,169 samples in binary PLINK format to another GWAS dataset with 528,969 SNPs and 11,950 samples, using a Linux system, a single core and 4 GB of RAM. Aligning the SHAPEIT2 results (25,169 and 19,321 variants on chromosome 1) to the Genome of The Netherlands imputation reference (499 samples, 1,536,126 SNPs on chromosome 1) [9] took 36 seconds using a single core and <1 GB of RAM.

## Comparison using PLINK alignment

We compared the alignment of ambiguous variants using GH to the alignment using the flip-scan option in PLINK. We performed this analysis by using the latest HapMap3 data. We randomly assigned the samples into two equally sized sets, henceforth denoted as set1 and set2. In set1 we randomly changed the strand of roughly 50% of the A/T and G/C variants.

Set1 was aligned using GH by using set2 as the reference using the default settings. We successful aligned 40,617 out of the 55,517 swapped variants, 14 (0.03%) variants were aligned to the incorrect strand. In total 29,801 A/T and G/C variants (27% of the total ambiguous variants) were excluded since there were not enough variants in LD for accurate alignment. There were no variants swapped by GH that were not flipped in our test set.

For the analysis using PLINK we denoted the samples in set1 as cases and set2 as controls; we merged both sets and used the flip-scan option using the default settings. PLINK does not actually report which variants should be swapped but instead provides a log with information on which the decision to swap a variant can be based. Since the PLINK manual does not provide a recommendation on how to select the variants to swap based on this file, we used the same criteria as those used by the GH, i.e. there need to be at least 3 variants in LD, and then we assessed if there were more positive than negative correlations. This

resulted in the successful alignment of 37,402 SNPs and the incorrect alignment of 54 SNPs (0.14%); 36,390 (33% of the total ambiguous variants) variants were excluded because of lack of variants in LD. We thus find that the number of incorrectly aligned SNPs increased by 40 SNPs and the number of excluded SNPs increased by 22% from 29,801 to 36,390 when using PLINK instead of GH.

Moreover, in one command GH covers many separate steps which require considerable manual work or scripting when using PLINK: manual alignment of non-ambiguous variants (which PLINK cannot do automatically), conversion of reference haplotypes to a PLINK supported format, merging the reference and study datasets, recoding using a fake phenotype file, running PLINK flip-scan to find swapped SNPs, and the selection and swapping of the SNPs on the wrong strand.

## Conclusions

We have shown that using Genotype Harmonizer we can provide near perfect alignment of ambiguous SNPs without any prior knowledge of the strands. Compared to PLINK we have improved the strand alignment and limited the number of manual steps without sacrificing run-time performance. Another advantage of GH over PLINK is our support of file formats storing haplotype phase or genotype probability information, which also makes our software useful to employ within an imputation workflow or on data that has already been imputed.

GH uses an advanced LD-based method to perform the alignment of ambiguous SNPs and supports many genotype file formats. The underlying Genotype IO API is part of the MOLGENIS open source suite [10], which is also used by several other genetic analysis tools, and we expect the number of supported formats to grow in the future. These enhancements will be made available in later releases of GH. We have used GH to harmonize over 15 imputations and GWAS datasets [11–14]. GH is now a standard part of our imputations and has been applied to over 25,000 samples (publications in preparation). We expect GH to be a major time saver for many research groups and to become a standard part of many analysis pipelines, as it alleviates manual steps when imputing data or when working with multiple GWAS datasets.

## Availability and requirements

- Project name: Genotype Harmonizer
- Project home page: www.molgenis.org/systemsgenetics
- Operating system(s): Platform independent
- Programming language: Java
- Other requirements: Java 1.6 or higher
- License: LGPLv3
- Any restrictions to use by non-academics: Free to use

## Acknowledgements

We thank Kate Mc Intyre and Jackie Senior for carefully reading and editing the manuscript and Alexandros Kanterakis for testing the software. Funding: The research leading to these

results received funding from BBMRI-NL, a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO project 184.021.007), to PD, MJB; the European Union Seventh Framework Programme (FP7/2007-2013) under grant 261433 (BioSHaRE-EU) to KJV; LifeLines/Target to EW; and TI Food and Nutrition (TIFN GH001) to MS.

## Author contributions

PD, MJB, LF, HJW, MS designed the software. PD, MJB, KJV, EW, DH implemented the software. PD, MJB, MS wrote the manuscript.

## References

1. Evangelou, E. *et al.* Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **14,** 379–89 (2013).
2. Marchini, J. *et al.* Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11,** 499–511 (2010).
3. Illumina. "TOP / BOT" Strand and "A / B" Allele. (2006). Available at: http://res.illumina.com/documents/products/technotes/technote_topbot.pdf.
4. Roshyara, N. *et al.* Impact of pre-imputation SNP-filtering on genotype imputation results. *BMC Genet.* **15,** 88 (2014).
5. Howie, B. *et al.* A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).
6. Willer, C. J. *et al.* METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–1 (2010).
7. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
8. Delaneau, O. *et al.* Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Genet.* **10,** 5–6 (2013).
9. The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46,** 818–825 (2014).
10. Swertz, M. A. *et al.* The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinformatics* **11,** S12 (2010).
11. Deelen, P. *et al.* Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *Eur J Hum Genet* **22,** 1321–1326 (2014).
12. Almeida, R. *et al.* Fine mapping of the celiac disease-associated LPP locus reveals a potential functional variant. *Hum. Mol. Genet.* **23,** 2481–2489 (2014).
13. Bonder, M. J. *et al.* Genetic and epigenetic regulation of gene expression in fetal and adult human livers. *BMC Genomics* **15,** 860 (2014).
14. Tigchelaar, E. F. et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: Study design and baseline characteristics BMJ Open **5**, (2015).

# 4

**Genome Medicine, 2015**

Patrick Deelen[1,2,*], Daria V. Zhernakova[1,*], Mark de Haan[1,2], Marijke van der Sijde[1], Marc Jan Bonder[1], Juha Karjalainen[1], K. Joeri van der Velde[1,2], Kristin M. Abbott[1], Jingyuan Fu[1,] Cisca Wijmenga[1], Richard J. Sinke[1], Morris A. Swertz[1,2,#], Lude Franke[1,#]

# Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels

# Abstract

## Background

RNA-sequencing (RNA-seq) is a powerful technique for the identification of genetic variants that affect gene expression levels, either through expression quantitative trait locus (eQTL) mapping or through allele specific expression (ASE) analysis. Given increasing numbers of RNA-seq samples in the public domain, we here studied to what extent eQTLs and ASE effects can be identified when using public RNA-seq data while deriving the genotypes from the RNA sequencing reads itself.

## Methods

We downloaded the raw reads for all available human RNA-seq dataset. Using these reads we performed gene expression quantification. All samples were jointly normalized and subjected to a strict quality control. We also derived genotypes using the RNA-seq reads and used imputation to infer non-coding variants. This allowed us to perform eQTL mapping and ASE analyses jointly on all samples that passed QC. Our results were validated using samples for which DNA-seq genotypes were available.

## Results

4,978 public human RNA-seq runs, representing many different tissues and cell-types, passed quality control. Even though this data originated from many different laboratories, samples reflecting the same cell-type clustered together, suggesting that technical biases due to different sequencing protocols are limited. In a joint analysis on the 1,262 samples with high quality genotypes, we identified *cis*-eQTLs effects for 8,034 unique genes (at a false discovery rate ≤ 0.05). eQTL mapping on individual tissues revealed that a limited number of samples already suffice to identify tissue-specific eQTLs for known disease-associated genetic variants. Additionally, we observed strong ASE effects for 34 rare pathogenic variants, corroborating previously observed effects on the corresponding protein levels.

## Conclusions

By deriving and imputing genotypes from RNA-seq data, it is possible to identify both eQTLs and ASE effects. Given the exponential growth of the number of publicly available RNA-seq samples, we expect this approach will become especially relevant for studying the effects of tissue specific and rare pathogenic genetic variants to aid clinical interpretation of exome and genome sequencing.



1 University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands
2 University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, The Netherlands
* These authors contributed equally to this work.
\# These authors contributed equally to this work.
Corresponding author: Lude Franke

## Background

Most disease-associated genetic variants in humans are regulatory and affect gene-expression levels [1–3]. With the availability of RNA-sequencing (RNA-seq) two strategies are now commonly used to identify these effects: (1) expression quantitative trait loci (eQTL) mapping to identify common genetic variants that affect gene-expression levels [4–8], and (2) allele-specific expression (ASE) analysis to ascertain whether one allele is more abundantly expressed than the other for heterozygous samples. ASE can reveal significant effects even if only a single sample is heterozygous, permitting investigation of rare and low-frequency variants in coding regions. On the other hand, eQTL analyses can be used for any genetic variant, but typically require the use of dozens of samples in order to have sufficient individuals in the different genotype classes [9–11]. Most eQTL studies so far have focused on a single tissue with large sample sizes [3,12,13] (thereby enabling identification of small effects and entire networks of downstream genes, i.e. *trans*-eQTLs) or on a few tissues with limited sample sizes [14–16] (enabling identification of tissue- and cell-type-specific *cis*-eQTLs). Although efforts are ongoing, for instance by the GTEx consortium, to investigate larger numbers of different tissues [17], still the number of samples studied remains limited. Ideally, eQTL data on many tissues and many different samples should be available, since this would permit eQTL mapping and ASE analyses on rare and low-frequency variants within different cell types. This is especially important for the functional interpretation of clinically important rare variants (particularly recessive Mendelian mutations, where the mutant alleles have appreciable frequencies in the general population [18]), but will also aid in the classification of variants of unknown significance [19].

Fortunately, the raw data of many RNA-seq experiments are being deposited in public databases, and the number of available human RNA-seq samples is growing exponentially, for example, in the European Nucleotide Archive (ENA) (Figure 1a). Since it has recently been shown that it is possible to derive reliable genotypes from RNA-seq reads [20], leveraging publicly available RNA-seq samples might be a viable strategy for obtaining the sample sizes required to perform eQTL mapping and ASE analyses on rare and low-frequency variants across multiple cell-types.

Here we present an approach to quantify, normalize and genotype a large number of heterogeneous RNA-seq samples. We show that it is possible to reliably identify eQTLs across many different tissues and also to obtain tissue-specific eQTLs by combining samples from a single tissue derived from many different experiments. We assessed allele-specific expression (ASE) in a large number of samples and identified rare and low-frequency (pathogenic) variants that affect gene-expression levels. We have made all our results freely available online (http://www.molgenis.org/ase), allowing for easy querying of genetic variants of interest.

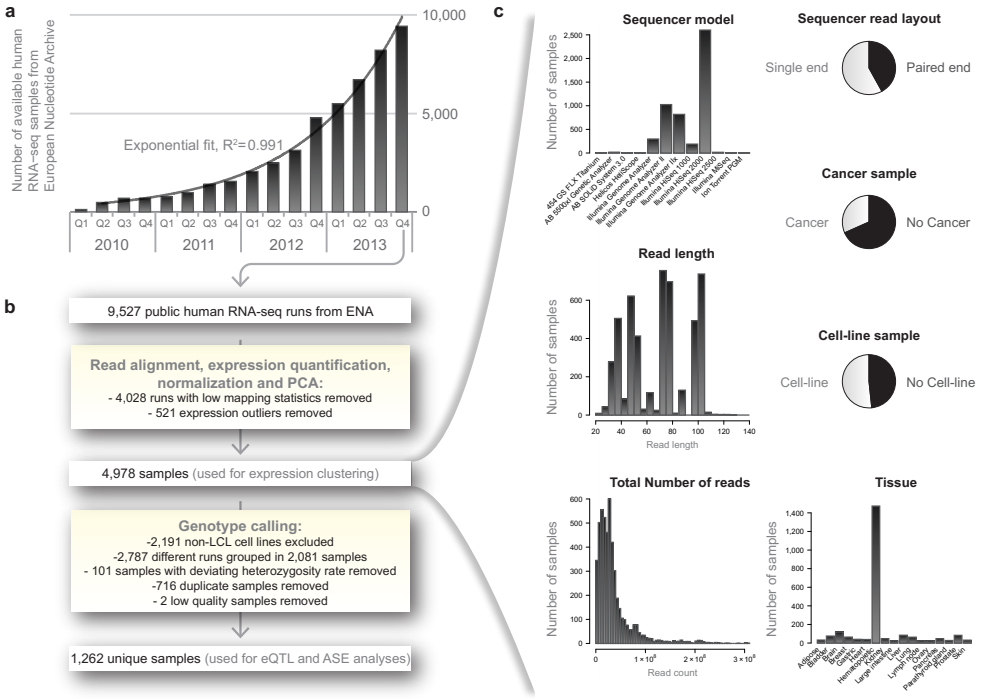



***Figure 1: Growth of publicly available RNA-seq and analysis workflow.*** *a) Over the past years the number of available public RNA-seq samples has increased exponentially (exponential fit r2 > 0.991) b) General overview of the steps taken to process, quality control and integrate all samples. c) Overview of diversity of 4,978 samples used for expression clustering. Three samples having read lengths >140 (365, 452, 151 bases) are omitted from the read length plot.*

# Methods

## Pipelines and QTL/ASE mapping software

We have made the pipeline and tools that we developed freely available as open source software. The pipelines are implemented in Molgenis compute [21] and can be downloaded at: http://github.com/molgenis/molgenis-pipelines. The eQTL/ASE mapping software is publicly available at: http://www.molgenis.org/systemsgenetics/QTL-mapping-pipeline

## Downloading public RNA-seq experiments

We downloaded the samples from the European Nucleotide Archive (ENA). The following filter criteria were used to download the data: Taxon: human (9606), Library strategy: RNA-Seq, Library source: Transcriptomic and Readcount: ≥ 500,000. This was performed on 16 January 2014 and resulted in 9,611 runs. We were able to download FASTQ files for 9,527 runs for which the md5sum was correct after the downloading.

## Read alignment

STAR 2.3.1l [22] was used to align the reads of the FASTQ files. It is known that read mapping to a common reference creates a mapping bias: more reads will be covering the reference allele than alternative allele [23]. To correct for this bias and allow the investigation of allelic imbalance, we aligned RNA-seq reads to the reference genome build 37 masked for SNPs with a MAF ≥ 1% in the Genome of The Netherlands (GoNL) data. Only uniquely mapping reads were included. We used a variable number of mismatches per run: for runs with a read length greater than 90 bases we allowed 4 mismatches, for a read length between 60 and 90 we allowed 3 mismatches, and for shorter reads we allowed 2 mismatches. The runs were filtered on their percentage of uniquely mapping reads. We selected 5,499 runs, each having at least 60% uniquely mapping reads. These filter criteria also ensured that all miRNA experiments were excluded.



## Gene level quantification

We used HTSeq-count 0.5.4 (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html) to quantify gene-expression levels. Ensembl version 71 was used as gene annotation database.

## Identification of gene-expression outliers

We performed quantile normalization and log2 transformation on the data from the 5,499 aligned runs. We then performed a principal component analysis (PCA) over the sample covariance matrix. This revealed 521 strong outliers for the first component (Figure S1). Close inspection of these 521 samples revealed that they included 3 samples that were in fact DNA-seq runs, 312 samples annotated as single-cell sequencing runs, 97 samples that specifically targeted the HLA region and 1 sample was a Geuvadis run[24] that did not cluster near the other Geuvadis samples. Based on this information we decided to remove these 521 runs leaving 4,978 runs. We then corrected the expression data for GC content and the total number of reads. After standardizing the expression levels for every gene we performed a new PCA (Figure 2, Figure S2). The raw and normalized expression data and the PCA results can be downloaded here: http://www.molgenis.org/ase. The annotations for each of these runs have been summarized in Table S1.

## Genotyping

After removing low quality samples we recalculated the PCA. The first 2 principal components show clear separation between primary tissues, cell lines and hematological tissues (Figure 2a). To select samples for eQTL and ASE analyses we decided to excluded all tumor-derived cell-line samples (where genotype calling is inherently difficult due to the presence of somatic copy number aberrations), by excluding all non-LCL samples with a principal score > 0 for PC2 (Figure 2a).

For the genotyping we used a combination of the GATK Unified Genotyper 2.8 [25] and imputation of the genotype likelihoods using Beagle 4 r1230 [26], which is identical to methods that have been proposed for low-coverage DNA sequencing [27]. There are many samples that were sequenced using multiple runs, in the cases where this had been specifically mentioned in the sample annotation, we merged all the aligned reads of the different runs to improve genotyping quality. We called genotypes for each sample individually for all 1000

Genomes, GoNL, and ClinVar [28] SNPs. We outputted all variants regardless of the calling quality or number of supporting reads. We excluded known RNA-editing sites, variants near splice junctions, and variants at repeat regions, as is recommend when calling variants in RNA-seq data [20].

The genotype likelihoods for the variants with a MAF ≥ 1% were used as input for imputation using Beagle 4 with version 5 of GoNL [29,30] as a reference. We performed imputation on all the samples merged together. The genotyping concordances of the Geuvadis samples were determined by calculating the correlation between the imputed RNA-seq dosages and the high-quality genotype calls of the Omni2.5 genotyping chips (as generated by the 1000 Genomes project).

For all samples, we calculated heterozygosity rates using the non-imputed genotypes while taking into account only SNPs with MAF ≥ 5% and a read coverage of at least 10 reads. We excluded 100 samples with a heterozygosity rate below 0.2 (suggesting the presence of chromosomal aberrations, uniparental disomies or strong inbreeding) or above 0.4 (suggesting potentially contaminated or pooled samples). This resulted in 1,980 genotyped samples.

## Removing duplicate samples

In order to identify duplicate samples that are not annotated as such by ENA we selected the high-quality imputed genotypes. We selected all variants with an estimated dosage $r^2$ above 0.95, a MAF of 0.05 and a genotyping rate of 0.95. We performed pruning using Plink 1.07 [31] (--indep --pairwise 1000 5 0.2) to select independent variants. We then calculated the pairwise genotype concordance for the remaining variants. Based on the resulting distribution we found that a cut-off of 78% was appropriate in order to deem samples duplicates (Figure S3).

If two or more samples were marked as duplicates we gave first priority to the Geuvadis samples, second priority to samples from tissues for which we had most other samples, and finally, those showing the highest number of expressed genes. Among the 1,264 unique samples that were eventually selected there were two samples that we excluded manually because they showed deviating expression levels from what we expected for their presumed tissue and because they had barely passed various other filter criteria. It is also worthwhile to note that these were among the 8 SOLID samples that had passed the rest of the QC. All our criteria finally resulted in 1,262 unique samples that we used for further analyses.

We subsequently investigated *XIST* gene-expression levels and overall chromosome Y expression levels and observed that the expression levels of these samples corresponded well to the gender annotations (available for 41% of the samples, Figure S4).

## Genotype PCA

The genotype PCA was performed on the 1,262 selected unique samples. The variant filtering and pruning was performed using the same settings as for removing the duplicates.

## eQTL mapping

Before we performed eQTL mapping, we selected the best RNA-seq run per sample by choosing the run with the highest number of expressed genes. The runs were normalized using Trimmed Mean of M-values (TMM) [32] and log2 transformation, centering and scaling. Finally we corrected our data for the number of mapped reads, the GC percentage, the first 4 genotype components, and the first 100 expression components. We grouped our samples using the genotype PCA into three different groups (Europeans (n = 948), Africans (n = 197) and Asians (n = 117)) and treated this as a meta-analysis when performing the eQTL mapping. We used our previously described eQTL mapping pipeline [3] and mapped *cis*-eQTL within 250 kb from the gene center. We only included variants with an expected dosage $r^2 \geq 0.8$ for the eQTL mapping.

To correct for multiple testing and in order to get reliable false discovery rates, we usually employ a permutation strategy where we define the null-distribution of eQTL effects by randomly assigning the genotype sample identifiers to expression sample identifiers and redoing the eQTL analysis. This is only effective when the genotype data has been generated independently from the expression data, no population stratification exists, and samples reflect the same cell-type or tissue. However, here, the genotype data and expression data have been derived from the same sample, and therefore a different permutation strategy is required, because if a gene is not expressed at all, no genotypes can be derived (and it could well be that subsequent imputation might not be able to resolve this as well either). It is therefore essential to only permute sample identifier labels within sets of samples that reflect the same cell-type or tissue. In order to do so, we permuted the sample identifiers within each of the different studies, because nearly all the studies concentrate on a single tissue. By using this approach we lower the chance that unknown confounders might cause false-positives. This is further alleviated by the fact that we have already accounted for most of the differences in expression between cell-types and tissues by correcting the expression data for 100 principal components.

The replication analysis was performed using the Geuvadis DNA-seq samples where we treated each Geuvadis population separately in a meta-analysis. For each replication analysis we only tested the most significant SNP for each significant gene.

We also performed tissue-specific eQTL mapping in four tissues. We selected only the samples coming from one population, which resulted in 42 European brain samples, 50 European breast samples, 42 European liver samples, and 45 Asian bladder samples. We ran eQTL mapping in the same manner as described above, with the exception that we performed a normal permutation since all samples were from the same tissue, and we tested whether the identified eQTLs were detectable in the Geuvadis LCL eQTL data.

## Allele-Specific Expression analysis

We performed allele-specific expression (ASE) analysis by fitting per SNP a binomial distribution using maximum likelihood estimation and subsequently assessed significance by using a likelihood ratio test. The FDR was controlled using the Benjamini–Hochberg procedure. During our initial ASE analysis (not shown) we observed a strong reference bias for low-frequency variants that had not been previously masked. We therefore again performed the

masking of the reference genome using all 1000G, GoNL and ClinVar variants and performed a new mapping of the 1,262 samples selected for the ASE analysis. We used Samtools mpileup 0.1.19 [33] and a custom script to obtain the read counts from these bam files, using only bases with a quality score of at least 17 (the use of a more stringent quality score of at least 30 resulted in fewer reads that we could use, and hence fewer significant ASEs that we could detect, but did not observe differences in the direction of the ASE effects, data not shown). We excluded variants in known RNA-editing sites, near splice junctions and in repeats regions in the same way as when we did the genotyping.

We checked for each sample genotype if the GATK deemed the individual heterozygous for this variant, thereby only using genotypes with a phred-scaled genotype quality (GQ) score above 30. For ASE analysis we selected the SNPs that were heterozygous in at least 5 samples, had at least 10 reads per allele, and at least 2% of all reads supporting each allele. We removed the sites that had a mappability score < 1 according to the USCS mappability track (CGR Alignability of 50mers) [24,34]. Using these criteria to select variants we tested for ASE in 56,825 SNPs when only interrogating the Geuvadis samples and tested for ASE in 225,562 SNPs when using all 1,262 samples.

## Results and discussion

### Expression quantification

We downloaded all publicly available human RNA-seq data from the ENA and aligned the reads for each of these samples. We identified 4,978 RNA-seq runs that passed our strict quality control (QC, see Methods). We performed several analyses to ascertain whether these sequence runs, produced in many laboratories around the world, jointly describe biologically coherent patterns. We first conducted principal component analysis (PCA) to obtain a global view on how the different samples clustered together. A PCA on the sample correlation matrix showed that components 1 and 2 permit near-perfect discrimination between primary tissues, cell lines, and hematopoietic tissues (Figure 2). Other components permit accurate identification of many tissue types such as brain (Figure 2b, components 4 and 10), liver (Figure 2c, components 14 and 11) and bladder (Figure 2d, components 4 and 38), even though the RNA-seq data for these tissues had been generated in at least six different laboratories, with often quite pronounced technical differences (e.g. in sequencer model, read layout, read length, and total number of reads, Figure 1c). Together, these results indicate that heterogeneous RNA-seq datasets that have been aligned, normalized and QC'ed in a systematic manner yield gene-expression profiles that very clearly describe biologically coherent phenomena. These results also indicate that researchers who would like to learn more about one specific tissue, could combine different (small-scale) RNA-seq data for that tissue into one large dataset.

### Genotyping and imputation

We then assessed whether genotypes could be accurately derived from the samples, which would permit eQTL and ASE analysis. After removing genetically identical samples and additional quality control (see Methods), we had a diverse data set of 1,262 unique individuals (Figure S5). We genotyped 321,415 common SNPs that had a GQ ≥ 30, a call-rate ≥ 80% and a MAF ≥ 0.05. We observed that the total number of high-quality genotype calls that could

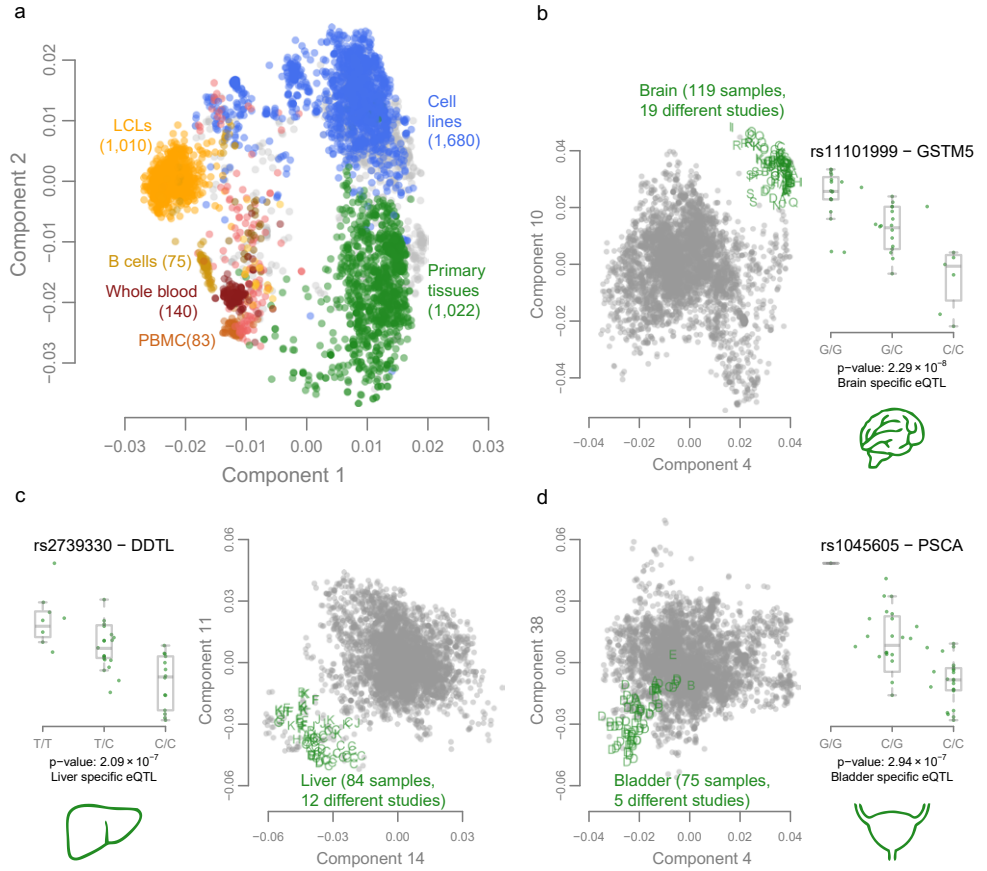



***Figure 2: PCA on expression data of all 4,978 samples that passed expression QC.*** *Panel a shows that the first 2 expression components show clear separation between primary tissue samples in green, cell lines (HeLa, K562, Hep G2, etc.) in blue and hematologic tissues and cell types in different shades of red and yellow. The other panels show the two best discriminating components for the different primary tissues. Each letter represents a sample from a distinct study showing that this clustering is not driven by a study specific effect. For each of these three primary tissues, we show an example eQTL effect specific to these tissues.*

be made per sample strongly correlated with the total number of sequenced bases per sample (Pearson $r^2$ = 0.85, Figure 3a, Figure S6a). As expected, genotypes could only be called in regions where genes are expressed (Figure 3b, Figure S7).

To ascertain the accuracy of the genotype calls, we compared the RNA-seq-derived genotypes with actual DNA-based genotype calls that were available for 459 Geuvadis [24] lymphoblastoid cell-line (LCL) samples that were part of the 1,262 samples. For the ASE analyses we only used high-quality genotype calls (GQ ≥ 30, see Methods), and for this subset of SNPs we observed a median concordance of 1 over all minor allele frequency (MAF) ranges (mean concordance = 0.96).

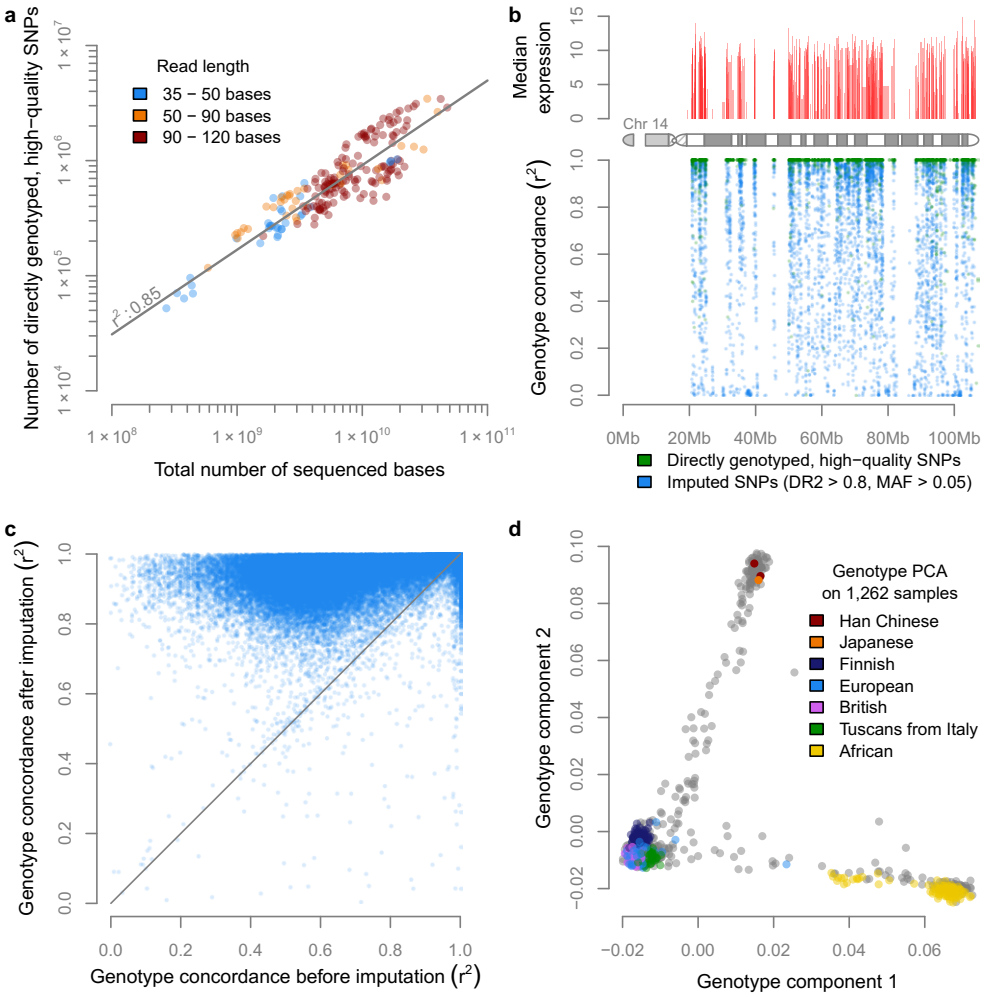



***Figure 3: Genotype concordance and genotype PCA on the 1,262 unique individuals that passed genotype QC.*** *a) We observe a strong correlation between the number of sequenced bases and the number of high-quality genotype calls. b) We show that genotyping is only possible in regions with gene-expression c) In 95% of the variants imputation increased genotyping concordance. d) PCA on the genotypes of all 1,262 samples reveals population structure with the expected European, Asian and African clusters.*

In order to perform eQTL analysis using non-coding SNPs as well, we used genotype imputation (see Methods) to increase the number of common SNPs to 1,081,155 (predicted dosage $r^2$ (DR2) ≥ 0.8 and MAF ≥ 0.05). Since most of the Geuvadis samples (used for determining the genotyping concordance) are part of the 1000 Genomes Project [35], we did not use the 1000 Genome reference panel, but used an independent panel – the Genome of the Netherlands (GoNL) [29,30] – to ensure that the genotype concordance measurements were not artificially inflated. The median genotype concordance $r^2$ for the 1,081,155 imputed SNPs for the European Geuvadis samples was $r^2$ = 0.92. When also including the African

Geuvadis samples, the genotype concordance decreased somewhat (median $r^2$ = 0.83), because the GoNL imputation reference panel only contained Dutch samples. The genotype concordance of directly genotyped common variants (irrespective of the genotype quality) showed an increased genotype concordance in 95% of the cases after imputation (Figure 3c). We also observed that, prior to imputation, there is a large difference in genotype concordance of variants in low-expressed genes compared to variants in highly expressed genes, where genotype calling is easier. However, this difference became much smaller after imputation, indicating that it is often possible to accurately call genotypes of SNPs that map within low-expressed genes by using imputation (Figure S8).

PCA on the imputed genotypes confirmed that the major components correctly captured the different ancestries of the individual samples (Figure 3d). These results also permitted us to stratify the samples into three different groups corresponding to European, African and Asian individuals, and to perform eQTL meta-analyses which are more robust than conducting an eQTL analyses on all samples combined in regions of the genome where allele frequencies differ substantially between populations.



## *Cis*-eQTL mapping

We then ascertained the reliability of conducting eQTL analysis when using genotypes derived solely from the RNA-seq data. To do so, we tested how many *cis*-eQTLs could be found in the Geuvadis LCL samples when using the RNA-seq derived and imputed genotypes, and also how far they could be replicated using the actual DNA-based genotypes that were available for these samples. An eQTL meta-analysis on the Geuvadis samples using the RNA-seq derived and imputed genotypes (see Methods) resulted in 8,765 unique genes with a significant *cis*-eQTL effect (at a false discovery rate (FDR) ≤ 0.05, Table 1). Of these, 95% could be replicated significantly using the actual DNA-based genotypes (99.95% with the same allelic direction), indicating that eQTL mapping using RNA-seq-derived genotypes is certainly possible for datasets that reflect one sequencing strategy (paired-end 75 bp reads) and one cell type.

We then performed eQTL analyses (all at FDR ≤ 0.05) on the non-Geuvadis samples while attempting to replicate the identified eQTLs in the Geuvadis data (DNA-based genotypes). We realized that these replication rates would be partly influenced by tissue-specific eQTL effects and first therefore investigated the non-Geuvadis LCL samples (n = 55). Given the sample size, we only identified 80 significant eQTL genes, but we could replicate 93% of these in the Geuvadis samples, all with the same allelic direction (Table 1). Subsequently we performed an eQTL mapping using all the hematological non-Geuvadis samples (n = 210), in which we identified 982 significant eQTL genes, of which 82% could be replicated in the Geuvadis samples (98.51% with identical allelic direction). Finally, we also included the primary tissue non-Geuvadis samples and identified 3,291 significant eQTL genes, of which 71% could be replicated in the Geuvadis LCL samples (98.34% with identical allelic direction).

We then performed an eQTL analysis on all the Geuvadis and non-Geuvadis samples, which identified significant *cis*-eQTLs for 8,034 unique genes (of which 84% were replicated in the Geuvadis DNA-seq-based eQTL data, 99.87% with identical allelic direction). This is fewer

| Dataset | No. European samples | No. African samples | No. Asian samples | No. unique eQTL genes | No. eQTLs replicated in Geuvadis LCLs, based on DNA derived genotypes | Replication | Percentage replicated with identical allelic direction |
|---|---|---|---|---|---|---|---|
| Geuvadis LCLs, RNA-seq derived genotypes | 371 | 88 | 0 | 8,765 | 8,301 | 95% | 99.95% |
| Non-Geuvadis LCLs, RNA-seq derived genotypes | 29 | 26 | 0 | 80 | 74 | 93% | 100% |
| Non-Geuvadis, all hematological samples (including LCLs) , RNA-seq derived genotypes | 129 | 81 | 0 | 982 | 803 | 82% | 98.51% |
| Non-Geuvadis, all samples, RNA-seq derived genotypes | 577 | 109 | 117 | 3,291 | 2,345 | 71% | 98.34% |
| All samples, RNA-seq derived genotypes | 948 | 197 | 117 | 8,034 | 6,728 | 84% | 99.87% |

***Table 1: Overview of identified eQTL genes (FDR < 0.05) that were significant in different subsets of the data.*** *eQTL expression quantitative trait locus; LCL lymphoblastoid cell-line*

than in the analysis on only the Geuvadis samples due to the fact we were dealing with many different tissues in the combined analyses: the small Geuvadis LCL-specific eQTL effects became diluted by the non-LCL samples, leading to fewer *cis*-eQTL effects.

On comparing the 8,765 eQTL genes identified in the Geuvadis samples to the 3,291 eQTL genes identified in the non-Geuvadis samples, we observed that 2,374 genes were identified in both datasets (Figure 4). As expected, the expression levels of the 903 genes that could not be identified in the Geuvadis data were significantly lower in the Geuvadis data (Wilcox p-value $4.05 \times 10^{-61}$) than in the non-Geuvadis samples.

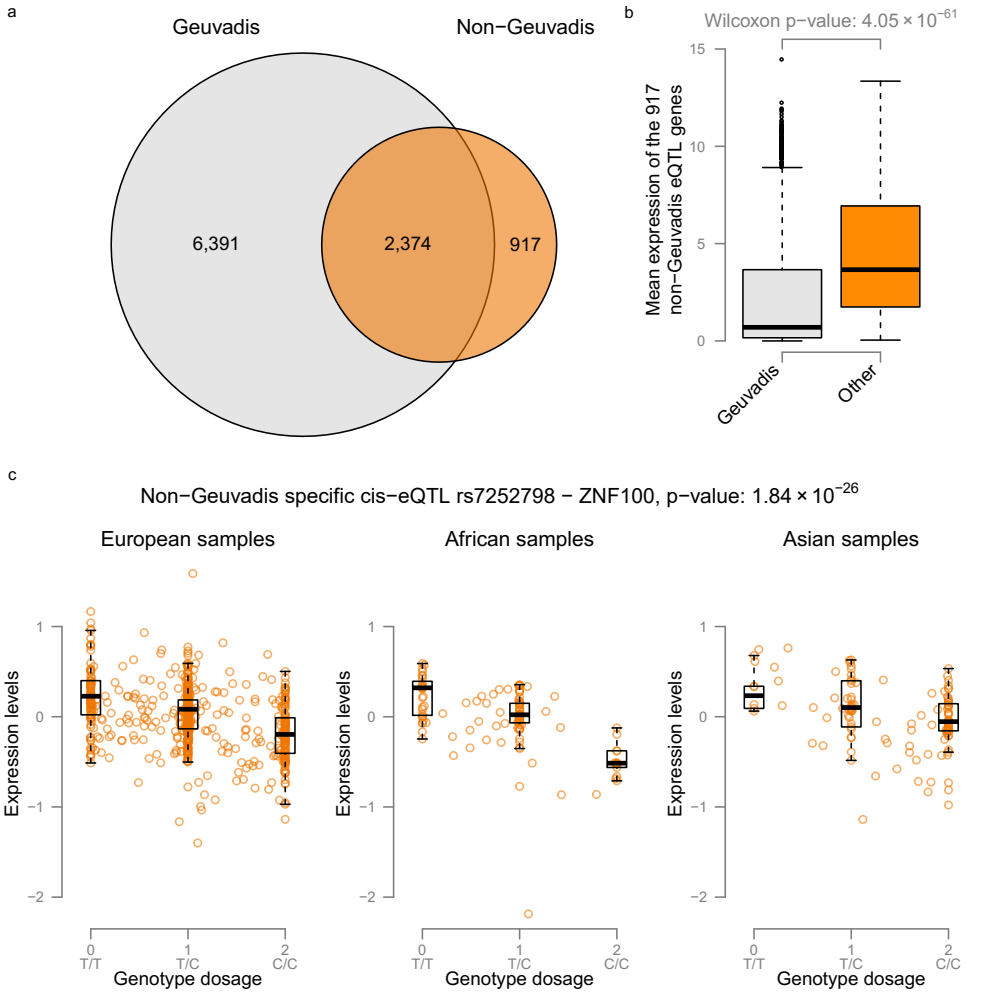



***Figure 4: Geuvadis eQTLs vs non-Geuvadis eQTLs.*** *a) When performing eQTL analysis on the non-Geuvadis samples we identify 3,559 significant eQTL genes (FDR < 0.05). 903 of these genes are not detected when using the Geuvadis samples. b) Comparing the expression levels of the genes not identified in the Geuvadis samples we observe that in general these genes are much more abundantly expressed in the non-Geuvadis samples. c) Example eQTL effect of rs7252798 affecting expression levels of ZNF100 that is only identified in the non-Geuvadis samples.*

## Tissue-specific eQTL mappings

Since the public RNA-seq data represents many different tissues, we assessed whether it is possible to perform tissue-specific eQTL mapping on samples of the same tissue generated by different laboratories. We performed separate eQTL mapping on four tissues: brain (42 samples from 7 studies), liver (42 samples from 8 studies), bladder (45 samples from 3 studies) and breast (50 samples from 4 studies), since they had eQTL data available on at least 40 samples. This resulted in 121 unique *cis*-regulated genes in brain (32% not detected in

Geuvadis), 86 genes in liver (37% not detected in Geuvadis), 65 genes in bladder (38% not detected in Geuvadis) and 43 genes in breast (19% not detected in Geuvadis). As expected, for genes with eQTLs that did not replicate in Geuvadis, we found that the expression in the respective tissues was higher (Figure S9). A representative example is shown for SNP rs11101999, which showed a *cis*-eQTL effect only in brain tissue, on glutathione S- transferase mu 5 (*GSTM5*), a gene that is specifically expressed in brain (Figure 2b, Figure S10a).

We saw that various GWAS disease-associated genetic variants showed tissue-specific eQTL effects: in the liver samples we found rs2739330 that significantly *cis*-regulates the D-dopachrome tautomerase-like gene (*DDTL*), which is known to be associated with concentrations of liver enzymes in plasma (Figure 2c, Figure S10b) [36]. Another example is rs1045605, which affects Prostate stem cell antigen (*PSCA*) gene-expression levels in bladder samples (Figure 2d, Figure S10c) and which is in near-perfect linkage disequilibrium (LD) with rs2294008 ($r^2$ = 0.98 and $D'$ = 0.998), a variant that is associated with both gastric [37] and bladder [38] cancers.

## Allele-specific expression

We mapped ASE by fitting a binomial distribution per SNP using maximum likelihood estimation and then assessed significance by using a likelihood ratio test. Similar to our study of the eQTLs, we first investigated the Geuvadis samples and identified 16,217 ASE SNPs (FDR ≤ 0.05) using the RNA-seq-derived high-quality genotypes (GQ ≥ 30). We compared these results to an ASE analysis using the actual DNA-based genotypes of the Geuvadis samples. 9,221 out of the 9,341 (99%) ASE SNPs that could be tested were replicated using DNA-based genotypes (99.87% with identical allelic direction). Vice versa, on using the Omni DNA genotypes to detect ASE effects, only 232 of them were not found when using RNA-seq genotyping. We next assessed the concordance with the eQTL results: since eQTL mapping and ASE analysis both test the association between genetic variation and gene-expression, we expected the same allele to be more highly expressed in both methods. Indeed, we observed that for 93% of the 1,552 SNPs that showed significant ASE and eQTL effects on the same gene, the allelic direction was consistent. This percentage is similar to another comparison of ASE and eQTL effects, in which 90% of the overlapping eQTL and ASE effects were in the same direction [39].

To gain maximum power to detect ASE effects we then performed an analysis on all 1,262 samples and identified 71,214 significant ASE SNPs (FDR ≤ 0.05), of which 4,781 pertained to rare SNPs with a MAF < 0.01 and to 9,018 low-frequency SNPs with a MAF between 0.01 and 0.05. We again compared these ASE SNPs to the eQTL mapping performed on all samples and observed that for 85% of the 1,956 SNPs that showed both significant ASE and eQTL effects on the same gene, the allelic direction was consistent.

It has been reported that nonsense SNPs show ASE with lower expression of the deleterious allele due to nonsense-mediated decay [9,24]. To investigate this, we annotated the ASE SNPs using SnpEff [40] and indeed found that, for nonsense SNPs, the alternative allele is often less expressed than the reference allele (Figure 5c), whereas for other variants we did not observe this bias (Wilcoxon p-value $2.19 \times 10^{-60}$). We also investigated the effect and expected functional impact of the ASE SNPs as predicted by SnpEff. We observed that the SNPs with an expected high functional impact according to SnpEff (Wilcoxon p-value $7.52 \times 10^{-3}$) and

those that introduce a stop codon (Wilcoxon p-value 3.66 x $10^{-3}$) again showed less expression of the alternative allele (Figure S11).

We further investigated the functional consequences of common ASE variants by assessing if they were present in the GWAS catalog [41]. We identified 5 ASE variants with GWAS associations (Table S2). For example, rs12203592 is located in the *IRF4* gene and the T allele increases the risk of non-melanoma skin cancers (NMSCs) [36] and increases expression levels (Figure 5a, Figure S12).

Since ASE mapping also permits the identification of rare and low-frequency variants, we were able to identify 34 variants known to be pathogenic in a Mendelian setting according to the ClinVar database (Table S2) [28]. One example is rs72550870, located in the *MASP2* gene, where we observed an ASE effect (Figure 5b). It has already been shown that the alternative C allele causes MASP2 deficiency with a recessive inheritance pattern and that heterozygous individuals have significantly lower MASP2 protein levels than individuals homozygous for the wild-type allele [42]. Our ASE results show exactly the same effect on gene-expression levels. Here it is important to note that the *MASP2* gene is predominantly expressed in

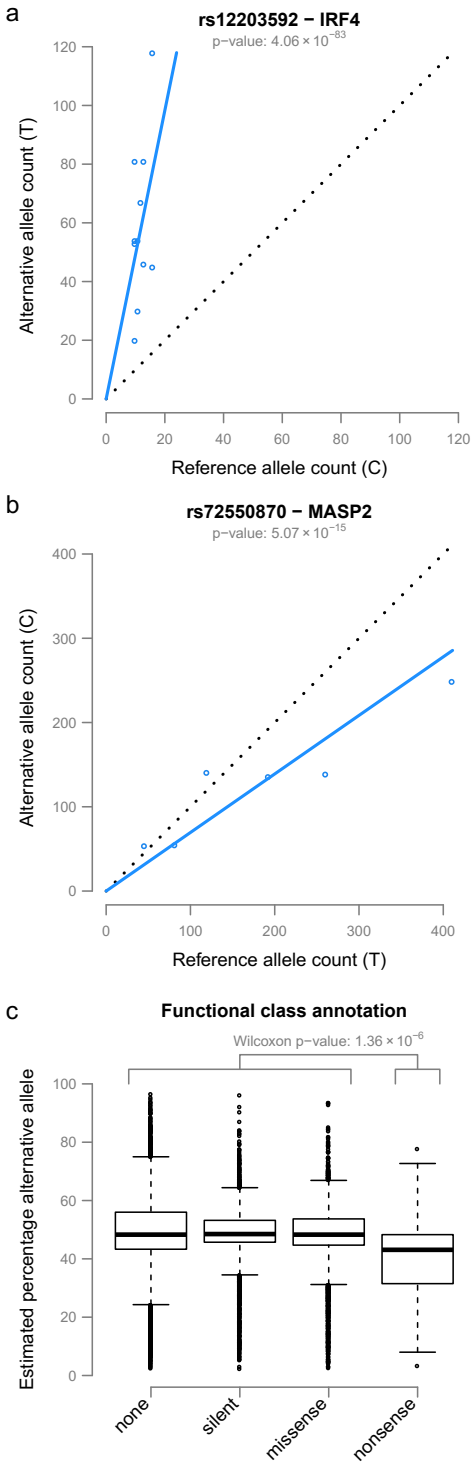



***Figure 5: Example ASE effects and direction of ASE effects over different functional classes.*** *a) ASE of rs12203592 located within the IRF4 gene. The T allele is more abundantly expressed and increases the risk of non-melanoma skin cancers (NMSCs). b) rs72550870, located in MASP2 shows lower expression for the alternative C allele, known to cause MASP2 deficiency (MASPD). c) All significant ASE SNPs were annotated with functional class information. As expected, nonsense mutations often lead to lower expression levels, in contrast to ASE effects in other functional classes.*

the liver (Figure S12a) and that all the samples showing this ASE effect were liver samples, demonstrating the power of a dataset containing multiple tissue types to target a variety of diseases.

The ASE effects obtained can be queried at http://molgenis.org/ase using the MOLGENIS software platform [43]. It is possible to query a specific variant or search for all ASE variants associated to a specific gene. All our data is available for downloading from this website, including the genotypes, expression data, principal components, eQTLs and ASE effects.

## Conclusions

We have shown that it is possible to reliably map eQTLs and perform ASE analyses by calling genotypes directly from RNA-seq data of 1,262 human samples, despite the fact that this data originated from different tissues, was obtained from different laboratories, and was generated using different sequencing techniques.

We called genotypes using GATK, and subsequently imputed using Beagle, while using the GoNL reference panel, to permit unbiased genotype concordance analyses for the Geuvadis samples. We observed that imputation improved the genotype concordance substantially. We find that it is more difficult to impute non-European samples, which is due to our imputation towards the GoNL reference panel. Although GoNL has been shown to yield high-quality imputation for European samples [29], its performance has not yet been assessed on non-European populations. It is therefore unreasonable to expect it to perform equally well for Asian or African samples. We expect that a more diverse reference panel will help to resolve this issue in the future.

In this study we had to annotate each sample manually because we found that the annotations available for the different sequence runs were typically limited and inconsistent in terminology. A second reason was that the sample annotations were scattered over multiple databases. Thirdly, although in general the ENA provided better-structured annotations, the information in the Sequence Read Archive was typically more extensive, providing a total of 572 different annotation fields, of which 16 could refer to the tissue of origin. We expect that with more consistency in sample annotation, future large-scale integration of public RNA-seq datasets using automated sample annotation will become feasible.

We have demonstrated that it is possible to run tissue-specific eQTL mapping in public RNA-seq data. We showed that when using only 42 liver samples (originating from eight different labs), it was possible to identify eQTLs that are liver-specific, some of which had been detected by earlier GWAS studies as associated with liver-specific traits. Although the concept of tissue-specific eQTLs is not new, our results demonstrate that different research groups investigating a specific disease in a particular tissue can combine their data in order to conduct joint eQTL mapping. This strategy will certainly prove useful for tissues that are difficult to obtain.

We were able to identify ASE effects for various rare disease-causing variants using only 1,262 samples. We expect our approach will also be useful for studying many other rare pathogenic variants in the near future, because the number of publicly available RNA-seq samples is growing exponentially: at the end of July 2014, the ENA contained 14,831 human

RNA-seq samples, which is over 1.5 times the number of samples that we investigated here. Additionally, the read-depth and read-length per sample are steadily increasing (Figure S6b), permitting more sensitive eQTL and ASE analyses (on less expressed genes) on the newly deposited samples. Although a subset of the 1,262 samples, used for ASE analysis, reflect cancer samples, we did not observe that inclusion or exclusion of cancer samples resulted in fewer significant SNPs that showed an ASE effect (proportional to the number of samples omitted), but not to differences in the direction of ASE effects (data not shown).

We anticipate that, with more samples available, eQTL and ASE effects will be detectable for many more (rare) variants. These will be of particular relevance for rare genetic variants that have been identified in patients by exome or genome sequencing but for which the clinical significance remains unknown. If such rare alleles are also present in any of the publicly available RNA-seq samples and they are seen to reduce expression levels strongly, this might suggest they have a loss-of-function effect, strongly warranting clinical follow-up. As such, our approach could well complement existing computational prediction algorithms (that have so far been based primarily on allele frequencies and conservation information), and help speed up the identification of disease-causing mutations, leading to better treatment options and well-informed decisions for patients and their families.



## Acknowledgements

We thank the Target project (http://www.rug.nl/target) for providing the compute infrastructure, the MOLGENIS developers Fleur D. L. Kelpin and Bart Charbon for help with the online query tool, Jackie Senior for editing the manuscript, and Ye Chun for helpful comments. This work was supported by the Netherlands Organization for Scientific Research [NWO-VENI grant 916.10.135 to LF and NWO-VIDI grant 016.146.374 to LF], and a Horizon Breakthrough grant from the Netherlands Genomics Initiative [grant 92519031 to LF]. The research leading to these results also received funding from the European Community's Health Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 259867. This study was financed in part by the SIA-raakPRO subsidy for the BioCOMP project, and in part by Rainbow grants 2 and 3 from BBMRI-NL, a research infrastructure financed by the Netherlands Organization for Scientific Research [NWO project 184.021.007 to LF, MS, PD, DVZ].

## Author contributions

PD and DVZ performed most of the data processing and analysis. PD, DVZ and LF designed the study and wrote the manuscript. LF and MAS supervised the work. MdH, MvdS, MJB, JK, JF and JvdV performed part of the analyses. KMA, CW and RJS contributed to the study design and discussion.

## Additional material

The following supplements are available with the on-line version of this paper.

Figure S1 PCA on expression values shows strong outliers that are removed from the analysis. The 521 outliers of the first component (left of the red line) where removed from our analyses.

Figure S2: Correlation of principal components vs different confounders.

Table S3: Table with sample annotations for the 4,978 samples that passed quality control.

Figure S4: Identification of duplicate samples. We used a cut-off of 78% identity to select duplicate markers. By using this cut-off level we could identify all the duplicates, which we expected based on the annotations. The reason that we have multiple peaks above this cut-off is due to the difference in genotyping quality among the samples. Panel b is the enlargement of the lower part of panel a.

Figure S5: Expression of *XIST* and chromosome Y genes. We show a clear separation of males and females using both *XIST* expression and chromosome Y expression. In two cases, the samples were annotated as male but clustered within the females; these are likely mis-annotations.

Figure S6: Overview of the properties of the 1,262 samples used for eQTL and ASE analyses. Here we show that the samples which we successfully genotyped and used for the eQTL and ASE analysis still show high heterogeneity in sequencer models (a), read layout (b), sampled tissue (c), cancer status (d), total number of reads (e), read length (f).

Figure S7: The relation between sequencing depth and the number of high quality genotypes. a) We observe a strong relation between the number of sequenced bases and the number of high quality genotypes, we do not observe that paired end sequencing improves genotyping. b) We observe that newer samples usually have more bases sequenced.

Figure S8: Overview of genotyping accuracy and gene-expression over all chromosomes.

Figure S9: Relation between gene-expression levels and genotype concordance before and after imputation. Genotype concordances in all Geuvadis samples (a) and European Geuvadis samples (b) of common SNPs (MAF ≥ 0.05, DR2 ≥ 0.8) before and after imputation grouped by the median expression levels of their genes.

Figure S10: Expression of tissue-specific *cis*-eQTL genes vs Geuvadis expression. We find that genes with tissue specific *cis*-eQTLs are more abundantly expressed in the respective tissues compared to the Geuvadis samples in which we did not observe the *cis*-eQTLs.

Figure S11: Expression of example tissue-specific eQTL in different tissues. Here we show three example tissue-specific eQTL genes: a) *GSTM5*, brain-specific. b) *DDTL*, liver-specific. c) *PSCA,* bladder specific.

Figure S12: Predicted functional impact of ASE variants. The annotation of ASE SNPs predicted impact and effect was performed using SnpEff. (a) Most of the high-impact SNPs have lower expression of the alternative allele. (b) The majority of the SNPs introducing a stop codon have lower expression of the alternative allele.

Table S13: Overview of detected ASE variants with ClinVar or GWAS annotation

Figure S14: Expression of *MASP2* gene is liver-specific and the expression of *IRF4* gene is hematopoietic-specific. (a) *MASP2* gene has higher expression in liver compared to other tissues. (b) *IRF4* gene has higher expression in hematopoietic cells compared to other tissues.

## References


1. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6,** e1000888 (2010).
2. Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **42,** 295–302 (2010).
3. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).
4. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–72 (2010).
5. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464,** 773–7 (2010).
6. Lappalainen, T. *et al.* Epistatic selection between coding and regulatory variation in human evolution and disease. *Am. J. Hum. Genet.* **89,** 459–63 (2011).
7. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* (2013). doi:10.1101/gr.155192.113
8. Zhernakova, D. V. *et al.* DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. *PLoS Genet.* **9,** e1003594 (2013).
9. Kukurba, K. R. *et al.* Allelic expression of deleterious protein-coding variants across human tissues. *PLoS Genet.* **10,** e1004304 (2014).
10. Heap, G. A. *et al.* Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Genet.* **19,** 122–34 (2010).
11. Pastinen, T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat. Rev. Genet.* **11,** 533–8 (2010).
12. Fehrmann, R. S. N. *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.* **7,** e1002197 (2011).
13. Wright, F. A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46,** 430–7 (2014).
14. Fu, J. *et al.* Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet* **8,** e1002431 (2012).
15. Nica, A. C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* **7,** e1002003 (2011).

16. Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325,** 1246–50 (2009).
17. The Gtex Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–5 (2013).
18. Papadopoulos, P. *et al.* Developments in FINDbase worldwide database for clinically relevant genomic variation allele frequencies. *Nucleic Acids Res.* **42,** D1020–D1026 (2014).
19. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508,** 469–476 (2014).
20. Breu, F. *et al.* Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* **93,** 641–651 (2013).
21. Byelas, H. *et al.* Scaling bio-analyses from computational clusters to grids. in *CEUR Workshop Proceedings* **993,** (2013).
22. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
23. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinforma.* **25,** 3207–3212 (2009).
24. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).
25. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43,** 491–498 (2011).
26. Browning, B. L. *et al.* Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194,** 459–71 (2013).
27. Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.* **44,** 631–5 (2012).
28. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980–D985 (2014).
29. Deelen, P. *et al.* Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *Eur. J. Hum. Genet.* **In press,** (2014).
30. The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46,** 818–825 (2014).
31. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
32. Robinson, M. D. *et al.* A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11,** R25 (2010).
33. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma.* **25,** 2078–2079 (2009).
34. Derrien, T. *et al.* Fast computation and applications of genome mappability. *PLoS One* **7,** e30377 (2012).
35. Howie, B. *et al.* Genotype imputation with thousands of genomes. *G3 genes - genomes - Genet.* **1,** 457–70 (2011).
36. Zhang, M. *et al.* Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans. *Hum. Mol. Genet.* **22,** 2948–2959 (2013).

37. Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. *Nat Genet* **40,** 730–740 (2008).
38. Wu, X. *et al.* Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. *Nat Genet* **41,** 991–995 (2009).
39. Lagarrigue, S. *et al.* Analysis of allele-specific expression in mouse liver by RNA-seq: a comparison with cis-eQTL identified using genetic linkage. *Genetics* **195,** 1157–1166 (2013).
40. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin).* **6,** 80–92 (2012).
41. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001–D1006 (2014).
42. Stengaard-Pedersen, K. *et al.* Inherited deficiency of mannan-binding lectin-associated serine protease 2. *N. Engl. J. Med.* **349,** 554–60 (2003).
43. Swertz, M. A. *et al.* The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinformatics* **11,** S12 (2010).

# 5

**Nature Medicine, 2016**

Yang Li[1,*], Marije Oosting[2,*], Patrick Deelen[1,3], Isis Ricaño-Ponce[1], Sanne Smeekens[2], Martin Jaeger[2], Vasiliki Matzaraki[1], Morris A. Swertz[1,3], Ramnik J. Xavier[4,5], Lude Franke[1], Cisca Wijmenga[1,#], Leo A.B. Joosten[2,#], Vinod Kumar[1,#], Mihai G. Netea[2,#]

# Inter-individual variability and genetic influences on cytokine responses against bacterial and fungal pathogens

## Abstract

Little is known about the inter-individual variation of cytokine responses to different pathogens in healthy individuals. To systematically describe cytokine responses elicited by distinct pathogens, and to determine the impact of genetic variation on cytokine production, we profiled cytokines produced by peripheral blood mononuclear cells from 197 individuals of European origin from the 200 Functional Genomics (200FG) cohort within the Human Functional Genomics Study (www.humanfunctionalgenomics.org), obtained over three different years. By comparing bacteria- and fungi-induced cytokine profiles, we show that most cytokine responses are organized around a physiological response to specific pathogens, rather than around a particular immune pathway or cytokine. We then correlated genome-wide SNP genotypes with cytokine abundance and identified six cytokine QTLs. Among them, a cytokine QTL at NAA35-GOLM1 locus markedly modulates IL-6 production in response to multiple pathogens, and associated with susceptibility to candidemia. Furthermore, the cytokine QTLs we identified are enriched among SNPs previously associated with infectious diseases and heart diseases. These data reveal and begin to explain the variability in cytokine production by human immune cells in response to pathogens.

1 University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands
2 Department of Internal Medicine and Radboud Center for Infectious Diseases, Radboud University Medical Center, Nijmegen, The Netherlands
3 University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, The Netherlands
4 Center for Computational and Integrative Biology and Gastrointestinal Unit, Massachusetts General Hospital, Harvard School of Medicine, Boston, MA 02114 USA
5 Broad Institute of MIT and Harvard University, Cambridge, MA 02142 USA
* Equal contributions
# Equal contributions
Corresponding authors: Vinod Kumar & Mihai G. Netea

## Introduction

Infections have shaped the human immune system [1,2], with genetic variability contributing to differential susceptibility to infections [3,4]. However, a specific genetic variant that may confer protection against one infection could prove deleterious for other infections, and this is determined by the local infection burden in different geographical locations [2]. Moreover, the shaping of the immune system by infections also has direct consequences for the susceptibility to autoimmune and inflammatory diseases [5–7]. Unravelling the interplay between environmental factors, such as infections, and the genetic variation in a population is crucial for understanding the pathogenesis of common autoimmune and infectious diseases, and for designing novel therapeutic strategies.

The study of healthy population-based cohorts in the context of appropriate microbial stimulations can be used to assess inter-individual variability and to identify genetic loci that regulate immune responses [8–12]. However, practically all the genome-wide studies done to date have emphasized the regulatory effect of genetic variation on gene expression by focusing on transcript abundance [8–12]. Since protein quantities are more precise regulators of cellular phenotypes [13], characterizing the genetic loci that regulate protein abundances and biological processes is a crucial next step towards mechanistic insights [14].



Here we stimulated peripheral blood mononuclear cells (PBMCs), rather than isolated immune cell populations, to capture interactions between different immune cell types (e.g. between monocytes and T cells) that are very important for the natural immune responses. We studied inter-individual and inter-stimulus variation in production of cytokines, and we identified independent genome-wide significant cytokine quantitative trait loci (cQTLs). The regulatory consequences of these cQTLs on downstream genes were characterized by performing the expression-QTL (eQTL) using stimulation-specific expression data [15]. By comparing the bacterial and fungal induced cytokine profiles and cQTLs, we show that the genetic variability in the immune genes/pathways is organized around a physiological response to specific pathogens, rather than a response aiming to modulate production of a specific cytokine. In addition, we have identified and validated a cytokine QTL that reveals a novel *trans*-regulatory network in the context of cytokine responses to important human pathogens.

## Results

### Stimulation increases inter-individual variability in cytokine levels

To systematically determine the impact of genetic variation on cytokine production, we obtained PBMCs from 197 individuals of European origin from the 200 Functional Genomics (200FG) cohort within the Human Functional Genomics Study (www.humanfunctionalgenomics.org) in three different years (Supplementary table 1), and profiled cytokines secreted in response to a variety of bacterial and fungal pathogens (Supplementary table 2). In the first study we measured seven cytokines induced by ten different stimuli in 73 healthy volunteers (year 2009 cohort). After stringent quality control of cytokine distributions (see Methods), we obtained a total of 62 (cytokine-stimuli pairs) different cytokine measurements (Supplementary table 3). Cytokine production follows a non-Gaussian or bimodal

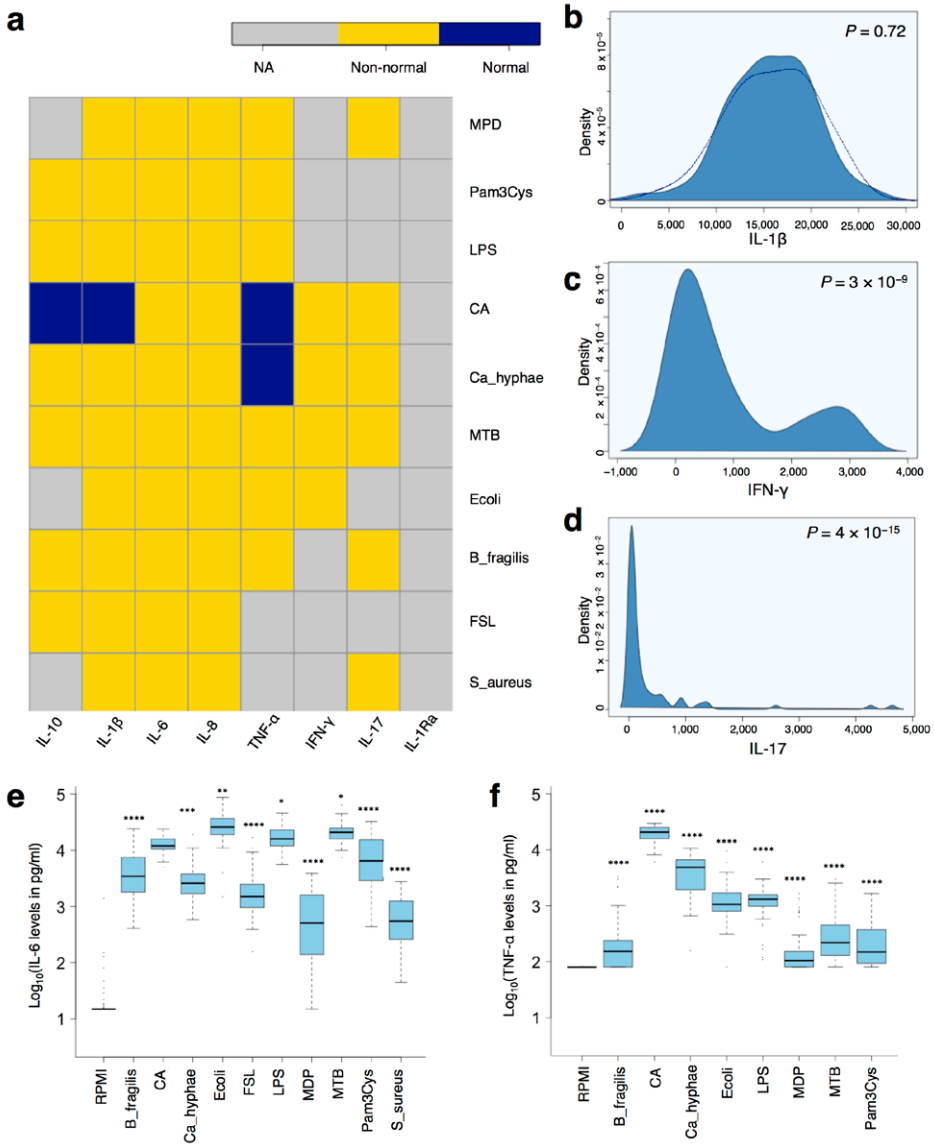



***Figure 1: Inter-individual variability in cytokine production upon PBMC stimulation.*** *(a) PBMCs were cultured with the indicated pathogen-related stimuli for 24 hours time period. Cytokine abundance was measured by ELISA. The distributions of raw cytokine levels from the 2009-cohort were tested using the Shapiro-Wilk normality test Blue indicates normal (P > 0.05); yellow indicates non-normal (p<0.05); and grey indicates distributions not tested due to unavailability of the measurements in 2009 dataset. (b) Distribution of Candida albicans-induced IL-1β. (c-d) Distributions of Candida albicans-induced IFN-γ (c) and Candida albicans-induced IL-17 (d). P values shown in the panels (b-d) were obtained from the Shapiro-Wilk normality test. (e-f) Log-transformed abundance of (e) IL-6 and (f) TNF-α …*

distribution, with a few exceptions (Figure 1a-d). Individuals exhibit significantly increased inter-individual variability ($P < 0.01$) in cytokine secretion upon stimulation, compared to basal unstimulated state (Figure 1e-f and Supplementary figure 1). We obtained similar results from the other two datasets measured in 2011 and 2013 (Supplementary figure 2).

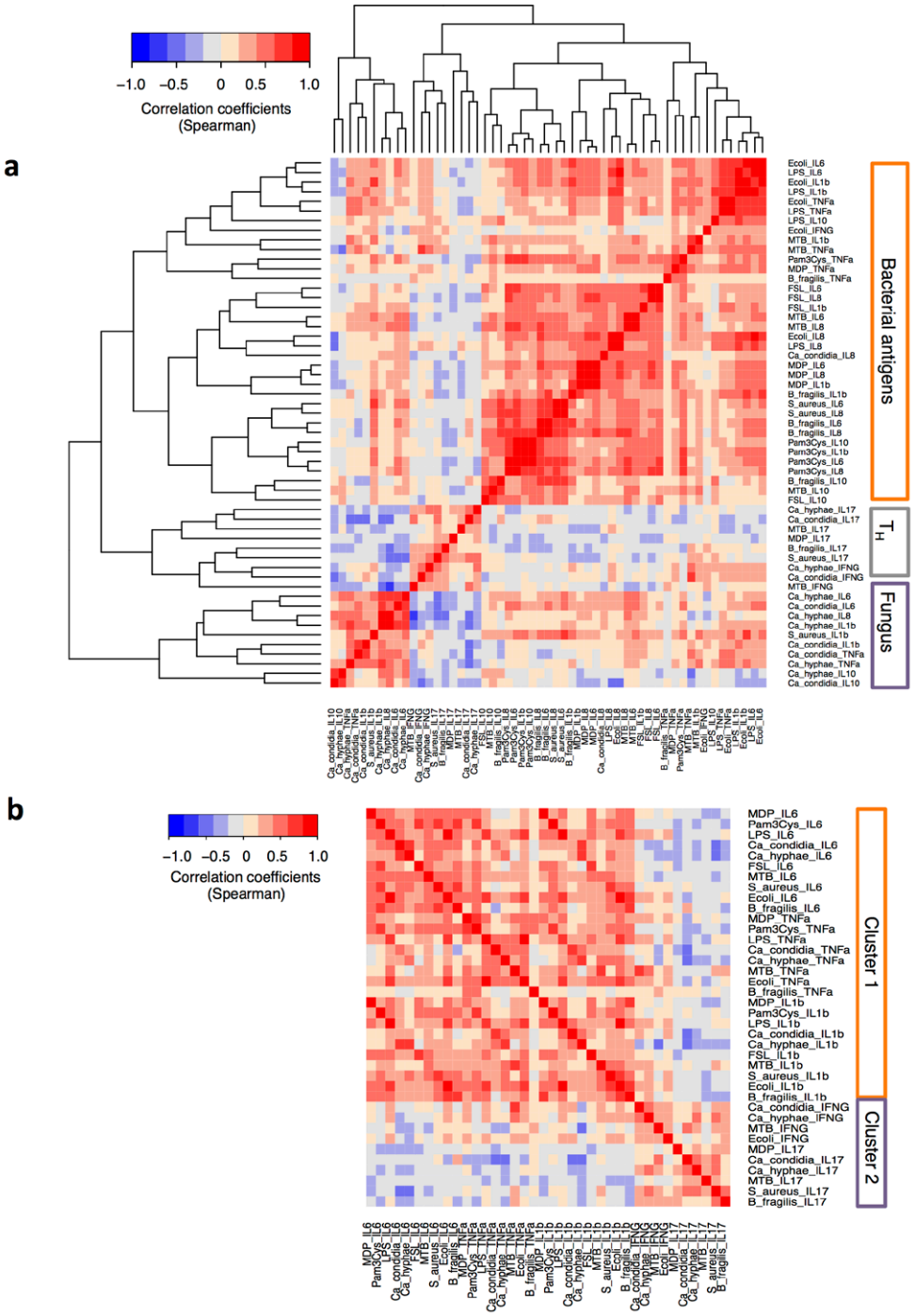
## Cytokine responses are organized in a pathogen-specific manner

It is possible that an individual could be either a high or low responder when considering all cytokines (e.g. TNF-α, IL-6 and IL-10) produced in response to one microorganism. Alternatively, an individual could be a high or low responder for a particular cytokine in response to stimulation with any type of pathogen. To examine this, we performed a unsupervised clustering of the cytokine responses induced by the various pathogens and microbial ligands. Correlations between levels of various cytokines were found in response to stimulation with a certain pathogen, rather than within a cytokine pathway, and this conclusion was validated by additional analyses in the cohort studies performed in 2011 and 2013 (Figure 2a) (Supplementary figures. 3 and 4). For example, bacterial (LPS, *E. coli* and *M. tuberculosis*) induced cytokines (TNF-α, IL-6 and IL-10) were strongly clustered together and were clearly separated from fungus (*C. albicans*) induced cytokine cluster (Figure 2a).



IFN-γ and IL-17 production were exceptions to this rule, however; the magnitude of IFN-γ or IL-17 production by PBMCs from any individual correlated independently of the identity of the pathogen stimulus (Figure 2a). This suggests an important evolutionary role for $T_H17$ responses, which may be a general host defence pathway for both bacteria and fungi. Secondly, the differentiation of naïve T cells into $T_H1$ or $T_H17$ effector lymphocytes is under the control of monocyte-derived cytokines [16]. Therefore, a strong or weak monocyte-dependent cytokine production capacity may be associated with strong or weak helper T cell responses. However, upon pair-wise correlation of cytokines we observed two clusters (Figure 2b) in which Cluster 1 consists of cytokines mainly produced by monocytes, while Cluster 2 consists of cytokines known to be released mainly by T-cells. Although concentrations of IL-12, IL-18 and IL-23 were very low in our system, and could not be used to assess correlations (data not shown), other monocyte-derived cytokines such as IL-1$\beta$ and IL-6, which have been reported to induce $T_H17$ responses [17], were easily detectable. However, there was a poor correlation between monocyte-derived cytokine production and T cell cytokine production (Figure 2b-e). Moreover, $T_H1$ and $T_H17$ responses did not strongly correlate with each other, although the correlation was somewhat stronger than between monocyte and lymphocyte responses. This was also consistent when we focused specifically on one type of stimulation. For example, we observed strong correlation between *Candida*-induced IL-6

*… produced upon indicated stimulation. The length of the box in the box-plot is interquartile range (=Q3-Q1). The whiskers indicate the range of one and a half times the length of the box from either end of the box. The equality of variance of cytokine levels before and after stimulation was tested using Levene's test. The stars on the box plots depict the significance (*, P < 0.01; **, P < 0.001; ***, P < 0.0001; ****, P <0.00001). RPMI, unstimulated state; Bfrag, Bacteroides fragilis; CA, Candida albicans; CAhy, Candida albicans hyphae; Ecoli, Escherichia coli; FSL, lipopeptide; LPS, lipopolysaccharide; MDP, muramyl dipeptide; MTB, Mycobacterium tuberculosis; Pam3Cys, a synthetic triacylated lipopeptide; Saureus, Staphylococcus aureus. The data shown is from one independent experiment from 2009 cohort.*


a
−1.0 −0.5 0 0.5 1.0
Correlation coefficients
(Spearman)
Ecoli_IL6
LPS_IL6
Ecoli_IL1b
LPS_IL1b
Ecoli_TNFa
LPS_TNFa
LPS_IL10
Ecoli_IFNG
MTB_IL1b
MTB_TNFa
Pam3Cys_TNFa
MDP_TNFa
B_fragilis_TNFa
FSL_IL6
FSL_IL8
FSL_IL1b
MTB_IL6
MTB_IL8
Ecoli_IL8
LPS_IL8
Ca_condidia_IL8
MDP_IL6
MDP_IL8
MDP_IL1b
B_fragilis_IL1b
S_aureus_IL6
S_aureus_IL8
B_fragilis_IL6
B_fragilis_IL8
Pam3Cys_IL10
Pam3Cys_IL1b
Pam3Cys_IL6
Pam3Cys_IL8
B_fragilis_IL10
MTB_IL10
FSL_IL10
Ca_hyphae_IL17
Ca_condidia_IL17
MTB_IL17
MDP_IL17
B_fragilis_IL17
S_aureus_IL17
Ca_hyphae_IFNG
Ca_condidia_IFNG
MTB_IFNG
Ca_hyphae_IL6
Ca_condidia_IL6
Ca_hyphae_IL8
Ca_hyphae_IL1b
S_aureus_IL1b
Ca_condidia_IL1b
Ca_condidia_TNFa
Ca_hyphae_TNFa
Ca_hyphae_IL10
Ca_condidia_IL10
Bacterial antigens
TH
Fungus
b
−1.0 −0.5 0 0.5 1.0
Correlation coefficients
(Spearman)
MDP_IL6
Pam3Cys_IL6
LPS_IL6
Ca_condidia_IL6
Ca_hyphae_IL6
FSL_IL6
MTB_IL6
S_aureus_IL6
Ecoli_IL6
B_fragilis_IL6
MDP_TNFa
Pam3Cys_TNFa
LPS_TNFa
Ca_condidia_TNFa
Ca_hyphae_TNFa
MTB_TNFa
Ecoli_TNFa
B_fragilis_TNFa
MDP_IL1b
Pam3Cys_IL1b
LPS_IL1b
Ca_condidia_IL1b
Ca_hyphae_IL1b
FSL_IL1b
MTB_IL1b
S_aureus_IL1b
Ecoli_IL1b
B_fragilis_IL1b
Ca_condidia_IFNG
Ca_hyphae_IFNG
MTB_IFNG
Ecoli_IFNG
MDP_IL17
Ca_condidia_IL17
Ca_hyphae_IL17
MTB_IL17
S_aureus_IL17
B_fragilis_IL17
Cluster 1
Cluster 2

and IL-8 and between *Candida*-induced IL-10 and TNF-α, while IL-17 showed poor correlation with any of the other cytokines (Supplementary figure 5a-c). This demonstrates that one particular individual could be a high responder in terms of one set of cytokines but a low responder for other cytokines.

## Genome-wide cQTL mapping identifies cell-count independent cQTLs

We generated both genotype and cytokine data for 107 individuals (Supplementary table 1) We used the 2013 dataset as a discovery cohort to identify genome-wide significant cQTLs since this cohort contained the largest numbers of individuals (n=79). Genotyping was performed using Illumina HumanOmniExpressExome SNP chip and was imputed to obtain genotypes at ~ 7 million SNPs. We selected ~ 4 million SNPs that showed minor allele frequency ≥ 5% and passed other standard quality filters. The cytokine and genotype data available enabled us to study cQTLs for three stimuli: a Gram-negative stimulus (LPS), a mycobacterium (*M. tuberculosis*; MTB) and a fungus (*C. albicans*), which provided 18 measurements (3 stimulations×6 cytokines; IL-6, IL-8, IL-10, IL1-Ra, IL-1*β*, TNF-α). IFN-γ and IL-17 measurements were not available for the 2013 dataset. Upon quality check for cytokine distributions, we obtained 17 stimulation-cytokine pairs (Supplementary figure 2) for which the data were reliable to correlate with genotypes at ~ 4 million SNPs. Raw cytokine levels were first log-transformed then mapped to genotype data using a linear regression model with age and gender as covariates. This analysis revealed six significant cQTLs ($P < 5 \times 10^{-8}$) (Supplementary tables 4-5). We identified two independent cQTLs for *C. albicans*-induced IL-6 levels (Figure 3a-c), two independent cQTLs for MTB-induced IL-8 levels (Figure 3d-f), one for *C. albicans*-induced TNF-α and one for LPS-induced IL-10 levels while no cQTLs were identified for IL-1*β* and IL-1Ra. The total number cQTLs at different thresholds are listed in Supplementary table 6.

We next tested whether different immune cell counts in PBMC preparations influence the cQTLs. For this we made use of the FACS assessment in the 500 Functional Genomics study (500FG cohort), in which cell populations are examined in detail (see Methods) and measured *Candida*-induced IL-6 and TNF-α levels. First, we analysed the correlation structure between cell counts and cytokine measurements (Supplementary figure 6) and observed weak correlations (mean correlation coefficients across five cell types = 0.062). For example, *Candida*-induced IL-6 levels in PBMCs showed a weak correlation with monocyte counts, but not with other cell types. We tested the association of *Candida*-induced cQTLs with cytokine levels in 500FG cohort upon correcting for age, gender and cell counts. Among those three cQTLs tested, SNP rs11141235, associated with *Candida*-induced IL-6 levels, showed a clear replication of association (Supplementary table 4; $P = 0.017$) even after correction for

***Figure 2: Cytokine responses are organized around a physiological response towards specific pathogens.*** *(a) Unsupervised hierarchical clustering of cytokine responses performed using Spearman correlation as the measure of similarity. Red depicts a strong positive correlation whereas blue indicates a strong negative correlation. $T_H$ cluster, cytokines derived from T helper cells; Fungus cluster, Candida albicans induced cytokines. (b) Pair-wise correlation coefficients of production of monocyte-derived cytokines and T lymphocyte-derived cytokines. Cluster 1, monocyte-derived cytokines; Cluster 2, $T_H1$ and $T_H17$-derived cytokines. The data shown is from one independent experiment from 2009 cohort.*

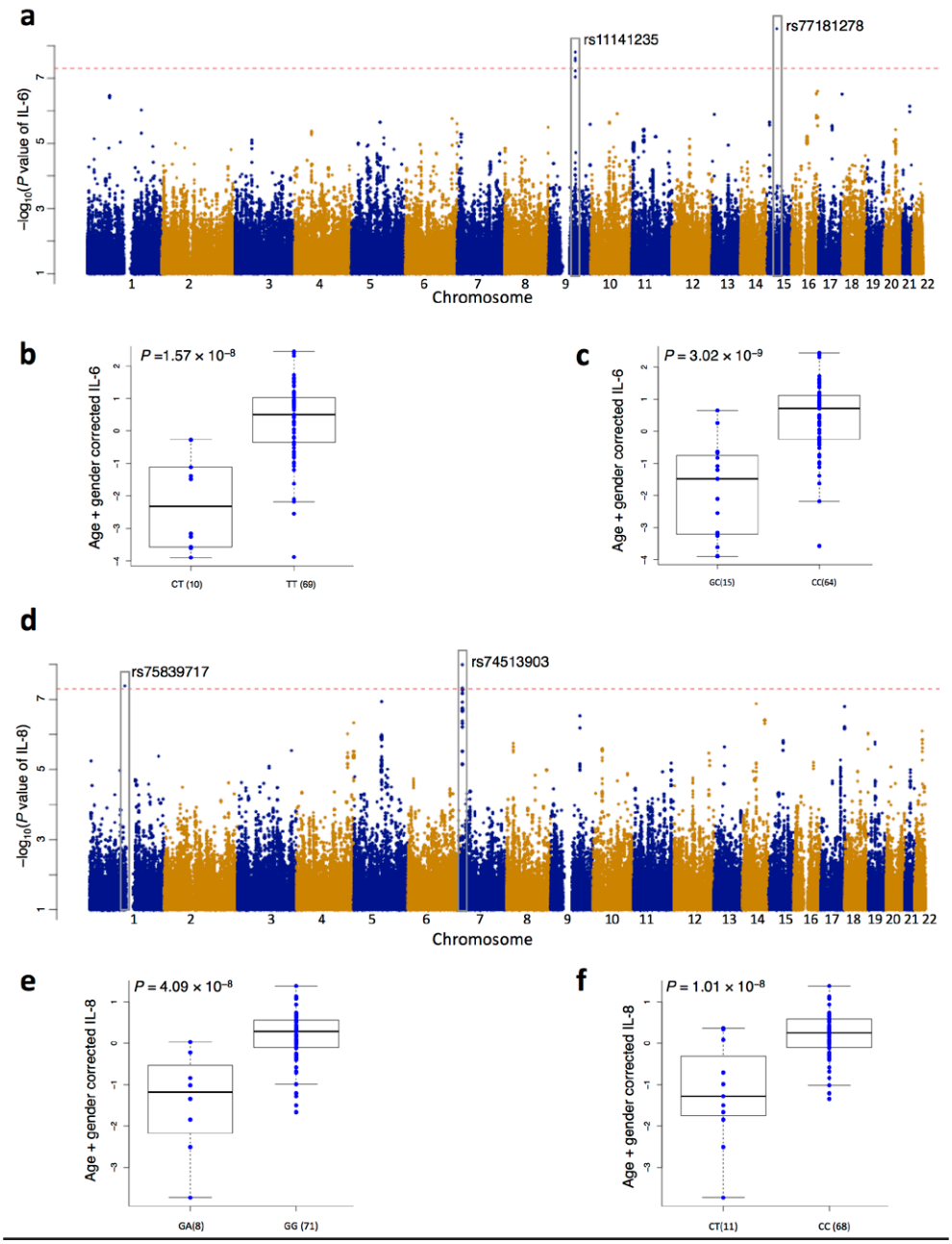



***Figure 3: Genome-wide cytokine QTL mapping identifies stimulation-induced cQTLs.*** *Manhattan plots showing the genome-wide QTL mapping results for (a) Candida albicans-induced IL-6 levels and (d) Mycobacterium tuberculosis-induced IL-8 levels. Horizontal dashed line corresponds to $P < 5 \times 10^{-8}$. Boxplots showing the association of genotypes at (b) chromosome 9 SNP rs11141235, (c) chromosome 15 SNP rs77181278 with Candida albicans induced IL-6 levels and (e) chromosome 1 SNP rs75839717, (f) chromosome 7 SNP …*

monocyte cell counts ($P$ = 0.030). In contrast, none of the 6 cQTLs were directly associated with cell counts (data not shown), suggesting the independent role of genetic variation on regulating cytokine production.

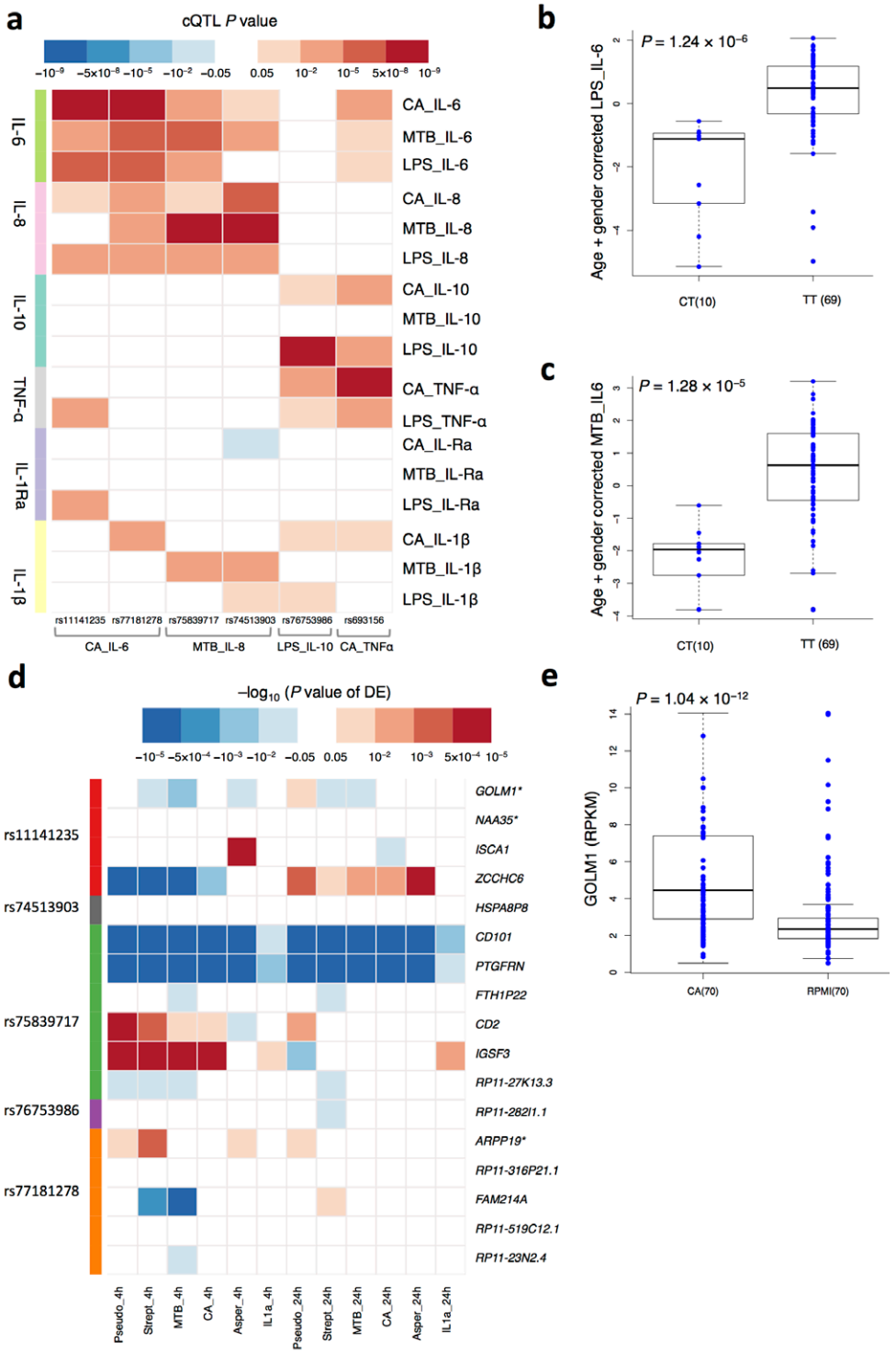
## Correlations between cytokine responses are partially genetically determined

The clustering of cytokine responses (Figure 2) showed correlations of cytokines induced by specific pathogens, as well as distinct clusters separating the monocyte-derived cytokine production from T cell-derived cytokines induced by various stimuli. To assess whether this observation may show a genetic component, we tested whether strong cQTLs of one pathogen-induced cytokine ($P < 1.0\times10^{-5}$) could also be associated with cytokine levels induced by other pathogens albeit with nominal significance ($P < 0.05$). QTLs of IL-10 were more likely to be pathogen-specific (Supplementary figure 7a), while QTLs of other cytokines (IL-6, IL-8 and TNF-α) were more likely to be shared genetic loci that respond to all pathogens (Supplementary figure 7b-c). We found similar results with different $P$ value thresholds (Supplementary figure 8). We also found that all six GWAS-significant cQTLs were associated with cytokines induced by other bacterial and fungal stimulations (Figure 4a). SNPs affecting fungus-induced IL-6 and IL-8 were also strongly associated with bacteria-induced IL-6 and IL-8 levels, but not with IL-10 and TNF-α (Figure 4a). This indicates that the SNPs associated with IL-6 and IL-8 levels are pathogen-independent, maybe because these SNPs are acting on genes/proteins that are downstream of pathogen recognition receptors and therefore are shared between pathogens. These results suggest that the correlation between monocyte-derived cytokines (Figure 2) may be partly genetically determined. The top association for *Candida*-induced IL-6 at the chr9q21 locus provides an illustrative example for a strong shared cQTL (Figure 4b-c), where the minor allele C at SNP rs11141235 was not only associated with lower *C. albicans*-induced IL-6 production (Figure 3b), but also with LPS-induced (Figure 4b) and MTB induced IL-6 production (Figure 4c). Importantly, the cQTLs identified in the 2013 cohort were all validated in the other cohorts, demonstrating the robustness of the associations identified (Supplementary figure 9).

## A cQTL gene *GOLM1* on chr9q21 modulates cytokine production

To identify the putative causal genes at six significant cQTLs, we tested the expression levels of all genes located within a 500kb *cis*-window of the 6 cQTLs in PBMCs stimulated with different microbial antigens (Figure 4d). Genes identified by this differential expression analysis were not cytokine genes, suggesting that the cQTLs identified are mainly *trans*-QTLs of regulatory genes modulating cytokine production.

*… rs74513903 with Mycobacterium tuberculosis induced IL-8 levels. The number of individuals per genotype is shown in parenthesis below each boxplot. The length of the box in the box-plot is interquartile range (=Q3-Q1). The whiskers indicate the range of one and a half times the length of the box from either end of the box. P values were from the linear regression analysis of cytokine on genotype data. The data shown is from one independent experiment from 2013 cohort.*


a
cQTL P value
−10−9 −5×10−8 −10−5 −10−2 −0.05 0.05 10−2 10−5 5×10−8 10−9
IL-6
IL-8
IL-10
TNF-α
IL-1Ra
IL-1β
CA_IL-6
MTB_IL-6
LPS_IL-6
CA_IL-8
MTB_IL-8
LPS_IL-8
CA_IL-10
MTB_IL-10
LPS_IL-10
CA_TNF-α
LPS_TNF-α
CA_IL-Ra
MTB_IL-Ra
LPS_IL-Ra
CA_IL-1β
MTB_IL-1β
LPS_IL-1β
rs11141235 rs77181278 rs75839717 rs74513903 rs76753986 rs693156
CA_IL-6 MTB_IL-8 LPS_IL-10 CA_TNFα
b
P = 1.24 × 10−6
Age + gender corrected LPS_IL-6
CT(10)
TT (69)
c
P = 1.28 × 10−5
Age + gender corrected MTB_IL6
CT(10)
TT (69)
d
−log10 (P value of DE)
−10−5 −5×10−4 −10−3 −10−2 −0.05 0.05 10−2 10−3 5×10−4 10−5
rs11141235
rs74513903
rs75839717
rs76753986
rs77181278
GOLM1*
NAA35*
ISCA1
ZCCHC6
HSPA8P8
CD101
PTGFRN
FTH1P22
CD2
IGSF3
RP11-27K13.3
RP11-282I1.1
ARPP19*
RP11-316P21.1
FAM214A
RP11-519C12.1
RP11-23N2.4
Pseudo_4h Strept_4h MTB_4h CA_4h Asper_4h IL1a_4h Pseudo_24h Strept_24h MTB_24h CA_24h Asper_24h IL1a_24h
e
P = 1.04 × 10−12
GOLM1 (RPKM)
CA(70)
RPMI(70)

The top associated cQTL rs11141235 on the chromosome 9q21 region was associated with *Candida*-induced IL-6 levels (Figure 3a). To identify the causal mechanism at this locus, we generated gene expression data by RNA sequencing in PBMCs from 70 individuals [18] with and without *Candida* stimulation. We reconfirmed the significant differential expression of *GOLM1* in response to *Candida* stimulation in this larger cohort (Figure 4e). Next, we mapped *Candida*-response eQTL at rs11141235 and at another SNP rs11141242 in the locus (D'=0.95), which is a more frequent polymorphism. The eQTL results indicated that rs11141242 was significantly ($P$ = 0.016) associated with the expression levels of *GOLM1* (Golgi membrane protein 1), where the minor allele was associated with lower levels of *GOLM1* (Figure 5a), while rs11141235 showed a similar trend (Figure 5b), suggesting the role of haplotypes in regulating *GOLM1* expression.

## *GOLM1* cQTL is associated with susceptibility to candidemia



*GOLM1* encodes a 73kDa Golgi protein and is upregulated in response to viral infection [19]. We assessed whether genetic variants in *GOLM1* locus could influence susceptibility to disseminated infection with *C. albicans* in a previously described cohort of 225 European patients with candidemia [20]. Since the genotype data at rs11141235 was not available from this cohort, we tested another variant, rs7036187, that is in linkage disequilibrium with rs11141235 (D' =1) in the *GOLM1* locus and found it to be associated with candidemia, where the risk allele A was more frequent in cases ($P$ = 0.016, Odds ratio= 2.36). To test whether the *GOLM1* cQTL affects candidemia through the IL-6 pathway we built a co-expression network around the *GOLM1* using gene expression data from PBMCs of 70 healthy volunteers either upon *Candida* stimulation (Figure 5c) or without stimulation (Supplementary figure 10). Pathway enrichment analysis on strongly co-expressed genes (r2 < 0.8) with *GOLM1* during *Candida* stimulation showed *GOLM1* co-expressed genes to be enriched for cytokine production pathways, which suggests that *GOLM1* is associated with cytokine signaling (Figure 5d). The enrichment of genes for cytokine signalling after stimulation could be also the consequence of the stimulation and not necessarily specific to GOLM1 co-expression. Therefore, we tested whether the extent of gene enrichment for IL-6 signaling

***Figure 4: Genome-wide significant cQTLs affect cytokine production induced by both bacterial and fungal stimulation.*** *(a) The P values of six significant cQTLs for other cytokine levels. The colour legend for the heat map indicates the range of P values from QTL mapping. P values were from the linear regression analysis of cytokine on genotype data. (b,c) Correlation of SNP rs11141235, genotype with IL-6 induced by LPS (b) and by Mycobacterium tuberculosis (c). The length of the box in the box-plot is interquartile range (=Q3-Q1). The whiskers indicate the range of one and a half times the length of the box from either end of the box. The number of individuals per genotype is shown in parenthesis below each boxplots. (d) P values for differential expression of genes (±250 kb around the SNP) selected from genome-wide significant cQTL loci upon different stimulations in human PBMCs (n=8). P values were from differential expression analysis using threshold of FDR=0.05 and fold change > 2. Genes were selected based on their physical positions which are within ±250 kb window around the SNP. PBMC stimulations were done for either 4 h or 24 h. The figure show results from both 4h and 24h. Red: up-regulation; Blue: down-regulation; *, genes with suggestive eQTLs in RNAseq data. e) GOLM1 expression levels upon C. albicans stimulation in PBMCs of 70 samples.*

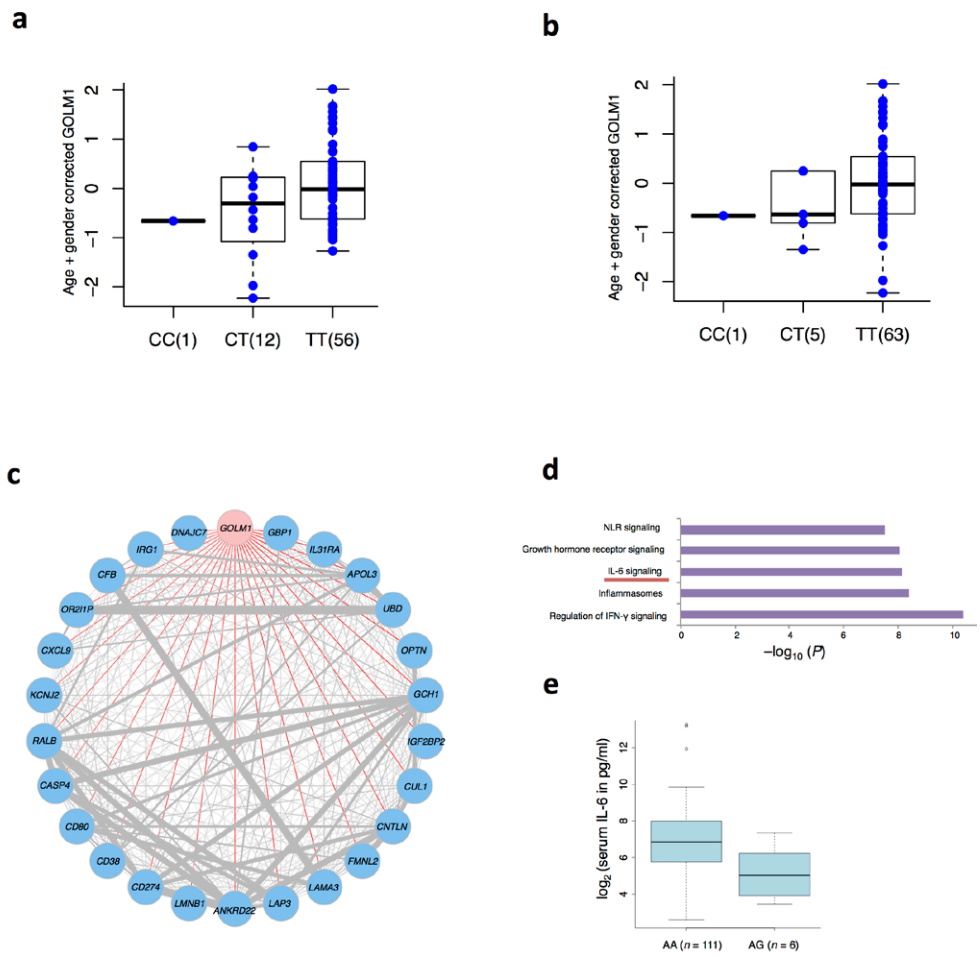



***Figure 5: GOLM1 is involved in IL-6 production.*** *Boxplots showing the correlation between gene expression levels of GOLM1 of 69 PBMC samples and SNPs (a) rs11141242 (P value = 0.017) and (b) rs11141235 upon C. albicans stimulation. c) Co-expression network for GOLM1 built using gene expression data using Spearman correlation from 70 PBMC samples stimulated with C. albicans. The red lines depict the correlation coefficient of more than 0.7 between other genes and GOLM1 in the network. (d) Pathway enrichment analysis on genes that are highly correlated with GOLM1 (Spearman correlation coefficient > 0.7) based on Reactome pathway database. (e) Correlation between secreted IL-6 levels and genotypes at rs7036187 of 117 candidemia patients (Student t test P = 0.015). The length of the box in the box-plot is interquartile range (=Q3-Q1). The whiskers indicate the range of one and a half times the length of the box from either end of the box. There are 111 patients with AA genotype and 6 patients with AG genotype.*

was specifically linked to *GOLM1* upon *Candida* stimulation (see Methods), when compared to randomly chosen differentially expressed genes in response to *Candida* stimulation. This analysis showed a significantly stronger enrichment of genes co-expressed with *GOLM1* for IL-6 signaling than randomly chosen genes (Supplementary figure 11a-b). In addition, in patients with candidemia, we also assessed the effects of the rs7036187 polymorphism, a SNP associated to susceptibility to disease, on serum IL-6 concentrations. This SNP was associated with circulating IL-6 concentration (Figure 5e), where AG genotypes were associated with lower levels of IL-6 ($P = 0.015$) suggesting that the polymorphisms in the *GOLM1* locus may influence *Candida*-induced cytokines and susceptibility to candidemia.

## Cytokine QTLs overlap with human disease associated SNPs

We tested whether SNPs previously associated with human diseases and particularly with infectious diseases are enriched with cQTLs. We extracted GWAS SNPs from the NHGRI GWAS catalog [21] and binned them into eight categories based on their association with different human phenotypes (See Methods). Next we identified all cQTLs that were associated with cytokine levels with a $P$ value < 0.05 (Supplementary table 6) and tested whether these cQTLs are linked to GWAS SNPs or their proxies. Sixty-one percent of infectious-disease-associated SNPs were also cQTLs, and 43% of immune-mediated-disease associated SNPs were also cQTLs (Figure 5a). We used height-associated SNPs as background SNPs (or null set of SNPs) to test whether cQTLs are more often associated with a particular human disease. 38.5% of height associated SNPs were also cQTLs. We observed a significant enrichment ($P < 9.99 \times 10^{-8}$) of cQTLs among infectious disease-associated SNPs. A proportion of heart disease-associated SNPs were also cQTLs, suggesting a role for cytokine pathways in the pathogenesis of cardiovascular diseases (Figure 6a). We found similar results when we selected cQTLs with a different $P$ value ($P < 0.01$, Supplementary table 6) threshold (Supplementary figure 12) to test for their enrichment among GWAS SNPs. However, the sensitivity of the enrichment results dropped when we used more stringent $P$ values to call putative cQTLs since the number of cQTLs available to perform enrichment analysis was reduced.

Furthermore, we tested whether infectious and autoimmune disease-associated SNPs are predominantly associated with increased or decreased cytokine production. We observed no significant difference between the numbers of autoimmune disease risk alleles associated with increased or decreased cytokine production (Supplementary figure 8). In contrast, risk alleles of infectious disease SNPs, with the exception of malaria- associated SNPs, are mostly associated with lower cytokine production capacity (Figure 6b). These patterns suggest that the genetic alterations associated with autoimmune diseases are correlated with both increased and decreased cytokine production capacity, whereas susceptibility to infections is associated with a lower capacity for cytokine production from monocytes or lymphocytes, depending on the type of infection. Inflammatory bowel disease (IBD) is a chronic immune-mediated disease of the human gastrointestinal tract. We observed a trend where high proportion of risk alleles of IBD-associated SNPs were associated with lower cytokine production capacity (Figure 6c), suggesting the role of infectious agents in IBD. These results again highlight the importance of response QTLs to understanding complex human diseases.

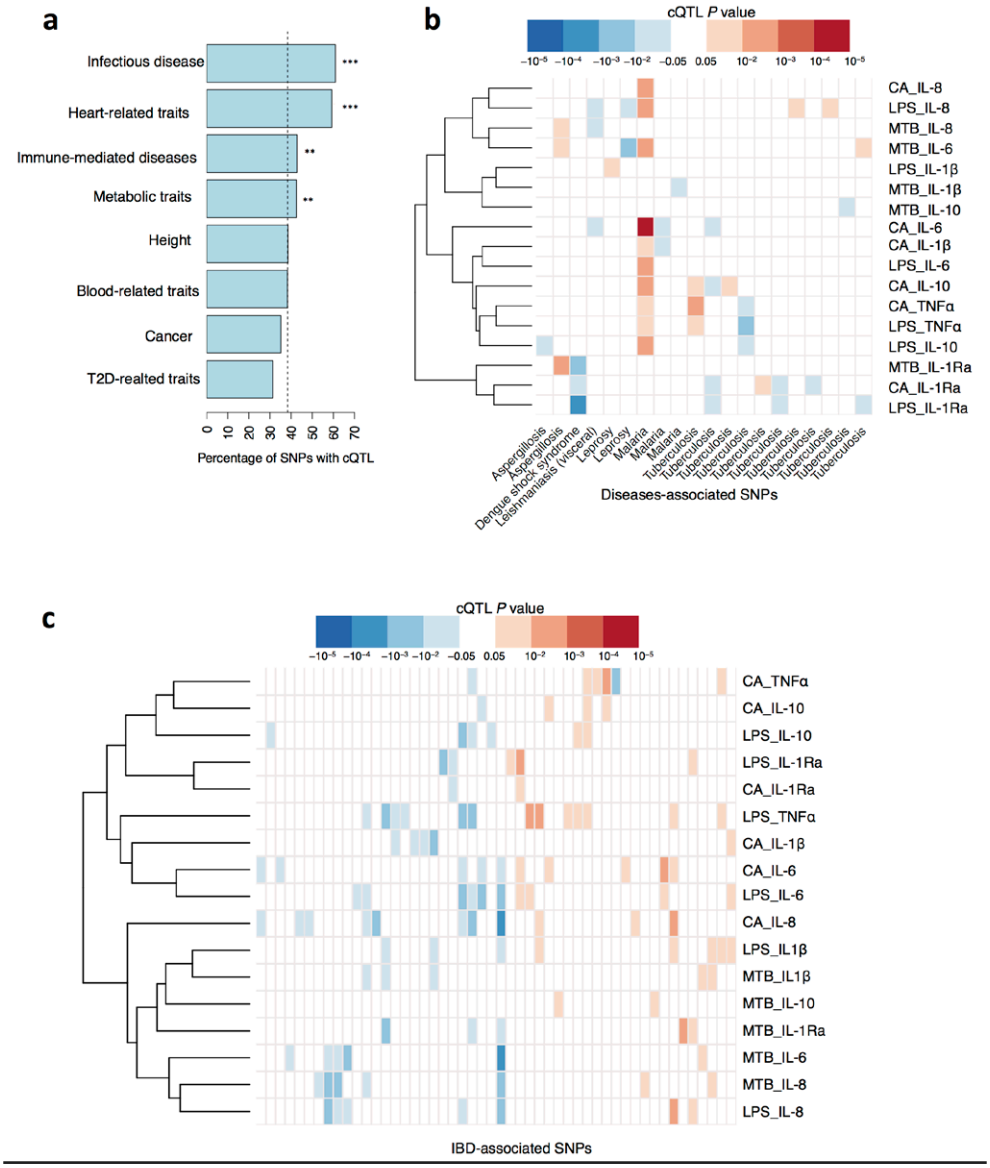



***Figure 6 Association of cQTLs with infectious diseases.*** *(a) The percentage of SNPs associated with each category of disease that also qualify as suggestive cytokine QTLs (P value <0.05). Dotted line indicates the percentage of cQTLs that overlapped with height-associated SNPs, which served as reference set (null set). Enrichment analysis from Fisher exact test are indicated by red "stars" (***, $P < 10^{-8}$ ; **, $P < 10^{-4}$ ; *, $P < 0.05$). (b-c) QTLs associated with indicated stimulus-cytokine pairs (rows) compared with SNPs associated with susceptibility to the indicated pathogens (columns) (b) and with IBD (c). The colours represent the –log10P values of cytokine QTLs. P values were obtained from linear regression model of cytokine levels on genotype data. Red and blue indicate association with upregulated or downregulated cytokine levels, respectively.*

## Discussion

While several recent studies investigated immune traits during steady-state conditions [22–24] or serologic responses to past infections [25], an important question remains regarding the architecture of the immune response and its relation to genetic variation. While several eQTL studies addressed this important question using different immune cell populations either without or upon stimulation [8,9,11,12], they have two intrinsic limitations: they only used stimulation of purified cell populations with purified ligands (with the exception of influenza virus stimulation in one of the studies), and eQTLs only interrogate transcript levels which are known to correlate only partially with protein abundance [13,26–29]. The present study therefore addresses the importance of understanding human immune responses to pathogens by assessing the architecture of one of the most important steps in the activation of the immune responses: cytokine production.



First, our study not only confirms the non-parametric nature of cytokine production distribution of monocyte-derived pro-inflammatory cytokines [30,31], but also extend this to T-lymphocyte derived cytokines. The non-Gaussian or bi-modal distribution of cytokine production identifies low- and high-producers, giving scope to the hypothesis that the cytokine synthesis phenotype may have a strong effect on susceptibility to immune-mediated diseases.

Second, the production capacity of various cytokines strongly correlates when cells were stimulated with a specific pathogen, while the correlation was poor when comparing bacterial versus fungal stimulation. This observation makes sense from both an evolutionary point of view, as immune responses mainly need to have plasticity to respond to specific infectious pressures in a certain geographic area [2], and from a biological point of view in which Toll-like receptors are the main receptor pathway recognizing bacteria, while C-type lectin receptors mainly recognize fungi. Importantly, regulation of the pathogen-specific cytokine responses is most likely only partially genetic, as some genetic polymorphisms regulate multiple cytokine responses to a certain pathogen (especially for monocyte-derived production), while others regulate the monocyte-derived production of cytokine responses due to multiple pathogens (see below). It is therefore likely that non-genetic external factors encountered during one's lifetime also play an important role in long-term modulation of cytokine responses, and epigenetic regulation may represent one of the molecular substrates for this process [32,33].

One remarkable exception to the rule of pathogen-centric responses is represented by specific lymphocyte responses such as IL-17 production, which represents a separate strongly correlated cluster independent of the type of pathogen and may be an important aspect of IL-17 biology. Th17 responses are crucial for mucosal host defence [34], and defects in this pathway lead to high susceptibility to both fungal and bacterial pathogens [35–37]. These data argue that Th17 responses are a crucial component of host defence against both bacteria and fungi.

Third, we have identified six novel genome-wide significant cQTLs that influence cytokine responses: IL-6 and IL-8 identify the highest number of cQTLs, while IL-1β and IL-1Ra show no cQTLs, suggesting that the immune response can buffer easier variation in IL-6 and IL-8, while the IL-1 pathway is highly conserved. Moreover, cQTLs are *trans*-QTLs that influence

cytokine production indirectly through regulatory loops, in line with the observation that cytokine responses are organized around regulation of pathogen-specific host responses, rather than towards regulation of specific cytokines. This is in line with a recent study showing that 90% of causal SNPs tend to occur near binding sites for master regulators of stimulus-dependent gene expression and map to enhancers which gain histone acetylation [38]. One of the strongest cQTLs that we identified influences expression of the *GOLM1* upon *Candida* stimulation, which in turn influences susceptibility to candidemia. GOLM1 encodes a Golgi phosphoprotein, also referred as GP73, which is known to respond to viral infections [19]. This molecule has also been tested as a useful circulating biomarker for several viral and non-viral induced liver diseases [39,40] and the serum GOLM1 levels was shown to correlate with serum IL-6 levels in hepatocellular carcinoma [41]. The *trans*-regulatory network of GOLM1 that we describe here provides further insights into the understanding of the GOLM1-mediated cytokine regulation, not only in cancer but also for infectious diseases.

There are also some limitations to the study. Firstly, we cannot exclude that cQTLs were missed and/or some of the cQTLs with rare allele frequencies are false positives in the analyses due to the relatively low number of volunteers. The same reason prevented us from performing a robust assessment of cQTLs of lymphocyte-derived cytokines. Secondly, the present study investigated cytokine levels induced by bacterial and fungal pathogens, but not viral stimuli. Thirdly, variation in the immune system can be driven by both heritable and non-heritable influences. Finally, the experimental set up of ex-vivo PBMCs stimulated for 24 hours provides the opportunity to study the interactions between immune cells such as monocytes, T cells and B cells in response to pathogens. However the time-dependent dynamic interactions at tissue level are only partially captured. Therefore, PBMCs alone may not fully provide the in-vivo picture of immune response since cell-cell interactions also occur at specific tissue locations. Some of these remaining questions will be addressed by the currently ongoing analyses of a larger cohort of 500 healthy volunteers within the Human Functional Genomics Study, in which more volunteers, a larger panel of stimuli, and external factors (e.g. diet, microbiome) will be included.

## Methods

### Ethics statement

Samples of venous blood were drawn after informed consent was obtained, and the study was approved by the Ethical Committee of Radboud University Nijmegen (nr. 42561.091.12). Experiments were conducted according to the principles expressed in the Declaration of Helsinki.

### 200FG cohort

Individuals in this study were foresters from the 'Geldersch Landschap', 'Hoge Veluwe', 'Twickel', and 'Kroondomein het Loo' in the Netherlands. Foresters were asked to donate blood in order to determine the serology against *Borrelia* bacteria, since Lyme disease occurs as an occupational disease. The cohort of individuals was chosen because of the good health reported by this general population. None of the volunteers included in the study had *Borrelia* infection. In this cohort, all individuals gave written informed consent to donate extra blood to use for research. Blood was drawn in 2009, 2011 and 2013. The foresters

were between 23-73 years old, and consisted of 77% males and 23% females. The cQTLs identified were additionally validated in a cohort of 500 healthy individuals of Dutch European ancestry from the Human Functional genomics Project (500FG cohort, www.humanfunctionalgenomics.org).

## PBMC collection and stimulation experiments

After obtaining informed consent, venous blood was drawn from the cubital vein of volunteers into 10 mL EDTA tubes (Monoject). Isolation of PBMCs was performed according standard protocols, with minor modifications. The PBMC fraction was obtained by density centrifugation of blood diluted 1:1 in pyrogen-free saline over Ficoll-Paque (Pharmacia Biotech). Cells were washed twice in saline and suspended in medium (RPMI 1640) supplemented with gentamicin 10 mg/mL, L-glutamine 10 mM and pyruvate 10 mM. Addition of antibiotics such as gentamycin is a standard methodology used in order to avoid contamination of cultures, and it does not influence the ability to induce cytokine production by PBMCs or macrophages (data not shown). The cells were counted in a Coulter counter (Coulter Electronics) and the number was adjusted to $5\times10^6$ cells/mL. Then $5\times10^5$ PBMCs in a 100 µL volume were added to round-bottom 96-wells plates (Greiner) and incubated with 100 µL of stimulus. After 24 h the supernatants were collected and stored at -20°C until assayed. The stimulation time periods were chosen based on extensive previous studies that showed that 24 h stimulation was best suited to assess monocyte-derived cytokines [42,43].



## Stimulation of PBMCs

### *Bacteria*

*Bacteroides fragilis* (NCTC 10584) grown anaerobically overnight at 37°C on blood agar plates (BD Biosciences, Franklin Lakes, NJ, USA) was inoculated in 20 mL pre-warmed and pre-reduced Brain Heart Infusion broth (BD Diagnostics, Basel, Switzerland) and again grown anaerobically overnight at 37 °C until reaching stationary growth phase mimicking growth conditions in abscesses. Bacterial suspensions were washed three times in phosphate-buffered saline (PBS; B. Braun Medical B.V., Melsungen, Germany) and heat-killed at 95°C for 30 min. Before heat-killing, aliquots of bacterial suspensions were taken to determine colony-forming unit (cfu) counts. Heat-killed bacteria were washed again and after adjusting the concentration in PBS to $1 \times 10^8$ cfu/mL, stored at −80 °C. *B. fragilis* was used in the stimulation experiments as $1\times10^6$/mL. *E. coli* ATCC 25922 was grown overnight in culture medium, washed three times with PBS, and heat-killed for 60 min at 80°C; *Staphylococcus aureus* strain ATCC 29213 was grown overnight in culture medium, washed twice with cold PBS, and heat-killed for 30 min at 100°C; both *E. coli* and *S. aureus* were used in a final concentration of $1\times10^6$/mL. Success of heat-inactivation was confirmed by cultures.

Cultures of H37Rv *Mycobacterium tuberculosis* (MTB) were grown to mid-log phase in Middlebrook 7H9 liquid medium (Difco, Becton-Dickinson) supplemented with oleic acid/albumin/dextrose/catalase (OADC) (BBL, Becton-Dickinson), washed three times in sterile saline, heat killed and then disrupted using a bead beater, after which the concentration was measured using a bicinchoninic acid (BCA) assay (Pierce, Thermo Scientific).

### *Fungi*

Heat-killed *C. albicans* blastoconidia (strain ATCC MYA-3573, UC 820) at a concentration of 106 CFU/mL were used throughout this study. To generate hyphae, live yeast forms of Candida were grown for 24 h at 37°C in RPMI 1640 (Gibco-BRL, Grand Island, NY), adjusted to pH 6.4 by using hydrochloric acid. After 24 h, more than 95% of blastoconidia were grown to hyphae, which were checked by microscope. Hyphae were heat killed for 45 min at 98°C and resuspended in RPMI 1640 to a hyphal inoculum size that originated from $10^6$/mL blastoconidia (referred to as $10^6$/mL hyphae).

Ligands – FSL-1 and Pam3Cys were purchased at EMC microcollections (L-7000, L-2000, respectively) and used in a final concentration of 1 µg/mL and 10 µg/mL.

### *Microbial ligands*

MDP (muramyl dipeptide) was purchased at Sigma (A-9519) and used at a final concentration of 10 µg/mL. LPS (*E. coli* serotype 055:B5) was purchased from Sigma and an extra purification step was performed as described previously [44]. Purified LPS was tested in TLR4−/− mice for the presence of contaminants and did not have any TLR4-independent activity [45].

A total of $5\times10^5$ PBMCs in a total volume of 200 µL per well were incubated at 37°C in round-bottom 96-well plates (Greiner) with the different stimuli, as indicated above. After 24 h (early cytokines IL-1β, TNF-α, IL-6, IL-8, and IL-10), or 7 days of incubation (IFN-γ and IL-17), supernatants were collected and stored at −20°C until assayed. When cells were cultured for 7 days, this was done in the presence of 10% human pooled serum.

## Cytokine measurements

Concentrations of human cytokines determined using specific commercial ELISA kits from R&D Systems: IL-1β (catalog number DLB50), IL-6 (D6050), IL-10 (D1000B), TNF-α (DTA00C), IL-17 (D1700), or IFN-γ (DIF50) in accordance with the manufacturers' instructions. Detection limits were 20 pg/mL, except for IFN-γ ELISA (12 pg/mL).

## IL-6 measurements in Candidemia cohort

Concentrations of human IL-6 in the serum of candidemia patients were determined using specific commercial ELISA kits (PeliKine Compact or R&D Systems), in accordance with the manufacturers' instructions. The data were available for 117 Caucasian candidemia patients. Candidemia patients were stratified on rs7036187 SNP genotype to obtain 111 AA and 6 AG patients. For each individual the median values of IL-6 levels measured across 15 days were used. The statistical difference was tested using a student *t* test (one-sided) by comparing the log2 transformed IL-6 values. P value less than 0.05 was considered significant.

## Cytokine clustering and variance analysis

Raw cytokine levels were first log-transformed, then cytokine measurements showing little/no variation across individuals were filtered out for the follow-up analysis. We excluded 9 cytokine measurements in 2009. The normality test was performed on both raw (Figure 1a) and log-transformed data (Supplementary figure 13) using Shapiro-Wilk normality test, respectively. P value > 0.05 was used as threshold for normal distribution. Unsupervised

hierarchical clustering was performed using Spearman correlation as the measure of similarity. In order to test the equality of variance of cytokine levels before and after stimulation, Levene's test was used.

## Genotyping, quality control and imputation

DNA samples of 112 individuals were genotyped using the commercially available SNP chip, Illumina HumanOmniExpressExome-8 v1.0. The genotype calling was performed using Opticall 0.7.0 [46] using the default settings. Four samples with a call rate ≤ 0.99 were excluded from the dataset as were variants with a HWE ≤ 0.0001, call rate ≤ 0.99 and MAF ≤ 0.001. Two samples were excluded as potential ethnic outliers identified by multi-dimensional scaling plots of samples merged with 1000 Genome data (Supplementary figure 14). This resulted in a dataset of 106 samples containing genotype information of 282,382 variants for further imputation. The strands and variant-identifiers were aligned to the reference Genome of The Netherlands (GoNL) [18] dataset using Genotype Harmonizer [47]. The data was phased using SHAPEIT2 v2.r644 [48] using the GoNL as a reference panel. Finally this data was imputed using IMPUTE2 [49] with the GoNL as the reference panel [50]. Post imputation provided 7512899 variants. We selected 3959389 SNPs that showed MAF ≥ 5%, INFO score ≥ 0.8 and 3 samples per genotype for downstream cytokine QTL mapping.

## Cytokine QTL mapping

Lack of either DNA or cytokine measurements for 90 individuals sampled in three different years restricted us to obtain both genotype and cytokine data for 107 individuals out of 197 individuals (Supplementary table 1). We used the 2013 dataset as a discovery cohort to identify genome-wide significant cQTLs since this cohort had the largest numbers of individuals (n=79). The 2009 (n=30) and 2011 datasets (n=78) were used as validation cohorts. We coded gender information either 0 for females or 1 for males. The actual age and coded gender information were included as co-variables in the linear model to correct the cytokine distributions for QTL mapping. We focused only on infectious stimulations such as LPS, Candida and MTB to map cQTLs, which provided 18 measurements (3 stimulations× 6 cytokines). Since it was difficult define a numerical cut-off to filter out the cytokine measurements that were not informative, we visually inspected all the cytokine distribution plots to check if the cytokine measurements provide clear variation across individuals (Supplementary figure 15). By manually checking the log-transformed cytokine distributions, we excluded one measurement (MTB-induced TNF-α) from further QTL mapping as this cytokine measurement showed very little variation (low production capacity in the majority of individuals), and was thus not informative (Supplementary figure 2). Raw cytokine levels were first log-transformed then mapped to genotype data using a linear regression model with age and gender as covariates. A $P$ value < $5\times10^{-8}$ was considered to be the threshold for significant cytokine QTLs. In order to check whether our QTL mapping indicated any significant inflation of test statistics due to population structure, we calculated genetic inflation factor lambda (observed vs. expected $P$ values) for all cytokine measurements. We found that the lambda values were around 1 (0.99 to 1.04) indicating there is no or very little population stratification (Supplementary figure 16). Since the different amounts of cytokine production can also be driven by different immune cell counts in PBMC preparations, we tested whether cQTLs can influence the cytokine levels independent of different cell counts. For this we focused on *Candida*-induced cQTLs and made use of the FACS assessment in the

500 Functional Genomics study (500FG cohort), in which cell populations are examined in detail. We obtained cell count data measured by FACS for total lymphocytes, T cells, B cells, monocytes and NK cells from 487 individuals from the 500FG cohort. We measured IL-6 and TNF-α levels in response to *Candida* stimulation to check if the *Candida*-induced cQTLs can be replicated after cell counts correction.

## RNA sequencing and expression analysis

Candidate genes from significant cytokine QTL loci were further tested if they responded to any of the pathogens using RNA seq data from PBMCs of eight individuals, which were stimulated by *Pseudomonas aeruginosa*, *Streptococcus pneumoniae*, *Mycobacterium tuberculosis*, *Candida albicans*, *Aspergillus fumigatus*, and IL-1alpha. The PBMCs from 70 individuals of the GONL cohort [18] were stimulated with or without *Candida albicans* as previously described [51]. Sequencing reads were mapped to the human genome using STAR (version 2.3.0) [52]. The aligner was provided with a file containing junctions from Ensembl GRCh37.71. Htseq-count of the Python package HTSeq (version 0.5.4p3) was used (The HTSeq package, http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html) to quantify the read counts per gene based on annotation version GRCh37.71, using the default union-counting mode. Differentially expressed genes were identified by statistics analysis using DESeq2 package from bioconductor [53]. The statistically significant threshold (FDR $P \leq 0.05$ and Fold Change $\geq 2$) was applied. Gender and age were included as known covariates in a linear model for assessing genotype effect. All eQTL mapping was done using Matrix-eQTL [54]. To build co-expression networks, we extracted top 25 co-expressed genes for top 20 differentially expressed genes in Candida-stimulation experiment and performed the pathway enrichment analysis on these gene-sets. Then, we compared whether these genes are also enriched for IL-6 signaling similar to *GOLM1* co-expressed genes (Supplementary figure 11a). In addition, we extracted the co-expressed genes with *GOLM1* and ranked them according to their co-expression correlation values with *GOLM1*. We performed pathway enrichment analysis on top 50 (r2 with *GOLM1* > 0.7), middle 50 (r2 with *GOLM1* between 0.7-0.4), bottom 50 (r2 with *GOLM1* < 0.2) and randomly chosen 50 genes. Then, we tested if the enrichment for IL-6 signaling is dependent on the strength of the co-expression with *GOLM1* (Supplementary figure 11b).

## Extraction of infectious disease associated SNPs

SNPs associated with a number of infectious diseases that showed P value < $9.99\times10^{-6}$ were extracted using the GWAS catalogue (http:/www.genome.gov/gwasstudies). As of December 2014, there were two studies on leprosy, two studies on malaria, four studies on tuberculosis, four studies on chronic hepatitis C infection, one study on HPV seropositivity, one study on Dengue shock syndrome and one study on meningococcal susceptibility. By systematic search in the literature, SNPs associated with susceptibility to additional infectious diseases not reported in the GWAS catalogue were also extracted. There were three studies on invasive aspergillosis and two studies on pneumococcal disease (Supplementary table 7). The infectious disease associated SNPs shown in Figure 6 are rs1519551, rs4833095, rs3132468, rs9271858, rs16948876, rs3764147, rs11036238, rs9940464, rs6755404, rs958617, rs6545883, rs1900442, rs8005962, rs1925714, rs4331426, rs2505675, rs160441, rs1948632, rs6538140, and rs9373523.

### GWAS SNP extraction and enrichment analysis

GWAS SNPs from the GWAS catalog [21] and their proxies ($r^2 \geq 0.8$ from 500kb window) were first extracted, which provided a list of SNPs associated to 122 different human traits and diseases. We selected diseases/traits for which at least 10 independent SNPs were reported to be associated. We then binned these GWAS SNPs into eight categories based on their association to closely related human phenotypes (cancer, immune-mediated diseases, infectious disease, heart-related traits, blood-related traits, metabolic traits, height and Type 2 diabetes related traits). Duplicated SNPs are removed from further analysis. We then intersected the SNPs of each category with cQTLs that showed $P < 0.05$ in our study. The Fisher exact test was applied to test the over-representation cQTL SNPs in infectious disease SNPs using the height associated SNPs as reference.

### Online database

All data used in this project have been meticulously catalogued and archived at BBMRI-NL data infrastructure (http://hfgp.bbmri.nl) using the MOLGENIS open source platform for scientific data. This allows flexible data querying and download, including sufficiently rich metadata and interfaces for machine processing (R statistics, REST API) and using FAIR principles to optimize Findability, Accessibility, Interoperability and Reusability.



## Acknowledgements

The authors thank all volunteers from the 200 Functional Genomics cohort of the Human Functional Genomics Study for participation in the study. The authors would like to thank Kate McIntyre for editing the final text. This study was partially supported by a ERC Consolidator Grant (3310372) to MGN and by the ERC Advanced Grant (FP/2007-2013/ERC grant 2012-322698 to CW), the Dutch Digestive Diseases Foundation (MLDS WO11-30 to CW and VK), the European Union's Seventh Framework Programme (EU FP7) TANDEM project (HEALTH-F3-2012-305279 to CW and VK), and Netherlands Organization for Scientific Research (NWO) VENI grant (863.13.011 to YL). This study made use of data generated by the 'Genome of the Netherlands' project, which is funded by the Netherlands Organization for Scientific Research (grant no. 184021007). The data were made available as a Rainbow Project of BBMRI-NL.

## Author contributions

M.G.N. and C.W. coordinated the recruitment of cohorts and data generation. M.G.N., V.K., L.A.B.J. and C.W. conceived and directed the study with input from all authors. Y.L. analysed and interpreted the data. P.D., I.R.-P., V.M. and V.K. performed genotyping and imputation. M.O., S.S. and M.J. conducted the stimulation experiments and cytokine quantification. M.A.S., R.J.X. and L.F. provided the computational framework for the study and critical inputs to the study design. M.G.N., V.K., C.W., Y.L. and M.O. wrote the manuscript with input from all authors.

## Additional material

The following supplements are available with the on-line version of this paper.

Figure S1: The increased variation in cytokine levels upon stimulation (2009 cohort).
Figure S2: The increased variation in cytokine levels upon stimulation (2013 cohort).
Figure S3: The cytokine responses are organized around a physiological response towards specific pathogens (2011 cohort).
Figure S4: The cytokine responses are organized around a physiological response towards specific pathogens (2013 cohort).
Figure S5: The cytokine responses are organized around a physiological response towards specific pathogens.
Figure S6: The correlation between cytokine responses and cell counts after age and gender correction.
Figure S7: Cytokine QTLs of one pathogen induced IL10 are more likely to be pathogen specific.
Figure S8: Cytokine QTLs of one pathogen induced IL10 are more likely to be pathogen specific at a different threshold.
Figure S9: Identification of a cQTL on chromosome 9 with very strong effects on IL-6 production capacity.
Figure S10: Autoimmune disease associated SNPs mostly up-regulate the cytokines.
Figure S11: Enrichment of IL6 signaling pathways.
Figure S12: Significant enrichment of cQTLs to be associated with infectious diseases
Figure S13: More cytokines follow normal distribution after transformation.
Figure S14: Multidimensional scaling plot for 200FG samples.
Figure S15: Non-informative cytokine distributions that were excluded from cytokine QTL mapping.
Figure S16: Genetic inflation factor lambda (observed vs. expected P values) for all cytokine measurements.
Table S1: Overview of samples with cytokine measurements and genotype information.
Table S2: Overview of cytokine measurements from stimulation experiments.
Table S3: Cytokine measurements of 2009 cohort included for clustering analysis.
Table S4: A list of genome-wide significant cytokine QTLs.
Table S5: Cytokine QTL p values of all measured phenotypes at genome-wide significant cQTLs SNPs from 2013 cohort.
Table S6: The number of cQTLs at different thresholds.
Table S7: Cytokine QTL at infectious disease associated SNPs from 2013 cohort.

## References

1. Fumagalli, M. *et al.* Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. *PLoS Genet.* **7,** e1002355 (2011).
2. Netea, M. G. *et al.* Genetic variation in Toll-like receptors and disease susceptibility. *Nat. Immunol.* **13,** 535–542 (2012).
3. Karlsson, E. K. *et al.* Natural selection and infectious disease in human populations. *Nat. Rev. Genet.* **15,** 379–393 (2014).
4. Hill, A. V. S. Evolution, revolution and heresy in the genetics of infectious disease susceptibility. *Philos. Trans. R. Soc. B Biol. Sci.* **367,** 840–849 (2012).

5. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–124 (2012).
6. Kumar, V. *et al.* Genetics of immune-mediated disorders: from genome-wide association to molecular mechanism. *Curr. Opin. Immunol.* **31,** 51–57 (2014).
7. Zhernakova, A. *et al.* Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. *Am. J. Hum. Genet.* **86,** 970–7 (2010).
8. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343,** 1246949 (2014).
9. Lee, M. N. *et al.* Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells. *Science (80-. ).* **343,** 1246980–1246980 (2014).
10. Berry, M. P. R. *et al.* An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* **466,** 973–977 (2010).
11. Ye, C. J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science (80-. ).* **345,** 1254665–1254665 (2014).
12. Raj, T. *et al.* Polarization of the Effects of Autoimmune and Neurodegenerative Risk Alleles in Leukocytes. *Science (80-. ).* **344,** 519–523 (2014).
13. Vogel, C. *et al.* Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13,** 227–232 (2012).
14. Naitza, S. *et al.* A Genome-Wide Association Scan on the Levels of Markers of Inflammation in Sardinians Reveals Associations That Underpin Its Complex Regulation. *PLoS Genet.* **8,** e1002480 (2012).
15. Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47,** 115–125 (2015).
16. Zambrano-Zaragoza, J. F. *et al.* Th17 Cells in Autoimmune and Infectious Diseases. *Int. J. Inflam.* **2014,** 1–12 (2014).
17. Mills, K. H. G. *et al.* The role of inflammasome-derived IL-1 in driving IL-17 responses. *J. Leukoc. Biol.* **93,** 489–497 (2013).
18. The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46,** 818–825 (2014).
19. Kladney, R. D. *et al.* GP73, a novel Golgi-localized protein upregulated by viral infection. *Gene* **249,** 53–65 (2000).
20. Kumar, V. *et al.* Immunochip SNP array identifies novel genetic variants conferring susceptibility to candidaemia. *Nat. Commun.* **5,** 4675 (2014).
21. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001–D1006 (2014).
22. Orrù, V. *et al.* Genetic Variants Regulating Immune Cell Levels in Health and Disease. *Cell* **155,** 242–256 (2013).
23. Roederer, M. *et al.* The Genetic Architecture of the Human Immune System: A Bioresource for Autoimmunity and Disease Pathogenesis. *Cell* **161,** 387–403 (2015).
24. Brodin, P. *et al.* Variation in the Human Immune System Is Largely Driven by Non-Heritable Influences. *Cell* **160,** 37–47 (2015).
25. Rubicz, R. *et al.* Genome-wide genetic investigation of serological measures of common infections. *Eur. J. Hum. Genet.* **23,** 1544–1548 (2015).

26. Ghazalpour, A. *et al.* Comparative Analysis of Proteome and Transcriptome Variation in Mouse. *PLoS Genet.* **7,** e1001393 (2011).
27. Jüschke, C. *et al.* Transcriptome and proteome quantification of a tumor model provides novel insights into post-transcriptional gene regulation. *Genome Biol.* **14,** r133 (2013).
28. Battle, A. *et al.* Impact of regulatory variation from RNA to protein. *Science (80-. ).* **347,** 664–667 (2015).
29. Wu, L. *et al.* Variation and genetic control of protein abundance in humans. *Nature* **499,** 79–82 (2013).
30. Endres, S. *et al.* Measurement of immunoreactive interleukin-1β from human mononuclear cells: Optimization of recovery, intrasubject consistency, and comparison with interleukin-1α and tumor necrosis factor. *Clin. Immunol. Immunopathol.* **49,** 424–438 (1988).
31. Endres, S. *et al.* In vitro production of IL 1β, IL 1α, TNF and IL 2 in healthy subjects: distribution, effect of cyclooxygenase inhibition and evidence of independent gene regulation. *Eur. J. Immunol.* **19,** 2327–2333 (1989).
32. Saeed, S. *et al.* Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science (80-. ).* **345,** 1251086–1251086 (2014).
33. Cheng, S.-C. *et al.* mTOR- and HIF-1 -mediated aerobic glycolysis as metabolic basis for trained immunity. *Science (80-. ).* **345,** 1250684–1250684 (2014).
34. Kayama, H. *et al.* Regulation of Intestinal Homeostasis by Innate Immune Cells. *Immune Netw.* **13,** 227 (2013).
35. van de Veerdonk, F. L. *et al. STAT1* Mutations in Autosomal Dominant Chronic Mucocutaneous Candidiasis. *N. Engl. J. Med.* **365,** 54–61 (2011).
36. Liu, L. *et al.* Gain-of-function human *STAT1* mutations impair IL-17 immunity and underlie chronic mucocutaneous candidiasis. *J. Exp. Med.* **208,** 1635–1648 (2011).
37. Milner, J. D. *et al.* Impaired TH17 cell differentiation in subjects with autosomal dominant hyper-IgE syndrome. *Nature* **452,** 773–776 (2008).
38. Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2014).
39. Kladney, R. D. *et al.* Expression of GP73, a resident Golgi membrane protein, in viral and nonviral liver disease. *Hepatology* **35,** 1431–1440 (2002).
40. Liu, X. *et al.* Golgi protein 73(GP73), a useful serum marker in liver diseases. *Clin. Chem. Lab. Med.* **49,** 1311–6 (2011).
41. Liang, H. *et al.* Interleukin-6 and oncostatin M are elevated in liver disease in conjunction with candidate hepatocellular carcinoma biomarker GP73. *Cancer Biomarkers* **11,** 161–171 (2012).
42. Netea, M. G. *et al.* A semi-quantitative reverse transcriptase polymerase chain reaction method for measurement of mrna for TNF-α and IL-1β in whole blood cultures: Its application in typhoid fever and exentric exercise. *Cytokine* **8,** 739–744 (1996).
43. van Crevel, R. *et al.* Disease-specific ex vivo stimulation of whole blood for cytokine production: applications in the study of tuberculosis. *J. Immunol. Methods* **222,** 145–153 (1999).
44. Hirschfeld, M. *et al.* Cutting Edge: Repurification of Lipopolysaccharide Eliminates Signaling Through Both Human and Murine Toll-Like Receptor 2. *J. Immunol.* **165,** 618–622 (2000).

45. Sutmuller, R. P. M. Toll-like receptor 2 controls expansion and function of regulatory T cells. *J. Clin. Invest.* **116,** 485–494 (2006).
46. Shah, T. S. *et al.* optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. *Bioinformatics* **28,** 1598–603 (2012).
47. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes* **7,** 901 (2014).
48. Delaneau, O. *et al.* Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Genet.* **10,** 5–6 (2013).
49. Howie, B. *et al.* Genotype imputation with thousands of genomes. *G3 genes - genomes - Genet.* **1,** 457–70 (2011).
50. Deelen, P. *et al.* Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of the Netherlands'. *Eur. J. Hum. Genet.* **22,** (2014).
51. Smeekens, S. P. *et al.* Functional genomics identifies type I interferon pathway as central for host defense against Candida albicans. *Nat. Commun.* **4,** 1342 (2013).
52. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
53. Love, M. I. *et al.* Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
54. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28,** 1353–1358 (2012).

# 6

**Nature Genetics, 2017**

Marc Jan Bonder[1,*], René Luijk[2,*], Daria V. Zhernakova[1,**], Matthijs Moed[2,**], Patrick Deelen[1,3,**], Martijn Vermaat[4,**], Maarten van Iterson[2], Freerk van Dijk[1,3], Michiel van Galen[3], Jan Bot[5], Roderick C. Slieker[2], P. Mila Jhamai[6], Michael Verbiest[3], H. Eka D. Suchiman[2], Marijn Verkerk[6], Ruud van der Breggen[2], Jeroen van Rooij[6], Nico Lakenberg[2], Wibowo Arindrarto[8], Szymon M. Kielbasa[7], Iris Jonkers[2], Peter van ’t Hof[7], Irene Nooren[5], Marian Beekman[2], Joris Deelen[2], Diana van Heemst[9], Alexandra Zhernakova[1], Ettje F. Tigchelaar[1], Morris A. Swertz[1,3], Albert Hofman[10], André G. Uitterlinden[6], René Pool[11], Jenny van Dongen[11], Jouke J. Hottenga[11], Coen D.A. Stehouwer[12,13], Carla J.H. van der Kallen[12,13], Casper G. Schalkwijk[12,13], Leonard H. van den Berg[14], Erik. W van Zwet[8], Hailiang Mei[7], Yang Li[1], Mathieu Lemire[15], Thomas J. Hudson[15,16,17], the BIOS Consortium[18], P. Eline Slagboom[2], Cisca Wijmenga[1], Jan H. Veldink[14], Marleen M.J. van Greevenbroek[12,13], Cornelia M. van Duijn[19], Dorret I. Boomsma[11], Aaron Isaacs[19,##], Rick Jansen[20,##], Joyce B.J. van Meurs[6,##], Peter A.C. ’t Hoen[4,#], Lude Franke[1,#], Bastiaan T. Heijmans[2,#]

# Disease variants alter transcription factor levels and methylation of their binding sites

## Abstract

Most disease-associated genetic variants are non-coding, making it challenging to design experiments to understand their functional consequences [1,2]. Identification of expression quantitative trait loci (eQTLs) has been a powerful approach to infer downstream effects of disease variants but the large majority remains unexplained [3,4]. The analysis of DNA methylation, a key component of the epigenome [5,6], offers highly complementary data on the regulatory potential of genomic regions [7,8]. Here, we show that disease variants have wide-spread effects on DNA methylation *in trans* that likely reflect differential occupancy of *trans*-binding sites by *cis*-regulated transcription factors. Using multiple omics data on 3,841 Dutch individuals, we identified 1,907 established trait-associated SNPs that affect methylation levels of 10,141 different CpG sites *in trans* (FDR<0.05). These included SNPs that affect both the expression of a nearby transcription factor (like *NFKB1*, *CTCF* and *NKX2-3*) and methylation of its respective binding site across the genome. *Trans*-meQTLs effectively expose downstream effects of disease-associated variants.



1 Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands
2 Molecular Epidemiology Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands
3 Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands
4 Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
5 SURFsara, Amsterdam, the Netherlands
6 Department of Internal Medicine, ErasmusMC, Rotterdam, The Netherlands
7 Sequence Analysis Support Core, Leiden University Medical Center, Leiden, The Netherlands
8 Medical Statistics Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands
9 Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands
10 Department of Epidemiology, ErasmusMC, Rotterdam, The Netherlands
11 Department of Biological Psychology, VU University Amsterdam, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands
12 Department of Internal Medicine, Maastricht University Medical Center, Maastricht, The Netherlands
13 School for Cardiovascular Diseases (CARIM), Maastricht University Medical Center, Maastricht, The Netherlands
14 Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands
15 Ontario Institute for Cancer Research, Toronto, Ontario, Canada M5G 0A3
16 Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada M5S 1A1
17 Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada M5S 1A1
18 A full list of members and affiliations appears in the Supplementary Note.
19 Genetic Epidemiology Unit, Department of Epidemiology, ErasmusMC, Rotterdam, The Netherlands
20 Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands
* Equal contributions
** Equal contributions
# Equal contributions
## Equal contributions
Corresponding authors: Lude Franke & Bastiaan T. Heijmans

To systematically study the role of DNA methylation in explaining downstream effects of genetic variation, we analysed genome-wide genotype and DNA methylation in whole blood from 3,841 samples from five Dutch biobanks [9–13] (Figure 1, Table S1, Supplemental Text). We found *cis*-meQTL effects for 34.4% of all 405,709 tested CpGs (n=139,566 at a CpG-level FDR of 5%, $P \leq 1.38x10^{-4}$), typically with a short physical distance between the SNP and CpG (median distance 10 kb, Supplementary Fig. 1). By regressing out primary meQTLs effect for each of these CpGs and repeating the *cis*-meQTL mapping, we observed up to 16 independent *cis*-meQTLs for these CpGs (SupplementaryTable 2) totalling 272,037 independent *cis*-meQTL effects. Few factors determine whether a CpG site shows a *cis*-meQTL effect except the variance in methylation level of the CpG site involved (Supplementary Fig. 2, Supplementary Fig. 3a). The proportion of methylation variance explained by SNPs, however, is typically small (Supplementary Fig. 3b). When accounting for this strong effect of CpG variation, we find only modest enrichments and depletions for *cis*-meQTL CpG sites for CpG island and genic annotation (Supplementary Fig. 3e) or when using annotations of biological function based on chromatin segmentations of 27 blood cell types (Figure 2a).



We contrasted these modest functional enrichments to CpGs whose methylation levels correlates with gene expression *in cis* (i.e. mapping expression quantitative trait methylations (eQTMs)), by generating RNA-seq data for 2,101 out of 3,841 individuals in our study. Using a conservative approach that maximally accounts for potential biases (see Methods), we identified 12,809 unique CpGs that correlated to 3,842 unique genes *in cis* (CpG-level FDR < 0.05). eQTMs were enriched for mapping in active regions, e.g. in and around active transcription start sites (TSSs) (3-fold enrichment, $P=1.8x10^{-91}$) and enhancers (2-fold enrichment, $P=1.1x10^{-139}$, Figure 2b). The majority of eQTMs showed the canonical negative correlation with transcriptional activity (69.2%) but a substantial minority of correlations was positive (30.8%) in line with recent evidence that DNA methylation does not always negatively correlate with gene expression [14]. As expected, negatively correlated eQTMs were enriched in active regions like active TSSs (3.7- fold enrichment, $P=9.5x10^{-202}$). Positive correlations primarily occurred in repressed regions (e.g. Polycomb repressed, 3.4-fold enrichment, $P=5.8x10^{-103}$) (Supplementary Fig. 4). The sharp contrast between positively and negatively associated eQTMs enabled us to predict the direction of the correlation. A decision tree trained on the strongest eQTMs (those with an FDR < $9.7x10^{-6}$, n=5,137) using data on histone marks and distance relative to gene, could predict the direction with an area under the curve of 0.83 (95% confidence interval, 0.78-0.87) (Figure 2d, e).

We next ascertained whether *trans*-meQTLs are biologically informative, since previous *trans*-eQTL mapping studies demonstrated that identifying *trans*-expression effects provide a powerful tool to uncover and understand downstream biological effects of disease-SNPs [3,15,16]. We focussed on 6,111 SNPs that were previously associated with complex traits and diseases ('trait-associated SNPs', see Methods and Table S3). We observed that one-third of these trait-associated SNPs (1,907 SNPs, 31.2%) affect methylation *in trans* at 10,141 CpG sites, totalling 27,816 SNP-CpG combinations (FDR<0.05, $P<2.6x10^{-7}$, Figure 3a). This represents a 5-fold increase in the number of CpG sites affected as compared with a previous *trans*-meQTL mapping study [17]. We evaluated whether the GWAS SNP themselves were likely underlying the trans-effects or that the associations could be attributed to another SNP in moderate LD. Of the 1,907 GWAS SNPs with *trans*-effects, 1,538 (87.2%) were in strong LD with the top SNP (R2 > 0.8), indicating that the GWAS SNPs indeed are the driving force behind

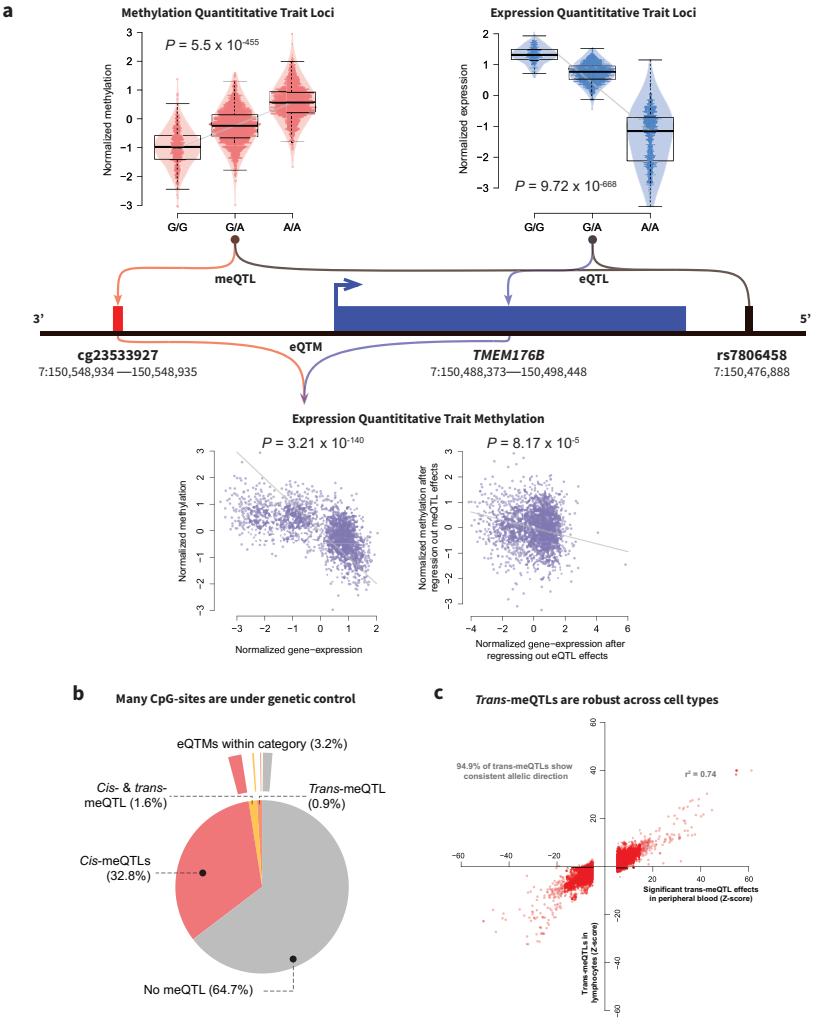



***Figure 1: Overview of a genomic region around TMEM176B and characteristics of CpGs associated to meQTLs and eQTMs.*** *In the illustration, the relations between a SNP, DNA methylation at nearby CpGs, and the associations with the gene itself are shown. Boxes show the median, the inter-quartile range (IQR). Whiskers show the outer quartile plus 1.5 times the IQR. The top left plot shows the observed methylation Quantitative Trait Locus (meQTL) between cg23533927 and rs7806458. The top right plot shows the observed expression Quantitative Trait Locus (eQTL) between TMEM176B and rs7806458. The observed methylation-expression association (eQTM) between TMEM176B and cg23533927, is shown below the gene. The left part shows the data before correction for the cis-eQTL and cis-meQTL, the eQTM effect after correction for cis-eQTLs and cis-meQTLs is shown on the right. b, Two overlaid pie charts. The inner chart indicates the proportion of tested CpGs harboring meQTLs. Over 35% of all tested CpGs show evidence for harboring a meQTL, either in cis or in trans. The outer chart indicates what CpGs are associated with gene expression in cis (in total 3.2%). c, Replication of peripheral blood trans-meQTLs in lymphocytes.*

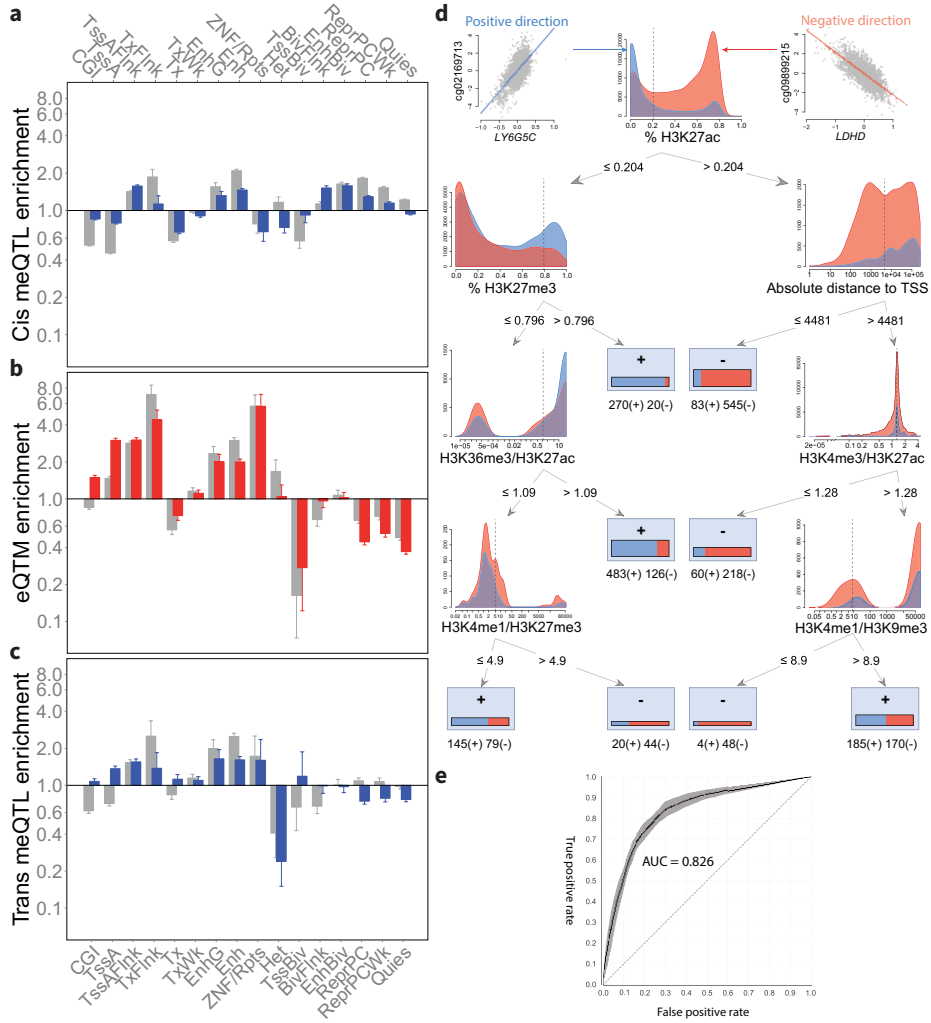



***Figure 2: Characterization of identified cis- and trans-meQTL and eQTM-effects.*** *a-c, Over- or underrepresentation of CpGs for predicted chromatin states for cis-meQTLs, trans-meQTLs and eQTMs. Grey bars reflect uncorrected enrichments, colored bars reflect enrichments after correction for factors influencing the likelihood of harboring a meQTL or eQTM, including methylation variability. Bar graphs show odds ratios and error bars (95% confidence interval). CGI: CpG island; TssA: Active TSS; TssAFlnk: Flanking active TSS; TxFlnk, Transcribed at gene 5' and 3'; Tx: Strong transcription; TxWk: Weak transcription; EnhG: Genic enhancer; Enh: Enhancer; ZNF/Rpts: ZNF genes and repeats; Het: Heterochromatin; TssBiv: Bivalent/ Poised TSS; BivFlnk: Flanking bivalent TSS/Enhancer; EnhBiv: Bivalent enhancer. d, Decision tree for predicting the effect direction of eQTMs. Each subplot shows the distributions for positive (blue) and negative (red) associations for that subset of the data. Dashed vertical lines indicate the optimal split used by the algorithm. The boxes in the leaves indicate the number of positive and negative effects in each of the leaves. e, Receiver operator characteristic curve showing the performance of the decision tree.*

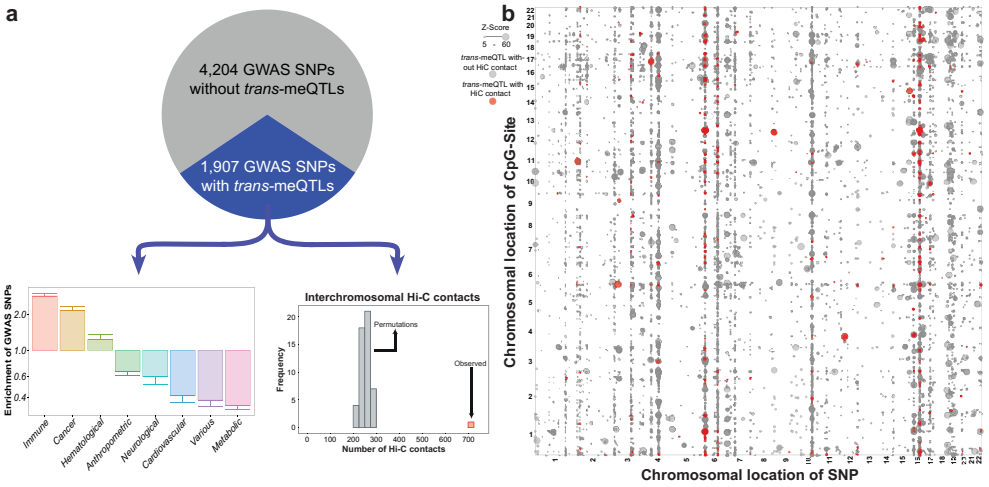



***Figure 3: trans-meQTLs are related to Hi-C interchromosomal contacts and enrichment of GWAS category of trans-meQTL SNPs.*** *a, Distribution of tested trait-associated SNPs influencing DNA methylation in trans. Over 1,900 SNPs (31.2%) of all tested SNPs have downstream effects on DNA methylation. For the associated GWAS SNPs we show the overrepresentation of SNPs with trans-meQTLs in different GWAS trait categories, where the y-axis shows the odds ratio (bottom left). Hi-C contacts are overrepresented among trans-meQTLs. Grey bars show the number of Hi-C contacts using permutated data, while the red bar reflects the actually observed number in our data (bottom right). b, Dot-plot depicting the trans-meQTLs. The effect strength is reflected by the size of the dot. Red dots indicate an overlap with a Hi-C contact. Several SNPs with widespread trans-meQTLs show inter-chromosomal contacts genome-wide, further implicating an important role for those SNPs in the development of the associated trait.*

many of the *trans*-meQTLs. Of note, due to the sparse coverage of the Illumina 450k array, the true number of CpGs in the genome that are altered by these trait associated SNPs will be substantially higher.

To validate our *trans*-meQTLs, we performed a replication analysis in a set of 1,748 lymphocyte samples [17]. Of the 18,764 overlapping *trans*-meQTLs, 94.9% had a consistent allelic direction (Figure 1E; Table S4). This indicates that the identified *trans*-meQTLs are robust and not caused by differences in cell-type composition. Further analysis of SNPs known to influence blood cell composition [18,19] showed no or only few *trans*-effects and alternative adjustments of the methylation-data corroborated the stability of *trans*-effects, both indicating a limited influence of cell type composition (Supplementary Results, Supplementary tables 5–7).

After the identification of the *trans*-meQTLs, we assessed if the *trans*-SNPs also affected expression of the genes associated with the *trans*-CpGs. By overlaying the *trans*-meQTLs and *cis*-eQTMs, we could link 436 SNPs to 850 genes, totalling 2,889 SNP-gene pairs. We found

significant associations (*trans*-eQTLs) (FDR < 0.05) for 8.4% of these effects, and 91% of these effects showed the expected direction of the effect, given the directions of the *trans*-meQTLs and *cis*-eQTMs (Table S8).

In contrast to *cis*-meQTL CpGs, *trans*-meQTLs CpGs show substantial functional enrichments: they are enriched around TSSs and depleted in heterochromatin (Figure 2c) and are strongly enriched for being an eQTM (1,913 CpGs (18.9%), 5.2-fold, $P=2.3x10^{-101}$). Among the 1,907 trait-associated SNPs that make up the *trans*-meQTLs there was an overrepresentation of GWAS-identified SNPs associated with immune- and cancer-related traits (Figure 3b). The large majority of *trans*-meQTLs were inter-chromosomal (93%, 9,429 CpG-SNP pairs) and included 12 *trans*-meQTLs SNPs (yielding 3,616 unique CpG-SNP pairs) that each showed downstream *trans*-meQTL effects across all of the 22 autosomal chromosomes (i.e. *trans*-bands, Figure 3d).



We subsequently studied the nature of these *trans*-meQTLs. Using high-resolution Hi-C data [20], we identified 720 SNP-CpG pairs (including 402 CpG sites and 172 SNPs) among the *trans*-meQTLs that overlapped with an inter-chromosomal contact, which is 2.9-fold more frequent than expected by chance ($P=3.7x10^{-126}$, Figure 3c, d). The enrichment for Hi-C inter-chromosomal contacts remained after removing SNPs that were responsible for *trans*-bands ($P=1.7x10^{-61}$). Hence, inter-chromosomal contacts may produce associations between SNPs and CpGs *in trans*. In order to characterize the 720 SNP-CpG pairs overlapping with inter-chromosomal contacts, we performed motif enrichments using three motif enrichment analyses (Homer, PWMEnrich, DEEPbind) [21–23]. These analyses revealed that the 402 CpG sites involved frequently overlapped with CTCF, RAD21 and SMC3 binding sites ($P=2.3x10^{-5}$, $P=3.5x10^{-5}$ and $P=5.1x10^{-5}$, respectively), factors known to regulate chromatin architecture [24,25]. An analysis of ChIP-Seq data on CTCF binding confirmed this finding (1.8-fold enrichment, $P=5.2x10^{-7}$).

We next tested whether the *trans*-meQTLs reflected the effect of differential transcription factor (TF) binding of TFs that map close to the SNPs. The rationale for this hypothesis is that binding of TFs has been linked to changes in local DNA methylation, primarily loss-of-methylation upon TF binding and gain-of-methylation after loss of TF occupancy [7,8]. This model suggests that *trans*-meQTLs may be attributed to SNPs affecting the expression of a TF *in cis* and that the SNP allele preferentially has a unidirectional effect on DNA methylation. In line with this prediction, we observed that if a SNP is associated with multiple CpGs sites *in trans* (at least 10, n=305), the direction of the association of the SNP was consistently skewed towards either increased or decreased DNA methylation. On average 76% of the CpGs per *trans*-meQTL SNP displayed the same direction of effect (expected 50%, $P=10^{-111}$; Figure 4a). A significant skew in direction of the allelic effect was present for 59.7% of the 305 individual SNPs with at least 10 *trans*-meQTL effects and increased to 95.2% for the 104 SNPs with at least 50 *trans*-meQTL effects (binomial test P<0.05), suggesting that differential TF binding may explain a substantial fraction of *trans*-meQTLs.

In order to explore this mechanism further, we combined ChIP-seq data on TF binding at CpGs and *cis*-expression effects of SNPs to directly examine the involvement of TFs in mediating *trans*-meQTLs. Among trait-associated SNPs influencing at least 10 CpGs *in trans* (n=305), we identified 13 *trans*-meQTL SNPs with strong support for a role of TFs (Figure 4a).

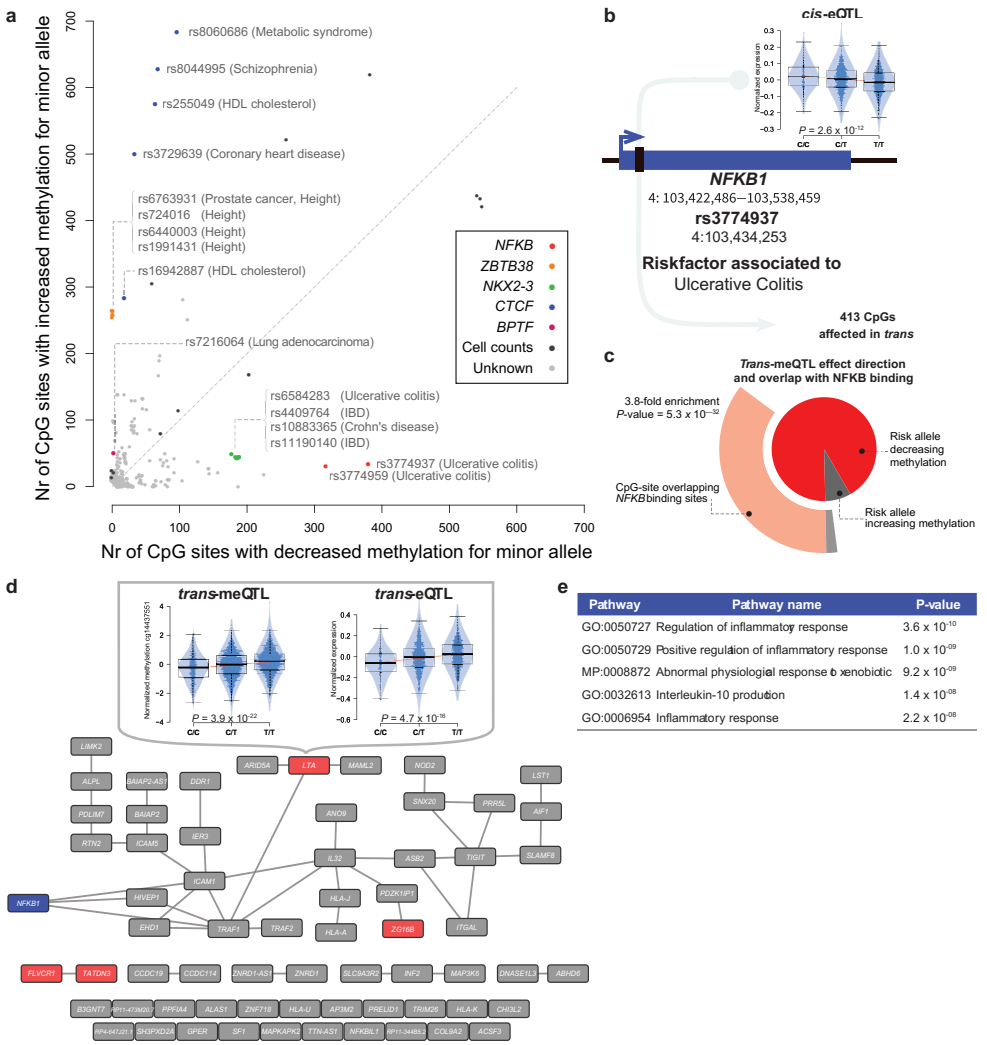



***Figure 4: An imbalance in effect direction of trans-meQTLs implies involvement of transcription factors.*** *a, Each dot represents a SNP with at least 10 trans-meQTL effects. The x-axis shows the number of trans-effects where the minor allele decreases methylation, whereas the y-axis shows an increase in methylation. SNPs with a multitude of effects of which many have the same allelic direction often exhibit evidence for a cis-eQTL on a transcription factor (colored dots), and an overrepresentation of trans-CpGs overlapping binding sites for that transcription factor. b, Depiction of the NFKB1 gene and rs3774937, associated with ulcerative colitis and an increased expression of NFKB1 for the risk and minor allele C. Boxes show the median and IQR. Whiskers show the outer quartile plus 1.5 times the IQR. c, In addition to influencing NFKB1 expression, rs3774937 also relates to DNA methylation at 413 CpGs in trans, decreasing methylation levels at 93% of affected CpG sites (dark grey). Many of the CpG sites (37.3%) overlap with NFKB binding sites (3.8-fold enrichment, P-value=5.3x10-32) (outer chart). d, Gene network of the eQTM genes associated with 72 of the …*

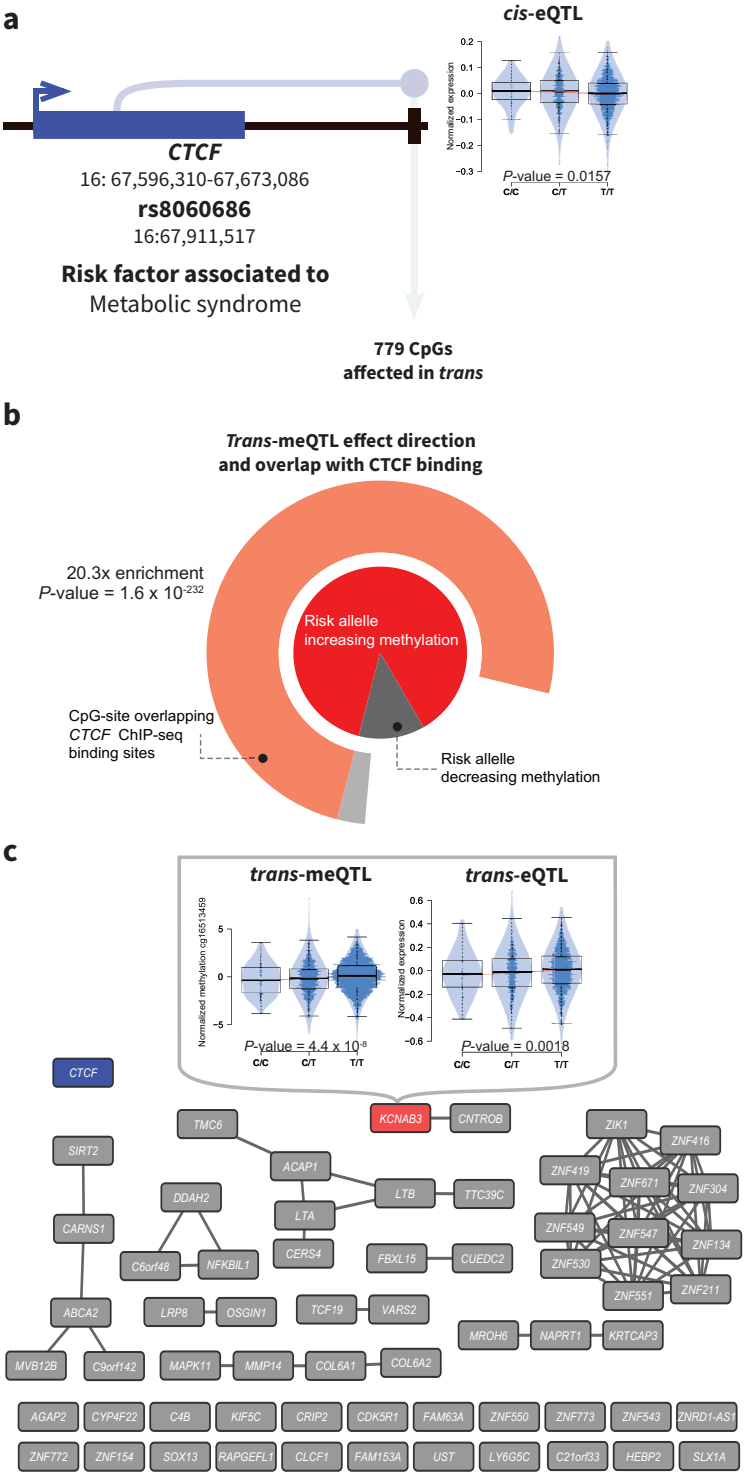
The most striking example was a locus on chromosome 4 (Figure 4b), where two SNPs (rs3774937 and rs3774959, in strong LD) were associated with ulcerative colitis (UC) [26]. Top SNP rs3774937 was associated with differential DNA methylation at 413 CpG sites across the genome, 92% of which showed the same direction of the effect, i.e. lower methylation associated with the minor allele (binomial $P=2.72x10^{-69}$). Of those 380 CpG sites with lower methylation, 147 (38.7%) overlap with a nuclear factor kappaB (NFKB) transcription factor binding site (2.75-fold enrichment, $P=5.3x10^{-32}$), as derived from ENCODE NFKB ChIP-seq data in blood cell types (Figure 4c). Three motif enrichment analyses (Homer, PWMEnrich, DEEPbind) [21–23] corroborated the enrichment of NFKB binding motifs for the 413 CpG sites (Figure 4c). Notably, SNP rs3774937 is located in the first intron of *NFKB1* and we found that the minor allele was associated with higher *NFKB1* expression (Figure 4a). Of the 413 *trans*-CpGs, 64 were eQTMs and revealed a coherent gene network (Figure 4d) that was enriched for immunological processes related to *NFKB1* function [27] (Figure 4e). Taken together, these results support the idea that the minor allele of rs3774937, which is associated with increased UC risk, decreases DNA methylation *in trans* by increasing *NFKB1* expression *in cis*.



The same analysis approach indicated that the 779 *trans*-methylation effects of rs8060686 (associated with various phenotypes including metabolic syndrome [28] and coronary heart disease[29]) were mediated by altered CTCF binding which mapped 315 kb from the *trans*-meQTL SNP. We observed a strong CTCF ChIP-seq enrichment with 603/779 *trans*-CpGs overlapping with CTCF binding ($P=1.6x10^{-232}$) and enrichment for CTCF motifs (Figure 5). Of these *trans*-CpGs, only 13 were observed previously in lymphocytes [17]. Hence, the minor allele of rs8060686 increased DNA methylation *in trans* which could be attributed to a lower *CTCF* gene expression *in cis.*

We found another example of this phenomenon: 228 *trans*-meQTL effects of 4 SNPs on chromosome 10, mapping near *NKX2-3* and implicated in inflammatory bowel disease [26], were strongly enriched for NKX2 transcription factor motifs and associated with *NKX2-3* expression. Again, a negative correlation was observed: the minor allele of rs11190140 decreased DNA methylation *in trans* at NKX2-3 binding sites and increased *NKX2-3* gene expression *in cis* (Supplementary Fig. 6).

*… 413 CpGs (17.4%), that are showing a trans-meQTL and an trans-eQTL (in red). NFKB1 is depicted in blue, illustrations of the observed trans-meQTL (left plot) and trans-eQTL effects (right plot) of rs3774937. e, Top pathways as identified by DEPICT for which the genes in d were overrepresented. Many of the identified pathways were inflammation-related, in line with the inflammatory nature of ulcerative colitis.*


a
cis-eQTL
CTCF
16: 67,596,310-67,673,086
rs8060686
16:67,911,517
Risk factor associated to
Metabolic syndrome
779 CpGs
affected in trans
P-value = 0.0157
C/C
C/T
T/T
b
Trans-meQTL effect direction
and overlap with CTCF binding
20.3x enrichment
P-value = 1.6 x 10-232
Risk allelle
increasing methylation
CpG-site overlapping
CTCF ChIP-seq
binding sites
Risk allelle
decreasing methylation
c
trans-meQTL
trans-eQTL
P-value = 4.4 x 10-8
P-value = 0.0018
CTCF
TMC6
KCNAB3
CNTROB
SIRT2
ACAP1
LTB
TTC39C
DDAH2
LTA
CARNS1
CERS4
C6orf48
NFKBIL1
FBXL15
CUEDC2
ABCA2
LRP8
OSGIN1
TCF19
VARS2
MVB12B
C9orf142
MAPK11
MMP14
COL6A1
COL6A2
ZIK1
ZNF416
ZNF419
ZNF671
ZNF304
ZNF549
ZNF547
ZNF134
ZNF530
ZNF211
ZNF551
MROH6
NAPRT1
KRTCAP3
AGAP2
CYP4F22
C4B
KIF5C
CRIP2
CDK5R1
FAM63A
ZNF550
ZNF773
ZNF543
ZNRD1-AS1
ZNF772
ZNF154
SOX13
RAPGEFL1
CLCF1
FAM153A
UST
LY6G5C
C21orf33
HEBP2
SLX1A

A height locus [30] harbouring 4 SNPs and is associated with 267 *trans*-CpGs implicated a role for *ZBTB38* in mediating *trans*-meQTL effects (Supplementary Fig. 7). In contrast to the aforementioned TFs that are all transcriptional activators, *ZBTB38* is a transcriptional repressor [31,32] and its expression was positively correlated with methylation *in trans*, which is in line with our observation that eQTMs in repressed regions are enriched for positive correlations. Finally, the *trans*-methylation effects of rs7216064 (64 *trans*-CpGs), associated with lung carcinoma [33], preferentially occurred at regions binding CTCF, while the SNP was located in the *BPTF* gene, which is known to occupy CTCF binding sites[34] (Supplementary Fig. 8).

The possibility to link *trans*-meQTL effects to an association of TF expression *in cis* and concomitant differential methylation *in trans* at the respective binding site is limited to TFs for which ChIP-seq data or motif information is available. In order to make inferences on TFs for which such data is not yet available, we ascertained whether *trans*-meQTLs SNPs were more often associated with TF gene expression *in cis* as compared with SNPs without a *trans*-meQTL effect. We observed that 13.1% of the GWAS SNPs that produced *trans*-meQTLs also affect TF gene expression *in cis*, whereas only 4.5% of the GWAS SNPs without a *trans*-meQTLs affects TF gene expression *in cis* (Fisher's exact $P=6.6x10^{-13}$).



Here we report that one third of known disease- and trait-associated SNPs has downstream methylation effects *in trans* and often are associated with multiple regions across the genome. Our data suggest that the biological mechanism underlying *trans*-meQTLs commonly involves a local effect on the expression of a nearby TF that influences DNA methylation at the distal binding sites of that particular TF. The direction of downstream methylation effects is remarkably consistent for each SNP and indicates that decreased DNA methylation is a signature of increased binding of transcriptional activators. As such, our study reveals previously unrecognized functional consequences of disease variants in non-coding regions. These can be looked up online (see URLs), and will provide leads for experimental follow-up.

***Figure 5: Trans-meQTL CpGs related to rs8060686 show overlap with CTCF binding sites.*** *a, Depiction of the CTCF gene and rs8060686, associated with metabolic syndrome. The plot shows an increased expression of NFKB1 for the risk allele C. b, In addition to influencing CTCF expression, rs8060686 also influences DNA methylation at 779 CpGs in trans, increasing methylation levels at 87.7% of affected CpG sites (dark grey). In addition, many of the CpG sites (77.4%) overlap with CTCF binding sites (20.3-fold enrichment, P-value = 1.6 x $10^{-232}$), shown in the outer chart. c, Gene network of the genes associated with 60 of the 779 CpGs (7.7%) with a trans-meQTL. In the top part of the figure, there is an illustration of overlapping trans-meQTL (left) and trans-eQTL effects (right) for rs8060686.*

# Methods

## Cohort descriptions

The five cohorts used in our study are described briefly below. The number of samples per cohort and references to full cohort descriptions can be found in Table S1.

### *CODAM*

The Cohort on Diabetes and Atherosclerosis Maastricht [10] (CODAM) consists of a selection of 547 subjects from a larger population-based cohort [35]. Inclusion of subjects into CODAM was based on a moderately increased risk to develop cardiometabolic diseases, such as type 2 diabetes and/or cardiovascular disease. Subjects were included if they were of Caucasian descent and over 40 years of age and additionally met at least one of the following criteria: increased BMI (>25), a positive family history of type 2 diabetes, a history of gestational diabetes and/or glycosuria, or use of anti-hypertensive medication.

### *LifeLines-DEEP*

The LifeLines-DEEP (LLD) cohort [9] is a sub-cohort of the LifeLines cohort [36]. LifeLines is a multi-disciplinary prospective population-based cohort study examining the health and health-related behaviours of 167,729 individuals living in the northern parts of The Netherlands using a unique three-generation design. It employs a broad range of investigative procedures assessing the biomedical, socio-demographic, behavioural, physical and psychological factors contributing to health and disease in the general population. A subset of 1,500 LifeLines participants also take part in LLD [9]. For these participants, additional molecular data is generated, allowing for a more thorough investigation of the association between genetic and phenotypic variation.

### *LLS*

The aim of the Leiden Longevity Study [11] (LLS) is to identify genetic factors influencing longevity and examine their interaction with the environment in order to develop interventions to increase health at older ages. To this end, long-lived siblings of European descent were recruited together with their offspring and their offspring's partners, on the condition that at least two long-lived siblings were alive at the time of ascertainment. For men the age criteria was 89 or older, for women age 91 or over. These criteria led to the ascertainment of 944 long-lived siblings from 421 families, together with 1,671 of their offspring and 744 partners.

### *NTR*

The Netherlands Twin Register [12,37,38] (NTR) was established in 1987 to study the extent to which genetic and environmental influences cause phenotypic differences between individuals. To this end, data from twins and their families (nearly 200,000 participants) from all over the Netherlands are collected, with a focus on health, lifestyle, personality, brain development, cognition, mental health, and aging.

*RS*

The Rotterdam Study [13] is a single-centre, prospective population-based cohort study conducted in Rotterdam, the Netherlands [13]. Subjects were included in different phases, with a total of 14,926 men and women aged 45 and over included as of late 2008. The main objective of the Rotterdam Study is to investigate the prevalence and incidence of and risk factors for chronic diseases to contribute to a better prevention and treatment of such diseases in the elderly.

## Genotype data

### *Data generation*

Genotype data was generated for each cohort individually. Details on the methods used can be found in the individual papers (CODAM: van Dam et al. [35]; LLD: Tigchelaar et al. [9]; LLS: Deelen et al. [39], 2014; NTR: Willemsen et al. [12]; RS: Hofman et al. [13]).



### *Imputation and QC*

For each cohort separately, the genotype data were harmonized towards the Genome of the Netherlands [40] (GoNL) using Genotype Hamonizer[41] and subsequently imputed per cohort using Impute2 [42] using GoNL [43] reference panel (v5). Quality control was also performed per cohort. We removed SNPs based on imputation info-score (<0.5), HWE ($P$<$10^{-4}$), call rate (<95%) and minor allele frequency (>0.05), resulting in 5,206,562 SNPs that passed quality control in each of the datasets.

## Methylation data

### *Data generation*

For the generation of genome-wide DNA methylation data, 500 ng of genomic DNA was bisulfite modified using the EZ DNA Methylation kit (Zymo Research, Irvine, California, USA) and hybridized on Illumina 450k arrays according to the manufacturer's protocols. The original IDAT files were generated by the Illumina iScan BeadChip scanner. We collected methylation data for a total of 3,841 samples. Data was generated by the Human Genotyping facility (HugeF) of ErasmusMC, the Netherlands (see URLs).

### *Probe remapping and selection*

We remapped the 450K probes to the human genome reference (HG19) to correct for inaccurate mappings of probes and identify probes that mapped to multiple locations on the genome. Details on this procedure can be found in Bonder et al. (2014) [44]. Next, we removed probes with a known SNP (GoNL, MAF > 0.01) at the single base extension (SBE) site or CpG site. Lastly, we removed all probes on the sex chromosomes, leaving 405,709 high quality methylation probes for the analyses.

### *Normalization and QC*

Methylation data was processed using a custom pipeline based on the pipeline developed by Tost & Toulemat [45]. First, we used methylumi [46] to extract the data from the raw IDAT files. Next, we removed incorrectly mapped probes and checked for outlying samples using the first two principal components (PCs) obtained using principal component analysis (PCA). None of the samples failed our quality control checks, indicating high quality data. Following

quality control, we performed background correction and probe type normalization as implemented in DASEN [47]. Normalization was performed per cohort, followed by quantile normalization on the combined data to normalize the differences per cohort. We used mix-up mapper [48] to identify sample mix-ups between genotype and DNA methylation data, detecting and correcting 193 mix-ups. Lastly, in order to correct for known and unknown confounding sources of variation in the methylation data and increase statistical power, we removed the first components which were not affected by genetic information(22 PCs) from the methylation data using methodology we have successfully used in *trans*-eQTL [3,49] and meQTL analyses [44].

## RNA sequencing

Total RNA from whole blood was deprived of globin using Ambion's GLOBIN clear kit and subsequently processed for sequencing using Illumina's Truseq version 2 library preparation kit. Paired-end sequencing of 2x50bp was performed using Illumina's Hiseq2000, pooling 10 samples per lane. Finally, read sets per sample were generated using CASAVA, retaining only reads passing Illumina's Chastity Filter for further processing. Data was generated by the Human Genotyping facility (HugeF) of ErasmusMC, the Netherlands (see URLs).

Initial QC was performed using FastQC,v0.10.1 (See URLs),, removal of adaptors was performed using cutadapt [50] (v1.1), and Sickle,v1.2 See URLs) was used to trim low quality ends of the reads (min length 25, min quality 20). The sequencing reads were mapped to human genome (HG19) using STAR [51] v2.3.125 . Gene expression quantification was performed by HTseq-count. The gene definitions used for quantification were based on Ensembl version 71, with the extension that regions with overlapping exons were treated as separate genes and reads mapping within these overlapping parts did not count towards expression of the normal genes.

Expression data on the gene level were first normalized using Trimmed Mean of M-values [52]. Then expression values were log2 transformed, gene and sample means were centred to zero. To correct for batch effects, PCA was run on the sample correlation matrix and the first 25 PCs were removed using methodology that we have used before [3,49], details are provided in Zhernakova et al. [53].

## *Cis*-meQTL mapping

In order to determine the effect of nearby genetic variation on methylation levels (*cis*-meQTLs, here defined as the relationship between a CpG and a SNP no further than 250kb apart), we performed *cis*-meQTL mapping using 3,841 samples for which both genotype data and methylation data were available. To this end, we calculated the Spearman rank correlation per cohort, followed by meta-analysis using a weighted Z-method described previously [3]. To detect all possible independent SNPs regulating methylation at a single CpG-site we regressed out all primary *cis*-meQTL effects and then performed *cis*-meQTL mapping for the same CpG-site to find secondary *cis*-meQTL. We repeated this in a stepwise fashion until no more independent *cis*-meQTL were found.

To filter out potential false positive *cis*-meQTLs caused by SNPs affecting the binding of a probe on the array, we filtered the *cis*-meQTLs effects by removing any CpG-SNP pair for which the SNP was located in the probe. In addition, all other CpG-SNP pairs for which the

SNP was outside the probe, but in LD ($R^2 > 0.2$ or D' > 0.2) with a SNP inside the probe were also removed. We tested for LD between SNPs in the probe and in the surrounding *cis* area in the individual genotype datasets, as well as in GoNL v5, in order to be as strict as possible in marking a QTL as true positive.

To correct for multiple testing, we empirically controlled the false discovery rate (FDR) at 5%. For this, we compared the distribution of observed *P*-values to the distribution obtained from performing the analysis on permuted data. Permutation was done by shuffling the sample identifiers of one data set, breaking the link between, e.g., the genotype data and the methylation or expression data. We repeated this procedure 10 times to obtain a stable distribution of *P*-values under the null distribution. The FDR was determined by only selecting the strongest effect per CpG [3] in both the real analysis and in the permutations (i.e. probe level FDR < 5%).



## Cis-eQTL mapping

For a set of 2,116 BIOS samples we had also generated RNA-seq data. We used this data to identify *cis*-eQTLs. *Cis*-eQTL mapping was performed using the same method as *cis*-meQTL mapping. Details on these eQTLs will be described in a separate paper [53].

## Expression quantitative trait methylation (eQTM) analysis

To identify associations between methylation levels and expression levels of nearby genes (*cis*-eQTMs), we first corrected our expression and methylation data for batch effects and covariates by regressing out the PCs and regressing out the identified *cis*-meQTLs and *cis*-eQTLs, to ensure identified associations between CpG sites and gene expression levels were not due to shared genetic effects. We mapped eQTMs in a window of 250Kb around the TSS of a transcript. Further statistical analysis was identical to the *cis*-meQTL mapping. For this analysis we were able to use a total of 2,101 samples for which both genetic, methylation and gene expression data was available. To correct for multiple testing we controlled the FDR at 5%, the FDR was determined by only selecting the strongest effect per CpG [3] in both the real analysis and in the permutations.

## *Trans*-meQTL mapping

To identify the effects of distal genetic variation with methylation (*trans*-meQTLs) we used the same 3,841 samples that we had used for *cis*-meQTL mapping. To focus our analysis and limit the multiple testing burden, we restricted our analysis to SNPs that have been previously found to be significantly correlated to traits and diseases. We extracted these SNPs from the NHGRI genome-wide association study (GWAS) catalogue, used recent GWAS studies not yet in the NHGRI GWAS catalogue and studies on the Immunochip and Metabochip platform that are not included in the NHGRI GWAS catalogue (Supplemental file 1). We compiled this list of SNPs in December 2014. Per SNP we only investigated CpG sites that mapped at least 5 Mb from the SNP or on other chromosomes. Before mapping *trans*-meQTLs, we regressed out the identified *cis*-meQTLs to increase the statistical power of *trans*-meQTL detection (as done previously for *trans*-eQTLs [3]) and to avoid designating an association as *trans* that may be due to long-range LD (e.g. within the HLA region). To

ascertain the stability of the *trans*-meQTLs we also performed the *trans*-mapping using uncorrected data cell-type proportions corrected methylation data. In addition, we performed meQTL mapping on SNPs known to influence the cell type proportions in blood[18,19].

To filter out potential false positive *trans*-meQTLs due to cross-hybridization of the probe, we remapped the methylation probes with very relaxed settings, identical to Westra et al. [3], with the difference that we only accepted mappings if the last bases of the probe including the SBE site were accurately mapped to the alternative location. If the probe mapped within our minimal *trans*-window, 5 Mb from the SNP, we removed the effect as being a false positive *trans*-meQTL.

We controlled the false-discovery rate at 5%, identical to the aforementioned *cis*-meQTL analysis.

## *Trans*-eQTL mapping

To check if the *trans*-meQTL effects also showed in gene expression levels, we annotated the CpGs with a *trans*-meQTL to genes using our eQTMs. Using the 2,101 samples for which both genotype and gene expression data were available, we performed *trans*-eQTL mapping, associating the SNPs known to be associated with DNA methylation in *trans* with their corresponding eQTM genes.

## Annotations and enrichment tests

Annotation of the CpGs was performed using Ensembl[54] (v70), UCSC Genome Browser [55] and data from the Epigenomics Roadmap Project [56]. We used the Epigenomics Roadmap annotation for the SBE site of the methylation site using 27 blood cell types. We used both the histone mark information and the chromatin marks in blood-related cell types only, as generated by the Epigenomics Roadmap Project. Summarizing the information over the 27 blood cell types was done by counting presence of histone-marks in all the cell types and scaling the abundance, i.e. if the mark is bound in all cell types the score would be 1 if it would be present in none of the blood cell types the score would be 0.

To calculate enrichment of meQTLs or eQTMs for any particular genomic context, we used logistic regression because this allowed us to account for covariates such as CpG methylation variation. For *cis*-meQTLs, we used the variability of DNA methylation, the number of SNPs tested, and the distance to the nearest SNP per CpG as covariates. For all other analyses we used only the variability in DNA methylation as a covariate.

We used transcription factor ChIP-seq data from the ENCODE-project for blood-related cell lines (narrow peak data). We overlapped CpG locations with ChIP-seq signals and performed a Fisher exact test to determine whether the *trans*-meQTL probes associated with a SNP were overlapping a ChIP-seq region more often than other *trans*-meQTL probes.

Enrichment of known sequence motifs among *trans*-CpGs was assessed by PWMEnrich [22] package in R, Homer [57] and DEEPbind [23]. For PWMEnrich, hundred base pair sequences around the interrogated CpG site were used, and as a background set we used the top CpGs from the 50 permutations used to determine the FDR threshold of the *trans*-meQTLs. For Homer the default settings for motif enrichment identification were used, and the same

CpGs derived from the permutations were used as a background. For DEEPbind we used both the permutation background like described for Homer and the permutations background as described for PWMEnrich.

Using data published by Rao et al. [20] we were able to intersect the *trans*-meQTLs with information about the 3D structure of the human genome using combined Hi-C data for both inter- and intra-chromosomal data at 1Kb and the quality threshold of E30 in the GM12878 lymphoblastoid cell line. Both the *trans*-meQTL SNP and *trans*-meQTL probes were put in the relevant 1Kb block, and for these blocks we looked up the chromosomal contact value in the measurements by Rao et al. Surrounding the *trans*-meQTLs SNPs, we used a LD window that spans maximally 250Kb from the *trans*-meQTL SNP and had a minimal $R^2$ of 0.8. If a Hi-C contact between the SNP block and the CpG-site was indicated, we flagged the region as a positive for Hi-C contacts. As a background, we used the combinations found in our 50 permutated *trans*-meQTL analyses, taking for each permutation the top *trans*-meQTLs that were similar in size to the real analysis.



## eQTM direction prediction

We predicted the direction of the eQTM effects using both a decision tree and a naive Bayes model (as implemented by Rapid-miner [58] v6.3). We built the models on the strongest eQTMs (FDR<$9.73x10^{-6}$). For the decision tree we used a standard cross-validation set-up using 20 folds. For the naive Bayes model we used a double loop cross-validation: performance was evaluated in the outer loop using 20-fold cross-validation, while feature selection (using both backward elimination and forward selection) took place in the inner loop using 10-fold cross-validation. Details about the double-loop cross-validation can be found in Ronde et al. [59]. During the training of the model, we balanced the two classes making sure we had an equal number of positively correlating and negatively correlating CpG-gene combinations, by randomly sampling a subset of the overrepresented negatively correlating CpG-gene combination group. We chose to do so to circumvent labelling al eQTMs as negative, since this is the class were the majority of the eQTMs are in.

In the models we used CpG-centric annotations: overlap with epigenomics roadmap chromatin states, histone marks and relations between the histone marks, GC content surrounding the CpG-site and relative locations from the CpG-site to the transcript.

## DEPICT

To investigate whether there was biological coherence in the *trans*-meQTLs identified for the *NFKB1* locus, we performed gene-set enrichment analysis for the genes near the *trans*-CpG sites of the UC genetic risk factor (which maps in the *NFKB1* locus). To do so, we adapted DEPICT [27], a pathway enrichment analysis method that we originally have developed for GWAS. Instead of defining loci with genes by using the top associated SNPs (as is done when analysing GWAS data), we used the eQTM information to empirically link *trans*-CpGs to genes (that map close to the CpGs). Within the DEPICT gene set enrichment, significance is determined by using a background set of genes. As a background in the adapted DEPICT enrichment analyses we matched our background to the results from the actual *trans*-meQTL and eQTM analyses: the matching was performed by generating a set of background CpGs (and corresponding correlating eQTM genes), by selecting an equal number of CpGs

for which we had found *trans*-meQTL effects with SNPs that map outside the *NFKB* locus. By doing so we ensure that the characteristics of these background CpGs are the same as the real *NFKB trans*-meQTL CpGs, both in terms of CpG variance and the requirement that they also show a significant correlation with expression levels of genes close to the CpG (i.e. a *cis*-eQTM), ensuring that the corresponding input genes for DEPICT have the same expression variation distribution in the actual *NFKB* analysis and in the background. Subsequent pathway enrichment analysis was conducted as described before [27], and significance was determined by controlling the false discovery rate at 5%.

## URLs

- All results can be queried using our dedicated QTL browser: www.genenetwork.nl/biosqtlbrowser
- Data was generated by the Human Genotyping facility (HugeF) of ErasmusMC, the Netherlands, see: www.glimDNA.org
- LifeLines: http://lifelines.nl/lifelines-research/general
- Leiden Longevity Study http://www.healthy-ageing.nl & http://www.leidenlangleven.nl
- Netherlands Twin Registry: http://www.tweelingenregister.org
- Rotterdam studies: http://www.erasmus-epidemiology.nl/research/ergo.htm, the Genetic Research in Isolated Populations program: http://www.epib.nl/research/genetice-pi/research.html#gip
- Codam study http://www.carimmaastricht.nl/
- PAN study: http://www.alsonderzoek.nl/
- FastQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- Sickle: https://github.com/najoshi/sickle

## Accession codes

All results can be queried using our dedicated QTL browser, see URLs. Raw data was submitted to the European Genome-phenome Archive (EGA), under accession EGAS00001001077.

## Acknowledgements

This work was performed 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). Samples were contributed by LifeLines, the Leiden Longevity Study, the Netherlands Twin Registry (NTR), the Rotterdam studies, the Genetic Research in Isolated Populations program, the Codam study and the PAN study. We thank the participants of all aforementioned biobanks and acknowledge the contributions of the investigators to this study (Supplemental Acknowledgements). This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. L.F. is supported by a grant from the Dutch Research Council (ZonMW-VIDI 917.14.374) and is supported by FP7/2007–2013, grant agreement 259867, and by an ERC Starting Grant, grant agreement 637640 (ImmRisk).

## Author contributions

BTH, PACtH, JBJvM, AI, RJ and LF formed the management team of the BIOS consortium. DIB, RP, JVD, JJH, MMJVG, CDAS, CJHvdK, CGS, CW, LF, AZ, EFG, PES, MB, JD, DvH, JHV, LH-vdB, CMvD, BAH, AI, AGU managed and organized the biobanks. JBJvM, PMJ, MV, HEDS, MV, RvdB, JvR and NL generated RNA-seq and Illumina 450k data. HM, MvI, MvG, JB, DVZ, RJ, PvtH, PD, IN, PACtH, BTH and MM were responsible for data management and the computational infrastructure. MJB, RL, MV, DVZ, RS, IJ, MvI, PD, FvD, MvG, WA, SMK, MAS, EWvZ, RJ, PACtH, LF and BTH performed the data analysis. MJB, RL, LF and BTH drafted the manuscript. D.V.Z, M.M., P.D. and M.V. contributed equally. A.I., R.J. and J.B.J.M. contributed equally

## Additional material

The following supplements are available with the on-line version of this paper.



| | |
|---|---|
| Figure S1: | Density of distances between CpG-site and strongest associated meQTL SNP |
| Figure S2: | Relation between methylation variation and meQTL associated CpGs |
| Figure S3: | Characterization of cis-meQTLs |
| Figure S4: | Characterization of cis-eQTMs in relation to the direction of the eQTM effect |
| Figure S5: | Trans-meQTLs identified for a risk factor for inflammatory bowel disease, rs11190140, and the overlap with NKX2-3 |
| Figure S6: | Trans-meQTLs identified for a risk factor for height, rs6763931, and the overlap with ZBTB38 |
| Figure S7: | Trans-meQTLs identified for a risk factor related to lung carcinoma, rs7216064, and overlap with BPTF |
| Table S1: | Descriptions and number of samples per cohort. |
| Table S2: | Number of independent cis-meQTLs per QTL mapping round |
| Table S3: | GWAS SNPs tested for trans-meQTLs |
| Table S4: | Replication of lymphocytes trans-meQTLs in blood and vice-versa |
| Table S5: | Results of trans-meQTL in non-corrected data |
| Table S6: | Results of trans-meQTL in blood-cell composition corrected data |
| Table S7: | Results of trans-meQTL mapping on Blood cell composition related SNPs |
| Table S8: | Trans-meQTL effects replicated in expression |
| Note S1: | Supplementary results & acknowledgements |

## References

1. Manolio, T. A. Genomewide Association Studies and Assessment of the Risk of Disease. *New Engl. J. Med.* **362,** 166–176 (2010).
2. Visscher, P. M. *et al.* Five years of GWAS discovery. *Am. J. Hum. Genet.* **90,** 7–24 (2012).
3. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).
4. Wright, F. A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46,** 430–7 (2014).
5. Bernstein, B. E. *et al.* The Mammalian Epigenome. *Cell* **128,** 669–681 (2007).

6. Mill, J. *et al.* From promises to practical strategies in epigenetic epidemiology. *Nat. Rev. Genet.* **14,** 585–594 (2013).
7. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2013,** e00523 (2013).
8. Tsankov, A. M. *et al.* Transcription factor binding dynamics during human ES cell differentiation. *Nature* **518,** 344–349 (2015).
9. Tigchelaar, E. F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5,** (2015).
10. van Greevenbroek, M. M. J. *et al.* The cross-sectional association between insulin resistance and circulating complement C3 is partly explained by plasma alanine aminotransferase, independent of central obesity and general inflammation (the CODAM study). *Eur. J. Clin. Invest.* **41,** 372–379 (2011).
11. Schoenmaker, M. *et al.* Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *Eur. J. Hum. Genet.* **14,** 79–84 (2006).
12. Willemsen, G. *et al.* The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection. *Twin Res. Hum. Genet.* **16,** 271–281 (2013).
13. Hofman, A. *et al.* The rotterdam study: 2014 objectives and design update. *Eur. J. Epidemiol.* **28,** 889–926 (2013).
14. Hu, S. *et al.* DNA methylation presents distinct binding sites for human transcription factors. *Elife* **2013,** 1–16 (2013).
15. Yao, C. *et al.* Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease phenotypes. *Circulation* **131,** 536–549 (2015).
16. Huan, T. *et al.* A meta-analysis of gene expression signatures of blood pressure and hypertension. *PLoS Genet.* **11,** e1005035 (2015).
17. Lemire, M. *et al.* Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nat. Commun.* **6,** 6326 (2015).
18. Orrù, V. *et al.* Genetic Variants Regulating Immune Cell Levels in Health and Disease. *Cell* **155,** 242–256 (2013).
19. Roederer, M. *et al.* The Genetic Architecture of the Human Immune System: A Bioresource for Autoimmunity and Disease Pathogenesis. *Cell* **161,** 387–403 (2015).
20. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).
21. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38,** 576–589 (2010).
22. D, S. R. and D. PWMEnrich: PWM enrichment analysis. R package version 4.6.0. (2015).
23. Alipanahi, B. *et al.* Supp:Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33,** 831–838 (2015).
24. Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci USA* **111,** 996–1001 (2014).
25. Splinter, E. *et al.* CTCF mediates long-range chromatin looping and local histone modification in the ??-globin locus. *Genes Dev.* **20,** 2349–2354 (2006).
26. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–124 (2012).

27. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* **6,** (2015).
28. Kristiansson, K. *et al.* Genome-wide screen for metabolic syndrome susceptibility loci reveals strong lipid gene contribution but no evidence for common genetic basis for clustering of metabolic syndrome traits. *Circ. Cardiovasc. Genet.* **5,** 242–249 (2012).
29. Lettre, G. *et al.* Genome-Wide association study of coronary heart disease and its risk factors in 8,090 african americans: The nhlbi CARe project. *PLoS Genet.* **7,** (2011).
30. Soranzo, N. *et al.* Meta-analysis of genome-wide scans for human adult stature identifies novel loci and associations with measures of skeletal frame size. *PLoS Genet.* **5,** (2009).
31. Filion, G. J. P. *et al.* A Family of Human Zinc Finger Proteins That Bind Methylated DNA and Repress Transcription A Family of Human Zinc Finger Proteins That Bind Methylated DNA and Repress Transcription. *Mol. Cell. Biol.* **26,** 169 (2006).
32. Sasai, N. *et al.* Many paths to one goal? The proteins that recognize methylated DNA in eukaryotes. *Int. J. Dev. Biol.* **53,** 323–334 (2009).
33. Shiraishi, K. *et al.* A genome-wide association study identifies two new susceptibility loci for lung adenocarcinoma in the Japanese population. *Nat. Genet.* **44,** 900–903 (2012).
34. Qiu, Z. *et al.* Functional Interactions between NURF and Ctcf Regulate Gene Expression. *Mol. Cell. Biol.* **35,** 224–37 (2015).
35. Van Dam, R. M. *et al.* Parental history off diabetes modifies the association between abdominal adiposity and hyperglycemia. *Diabetes Care* **24,** 1454–1459 (2001).
36. Scholtens, S. *et al.* Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int. J. Epidemiol.* **44,** 1172–80 (2015).
37. Boomsma, D. I. *et al.* Netherlands Twin Register: a focus on longitudinal research. *Twin Res* **5,** 401–406 (2002).
38. Boomsma, D. I. *et al.* Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. *Eur. J. Hum. Genet.* **16,** 335–342 (2008).
39. Deelen, J. *et al.* Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Hum. Mol. Genet.* **23,** 4420–32 (2014).
40. Francioli, L. C. *et al.* Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46,** (2014).
41. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes* **7,** 901 (2014).
42. Howie, B. *et al.* A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).
43. Deelen, P. *et al.* Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *Eur. J. Hum. Genet.* **22,** 1321–1326 (2014).
44. Bonder, M. J. *et al.* Genetic and epigenetic regulation of gene expression in fetal and adult human livers. *BMC Genomics* **15,** 860 (2014).
45. Touleimat, N. *et al.* Complete pipeline for Infinium(®) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* **4,** 325–41 (2012).

46. Davis, S. *et al.* Methylumi: Handle Illumina methylation data. *R Packag. version 2.2.0.* (2012).
47. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14,** 293 (2013).
48. Westra, H.-J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* **27,** 2104–11 (2011).
49. Fehrmann, R. S. N. *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.* **7,** e1002197 (2011).
50. Martin, M. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011. Date of access 05/08/2015. . **17,** 10–12 (2011).
51. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
52. Robinson, M. D. *et al.* A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11,** R25 (2010).
53. Zhernakova, D. V *et al.* Hypothesis-free identification of modulators of genetic risk factors. *bioRxiv* 1–25 (2015). doi:10.1101/033217
54. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41,** 48–55 (2013).
55. Kent, W. J. *et al.* The Human Genome Browser at UCSC W. *J. Med. Chem.* **19,** 1228–1231 (1976).
56. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
57. Heinz, S. *et al.* Effect of natural genetic variation on enhancer selection and function. *Nature* **503,** 487–492 (2013).
58. Hofmann, M. *et al. Rapid Miner Data Mining Use Cases and Business Analytics Applications*. (Chapman & Hall/CRC, 2013).
59. De Ronde, J. J. *et al.* Breast cancer subtype specific classifiers of response to neoadjuvant chemotherapy do not outperform classifiers trained on all subtypes. *PLoS One* **9,** e88551 (2014).

# 7

**Nature Genetics, 2017**

Daria V. Zhernakova[1,*], Patrick Deelen[1,2,*], Martijn Vermaat[3,*], Maarten van Iterson[4,*], Michiel van Galen[3], Wibowo Arindrarto[5], Peter van 't Hof[5], Hailiang Mei[5], Freerk van Dijk[1,2], Harm-Jan Westra[6,7,8], Marc Jan Bonder[1], Jeroen van Rooij[9], Marijn Verkerk[9], P. Mila Jhamai[9], Matthijs Moed[4], Szymon M. Kielbasa[4] , Jan Bot[10], Irene Nooren[10], René Pool[11], Jenny van Dongen[11], Jouke J. Hottenga[11], Coen D.A. Stehouwer[12], Carla J.H. van der Kallen[12], Casper G. Schalkwijk[12], Alexandra Zhernakova[1], Yang Li[1], Ettje F. Tigchelaar[1], Niek de Klein[1], Marian Beekman[4], Joris Deelen[4], Diana van Heemst[13], Leonard H. van den Berg[14], Albert Hofman[15], André G. Uitterlinden[9] , Marleen M.J. van Greevenbroek[12], Jan H. Veldink[16], Dorret I. Boomsma[11], Cornelia M. van Duijn[17], Cisca Wijmenga[1], P. Eline Slagboom[4], Morris A. Swertz[1,2], Aaron Isaacs[17,18], Joyce B.J. van Meurs[9], Rick Jansen[19], Bastiaan T. Heijmans[4,#], Peter A.C. 't Hoen[3,#], Lude Franke[1,#]

# Identification of context-dependent expression quantitative trait loci in whole blood

## Abstract

Genetic risk factors often localize to noncoding regions of the genome with unknown effects on disease etiology. Expression quantitative trait loci (eQTLs) help to explain the regulatory mechanisms underlying these genetic associations. Knowledge of the context that determines the nature and strength of eQTLs may help identify cell types relevant to pathophysiology and the regulatory networks underlying disease. Here we generated peripheral blood RNA-seq data from 2,116 unrelated individuals and systematically identified context-dependent eQTLs using a hypothesis-free strategy that does not require previous knowledge of the identity of the modifiers. Of the 23,060 significant cis-regulated genes (false discovery rate (FDR) ≤ 0.05), 2,743 (12%) showed context-dependent eQTL effects. The majority of these effects were influenced by cell type composition. A set of 145 cis-eQTLs depended on type I interferon signaling. Others were modulated by specific transcription factors binding to the eQTL SNPs.



1 University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands
2 University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, the Netherlands
3 Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands
4 Molecular Epidemiology Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands
5 Sequence Analysis Support Core, Leiden University Medical Center, Leiden, the Netherlands
6 Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, USA
7 Partners Center for Personalized Genetic Medicine, Boston, USA
8 Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, USA
9 Department of Internal Medicine, ErasmusMC, Rotterdam, the Netherlands
10 SURFsara, Amsterdam, the Netherlands
11 Department of Biological Psychology, VU Amsterdam, Neuroscience Campus Amsterdam, Amsterdam, the Netherlands
12 Department of Internal Medicine and School for Cardiovascular Diseases (CARIM), Maastricht University Medical Center, Maastricht, the Netherlands
13 Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, the Netherlands
14 Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, the Netherlands
15 Department of Epidemiology, ErasmusMC, Rotterdam, The Netherlands
16 Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, the Netherlands
17 Genetic Epidemiology Unit, Department of Epidemiology, ErasmusMC, Rotterdam, the Netherlands
18 CARIM School for Cardiovascular Diseases and Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, the Netherlands
19 Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, the Netherlands
* Equal contributions
# Equal contributions
Corresponding authors: Lude Franke, Peter A.C. 't Hoen & Bastiaan T. Heijmans

Genetic risk factors often localize in non-coding regions of the genome with unknown effects on disease etiology [1,2]. Expression quantitative trait loci (eQTLs) help to explain the regulatory mechanisms underlying these genetic associations [3–6]. Knowledge of the context that determines the nature and strength of eQTLs may help identify cell types relevant to the pathophysiology and the regulatory networks underlying disease [7–17]. Here, we generated peripheral blood RNA-seq data from 2,116 unrelated individuals and systematically identified context-dependent eQTLs using a hypothesis-free strategy that does not require prior knowledge of the identity of the modifiers. Of the 23,060 significant *cis*-regulated genes (false discovery rate (FDR) ≤ 0.05), 2,743 (12%) showed context-dependent eQTL effects. The majority of those were influenced by cell type composition. A set of 145 *cis*-eQTLs depended on type I interferon signaling. Others were modulated by specific transcription factors binding to the eQTL SNPs.



We created the Biobank-based Integrative Omics Study (BIOS) dataset by sequencing whole peripheral blood mRNA in 2,116 healthy adults from four Dutch cohorts [18–21] (Supplementary table 1, Supplementary note chapter 1, https://www.ebi.ac.uk/ega/datasets/EGAD00001001623). We quantified gene and exon expression, as well as exon ratios (the proportion of expression of an exon relative to the total expression of all exons of a gene) and polyA ratios (the ratio of the expression in upstream and downstream parts of the 3'-UTRs separated by annotated polyadenylation (polyA) sites) and performed *cis*-eQTL mapping for all of these (Supplementary note chapter 2). We detected *cis*-eQTL effects for 66% of the protein coding genes tested and 19% of the non-coding genes tested. In total, we identified eQTL effects for 23,060 different genes (false discovery rate (FDR) ≤ 0.05). These eQTLs replicated well in earlier microarray-based datasets from blood samples [22] and an RNA-seq based from lymphoblastoid cell lines (LCL) [23] (Supplementary note chapter 3), but also substantially extended the list of genes that are known to be under genetic regulation (replication results in Supplementary note chapter 3, Supplementary table 2). In addition to detected gene-level eQTLs, we identified 21,888 different genes with one or more exon-level QTL effect and 9,777 and 2,322 genes where SNPs affected the inclusion rate of exons and the usage of polyA sites, respectively (Supplementary table 3). All eQTLs can be found using our eQTL browser at http://genenetwork.nl/biosqtlbrowser. Multiple unlinked SNPs in the same locus may independently influence expression or mRNA processing of the same gene [24]. This was observed for more than half of the *cis*-regulated genes (Figure 1a, Supplementary figure 1).

The gene level *cis*-eQTL SNPs were strongly enriched for DNase I footprints, various histone marks and binding sites of multiple transcription factors [25], (Supplementary table 4, Supplementary note chapter 4) suggesting likely detection of causal regulatory variants. Moreover, top eQTL SNPs were significantly enriched for general and blood-cell type-specific enhancers (taken from [26]), but not for non-blood tissue-specific enhancers (Supplementary table 5). Evidence for the functionality of exon ratio and polyA ratio QTLs in mRNA splicing and polyadenylation is also presented in Supplementary note chapter 4.

One third (2,064 or 32.7%) of previously established genetic risk factors for disease or complex traits (derived from the NHGRI GWAS catalog and a set of reported ImmunoChip associations, $P \leq 5 \times 10^{-8}$, Supplementary table 6) were in strong linkage disequilibrium (LD $r^2 \geq 0.8$) with a top eQTL SNP (Supplementary table 7, Figure 1b). As expected, eQTL effects were

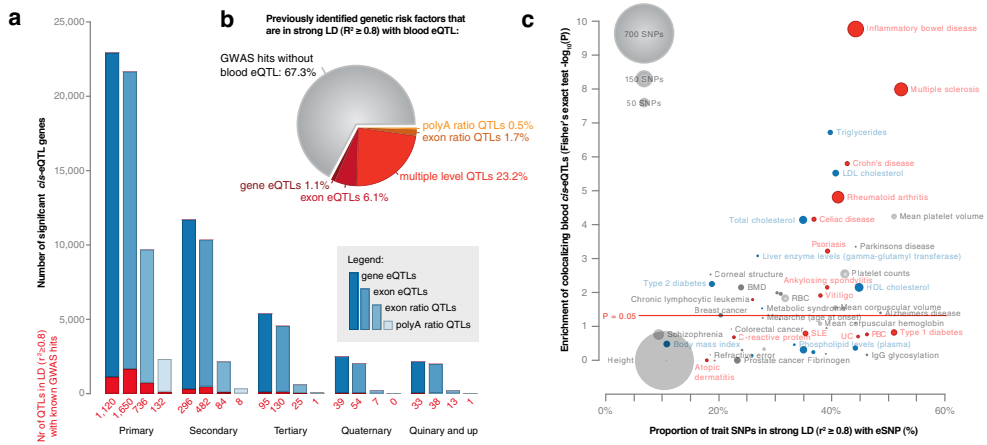



***Figure 1: Over 20,000 genes are regulated by cis-eQTLs overlapping with 33% of the entries in the GWAS catalog.*** *(a) Numbers of cis-regulated genes having one, two, three, four, and ≥ five independent eQTL effects (FDR ≤ 0.05). The number of eQTLs overlapping with SNPs in the GWAS catalog ($r^2 \geq 0.8$) are indicated in red. (b) Distribution of GWAS catalog variants over the different types of eQTLs. Of the GWAS catalog SNPs, 8% affected exon-level QTLs or polyA ratio QTLs, but not overall gene expression levels. (c) Auto-immune disorders and traits related to blood showed a higher co-localization with eQTLs compared to anthropometric traits and diseases without an immune or hematological component. The x-axis shows the percentage of GWAS hits co-localizing with eQTLs and the y-axis shows the enrichment of overlap (calculated at the gene level), using genes associated with height as a conservative background.*

predominantly found for SNPs associated with hematological, lipid or immune-related traits. We observed a highly significant enrichment of co-localization of eQTL and GWAS SNPs (LD $r^2 \geq 0.8$) for many immune disorders compared to the 10% overlap found for height, which we considered as a conservative background (Figure 1c, Supplementary note chapter 5). This indicates that our blood *cis*-eQTLs are highly informative for diseases such as inflammatory bowel disease, multiple sclerosis and rheumatoid arthritis.

Effect sizes for eQTLs often depend on the cell type or tissue under investigation [8–11,27,28], and may be modified by external and environmental factors [14–17,29,30]. We developed a hypothesis-free strategy to identify which of the observed eQTLs were dependent on intrinsic or extrinsic factors (context-dependent eQTLs) (Figure 2a,b, Supplementary figure 2, online methods). Instead of using known factors, such as the percentage of neutrophils in blood in a gene by environment interaction model [13], we used the expression levels of other genes as interaction factors. We call these genes 'proxy genes', as they may reflect the abundance of a cell type or the activity of signaling pathways.

We identified 10 modules of in total 1,842 eQTLs independently affected by 10 largely uncorrelated proxy genes (Figure 2c, Supplementary table 8). eQTLs with context-dependent effects can be obtained from our BIOS eQTL browser. An example is shown in Figure 2b, where we found an eQTL effect of SNP rs1981760 (a SNP associated with leprosy susceptibility)

on *NOD2* expression. The first top proxy gene, *STX3*, had a significant interaction with this eQTL. Samples with very low expression of *STX3* showed only a very weak eQTL effect on *NOD2*, whereas samples with very high *STX3* expression showed a stronger eQTL effect size. Further analysis demonstrated that *STX3* expression was strongly correlated (Pearson r = 0.74) with the percentage of neutrophils in the blood, indicating that *STX3* is a proxy for neutrophil levels in blood.

It can be challenging to understand what the proxy genes represent. We first assessed whether they are markers for specific cell types, and correlated them with blood cell counts measured in our samples (neutrophils, lymphocytes, eosinophils, basophils and monocytes, and baseline gene expression levels in purified blood cells from the BLUEPRINT consortium [31] (Figure 2c, Supplementary figure 3). Eight out of ten proxy genes likely represent specific cell type levels in blood (Supplementary note chapter 6). Analysis of eQTL gene expression in BLUEPRINT (Supplementary figure 4a) and eQTL interactions with measured blood cell counts confirmed the cell type-dependent effects of neutrophils and eosinophils (Supplementary figure 5, Tables S9, 10), but our unbiased analysis also identified effects for cell types for which actual cell counts were not available (erythroblasts, CD4+ T-cells and NK cells/CD8+ T-cells). Replication of our cell type-dependent eQTLs in eQTLs datasets with purified cell types supported these observations (Supplementary table 11, Supplementary figure 4b-c).

Cell type-specific eQTL genes were enriched in cell type-specific signaling pathways (Figure 2c, Supplementary table 12). For example, genes for which the *cis*-eQTL effects were particularly strong in erythroblasts (represented by proxy gene *TSPAN5*) are enriched for erythrocyte-specific functions. They were also enriched in binding sites for transcription factors involved in erythrocyte development based on ENCODE ChIP-seq data (GATA1, TAL1, GATA2 and MafK, each with enrichment p-values $\leq 10^{-5}$) [32–34]. A well-established *cis*-eQTL for *SMIM1*, an erythrocyte-specific gene encoding a protein that determines the Vel blood group[4], was contained in the set of eQTLs affected by *TSPAN5* expression. For eQTLs affected by other proxy genes, we also identified specific transcription factors with established functions in the corresponding cell types (Supplementary table 13).

In Supplementary figure 6 and Supplementary note chapter 7, we show examples of how eQTLs can be used to gain insights into five autoimmune disorders. Clustering of the eQTL genes based on co-expression revealed sets of genes hinting at specific cell types and biological functions. For inflammatory bowel disease (IBD), for instance, the clustering revealed a T-cell cluster and a neutrophil cluster. Adding the cell type-dependent eQTLs further corroborated the cell type annotations of the clusters. In total we found 138 context-dependent eQTLs for GWAS variants (Supplementary table 14).

The identified interaction modules are not restricted to cell type-specific effects. One of the proxy genes, *SP140*, is not a proxy for cell type, but for type I interferon response, as demonstrated by pathway enrichment of genes that correlated positively with *SP140* expression levels (Supplementary note chapter 8). Genes that correlated negatively with *SP140* are involved in anti-bacterial response and inflammation (Figure 3a). Likewise, the affected eQTL genes can be divided into two groups: those that were positively and those negatively correlated with *SP140* expression (Figure 3b). Gene annotations from the interferome database

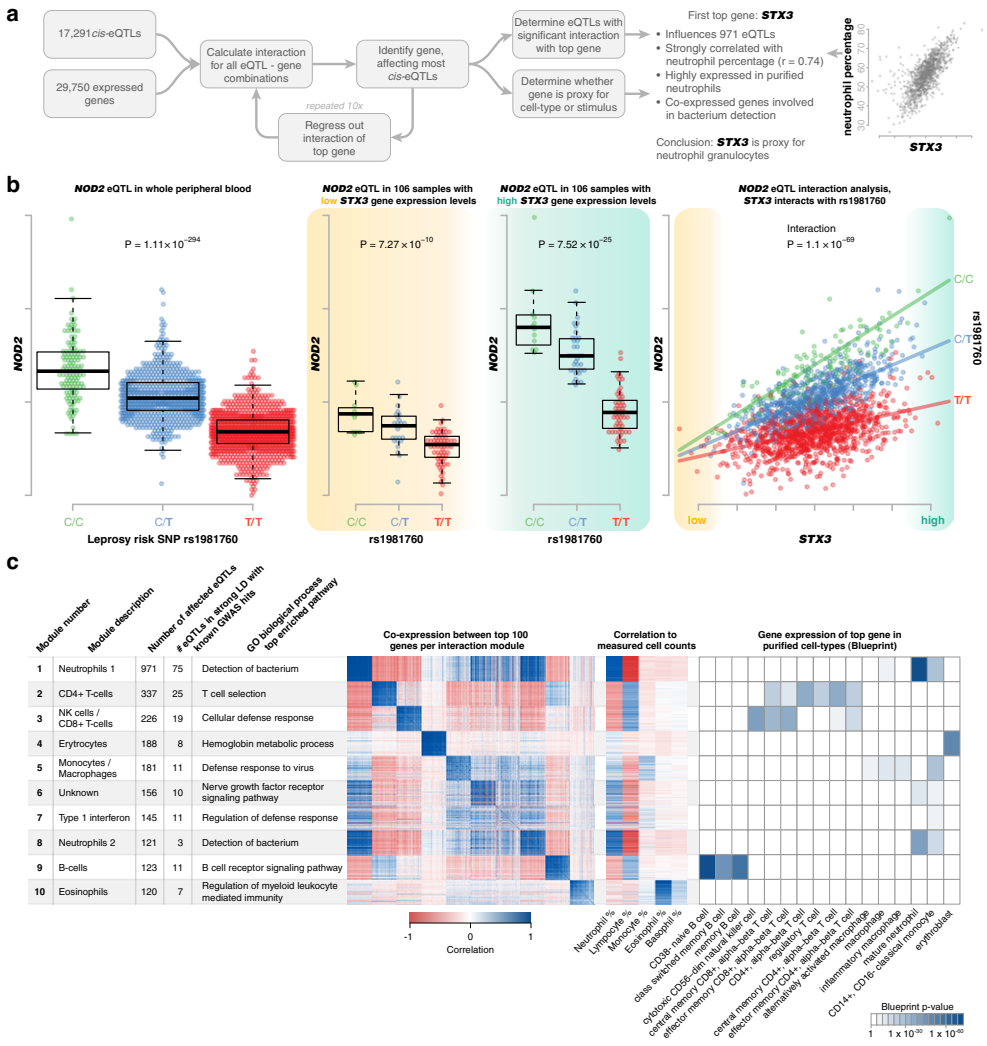



***Figure 2: Identification of the strongest modifiers of eQTL effects.*** *(a) Overview of method used to annotate the eQTLs as context-dependent: each of the highly expressed genes (at least one read in all samples) was tested for its ability to modify each of the 17,291 eQTL affecting highly expressed genes. For each of these proxy genes, we determined the overall strength of the interaction effects with all eQTLs. We selected the strongest proxy gene and regressed its effects from the data. We did this for 10 iterations allowing the identification of 10 independent proxies that affected the strength of eQTLs. (b) An example of a context-dependent eQTL effect is rs1981760, a strong eQTL for the NOD2 gene. This SNP was in strong LD ($r^2$ = 0.99) with rs9302752, a variant associated with leprosy susceptibility. The leprosy risk allele resulted in decreased expression of NOD2. In samples with low STX3A expression, only a weak eQTL effect was observed. In samples with high STX3A expression, a strong eQTL effect was observed. In accordance with this, using this STX3A gene as a covariate in an interaction model revealed a very strong interaction effect. STX3A is the first proxy gene …*

[35] confirmed that the up-regulated eQTL genes are indicative of type I, but not of type II, interferon response (Figure 3c). In support of the modifying effects of viral cues on this set of eQTLs, eQTL genes that were recently reported as rhinovirus-response QTLs[15] typically demonstrated higher *SP140* interaction effects than other eQTL genes (Wilcoxon p-value = 0.02).

Each of the aforementioned ten proxy genes demonstrated effects on many (>120) eQTLs. However, some other factors may also exist that affect more limited numbers of eQTLs. To identify these, we first corrected the expression data for the 10 proxy genes and their eQTL interaction effects and then ascertained for each gene-level eQTL whether the eQTL effect size was significantly dependent on the expression of any other gene. This resulted in the identification of an additional set of 901 context-dependent eQTLs (FDR ≤ 0.05) (Supplementary table 15). Of these eQTL interactions, 113 could also be detected in Geuvadis LCLs (FDR ≤ 0.05, 94% with the same interaction direction) (Supplementary table 16). These LCLs are homogeneous cell populations, so any interaction effect that replicates is unlikely due to cell type-specific eQTL effects, but rather reflects an external stimulation or activation of core biological processes. A few of these context-dependent eQTLs enable inference of regulatory networks.



An example is the *cis*-eQTL (rs968567) effect on the lipid biosynthesis-related gene *FADS2* that is modified by the expression of the sterol regulatory element binding transcription factor gene *SREBF2* (p-value = $4.1 \times 10^{-14}$, p-value in Geuvadis = 0.002) (Figure 4a,b). The eQTL SNP is in close proximity to an SREBF2 binding site (ENCODE ChIP-seq data, Figure 4c) and it is therefore likely that the SNP modifies the affinity of the *FADS2* promoter for SREBPF2. *SREBF2* showed a significant negative correlation with HDL cholesterol (Pearson r = -0.18, p-value = $5.1 \times 10^{-6}$) and a positive correlation with lymphocyte percentage (Pearson r = 0.19, p-value = $1.6 \times 10^{-6}$). A partial correlation analyses revealed that the correlation with HDL cholesterol is independent of the correlation to the lymphocyte percentage (Pearson r on residuals of HDL after correcting for lymphocytes: -0.17, p-value = $2.7 \times 10^{-5}$), showing that the correlation to HDL is not driven by cell type composition. We propose a model where extracellular (HDL) cholesterol levels modify SREBF2 binding to the *FADS2* promoter, which, in turn has effects on the expression of *FADS2* and lipid desaturase activity in the cell. This SNP

*… we identified and correlated strongly with neutrophil percentage (Pearson r = 0.72). Gene enrichment analysis of STX3A and other genes showing similar interaction patterns, revealed involvement in antibacterial response. Furthermore, individuals carrying the leprosy risk allele had significantly weaker NOD2 up-regulation in neutrophils compared to non-carriers. This is in line with earlier reports showing this eQTL to be stronger in FACS-sorted neutrophils compared to monocytes [27]. (c) We annotated each of our 10 proxy genes using the top 100 proxy genes per module with similar effects, and showed that, as expected, these top 100 genes were strongly correlated per module. This top 100 was used for gene function enrichments (for full results see Supplementary table 12) and was correlated to known cell proportions. We used BLUEPRINT expression data for sorted populations of blood cells to validate cell type-specific expression in each module. N=2,116 individuals were used in the analysis.*

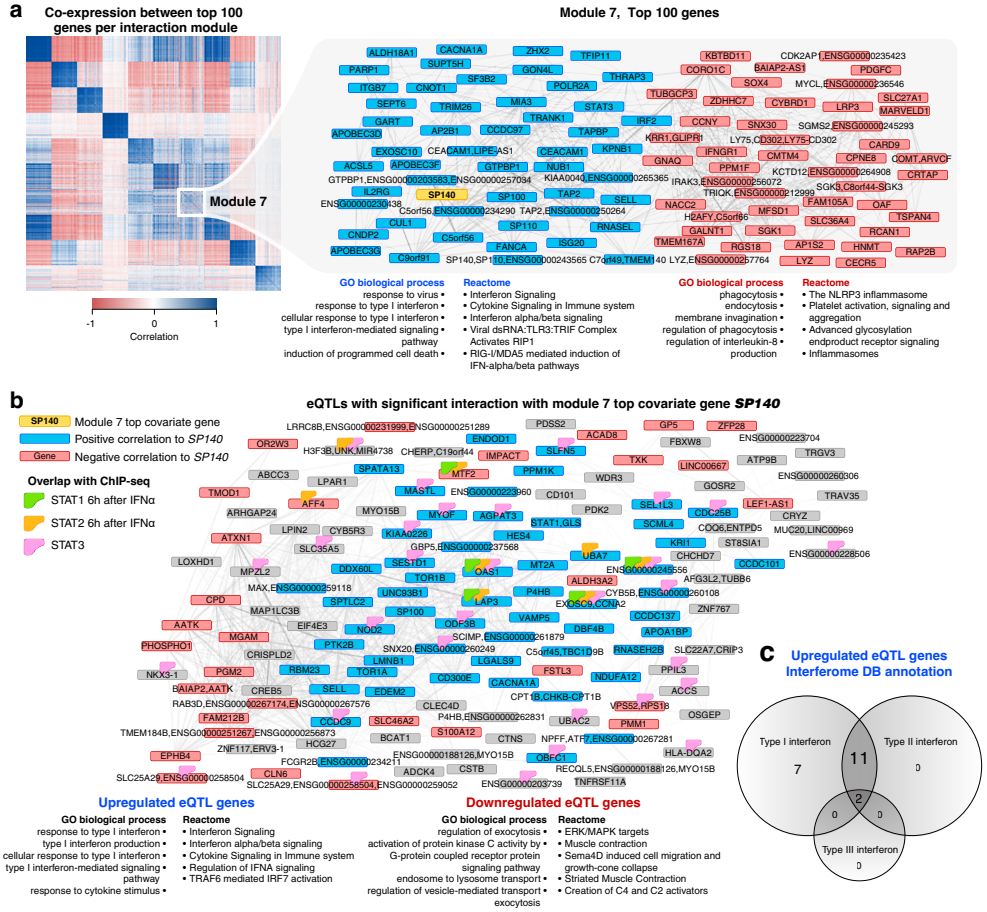



***Figure 3: eQTLs modified by type I interferon signaling.*** *(a) Expression-level based clustering of genes positively correlated with proxy gene SP140 (blue) and those negatively correlated with SP140 (red). Enrichment analysis of these two clusters showed distinct biology: the up-regulated genes were enriched for type I interferon response and response to viruses whereas the down-regulated genes indicated an anti-bacterial inflammatory response. Type I interferon signaling is activated in a viral response and type II interferon signaling is activated upon bacterial response [36]. The positively correlated genes were enriched for up-regulated genes upon rhinovirus stimulation [15] (Fisher exact p-value 1.14 x $10^{-9}$), in line with their involvement in the type I interferon response. In contrast, the negatively correlated genes were enriched for genes up-regulated upon LPS stimulation (Fisher exact p-value 0.02) and interferon-gamma stimulation (Fisher exact p-value 8.72 x $10^{-4}$) [14], supporting the anti-bacterial function of these negatively correlated genes. (b) The eQTLs affected by SP140 expression can also be divided into those genes positively or negatively correlated with SP140 expression. The significantly positively correlated eQTL genes were also enriched for type I interferon response, whereas the negatively correlated eQTL genes did not show strong enrichment for biological functions. Genes bound by STAT transcription factors, as identified by in ENCODE ChIP-seq data from LCL, are labeled. Both type I and type II interferon …*

also increases risk for rheumatoid arthritis, blood metabolite levels and lipid levels and using our method we now implicate altered binding of SREBF2 as a possible functional mechanism behind these associations.

Another example is a *cis*-eQTL effect on the *MYBL2* gene, encoding a known transcription factor that controls cell division and a tumor suppressor [39] (Figure 5a-c). According to ENCODE ChIP-seq data, the top eQTL SNP, rs285205, was located in an EBF1 binding site (Figure 5d). EBF1 is a known player in B-cell differentiation and proliferation. Although *FCRLA* expression was the strongest modifier of the eQTL, *EBF1* was highly correlated with *FCRLA* expression and revealed a significant interaction effect on the *MYBL2* eQTL (p-value = 1.8 x $10^{-14}$) (Figure 5c). The eQTL SNP, therefore, likely affects the binding affinity of EBF1.

In conclusion, we greatly expanded the catalog of SNPs that have a known regulatory function. To gain a better understanding of the biology behind these regulatory variants, we assessed the context-dependency of the eQTLs and determined 2,743 to be context-dependent. With future increases in sample size, we expect it will become possible to identify more unanticipated intrinsic factors and external stimuli that modify the downstream effects of genetic risk factors. As such, our approach complements perturbation experiments to gain better insight into regulatory networks and their stimuli and it can easily be applied to other tissues. A caveat of our hypothesis-free approach is that it is not always straightforward to understand which internal or external cues the proxy genes represent. Integration with other expression or transcription factor binding data, as we have done here, is, therefore, instrumental for the interpretation of context-dependent eQTLs.



*… signaling result in binding of heterodimers of the STAT1 transcription factor. Unique to type I interferon is that STAT1 forms a complex with STAT2 and IRF9, resulting in the activation of viral response genes. STAT3 activation is also unique to the type I response, resulting in the down-regulation of inflammatory pathways [37]. The eQTLs were enriched for STAT1 (p-value = 4.82 x $10^{-4}$), STAT2 (3.12 x $10^{-4}$) and STAT3 (4.72 x $10^{-5}$) binding sites (based on ENCODE ChIP-seq experiments) (Supplementary table 13). Motif enrichment analysis [38] on the 25 bp flanking regions of the eQTL SNPs confirmed the enrichment of STAT-binding motifs (Wilcoxon rank-sum test, p-value = 9.61 x $10^{-5}$). (c) Interferome DB annotation of the up-regulated eQTL genes confirmed their role in type I (and not type II or III) interferon signaling. N=2,116 individuals were used in all eQTL analyses.*

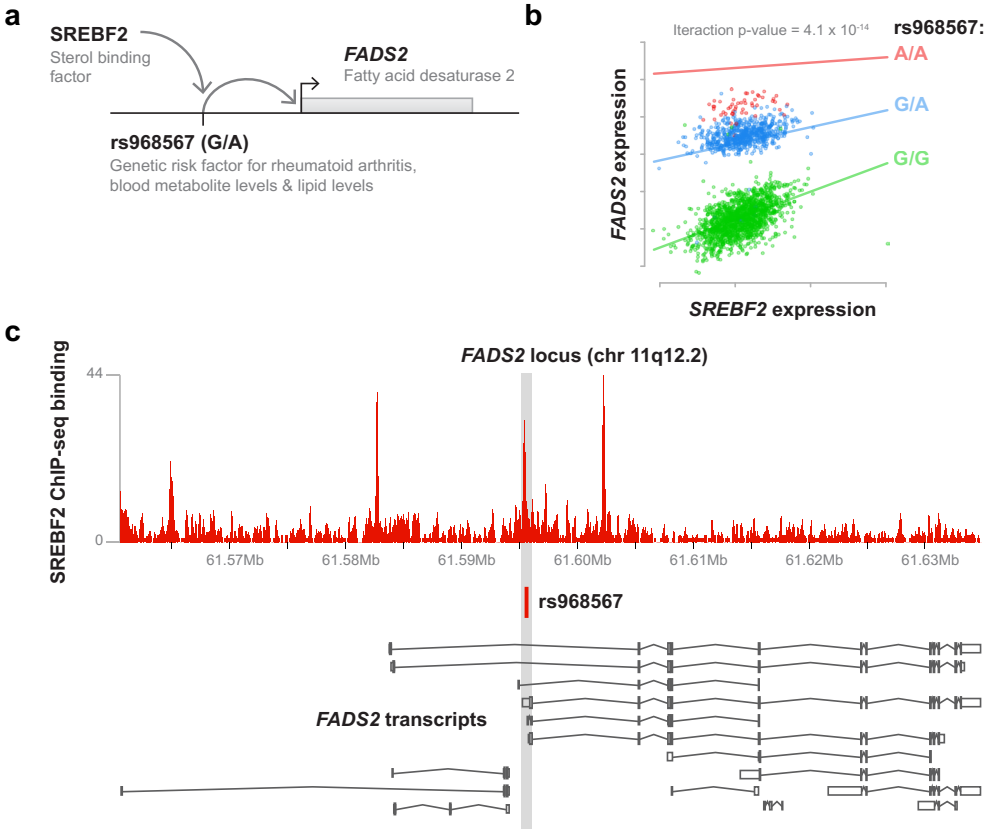



***Figure 4: FADS2 eQTL modulated by SREBF2 expression.*** *(a) The eQTL SNP rs968567 is located in a SREBF2 binding site in the FADS2 promoter. (b) The eQTL was modulated by SREBF2 expression and stronger in samples with low SREBF2 expression. The nominal p-value for the interaction effect is indicated. N=2,116 individuals were used in the eQTL analysis. (c) rs968567 is located in an ENCODE ChIP-seq peak of SREBF2 binding.*

# Methods

## Cohort description

The four cohorts used in our BIOS (Biobank-based Integrative Omics Study) study are briefly described below. The age range of the individuals differed for the different biobanks (Supplementary figure 7). The number of samples per cohort used in our study can be found in Supplementary table 1.

### *CODAM*

The Cohort on Diabetes and Atherosclerosis Maastricht [18] (CODAM) consists of a selection of 547 subjects from a larger population-based cohort [42]. Inclusion of subjects into CODAM was based on a moderately increased risk of developing cardiometabolic diseases such as type 2 diabetes and/or cardiovascular disease. Subjects were included if they were of Caucasian

descent, over 40 years of age and additionally met at least one of the following criteria: increased BMI (>25), a positive family history of type 2 diabetes, a history of gestational diabetes and/or glycosuria, or use of anti-hypertensive medication.

### *LLD*

The LifeLines-DEEP (LLD) cohort [19] is a sub-cohort of the LifeLines cohort [43] with additional molecular data on 1,500 participants. LifeLines is a multi-disciplinary prospective population-based cohort study examining the health and health-related behaviors of 167,729 individuals living in the northern parts of the Netherlands using a unique three-generation design. It employs a broad range of investigative procedures assessing the biomedical, socio-demographic, behavioral, physical and psychological factors contributing to health and disease in the general population, with a special focus on multi-morbidity and complex genetics.



### *LLS*

The aim of the Leiden Longevity Study [20] (LLS) is to identify genetic factors influencing longevity and examine their interaction with the environment in order to develop interventions to increase health at older ages. To this end, long-lived siblings of European descent were recruited together with their offspring and their offspring's partners, on the condition that at least two long-lived siblings were alive at the time of ascertainment. For men the age criteria was 89 or older, for women age 91 or over. These criteria led to the ascertainment of 944 long-lived siblings from 421 families, together with 1,671 of their offspring and 744 partners.

### *RS*

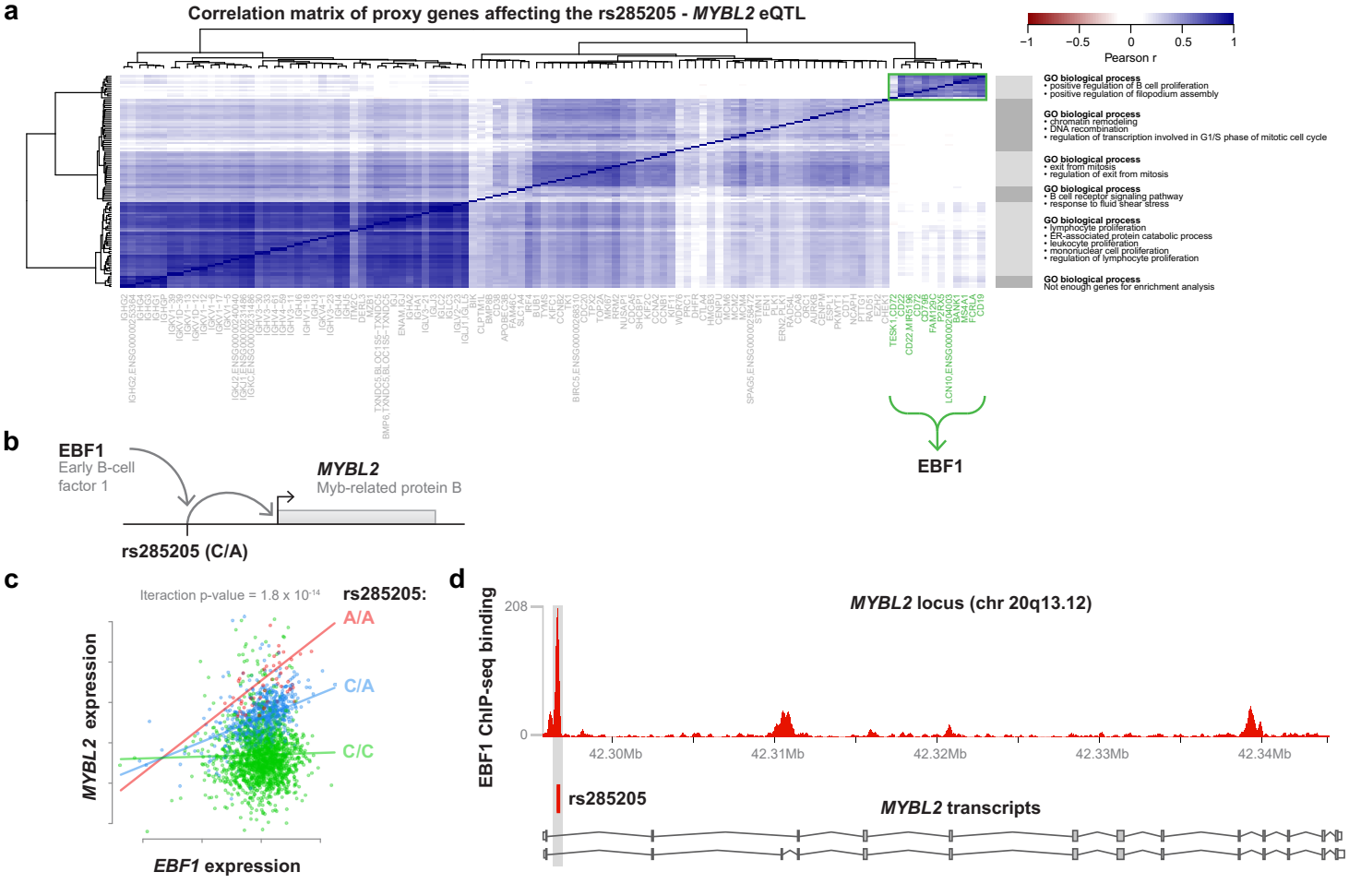
The Rotterdam Study [21] is a single-center, prospective population-based cohort study conducted in Rotterdam, the Netherlands. Subjects were included in different phases from the start of the study in 1998, with a total of 14,926 men and women aged 45 and over included as of late 2008. The main objective of the Rotterdam Study is to investigate the prevalence and incidence of and risk factors for chronic diseases to contribute to a better prevention and treatment of such diseases in the elderly.

### *Ethical approval*

The ethical approval for this study lies with the individual participating cohorts, CODAM, LLD, LLS and RS and can be found in references [21].

## RNA data preparation and sequencing

Total RNA from whole blood was deprived of globin using Ambions GLOBINclear kit and subsequently processed for sequencing using Illumina's Truseq version 2 library preparation kit. Paired-end sequencing of 2x50bp was performed using Illumina's Hiseq2000, pooling samples at 10 per lane, and aiming for >15M read pairs per sample. Finally, read sets per sample were generated using CASAVA, retaining only reads passing Illumina's Chastity Filter for further processing.


a
Correlation matrix of proxy genes affecting the rs285205 - MYBL2 eQTL
Pearson r
−1
−0.5
0
0.5
1
GO biological process
• positive regulation of B cell proliferation
• positive regulation of filopodium assembly
GO biological process
• chromatin remodeling
• DNA recombination
• regulation of transcription involved in G1/S phase of mitotic cell cycle
GO biological process
• exit from mitosis
• regulation of exit from mitosis
GO biological process
• B cell receptor signaling pathway
• response to fluid shear stress
GO biological process
• lymphocyte proliferation
• ER-associated protein catabolic process
• leukocyte proliferation
• mononuclear cell proliferation
• regulation of lymphocyte proliferation
GO biological process
Not enough genes for enrichment analysis
EBF1
b
EBF1
Early B-cell
factor 1
MYBL2
Myb-related protein B
rs285205 (C/A)
c
Iteraction p-value = 1.8 x 10-14
rs285205:
A/A
C/A
C/C
MYBL2 expression
EBF1 expression
d
MYBL2 locus (chr 20q13.12)
EBF1 ChIP-seq binding
208
0
42.30Mb
42.31Mb
42.32Mb
42.33Mb
42.34Mb
rs285205
MYBL2 transcripts

### *Preprocessing*

The quality of the raw reads was checked using FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/]. The adaptors identified by FastQC (v0.10.1) were clipped using cutadapt (v1.1) applying default settings (min overlap 3, min length). Sickle (v1.200) [https://github.com/najoshi/sickle] was used to trim low quality ends of the reads (min length 25, min quality 20).

### *Alignment*

Read alignment was performed using STAR 2.3.0e [44]. To avoid reference mapping bias, all GoNL SNPs with MAF > 0.01 in the reference genome were masked with N [45]. Read pairs with at most 8 mismatches, mapping to at most 5 positions, were used.

### *Alignment statistics*

Mapping statistics from the BAM files were acquired through Samtools flagstat (v0.1.19-44428cd). The 5' and 3' coverage bias, duplication rate and insert sizes were assessed using Picard tools (v1.86).

### *Expression quantification*

We estimated expression on the gene, exon, exon ratio and polyA ratio levels using Ensembl v.71 annotation (which corresponds to GENCODE v.16).

Overlapping exons (on either of the two strands) were merged into meta-exons and expression was quantified for the whole meta-exon. For that, custom scripts were developed which use coverage per base from coverageBed and intersectBed from the Bedtools suite (v2.17.0) [46] and R (v2.15.1). This resulted in base counts per exon or meta-exon.

***Figure 5: MYBL2 eQTL is modulated by B-cell proliferation gene EBF1.*** *(a) Heatmap of the co-expression of 109 proxy genes that modulated the eQTL effect on MYBL2 expression. Gene function enrichment analyses on the genes in these clusters revealed that all were related to proliferation or cell cycle checkpoints. Interestingly, only one cluster increased the magnitude of the MYBL2 eQTL effect in contrast to the other clusters, which all repressed this eQTL. This eQTL activating cluster was strongly enriched for "positive regulation of B-cell proliferation" (p-value = 1 x $10^{-7}$), and the strongest proxy gene in this cluster was FCRLA, which is known to be highly expressed in proliferating B-cells residing in the germinal center of the lymph nodes (centroblasts) [40,41]. (b) Regulation of MYBL2 by the different cell-cycle clusters is likely modulated via EBF1 and rs285205. In our analysis we had initially only considered genes that where expressed in each of our individuals (see methods), and therefore had not studied low-abundant transcription factor genes. When also including these genes, we observed that this cluster of genes is strongly co-expressed with EBF1, a gene encoding a transcription factor that binds at the site of the eQTL SNP, suggesting that EBF1 might drive the eQTL interaction effect for MYBL2. EBF1 is a known player in B-cell differentiation and proliferation and is positively correlated with both MYBL2 (r = 0.11, p-value = 6.99 x $10^{-7}$) and FCRLA expression (r = 0.8, p-value ≤ 2.2 x $10^{-16}$). (c) Interaction plot showing that EBF1 expression modified the eQTL effect of rs285205. The nominal p-value is indicated. (d) ENCODE ChIP-seq data in LCLs showed strong binding of EBF1 at rs285205. N=2,116 individuals were used in all eQTL analyses.*

Gene expression, as base count per gene, was calculated as the sum of expression values of all exons of each gene (excluding meta-exons). Overlapping gene parts are counted separately from the unique gene parts throughout this manuscript.

Expression of exons relative to their gene (exon ratio) was calculated by dividing the exon base counts by the summed base counts for all exons of the same gene. Meta-exons overlapping with multiple genes were discarded.

Overlapping 3' UTRs for the same gene, as annotated in Ensembl, were merged by gene. A collection of polyadenylation sites was retrieved from PolyA_DB and the annotated 3' ends of transcripts from Ensembl. These polyadenylation sites were used to split the merged 3' UTRs into bins. To avoid small bins that tend to give noisy ratios, we applied some filtering on the polyA sites. PolyA sites located no more than 10bp from the start or from the end of the 3'UTR were discarded. Additionally, sites that are no more than 10 bp apart were merged (if their number was even, the first site downstream was used). For all genes with at least two bins (corresponding to at least two potential sites of polyadenylation), we calculated the ratio of base counts between every two neighboring bins (polyA ratios).

## Genotype data

### *Data generation*

Genotype data was generated for each cohort individually. Details on the methods used can be found in the individual papers (CODAM: van Dam et al. [42]; LLD: Tigchelaar et al. [47]; LLS: Deelen et al. [48]; RS: Hofman et al. [21]).

### *Imputation and QC*

The genotype data were harmonized to the Genome of the Netherlands [49] (GoNL) using Genotype Hamonizer [50] and subsequently imputed per cohort with Impute2 [51] using the GoNL reference panel [52] (v5). Quality control was also performed per cohort. We removed SNPs with an imputation info-score below 0.5, a HWE *P*-value smaller than $10^{-4}$, a call rate below 95% or a minor allele frequency smaller than 0.05. In total, 9,333,740 SNPs passed QC in at least one dataset.

## Quality Control

To identify low quality samples, we applied several quality metrics and used a combination of them to decide whether to exclude a sample from further analyses.

### *Read counts*

For each sample, the total number of mapped reads was used as a quality measure. Those samples for which these counts were less than 70% were flagged and excluded from the analysis.

### *Exon and gene expression correlation*

For each pair of samples, Spearman correlation of their expression was calculated on gene and exon levels. From these values, the median Spearman correlation for each sample was calculated (D-statistic). Samples with D-statistics lower than 0.85 were excluded from the analysis.

### *Genotype concordance*

As an extra quality control step we compared imputed genotypes to RNA-seq derived genotypes. Concordance is expected to be low in cases of bad quality RNA-seq or imputed genotype data or in cases of sample mix-ups.

RNA-seq genotypes were called using samtools mpileup [53] (with the following parameters: -A -B -Q 0 -s -d10000000, calling only GoNL SNPs with MAF > 0.01) and SNVMix2 [54]. Only the genotypes with posterior probabilities higher than 0.8 were included. We determined genotype concordance per sample as genotype correlation of high-confidence SNPs (the SNPs which had a mean genotype correlation across all samples that was not less than 0.9). Outlier samples, for which the genotype concordance was less than 0.9, were flagged and excluded from the analysis.

### *Heterozygosity rate*

A maximum heterozygosity rate of 0.52 was used to exclude contaminated RNA-seq samples. This was calculated using the same high quality genotypes used for the genotype concordance calculations.

### *Mix-up mapping*

Previously we showed that sample mix-ups occur frequently in genetical genomics datasets, introducing noise in subsequent analyses[55]. We checked the data for mix-ups using this published method and flagged possible mixed samples.

## QTL mapping

We used our previously described pipeline [22] to perform eQTL mapping. We mapped QTLs using a Spearman's rank correlation on imputed genotype dosages in each cohort and then ran a meta-analysis combining the results by weighted z-score method. To control the false discovery rate (FDR) at 0.05, we created a null distribution by permuting sample labels of the expression data, repeating this process 10 times.

### *Expression data normalization*

Expression data on the gene and exon level were first normalized using Trimmed Mean of M-values (TMM) [56]. Expression values were then $\log_2$ transformed, probe and sample means centered to zero and their standard deviation was scaled to one. To correct for batch effects, principal component analysis (PCA) was run on the sample Spearman correlation matrix and the first 25 PCs were removed [22]. We observed that removing these PCs resulted in the detection of the highest number of eQTLs. To ascertain that none of these 25 PCs are under genetic control, we ran separate QTL mapping on each principal component and ensured that there were no SNPs associated with them.

Exon ratio and polyA ratio expression data were not normalized, as ratios are not dependent on the library size and we used non-parametric statistics.

### *cis-QTL mapping*

For running cis-QTL mapping we tested genes (or exons, exon ratios and polyA ratios) and SNPs located within 250kb from the gene (or exon) center. Only SNPs with minor allele frequency (MAF) ≥ 0.05, call rate (CR) ≥ 0.95 and Hardy-Weinberg equilibrium p-value ≥ 0.001

were included. We identified independent QTL effects by stepwise regression: we found secondary QTLs by regressing out the primary QTLs and we identified tertiary QTLs by regressing primary and secondary QTLs. This procedure was repeated until no more independent effects were found. We acknowledge that it might be possible that some of the identified independent effects might actually tag untyped variants.

## Set of background SNP for functional enrichment analyses.

For assessment of functional enrichment of eSNPs on each QTL level, we created a background list of SNPs which we compared to the real set. For each eQTL SNP, we selected the variants within a 50,000 bp window, with a MAF differing not more than 0.05 point from the eQTL SNP, and a linkage disequilibrium $r^2 \leq 0.5$. From the variants that match these criteria, we selected the variant that is physically closest to the eQTL SNP as a background SNP.

## Replication of cis-eQTLs

The first replication dataset was Geuvadis RNA-seq data of lymphoblastoid cell cines (LCLs) [23]. For replication, we took raw RNA-seq reads of 373 European samples and processed them using the same alignment and quantification pipeline as we used on the BIOS data. For eQTL mapping, we regressed out the first 20 PCs from the expression data (due to the smaller sample size of Geuvadis dataset). To replicate BIOS eQTLs in Geuvadis, we took all significant eQTLs (top SNP per gene) from BIOS and ran eQTL mapping in Geuvadis, testing only these eQTLs. We then checked how many eQTLs out of all those tested were replicated and for how many of the replicated eQTLs the allelic direction was opposite. We did the same in the other direction, testing how many of the Geuvadis eQTLs were replicated in the BIOS data.

The second dataset was a meta-analysis of 5,311 microarray peripheral blood samples published by Westra et al. [22]. As raw data were not available for these data, we used all significant eQTLs (FDR < 0.05) identified in the meta-analysis, mapped the microarray probes to gene and exons using Ensembl v.71 gene annotation, and then tested these SNP-gene and SNP-exon combinations in the BIOS data.

## GWAS annotation

For annotating the eQTLs with known disease/trait associations, we used a set of 6,321 SNPs derived from the NHGRI GWAS catalog and a set of reported ImmunoChip associations, each with reported p-value $\leq 5 \times 10^{-8}$ (Supplementary table 6).

## Interaction analysis

For an overview of the method used for the interaction analysis see Supplementary figure 2. The interaction analysis was performed using the following linear model:

$$Y \approx I + \beta_1 G + \beta_2 P + \beta_3 P \cdot G$$

where $Y$ is the eQTL gene expression, $G$ is the eQTL SNP genotype, $P$ is the proxy gene, $P \cdot G$ is the interaction term between the proxy gene and the genotype, $I$ is the intercept, and $\beta_1$, $\beta_2$, and $\beta_3$ are regression coefficients.

As a linear model is parametric, and thus more sensitive to outliers and non-normal distributions than our non-parametric eQTL model, we performed stricter quality control. We found that several metrics introduced outliers in our data that confounded the linear regression analyses. These were percentage of coding bases, median 3' bias, percentage of uniquely mapped reads and percentage of mRNA bases (Supplementary figure 8). Based on these metrics we removed 75 samples and used the remaining 2,041 samples in the interaction analyses. We confined the interaction analysis to genes with at least one mapped read in all samples; this criterion was used for both the proxy genes and the eQTL genes. As a result, we tested 29,750 genes as potential proxies and 17,291 eQTL effects.

The normalization for the expression levels of the eQTL genes requires different normalization than the expression of the proxy genes. The gene expression data of the eQTL genes was corrected using covariates for the source biobank, the first 25 PCs, gender, median 3' bias, median 5' bias, GC content and the percentage of intronic bases. In order to detect biologically meaningful interaction effects, we also regressed out the interaction effects for gender, median 3' bias, median 5' bias, GC content and the percentage of intronic bases. The expression data used in the interaction term was processed in a similar manner, with the exception that we did not correct for the principal components, as this would have removed correlations to cell types, and we did not correct for interactions with the technical covariates.



We excluded interactions where the eQTL SNP showed a significant eQTL effect on the tested proxy gene, as we wanted to exclude the cases in which the gene giving the interaction effect was in the same locus as the tested eQTL gene.

We then performed an iterative interaction analysis by regressing the top covariate in a stepwise manner. After the first round of interaction analysis, we identified the covariate having the highest chi2sum over all interaction z-scores (I.E. the chi2sum is per covariate the sum of the squared interaction z-score of all eQTLs). We regressed out this covariate from covariate and gene expression data and repeated the interaction analysis. This procedure was repeated 10 times. For each top covariate, we identified a set of covariates (module) with a similar interaction pattern by taking the top 100 covariates having the highest chi2sum difference between the current interaction analysis step and the previous step (effectively identifying co-expressed genes). These covariates are mostly highly coexpressed with the top covariate in the module (Figure 2c).

To determine the significance level of interactions, we permuted genotype sample labels and ran the interaction analysis. This enabled us to determine which eQTLs significantly interact with the top covariate of the module with a FDR ≤ 0.05.

We ran the interaction analysis on exon and exon ratio levels in a similar manner as for the gene level.

The implementation and manual of our method can be found here: https://github.com/molgenis/systemsgenetics/wiki/Discovery-of-hidden-confounders-of-QTLs

### *Interaction module functions*

To find the prevalent cell type for each module, we used several sources of information. Some of the BIOS biobanks had cells counts available, making it possible to correlate the top 100 covariates of each module with cell type percentages.

As an additional source of evidence, we used expression profiles for isolated populations of 17 of the major cell types in blood generated by the BLUEPRINT consortium [31].

To determine the putative function of each module, we performed pathway enrichment analysis using GeneNetwork [57,58] on the top 100 covariates in the module and on all eQTL genes having a significant interaction with the top covariate of the module.

To gain more insight into the function of the modules we identified, we overlapped the interaction results with those of several papers which studied stimulated cells and response QTLs (reQTLs): a study of PBMCs infected with rhinovirus [15] and a study of monocytes infected by LPS (at two time points: after 2 and 24 hours) and IFN [14]. To investigate whether our interaction modules represent anti-viral or anti-bacterial response, we checked for enrichment of differentially expressed genes reported for each stimulation (with $-1 < \log FC < 1$) within the top 100 covariates from each interaction module by performing a one-tailed Fisher's exact test to determine the significance. We also checked whether the reported reQTLs showed significantly stronger interaction with the top covariate of each module by performing a Wilcoxon rank-sum test on interaction z-scores.

We also checked whether we see an enrichment of binding of particular transcription factors (TFs) using ChIP-seq data from ENCODE [59]. First, we determined which TFs overlapped with the eQTL SNP or a variant in very strong LD ($r^2 \geq 0.99$). Then, using a Fisher exact test, we determined if we found any enrichment in overlap between the genes assigned to a module and the genes not significantly assigned to this module.

### *Using interaction modules to better understand disease mechanisms*

We extracted the genes regulated by any type of top QTL variant in strong LD ($r^2 \geq 0.8$) with the top GWAS hits. Co-expression was calculated in our data for these genes and Cytoscape 3.2.1 [60] was used to create network plots. Assignment to specific clusters was performed using the R implementation of Affinity Propagation [61,62]. Cell-type-specific expression levels were based on the RNA-seq generated by the BLUEPRINT consortium[31] and plotted using gplots. We performed gene function enrichment analysis using GeneNetwork [57].

### *Cell-type-specific eQTL mapping*

The cell-type-specific eQTL were identified using the same method we used for the gene-based interaction analyses. However, here we used the cell-type percentages instead of the expression of other genes. As not all cohorts measured cell counts, we estimated them for these cohorts. RNA-seq data and cell count measurements in 628 samples from the LL and 650 samples from the LLS cohorts were used to build prediction models for cell counts using an in house predictor for neutrophils, lymphocytes, monocytes, eosinophils and basophils. We evaluated this method by using cross validations (Supplementary figure 9).These models were applied to RNA-seq data of 185 samples from the CODAM cohort and 14 samples from the LLS cohort for prediction of cell counts of the five cell types. In addition, the prediction

models were applied to estimate cell counts for neutrophils, eosinophils and basophils, using RNA-seq data from 652 samples from the RS cohort in which the cell count measurements of lymphocytes and monocytes were available.

### *BLUEPRINT tissue-specific expression data analysis*

BLUEPRINT data was downloaded from their ftp site (ftp://ftp.ebi.ac.uk/pub/databases/blueprint). All venous blood-, myeloid cell- and erythroblast-derived RNA-seq data was downloaded. Read counts were obtained according to the gene quantification performed by the Center for Genomic Regulation. Subsequently, TMM normalization [56] was performed. The averaged normalized log-counts per million per cell type were used for drawing the heatmaps. For each module, we extracted the corresponding genes based on their ENSEMBL gene identifier (for 'meta-exons' we used the first Ensembl id and three noncoding RNAs could not be extracted from the BLUEPRINT data). Furthermore, the R-package pheatmap (1.0.7) was used to generate the heatmaps.



## Data Availability Statement

Access to the raw RNA-seq data can be obtained via the EGA under accession: EGAD00001001623.

Genotype data is available via the respective biobanks.

| Biobank | Website | Contact e-mail address |
|---|---|---|
| LLS | http://www.leidenlangleven.nl/en/home | m.beekman@lumc.nl |
| LifeLines | https://lifelines.nl/lifelines-research/access-to-lifelines | LLscience@umcg.nl |
| CODAM | - | m.vangreevenbroek@maastrichtuniversity.nl |
| RS | http://www.epib.nl/research/ergo.htm | m.a.ikram@erasmusmc.nl |

eQTL results can be accessed via: http://genenetwork.nl/biosqtlbrowser/

## Acknowledgements

This work was performed 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). Samples were contributed by LifeLines (http://lifelines.nl/lifelines-research/general), the Leiden Longevity Study (http://www.healthy-ageing.nl; http://www.leidenlangleven.nl), the Rotterdam studies (http://www.erasmus-epidemiology.nl/rotterdamstudy) and the CODAM study (http://www.carimmaastricht.nl). We thank the participants of all aforementioned biobanks and acknowledge the contributions of the investigators to this study (Supplemental Acknowledgements). This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative and the Groningen Center for Information Technology (Strikwerda, W. Albers, R. Teeninga, H. Gankema and H. Wind) and Target storage (E. Valentyn and R. Williams). Target is supported by Samenwerkingsverband Noord Nederland, the European Fund for Regional Development,

the Dutch Ministry of Economic Affairs, Pieken in de Delta and the provinces of Groningen and Drenthe. This work is supported by a grant from the European Research Counsil (ERC Starting Grant agreement number 637640 ImmRisk) to Lude Franke. The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization 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. 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 GWAS genotype data for the Rotterdam Study is supported by the Netherlands Organization of Scientific Research NWO Investments (nr. 175.010.2005.011, 911-03-012). This study is funded by the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) project nr. 050-060-810. We thank Pascal Arp, Mila Jhamai, Marijn Verkerk, Lizbeth Herrera and Marjolein Peters for their help in creating the GWAS database. Work on cell count estimation was funded by NWO 863.13.011.

## Author contributions

BTH, PACtH, JBJvM, AI, RJ and LF formed the management team of the BIOS consortium. DIB, RP, JvD, JJH, MMJVG, CDAS, CJHvdK, CGS, CW, LF, AZ, EFG, PES, MB, JD, DvH, JHV, LHvdB, CMvD, AH, AI, AGU managed and organized the biobanks. JBJvM, PMJ, MV, JvR and NL generated RNA-seq data. HM, MvI, MvG, WA, JB, DVZ, RJ, PvtH, PD, MV, IN, MaS, PACtH, BTH and MM were responsible for data management and the computational infrastructure. DVZ, PD, MV, MvI, FvD, MvG, WA, MJB, NdK, HJW, SMK, JL, MAS, PACtH and LF performed the data analysis. DVZ, PD, PACtH and LF drafted the manuscript.

## Additional material

The following supplements are available with the on-line version of this paper.

Figure S1: Number of independent eQTLs in gene-level, exon-level, exon ratio and polyA ratio eQTL mapping
Figure S2: Detailed workflow of interaction analysis
Figure S3: Heatmap of expression of proxy genes per module in the BLUEPRINT data
Figure S4: Expression of module eQTL genes in BLUEPRINT data and their replication in previously reported cell-type-specific datasets
Figure S5: Comparison of interaction Z-score obtained by using cell counts and using proxy genes.
Figure S6: eQTLs associated with different autoimmune diseases.
Figure S7: Histogram of RNA blood sampling age distribution per biobank
Figure S8: Plots of picard metrics results and the outlier samples removed based on these metrics
Figure S9: Accuracy of cell count prediction method
Table S1: Number of samples before and after QC
Table S2: Replication of BBMRI and Geuvadis eQTLs

Table S3: Total number of primary cis-eQTLs
Table S4: GWAVA functional annotation of eSNPs
Table S5: Enhancer enrichment
Table S6: The set of trait/disease-associated variants used for eQTL annotation
Table S7: eQTLs associated with diseases and complex traits
Table S8: Top 100 proxy genes and corresponding eQTLs for the top 10 interaction modules
Table S9: Correlation of eQTL interaction z-scores with cell-type-specific interaction analysis
Table S10: Overlap of eQTLs significantly interacting with identified modules and those significantly interacting with measured cell counts.
Table S11: Replication of interactions in Geuvadis
Table S12: Pathway enrichment analysis results for the cell-type-specific eQTL genes of the top 10 interaction modules for gene, exon and exon ratio level analysis
Table S13: TF enrichment analysis of significant context-specific eQTLs
Table S14: GWAS hits for eQTLs significantly interacting with top 10 modules
Table S15: Interactions remaining after correcting for the first ten proxy genes
Table S16: Replication of interactions not falling into the top 10 modules in Geuvadis
Note S1: Supplementary results



## References


1. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22,** 1748–1759 (2012).
2. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 9362–7 (2009).
3. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466,** 714–9 (2010).
4. Cvejic, A. *et al.* SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nat. Genet.* **45,** 542–5 (2013).
5. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507,** 371–5 (2014).
6. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.* **373,** 895–907 (2015).
7. Fu, J. *et al.* Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet* **8,** e1002431 (2012).
8. Fairfax, B. P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44,** 502–10 (2012).
9. Andiappan, A. K. *et al.* Genome-wide analysis of the genetic regulation of gene expression in human neutrophils. *Nat. Commun.* **6,** 7971 (2015).
10. Francesconi, M. & Lehner, B. The effects of genetic variation on gene expression dynamics during development. *Nature* **505,** 208–11 (2014).
11. Powell, J. E. *et al.* Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. *Genome Res.* **22,** 456–66 (2012).

12. Deelen, P. *et al.* Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med.* **7,** 30 (2015).
13. Westra, H.-J. *et al.* Cell Specific eQTL Analysis without Sorting Cells. *PLoS Genet.* **11,** e1005223 (2015).
14. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343,** 1246949 (2014).
15. Çalışkan, M., Baker, S. W., Gilad, Y. & Ober, C. Host genetic variation influences gene expression response to rhinovirus infection. *PLoS Genet.* **11,** e1005111 (2015).
16. Lee, M. N. *et al.* Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells. *Science (80-. ).* **343,** 1246980–1246980 (2014).
17. Barreiro, L. B. *et al.* Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 1204–9 (2012).
18. van Greevenbroek, M. M. J. *et al.* The cross-sectional association between insulin resistance and circulating complement C3 is partly explained by plasma alanine aminotransferase, independent of central obesity and general inflammation (the CODAM study). *Eur. J. Clin. Invest.* **41,** 372–379 (2011).
19. Tigchelaar, E. F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5,** (2015).
20. Schoenmaker, M. *et al.* Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *Eur. J. Hum. Genet.* **14,** 79–84 (2006).
21. Hofman, A. *et al.* The rotterdam study: 2014 objectives and design update. *Eur. J. Epidemiol.* **28,** 889–926 (2013).
22. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).
23. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).
24. Wood, A. R. *et al.* Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. *Hum. Mol. Genet.* **20,** 4082–92 (2011).
25. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods* **11,** 294–6 (2014).
26. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–61 (2014).
27. Naranbhai, V. *et al.* Genomic modulators of gene expression in human neutrophils. *Nat. Commun.* **6,** 7545 (2015).
28. Raj, T. *et al.* Polarization of the Effects of Autoimmune and Neurodegenerative Risk Alleles in Leukocytes. *Science (80-. ).* **344,** 519–523 (2014).
29. Idaghdour, Y. *et al.* Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nat. Genet.* **42,** 62–7 (2010).
30. Yao, C. *et al.* Sex- and age-interacting eQTLs in human complex diseases. *Hum. Mol. Genet.* **23,** 1947–56 (2014).
31. Adams, D. *et al.* BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotechnol.* **30,** 224–226 (2012).

32. Dore, L. C. & Crispino, J. D. Transcription factor networks in erythroid cell and megakaryocyte development. *Blood* **118,** 231–239 (2011).
33. Hall, M. A. *et al.* The critical regulator of embryonic hematopoiesis, SCL, is vital in the adult for megakaryopoiesis, erythropoiesis, and lineage choice in CFU-S12. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 992–7 (2003).
34. Pevny, L. *et al.* Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature* **349,** 257–60 (1991).
35. Rusinova, I. *et al.* Interferome v2.0: an updated database of annotated interferon-regulated genes. *Nucleic Acids Res.* **41,** D1040-6 (2013).
36. Platanias, L. C. Mechanisms of type-I- and type-II-interferon-mediated signalling. *Nat. Rev. Immunol.* **5,** 375–86 (2005).
37. Ivashkiv, L. B. & Donlin, L. T. Regulation of type I interferon responses. *Nat. Rev. Immunol.* **14,** 36–49 (2014).
38. McLeay, R. C. & Bailey, T. L. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11,** 165 (2010).
39. Heinrichs, S. *et al.* MYBL2 is a sub-haploinsufficient tumor suppressor gene in myeloid malignancy. *Elife* **2,** e00825 (2013).
40. Facchetti, F., Cella, M., Festa, S., Fremont, D. H. & Colonna, M. An unusual Fc receptor-related protein expressed in human centroblasts. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 3776–81 (2002).
41. Rosén, A. *et al.* Lymphoblastoid cell line with B1 cell characteristics established from a chronic lymphocytic leukemia clone by in vitro EBV infection. *Oncoimmunology* **1,** 18–27 (2012).
42. Van Dam, R. M., Boer, J. M. a, Feskens, E. J. M. & Seidell, J. C. Parental history off diabetes modifies the association between abdominal adiposity and hyperglycemia. *Diabetes Care* **24,** 1454–1459 (2001).
43. Scholtens, S. *et al.* Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int. J. Epidemiol.* **44,** 1172–80 (2015).
44. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
45. Liu, Z. *et al.* Comparing Computational Methods for Identification of Allele-Specific Expression based on Next Generation Sequencing Data. *Genet. Epidemiol.* (2014). doi:10.1002/gepi.21846
46. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–2 (2010).
47. Tigchelaar, E. F. *et al. An introduction to LifeLines DEEP: study design and baseline characteristics*. 0–21 (2014). doi:10.1101/009217
48. Deelen, J. *et al.* Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Hum. Mol. Genet.* **23,** 4420–32 (2014).
49. The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46,** 818–825 (2014).
50. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes* **7,** 901 (2014).

51. Howie, B., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).
52. Deelen, P. *et al.* Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *Eur J Hum Genet* **22,** 1321–1326 (2014).
53. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma.* **25,** 2078–2079 (2009).
54. Goya, R. *et al.* SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* **26,** 730–6 (2010).
55. Westra, H.-J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* **27,** 2104–11 (2011).
56. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11,** R25 (2010).
57. Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47,** 115–125 (2015).
58. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* **6,** (2015).
59. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22,** 1813–31 (2012).
60. Cline, M. S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2,** 2366–82 (2007).
61. Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315,** 972–6 (2007).
62. Bodenhofer, U., Kothmeier, A. & Hochreiter, S. APCluster: an R package for affinity propagation clustering. *Bioinformatics* **27,** 2463–4 (2011).

# 8

**BioRxiv, 2018**

Patrick Deelen[1,2,*], Sipko van Dam[1,*], Johanna C. Herkert[1,**], Juha M. Karjalainen[1,**], Harm Brugge[1,**], Kristin M. Abbott[1], Cleo C. van Diemen[1], Paul A. van der Zwaag[1], Erica H. Gerkes[1], Evelien Zonneveld-Huijssoon[1], Jelkje J. Boer-Bergsma[1], Pytrik Folkertsma[1], Tessa Gillett[1], K. Joeri van der Velde[1,2], Roan Kanninga[1,2], Peter C. van den Akker[1], Sabrina Z. Jan[1], Edgar T. Hoorntje[1,3], Wouter P. te Rijdt[1,3], Yvonne J. Vos[1], Jan D.H. Jongbloed[1], Conny M.A. van Ravenswaaij-Arts[1], Richard Sinke[1], Birgit Sikkema-Raddatz[1], Wilhelmina S. Kerstjens-Frederikse[1], Morris A. Swertz[1,2], Lude Franke[1]

# Improving the diagnostic yield of exome-sequencing, by predicting gene-phenotype associations using large-scale gene expression analysis

## Abstract

The diagnostic yield of exome and genome sequencing remains low (8-70%), due to incomplete knowledge on the genes that cause disease. To improve on this, we have created GeneNetwork Assisted Diagnostic Optimization (GADO), which uses RNA-seq data from 31,499 samples to predict which genes cause specific disease phenotypes. We show that this unbiased method, which does not rely upon specific knowledge on individual genes, is effective in both identifying new disease genes, and flagging genes that have previously been incorrectly implicated in disease. GADO can be run on www.genenetwork.nl by supplying HPO-terms and a list of genes that contain candidate variants. Finally, applying GADO to a cohort of 61 patients where exome-sequencing analysis had not resulted in a genetic diagnosis, yielded likely causal genes for ten cases.

1 University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands
2 University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, the Netherlands
3 Netherlands Heart Institute, Utrecht, the Netherlands
* Equal contributions
** Equal contributions
Corresponding author: Lude Franke

## Introduction

Diagnostic yield is steadily increasing with the increasing use of whole-exome sequencing (WES) and whole-genome sequencing (WGS) to diagnose patients with a suspected genetic disorder [1]. Although many genes have been associated to Mendelian diseases the diagnostic yield of genome sequencing remains limited, varying from 8% to 70% [2].

Tools exist that can help prioritize candidate genes based on existing knowledge, of which some use human phenotype ontology (HPO) terms [3] to denote the phenotype of a patient. However, these methods are often limited in their ability to identify previously unknown disease-gene associations [4]. For instance, AMELIE prioritizes candidate genes using an automated literature analysis, but cannot pinpoint genes, unknown to cause a certain disease [5]. In contrast, Exomiser can also predict new disease genes by using existing (knock out) annotations for genes or orthologues in other organisms [6]. Also, the tissue specificity of gene expression has been shown to be informative for predicting disease relevance [7]. While each of these methods have proven highly valuable, one challenge remains: for most protein-coding and non-coding genes very little is known, making it also very challenging to make inferences whether a mutation in those genes might cause a specific phenotype.



Another problem is that some genes or variants that have previously been implicated in the prevalence of a specific disease are now reported as either being false positive associations or having a limited penetrance [8,9]. Often these likely false associations are identified because the presumed causative variant alleles turn out to be too common in large populations, such as present in ExAC [10,11]. Alternatively, the effects of variants in some genes could not be replicated in population based biobanks [12]. Although only few genes have been definitely refuted in literature, it has been shown that many genes reported in rare disease databases only have limited evidence to link the gene to the disease [13].

Here we present a new method to overcome some of these challenges: by using 31,499 RNA-sequencing (RNA-seq) of a wide range of tissues and cell types we can predict gene functions and disease associations using gene co-regulation, while not being biased towards existing gene annotations by using a leave-one-out procedure. This allowed us to accurately predict gene functions and to prioritize candidate diseases genes with high accuracy. This is possible because if genes are known to cause a specific disease or disease symptom they will often have similar molecular functions or be involved in the same biological process or pathway [14]. When the reported disease associations cannot be predicted this may indicate false positive associations.

We have developed a user-friendly web-based tool called GADO (GeneNetwork Assisted Diagnostic Optimization, available at www.genenetwork.nl.) that can prioritize variants in known *and* unknown genes using HPO-terms to describe a patient's phenotype. GADO ranks variants using HPO terms to describe a patient's phenotype. To validate our prioritization method, we tested how well our method predicts disease-causing genes based on HPO-terms described for each of the genes in the OMIM database. We then used exome sequencing data of patients with a known genetic diagnosis to benchmark GADO. Finally, we applied our methodology to previously inconclusive WES data and identified several genes

that contain variants that likely explain the phenotype of the respective patients. Thus, we show that our methodology is successful in identifying variants in novel, likely relevant genes explaining the patient's phenotype.

## Results

### Gene prioritization using GADO

We have developed GADO, a method that can perform gene prioritizations, which uses as input a list of phenotypes (described using HPO terms [15]) that have been observed in a patient. In combination with a list of candidate genes (i.e. genes harboring rare and possibly damaging variants), GADO reports a ranked list of genes with the most likely candidate genes on top (Figure 1a). These gene prioritizations are based on the predicted involvement of the candidate genes for the specified set of HPO terms. These predictions are made by analyzing public RNA-seq data from 31,499 samples (Figure 1b), resulting in a gene prioritization Z-score for each HPO term. These predictions are solely based on observed co-regulation of genes annotated to a certain HPO term with other genes. This makes it possible to also prioritize genes that currently lack any biological annotation.

### Public RNA-seq data acquisition and quality control

To predict functions of genes and HPO term associations, all human RNA-seq samples that were publicly available in the European Nucleotide Archive (accessed June 30, 2016) were downloaded (supplementary table 1) [16]. Gene-expression was quantified by using Kallisto [17] and samples for which a limited number of reads are mapped, were removed. A principal component analysis (PCA) on the correlation matrix was used to remove low quality samples and to remove samples that were falsely annotated as RNA-seq but turned out to be DNA-seq. Finally 31,499 samples were included and gene expression levels for 56,435 genes (of which 22,375 are protein-coding) were quantified.

Although these samples are generated in many different laboratories, we previously observed that, after correcting for technical biases, it is possible to integrate these samples into a single expression dataset [18]. We validated that this is also true for our dataset by visualizing the data using t-Distributed Stochastic Neighbor Embedding (t-SNE). We labeled the samples based on cell-type or tissue and we observed that samples cluster together based on cell-type or tissue origin (Figure 2). Technical biases, such as whether single-end or paired-end sequencing had been used, did not lead to erroneous clusters, which suggests that this heterogeneous dataset can be used to ascertain co-regulation between genes and can thus serve as the basis for predicting the functions of genes. (Supplementary methods 1)

### Prediction of gene HPO associations and gene functions

To predict HPO term associations and putative gene functions (Figure 1b), we used a co-regulation method that we had previously developed and applied to public expression microarrays [14]. However, since microarrays only cover a subset of the protein-coding genes (n = 14,510), we decided to use public RNA-seq data here instead. This allows for more accurate quantification of lower expressed genes and the expression quantification of many more genes, including a large number of non-protein-coding genes [19]. Our method uses principal

component analysis to identify a set of components that describe co-regulation between genes. While some of this co-regulation between genes is determined by pairs of genes that are specifically expressed in certain tissues (i.e. tissue-specific expression), a considerable proportion of this co-regulation reflects pairs of genes that are involved in the same biological pathways.

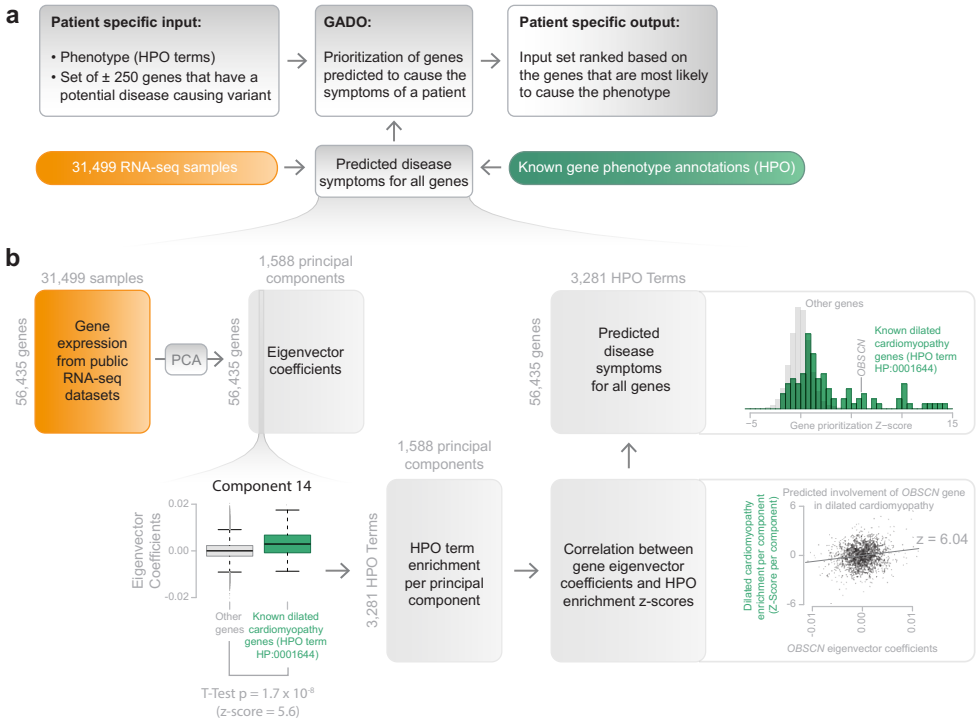



***Figure 1: Schematic overview of GADO.*** *(a) Per patient, GADO requires a set of phenotypic features (encoded using HPO terms) and a list of candidate genes (gene names either entered using HGNC symbols or Ensembl IDs). This gene list should contain genes in which rare variants have been observed for the patient. It then ascertains whether any of these genes have been predicted to cause the phenotypic features, observed in the patient. These HPO phenotypes predictions per gene are based on observed co-regulation with sets of genes that are already known to be associated with these phenotypes. (b) Overview of how disease symptoms are predicted using gene expression data from 31,499 human RNA-seq samples. A principal component analysis on the co-expression matrix results in the identification of 1,588 significant principal components. For each HPO term we investigate every component: per component we test whether there is a significant difference between eigenvector coefficients of genes known to cause a specific phenotype and a background set of genes. This results in a matrix that indicates which principal components are informative for every HPO term. By correlating this matrix to the eigenvector coefficients of every individual gene, it is possible to infer the likely HPO disease phenotype term that would be the result of a pathogenic variant in that gene.*

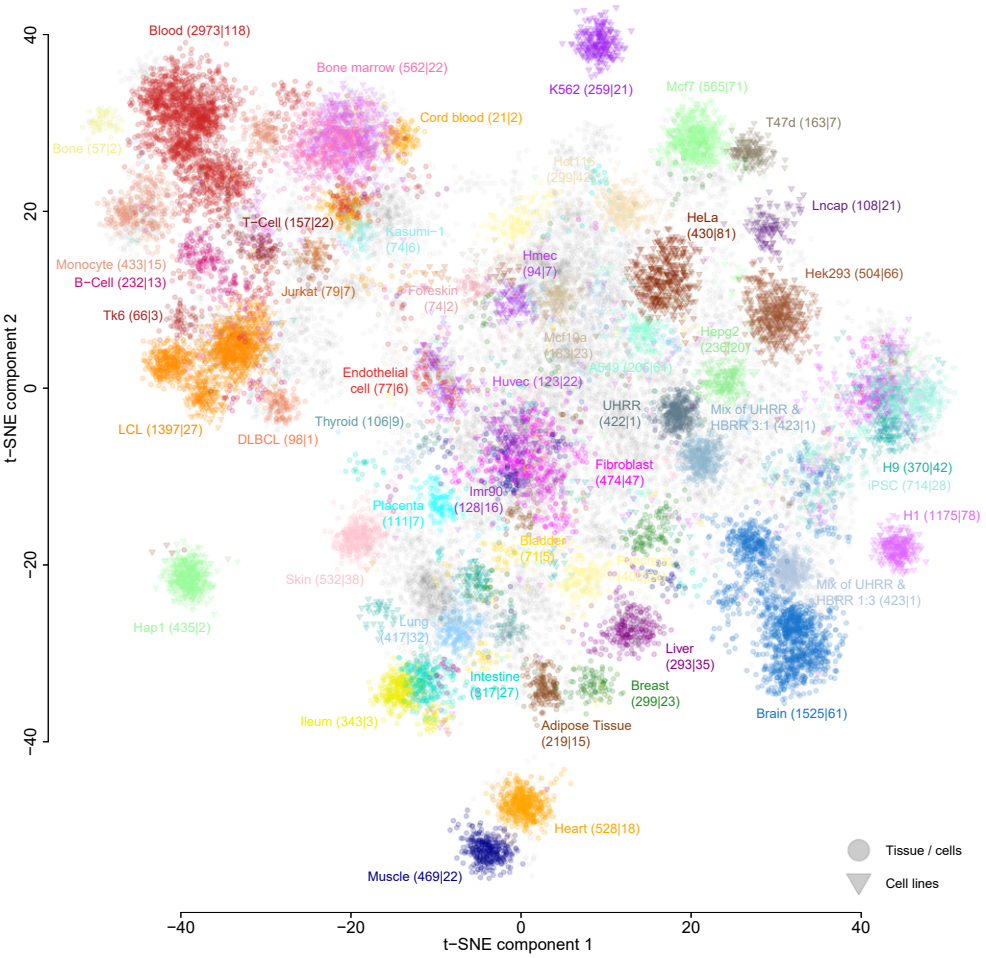



***Figure 2: A compendium of gene expression profiles that can be used for gene function prediction.*** *We downloaded 31,499 RNA-seq samples from ENA. These samples come from many different studies. They show coherent clustering after correcting for technical biases. Generally, samples originating from the same tissue, cell-type or cell-line cluster together. The two axes denote the first t-SNE components. The number of samples per tissue or cell-type are mentioned, and after the colon the number of unique studies is mentioned, indicating that samples cluster by tissue or cell-type, and that this clustering is not due to systematic technical confounding due to the fact that for a given tissue, samples come from only a single laboratory.*

We applied this prediction methodology [14] to the HPO gene sets and also to Reactome [20], KEGG pathways [21], Gene Ontology (GO) molecular function, GO biological process and GO cellular component [22] gene sets. For 5,088 of the 8,657 gene sets (59%) with at least 10 genes annotated, the gene function predictions had significant predictive power (see

materials and methods). For the 8,657 gene sets with at least 10 genes annotated, the median predictive power, denoted as Area Under the Curve (AUC), ranged between 0.73 (HPO) to 0.87 (Reactome) (Table 1).

## Prioritization of known disease genes using the annotated HPO terms

Once we had calculated the prioritization Z-scores of HPO disease phenotypes, we leveraged these scores to prioritize genes found by sequencing the DNA of a patient. For each individual HPO term–gene combination, we calculated a prioritization Z-score that can be used to rank genes. In practice, however, patients often present with not one feature but a combination of multiple phenotypic features. Therefore, we combined the prioritization Z-scores for each HPO term to generate an overall prioritization Z-score that explains the full spectrum of features in a patient. GADO uses these combined prioritization Z-scores to prioritize the candidate genes: the higher the combined prioritization Z-score for a gene, the more likely it explains the patient's phenotypes.



Because many HPO terms have fewer than 10 genes annotated, and since we were unable to make significant predictions for some HPO terms, certain HPO terms are not suitable to use for gene prioritization. To overcome this problem we take advantage of the way HPO terms are structured: each term has at least one parent HPO term that describes a more generic phenotype and thus has also more genes assigned to it. Therefore, if an HPO term cannot be used, GADO will make suggestions for suitable parental terms (supplementary figure 1).

To benchmark our prioritization method, we used the OMIM database [23]. Due to our leave-one-out approach (see methods) we could directly test how well our method was able to retrospectively rank disease-causing genes listed in OMIM based on the annotated symptoms of these diseases. For each OMIM disease gene (n = 3,382) we used the associated disease features (on average 15 HPO-terms per gene) as input for GADO. We found that GADO ranks the causative gene in the top 5% for 49% of the diseases (Figure 3a, supplementary figure 2). However, in clinical

| Database | Number of gene sets | Gene sets ≥ 10 genes | Gene sets with significant predictive power | Median AUC |
|---|---|---|---|---|
| Reactome | 2,143 | 1,388 | 1,150 | 0.87 |
| GO molecular function | 4,070 | 726 | 398 | 0.82 |
| GO biological process | 11,753 | 2,576 | 1,115 | 0.82 |
| GO cellular component | 1,609 | 500 | 370 | 0.84 |
| KEGG | 186 | 186 | 168 | 0.84 |
| HPO | 7,920 | 3,281 | 1,887 | 0.73 |

***Table 1: Gene function prediction accuracy.*** *Gene co-expression information of 31,499 samples is used to predict gene functions. We show the prediction accuracy for gene sets from different databases. AUC, Area Under the Curve, GO, Gene Ontology, HPO, Human Phenotype Ontology.*

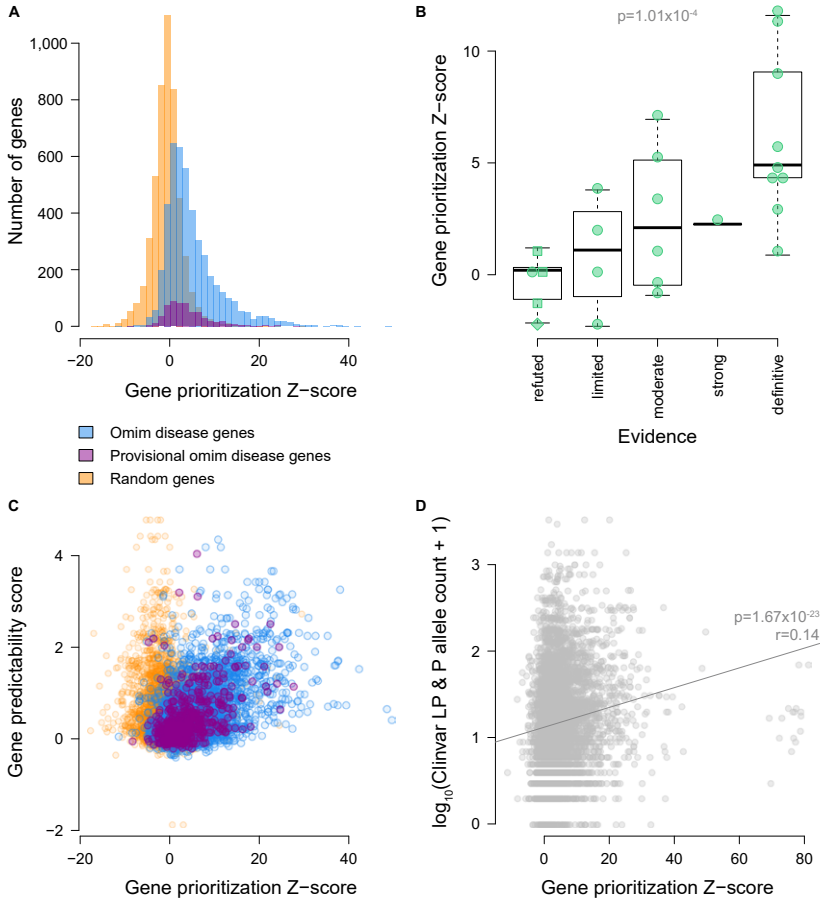



***Figure 3: Performance of disease gene prioritization compared to random permutation.*** *(a) OMIM disease genes and provisional disease genes have significantly stronger prioritization Z-scores compared to permuted disease genes (T-test p-values: $2.16\times10^{-532}$ & $5.38\times10^{-80}$, respectively). We also observe that the predictions of the provisional OMIM genes are, on average, weaker than the other OMIM disease genes (T-test p-value: $1.89\times10^{-7}$). Because we use a leave-one-out strategy when calculating prioritization Z-scores for genes that have already been associated to an HPO term, there is no prediction bias towards known associations. Therefore, this benchmark is informative of the power to predict novel associations (see methods). (b) We observe a significant relation (Spearman p-value: $1.01\times10^{-4}$) between the burden of evidence that a gene is associated to a disease and the GADO prioritization Z-score. Most genes are scored by [13] some additional refuted genes, denotated as squares or diamonds, are reported by [8] and [12] (c) We observe a clear relation between the prioritization Z-scores and the gene predictability scores (Pearson r = 0.54). We don't observe this relation in the permuted results. (d) Our gene prioritization Z-scores are significantly correlated (Pearson p-value: $1.66\times10^{-24}$) to the number of likely pathogenic (LP) and pathogenic (P) variants reported for a gene in ClinVar.*

practice it is not uncommon that only a subset of the features of a patient have been recorded. We therefore repeated this analysis while randomly selecting at most 5 HPO terms per disease. We found that the GADO scores remained stable and are strongly correlated (Pearson correlation r = 0.86) compared to using all HPO terms (supplementary figure 3).

## Gene predictability scores explain performance differences between genes

For some diseases in OMIM, GADO could not predict gene-phenotype combinations, as indicated by a prioritization Z-scores close to 0 or below 0 (Figure 3a). For example, variants in *SLC6A3* are known to cause infantile Parkinsonism-dystonia (MIM 613135) [24–26], but GADO was unable to predict the annotated HPO terms related to the Parkinsonism-dystonia for this gene. This may, however, be due to very low expression levels of SLC6A3 in most tissues except specific brain regions 27.



To better understand why we cannot predict HPO terms for all genes, we used the Reactome, GO and KEGG prioritization Z-scores. Jointly these databases comprise thousands of gene sets. Since these databases describe such a wide range of biology, we assumed that if a gene does not show any prediction signal for any gene set in these databases, gene co-expression is probably not informative for this gene. To quantify this, we calculated, per gene, the average skewness of the pathway prioritization Z-score distribution of the Reactome, GO and KEGG gene sets. This average we use as the 'gene predictability score' for every gene that is independent of whether this gene is already known to play a role in any a disease or pathway (Figure 3c, supplementary figure 2). We then ascertained whether these 'gene predictability scores' are correlated with the HPO-based prioritization Z-score of the OMIM diseases, and found a strong correlation (r: 0.54, p-value: $1.14\times10^{-332}$) between the gene predictability scores and GADO's ability to identify a known disease gene (Figure 3c, supplementary table 2).

## Disease associated genes with limited evidence for association have lower prioritization Z-scores

We used a set of disease genes that had been systematically studied by Strande et al. [13] to ascertain the burden of evidence that exists for these genes, and complemented this list with a set of refuted genes [8,12]. We observed that the GADO prioritization scores are related to this burden of evidence: refuted genes and genes with limited evidence have significantly lower prioritization Z-scores, compared to genes with more supporting evidence (Spearman p-value: $1.01\times10^{-4}$) (Figure 3b). Our prioritization Z-scores are also correlated to the number of times an allele within a gene has been reported to be pathogenic or likely pathogenic in ClinVar [28] (r: 0.14 p-value: $1.67\times10^{-23}$) (Figure 3d), which indicates that if many independent submissions have implicated the same gene in disease, that gene is more likely to be a true disease-causing gene. This is corroborated by the significant correlation between the ExAC missense constraint score [10] (a metric denoting a depletion of missense variation in a gene) and the number of submissions to ClinVar (r: 0.12 p-value: $8.81\times10^{-17}$) (Supplementary figure 4a). Interestingly, we do not observe a correlation between our prioritization Z-scores and the ExAC missense constraints (Supplementary figure 4b). A linear model to explain the number of ClinVar submissions using both our prioritization Z-scores together with the ExAC constraints performs significantly better than when solely using the ExAC constraints

to predict the number of pathogenic or likely pathogenic in ClinVar (r: 0.21 vs r: 0.12, ANOVA p-value: $1.24\times10^{-34}$). This indicates that GADO is informative for predicting the involvement of genes in disease, independent from ClinVar and ExAC.

A set of genes known to cause cardiomyopathy was scored for the amount of evidence in literature that these genes are involved in cardiomyopathy. Here, we again observe that genes with limited evidence have lower prioritization Z-scores (spearman p-value: $8.71\times10^{-04}$) (supplementary figure 5), suggesting these could potentially reflect false-positive associations.

We were somewhat worried that such false-positive associations could detrimentally affect our gene – phenotype predictions. To ascertain this, we randomly added 10% more genes to each HPO-term and recalculated the predictions. We then observed that our predictions were robust, and that AUC values (indicating to what extent gene co-regulation can predict gene – phenotype associations) were very similar to the original AUC values (Pearson correlation r = 0.97, Supplementary figure 5).

## Benchmarking GADO using solved cases with realistic phenotyping

Although these *in silico* benchmarking demonstrated the potential of GADO, it used all annotated HPO terms for a disease. In practice, however, patients may only present with a limited number of the annotated features of a disease. To perform a validation that was a realistic reflection of clinical practice, we used exome sequencing data of 83 patients with a known genetic diagnosis. We used their phenotypic features as listed in their medical records prior to when the genetic diagnosis had been made (supplementary table 3). Per patient, our exome-sequencing pipeline GAVIN [29] returned a median of 55 possible disease-causing genes with variants that are rare and predicted to be deleterious. We then ran GADO and observed that for 41% of these patients the actual causative gene ranked in the top 3 (median rank was 6.5 for all 83 patients, supplementary figure 6). Using a stringent threshold (prioritization Z-score ≥ 5), which we also used for the prioritization of unsolved cases (see below), to select strong candidate genes, we identifed the causative gene for 17 cases (20%) while only needing to follow-up a single variant (range 0-5) per patient on average.

Because of our leave-one-out procedure when calculating prioritization Z-scores for known disease genes (see methods), our performance in solved cases is indicative of the power of GADO to prioritize novel disease-associated genes without prior annotations or associations. However, these unbiased predictions can sometimes cause problems when using GADO in clinical practice, because GADO cannot predict every known gene-HPO combination accurately. As such some of these known gene-HPO combinations might have rather insignificant Z-scores. To make sure GADO is also suited for cases with variants in currently known disease associated genes, we adjusted our prediction matrix to ensure that known HPO-term associations for genes are also prioritized (see methods). This does not affect GADO's ability to prioritize novel diseases genes, but solely helps the prioritization performance of known disease genes, but ensures that users of the GADO website will see these known disease-phenotype as top ranking genes . By doing this we achieved a similar prioritization performance as compared to Exomiser (Supplementary table 4, Figure 4a). For this comparison, we used both methods to rank the on average 663 variants that are selected by Exomiser. For Exomiser, we used the default 'combined prioritization' strategy that is based

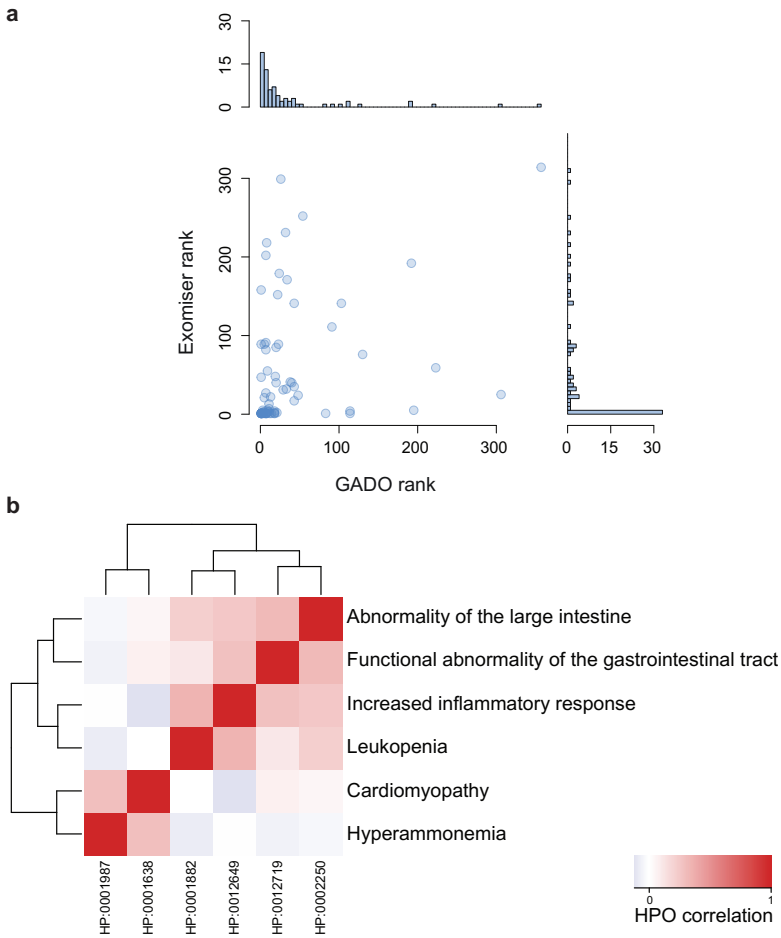



***Figure 4: Performance of GeneNetwork on solved cases.*** *(a) Comparison between using GADO and Exomiser to rank candidate variants. (b) Our cohort contained a case with two distinct conditions, and clustering showed the HPO terms of the same disease are closest to each other. Note, the HPO term "Inflammation of the large intestine" did not yield a significant prediction profile and therefore the parent terms "Abnormality of the large intestine", "Increased inflammatory response" and "Functional abnormality of the gastrointestinal tract" were used for this case.*

on the variant score and the gene score, whereas in GADO we solely used the prioritization Z-scores (Supplementary methods 2). Although our median rank of the causative gene is better compared to Exomiser (GADO: 12.5 vs Exomiser: 21), Exomiser on the other hand, is able to rank more genes in the top 3 (Exomiser: 28 vs GADO: 14).

## Clustering of HPO terms

In addition to ranking potentially causative genes based on a patient's phenotype, GADO can be used to cluster HPO terms based on the genes that are predicted to be associated to these HPO terms. This can help to identify pairs of symptoms that often occur together, as well as symptoms that rarely co-occur. In a patient diagnosed with a glycogen storage disease, GSD type Ib, caused by compound heterozygous variants in *SLC37A4* (MIM 602671) and Dilated Cardiomyopathy (DCM) that is probably caused by a truncating variant in *TTN* (MIM 188840) HPO terms related to GSD type Ib ('leukopenia' (HP:0001882) and 'inflammation of the large intestine' (HP:0002037)) cluster together, while Cardiomyopathy (HP:0001638) was only weakly correlated to these specific features (Figure 4b).

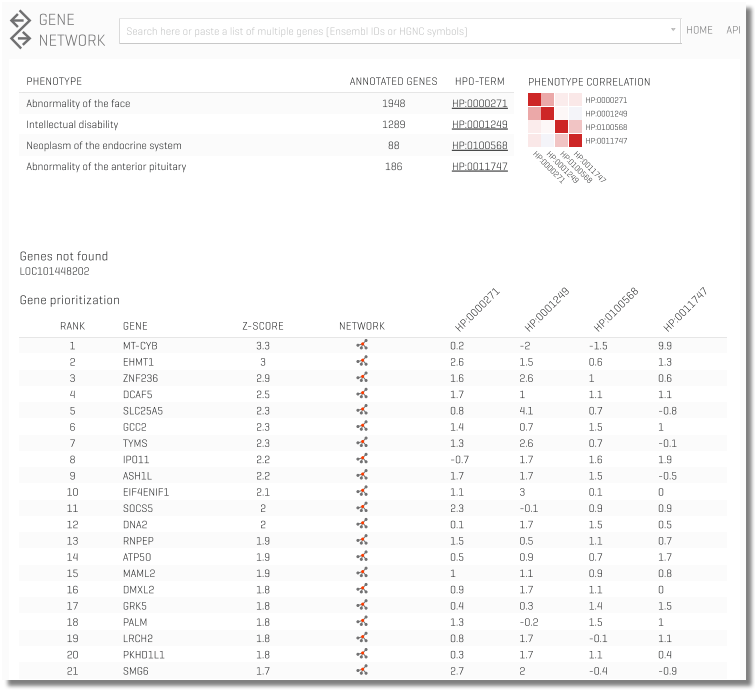
## Reanalysis of previously unsolved cases

To assess GADO's ability to discover new disease genes, we applied it to data from 61 patients who are suspected to have a Mendelian disease but who did not receive a genetic diagnosis. All patients had undergone prior genetic testing (WES with analysis of a gene panel according to their phenotype, supplementary table 5). On average GADO reported 2.9 genes with a prioritization Z-score ≥ 5 (which we used as an arbitrary cut-off and that corresponds to a p-value ≤ $5.7\times10^{-7}$) and which were further assessed. In ten cases, we identified variants in genes not associated to a disease in OMIM or other databases, but for which we could find literature or for which we gained functional evidence implicating their disease relevance (Table 2). For example, we identified two cases with DCM with rare compound heterozygous variants in the *OBSCN* gene (MIM 608616) that are predicted to be damaging. In literature, inherited variant(s) in *OBSCN*, encoding obscurin, are associated with hypertrophic CM [30] and DCM [31]. Furthermore, obscurin is a known interaction partner of titin (TTN), a well-known DCM-related protein [30]. Another example came from a patient with ichthyotic peeling skin syndrome, which is caused by a damaging variant in *FLG2 (*MIM 616284). We recently published this case where we prioritized this gene using an alpha version of GADO [32].

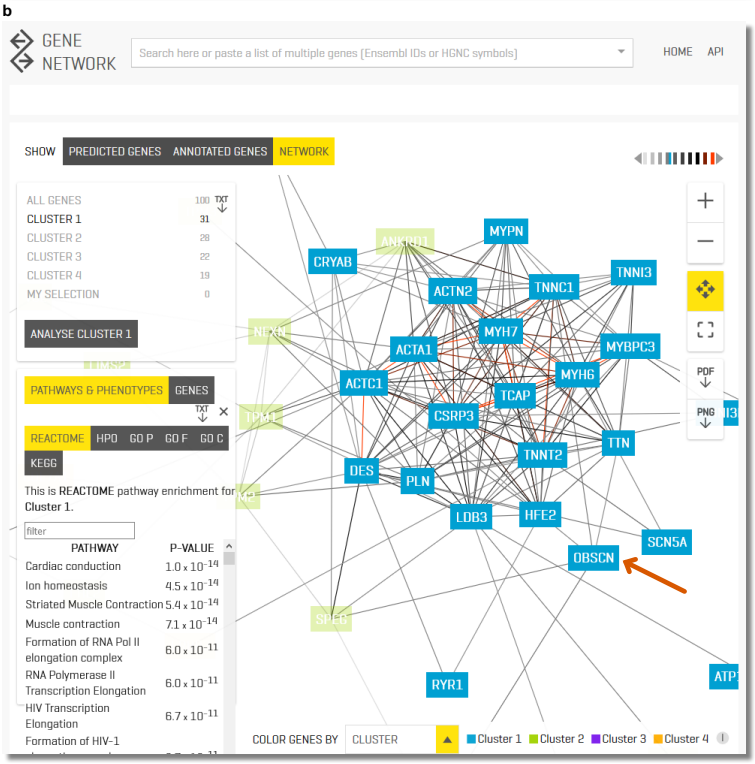
***Table 2: unsolved cases with new candidate genes.*** *In 10 out of 61 unsolved patients we identified new likely causal genes. For these genes we found literature that indicates these genes fit the phenotype of these patients or we gained functional evidence implicating their disease relevance. HP:0001644=Dilated cardiomyopathy; HP:0008066=Abnormal blistering of the skin; HP:0008064=Ichthyosis; HP:0001263=Global developmental delay; HP:0001249=Intellectual disability; HP:0000717=Autism; HP:0000708=Behavioral abnormality; HP:0002167=Neurological speech impairment; HP:0002360=Sleep disturbance; HP:0000664=Synophrys; HP:0001638=Cardiomyopathy; HP:0004322=Short stature; HP:0001249=Intellectual disability; HP:0003493=Antinuclear antibody positivity; HP:0002583=Colitis; HP:0012649=Increased inflammatory response; HP:0001890=Autoimmune hemolytic anemia. *These variants where pre-filtered for family segregation. **The variants in these genes do not fully explain the phenotype but are likely contributing to the phenotype.*

| HPO terms used | Number of genes with candidate variant | Number of genes with Z ≥ 5 | Candidate gene | Variants | CADD scores | GnomAD minor allele frequency | Supporting papers | Expression in relevant tissue |
|---|---|---|---|---|---|---|---|---|
| HP:0001644 | 247 | 5 | OBSCN | NM_001098623.2:c.[15037C>T];[20963delC] | 24.8<br>25.2 | $8.0 \times 10^{-5}$<br>$1.7 \times 10^{-3}$ | [30,31] | Yes |
| HP:0001644 | 226 | 3 | OBSCN | NM_001098623.2:c.[5545C>T];[22384+3_22384+21del] | 14.7<br>7.8 | $3.2 \times 10^{-4}$<br>0 | [30,31] | Yes |
| HP:0008066<br>HP:0008064 | 359 | 3 | FLG2 | NM_001014342.2:c.[632C>G];[632C>G] | 35.0 | $1.1 \times 10^{-5}$ | [33] | Yes |
| HP:0001263<br>HP:0001249<br>HP:0000717<br>HP:0000708<br>HP:0002167<br>HP:0002360<br>HP:0000664 | 206 | 12 | INO80 | NM_017553.2:c.[898C>T] | 34 | 0 | [34,35] | Yes |
| HP:0001644 | 346 * | 2 | MB | NM_00203377.1:c.[214G>A] | 22.4 | $3.6 \times 10^{-5}$ | [36] | Yes |
| HP:0001644 | 126 * | 1 | SYNPO2L ** | NM_001114133.2:c.[473G>A] | 24.1 | $5.4 \times 10^{-4}$ | [37] | Yes |
| HP:0001638 | 336 | 4 | NRAP ** | NM_001261463.1:c.[4648C>T] | 20.4 | $8.7 \times 10^{-4}$ | [38] | Yes |
| HP:0004322<br>HP:0001249 | 381 | 10 | CCNB2 | NM_004701.3:c.25-3_25delCAGG | 24.5 | 0 | [39] | Yes |
| HP:0003493<br>HP:0002583 | 246 | 6 | LY75 | NM_002349.2:c.3476C>T(;)23C>G | 22.7<br>24.1 | $3.2 \times 10^{-3}$<br>$2.6 \times 10^{-3}$ | [40] | Yes |
| HP:0012649<br>HP:0002583<br>HP:0001890 | 318 | 8 | AGAP2 | NM_001122772.1:c.421delC | 27.2 | 0 | [41] | Yes |

**a**


GENE NETWORK
Search here or paste a list of multiple genes (Ensembl IDs or HGNC symbols)
HOME
API
PHENOTYPE
ANNOTATED GENES
HPO-TERM
PHENOTYPE CORRELATION
Abnormality of the face 1948 HP:0000271
Intellectual disability 1289 HP:0001249
Neoplasm of the endocrine system 88 HP:0100568
Abnormality of the anterior pituitary 186 HP:0011747
Genes not found
LOC101448202
Gene prioritization
RANK GENE Z-SCORE NETWORK HP:0000271 HP:0001249 HP:0100568 HP:0011747
1 MT-CYB 3.3 0.2 -2 -1.5 9.9
2 EHMT1 3 2.6 1.5 0.6 1.3
3 ZNF236 2.9 1.6 2.6 1 0.6
4 DCAF5 2.5 1.7 1 1.1 1.1
5 SLC25A5 2.3 0.8 4.1 0.7 -0.8
6 GCC2 2.3 1.4 0.7 1.5 1
7 TYMS 2.3 1.3 2.6 0.7 -0.1
8 IPO11 2.2 -0.7 1.7 1.6 1.9
9 ASH1L 2.2 1.7 1.7 1.5 -0.5
10 EIF4ENIF1 2.1 1.1 3 0.1 0
11 SOCS5 2 2.3 -0.1 0.9 0.9
12 DNA2 2 0.1 1.7 1.5 0.5
13 RNPEP 1.9 1.5 0.5 1.1 0.7
14 ATP5D 1.9 0.5 0.9 0.7 1.7
15 MAML2 1.9 1 1.1 0.9 0.8
16 DMXL2 1.8 0.9 1.7 1.1 0
17 GRK5 1.8 0.4 0.3 1.4 1.5
18 PALM 1.8 1.3 -0.2 1.5 1
19 LRCH2 1.8 0.8 1.7 -0.1 1.1
20 PKHD1L1 1.8 0.3 1.7 1.1 0.4
21 SMG6 1.7 2.7 2 -0.4 -0.9

**b**


GENE NETWORK
Search here or paste a list of multiple genes (Ensembl IDs or HGNC symbols)
HOME
API
SHOW PREDICTED GENES ANNOTATED GENES NETWORK
ALL GENES 100
CLUSTER 1 31
CLUSTER 2 28
CLUSTER 3 22
CLUSTER 4 19
MY SELECTION 0
ANALYSE CLUSTER 1
PATHWAYS & PHENOTYPES GENES
REACTOME HPO GO P GO F GO C
KEGG
This is REACTOME pathway enrichment for Cluster 1.
filter
PATHWAY P-VALUE
Cardiac conduction 1.0 x 10-14
Ion homeostasis 4.5 x 10-14
Striated Muscle Contraction 5.4 x 10-14
Muscle contraction 7.1 x 10-14
Formation of RNA Pol II elongation complex 6.0 x 10-11
RNA Polymerase II Transcription Elongation 6.0 x 10-11
HIV Transcription Elongation 6.7 x 10-11
Formation of HIV-1
MYPN
CRYAB
TNNI3
ACTN2
TNNC1
MYH7
ACTA1
MYBPC3
ACTC1
MYH6
TCAP
CSRP3
TTN
TNNT2
DES
PLN
LDB3
HFE2
SCN5A
OBSCN
RYR1
PDF
PNG
COLOR GENES BY CLUSTER
Cluster 1 Cluster 2 Cluster 3 Cluster 4

We compared GADO to Exomiser, ENDEAVOR [42] and ToppGene [43] on our unsolved cases for which we identify a strong candidate (Supplementary methods 3). Exomiser could be run directly using the HPO terms. The other tools required a list of training genes (i.e. the genes known to cause a specific HPO term), but provided no options to integrate the results of multiple sets of training genes. We therefore only used ENDEAVOR and ToppGene for those cases with a single reported HPO term. ENDEAVOR supported a maximum of 200 input genes in the training set (i.e. those genes known to cause a specific HPO term) and at most 200 genes to prioritize (i.e. those genes in which rare variants had been observed). If for an HPO term over 200 genes were known, we selected a random subset of 200 genes. If a patient had candidate variants in more than 200 genes, we trimmed this set to 200 genes by randomly removing genes, while ensuring that the known causal gene was retained. The median rank of these genes was 3 for GADO, 68.5 for Exomiser, 7.5 for ENDEAVOR and 24 for ToppGene (Supplementary table 6). The Exomiser ranks however, are not directly comparable since Exomiser does its own variant select which yields more variants than GAVIN, the method we used prior to running GADO, ENDEAVOR and ToppGene. To overcome this, we also calculated the percentile of the candidate gene among the total genes selected either by GAVIN or Exomiser, the median percentile for GADO was: 1.2 and for Exomiser: 7.9.



## www.genenetwork.nl

All analyses described in this paper can be performed using our online toolbox at www.genenetwork.nl. Users can perform gene prioritizations using GADO by providing a set of HPO terms and a list of candidate genes (Figure 5a). Per gene, it is also possible to download all prioritization Z-scores for the HPO terms and pathways. Furthermore, the predicted pathway and HPO annotations of genes can be used to perform function enrichment analysis (Figure 5b). We also support automated queries to our database using a http+JSON api.

***Figure 5: www.genenetwork.nl*** *(a) Prioritization results of one of our previously solved cases. This patient was diagnosed with Kleefstra syndrome. The patient only showed a few of the phenotypic features associated with Kleefstra syndrome and additionally had a neoplasm of the pituitary (which is not associated with Kleefstra syndrome). Despite this limited overlap in phenotypic features, GADO was able to rank the causative gene (EHMT1) second. Here, we also show the value of the HPO clustering heatmap, the two terms related to the neoplasm cluster separately from the intellectual disability and the facial abnormalities that are associated to Kleefstra syndrome. (b) Clustering of a set of genes allowing function / HPO enrichment of all genes or specific enrichment of automatically defined sub clusters. Here we loaded all known DCM genes and OBSCN, and we focus on a sub-cluster of genes containing OBSCN (highlighted by the arrow). We see that it is strongly co-regulated with many of the known DCM genes. Pathway enrichment of this sub-cluster reveals that these genes are most strongly enriched for the muscle contraction Reactome pathway. DCM, Dilated Cardiomyopathy.*

## Discussion

The identification of new disease-causing genes is a daunting process. GADO can aid in the discovery of these unknown disease genes. The main advantage of our methodology is that it does not rely on any prior knowledge about the genes that we prioritize and can therefor also detect genes for which nothing is known. Instead, we used predicted gene functions based on co-regulation networks extracted from a large compendium of publicly available RNA-seq samples allowing accurate expression quantification of many genes, including lowly expressed genes and non-coding genes [18]. A realistic benchmark using real cases and features listed in the medical records allowed us to identify the causative genes for 20% of the cases, while only requiring us to follow-up on average only a single gene per patient.

GADO is trained in such a way that for each gene – phenotype combination that is already known, this knowledge is not used when using co-regulation information to make inferences on that specific gene-phenotype association. A major advantage of this is that our gene – phenotype predictions are not biased towards known associations. However, since we do not incorporate these known disease associations into our model, the performance of GADO is lower when studying patients with mutations in well-established genes, as compared to methods that explicitly use these known gene – phenotype associations. To accommodate this issue, we have added these known gene – phenotypes to GADO, to ensure GADO users will not miss out on known associations. This is useful for genes with a low predictability score indicating that gene expression data is not informative for its function predictions and for genes such *TTR* that act in a unique manner compared to other genes that give rise to CM. *TTR* is implicated in hereditary amyloidosis (MIM 105210) [44] and there is a large amount of evidence linking this gene to CM. Mutations in *TTR* cause accumulation of the transthyretin protein in different organ systems, including the heart, resulting in CM. However, this gene is primarily expressed in the liver. Therefore, its disease mechanism is different from other mechanisms resulting in CM, as many inherited CMs are caused by deleterious variants in genes highly expressed in the heart and directly affecting the function of the cardiac sarcomere [45]. Because this gene is expressed in a different tissue than all other CM genes co-expression is not informative and as a result the phenotypic function prediction for this gene is worse than we would expect based on the predictability score.

Finally, we used GADO on 61 unsolved cases and identified for 10 cases (16.4%) the potential novel disease genes that are strong candidates based on literature or functional evidence. All these samples already went through an extensive diagnostic procedure so these finding are on top of the normal diagnostic yield. When applying GADO, we could identify a very likely causal gene for 16.4% of these unsolved cases, based on the existence of circumstantial evidence in literature on these genes. This is only a bit lower than what we observed for solved cases where the causal gene is known: when we assumed that the causal gene was not yet known, GADO identified the causative gene for 20% of the cases. We should note that this 16.4% yield in unsolved cases might actually be an underestimate: GADO also had prioritized genes with a high significance score for some of the other unsolved cases, of which some are likely to be responsible for the phenotypes observed in these patients. Regretfully, for these genes no literature currently exists that supports their role in the diseases of these patients. This is one of the pertinent issues when it comes down to diagnosing

patients. Additional repositories that use orthogonal data to make inferences on the phenotypic consequences of mutations in genes and initiatives like Genematcher [46] therefore remain urgently needed, in order to increase the diagnostic yield.

Given that nearly 5% of patients with a Mendelian disease have another genetic disease [47], it is important to consider that multiple genes might each contribute to specific phenotypic effects. Clinically, it can be difficult to assess if a patient suffers from two inherited conditions, which may hinder variant interpretation based on HPO terms. We showed that GADO can disentangle the phenotypic features of two different diseases manifesting in one patient by correlating and subsequently clustering the profiles of HPO terms describing the patient's phenotype. If the HPO terms observed for a patient do not correlate, it is more likely that they are caused by two different diseases. An early indication that this might be the case for a specific patient can simplify subsequent analysis because the geneticist or laboratory specialist performing the variant interpretation can take this in consideration. GADO also facilitates separate prioritizations on subsets of the phenotypic features.



We compared GADO to Exomiser, which is closely related to GADO as it prioritizes genes based on specified HPO terms and also infers HPO annotation for unknown genes [6]. The gene prioritization by Exomiser is based on the effects of orthologs in model organisms and applies a guilt-by-association method using protein-protein associations provided by STRING [48]. Exomiser performs similar to GADO in ranking known disease-causing genes (supplementary figure 7, supplementary table 4) and is also able to identify potential new genes in human disease. However, only a subset of the protein-coding genes have orthologous genes in other species for which a knockout model also exists and the used STRING interactions are biased towards well studied genes and rely heavily on existing annotations to biological pathways (supplementary figure 8). There are however, still 3,922 protein-coding genes that are not currently annotated in any of the databases we used, and there are even more non-coding genes for which the biological function or role in disease is unknown. Since GADO does not rely on prior knowledge, it can be used to prioritize variants in both coding *and* non-coding genes (for which no or limited information is available). GADO thus enables the discovery of novel human disease genes and can complement existing tools in analyzing the genomic data of patients who have a broad spectrum of phenotypic abnormalities.

Others tools such as, ENDEAVOR [42], ToppGene [43] and Suspects [49], that have been used successfully before to prioritize candidate genes are not directly comparable to GADO, since these tools work by either supplying a single HPO term or a set of training genes. However, these tools can be used to successfully prioritize disease genes [50]. In some cases, a single HPO term might be sufficient or a custom gene can be useful when a specific syndrome is suspected and already several other genes have already been implicated for this syndrome. Unfortunately, in clinical practice often multiple HPO terms are needed to describe a patient's phenotype (e.g. for our set of solved cases we used two HPO terms on average). Morepver, it is also often unclear which syndrome a patient has, which inhibits the ability to prioritize genes based on already associated genes to a syndrome.

We found that for some disease genes GADO is unable to predict the already known phenotypic consequences. This is partially explained by genes for which gene-expression data is not informative for function predictions. For instance, because a gene has very low gene

expression, because different splice variants have different functions, or because the regulation of a gene its function relies heavily on post-translation modification. We have defined an empirical measurement called 'gene predictability' that indicates how informative gene expression is for function prediction of individual genes. We found a strong correlation between this predictability metric and our ability to predict known phenotypic consequences of disease associated genes. This however does not fully explain our inability to predict known phenotypicconsequences, in some cases this can simply be due to an alternative disease mechanism.

GADO can also point to genes that may have been falsely associated to a disease. Genes for which there is limited evidence to link them to a disease have, on average, lower prioritization Z-scores compared to well established genes and genes that have been refuted in literature have even lower scores. In addition, we found a statistically significant association between the prioritization Z-scores of known disease-gene combinations and the number of pathogenic or likely pathogenic alleles reported in ClinVar, thereby assuming that the genes with many submissions are more likely to be truly related to human disease. We also observed a statistically significant correlation between the ExAC missense constraint and the number of alleles submitted to ClinVar. Interestingly, the ExAC missense constraints are not correlated to our prediction scores showing that both can be used as independent predictors of potential false-positive disease associations.

The median prediction performance of HPO terms is lower compared to the other gene sets databases used in our study, such as Reactome. This may be due to the fact that phenotypes can arise by disrupting multiple distinct biological pathways. For instance, DCM can be caused by variants in sarcomeric protein genes, but also by variants in calcium/sodium handling genes or by transcription factor genes [45]. As our methodology makes guilt-by-association predictions based on whether genes are showing correlated gene expression levels, the fact that multiple separately working processes can cause the same phenotype can reduce the accuracy of the predictions (although it is often still possible to use these predictions, e.g. the DCM HPO phenotype prediction performance AUC = 0.76). We envision that by creating sub-clusters, based on these different pathways, and redoing our gene-expression based predictions, it might be possible to further improve the performance of HPO based prioritizations in the future. Insufficient statistical power might also hinder accurate predictions for HPO terms. This may specifically be true for genes that are poorly expressed or expressed in only a few of the available RNA-seq samples. The latter issue we expect to overcome in the near future as the availability of RNA-seq data in public repositories is rapidly increasing. Initiatives such as Recount [51] or SkyMap [52] enable easy analysis on these samples, allowing us to update our predictions in the future, thereby increasing our prediction accuracy.

We have developed GADO, a novel approach that can aid users in prioritizing genes using multiple patient-specific HPO terms. We performed our GADO benchmarking while using GAVIN for the selection of genes that contain (likely pathogenic) rare variants. However, GADO can work with any other methodology for identification of genes harboring rare and potentially pathogenic alleles. GADO prioritizes variants in coding *and* non-coding genes, including genes for which there is no current knowledge about their function and those that have not been annotated in any ontology database. This gene prioritization is based on

co-regulation of genes identified by analyzing 31,499 publicly available RNA-seq samples. Therefore, in contrast to many other existing prioritization tools, GADO has the ability to identify novel genes involved in human disease. By providing a statistical measure of the significance of the ranked candidate variants, GADO can provide an indication for which genes its predictions are reliable. GADO can also detect phenotypes that do not cluster together, which can alert users to the possible presence of a second genetic disorder and facilitate the diagnostic process in patients with multiple non-specific phenotypic features. GADO can easily be combined with any filtering tool to prioritize variants within WES or WGS data and can also be used in gene panels such as PanelApp [53]. Finally, GADO can aid in the identification of genes falsely associated to diseases. GADO is freely available at www.genenetwork.nl to help guide the differential diagnostic process in medical genetics.



## Materials and Methods

### Gene co-regulation and function predictions

We used publicly available RNA-seq samples from the European Nucleotide Archive (ENA) database [54] to predict gene functions and gene-HPO term associations. After processing and quality control we included 31,499 sample for which we have expression quantification on 56,435 genes (in-depth details are provided in supplementary methods 1). We subsequently performed a PCA on the gene correlation matrix and selected 1,588 reliable principal components (PCs) (Cronbach's Alpha ≥ 0.7).

We used the eigenvectors of these 1,588 PCs to predict gene functions using a method we published earlier [55] (Figure 1). Per PC we used the eigenvector coefficients for the genes that are part of a gene-set and the eigenvector coefficients of the background genes that are not in the current gene-set. We used a student's T-test to compare the eigenvector coefficients of the genes in the gene-set to the eigenvector coefficients of the background genes. We then calculated a T-test p-value and converted this to a Z-score. This resulted in a matrix where for each gene-set for each of the 1,588 PCs a Z-score had been calculated, and these Z-scores reflect the importance of a specific component for predicting which genes are part of a specific geneset. In order to finally predict which genes are part of a specific geneset, we calculate the correlation between the 1,588 T-test Z-scores for a given geneset and the 1,588 eigenvector coefficients of each gene. The rationale here is that if the same components are relevant for an individual gene (as determined through the eigenvector coefficients) and also for a specific pathway (large Z-score from the T-test) then this indicates that the expression regulation of that gene is similar to the expression regulation pattern of that pathway. The p-value that belongs to this correlation was subsequently transformed to a Z-score and was used as the prioritization Z-score (where a high score makes it more likely that a gene is part of a gene-set).

### Leave-one-out procedure

However, there is one exception to this procedure when we want to calculate the prioritization Z-score for a gene – geneset combination when that gene – geneset is already known: If we would include this gene when conducting the 1,588 T-tests and subsequent Z-scores (for determining the importance of each component for predicting this geneset), a positive correlation between the 1,588 eigenvector coefficients and the 1,588 Z-scores is expected,

which leads to a bias in the predictions towards genes with a known HPO annotation. To prevent this bias, we used a leave-one-out procedure where we always exclude the current gene from the gene set and recalculate the Z-scores derived from the T-tests before correlating the profile of a gene-set to the eigenvector coefficients of this gene. This ensures that there is no inflation of prioritization Z-score for genes that already have been annotated to the corresponding gene-set. It also allows use to calculate reliable AUC based on the current annotations to a gene-set [55].

To determine the accuracy of our predictions we assessed our ability to predict known gene-set annotations: for each gene-set, we calculated an Area Under the Curve (AUC) using the prioritization Z-scores of the genes that are part of a set versus those that are not part of a set. We used a Mann–Whitney U test to calculate if the prioritization Z-score of currently annotated genes are significantly larger than the genes not annotated to this gene-set. If this is not the case, we concluded that we could not make meaningful prioritizations for this gene-set by using the 1,588 principal components.

We applied this methodology to the gene-sets described by terms in the following databases: Reactome and KEGG pathways, Gene Ontology (GO) molecular function, GO biological process and GO cellular component terms and finally to HPO terms. We excluded gene-sets with fewer than 10 annotated genes and with a p-value ≤ 0.05 (Bonferroni corrected for the number of pathways in a database).

## Gene predictability scores

To explain why for some genes we cannot predict known HPO annotation, we have established a gene predictability score. We have calculated this gene predictability using the prioritization Z-scores based on Reactome, GO and KEGG. For each gene and for each database we calculated the skewness in the distribution of the pathway prioritization Z-scores of the gene sets. We used the average skewness as the gene predictability score.

## GADO predictions

To identify potential causative variants in patients, we used HPO terms to describe a patient's features. We only used the HPO terms which have significant predictive power (Bonferroni corrected p-value of U test to calculate the AUC ≤ 0.05). If the predictions for a patient's HPO term were not significant, the parent/umbrella HPO terms were used (supplementary figure 1). The online GADO tool suggests the parent terms from which the user can then select which terms should be used in the analysis. The gene prioritization Z-scores for an HPO term were used to rank the genes. If a patient's phenotype was described by more than one HPO term, a meta-analysis was conducted to integrate the predictions of the used HPO terms. In these cases a combined prioritization Z-score was calculated using the Z-transform test [56]. This was done by adding the prioritization Z-scores for each of the patient's HPO terms and then dividing by the square root of the number of HPO terms. This will result in a combined prioritization Z-score reflecting the predictions of all the supplied HPO terms. The genes with the highest prioritization Z-scores are predicted to be the most likely candidate causative genes for a case.

In addition to the predictions described above, we have created a GADO option which ensures any HPO term associated to a gene obtains a minimum prioritization Z-score of 3 for this gene. This option is not used for the benchmark results shown within this manuscript with the exception of the comparison against Exomiser using previously solved cases which was ran once with, and once without this option.

Gene prioritization analysis using HPO terms and a list of candidate genes can be performed at www.genenetwork.nl.

## Validation of disease-gene predictions

To benchmark our method we used the OMIM morbid map [23] downloaded on March 26, 2018, containing all disease-gene-phenotype entries. From this list, we extracted the disease-gene associations, excluding non-disease and susceptibility entries. We extracted the provisional disease-gene associations separately. For each disease in OMIM, we used GADO to determine the rank of the causative gene among all genes in the OMIM morbid map. For this we used all phenotypes annotated to the OMIM disease. If any of the HPO terms did not have significant predictive power, the parent terms were used.



To determine if these distributions were significantly different from what we expect by chance, we permuted the data. We replaced the existing gene-OMIM annotation but assigned every gene to a new disease (keeping the phenotypic features for a disease together), assuring that the randomly selected gene was not already annotated to any of the phenotypes of the original gene.

## Cohort of previously solved cases

To test if GADO could help prioritize genes that contain the causative variant, we used 83 samples of patients who were previously genetically diagnosed through whole exome analysis or gene panel analysis. These samples encompass a wide variety of different Mendelian disorders (supplementary table 3). To assess which genes harbor potentially causative variants, we first called and annotated the variants from the exome sequencing files (Supplementary methods 4). For 11 of the previously solved cases, GAVIN did not flag the causative variant as a candidate. Since this is the result of the specificity and sensitivity tradeoff made by GAVIN, we added the causative genes that had been missed by GAVIN for these 11 cases, so that we could still benchmark GADO on these patients.

The phenotypic features of a patient were translated into HPO terms, which were used as input to GADO. Here we only used features reported in the medical records prior to the molecular diagnosis. If any of the HPO terms did not have significant predictive power, the parent terms were used. From the resulting list of ranked genes, the known disease genes harboring a potentially causative variant were selected. Next, we determined the rank of the gene with the known causative variant among the selected genes. If a patient harbored multiple causative variants in different genes, in case of di-genic inheritance or two inherited conditions, the median rank of these genes was reported (supplementary table 3).

### Unsolved cases cohorts

In addition to the patients with a known genetic diagnosis, we tested 61 unsolved cases (supplementary table 4). These are patients with mainly cardiomyopathies or developmental delay. All patients were previously investigated using exome sequencing, by analyzing a gene panel appropriate for their phenotype. To allow discovery of potential novel disease genes, we used GADO to rank genes with candidate variants (Supplementary methods 4). For genes with a prioritization Z-score ≥ 5, a literature search for supporting evidence was performed to assess whether these genes are likely candidate genes.

### Website

To make our method and data available we have developed a website available at www.genenetwork.nl that can be used to run GADO, lookup gene functions predictions, visualize networks using co-regulations scores and perform function enrichments of sets of genes (Supplementary methods 5).

### GADO prediction of false positives

Gene confidence annotations were retrieved from previous studies [13]. We used annotations from [13] in our figure. We added an additional 4 genes from to the refuted category as the variants associated to the diseases have been found to be to common [8,12]. We assigned a score of 1 to the “refuted” genes, 2 to “limited” genes, 3 to “moderate” genes, 4 to “strong” genes and 5 to “definitive” genes. Next we calculated the spearman-rank correlation between these values and the prioritization Z-scores for the corresponding genes (figure 3b).

## Acknowledgements

This work is supported by a grant from the European Research Counsil (ERC Starting Grant agreement number 637640 ImmRisk) to Lude Franke and two VIDI grants (917.14.374 and 917.16.455) from the Netherlands Organisation for Scientific Research (NWO) to Lude Franke and Morris Swertz. This work was supported by BBMRI-NL, a research infrastructure financed by the Dutch government (NWO 184.021.007). Wouter P. te Rijdt is supported by Young Talent Program (CVON PREDICT) grant 2017T001 from the Dutch Heart Foundation. Netherlands Heart Institute, Utrecht, the Netherlands.

We are grateful for the participation of the patients and their parents in this study. We thank Kate Mc Intyre for editing the manuscript and Marieke Bijlsma, Gerben van der Vries, Sido Haakma and Pieter Neerincx for support with the computational analyses. This work was carried out on the computer cluster of the Genomics Coordination Center, hosted at the University of Groningen Center for Information Technology (Strikwerda, W. Albers, R. Teeninga, H. Gankema and H. Wind) and Target storage (E. Valentyn and R. Williams). Target is supported by Samenwerkingsverband Noord Nederland, the European Fund for Regional Development, the Dutch Ministry of Economic Affairs, Pieken in de Delta and the provinces of Groningen and Drenthe.

## Author contributions

PD, SD, JH & LF wrote the manuscript. JK, HB, KA, CD, PZ, EG, PA, JJ, CR, RS, BS, WK, MS, LF edited the manuscript. LF conceived the method. PD, SD, JK, LF developed the statistical methods. PD, SD, JK, HB, PF, TG, LF wrote the software. PD, JH, KA, CD, PZ, EG, KV, RK, PA, SJ, EH, WR, YV, JJ, CR, RS, BS, WK, EZ, JB processed, analyzed and interpreted the solved and unsolved cases.

## Additional material

The following supplements are available with the on-line version of this paper.

Methods S1: Processing and quality control of public RNA-seq data
Methods S2: Comparing GADO and Exomiser on cases with known disease genes
Methods S3: Using alternative tools to prioritize the candidate genes found using GADO
Methods S4: Variant calling and processing of benchmark samples
Methods S5: GeneNetwork website
Figure S1: Selection of parent HPO term if GADO does not have significant predictive power for query term.
Figure S2: Performance of disease gene prioritization compared to random permutation.
Figure S3: The prioritization Z-score when using a maximum of 5 random HPO terms to predict known diseases genes are strongly correlated to using all annotated HPO terms.
Figure S4: Correlation between the GADO prioritization Z-scores and the ExAC missense constraint.
Figure S5: Comparison of GADO performance with the level of evidence for each cardiomyopathy-related gene.
Figure S6: Including 10% random genes when predicting HPO-terms has a marginal effect on prediction accuracy
Figure S7: Rank of the known causative gene among the candidate disease causing variants.
Figure S8: Correcting for biases in co-expression networks.
Figure S9: Histogram of the gene types included in our analyses.
Figure S10: PCA plot of 36,761 samples.
Figure S11: Investigation of principal components capturing technical biases.
Figure S12: Variance explained by first 1,588 PCs.
Figure S13: Visualization of PC1 to PC 10 of PCA over gene correlation matrix.
Figure S14: Outlier genes in PC 8 and PC 9 of PCA over gene correlation matrix.
Figure S15: PC sample scores to distinguish different tissues.
Figure S16: Outlier samples in PC sample scores of PC 8 and PC 9.
Table S1: A list of samples annotated in the European Nucleotide Archive June 30, 2016.
Table S2: OMIM disease gene relations, prioritization Z-score and predictability scores.
Table S3: A list of 83 diagnosed patients with Mendelian disorders and corresponding predictions with GADO.

Table S4: Comparison between GADO and Exomiser predictions using a list of 83 diagnosed patients with Mendelian disorders.
Table S5: A list of 61 undiagnosed patients with suspected Mendelian disorders.
Table S6: Prioritization performance comparison in unsolved cases

## References

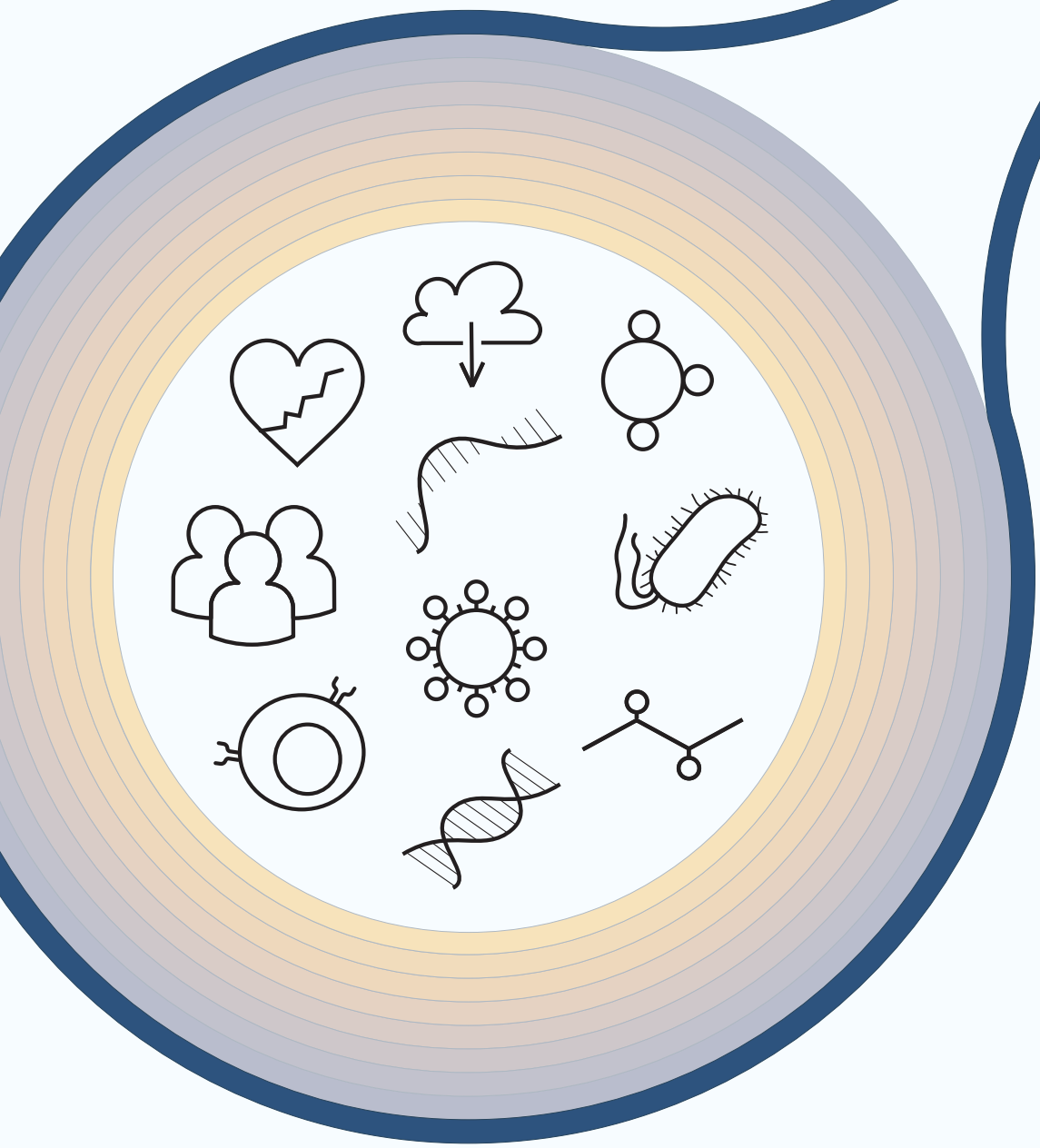

1. Brown, T. L. & Meloche, T. M. Exome sequencing a review of new strategies for rare genomic disease research. *Genomics* **108,** 109–114 (2016).
2. Wright, C. F., FitzPatrick, D. R. & Firth, H. V. Paediatric genomics: diagnosing rare disease in children. *Nat. Rev. Genet.* **19,** 253–268 (2018).
3. Smedley, D. & Robinson, P. N. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Med.* **7,** 81 (2015).
4. Eilbeck, K., Quinlan, A. & Yandell, M. Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.* **18,** 599–612 (2017).
5. Birgmeier, J. *et al.* AMELIE accelerates Mendelian patient diagnosis directly from the primary literature. *bioRxiv* 171322 (2017). doi:10.1101/171322
6. Bone, W. P. *et al.* Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genet. Med.* **18,** 608–17 (2016).
7. Feiglin, A., Allen, B. K., Kohane, I. S. & Kong, S. W. Comprehensive Analysis of Tissue-wide Gene Expression and Phenotype Data Reveals Tissues Affected in Rare Genetic Disorders. *Cell Syst.* **5,** 140–148.e2 (2017).
8. Nouhravesh, N. *et al.* Analyses of more than 60,000 exomes questions the role of numerous genes previously associated with dilated cardiomyopathy. *Mol. Genet. genomic Med.* **4,** 617–623 (2016).
9. Shah, N. *et al.* Identification of Misclassified ClinVar Variants via Disease Population Prevalence. *Am. J. Hum. Genet.* **102,** 609–619 (2018).
10. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
11. Tarailo-Graovac, M., Zhu, J. Y. A., Matthews, A., van Karnebeek, C. D. M. & Wasserman, W. W. Assessment of the ExAC data set for the presence of individuals with pathogenic genotypes implicated in severe Mendelian pediatric disorders. *Genet. Med.* **19,** 1300–1308 (2017).
12. Wright, C. F. *et al.* Assessing the pathogenicity, penetrance and expressivity of putative disease-causing variants in a population setting. *bioRxiv* 407981 (2018). doi:10.1101/407981
13. Strande, N. T. *et al.* Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. *Am. J. Hum. Genet.* **100,** 895–906 (2017).
14. Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47,** 115–125 (2015).
15. Köhler, S. *et al.* The Human Phenotype Ontology in 2017. *Nucleic Acids Res.* **45,** D865–D876 (2017).
16. Leinonen, R. *et al.* The European Nucleotide Archive. *Nucleic Acids Res.* **39,** D28-31 (2011).

17. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34,** 525–527 (2016).
18. Deelen, P. *et al.* Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med.* **7,** 30 (2015).
19. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS One* **9,** e78644 (2014).
20. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46,** D649–D655 (2018).
21. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45,** D353–D361 (2017).
22. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45,** D331–D338 (2017).
23. McKusick-Nathans Institute of Genetic Medicine & Johns Hopkins University. Online Mendelian Inheritance in Man, OMIM. Available at: https://omim.org/.
24. Kurian, M. A. *et al.* Homozygous loss-of-function mutations in the gene encoding the dopamine transporter are associated with infantile parkinsonism-dystonia. *J. Clin. Invest.* **119,** 1595–603 (2009).
25. Puffenberger, E. G. *et al.* Genetic Mapping and Exome Sequencing Identify Variants Associated with Five Novel Diseases. *PLoS One* **7,** e28936 (2012).
26. Kurian, M. A. *et al.* Clinical and molecular characterisation of hereditary dopamine transporter deficiency syndrome: an observational cohort and experimental study. *Lancet Neurol.* **10,** 54–62 (2011).
27. The Gtex Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–5 (2013).
28. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980–D985 (2014).
29. van der Velde, K. J. *et al.* GAVIN: Gene-Aware Variant INterpretation for medical sequencing. *Genome Biol.* **18,** 6 (2017).
30. Arimura, T. *et al.* Structural analysis of obscurin gene in hypertrophic cardiomyopathy. *Biochem. Biophys. Res. Commun.* **362,** 281–287 (2007).
31. Marston, S. *et al.* OBSCN Mutations Associated with Dilated Cardiomyopathy and Haploinsufficiency. *PLoS One* **10,** e0138568 (2015).
32. Bolling, M. C. *et al.* Generalized Ichthyotic Peeling Skin Syndrome due to FLG2 Mutations. *J. Invest. Dermatol.* (2018). doi:10.1016/j.jid.2018.01.038
33. Alfares, A., Al-Khenaizan, S. & Al Mutairi, F. Peeling skin syndrome associated with novel variant in *FLG2* gene. *Am. J. Med. Genet. Part A* **173,** 3201–3204 (2017).
34. Alazami, A. M. *et al.* Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families. *Cell Rep.* **10,** 148–61 (2015).
35. Runge, J. S., Raab, J. R. & Magnuson, T. Identification of Two Distinct Classes of the Human INO80 Complex Genome-Wide. *G3 (Bethesda).* **8,** 1095–1102 (2018).
36. Meeson, A. P. *et al.* Adaptive mechanisms that preserve cardiac function in mice without myoglobin. *Circ. Res.* **88,** 713–20 (2001).

37. van der Harst, P. *et al.* 52 Genetic Loci Influencing Myocardial Mass. *J. Am. Coll. Cardiol.* **68,** 1435–1448 (2016).
38. Truszkowska, G. T. *et al.* Homozygous truncating mutation in NRAP gene identified by whole exome sequencing in a patient with dilated cardiomyopathy. *Sci. Rep.* **7,** 3362 (2017).
39. Thiel, C. T. *et al.* Severely incapacitating mutations in patients with extreme short stature identify RNA-processing endoribonuclease RMRP as an essential cell growth regulator. *Am. J. Hum. Genet.* **77,** 795–806 (2005).
40. Mukawa, K. *et al.* Lymphocyte Antigen 75 Polymorphisms Are Associated with Disease Susceptibility and Phenotype in Japanese Patients with Inflammatory Bowel Disease. *Dis. Markers* **2016,** 1–7 (2016).
41. Kim, S.-E. *et al.* Genome-wide analysis identifies colonic genes differentially associated with serum leptin and insulin concentrations in C57BL/6J mice fed a high-fat diet. *PLoS One* **12,** e0171664 (2017).
42. Tranchevent, L.-C. *et al.* Candidate gene prioritization with Endeavour. *Nucleic Acids Res.* **44,** W117–W121 (2016).
43. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37,** W305–W311 (2009).
44. Benson, M. D. Inherited amyloidosis. *J. Med. Genet.* **28,** 73–8 (1991).
45. Posafalvi, A. *et al.* Clinical utility gene card for: dilated cardiomyopathy (CMD). *Eur. J. Hum. Genet.* **21,** (2013).
46. Sobreira, N., Schiettecatte, F., Valle, D. & Hamosh, A. GeneMatcher: A Matching Tool for Connecting Investigators with an Interest in the Same Gene. *Hum. Mutat.* **36,** 928–930 (2015).
47. Posey, J. E. *et al.* Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *N. Engl. J. Med.* **376,** 21–31 (2017).
48. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45,** D362–D368 (2017).
49. Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J. & Pickard, B. S. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics* **6,** 55 (2005).
50. Moreau, Y. & Tranchevent, L.-C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.* **13,** 523–536 (2012).
51. Collado-Torres, L., Nellore, A. & Jaffe, A. E. recount workflow: Accessing over 70,000 human RNA-seq samples with Bioconductor. *F1000Research* **6,** 1558 (2017).
52. Tsui, B. Y., Dow, M., Skola, D. & Carter, H. Extracting allelic read counts from 250,000 human sequencing runs in Sequence Read Archive. *bioRxiv* 386441 (2018). doi:10.1101/386441
53. Genomics England. PanelApp. Available at: https://panelapp.genomicsengland.co.uk.
54. Silvester, N. *et al.* Content discovery and retrieval services at the European Nucleotide Archive. *Nucleic Acids Res.* **43,** D23–D29 (2015).
55. Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47,** 115–125 (2015).

56. Whitlock, M. C. Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* **18,** 1368–1373 (2005).

# General discussion



The application of genetics in medical research and healthcare is developing rapidly. For complex diseases, we now know of thousands of loci that modulate disease risk and for many Mendelian disorders, we now know the causative genetic variants. However, behind these successes there are many things that we do not yet know. We generally do not know by which mechanisms disease risk loci contribute to disease development, let alone how to combat these adverse effects. It is also often unclear what the relevant environmental factors are and how they contribute to disease development. For Mendelian disorders, despite all the progress in sequencing technologies and variant interpretation, we are still unable to provide a genetic diagnosis for 30% to 92% of cases, depending on disease type [1]. In this final chapter, I discuss how the work in this thesis contributes to our understanding of complex diseases and of the rare variants underlying Mendelian diseases. Finally, I will share my vision on how genetics can help shape the future of healthcare and how the work in this thesis has contributed to these developments.

## Improving genotype imputation

Initially, HapMap data was used as a reference panel for imputation in GWAS because these samples had been more densely genotyped than the average study data [2]. Later, once the cost of DNA-seq had decreased, the Thousand Genomes Project (1000G) was started. In 1000G a large number of samples were subjected to genome sequencing, and 1000G soon became the standard to use as a reference panel [3]. It was, however, shown that the reliability of genotype imputation improved if the study data had a similar ancestry to the reference panel [4]. This was one of the reasons for BBMRI-NL to initiate the Genome of the Netherlands (GoNL) project [5] in which we performed genome sequencing of 250 parent-offspring families.

In **chapter 2** of this thesis we show how the GoNL reference panel can be used to improve genotype imputation quality of genetic datasets. We show that the GoNL reference panel outperforms the 1000G dataset, not only for samples of Dutch descent, but also for samples from other European populations. We also show that merging the GoNL with the 1000G data into a single, larger panel yields the best results because imputation becomes more accurate if the reference panel is larger. The GoNL reference panel has now been used to impute genotypes for many Dutch biobanks and other cohorts [6–11]. Unfortunately, imputation of rare genetic variants remained relatively unreliable. To overcome this, others have taken the next step of combining as many samples as possible within the Haplotype Reference Consortium (HRC). The HRC has combined 1000G, GoNL and many other genome sequencing datasets into one large reference panel to further improve genotype imputation [12].

In principle, imputation reference panels of one population can be used to impute a different population. This allows imputation of Dutch genetic data using, for instance, the 1000G data, which does not contain Dutch samples. However, if a specific haplotype is unique to a population, or simply more common in that population, it may not be properly represented by the reference panel from a different population. This will result in poor imputation of variants on this haplotype and will hinder downstream analysis. Imputation using a population-specific panel can be used to overcome these limitation, as was shown for a cholesterol GWAS imputed using GoNL [8].

For samples of a different ancestry than that of the reference panel, this problem is more substantial. There will be many unique haplotypes and larger differences in haplotype prevalence. Unfortunately, the current HRC release consists mostly of samples of European decent and is therefore less suited for the imputation of samples from people of other ethnicities. It is, however, important to conduct GWAS in different populations to gain a better understanding of the genetic architecture of a trait or disease and to make sure that polygenic scores can also be used for people who are not of European descent. Currently, most GWASs are conducted on people of European descent, which means that the polygenic scores created using these GWAS data perform poorly in individuals of different ancestries [13]. This problem can be solved by performing more GWAS studies on these other populations. To make sure these new GWAS have optimal power, it is important to extend the HRC with reference haplotypes from all populations to make sure that the variant imputation is reliable for all samples.



When performing genotype imputation, it is also important that the data you want to impute matches the reference data format. For instance, individual genetic variants must be coded using the same genomic strand as the imputation reference panel to be used. If it is known what strand is used for the genotype data, this is an easy adjustment. However, in practice, this is not always immediately evident, which makes imputation more complicated. While some tools did exist to do some automatic adjustment of genomic strand, these required several pre- and post-processing steps. To simplify this process, we developed Genotype Harmonizer, which we presented in **chapter 3**. Genotype Harmonizer works directly on the file formats commonly used in the imputation process and improves the accuracy of strand alignment over other tools. We have used Genotype Harmonizer in several chapters of this thesis, and it is now being used in various ongoing projects.

## Understanding the downstream consequences of genetic risk factors

To shed light on the effect of regulatory genetic variation on different molecular modalities, we performed population-based studies focused on gene expression levels (**chapters 4 to 7**), the effect on cytokine response (**chapter 5**) and DNA methylation (**chapter 6**).

### Effects of genetic variants on gene expression levels across different cell types

Despite having near-identical DNA, the different cells in our body can have completely distinct functions and morphology. This is due to the orchestrated regulation of many different genes, many of them specifically expressed or regulated in certain cell types. Thus, in order to identify the downstream molecular consequences of genetic risk factors, we should ideally study all the relevant tissues and cell types per disease or trait. Unfortunately, the majority of the expression quantitative trait locus (eQTL) mapping studies that investigate the downstream effects of genetic risk factors on, for instance, gene expression levels are conducted on easily obtainable cell material such as whole blood or specific blood cell types. While this has greatly helped to identify downstream effects for some genetic risk factors, we know that the presence and magnitude of these eQTLs can be context-dependent and differ among tissues and cell types [14].

We partially overcame this in **chapter 4** of this thesis, where we downloaded all publicly available RNA sequencing (RNA-seq) data that we could access in an attempt to investigate *cis*-eQTL effects that are not found in blood-related datasets. This was possible because we were able to derive genotypes for these samples based on the RNA-seq data (Figure 1). We show that it is feasible to integrate RNA-seq from many different sources and to detect tissue-specific eQTLs that are not detected in blood using genotypes derived from the RNA-seq data. This is important because these tissue-specific eQTLs overlap with genetic risk factors that could not be interpreted using blood eQTLs.

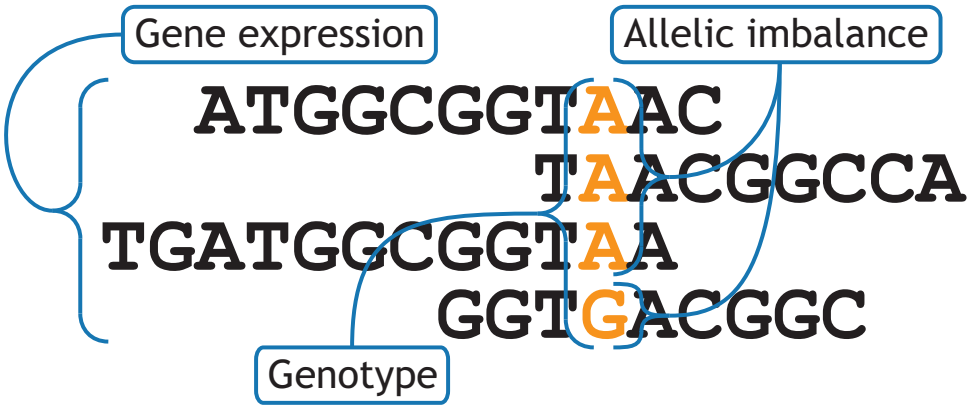



***Figure 1: RNA sequencing.*** *Since RNA molecules are a copy of the DNA, they also contain the variation present in the genome. That means that information on the genotypes of a sample is also present in the reads obtained by RNA-seq, which makes it possible to call variants based on the RNA-seq reads on top of the expression quantification for which RNA-seq is normally used. Additionally, by assessing allelic imbalance, it is also possible to assess which of the two copies of a gene is more abundantly expressed.*

The Genotype-Tissue Expression (GTEx) project also aimed to perform eQTL studies in different tissues and collected gene expression data of 44 different tissues from the same individuals. This resulted in new insights into the downstream gene expression consequences of genetic risk factors across many tissues [15]. However, while the GTEx is a tremendous resource, it too is limited in sample size, and this limits its power to detect genetic variants that affect gene expression levels on other chromosomes (*trans*-eQTL effects) because these effects are typically very small [16]. As public RNA-seq repositories continue to grow exponentially (Figure 2), I expect that public samples will remain valuable as a complement to GTEx for detecting tissue-specific eQTL effects.

## Effects of genetic variants on methylation levels

In addition to using expression data to investigate the downstream effects of genetic risk factors, it is also possible to investigate the downstream effects on DNA methylation, called meQTLs (methylation quantitative trait loci). In **chapter 6** we investigated if genetic risk factors have effects on DNA methylation, and found several examples of risk variants near or in transcription factors (TFs) that altered methylation around or within genes that are

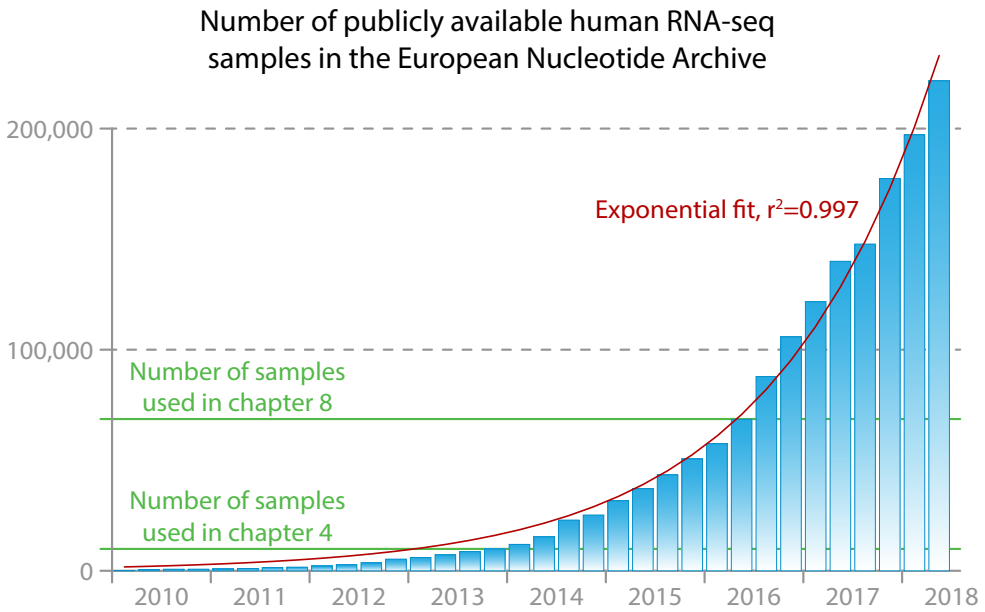



***Figure 2: Growth of publicly available human RNA-seq data.*** *Public RNA-seq repositories are growing at an exponential rate. We had already observed this exponential growth in chapter 4, and the trend continued in subsequent years. This allowed us to use a much larger dataset for chapter 8 and promises many more possibilities for the future.*

known targets of these TFs. It has been shown that TFs can associate with other proteins that can methylate or demethylate DNA, providing a potential biological explanation for the observed phenomenon [17]. Interestingly, not all known TF binding sites show altered methylation and, while this might be due to technical or power limitations, it could also be that only a subset of targets of these TFs are affected by these risk factors. It might be that these meQTL are actually cell-type-specific and that it is only the TF targets active within this cell type that show altered methylation levels. Alternatively, it could be that the methylases or demethylases are not effective at all places where the TFs bind. However, while these questions remain to be resolved, we have shown that *trans*-meQTL mapping provides another strategy to identify the molecular downstream consequences of genetic variation.

## Understanding the interplay of genetic risk factors and environmental exposure

We know that, next to genetic factors, environmental factors such as food intake, lifestyle, pollution and pathogens play an important role in the development of many complex diseases. The effect of environment on disease predisposition can be mediated by genetic variation. For instance, in the case of celiac disease, the stimulation of gluten in combination with specific variants in human leukocyte antigen region is a major trigger of disease development. However, when investigating the role of genetic variants on gene expression and DNA methylation in **chapters 4 & 6,** we did not take such interaction effects into account.

For the public RNA-seq data, we simply did not have any environmental information available. For the BIOS data used to map the meQTL, we had some information available that could, in principle, be used to test for interactions between meQTL effects and environmental factors. These analyses, however, are non-trivial. Not only do they require tremendous statistical power and complex models, the interpretation is often very difficult as many environmental factors are correlated to each other. Instead of using natural occurring stimulations, it is also possible to perform a systematic analysis by directly stimulating cell cultures and measuring the effect. Such perturbations are expensive and time consuming as the cells need to be cultured and profiling needs to be done before and after the stimulations.

## Effects of genetic variants and pathogenic stimulations on cytokine levels

In **chapter 5** we used *in vitro* stimulations to investigate how the interplay between genetic variants and stimulation with pathogens affects cytokine abundance. We found several examples of cytokine QTLs (cQTLs) that are only observed upon pathogenic stimulations. Despite the limited power to detect significant cQTLs due to the limited number of available samples (n=197), we did find an enrichment for cQTLs among loci associated to infectious diseases and autoimmune disorders. We found that infectious disease risk alleles mostly lower the cytokine response, and this is in contrast to what we saw for the autoimmune loci, which do not favor a specific effect direction. This shows how stimulations can reveal the altered regulatory consequences of genetic variation. However, it also shows that many more samples are needed to fully map the regulatory response to stimuli.

## Effects of genetic variants and environmental stimulation on gene expression levels

*In vitro* perturbation studies, where cells are exposed to different stimuli, have shown that the strength at which genetic variation can affect gene expression levels can be strongly dependent on external stimulations [18]. Unfortunately, it is not feasible to perform such experiments on a large number of samples. In **chapter 7** we partially overcome this by using gene expression levels as a proxy for environmental stimulations and ascertaining whether these can mediate the effect size of eQTLs. The principle is as follows: since we know that external stimulations such as infections alter the gene expression levels of specific genes, these genes can be used, at least to some extent, to measure the stimuli to which individual samples have been exposed. We therefore developed a computational method that allowed us to identify the 10 largest factors that modulate eQTL effect strength. We found that the biggest influencers of eQTLs was the compositions of blood cell-types, as expected the eQTLs modulated by these influencers showed high overlap with known cell-type-dependent eQTLs. This also allowed us to identify cell-type-dependent eQTLs for cell types for which no purified eQTL data is available, such as erythrocytes. In addition, we also found a large group of eQTLs that reflect type 1 interferon genes whose effect sizes were strongly affected by genes that are proxies for type 1 interferon response, suggesting that these eQTLs are modulated by viral stimulation. This was confirmed when we compared eQTL data of *in vitro* rhinovirus stimulated cells and observed a significant overlap with our results.

An additional benefit of our method is that we were able to correct for the major influencers of eQTL effects. This enabled the detection of more subtle modulators of eQTL effects, which allowed for the dissection of regulatory networks. We found examples where the

abundance of specific TFs, those that are known to bind at the DNA where the eQTL SNP resides, influences the effect that a genetic variant has on expression levels of the target gene. While it makes conceptual sense that some eQTL variants work by influencing TF binding efficiency, and that these eQTLs are only active if the TF is present, our method allowed us to actually identify eQTLs where this happens. This also aids in the construction of regulatory networks as this information allows genes to be placed upstream or downstream of each other within a regulatory network.

Even though it is possible to detect the effect of external factors on regulation using expression levels as proxies, we were unable to link all these effects to specific environmental stimuli. It is therefore also worthwhile to collect and harmonize actual environmental information to use directly in the interaction model or to use as an aid in the interpretation of identified effects. Based on our results, future studies should carefully consider if *in vitro* stimulations are necessary or if it is more sensible to measure a specific exposure in a population cohort or use a proxy measurement for an environmental factor. An intermediate design might also be sensible where the effects are first discovered using population-based cohorts and then validated using *in vitro* perturbations. Fewer samples would then be needed for the *in vitro* experiment since it is used to replicate the identified associations.



## Future perspectives

### Future directions to gain better insight into the downstream consequences of genetic variants and environmental stimuli

Knowledge about the effects of stimulations is important when studying complex diseases because we know that genetic variation usually explains only a part of disease risk, and that there is a large environmental component [19]. However, it is often not clear which environmental stimuli affect disease development, or in which direction and by which biological mechanisms they do so. I hypothesize that some of these environmental effects on complex diseases work by mediating the eQTLs of genetic risk factors for disease. In other words, genetic risk variants might influence how genes are regulated, but only do so to the full extent if specific environmental factors are present or absent. Knowledge of which factors influence eQTLs might then help to identify which of these factors are related to disease development. Hopefully this will eventually reveal new ways to prevent complex diseases not only by controlling for currently unknown environmental factors, but also by identifying drugs that shield the eQTL effects of genetic risk factors from the environmental stimulations needed for disease development.

We have shown that it is possible to identify modulators such as cell type composition and stimulation differences in whole blood samples, even though our study only had limited power to detect these effects. Ideally, we should increase the number of samples for which we have gene expression measurements combined with other molecular measurements, phenotypic information, and lifestyle/environmental information, so that we do not have to rely on genes whose expression levels are a proxy for the environmental exposures. This is now becoming a reality, as many biobanks are currently collaborating on large-scale eQTL analyses, which enables us to start studying this in over 32,000 samples [20]. I expect that

large, particularly longitudinal, biobanks will be instrumental for these studies because longitudinal information can provide insight into the effects of environmental exposures on participants prior to disease development.

Complementary to the large-scale bulk expression studies, single-cell-expression quantification is also expected to yield new and unique insights [21]. By profiling each cell separately, it is possible to obtain the expression per cell type, which allows for the identification of cell-type-specific eQTLs [22]. This paves the way to further investigate how environmental exposures affect the magnitude of eQTLs in a cell-type-dependent manner.

In principle it is also possible to directly test for interaction of genetic variants and environment on disease risk. However, because these interaction effects are often small, it is very difficult to reliably detect them. Because the effects of variants on gene expression or other molecular traits are typically more profound, environmental interactions are easier to detect. I expect that future interaction studies on diseases will specifically test interaction already identified on a molecular level. This might also reveal new disease-associated variants that would normally have gone unnoticed when only the main effects for these variants for association with disease would have been studied.

However, for any experiment studying the effect of stimulations, we should realize that the samples used are already the product of a lifetime of different stimulations. This might seem like a hypothetical limitation, but it is possible, for instance due to bistability of regulatory networks [23] (Box 1), that the effect of a stimulation remains years after the actual stimulation has dissipated (e.g. prior viral infections that have triggered the immune system). Even *in vitro* perturbations can be confounded by these past stimulations. It is therefore important to realize that all humans are unique; on top of our genetic differences we all have a distinct history of environmental influences. So, even if it would be feasible to try all possible stimulations in all possible combinations, we will at some point run into the limitations of what can be done by association studies since other humans will not always be informative for a specific individual. This means that, on top of large cohorts with stimulations and association analysis, we also need to make sure that we understand the mechanisms behind regulation and how these mechanisms are affected by external factors.

## Perspectives on applying genetic information in the diagnostic process

In current diagnostic practice, the focus for patients mainly lies in physical examination and a compendium of different scans and assays, and molecular genetic testing is only performed if a Mendelian disorder is suspected at some point during the diagnostic process. With the current decline in cost and increase in knowledge about genetic testing, I expect that genetics will gain more emphasis and be used earlier in the diagnostic process. We already see this trend in very specific cases. For instance, rapid sequencing is now conducted as an early diagnostic test for severely ill newborns who are admitted to the intensive care [24] because it can help assist in diagnosis and has proven relatively cost-effective [25].

Often, a known pathogenic variant is identified in a gene that we know is linked to phenotypic presentation of a patient. However, providing a genetic diagnosis for a suspected Mendelian disorder is still often difficult. It can be difficult to draw a definitive conclusion, for instance, if a patient has a previously undescribed variant or an atypical phenotypic

***Box 1: Bistability of regulatory networks.*** *A simple regulatory network will respond to a specific input, but will return to its original state after these inputs return to normal. In the case of bistability, a network can get stuck in an alternative state even if the original input that triggered the change no longer exists.*



presentation for a known disease gene harboring the variant. Even more difficult are cases where the causative gene has not yet been linked to a disease. Typically, exome sequencing yields approximately 875 candidates variants that might be relevant to disease development [26]. Although many strategies exist to filter this list further, final interpretation of the identified candidate gene/variant remains time-consuming. In **chapter 8** we presented GADO, a method that can aid in this process by using public RNA-seq data (Figure 2). Our GADO method predicts the most likely phenotypic consequences of genes and can rank the genes with candidate variants based on the phenotypic features of patients. In chapter 8, we show how this can optimize the diagnostic process and aid in the discovery of new disease genes.

To expand the future role of genetic testing within diagnostics, we need to understand how a variant can affect a person's phenotype even when the variant has never been observed before. It is important here to realize that Mendelian diseases and complex diseases are not distinct classes, but rather a continuum [27]. The variants that underlie Mendelian diseases simply have a very high penetrance, and this does not preclude there being other variants that can modulate the severity of the disease or even, in rare situations, provide rescue from it [28,29]. The reason that we are unable to provide a genetic diagnosis for some cases might be because the disease is driven by a combination of a few variants with incomplete penetrance or by variants that modulate penetrance [30]. For instance, this may be due to a non-coding variant that upregulates the disease-causing allele, resulting in a disease that normally is considered recessive (Figure 3). Increasing our fundamental knowledge of how gene regulation works can aid in diagnosing patients with these more complicated forms of rare diseases.

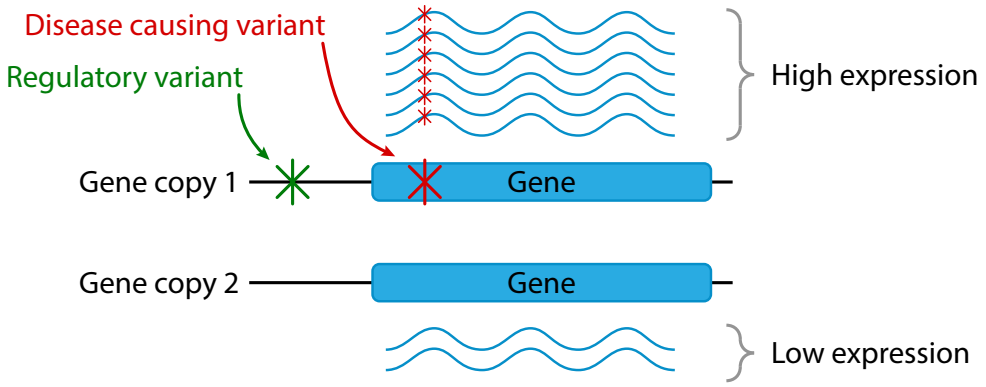



***Figure 3: Concept of a regulatory variant affecting the penetrance of a pathogenic variant.*** *Normally people have two copies of a gene, one from each parent. In the case of recessive diseases, both genes need to be damaged by a pathogenic variant before a person gets sick. If only one gene copy is damaged, that person is only a carrier of a disease. There are many scenarios imaginable in which regulatory variants might modify the penetrance of disease-causing variant. Here we show a scenario where a person is carrier of a variant that causes a recessive disease. Normally this person would not be affected by the disease, however there is also a regulatory variant near the gene copy with the disease-causing variant. In this scenario, this regulatory variant causes upregulation of the gene. The increase in expression of the gene copy affected by this regulatory variant might normally be perfectly benign. In this case, however, it results in higher expression of the damaged gene. It is thereby possible that this causes the disease, or a milder version of it.*

## Proliferation of large-scale genotyping in general health care

I envision that in the near future genetic information will become more and more important in all areas of healthcare, and the challenge ahead of us is how to enable clinicians to use all this information. Traditionally, genetic testing is driven by medical complaints and the relevance of the variants identified is weighted in light of a patient's phenotype. Interpretation of a genetic profile without a specific medical question is much more difficult. This is partially due to the fact that many variants have been implicated in Mendelian diseases, but not all of them are fully penetrant [31,32]. This means that if you find the implicated variant in a patient with the disease, then that variant it is very likely causing the disease. However, if you find the same variant in a healthy individual it does not necessarily mean that this person will get the disease. This means that other, possibly genetic, factors are also relevant to disease development, thus complicating pre-symptomatic variant interpretation.

Despite these challenges, there are now many examples that show the added value of screening healthy individuals. One good example is familial hypercholesterolemia (FH). If left untreated, FH can lead to Coronary Artery Disease (CAD), which in turn may lead to heart attacks as well as other health consequences. FH has an estimated prevalence of 1 in 256, and a study in the United States found that only 15% of carriers were diagnosed with FH prior to genetic testing. Also, only 24% of the carriers of FH variants met the clinical criteria for a probable or definite FH diagnosis [33]. This shows that many individuals who

have FH or are at risk for FH do not receive treatment or screening as they are unaware of their disease. Population-based genetic screening can help identify the carriers of variants implicated to be causative for FH so that the lipid levels of these carriers can be regularly tested and, if needed, controlled by medications. Many other actionable genes exist, and the American College of Medical Genetics and Genomics (ACMG) maintains a list of them [34]. An on-going study on routine genomic screening estimates that 3.5% of the volunteers will receive results for one of the 76 actionable genes tested within this project.

Genetic profiles can also be used to stratify patients into high and low risk for developing a complex disease [35]. This can be performed by calculating polygenic scores that sum up a set of variants using weights derived from a previously conducted GWAS. Individuals with a high score might benefit from additional screening or even pre-symptomatic treatment. An interesting example here is again CAD, which can also be the result of a complex genetic cause, rather than FH [36]. It turns out that, based on the polygenic scores, it is possible to identify individuals with a risk for CAD that is similar to that of people with the rare variants associated with FH. However, the chance of having such a high polygenic score for CAD is 20-fold higher than the chance of having an FH variant. Since these high-risk individuals would benefit from the same lipid lowering medication as FH variant carriers, the potential health benefits of these polygenic risk scores is very clear.



Another obvious application of genetic screening is pharmacogenetics [37]. It is well established that people respond differently to drugs: a drug might not be effective for all cases and some people have severe adverse reactions to specific drugs. Drug usage and drug dosage therefore need to be tailored to patients. Using genetic profiles, it is possible to predict the effectiveness, the optimal dosage, and even the severity of the side effects [38].

By overlapping genetic profiles of prospective parents, it is also possible to identify risk for future children. Through this preconception carrier screening it is possible to identify diseases for which both parents are carriers [39]. If both parents are carriers of a recessive disease, then children will have a 25% chance of having the disease. Roughly 1 in 600 children is born with a lethal recessive disease for which no treatment is available [40]. By identifying this risk prior to conception, *in vitro* fertilization in combination with embryo selection or genetic testing of the fetus can prevent needless suffering.

At some point in the future, genetic screening might simply become part of the neonatal heel prick screening. This screening already tests for a set of predetermined diseases that benefit from early interventions. However, this does raise the ethical question of whether we want to burden children or parents with knowledge about variants that give high risk for late-onset diseases that won't manifest for decades and don't require intervention during childhood or early adolescence. This is a discussion that has been summarized by the ACMG and they currently recommend to always report these variants regardless of age [34].

Ideally, we would use genome-sequencing or at least exome-sequencing to generate these genetic profiles. However, as discussed in the introduction, these sequencing techniques are currently relatively expensive and their wide-spread use would impose a huge burden on the healthcare budget. Initially, a more pragmatic approach using genotyping chips might allow wider application of genetic profiles even though these will be less powerful. Genotyping chips such as the Global Screening Array and UK Biobank Axiom Array contain many

variants associated to Mendelian diseases and will also enable the creation of genetic risk profiles. However, it is then important to realize that these chips can never cover all disease-associated variants. This will mean that not all cases of, for instance, FH will be picked up. Preconception screening will also be less reliable when using genotyping chips compared to sequencing techniques. However, if using the cheaper chips will enable screening of more individuals and more couples, then the net effect on healthcare of using chips could be larger compared to sequencing on a smaller scale.

Clearly, since genetic data is very privacy sensitive [41], major hurdles need to be overcome before we can use genetic data more widely. We need to make sure that we properly account for the legal and privacy issues, and we need technical solutions to safely store the data while maintaining accessibility for the parties that need to use the data. Given that current medical information is equally sensitive and there are solutions for dealing with it, I am confident that this will also be possible for genetic data in the future. It is here important to note that there has been public resistance to electronic health records, and it is likely that not everyone will be comfortable with electronic genetic records. We should therefore make sure to emphasize to everyone that genetic profiling is optional, that a patient can choose which medical providers have access to their data (just as with medical records), and that patients will always have the right to have their data removed.

For research purposes, several parties are already successfully sharing genetic data in central databases [42–45] and, in a more clinical settings, initiatives such as the variant sharing of the Dutch *Vereniging Klinisch Genetische Laboratoriumdiagnostiek* aim to improve genetic diagnostics of rare diseases [46]. The system I envision would enable a bigger role for genetics in daily medical practice (Figure 4). While I think that this data could be stored in a central system, this does not mean that anyone can access this data. It is not even necessary for all medical practitioners to have direct access to the genotype data or have deep knowledge about genetics. Medically relevant information derived from genetic data can be shared using clinical decision support (CDS) systems [47]. A general practitioner could query such a CDS for a list of his patients for whom regular lipid screening is recommended, for instance because they have FH or a high polygenic score for CAD. In another application, a pharmacist could get an automatic warning if a prescribed drug is incompatible with the patient's genotype.

Interestingly, applying large-scale genotyping approaches can substantially boost the added value of genotyping. All patients who are subjected to genotyping could be asked if this data can be used for research and if this can be linked to medical records or existing biobank participation. Life-science researchers can use such a system to greatly expand our knowledge of genetics in a completely anonymous manner [48]. We can use it to evaluate current predictive models and identify new ways to predict disease development and treatment outcome. Naturally an opt-in or opt-out system is needed for each of these applications to accommodate individual's preferences and convictions.

Although it is not yet feasible to offer genotyping to everyone, population-based biobanks that perform genotyping on their participants continue to grow. For instance, the UK Biobank now has genetic information available for 500,000 participants. We are also now taking the next step with the Lifelines biobank [6]. Lifelines is a longitudinal study of 167,000

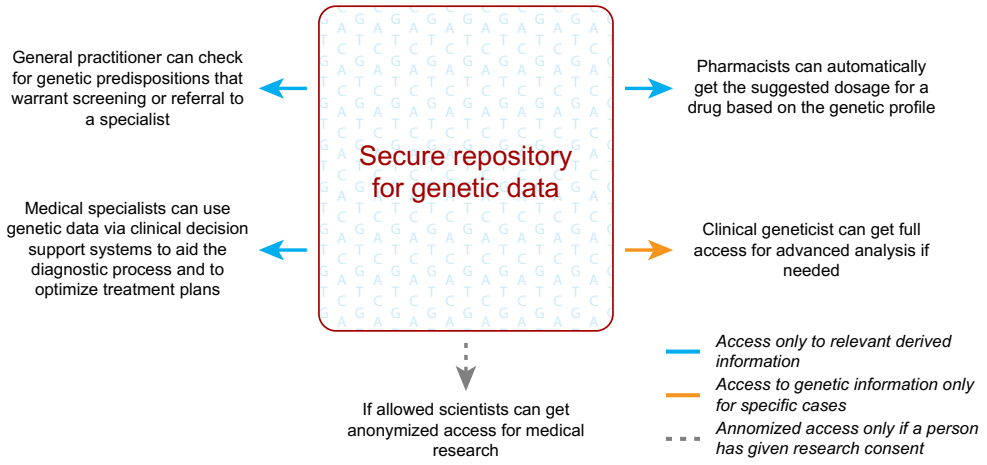



***Figure 4: A possible system to store and use genetic data of patients.*** *Genetic data generated from a patient can be stored in a secure location. Even though the data is stored in a central system, the patient remains in control of his own data and can choose who can access his genotype information.*

individuals living in the three northern provinces of the Netherlands. Roughly 15,000 participants have been genotyped and current efforts aim to genotype an additional 60,000 individuals using the Global Screening Array (GSA) [49]. With the GSA it is possible to genotype many clinically relevant markers, and it allows accurate imputation of low-frequency alleles. By using the vast collection of phenotypic, lifestyle and medication information available for the Lifelines participants, we expect to gain new insight into how the effects of genetic risk factors are modulated by external factors.

### The next steps in personalized healthcare

On top the genetic information, there are specific cases in which it may be necessary to measure molecular phenotypes directly in a patient [50]. Examples of this include using RNA-seq to prioritize genes with aberrant expression or splicing to aid in the diagnostic process of Mendelian diseases [51,52], using metabolomics to diagnose mitochondrial diseases [53], and using gut metagenomics to reveal biomarkers that discriminate between ulcerative colitis and Crohn's disease [54]. A major advantage of molecular phenotypes is that they also reflect environmental exposures. With the exception of patterns in acquired mutations, this is not captured by genetic screens. Additionally, molecular phenotyping will be essential for personalized medicine for the more complex and rarer cases in which knowledge gained from biobanks is less informative, although it is important to note that these methods still rely heavily on biobanks to characterize "normal variation" in the population.

## Final remarks

I think the scientific community is making great progress in understanding the human genome and applying this knowledge to improve healthcare. Although it might seem that we are taking only baby steps each time, and making only small improvements, taking all the

research combined, we have vastly improved our knowledge and capabilities. In the coming years, I expect that the increase in sample sizes will yield insights into many more subtle effects. These larger cohorts will also enable research on ever rarer variants and on interactions between genetic variation and lifestyle and environment. In this thesis, we have already shown in **chapter 8** how the more fundamental research of the earlier chapters can find its way to become added value for clinical practice. Over the coming decades I expect that genetics will have an increasingly important role in medicine as we gain more and more insight into the workings of our genome.

## References


1. Wright, C. F. *et al.* Paediatric genomics: diagnosing rare disease in children. *Nat. Rev. Genet.* **19,** 253–268 (2018).
2. International HapMap Consortium, T. I. H. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–61 (2007).
3. Gibbs, R. A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
4. Howie, B. *et al.* Genotype imputation with thousands of genomes. *G3 genes - genomes - Genet.* **1,** 457–70 (2011).
5. The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46,** 818–825 (2014).
6. Scholtens, S. *et al.* Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int. J. Epidemiol.* **44,** 1172–80 (2015).
7. van der Lee, S. J. *et al.* PLD3 variants in population studies. *Nature* **520,** E2-3 (2015).
8. van Leeuwen, E. M. *et al.* Genome of the Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels. *Nat. Commun.* **6,** (2015).
9. de Vries, P. S. *et al.* Genetic variants in the ADAMTS13 and SUPT3H genes are associated with ADAMTS13 activity. *Blood* **125,** (2015).
10. Atanasovska, B. *et al.* GWAS as a Driver of Gene Discovery in Cardiometabolic Diseases. *Trends Endocrinol. Metab.* **26,** 722–732 (2015).
11. Shah, S. *et al.* Improving Phenotypic Prediction by Combining Genetic and Epigenetic Associations. *Am. J. Hum. Genet.* **97,** 75–85 (2015).
12. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48,** 1279–1283 (2016).
13. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100,** 635–649 (2017).
14. Peters, J. E. *et al.* Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease. *PLOS Genet.* **12,** e1005908 (2016).
15. GTEx Consortium *et al.* Genetic effects on gene expression across human tissuesGTEx. *Nature* **550,** 204–213 (2017).
16. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).
17. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science (80-. ).* **356,** eaaj2239 (2017).

18. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343,** 1246949 (2014).
19. Visscher, P. M. *et al.* Heritability in the genomics era — concepts and misconceptions. *Nat. Rev. Genet.* **9,** 255–266 (2008).
20. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv* 447367 (2018). doi:10.1101/447367
21. Papalexi, E. *et al.* Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* (2017). doi:10.1038/nri.2017.76
22. Wijst, M. G. P. van der *et al.* Single-cell RNA sequencing reveals cell-type specific cis-eQTLs in peripheral blood mononuclear cells. *bioRxiv* 177568 (2017). doi:10.1101/177568
23. Pfeuty, B. *et al.* The combination of positive and negative feedback loops confers exquisite flexibility to biochemical switches. *Phys. Biol.* **6,** 046013 (2009).
24. van Diemen, C. C. *et al.* Rapid Targeted Genomics in Critically Ill Newborns. *Pediatrics* **140,** e20162854 (2017).
25. Ostrander, B. E. P. *et al.* Whole-genome analysis for effective clinical diagnosis and gene discovery in early infantile epileptic encephalopathy. *npj Genomic Med.* **3,** 22 (2018).
26. Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312,** 1870–9 (2014).
27. Badano, J. L. *et al.* Beyond Mendel: an evolving view of human genetic disease transmission. *Nat. Rev. Genet.* **3,** 779–789 (2002).
28. Holmans, P. A. *et al.* Genetic modifiers of Mendelian disease: Huntington's disease and the trinucleotide repeat disorders. *Hum. Mol. Genet.* **26,** R83–R90 (2017).
29. Priya, S. *et al.* Bardet-Biedl syndrome: Genetics, molecular pathophysiology, and disease management. *Indian J. Ophthalmol.* **64,** 620–627 (2016).
30. Castel, S. E. *et al.* Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat. Genet.* **50,** 1327–1334 (2018).
31. Shawky, R. M. Reduced penetrance in human inherited disease. *Egypt. J. Med. Hum. Genet.* **15,** 103–111 (2014).
32. Tarailo-Graovac, M. *et al.* Assessment of the ExAC data set for the presence of individuals with pathogenic genotypes implicated in severe Mendelian pediatric disorders. *Genet. Med.* **19,** 1300–1308 (2017).
33. Abul-Husn, N. S. *et al.* Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science* **354,** aaf7000 (2016).
34. Green, R. C. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* **15,** 565–574 (2013).
35. Chatterjee, N. *et al.* Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17,** 392–406 (2016).
36. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* 1 (2018). doi:10.1038/s41588-018-0183-z
37. Wang, L. *et al.* Genomics and Drug Response. *N. Engl. J. Med.* **364,** 1144–1153 (2011).
38. Swen, J. J. *et al.* Pharmacogenetics: From Bench to Byte— An Update of Guidelines. *Clin. Pharmacol. Ther.* **89,** 662–673 (2011).

39. Metcalfe, S. A. Carrier screening in preconception consultation in primary care. *J. Community Genet.* **3,** 193–203 (2012).
40. Plantinga, M. *et al.* Population-based preconception carrier screening: how potential users from the general population view a test for 50 serious diseases. *Eur. J. Hum. Genet.* **24,** 1417–23 (2016).
41. Regulation (EU) 2016/679. of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da. *Off. J. Eur. Union* 1–88 (2016). doi:L:2016:119:TOC
42. Leinonen, R. *et al.* The European Nucleotide Archive. *Nucleic Acids Res.* **39,** D28-31 (2011).
43. Joly, Y. *et al.* Analysis of five years of controlled access and data sharing compliance at the International Cancer Genome Consortium. *Nat. Genet.* **48,** 224–225 (2016).
44. Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39,** 1181–6 (2007).
45. Thompson, R. *et al.* RD-Connect: An Integrated Platform Connecting Databases, Registries, Biobanks and Clinical Bioinformatics for Rare Disease Research. *J. Gen. Intern. Med.* **29,** 780–787 (2014).
46. Weiss, M. M. *et al.* Best Practice Guidelines for the Use of Next-Generation Sequencing Applications in Genome Diagnostics: A National Collaborative Study of Dutch Genome Diagnostic Laboratories. *Hum. Mutat.* **34,** 1313–1321 (2013).
47. Overby, C. L. *et al.* Opportunities for genomic clinical decision support interventions. *Genet. Med.* **15,** 817–823 (2013).
48. Jagadeesh, K. A. *et al.* Deriving genomic diagnoses without revealing patient genomes. *Science (80-. ).* **357,** (2017).
49. Illumina. Illumina Announces Initial Customer Orders for the Global Screening Array. (2016). Available at: http://www.illumina.com/company/news-center/press-releases/press-release-details.html?newsid=2178011. (Accessed: 30th December 2016)
50. Karczewski, K. J. *et al.* Integrative omics for health and disease. *Nat. Rev. Genet.* **19,** 299–310 (2018).
51. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9,** eaal5209 (2017).
52. Kremer, L. S. *et al.* Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* **8,** 15824 (2017).
53. Esterhuizen, K. *et al.* Metabolomics of mitochondrial disease. *Mitochondrion* **35,** 97–110 (2017).
54. Pascal, V. *et al.* A microbial signature for Crohn's disease. *Gut* **66,** 813–822 (2017).

# Appendices

## Summary

When comparing individuals, it is clear that there are many places throughout the genome where people can differ from each other. These differences are caused by DNA mutations that have randomly occurred over thousands of generations. Due to the accumulation of these genetic variants, on average, 15 million bases will be different when comparing two humans.

While the functional consequences of most of these genetic differences are benign, some genetic variants can cause a disease or increase the chance of developing a disease and some genetic variants are actual beneficial. For example, when a variant maps inside a gene it might result in a different, non-functioning protein. It is also possible that a variant might have a regulatory effect, for example on the amount of RNA that is being produced, that results in altered proteins levels that might cause problems. Regretfully, it is difficult to predict the functional or detrimental effect of a specific variant. Therefore, for most variants, we currently do not know if they will cause disease. However, thanks to the large-scale execution of Genome Wide Association Studies (GWAS) in the last 14 years, it is now becoming clear that thousands of variants are associated to all kinds of common complex diseases, including autoimmune diseases and neurological disorders.

However, it remains very difficult to interpret these results because of the limited resolution of GWAS - the DNA chips typically used for GWAS usually interrogate fewer than a million genetic variants. Fortunately, it is possible to make inferences about many more variants through a statistical method called 'imputation' that works by using information from an appropriate 'reference dataset'. One such set is the Genome of the Netherlands (GoNL) in which 250 Dutch parent-offspring families were whole-genome sequenced by the BBMRI-NL consortium. In **chapter 2,** I discuss how the GoNL data can be used for this imputation process and show how use of the GoNL dataset can improve the resolution of GWAS in studies that concentrate on European samples. I also show that integrating GoNL data with the existing reference dataset further improves the quality of the imputation process. In **chapter 3,** I describe a software package called Genotype Harmonizer that can facilitate this imputation process.

Another challenge in the interpretation of variants associated to disease is that most variants are located outside of genes. This means that these variants likely have some kind of regulatory effect on a gene or on multiple genes, and that this altered regulation could cause the disease. It is, however, very difficult to predict which genes will be affected by an individual variant. In some cases the gene closest to a variant will be affected, but it is very common for variants to affect more distal genes as well. By mapping expression quantitative trait loci (eQTLs) it is possible to discover which genetic variants alter the expression levels of specific genes. Knowledge about these genes helps substantially improve our understand of the biological processes that underlie each of these variants.

While many eQTL studies have now been performed in blood, they have some limitations. First of all, not all genes are expressed in blood. Second, we know that the way genetic variants affect gene expression levels can be highly cell-type- and tissue-specific. For instance, it is possible that a variant has no regulatory effect in blood yet strongly affects gene

expression levels in the liver. While some eQTL studies have focused on specific tissues, these have usually had limited sample sizes and therefore limited statistical power to identify eQTL effects. To overcome this, in **chapter 4,** I present a strategy that uses publicly available data to detect tissue-specific regulatory effects, which yields insight into the regulatory consequences of genetic variants in many tissues. I also use this data to show how some protein-coding variants that cause rare diseases affect expression levels in the disease-relevant tissue.

In addition to looking at gene expression effects, it is also possible to look at other types of biological molecules to gain insight into the consequences of genetic variants. In **chapter 5**, I investigate the effect of genetic variants on cytokine levels. This is relevant because cytokines are important regulators of immune responses; they play key roles in several autoimmune disorders and in the defense against infectious diseases. In **chapter 6,** I study effects on DNA methylation, investigating how disease-causing variants located on a specific chromosome affect methylation levels on other chromosomes (i.e. *trans*-meQTL effects). This work provides insight into specific regulatory genes that are important in disease development, as well as insights into the downstream consequences of these genes.

These studies clearly show that genetic variants have effects on many different molecular layers. However, environment also has a strong effect on molecular mechanisms and disease-development. Environmental effects can interact with genetic variants and alter the molecular consequences of a variant. In **chapter 7**, I describe a method that I developed to study these interaction effects. By applying this method to a large-scale eQTL dataset comprising data from multiple biobanks, such as Lifelines, unified by BBMRI-NL, it became clear that cell-type-composition differences and viral stimuli strongly alter the molecular consequences of disease-associated variants. By correcting for these strong influencers, we could also detect more subtle effects that revealed how the regulatory effects of transcription factors are influenced by genetic variants.

Although GWAS studies have been very successful in the identification of variants contributing to common diseases, they are underpowered to identify the rare variants that underlie rare diseases. In **chapter 8**, I describe a novel method that uses public RNA-seq data to predict candidate genes that could explain the symptoms of a patient with suspected rare disease-causing variants. I show that, by applying this method, it is possible to provide more patients with a diagnosis. I also show that it is possible to flag genes that may have been falsely associated to a disease.

The work presented here is part of a rapidly growing field and has been placed in context in **chapter 9**. Here I discuss how the advances I made contribute to the field of human genetics and what future challenges await. I also discuss how the healthcare system can benefit from further implementing genetics and speculate on how this might be done. Finally, I explain the need to measure molecular phenotypes in complex patients to fully explain their phenotype.

## Samenvatting

Als je het genoom van individuen met elkaar vergelijkt zijn er veel plaatsen waar mensen van elkaar kunnen verschillen. Deze verschillen worden veroorzaakt door mutaties die zich hebben opgestapeld over een periode van duizenden generaties. Door de accumulatie van deze mutaties zullen twee mensen gemiddeld op 15 miljoen basen verschillen als je ze met elkaar vergelijkt. Hoewel de meeste van deze genetische varianten geen positieve of negatieve gevolgen hebben, kunnen sommige van deze varianten een ziekte veroorzaken of de kans op het ontwikkelen van een ziekte vergroten. Wanneer een variant bijvoorbeeld binnen een gen zit kan dit resulteren in een ander, niet-functionerend eiwit. Het is ook mogelijk dat een variant een effect kan hebben op de hoeveelheid RNA die wordt geproduceerd, wat resulteert in gewijzigde eiwitniveaus die problemen kunnen veroorzaken.

Helaas is het moeilijk te voorspellen wat de functionele of schadelijke effecten van specifieke varianten zijn. Daarom weten we voor het overgrote deel van de varianten momenteel niet of ze ziekten zullen veroorzaken. Dankzij de beschikbaarheid van Genome Wide Association Studies (GWAS) in de afgelopen 14 jaar, wordt nu echter duidelijk dat duizenden varianten worden geassocieerd met allerlei veelvoorkomende complexe ziekten, zoals auto-immuunziekten en neurologische aandoeningen.

Het blijft echter zeer moeilijk om deze resultaten te interpreteren vanwege de beperkte resolutie van GWAS, aangezien de DNA-chips die doorgaans hiervoor worden gebruikt gewoonlijk minder dan een miljoen genetische varianten ondervragen. Ondanks de lage resolutie is het wel mogelijk om conclusies te trekken over de andere varianten via een statistische methode genaamd ‘imputatie’, dat werkt met behulp van informatie uit een ‘referentie dataset’. Eén zo’n set is het Genoom van Nederland (GoNL), bestaande uit 250 Nederlandse ouders-kind families waarvan het hele genoom werd gesequenced door het BBMRI-NL consortium. In **hoofdstuk 2** bespreek ik hoe het GoNL gebruikt kan worden voor dit imputatieproces. Ik laat zien hoe dit de resolutie van GWAS kan verbeteren in onderzoeken die zich concentreren op Europese monsters. Ik laat ook zien dat integratie van de GoNL met andere, al bestaande, referentie datasets de kwaliteit van het imputatieproces verder verbetert. In **hoofdstuk 3** beschrijf ik een softwarepakket genaamd Genotype Harmonizer dat dit imputatieproces kan faciliteren.

Een andere uitdaging in de interpretatie van varianten geassocieerd met ziekte, is dat de meeste varianten zich buiten de genen bevinden. Dit betekent dat deze varianten waarschijnlijk een soort regulerend effect hebben op een gen of op meerdere genen en dat deze gewijzigde regulatie de ziekte kan veroorzaken. Het is echter erg moeilijk om te voorspellen welke genen door een individuele variant zullen worden beïnvloed. In sommige gevallen zal het gen dat het dichtst bij een variant ligt, worden beïnvloed door een variant, maar het is heel gebruikelijk dat varianten ook meer distale genen beïnvloeden. Door expressie quantitative trait-loci (eQTL’s) in kaart te brengen, is het mogelijk om te ontdekken welke genetische varianten de expressieniveaus van specifieke genen veranderen. Kennis over deze genen helpt de biologische processen die ten grondslag liggen aan elk van deze varianten aanzienlijk beter te begrijpen.

Hoewel veel van deze eQTL-onderzoeken in bloed zijn uitgevoerd, bestaan er enkele beperkingen: ten eerste worden niet alle genen in bloed tot expressie gebracht. Ten tweede weten we ook dat de manier waarop genetische varianten genexpressieniveaus beïnvloeden, zeer celtype en weefselspecifiek kan zijn. Het is bijvoorbeeld mogelijk dat een variant geen regulerend effect heeft in het bloed, maar dat de variant de genexpressieniveaus in de lever wel sterk beïnvloedt. Hoewel sommige eQTL-onderzoeken zich op specifieke weefsels concentreerden, hadden ze meestal beperkte steekproefomvang en daardoor beperkte statistische kracht om eQTL-effecten te identificeren. Om dit te verhelpen, besprak ik in **hoofdstuk 4** een strategie die publiek beschikbare gegevens gebruikt om weefselspecifieke regulerende effecten te detecteren, die het mogelijk maken om inzicht te krijgen in de gevolgen op regulatie van genetische varianten in veel weefsels. Ik heb deze gegevens ook gebruikt om te laten zien hoe sommige eiwitcoderingsvarianten, die zeldzame ziekten veroorzaken, expressieniveaus beïnvloeden in het voor de ziekte relevante weefsel.

Naast het bestuderen van genexpressie-effecten, is het ook mogelijk om naar andere typen biologische moleculen te kijken om inzicht te krijgen in de gevolgen van genetische varianten. In **hoofdstuk 5** onderzocht ik het effect van genetische varianten op cytokineniveaus. Dit is relevant omdat cytokines belangrijke regulatoren zijn van immuunreacties en ze een sleutelrol spelen bij verschillende auto-immuunziekten en de verdediging tegen infectieziekten. In **hoofdstuk 6** bestudeerde ik ook effecten op DNA-methylatie. Ik heb onderzocht hoe ziekteverwekkende varianten, die zich op een specifiek chromosoom bevinden, de methylatie-niveaus op andere chromosomen beïnvloedden (d.w.z. *trans*-meQTL-effecten). Dit resulteerde in inzicht in specifieke regulerende genen die belangrijk zijn bij de ontwikkeling van ziekten, en gaf inzicht in de gevolgen die voortvloeien uit deze genen.

Deze studies tonen duidelijk aan dat genetische varianten effecten hebben op veel verschillende moleculaire lagen. De omgeving heeft echter ook een sterk effect op moleculaire mechanismes en de ontwikkeling van ziektes. Deze omgevingseffecten kunnen interageren met genetische varianten en de moleculaire gevolgen van een variant wijzigen. In **hoofdstuk 7** beschrijf ik een methode die ik heb ontwikkeld die deze interactie-effecten kan bestuderen. Door deze methode toe te passen op een grootschalige eQTL-dataset van verschillende bio-banken, zoals Lifelines, verenigd in BBMRI-NL, werd het duidelijk dat celtype-samenstellingsverschillen en virale stimuli de moleculaire gevolgen van ziekte gerelateerde varianten sterk veranderen. Door statistisch te corrigeren voor deze sterke effecten, konden we ook meer subtiele effecten detecteren die onthulden hoe de regulerende effecten van transcriptiefactoren worden beïnvloed door genetische varianten.

Hoewel GWAS-onderzoeken zeer succesvol zijn geweest in de identificatie van varianten die bijdragen aan veel voorkomende ziekten, hebben ze te weinig kracht om zeldzame varianten te identificeren die ten grondslag liggen aan zeldzame ziekten. In **hoofdstuk 8** beschrijf ik een nieuwe methode die openbare RNA-seq data gebruikt om kandidaat-genen te voorspellen die de symptomen zouden kunnen verklaren van een patiënt met vermoedelijke varianten die zeldzame ziekten veroorzaken. Ik laat zien dat het met deze methode mogelijk is om meer patiënten een diagnose te geven. Bovendien kan ik ook genen markeren die mogelijk vals zijn geassocieerd met een ziekte.

Het hier gepresenteerde werk maakt deel uit van een snelgroeiend veld en is in **hoofdstuk 9** in de juiste context geplaatst. Hier bespreek ik hoe de vorderingen die hier worden gemaakt bijdragen aan het gebied van de menselijke genetica en welke toekomstige uitdagingen wachten. Ik bespreek ook hoe de zorg kan profiteren van verdere implementatie van genetica en speculeren over hoe dit kan worden geïmplementeerd. Ten slotte leg ik uit waarom het nodig is om moleculaire fenotypen te meten bij complexe patiënten om hun ziekte volledig te verklaren.

## Acknowledgements

Dear family, friends and colleagues, while it is nearly impossible to acknowledge everyone by name, I would like to thank everyone for their support and collaboration.

First of all, I would like to thank my parents Bert & Ingrid for all their support and enabling me to develop.

I would like to thank my loving wife Marieke for being there for me. I could not have done it without your patience and support.

John, you are only 3-years-old now, but I would like to thank you for the joy you bring and for once in a while allowing me to type on my laptop without joining in.

Eline, thank you for designing the cover of this thesis and for being my sister.

Morris, I want to thank you for your guidance and mentoring me during my PhD.

Lude, I really enjoyed your input and our brainstorming sessions.

Cisca, I'm very grateful for your critical input and the opportunities you have given me.

Marc Jan, we have known each other since our bachelors, and I have benefited a lot from our shared time during our studies and PhD projects. I'm honored you are coming back to the Netherlands to be a paranymf during my defense.

Pieter, you have been a great college and I'm happy you will support me during my defense as a paranymf.

Dasha, we co-authored several papers and it was always a pleasure working with you.

Jingyuan, you where the supervisor of my first internship in human genetics, thank you for introducing me to the field and for your patience.

Jan Jakob Schuringa, I want to thank you and your colleagues for a very interesting internship at the UMCG hematology department.

I would like to thank Ramnik Xavier and Daniel MacArthur for hosting me at the Broad institute and all the members of their labs for all the interesting discussions.

Jackie & Kate, you both have been invaluable in making my writing readable and professional, thank you.

I enjoyed working with everyone in the FrankeSwertzLab, where there is room for feedback and room for encouragement. I feel we really conduct science as group instead of as competitive individuals.

I want to thank all the members of the GCC for their support with Molgenis and programming, for finding all the samples on the cluster, and for the research projects we worked on together.

It was also a pleasure collaborating with all the other researchers in the genetics department. It was really nice to get together with people from different disciplines and jointly work on projects.

I'm also grateful for the collaboration with the clinical geneticists and everyone from diagnostics, for collecting patient material and collaborating on rare disease projects.

I want to thank all the members of the GoNL consortium, the BIOS consortium and the eQTL consortium for their contributions in doing the big things we could have never have done on our own.

I have supervised several interns, and I would like to thank Matthieu Beukers, Harmen Boers, Marije van der Geest, Adriaan van der Graaf, Rahul Gannamani and Sophie Mulcahy Symmons for their efforts.

Without our computer clusters my work would not have been possible, at least not within a normal human lifespan, so I want to thank the CIT and everyone involved in keeping our computer infrastructure running.

I am also grateful to everyone working in the wet lab who generated the data I worked on.

I also want to thank the support staff, including the secretaries, facility management and financial managers for keeping the genetics department running smoothly.

I would like to thank my teachers from both the Hanze University and the VU for teaching me the skills needed for my PhD project.

Finally, I would like to thank all the biobank participants and patients that have allowed us to work with their data.

# Curriculum vitae

Patrick Deelen completed his BASc in bioinformatics at the Hanze University Groningen in 2010, following a one-year internship and the genetics department of University Medical Center Groningen (UMCG) under supervision of Jingyuan Fu and Lude Franke. He then continued his studies by pursuing a bioinformatics masters at the VU University Amsterdam, during which he had a 6-month internship in the hematology department of the UMCG under Jan Jacob Schuringa and again at the UMCG genetics department, now under Morris Swertz. In 2012, he received his MSc with honors, after which he started his PhD project in the genetics department of the UMCG under the guidance of Morris Swertz, Lude Franke and Cisca Wijmenga. During his PhD, Patrick spent 10 weeks at the Broad institute of MIT and Harvard visiting the lab of Ramnik Xavier and Daniel MacArthur. During his PhD studies, Patrick gave two oral presentations at European Society of Human Genetics conferences as well as presenting at several national conferences. He was twice awarded the prize for best poster and won the third prize for an oral presentation. In 2014, he was awarded a €4,000 De Cock grant and, in 2015, a €25,000 Gratama Foundation grant. He is currently a post-doc in the UMCG Genetics department working on the personalized medicine project of the UMCG.

## First author publications

1. Deelen, P. *et al.* Improving the diagnostic yield of exome-sequencing, by predicting gene-phenotype associations using large-scale gene expression analysis. *bioRxiv* 375766 (2018). doi:10.1101/375766
2. Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49,** 139–145 (2017).
3. Kanterakis, A. *et al.* Molgenis-impute: imputation pipeline in a box. *BMC Res. Notes* **8,** 359 (2015).
4. Deelen, P. *et al.* Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med.* **7,** 30 (2015).
5. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes* **7,** 901 (2014).
6. Deelen, P. *et al.* Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *Eur. J. Hum. Genet.* **22,** 1321–1326 (2014).

## Second author publications

7. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv* 447367 (2018). doi:10.1101/447367
8. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49,** (2017).
9. Li, Y. *et al.* Inter-individual variability and genetic influences on cytokine responses to bacteria and fungi. *Nat. Med.* **22,** 952–960 (2016).

## Co-author publications


10. Zhernakova, D. V. *et al.* Individual variations in cardiovascular-disease-related protein levels are driven by genetics and gut microbiome. *Nat. Genet.* 1 (2018). doi:10.1038/s41588-018-0224-7
11. van der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0089-9
12. Stolle, S. *et al.* Running-wheel activity delays mitochondrial respiratory flux decline in aging mouse muscle via a post-transcriptional mechanism. *Aging Cell* **17,** (2018).
13. Macé, A. *et al.* CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. *Nat. Commun.* **8,** (2017).
14. Jankipersadsing, S. A. *et al.* A GWAS meta-analysis suggests roles for xenobiotic metabolism and ion channel activity in the biology of stool frequency. *Gut* **66,** (2017).
15. Oosting, M. *et al.* Functional and Genomic Architecture of Borrelia burgdorferi-Induced Cytokine Responses in Humans. *Cell Host Microbe* **20,** (2016).
16. Graham, D. B. *et al.* TMEM258 Is a Component of the Oligosaccharyltransferase Complex Controlling ER Stress and Intestinal Inflammation. *Cell Rep.* **17,** (2016).
17. Slieker, R. C. *et al.* Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. *Genome Biol.* **17,** (2016).
18. Li, Y. *et al.* A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans. *Cell* **167,** (2016).
19. Bonder, M. J. *et al.* The effect of host genetics on the gut microbiome. *Nat. Genet.* **48,** (2016).
20. Iotchkova, V. *et al.* Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps. *Nat. Genet.* **48,** (2016).
21. Hehir-Kwa, J. Y. *et al.* A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.* **7,** (2016).
22. Visschedijk, M. C. *et al.* Pooled resequencing of 122 ulcerative colitis genes in a large Dutch cohort suggests population-Specific associations of rare variants in MUC2. *PLoS One* **11,** (2016).
23. Kleinloog, R. *et al.* RNA Sequencing Analysis of Intracranial Aneurysm Walls Reveals Involvement of Lysosomes and Immunoglobulins in Rupture. *Stroke* **47,** (2016).
24. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science (80-. ).* **352,** 565–569 (2016).
25. Ricaño-Ponce, I. *et al.* Refined mapping of autoimmune disease associated genetic variants with gene expression suggests an important role for non-coding RNAs. *J. Autoimmun.* **68,** (2016).
26. Van Leeuwen, E. M. *et al.* Population-specific genotype imputations using minimac or IMPUTE2. *Nat. Protoc.* **10,** (2015).
27. Tigchelaar, E. F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5,** (2015).
28. Farlow, J. L. *et al.* Lessons Learned from Whole Exome Sequencing in Multiplex Families Affected by a Complex Genetic Disorder, Intracranial Aneurysm. *PLoS One* **10,** e0121104 (2015).

29. van Leeuwen, E. M. *et al.* Genome of The Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels. *Nat. Commun.* **6,** 6065 (2015).
30. Bonder, M. J. *et al.* Genetic and epigenetic regulation of gene expression in fetal and adult human livers. *BMC Genomics* **15,** (2014).
31. Francioli, L. C. *et al.* Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46,** 818–825 (2014).
32. Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur. J. Hum. Genet.* (2013). doi:10.1038/ejhg.2013.118
33. Almeida, R. *et al.* Fine mapping of the celiac disease-associated LPP locus reveals a potential functional variant. *Hum. Mol. Genet.* **23,** (2014).
34. Byelas, H. *et al.* Scaling bio-analyses from computational clusters to grids. in *CEUR Workshop Proceedings* **993,** (2013).
35. Bonardi, F. *et al.* A proteomics and transcriptomics approach to identify leukemic stem cell (LSC) markers. *Mol. Cell. Proteomics* **12,** (2013).
36. Fu, J. *et al.* Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.* **8,** (2012).
37. Fehrmann, R. S. N. *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.* **7,** e1002197 (2011).
38. Szperl, A. *et al.* Exome sequencing in a family segregating for celiac disease. *Clin. Genet.* **80,** (2011).

## Consortium banner publications

I have collaborated within several consortiums. The following papers have been published with a banner author for these consortia.

### *Genome of the Netherlands (GoNL) Consortium*

The GoNL consortium performed whole genome sequencing on 769 individuals of Dutch families. This data was, among other applications, used to identify rare variants found in the Dutch population, to investigate demographic history, to study properties of *de novo* variants, and to improve genotype imputation.

39. Francioli, L. C. *et al.* A framework for the detection of de novo mutations in family-based sequencing data. *Eur. J. Hum. Genet.* **25,** 227–233 (2017).
40. Li, M. *et al.* Transmission of human mtDNA heteroplasmy in the genome of the Netherlands families: Support for a variable-size bottleneck. *Genome Res.* **26,** (2016).
41. Palamara, P. F. *et al.* Leveraging Distant Relatedness to Quantify Human Mutation and Gene-Conversion Rates. *Am. J. Hum. Genet.* **97,** 775–789 (2015).
42. Francioli, L. C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47,** 822–826 (2015).
43. Kloosterman, W. P. *et al.* Characteristics of de novo structural changes in the human genome. *Genome Res.* **25,** (2015).
44. Kiezun, A. *et al.* Deleterious Alleles in the Human Genome Are on Average Younger Than Neutral Alleles of the Same Frequency. *PLoS Genet.* **9,** e1003301 (2013).

*Biobank-based integrative omics study (BIOS) Consortium*

The BIOS consortium used samples from different Dutch biobanks and generated blood RNA-seq and blood DNA methylation data. This data was processed and integrated with genotype information. This was primarily used to investigate the effects of genetic variation on molecular phenotypes and to study the relation between expression and methylation.


45. Moore, R. *et al.* A linear mixed model approach to study multivariate gene-environment interactions. *bioRxiv* 270611 (2018). doi:10.1101/270611
46. Nedeljkovic, I. *et al.* Understanding the role of the chromosome 15q25.1 in COPD through epigenetics and transcriptomics. *Eur. J. Hum. Genet.* **26,** 709–722 (2018).
47. Jadhav, B. *et al.* RNA-Seq in 296 phased trios provides a high resolution map of genomic imprinting. *bioRxiv* 269449 (2018). doi:10.1101/269449
48. Luijk, R. *et al.* Genome-wide identification of directed gene networks using large-scale population genomics data. *Nat. Commun.* **9,** 3097 (2018).
49. Nedeljkovic, I. *et al.* COPD GWAS variant at 19q13.2 in relation with DNA methylation and gene expression. *Hum. Mol. Genet.* **27,** 396–405 (2018).
50. Wahl, A. *et al.* IgG glycosylation and DNA methylation are interconnected with smoking. *Biochim. Biophys. Acta - Gen. Subj.* **1862,** 637–648 (2018).
51. Tobi, E. W. *et al.* DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. *Sci. Adv.* **4,** (2018).
52. van der Plaat, D. A. *et al.* Occupational exposure to pesticides is associated with differential DNA methylation. *Occup. Environ. Med.* **75,** 427–435 (2018).
53. Xu, C.-J. *et al.* DNA methylation in childhood asthma: an epigenome-wide meta-analysis. *Lancet Respir. Med.* **6,** 379–388 (2018).
54. van Iterson, M. *et al.* omicsPrint: detection of data linkage errors in multiple omics studies. *Bioinformatics* **34,** 2142–2143 (2018).
55. van Dongen, J. *et al.* Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat. Commun.* **7,** 11115 (2016).
56. Linnér, R. K. *et al.* An epigenome-wide association study of educational attainment (n = 10,767). *bioRxiv* 114637 (2017). doi:10.1101/114637
57. Richard, M. A. *et al.* DNA Methylation Analysis Identifies Loci for Blood Pressure Regulation. *Am. J. Hum. Genet.* **101,** 888–902 (2017).
58. Mandaviya, P. R. *et al.* Homocysteine levels associate with subtle changes in leukocyte DNA methylation: an epigenome-wide analysis. *Epigenomics* **9,** 1403–1422 (2017).
59. van Iterson, M. *et al.* Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol.* **18,** (2017).
60. Baselmans, B. M. *et al.* Multivariate Genome-wide and integrated transcriptome and epigenome-wide analyses of the Well-being spectrum. *bioRxiv* 115915 (2017). doi:10.1101/115915
61. Ferreira, M. A. *et al.* Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat. Genet.* **49,** 1752–1757 (2017).
62. Wain, L. V. *et al.* Novel Blood Pressure Locus and Gene Discovery Using Genome-Wide Association Study and Expression Data Sets From Blood and the KidneyNovelty and Significance. *Hypertension* **70,** e4–e19 (2017).
63. Dekkers, K. F. *et al.* Blood lipids influence DNA methylation in circulating cells. *Genome Biol.* **17,** 138 (2016).

64. Ligthart, S. *et al.* Tobacco smoking is associated with DNA methylation of diabetes susceptibility genes. *Diabetologia* **59,** (2016).
65. Shah, S. *et al.* Improving Phenotypic Prediction by Combining Genetic and Epigenetic Associations. *Am. J. Hum. Genet.* **97,** 75–85 (2015).
66. Zilhão, N. R. *et al.* Epigenome-Wide Association Study of Tic Disorders. *Twin Res. Hum. Genet.* **18,** 699–709 (2015).
67. Baselmans, B. M. L. *et al.* Epigenome-Wide Association Study of Wellbeing. *Twin Res. Hum. Genet.* **18,** 710–719 (2015).
68. van Dongen, J. *et al.* Epigenome-Wide Association Study of Aggressive Behavior. *Twin Res. Hum. Genet.* **18,** 686–698 (2015).

### *eQTLGen consortium*

The eQTLGen consortium aims to better understand the regulatory consequences of disease-associated genetic variation. This is done by integrating as many blood eQTL datasets as possible.

69. Porcu, E. *et al.* Mendelian Randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *bioRxiv* 377267 (2018). doi:10.1101/377267
70. Qi, T. *et al.* Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat. Commun.* **9,** (2018).
71. Timmers, P. R. *et al.* Genomic underpinnings of lifespan allow prediction and reveal basis in modern risks. *bioRxiv* 363036 (2018). doi:10.1101/363036
72. Linnér, R. K. *et al.* Genome-wide study identifies 611 loci associated with risk tolerance and risky behaviors. *bioRxiv* 261081 (2018). doi:10.1101/261081
73. Lepik, K. *et al.* C-reactive protein upregulates the whole blood expression of CD59 - an integrative analysis. *PLOS Comput. Biol.* **13,** e1005766 (2017).
74. Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* **9,** 2941 (2018).
75. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50,** 668–681 (2018).

## Symbols explained

DNA - for the use of genetic data 

Cloud - for the data from public repositories 

Methyl group - for the DNA-methylation data 

Heart - for the data of patients that we used 

RNA - for the gene expression data 

Bacterium - for bacterial simulations 

People - for the cohort studies used 

Virus - for the environmental effects 

Peptide - for the cytokine data 

Cell - for the context-specific regulation