University of Groningen

A stew of mixed ingredients: Observational omics in the post-GWAS era

Bakker, Olivier

DOI:
10.33612/diss.231386136

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2022

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Bakker, O. (2022). *A stew of mixed ingredients: Observational omics in the post-GWAS era*. University of Groningen. https://doi.org/10.33612/diss.231386136

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 12-10-2022

A stew of mixed ingredients:
observational omics in the post-GWAS era

Olivier Berend Bakker

Paranymphs

Aarón Daniel Ramírez-Sánchez

Esteban Alexander Lopera Maya

Design and printing

Printing of this thesis was financially supported by the University of Groningen and the University Medical Center Groningen.

Printed by: Gildeprint

Cover design & book layout: Olivier Berend Bakker

Copyright

Copright © 2022 Olivier Berend Bakker. All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means without permission of the author.

rijksuniversiteit
groningen

A stew of mixed ingredients: observational omics in the post-GWAS era

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. C. Wijmenga
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
Woensdag 28 September 2022 om 12.45 uur

door

Olivier Berend Bakker

geboren op 15 April 1996
te Swifterbant

Promotor

Dr. S. Withoff

Copromotores

Dr. I.H. Jonkers
Prof. Dr. Y. Li

Beoordelingscommissie

Prof. dr. C. Gilissen
Prof. dr. D. Duffy
Prof. dr. J.D. Laman

“The truth is rarely pure and never simple”

- Oscar Wilde , The Importance of Being Earnest -

Propositions

1. Co-expression networks generated using bulk multi-tissue RNA sequencing provide broad insights into which pathways are impacted by the genetic basis of complex traits. However, these networks lack both directionality and specificity, which hampers interpretation. This is especially true for complex traits whose key tissues or cell-types are under-represented in the data used to construct the co-expression network (Chapter 3).

2. The search for core genes may prove fruitful for finding disease-relevant genes that have a big enough impact on disease risk to be effectively targeted by drugs but that are also not critical for cell survival (Chapter 3).

3. Expression quantitative trait loci exert their effects in a cell-type- and context-specific manner (Chapters 4 & 5). Some of these cell-type-specific effects can be reconstructed from bulk RNA sequencing data using cell-type deconvolution techniques (Chapter 4).

4. Cytokine production by immune cells is determined by genetics to a varying degree, depending on the biological process central to inducing production (Chapters 5 & 6). This genetic basis may have been shaped by evolutionary pressures (Chapter 7).

5. Gluten-specific T cells display a dynamic response profile to *in vitro* stimulation that shows overlap with *in vivo* activated gluten-specific T cells. This profile may be modulated by genetic factors causal for coeliac disease (Chapter 8).

6. Fine-mapping of the genetic factors causal for coeliac disease, or indeed any complex trait, should be performed using multiple assays, each with different fundamental assumptions (Chapter 9).

7. Advances in our theoretical understanding of complex trait genetics will be increasingly dependent on landmark advances in technology and experimental design.

8. “*Genetics and Statistics, then, have in common that each in its own field represents a distinctive point of view, which profoundly influences the intellectual processes with which scientific work is approached.*”

- R.A. Fisher

Contents

Chapter		Page
1	Introduction	12
2	A practical view of fine-mapping and gene prioritization in the post-GWAS era	26
3	Linking common and rare disease genetics through gene regulatory networks	48
4	Deconvolution of bulk blood eQTL effects into immune cell subpopulations	80
5	Shared genetics of cytokine production upon diverse stimulations	114
6	Integration of multi-omics data and deep phenotyping enables prediction of cytokine responses	134
7	Evolution of cytokine production capacity in ancient and modern European populations	162
8	Potential impact of coeliac disease genetic risk factors on T cell receptor signaling in gluten-specific CD4+ T cells	184
9	Fine-mapping of Coeliac disease associated genetic variants using SuRE-SNP implicates role for genetics in epithelial cells	210
10	General discussion	246
	Appendices	270

ATG TCT CTA CTC GCT CAC
CTG TCC AAG TTA TCG TGG
CCC TCA CAT ACT AGG TTA
ACC AAG CAA AAT ACG GAC
GTC CAG GTT TTC CAA AAA
ACA AGG CGT TGG TAC CTC
TTT CGG ACT GCC CGG AAA
ACC GAT TTG AAC TGT AAG
ATA CGT GAC AGA CGC CAC
GTA GGG ACT GTC TTC CAC
CTT CGA CTC ATA GTG GAG
CAG ATT CAC GGC CAA CAT
TTG GTG ACA ATC ACT GGA
AGC CGT CTG CAA GGC CGG
TTC ACT ACT GTC TGT TTT
CTC GAG GTT GCT CGA TCT
CCC TCC AAG GCC GAT AAG
GGG GAT TTT TTG AAG GCC
GAT AAA AAC AAC CTT CAG
GAC AGT TCA TTG TTT TTA
GGG CCG CCG CGC ACT TCC
TTG TTG GTC ATA CTC ATT
ATG CGC ATC CTC GAA TTA
AAC CCC TCC TCC TTG TCT
GGT GCT AAT ACG CCT ATG
GCA GCT GCG GGT CAA CCG
GCA GTT TCT CCA GAC TTG
CAG CGC CAA GTA CCC CGC
GAA CCA CTC ACC GCA AGT
AAA ACC CGC CCT TAC CTA
GCC GTC TGT CAA CCT ACA

ATG TCT CTA CTC GCT CAC
CTG TCC AAG TTA TCG TGG
CCC TCA CAT ACT AGG TTA
ACC AAG CAA AAT ACG GAC
GTC CAG GTT TTC CAA AAA
ACA AGG CGT TGG TAC CTC
TTT CGG ACT GCC CGG AAA
ACC GAT TTG AAC TGT AAG
ATA CGT GAC AGA CGC CAC
GTA GGG ACT GTC TTC CAC
CTT CGA CTC ATA GTG GAG
CAG ATT CAC GGC CAA CAT
TTG GTG ACA ATC ACT GGA
AGC CGT CTG CAA GGC CGG
TTC ACT ACT GGC TGT TTT
CTC GAG GTT GCT CGA TCT
CCC TCC AAG GCC GAT AAG
GGG GAT TTT TTG AAG GCC
GAT AAA AAC AAC CTT CAG
GAC AGT TCA TTG TTT TTA
GGG CCG CCG CGC ACT TCC
TTG TTG GTC ATA CTC ATT
ATG CGC ATC CTC GAA TTA
AAC CCC TCC TCC TTG TCT
GGT GCT AAT ACG CCT ATG
GCA GCT GCG GGT CAA CCG
GCA GTT TCT CCA GAC TTG
CAG CGC CAA GTA CCC CGC
GAA CCA CTC ACC GCA AGT
AAA ACC CGC CCT TAC CTA
GCC GTC TGT CAA CCT ACA

Joi:
*“Mere data makes a man. A and C and T and G. The alphabet of you.
All from four symbols. Where I am only two. 1 and 0.”*
K:
“Half as much, but twice as elegant sweetheart.”

- From *Blade Runner 2049* (2019) by Denis Villeneuve -

Part I

A broad view of the genetics underlying disease

Chapter 1

General introduction

Prologue: Observational omics and the stew of mixed ingredients

Around 2000 years ago, the Roman emperor and stoic philosopher Marcus Aurelius wrote down his inner thoughts in a work now known as the Meditations [I]. In his Meditations, Marcus Aurelius often reflects on the nature of the universe and our place within it. One of these meditations stuck with me, as I feel it almost perfectly expresses the complexity of biology and life as whole:

> *"Either an ordered universe, or a stew of mixed ingredients, yet still coherent order. Otherwise, how could a sort of private order subsist within you, if there is disorder in the Whole? Especially given that all things, distinct as they are, nevertheless permeate and respond to each other"*
>
> – Marcus Aurelius, Meditations 4.27

Here Marcus Aurelius reflects on the materialistic idea of the universe being an infinite set of atoms randomly bumping into each other, stating that there must be some sort of order in the chaos if we are going to be able to take control of our own mind. I find that this line of thinking applies perfectly to the world of biology, which can get so chaotic and complex that a sense of nihilism starts to set in, "Are we ever going to understand this vast complexity?" The analogy of a stew is particularly applicable to interpreting the biological processes foundational to complex disease. Stew is an ever-moving liquid that constantly changes state, in which ingredients interact with each other to form something new. Yet there is still order to a stew. Follow the recipe, and you get the same stew. Add in different ingredients, and the stew changes. In computational biology, it often feels like we are chefs trying to discover the recipe of our favourite stew by taste alone. Furthermore, I find Marcus Aurelious's notion that "*all things distinct as they are, nevertheless permeate and respond to each other*" very fitting when it comes to interpreting the relationships between human phenotypes, genes and diseases.

In this thesis the genetic basis for a variety of complex traits and diseases is studied from different perspectives, with the overarching goal of better understanding the fundamental principles with which genetic variants impact us. We have made use of **observational omics** data as well as more bespoke experimental setups to study: 1) how disease-associated genetic variants work together to affect gene expression (Chapters 2 and 3), 2) how the cellular context affects the discovery of genetic factors that impact gene expression (Chapter 4) and 3) how genetic factors impact our immune system and how this may have evolved over time (Chapters 5, 6 and 7). The third part of this thesis assesses how the genetic variants associated to a common auto-immune disorder, coeliac disease, might be mediating their effects on disease risk in specific cellular contexts (Chapters 8 and 9). Finally, in Chapter 10, the work in this thesis is reflected upon in the contect of the fields and future perspectives discussed.

[I] The key idea in Roman stoic philosophy is to 'live in accordance with Nature' which manifests itself as the recognition and acceptance of factors that are beyond your control and the rationalization of emotion. Stoics are commonly viewed as cold and completely repressing emotion, an alternative interpretation is that it is about accepting factors beyond your control, and not letting them control your decision making.

Observational omics: A type of data that is based on measuring an '-ome' such as the genome (ome referring to the totality of the measurement). The field of study for this -ome would be its corresponding '-omic', for instance genomics. Observational: In the sense in that observations are done, on for instance, a patient population, but there is no functional intervention either *in vivo* or *in vitro*.

Interpreting the genetics of complex traits

Perhaps the best example of a field where observational omics research has been successful is that of complex trait genetics. When the first draft of the human genome project was completed in 2000, there were grand ambitions that this would shed light on the nature of all diseases. While we have learned a great deal from this effort, twenty years later we now know that the truth is far more complex than originally thought. The field, while still young, has progressed rapidly and massively, from the first sequence taking more than a decade to complete to it now being a more or less routine effort to sequence thousands of genomes, even in single cells.

The first enterprises in modern disease genetics looked within families where disease-causing mutations were clearly inherited and followed a **Mendelian** inheritance pattern. These studies confined themselves to specific loci, and the disease-causing mutation could be linked back to a single gene [1]. It was later shown that loci could be screened for disease-causing mutations using the pattern of **linkage disequilibrium** (LD) between alleles [2,3], paving the way towards genome-wide assessment of disease-associated variants in genome-wide association studies (GWASs). The first GWAS was completed in 2002 and, in the 20 years since, there has been an explosion of GWASs examining a plethora of traits [4]. This explosion has been facilitated by the invention of genotyping arrays that measure genetic variation on specific sites for a fraction of the cost of sequencing. These many GWASs have provided invaluable insights into many traits and diseases [5] (for the sake of brevity, I will be considering disease a trait from here onward). For example, we now know that complex traits are extremely **polygenic** and likely result from a complex interplay of environment and genetics. The major limiting factor of GWASs is that their results have proven challenging to interpret, for several reasons.

Firstly, as the majority of the genome consists of non-coding regions, most variants associated to disease in GWASs fall within these non-coding regions. These variants don’t directly alter the structure or function of a protein, as opposed to coding variants which can have an impact on the protein product of the gene they are located in. For non-coding variants, it is not obvious which genes and proteins are affected by the trait-associated variants. This is compounded by the LD between trait-associated alleles, making it hard to know which exact allele within an **LD block** is causal for the trait. In a process called fine-mapping, attempts are made to integrate the associations obtained from GWASs with other functional omics in order to prioritise likely causal variants within disease loci (as in Chapter 2).

An additional challenge is that, for many GWAS traits, much of the heritability still cannot be fully explained by GWASs when compared to observations made in family-based study designs [6]. This is referred to as the missing heritability problem. There are many potential explanations for this issue, including detection power, missing low frequency variants (minor allele frequency (MAF)<1%), missing structural variants, variant–variant interactions and inaccurate phenotyping. This issue may impact inter-

---

Mendelian: Referring to Gregor Mendel’s principles of inheritance.

Linkage disequilibrium (LD): A non-random association between two alleles. A way to think of LD is as the correlation between the alleles in the population.

Polygenic: Involving multiple genes, usually meaning hundreds to thousands.

LD block: A genomic region containing a set of genetic variants which are in linkage disequilibrium.

pretation of GWAS results, as underpowered GWASs are especially likely to miss associations near genes relevant to trait biology, thereby potentially giving a skewed view of the true genetic basis of a trait. Indeed, it is well known that as power for a GWAS increases, so do the number of associated loci [7].

However, even if fine-mapping reveals all the causal variants, it still leaves the question of which gene is affected by the variant. There are several ways of prioritising the genes affected by GWAS variants or loci. Ultimately, most of these approaches operate under the assumptions of the central dogma of molecular biology, hence the most logical place to search for the connections between disease variants and genes is between the DNA and RNA levels. This is done in so-called expression quantitative trait loci (eQTL) studies. eQTL are genetic variants that are associated with the (mRNA) expression of genes. eQTLs can affect the expression of genes locally (*cis*-eQTLs) or of genes far away through a regulatory cascade (*trans*-eQTLs). One important way the eQTL can do this is by disrupting the binding of transcription factors (TFs) in enhancer or promoter regions, leading to a disruption of the mRNA expression (see Chapter 2).

While identification of eQTLs has proven useful to prioritise disease genes in some specific contexts, substantial gaps remain when using current eQTL resources to explain the links between GWAS variants and genes (see Chapter 10). While every non-coding GWAS variant that has a causal effect must impact expression in some way (either by affecting enhancers, promoters, splicing, chromatin or methylation), we currently cannot detect these effects accurately enough to explain GWAS loci. There are several reasons for this. Firstly, there are several post-transcriptional mechanisms in place that can buffer the effects on mRNA expression, so the protein levels are not altered, leading to eQTL that do not have a clear functional effect. Secondly, it has been suggested that studying the correct context and cell types for the traits under study is of paramount importance if the eQTL is going to be informative. Thirdly, due to technical limitations, *trans*-eQTL studies are currently rare and generally underpowered, even though it has been suggested that this is the most informative type of eQTL to study [8,9].

A final hurdle to interpreting GWASs lies in the fact that the effect sizes of the individual variants are very small, which means they are harder to interpret. In fact, the effect sizes of variants are getting so small that it has been suggested that any variant in the genome that influences an expressed gene is causal to a trait in some way [7,8]. This idea was first proposed by Ronald Fischer in 1918, when he described that a quantitative trait could result from an infinite number of genes, each influencing the trait with an infinitely small effect size. In 2017, Boyle et al. expanded upon this idea in the omnigenic model, which postulates that all genetic variants affecting a gene that is expressed are causal for a trait [8]. These authors further describe that these variants are on the edge of a gene regulatory network and mostly influence peripheral genes not directly involved in causing disease. Together, these peripheral genes affect core genes that are directly causal for disease. However, this model has not yet been empirically proven, and for now, remains theoretical.

Evolution and complex traits

There are many traits for which there is evidence that important genes have been impacted by natural selection. For example, variants in fatty acid desaturase enzymes have been selected for in Inuit populations, allowing them to adjust to the arctic environment and a diet high in polyunsaturated fatty acids [10]. Another well-known example of recent **local adaptation** is the ability to process lactose in adulthood. Around 5,000–10,000 years ago the domestication of cattle in the Middle East, North Africa and Europe introduced a strong selective pressure, leading to selection for variants in the lactase locus [11]. While these are both fairly straightforward and explainable traits and evidence for selection exists for complex traits like BMI, skin pigmentation and height [12], the evolution of the genetic basis for many complex traits remains unknown. Understanding which loci are impacted by selection can be helpful in pinpointing which genes might be key for trait biology [13]. For instance, variants that occur in highly **conserved** genes are more likely to lead to Mendelian disease [13]. Furthermore, positive selection can be used to identify cancer driving genes and mutations [14]. For complex traits, it is also hypothesised that variants that may have been beneficial in the past are now detrimental due to our vastly different environment. For example, variants that may have given increased protection against infection in the past may now modulate susceptibility to immune disease [15]. This is in line with epidemiological evidence showing a substantial increase in auto-immune and allergic diseases in recent years [16].

Much as with the progression to the polygenic view of disease genetics, so too have new models been developed to explain the massive complexity of complex traits in the evolutionary context [17,18]. As opposed to the classic hard or soft sweep models (**Fig. 1A, B**), where standing or *de novo* variants are strongly selected for and only a few preferential haplotypes remain in the population, the polygenic adaptation model allows for a more gradual gradient of different haplotypes to exist (**Fig. 1C**). This is also more in line with how one would expect variants acting on complex traits to develop, as they usually do not cause mortality before the age of reproduction due to the small effect of individual variants. Most tests that assay selection either operate by assaying changes in allele frequency between modern populations or look at the length of haplotypes around a variant in modern populations. However, these methods have difficulties to detect polygenic adaptation or soft-sweeps because of the very minor shifts in allele frequency that occur.

The recent development of protocols to extract (relatively) high-quality **ancient DNA** from human remains has opened a wealth of knowledge on our past [19]. This breakthrough has allowed for the direct observation of the scope of genetic variation in ancient individuals and made it possible to call genetic variants for thousands of ancient individuals [20]. As such, these genetic profiles could be integrated with modern GWAS data to see how trait-associated variants have changed over time. There are however major challenges with ancient DNA data, such as the sparsity of samples, significant DNA damage and the potentially different haplotype structure compared to modern

---

Local adaptation: The process of a population adapting to the local environment in the precense of a selective pressure**.**

Conserved: Referring to sections of DNA that remain highly similar between, or within, species, even in the precense of evolutionary pressures.

Ancient DNA: Referring to DNA extracted from ancient (pre-historic) human remains.

populations. In spite of these, such approaches could give new insight into how the genetic basis of complex traits evolved. Furthermore, by studying selection, we can gain insight into what variants are causal and help us understand what processes are foundational to shaping complex traits.

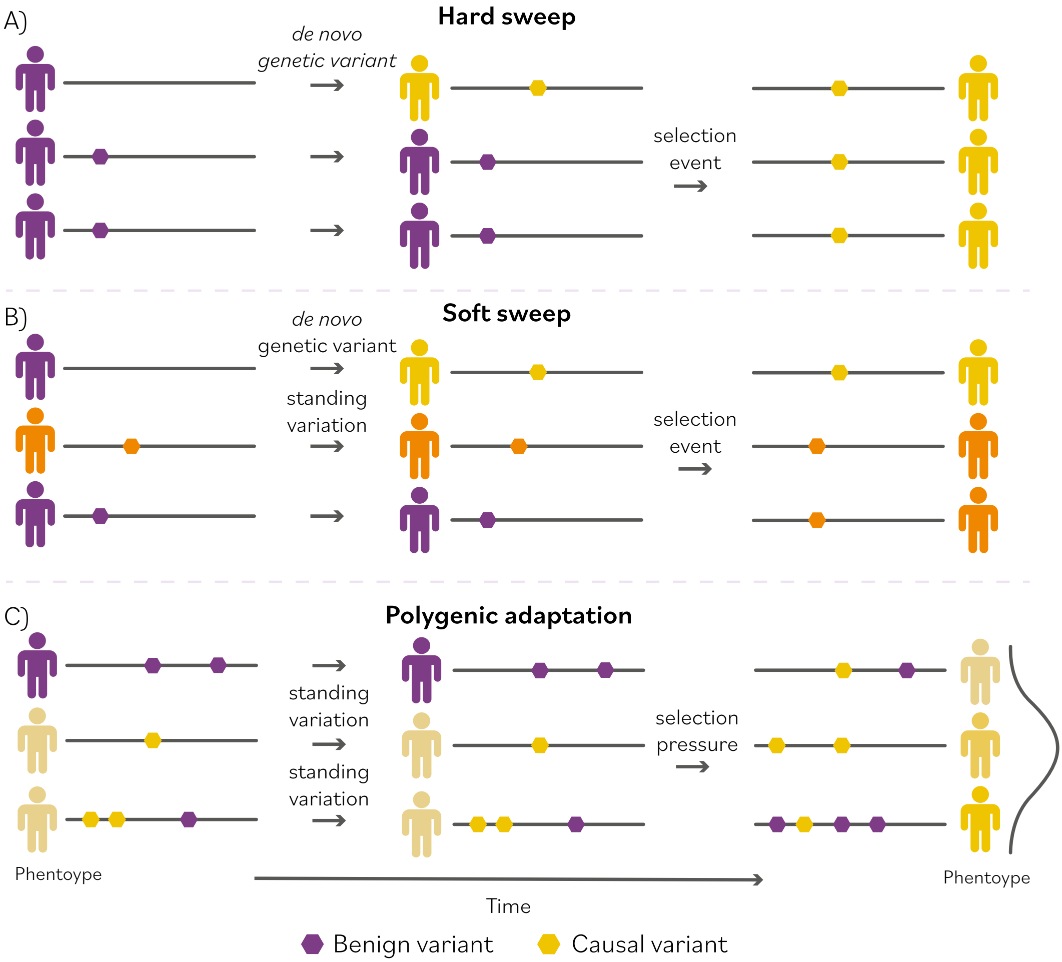


Fig. 1. illustrations of selection.
A) In the hard sweep model, a *de novo* genetic variant is introduced and, after a selection event, the entire population carries the beneficial allele. **B)** In the soft sweep model, either a standing genetic variant is present or a *de novo* genetic variant is introduced. When the selection event occurs, one of the possible alleles affecting the gene is kept. **C)** In polygenic adaptation, a complex standing variation exists in the population and minor changes are made slowly over time due to selective pressure, leading to a spectrum of possible alleles, leading to the phenotype. Figure inspired by Fan et al. *Science* 2016 [12] and Lluís Quintana-Murci et al. *Nat. Rev. Immunol.* 2013 [13].

The (genetic) factors impacting human immune variation

While building a fundamental understanding of the genetic mechanisms underlying all complex traits provides a foundation to understand any specific trait, each trait has its idiosyncrasies. Applying and adapting the knowledge gained from the fundamental models to a specific trait is just the first of many steps in translating genetic insight to treatment.

One complex trait that exemplifies the concept of a "stew of mixed ingredients" is the human immune response. Human immune responses are very complex, and variation exists on many levels within them. Variation in immune responses generally classifies into two categories. Intra-individual variation occurs when immune activity changes over time within the same individual, for example during an infection or because of ageing. Inter-individual variation describes the substantial variation in immune function between individuals, both in health and disease [21]. The disruption of immune homeostasis by environmental or genetic factors can have wide-reaching consequences for the individual. Simply put, if the immune system is underactive, an individual is more susceptible to infection, whereas an individual with an over-active immune system would be more susceptible to developing auto-immune disease.

In principle, foreign substances are first recognised by innate immunity, which is made up of monocytes, macrophages and granulocytes. The substances are taken up, processed by antigen-presenting cells which activate adaptive immunity. The adaptive immune response (consisting of T and B cells) activates upon presentation of antigens and the activation of co-stimulatory checkpoints, which ultimately leads to it removing the foreing substance. During the activation of immune cells, many pathways are activated, leading to the activation of downstream target genes that encode for important immune proteins. For example, immune cells release signalling molecules, including cytokines, that are used to control the inflammatory response by, for example, recruiting more immune cells to the site of inflammation or signalling to the target cells to proliferate or regulate cell survival [22]. Besides cytokines, many other levels are central in determining the immune response to foreign substances, for instance, antibodies, TFs, epigenetic immune memory, chemokines, the immune receptor repertoire and many more. Cytokine levels are, a good trait to study because they can be accurately quantified and are a good indication of inflammatory state and immune function.

There are many cytokines, each with their own function and role, but they generally fulfil one of the three roles outlined above (recruitment, proliferation and survival of target cells). Some cytokines are preferentially produced by certain immune cell subtypes, such as interleukin-17 by T helper 17 cells. As such, the levels of certain cytokines are good indicators for active inflammation and can point to which parts of the immune system are activated. However, the context in which immune cells are activated is key, as different immune cells can excrete the same cytokines, but these can have different functions, depending on the context [23].

Cytokine levels are influenced by the abundance of the immune cells that produce them, and the abundance of immune cells is, in turn, influenced by the immune context, e.g. an inflamed state [21,24]. Hence it is critical to account for cell-type-abundance when studying cytokine function, or indeed, any other immune function. Another source of inter-individual variability in the cytokine response is the genetics of the individual.

Recent work has suggested that the cytokine responses to stimulation, as well as the immune cell proportions, have a strong genetic basis [24-27]. The full extent of the polygenicity behind regulation of cytokine responses and how this interacts with the environmental influences is, however, still poorly understood. This extends to other components of the immune system and to its response to antigens. Importantly, deregulation of these components may lead to either infectious or auto-immune disease, making it essential to unravel these fundamental processes.

The genetics of coeliac disease

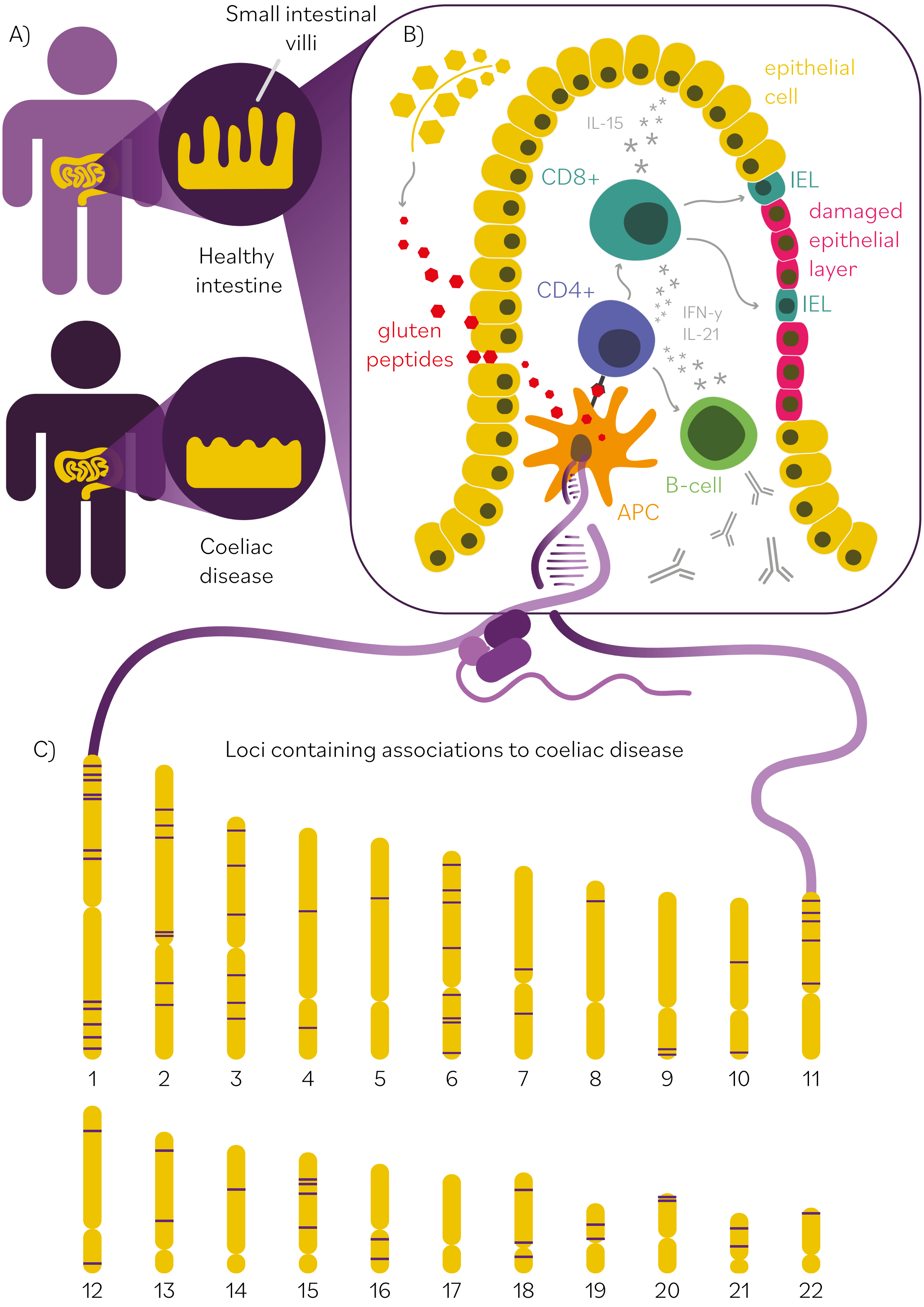
While variation in immune response caused by a combination of genetics and environmental factors is often balanced and innocuous, there are many examples of complex immune-mediated disorders. Viewed from the genetic perspective, there are many loci that confer risk for different diseases, potentially stemming from the fact that the cell types and processes involved in these diseases are partly shared, as well as from selective pressures [28]. An example of a complex trait that shares a common genetic basis with several other immune disorders is coeliac disease (CeD). However, like all complex traits, CeD has its own specific characteristics that complicate translation of general models to actionable findings.

CeD is an auto-immune disorder where an immune response to gluten triggers the atrophy of the small intestinal lining (**Fig. 2**). In CeD, gluten peptides present in wheat-, barley- and rye-derived foodstuffs are first modified by *trans*-glutaminase 2. These modified peptides are then presented to **CD4+ gluten-specific T cells** by **HLA-DQ2**- or **DQ8**-positive antigen-presenting cells [29]. This triggers a downstream immune response where B cells are activated to produce autoantibodies to *trans*-glutaminase 2 and gluten. In addition, cytotoxic lymphocytes are activated that damage the epithelial barrier of the small intestine [29,30]. This damaged epithelium can cause CeD patients to suffer from malnutrition and various gastrointestinal complaints. By far the strongest genetic signals for CeD are the HLA-DQ2 and DQ8 haplotypes, as these are necessary but not sufficient for someone to develop CeD. Besides the HLA complex, there are 43 currently known genetic loci that are associated with CeD risk [31,32].

Genes in these loci – *CD28*, *IL2* and *TAGAP* – have indicated a strong role for T and B cell biology [33,34]. Although CeD pathophysiology is fairly well understood, many outstanding questions remain. For example, the way in which genetic loci play a role in the disease process remains unclear. Moreover, the interplay between the epithelial barrier and the various immune cells that play a role in CeD is also still unclear. As we have a good understanding of the cause and effect with which CeD develops, it is a good model to study how genetic variation plays a role in a complex trait that has its own specific characteristics, context and components.

CD4+ gluten-specific T cells: A immune cell subtype that has a T cell receptor that recognizes processed gluten peptides presented by antigen presenting cells.

HLA-DQ2 and DQ8: Cell surface receptors expressed by antigen presenting cells that bind to the T cell receptor of CD4+ T cells (T helper) during antigen presentation. There are many other human leukocyte antigen (HLA) isoforms besides the DQ2 and DQ8, but these are required for the development of coeliac disease.


A)
Small intestinal villi
Healthy intestine
Coeliac disease
B)
epithelial cell
IL-15
CD8+
IEL
damaged epithelial layer
CD4+
IFN-y
IL-21
IEL
gluten peptides
B-cell
APC
C)
Loci containing associations to coeliac disease
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

Thesis outline

In this thesis the genetic underpinnings of complex disease are discussed from several perspectives. It is structured in four parts. Part I, takes a broader view on interpreting the genetics of complex traits and assess current techniques for arriving at disease mechanisms. It also assesses how the genetics of complex traits impact gene regulatory networks and link our observations to the recently introduced omnigenic model. Part II describes how both environmental and genetic factors can impact immune function, and how immune function can impact the discovery of genetic factors. Addtionally, it discusses how the genetic basis of cytokine production may have been influenced by selective pressures. Part III, dives deeply into the genetics of CeD, a common auto-immune disorder, and tries to identify the functional genetic elements that govern CeD risk in a cell-type-specific manner. Finally, in Part IV, reflects on the work presented and places it into the grander scientific context.

Part I: A broad view of the genetics underlying disease.

In Chapter 2, we discuss current approaches for fine-mapping the genetic underpinnings of complex disease as well as those methods aiming to prioritise relevant disease genes. As it is becoming clear that the role of genetics in complex disease is more difficult to interpret than originally thought, we place the outcomes in the bigger picture of their relevance to interpreting the genetics of complex disease.

In Chapter 3, we take a broad view of the genetics underlying both complex and Mendelian disease and attempt to link them together by integrating GWAS results with co-expression networks. We observe that the gene regulatory networks associated with complex disease are highly inter-connected and that the genes located in the centre of these networks are more likely to harbour Mendelian variants. We further discuss challenges associated with the re-construction of such networks and highlight a roadmap of how these could be overcome in future.

Part II: Genetics and human immune variation

In Chapter 4, we develop a new method, Decon-2, to identify cell-type-specific eQTL effects using bulk gene expression data. By jointly modelling the cell-type proportions and the eQTL effects, cell-type-specific eQTL effects can be identified. We extensively validate these effects in eQTL data from purified cell types and single-cell eQTL data.

In Chapter 5, we assess the shared genetic basis for *ex vivo* cytokine responses to stimulation by jointly modelling cytokine QTL effects on correlated cytokine levels. This multivariate strategy increases the power to detect genetic effects. We subsequently

Left: Fig. 2. Schematic view of the diseasse process and immune response in coeliac disease.
A) Ingestion of gluten leads to severe epithelial damage. **B)** Schematic representation of the currently known immune regulatory cascade in coeliac disease. Gluten in the lumen is digested by proteases into gluten peptides that make their way into the lamina propia, where they are further digested by tissue tranglutaminase 2 (TG2) and subsequently taken up by HLA DQ2- or DQ8-positive antigen-presenting cells (APC). These APC subsequently present the digested gluten peptides to gluten-specific CD4+ T cells, which start to produce a host of cytokines, which in turn activates CD8+ cytotoxic T cells, inter-epithelial lymphocytes (IEL) and B cells. The cytotoxic T cells then start to degrade the epithelial cells, and the B cells start to produce antibodies against TG2. **C)** Overview of the currently known genetic loci associated with coeliac disease. Purple bars indicate loci on the chromsome that contain associations to coeiliac disease Panel B is based on Moerkens & Mooiweer et al. [35]

link the cytokine QTL effects to stimulation-specific eQTL effects and complex disease loci.

In Chapter 6, we describe the genetic and environmental underpinnings of the cytokine response to stimulation. We find that we can explain varying levels of variation using a multitude of host factors, such as genetics, circulating metabolite levels and immune markers, gut microbiome and immune cell proportions. The accuracy of the predictions depends on the stimulation used, with some responses more accurately predicted than others.

In Chapter 7, we assess if the genetic basis for immune variation has been shaped by selective pressures acting in a polygenic manner. We apply an approach based on polygenic risk scores to identify how the collective of genetic effects changed over time in ancient individuals. This has been made possible due to relatively good quality DNA that can be extracted from the wealth of archaeological finds. We applied these scores for immune traits and immune-mediated disorders. We observed a switch towards tolerance against intracellular pathogens and inflammatory responses to extracellular pathogens at the start of the Neolithic period.

Part III: The genetics of CeD in different contexts

In Chapter 8, we study the activation of a specific immune cell type, CD4+ gluten-specific T cells, and their role in the CeD-specific immune response. We describe the activation of these gluten-specific T cells in terms of mRNA expression, protein excretion and chromatin state and attempt to link this to known genetic factors influencing CeD risk.

In Chapter 9, we take a deep dive into the genetics of CeD and fine-map genetic factors that may influence epithelial cells in the small intestine. To do so, we assess the effect of CeD-associated genetic variants using a LD-independent assay that quantifies allele-specific enhancer and promoter activity.

Part IV: Reflections

Chapter 10, reflects on the work presented in this thesis. It discusses several aspects of interpreting the genetic basis of complex disease, and the impact on interpretability of the approaches applied in this thesis. It reflects on some of the limits of, and future opportunities for, the models that are commonly used in the field and discusses the inherent limitations of observational data where causality is concerned. Finally, a more philosophical view is taken to test what (cognitive) biases are present in this work and what is driving them.

References

1. Chong, J. X. et al. The genetic basis of mendelian phenotypes: discoveries, challenges, and opportunities. Am. J. Hum. Genet. 97, 199–215 (2015).
2. Hall, J. M. et al. Linkage of early-onset familial breast cancer to chromosome 17q21. Science 250, 1684–1689 (1990).
3. Houwen, R. H. J. et al. Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. Nat. Genet. 8, 380–386 (1994).
4. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 47, D1005–D1012 (2019).
5. Visscher, P. M. et al. 10 Years of GWAS Discovery: biology, function, and translation. Am. J. Hum. Genet. 101, 5–22 (2017).
6. Manolio, T. A. et al. Finding the missing heritability of complex diseases. Nature 461, 747–753 (2009).
7. Tam, V. et al. Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. 20, 467–484 (2019).
8. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. Cell 169, 1177–1186 (2017).
9. Võsa, U. et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. 53, 1300–1310 (2021).
10. Fumagalli, M. et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. Science 349, 1343–1347 (2015).
11. Bersaglieri, T. et al. Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. 74, 1111–1120 (2004).
12. Fan, S., Hansen, M. E. B., Lo, Y. & Tishkoff, S. A. Going global by adapting local: A review of recent human adaptation. Science 354, 54–59 (2016).
13. Quintana-Murci, L. & Clark, A. G. Population genetic tools for dissecting innate immunity in humans. Nat. Rev. Immunol. 13, 280–293 (2013).
14. Martínez-Jiménez, F. et al. A compendium of mutational cancer driver genes. Nat. Rev. Cancer 20, 555–572 (2020).
15. Sironi, M. & Clerici, M. The hygiene hypothesis: an evolutionary perspective. Microbes Infect. 12, 421–427 (2010).
16. Bach, J.-F. Infections and autoimmune diseases. J. Autoimmun. 25 Suppl, 74–80 (2005).
17. Field, Y. et al. Detection of human adaptation during the past 2000 years. Science (2016) 354(6313):760-764.
18. Speidel, L., Forest, M., Shi, S. & Myers, S. A method for genome-wide genealogy estimation for thousands of samples. bioRxiv (2019) doi:10.1101/550558.
19. Slatkin, M. & Racimo, F. Ancient DNA and human history. Proc. Natl. Acad. Sci. 113, 6380–6387 (2016).
20. Allen Ancient DNA Resource (AADR): Downloadable genotypes of present-day and ancient DNA data | David Reich Lab. https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data.
21. Brodin, P. & Davis, M. M. Human immune system variation. Nat. Rev. Immunol. 17, 21–29 (2016).
22. Turner, M. D., Nedjai, B., Hurst, T. & Pennington, D. J. Cytokines and chemokines: At the crossroads of cell signalling and inflammatory disease. Biochim. Biophys. Acta BBA - Mol. Cell Res. 1843, 2563–2582 (2014).
23. Cytokines in the balance. Nat. Immunol. 20, 1557–1557 (2019).
24. Aguirre-Gamboa, R. et al. differential effects of environmental and genetic factors on T and B cell immune traits. Cell Rep. 17, 2474–2487 (2016).
25. Orrù, V. et al. Complex genetic signatures in immune cells underlie autoimmunity and inform therapy. Nat. Genet. 52, 1036–1045 (2020).
26. Li, Y. et al. A functional genomics approach to understand variation in cytokine production in humans. Cell 167, 1099-1110.e14 (2016).
27. Vuckovic, D. et al. The polygenic and monogenic basis of blood traits and diseases. Cell 182, 1214-1231.e11 (2020).
28. Ramos, P. S., Shedlock, A. M. & Langefeld, C. D. Genetics of autoimmune diseases: insights from population genetics. J. Hum. Genet. 60, 657–664 (2015).
29. Lindfors, K. et al. Coeliac disease. Nat. Rev. Dis. Primer 5, 1–18 (2019).
30. Jabri, B. & Sollid, L. M. T cells in coeliac disease. J. Immunol. 198, 3005–3014 (2017).
31. Dubois, P. C. A. et al. Multiple common variants for coeliac disease influencing immune gene expression. Nat. Genet. 42, 295–302 (2010).
32. Trynka, G. et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in coeliac disease. Nat. Genet. 43, 1193–1201 (2011).
33. van der Graaf, A. et al. Systematic prioritization of candidate genes in disease loci identifies TRAFD1 as a master regulator of ifnγ signaling in coeliac disease. Front. Genet. 11, 562434 (2020).
34. Kumar, V. et al. Systematic annotation of coeliac disease loci refines pathological pathways and suggests a genetic explanation for increased interferon-gamma levels. Hum. Mol. Genet. 24, 397–409 (2015).
35. Moerkens, R., Mooiweer, J., Withoff, S. & Wijmenga, C. Coeliac disease-on-chip: Modeling a multifactorial disease in vitro. United Eur. Gastroenterol. J. 7, 467–476 (2019).

Chapter 2

A practical view of fine-mapping and gene prioritization in the post-GWAS era

R.V. Broekema [1], O.B. Bakker [1*] and I.H. Jonkers [1]*

Open Biology, 2020

1 University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands
** These authors contributed equally*

Abstract

Over the past 15 years genome-wide association studies (GWASs) have enabled the systematic identification of genetic loci associated with traits and diseases. However, due to resolution issues and methodological limitations, the true causal variants and genes associated to traits remain difficult to identify. In this post-GWAS era, many biological and computational fine-mapping approaches now aim to solve these issues. Here we review fine-mapping and gene prioritization approaches that, when combined, will improve understanding of the underlying mechanisms of complex traits and diseases. Fine-mapping of genetic variants has become increasingly sophisticated: Initially, variants were simply overlapped with functional elements, but now the impact of variants on regulatory activity and direct variant-gene 3D-interactions can be identified. Moreover, gene manipulation by CRISPR/Cas9, the identification of expression quantitative trait loci, and the use of co-expression networks have all increased our understanding of the genes and pathways affected by GWAS loci. However, despite this progress, limitations including the lack of cell-type- and disease-specific data and the ever-increasing complexity of polygenic models of traits pose serious challenges. Indeed, the combination of fine-mapping and gene prioritization by statistical, functional and population-based strategies will be necessary to truly understand how GWAS loci contribute to complex traits and diseases.

Keywords

GWAS, fine-mapping, causal variants and genes, SNPs, complex traits, polygenic diseases

Introduction

Most, if not all, phenotypic traits and diseases have a genetic component that influences their development, susceptibility, or characteristics. Which genetic regions (loci) are linked to phenotypic traits has largely been determined by genome-wide association studies (GWASs) (**Fig. 1A**). GWASs compare and associate millions of relatively common genetic variants, usually single nucleotide polymorphisms (SNPs), between a baseline (healthy) population and one with a trait of interest such as type 1 diabetes [1], coeliac disease [2] or height [3]. The trait-associated genetic loci obtained by GWASs are marked by specific variants referred to as marker- or top-variants. Each marker-variant signifies a haplotype containing many nearby variants that are in high linkage disequilibrium (LD), indicating that they are most likely inherited together [4] (**Fig. 1B**). Over 4000 GWASs have been published since 2002 [5], yielding almost 150,000 marker-variant associations to hundreds of traits [6]. However, despite the method's great initial promise, GWASs have not provided immediate insights into the underlying biological mechanisms of each trait due to two major complicating factors.

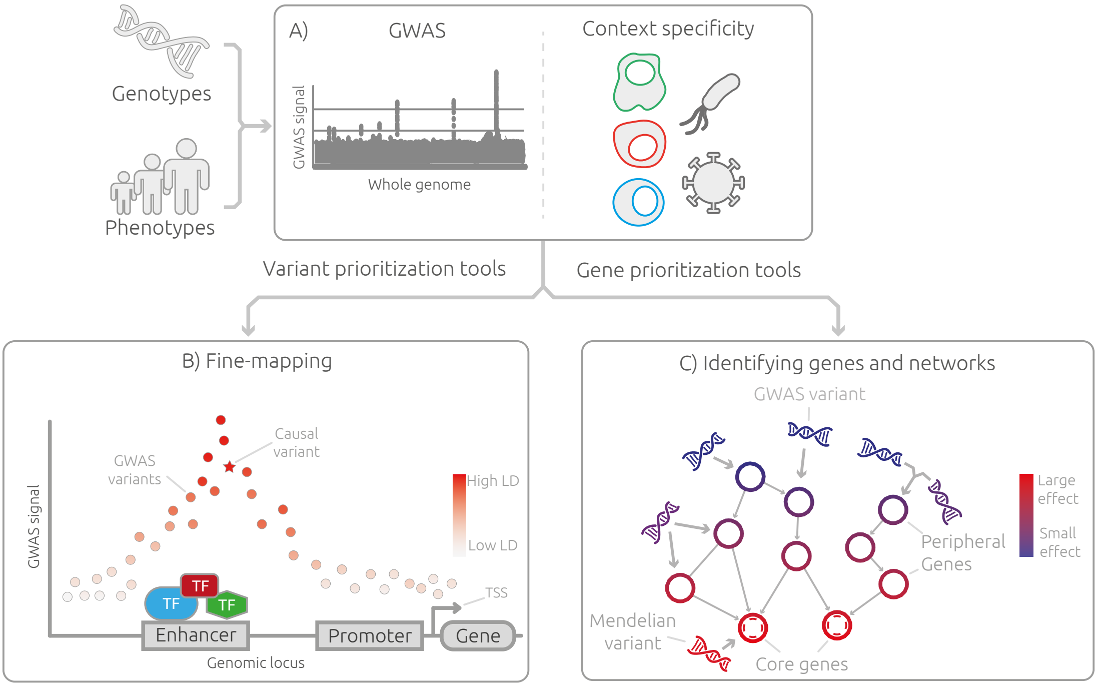


Fig. 1. Outline of the current post-GWAS workflow. A) Firstly, the correct context needs to be identified for the trait under study. **B)** Subsequently, causal variants can be fine-mapped to better understand the fundamental mechanisms of transcription. Here the causal variant (star) is not the strongest GWAS signal, but rather a variant in strong LD with the top effect located in an active enhancer region **C)** To gain insights into the biological processes leading to the phenotype, genes can be prioritized and causal networks constructed. GWAS variants are generally common in the population and have smaller effect-sizes (blue). Thus the genes that they impact are more likely to have a small effect on the phenotype as well (peripheral genes). The genes on which many peripheral genes converge (core genes) generally have stronger effects (red) on the phenotype. As such the variants that affect core genes are more likely to be Mendelian disease variants.

Firstly, GWASs cannot distinguish the marker-variant signal from that of the other varaints that are in high LD. Over 95% of the variants in high LD (R2>0.8) are located outside of genes in the non-coding DNA 7 and can be located up to 500 kilobase-pairs (kb) apart [8]. Consequently, any of them could be the actual causal variant (**Fig. 1B**).

Secondly, the effects of non-coding causal variants can be highly cell-type-, context- and disease-specific [9]. Non-coding DNA contains regulatory regions—enhancers and promoters—that can bind transcription factor (TF) proteins and regulate gene expression [10]. Which enhancers and promoters are utilized depends on the cell-type-specific abundance of ~1600 human TFs and their epigenetically regulated accessibility to a given regulatory region [11]. Variants can disrupt the binding of any of these TFs, resulting in changed enhancer or promoter activity. This, in turn, affects gene expression [12] and cellular pathways [13]. Thus, the cell-type and tissue- or disease-specific micro-environment greatly affect which variants, TFs, genes and pathways are involved (**Fig. 1**). These complexities make it difficult to understand how GWAS loci contribute to their associated traits and have significantly hampered the interpretation and application of GWAS results. To address this, many different fine-mapping approaches have been developed in the post-GWAS era with the aim of identifying the important variants and genes and interpreting their biological impact on diseases and traits [14-17].

Important to note, is that to reduce fine-mapping complexity, most approaches assume that only a single variant per locus contributes to a trait. This is however not a proper reflection of reality as multiple variants within a single GWAS locus can have an effect on a single gene's expression. This can occur in one of two ways, either the effect of the variants adds up in a linear way (additive effect) or an interaction between two or more variants is required to affect gene expression (epistatic effect) [18,19]. Thus, multiple variants may play a role in a single locus, either within a single cell-type, or in a context- and cell-type-specific manner [18]. This further complicates performing and interpreting fine-mapping and gene prioritization approaches. For simplicity, throughout this review we continue to address variants that affect gene regulation and pathways in association to a GWAS trait in any way as causal, even though a collective of smaller contributing effects acting in unison per locus may be necessary to elicit a functional effect on a GWAS trait.

Here we assess fine-mapping and gene prioritization approaches that have been used to translate GWAS loci to a functional understanding of the associated trait, while taking cell-type- and disease-specific context into account. Specifically, we review the genetics of lower effect-size common variants identified through GWASs rather than high effect-size mendelian disease variants (**Fig. 1C**). Moreover, we discuss the impact of the recent paradigm shift towards polygenic models and how these can be used to aid in the identification of gene networks that highlight core disease genes (**Fig. 1C**).

2) Fine-mapping from the variant perspective

Fine-mapping variants in GWAS loci requires an understanding of the underlying mechanism by which a variant can contribute to a trait. Overcoming LD and identifying the context-specific variants that are causal to a trait is imperative for understanding disease mechanisms and confidently identifying which downstream genes and pathways are affected. Many functional and computational (high-throughput) fine-mapping methods have been developed and applied for this purpose. Below we review several fine-mapping methods according to their increasing ability to describe the complex role of variants in GWAS traits and diseases.

2.1 Identifying overlap with functional elements

The most straightforward fine-mapping approach is to overlap GWAS variants in high LD with functional elements such as promoters and enhancers (**Fig. 2A**). Currently, the best resource for functional elements has been compiled by the epigenome roadmap consortium [20] (electronic supplementary material, **Suppl. Tab. 1**), which used ChIP-seq (electronic supplementary material, **Suppl. Tab. 2**) to measure histone marks to determine the location of functional elements in 127 different cell and tissue types [20,21]. Fine-mapping of GWAS variants from 21 autoimmune diseases using the epigenome roadmap and similar data estimated that ~60% of candidate causal variants map to immune cell enhancers, and another ~8% to promoters [12]. This was also reflected in the tissue-specific enrichment of type 1 diabetes susceptibility variants in lymphoid gene enhancers [22]. Moreover, candidate causal variants were enriched in enhancers defined by the histone mark H3K 27Ac in specific subsets of CD4+ T cells, CD8+ T cells and B cells [12]. This was also the case in another study in monocytes, neutrophils and CD4+ T cells [23]. Other studies have also identified tissue-specific enrichments of disease-associated variants via overlap with functional elements, showing that this approach can help specify which variants play a role in certain cell-types [23,24].

Other ways of detecting regulatory regions that can be used to fine-map GWAS variants are either based on DNA accessibility, e.g. ATAC-seq [25] and DNase-seq [26] (electronic supplementary material, **Suppl. Tab. 2**), or identify the inherent transcriptional activity of enhancers and promoters [27,28], e.g. GRO-seq [29], PRO-seq [30] and CAGE [31] (electronic supplementary material, **Suppl. Tab. 2**). Collective public databases utilizing these techniques—like the epigenome roadmap consortium [20], ENCODE [32], FANTOM5 [33] and the IHEC consortium [34]—are indispensable context-specific resources (electronic supplementary material, **Suppl. Tab. 1**). However, it appears to be more difficult than originally anticipated to specify the exact location of regulatory regions since all these methods show different sensitivities and accuracies in the mapping of active regulatory regions [35]. Moreover, overlap of a variant with an active regulatory region may not result in functional disruption of these elements, and thus does not definitively point to causality. This uncertainty limits the accuracy of fine-mapping through overlap with functional elements and still leaves us with a multitude of candidate causal variants.

2.2 Inferring allele-specific variant effects

In high-throughput methods such as ATAC-seq, the sequencing reads containing a variant can be separated based on its allele. The allele-specific abundance of sequencing reads can then directly inform us about the functionality of this variant on the open

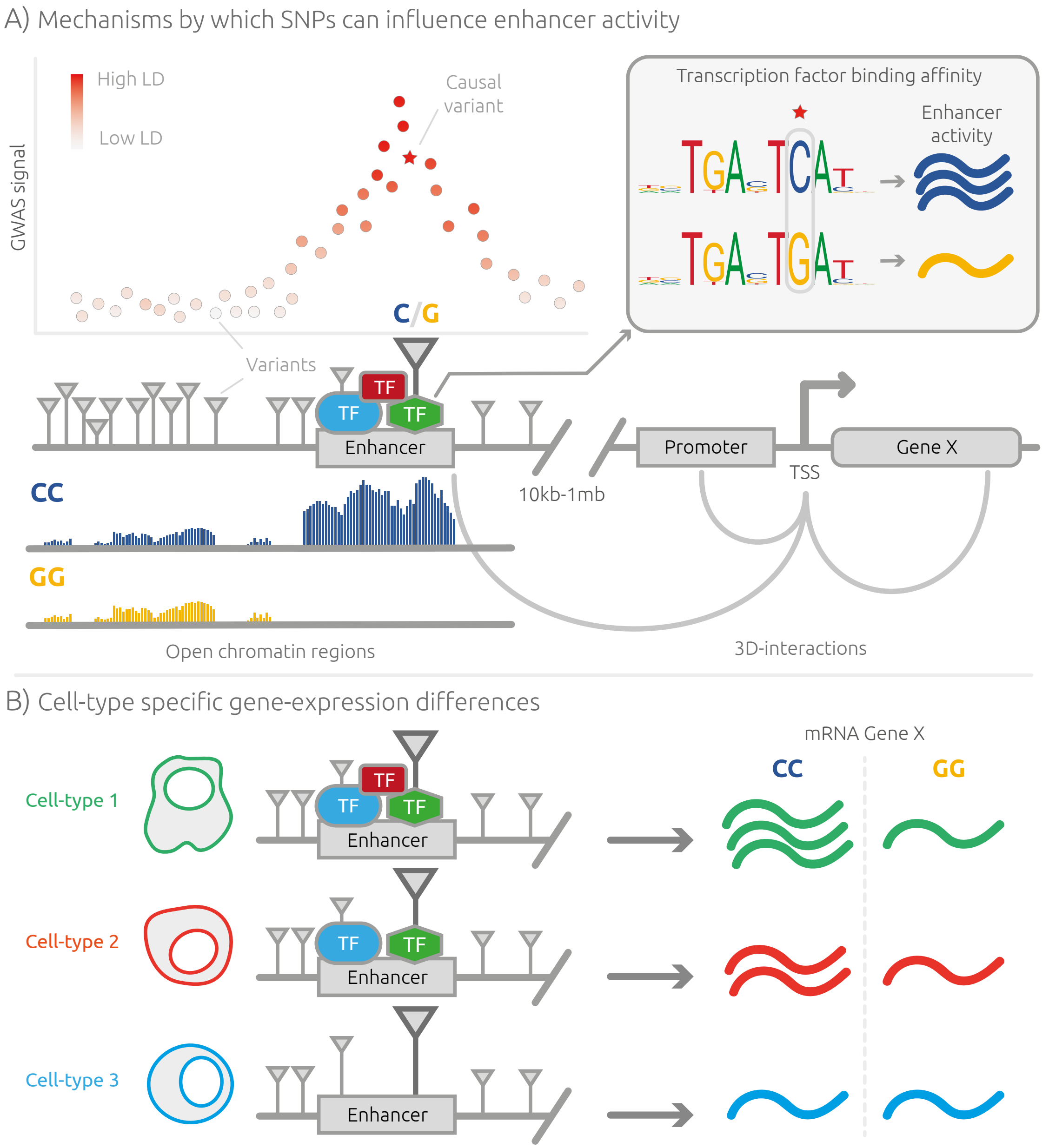


Fig. 2. An illustrative depiction of a GWAS locus showing example mechanisms by which variant effects on enhancer activity and gene expression can be detected. A) Many trait-associated variants are shown with varying linkage-disequilibrium strength as compared to the GWAS-identified marker-variant (in black). In this example the causal variant is located in an allele-dependent active enhancer (C-allele, caQTL) as shown by the open chromatin regions of the same locus (peak-density plot below the variant). The variant affects the transcription factor (TF) binding site of the green TF with a strong binding preference for the C-allele, as shown by the enhancer activity in the 'Transcription factor binding affinity' box. In addition, using 3D-interactions (gray arches connecting the gene, promoter, and enhancer), physical contact with the nearby 'Gene X' indicates the enhancer affects the gene's expression. **B)** To highlight cell-type-specific effects, the influence of the causal variant is depicted in three cell-types with varying TF availability. The mRNA expression of 'Gene X' is stronger for the CC-genotype compared to the GG-genotype because of the increased TF-binding affinity to the green TF (as shown in part A of this figure). This mRNA expression remains low but stable for the GG-genotype in all three cell-types regardless of the TF availability but decreases for the CC-genotype in cell-types with reduced TF availability, which reduces cooperative TF-binding.

chromatin region. Variants that cause allelic imbalance in regulatory regions are called chromatin accessibility quantitative trait loci (caQTLs, **Fig. 2A**) [25,36]. Many caQTLs were identified in primary CD4+ T cell ATAC-seq peaks, and these showed a strong enrichment in candidate causal autoimmune variants [36]. Similarly, the existence of variants or histone-QTLs that affect regulatory regions by altering enhancer-associated H3K27Ac or H3K4Me1 histone peaks also implies that these variants have an effect on cell-type-specific enhancer activity [23]. Due to their functional effect on DNA accessibility and epigenetic marks, these variants are more likely to be causal variants for GWAS traits.

Another mechanism by which non-coding GWAS variants can have an allelic effect on gene expression is alternative splicing of genes. GWAS-associated variants have the potential to induce cell-type specific alternative splicing (sQTL) or could affect *trans*-acting splicing regulation genes [37,38]. This was shown in a genome-wide approach where 622 exons with intronic sQTLs were identified. 110 of these exons harbored variants in linkage disequilibrium with GWAS marker-variants [37]. In a more specific example, the multiple sclerosis associated *PRKCA* gene is seemingly affected by an intronic sQTL that increases the expression of a gene isoform more prone to nonsense-mediated decay, thereby reducing the likely protective PRKCA mRNA levels post-transcriptionally [39]. However, sQTLs appear to also act through more complex mechanisms such as indirectly through caQTLs [40], or by inducing alternative upstream transcription start sites [41]. These, and many other examples [38] suggest that sQTLs may be an important but complex mechanism by which GWAS-associated variants affect a trait.

2.3 Identifying variants that disrupt underlying TF binding sites

Further prioritization of variants in regulatory regions that show allelic-imbalances can be done by computational or functional analysis of the underlying TF binding sites (TFBS) or motifs. Regulatory regions consist of both very strict and more degenerate DNA motifs [42] to which TFs can bind in order to initiate local transcription (e.g. enhancer RNAs) and regulate nearby or distant genes [10,27]. Variants can change the TFBS, altering the binding affinity of the TF and changing the activity of a regulatory region (**Fig. 2A**) [18,43,44]. The specificity and location of potential TFBSs have been collected for many cell-types in large databases such as JASPAR [45], FANTOM5 [33] and ENCODE [32] (electronic supplementary material, **Suppl. Tab. 1**), mostly using ChIP-seq and HT-SELEX [46] (electronic supplementary material, **Suppl. Tab. 2**).

An enrichment of TFBS disruption by putatively causal variants has been identified for 44 families of TFs [18]. For TFs like AP-1 and the ETS TF-family, regulatory regions containing these disrupted TFBSs also show effects on chromatin accessibility, indicating that the effect of variants on TF binding affinity leads to caQTLs [18]. Similarly, upon identification of nearly 9,000 DNase-seq locations affected by allelic-imbalances, it was found that the alleles associated with more accessible chromatin were also highly associated with increased TF binding [43]. In a more specific case, TFBS disruption analyses and *in vitro* confirmation by ChIP-seq led to the identification of rs17293632 as a likely causal SNP that increases Crohn's disease risk by disrupting an AP-1 TFBS [12]. Interestingly, this effect on AP-1 TFBSs was stimulation-specific: H3K 27Ac peaks with affected AP-1 TFBSs were enriched in stimulated CD4+ T cells compared to non-stimulated cells [12]. This highlights the importance of context-specificity and the need for tissue- and disease-relevant stimulations in experimental setups (**Fig. 2B**) [12,47]. Finally, in a study of leukaemia patients, a small DNA insertion resulting in a TFBS for MYB created an enhancer near TAL1, this led to activation of this oncogene and the onset of leukaemia [48]. Thus, decreased or increased affinity of TFs due to genetic variants or small DNA changes can have far-reaching effects.

Currently, only 10-20% of the potentially causal non-coding GWAS variants defined by allelic imbalances within a regulatory region can be shown to disrupt a known TFBS [12]. Therefore, the actual causal variants may potentially act through a different mechanism, or our understanding of TF binding may still be insufficient [49]. One complicating factor here is the potential cooperative binding of more than one TF at an overlapping TFBS. Detection of these cooperative binding motifs is currently being improved by both biological methods (such as SELEX-seq [50]) and computational methods (such as No Read Left Behind (NRLB) [44]) (electronic supplementary material, **Suppl. Tab. 3**). A striking example of context-specific cooperative binding of TFs is illustrated by an increased TFBS enrichment of p300, RBPJ and NF-kB in risk loci of GWAS traits as a consequence of the presence of Epstein-Barr virus (EBV) EBNA2 protein [51]. In this study, ChIP-seq data from EBV-transformed B cell lines was used, together with the RELI algorithm (electronic supplementary material, **Suppl. Tab. 3**), to systematically estimate enrichment of variants in TFBS [51]. In six out of the seven autoimmune disorders tested, RELI identified that 130 out of 1,953 candidate causal variants [12] overlapped with EBNA2 binding sites in B cell lines identified by ChIP-seq [51]. Interestingly, many autoimmune diseases, including coeliac disease and multiple sclerosis [52,53], are thought to be partially triggered by viral infections, suggesting that variants may only be causal when viral factors are also present. Moreover, TF motifs can be highly degenerate, and a small change in TF binding affinity can induce a subtle dosage effect on the activity of a regulatory region [44]. While this effect may be subtle, downstream genes could be affected sufficiently [44] to induce or affect a trait. Thus, a better understanding of how TF binding affinity to DNA motifs is mediated is necessary to comprehend how variants affect the functionality of a regulatory region.

2.4 Fine-mapping by detection of regulatory region activity

A more immediate fine-mapping approach is to directly measure the effect a variant can have on the strength of a regulatory region. Active promoters and enhancers have transcription start sites (TSSs), and the activity of an enhancer or promoter is directly correlated with the active transcription from these TSSs [27]. However, some promoter RNAs, and most enhancer RNAs, are very short lived, making them difficult to detect with most RNA sequencing methods [10,27]. CAGE (electronic supplementary material, **Suppl. Tab. 2**) does allow for identification of exact TSS locations, as well as expression levels of genes, by sequencing 5'-capped transcripts regardless of their stability [30]. CAGE has identified promoter and enhancer effects and showed that 52% of the effects observed in promoter regions were in secondary CAGE peaks, highlighting that genes can have multiple active promoters depending on the genotype [54]. CAGE QTLs have been observed for loci associated with systemic lupus erythematous (SLE) and inflammatory bowel disorder [54], supporting their relevance in immune disease.

Reporter-plasmid assays can also be applied to directly measure the effects of variants on enhancer or promoter TSS activity by moving variant-containing DNA fragments from their natural environment to a plasmid and transfecting these into a cell-type of interest. The most traditional reporter-plasmid assay, the luciferase assay (electronic supplementary material, **Suppl. Tab. 2**), was used to confirm a functional effect of rs1421085, which is associated to obesity risk, by showing that the risk-allele induces an increase in enhancer activity [55]. However, high-throughput reporter assay methods with high resolution are required to fine-map all potentially causal variants within entire GWAS loci based on regulatory region activity.

One such method, the massively parallel reporter assay (MPRA, electronic supplementary material, **Suppl. Tab. 2**), can test over 30,000 candidate variants by synthetically creating 180Bp DNA-fragments containing both alleles of a variant with a unique barcode and integrating these into GFP-reporter plasmids that are subsequently transfected into different cell lines [56]. An MPRA was used to identify expression of 12% (3,432) of the 30,000 candidate DNA fragments in three cell lines, with 842 showing allelic imbalances caused by SNPs. Indeed, [53] of these SNPs had previously been associated to GWAS traits [56]. Similar high-throughput fine-mapping methods that use patient-derived DNA instead of synthetically generated DNA sequences are STARR-seq [57] and SuRE [58] (electronic supplementary material, **Suppl. Tab. 2**). Using a whole-genome approach, the SuRE method managed to screen 5.9 million SNPs in the K562 red blood cell line, identifying over 30,000 SNPs that affect regulatory regions and allowing for in-depth fine-mapping of SNPs for 36 blood-cell-related GWAS traits [59]. Follow-up research on these reporter assays has identified a causal SNP (rs9283753) in ankylosing spondylitis [56] and another (rs4572196) in potentially up to 11 red blood cell traits [59]. Despite the obvious advantages of high-throughput fine-mapping screens, a major drawback is that these methods are usually applied in cancer or EBV-transformed cell lines. These cell lines can be significantly different from trait-specific tissue-derived cell-types [60] and have often accumulated many somatic mutations as a consequence of years of culturing [61]. Thus, the wrong variants may be identified as causal because the relevant cell-type and context-specific effects have not been considered [62].

2.5 From causal variant to gene using the 3D-interactome

When a causal variant has been identified, the gene expression effects of that variant can be directly assessed by mapping the necessary physical interaction of the regulatory region it affects with its target genes (**Fig. 2A**) [63,64]. For example, H3K 27Ac regions containing autoimmune-disease-prioritized variants were linked to the TSS of genes using HiChIP (electronic supplementary material, **Suppl. Tab. 2**) and shown to contain cell-type-specific interactions between the TSS of the *IL2* gene and rs7664452 in Th17 cells and between rs2300604 and target gene *BATF* in memory T cells [63]. Interestingly, for 684 autoimmune-disease-associated variants assessed with HiChIP, 2,597 gene–variant interactions were identified, indicating that autoimmune disease variants can regulate a multitude of genes. Moreover, only 14% (367) of these gene–variant interactions were with the gene closest to the variant [63]. Another example of a long-range interaction of a causal variant is that of the previously mentioned rs1421085, which is associated with obesity risk and located in an intron of *FTO*. TFBS disruption analyses have shown that rs1421085 disrupts the ARID5B TF-binding motif and affects the activity of an enhancer that regulates *IRX3* and *IRX5*, genes located 1.2Mbp upstream, instead of the initially expected co-localized *FTO* gene itself [55,65]. Thus, fine-mapping and interaction analysis has identified additional causal genes in this obesity-associated risk locus.

Hi-C (electronic supplementary material, **Suppl. Tab. 2**) is another high-throughput method for identifying specific promoter and enhancer gene interactions [19,66-68]. For example, Hi-C was used to prioritize four rheumatoid arthritis genes by overlapping promoter–gene interactions of various primary immune-cells with rheumatoid arthritis GWAS variants [19]. Another study analysed Hi-C datasets of 14 primary human tissues and showed that frequently interacting regions (FIREs) are enriched for disease-associated GWAS variants [68]. However, the resolution limitations of Hi-C and other interaction data make it difficult to precisely pin-point the causal variant within a regulatory region [63,64,68]. In addition, cell-type and environmental effects influence regulatory region interactions with genes, as shown by the fact that 38.8% of FIREs were identified in only one tissue or cell-type [68]. Thus, multiple strategies as described here and collected in databases such as the EnhancerAtlas2.0 [69] (electronic supplementary material, **Suppl. Tab. 1**) should be combined to confidently fine-map causal variants and link them to genes that play a role in GWAS traits.

3) Gene prioritization using GWAS traits

Traditional fine-mapping approaches focus on identifying the causal variants that affect a trait of interest. While very important, knowing which variants are causal does not identify the downstream effects of the variant on the trait. One way to gain such insights is by identifying the genes that are affected by each GWAS locus. Moreover, if the causal genes affected by a locus are known, this can reduce the credible set of potentially causal variants. Recent efforts in systems biology have focused on identifying such causal genes and their downstream effects.

3.1 Gene prioritization using expression quantitative trait loci

A more comprehensive approach to identifying the genes affected by a GWAS locus is through the use of quantitative trait loci (QTL, **Fig. 3A**). While caQTLs are often indicative of a causal variant or regulatory region, a specific subset of QTLs called expression QTLs (eQTL) can be used to identify the genes affected by a GWAS locus [70-72]. The simplest way to perform gene prioritization using eQTL analysis is simply to overlap the marker-variant of a GWAS locus with the top eQTL variant. An example of this is a SLE risk variant that is also a *cis*-eQTL for the TF IKF1. The eQTL on *IKF1* affected the transcription of ten genes in *trans* that are all regulated by *IKF1* [70], highlighting this gene as a likely candidate causal gene for SLE. Additionally, these types of effects can be context-specific, as was shown for a *cis*-eQTL on *TLR1* after stimulation of peripheral blood mononuclear cells (PBMCs) with *E. coli* [73]. This *cis*-eQTL was also a strong *trans* regulator of the *E. coli*-induced response network, regulating another 105 genes [73], showing that an eQTL can strongly influence the immune response to pathogens.

However, the top eQTL variant might not always be the same as, or in LD with, the top GWAS marker-variant due to noise in the eQTL data [74] or to multiple causal effects on a gene or disease in a locus [75]. As a result, many statistical frameworks have been created to give more accurate estimates of overlap or causality between a GWAS locus and a QTL locus, including FUMA [76], COLOC [77] and Mendelian Randomization (MR, electronic supplementary material, **Suppl. Tab. 3**). MR is commonly used to estimate causality between GWAS and QTL profiles [78-84] and has been successfully applied to identify genes causally linked with complex traits [3,79-81]. For example, MR studies were able to identify a causal role for *SORT1* on cholesterol levels [79,81], a role which has been experimentally validated [85]. Still, MR can be challenging as multiple variants in LD can affect the same gene (linkage), and several genes can be affected by the same causal variants (pleiotropy) [70,73,86]. More recent work on MR has focussed on more accurately controlling for pleiotropy and linkage [79,81,82,84]. Independent variant selection for MR is currently done by either LD-based clumping or some form of stepwise regression using tools like GCTA's COJO [75] (electronic supplementary material, **Suppl. Tab. 3**), which only select for independence and not causality. Accurate fine-mapping can potentially help these efforts by improving the independent variant selection for MR since fine-mapping can reveal the true causal variants independent of linkage.

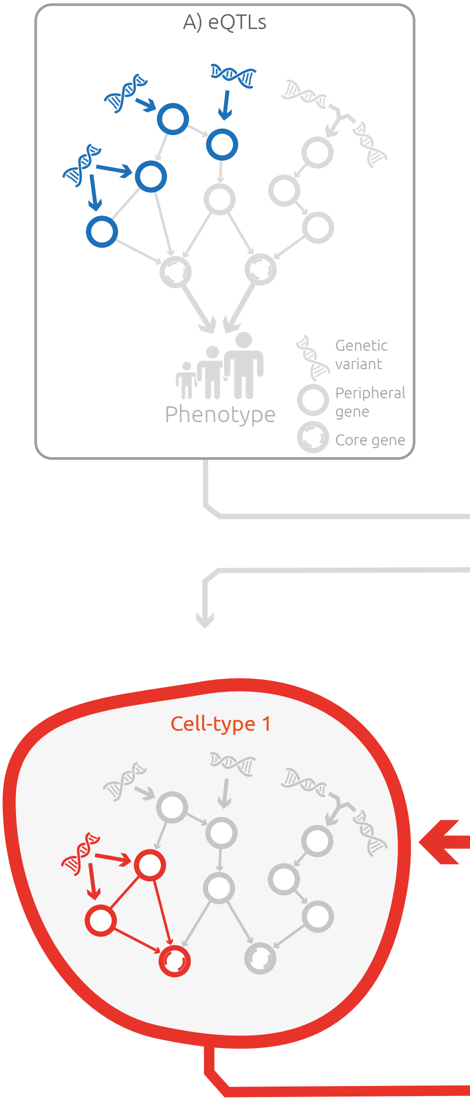
Recently, it has been suggested that ~70% of the heritability in mRNA expression is due to *trans*-eQTLs [87,88], which highlights the importance of *trans*-eQTL relationships. While *trans*-eQTLs have the potential to further our understanding of complex traits, the multiple testing burden is very large due to the large number of comparisons that have to be made when doing genome-wide *trans*-eQTL mapping (in the worst case millions of variants times ~60,000 genes) [70,72]. Therefore, many eQTL studies opt to only map *cis*-eQTL effects genome-wide, as this dramatically reduces the number of comparisons that have to be made [70-72,74]. Another approach is to limit the number of comparisons by only mapping *trans* effects for a predefined subset of variants or genes [70,72,73,86]. However, since a full *trans*-eQTL mapping dataset is rarely available, overlap between *trans* acting genes and GWAS loci will be missed.

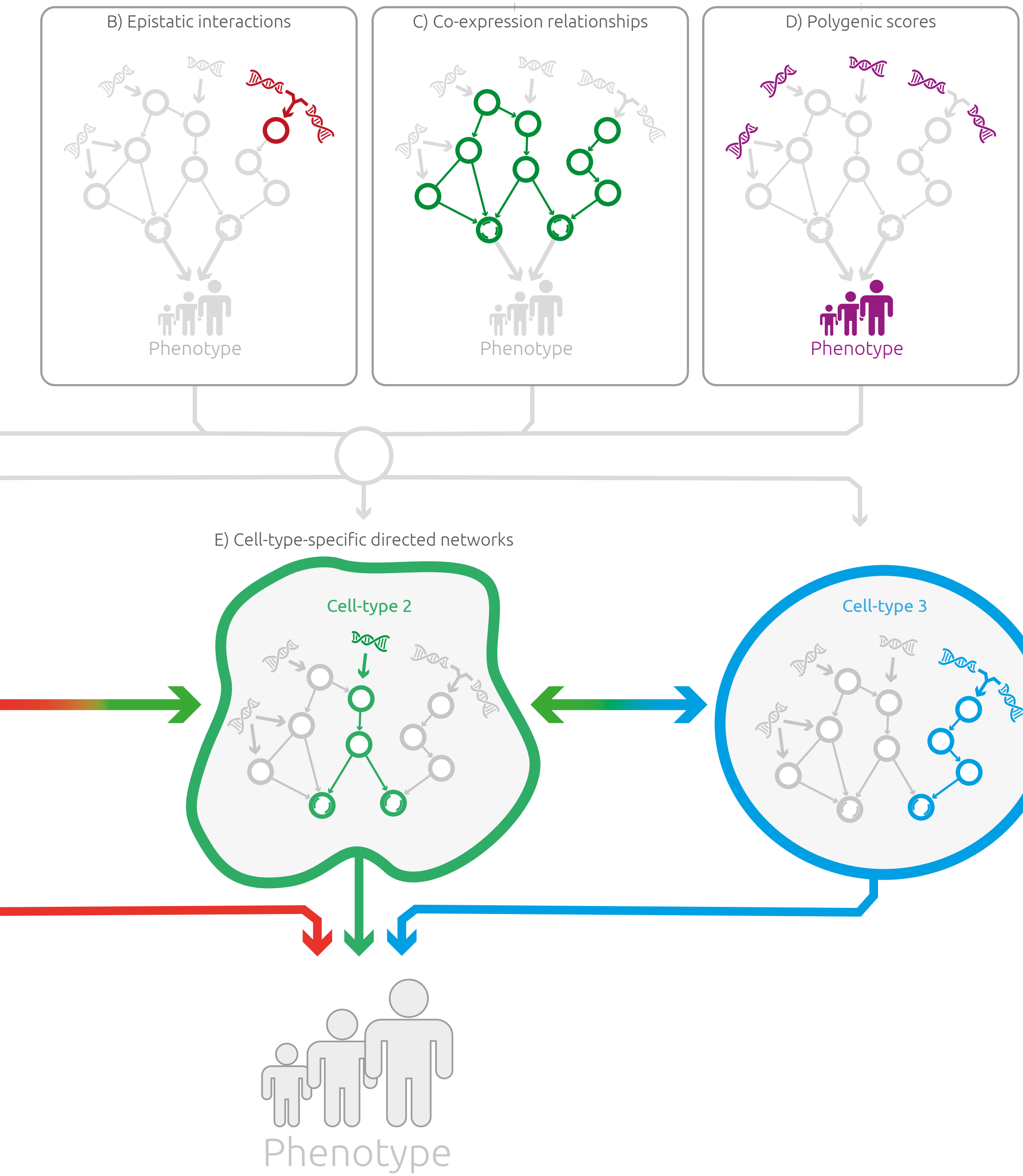
An additional challenge with QTL-based gene prioritization approaches lies in the context-specificity of the QTL data used, as different tissues, cell-types, time points and stimulation conditions can induce many different expression patterns and different interactions with the variants in a GWAS locus [23,73,89-92]. Consequently, the QTL information that is available might not be informative for the trait under study. This is especially challenging when studying traits that are present in a tissue other than blood, as is the case for neurological disorders [93,94], because sufficiently powerful cell-type- or context-specific QTL studies are usually not available. However, with the advent of single-cell RNA sequencing (scRNAseq) and the increasing availability of large-scale datasets for tissues other than blood, some of these challenges are being overcome [70,72,90,91]. scRNAseq (electronic supplementary material, **Suppl. Tab. 2**) allows for high-throughput eQTL analysis in individual cell-types instead of a bulk population, as shown for PBMCs [90]. This allows for an increase in resolution and can help to assess only the trait-relevant cell-types [91], as shown for eQTLs on *TSPAN13* and *ZNF414*, which were only present in CD4+ T cells and not in bulk or other specifically assessed cell-types [90]. Consortia that are amassing single-cell data at a large scale in many different tissues—like the human cell atlas [95], single cell eQTLgen [96] and the Lifetime consortium [97] (electronic supplementary material, **Suppl. Tab. 1**)—will facilitate the use of single-cell sequencing data for traits where bulk RNA-seq obtained from blood is not informative.

3.2 Identifying downstream effects of GWAS loci using other QTLs

Beyond gene-expression-based eQTL, a plethora of other QTL types exist that affect the abundance of proteins (pQTL) [98,99], metabolites (mQTL) [100], DNA methylation (meQTL) [101], microbiota (miQTL) [102] and cells (cell-count or ccQTL) [103,104]. Naturally, these can all be overlapped with GWAS loci to obtain insights into their pathology. For example, the *ex vivo* cytokine response to stimulation has been shown to have strong genetic regulators [99]. Interestingly, all the associated effects found were *trans* (i.e. not in proximity to

Fig. 3. Aspects of fine-mapping genes from GWAS loci. **A)** Using eQTLs (dark blue) and CRISPRi/a-based assays, GWAS loci can be linked to genes when using the correct context. **B)** Not every relationship between genetics and expression can be described additively. Epistatic effects (dark red) describe a relationship where two (or more) mutations are needed to arrive at the phenotype. **C)** Using co-expression, regulatory relationships between genes can be quantified, but the specific role of genetics in these relationships is unknown. **D)** Using polygenic scores, the joint effects of GWAS loci can be assessed, sacrificing resolution to obtain higher-level insights into the pathways affected by the genetics associated with a phenotype. **E)** When assessed at single cell resolution, the total network can be deconstructed into the cell-type relevant components. Affected cells can subsequently display an altered interaction with other cells within a tissue or individual, leading to a changed tissue- or individual wide outcome for a phenotype.




B) Epistatic interactions
Phenotype
C) Co-expression relationships
Phenotype
D) Polygenic scores
Phenotype
E) Cell-type-specific directed networks
Cell-type 2
Cell-type 3
Phenotype

the cytokine genes), suggesting that the release of cytokines is controlled by genes in the receptor's pathways rather than being directly controlled by the mRNA levels of the cytokine. Moreover, context-specificity is important, as QTLs affecting cytokines from T cells were found to be enriched in autoimmune GWAS loci, whereas QTLs affecting cytokines from monocytes were more enriched in infectious-disease-associated loci [99]. Thus, the effects of genetics on traits should not only be studied at the level of gene expression, but also at levels more directly related to a phenotype.

*3.3 Functional approaches to mapping genetic effects on expres*sion

While eQTL analysis provides invaluable insights into the genes that affect a trait or disease, context- and cell-type-specific biases in the expression data and LD structure in GWAS loci cause potential errors in gene prioritization. With the recent introduction of CRISPR/Cas9-based screens [105] (electronic supplementary material, **Suppl. Tab. 2**), it is now possible to functionally validate eQTL effects in a high-throughput manner independent of LD structure and in a cell-type relevant to the trait of interest.

CRISPR-based assays use guide RNAs to bind specific regions of the genome and either activate (CRISPRa) or interfere (CRISPRi) with the transcription of genes or enhancers [106]. Recent advances in both scRNAseq and CRISPRi/a have facilitated methodologies that evaluate enhancer effects on genes in single cells [107]. For example, a recent effort evaluated the effects of 5,920 candidate enhancers on gene expression using CRISPRi [107]. Strikingly, 664 showed a significant effect on gene expression in K562 cells. Thus, CRISPRi-based assays are capable of identifying enhancer–gene pairs in a high-throughput manner. However, as only ~10% of candidate enhancers were actually found to affect gene expression, identifying which enhancers are active based on already available data might not always be straightforward, even for a very well characterized cell-line such as K562 [20,32,34,58,59].

In addition to mapping active enhancer gene pairs, CRISPRi/a-based assays can be used to identify epistatic interactions between genes and to generate gene networks based on changes in co-expression in perturbed versus non-perturbed cells (**Fig. 3B**). Genes that are strongly co-expressed are likely to be regulated by a shared mechanism [86]. Therefore, identifying such genes can help reveal the gene network that leads to a disease-associated trait [94,108,109]. Indeed, a CRISPRi screen that targeted 12 TFs, chromatin modifying factors and non-coding RNAs was able to identify epistatic effects in cells perturbed by two guide RNAs [110]. In these cells, chromatin accessibility remained relatively stable in loci associated to auto-immune disease in cells with one perturbed TF. However, significant changes were observed when evaluating the chromatin accessibility for the same loci in cells also perturbed for *NFKB1*. This again highlights the importance of taking the entire context of a trait into account when fine-mapping or interpreting the role of a GWAS locus.

A major drawback of the majority of CRISPRi/a screens is that they are very laborious and therefore usually performed in easily manipulated, but also highly modified, cancer cell-lines [61]. Fortunately, recent studies have shown that CRISPRi screens can be applied to primary T cells [111,112]. This, while challenging, needs to be extended to other tissues and model systems. These studies will greatly assist variant, regulatory region and gene fine-mapping efforts because they directly identify the active enhancer–gene pairs and the downstream gene network affected in specific cell-types. In addition, future work could focus on performing CRISPRi/a screens in patient-derived cells that contain relevant risk genotypes to fully reach variant-level resolution.

3.4 Mapping gene–gene regulatory interactions using population data

Co-expression can also be modelled based on inter-individual variation in expression, which can be used to prioritize disease genes and make inferences about the downstream consequences of diseases (**Fig. 3C**) [94,108,109,113]. For example, DEPICT (electronic supplementary material, **Suppl. Tab. 3**), integrates gene co-regulation with GWAS data to provide likely causal genes and pathways relevant for the trait [113]. Moreover, the GADO tool (electronic supplementary material, **Suppl. Tab. 3**) correctly identified causal genes in 41% of a cohort of 83 patients with varying mendelian disorders, and prioritized several novel causal candidate genes by combining trait-specific gene sets with a co-expression network [109]. Finally, eMAGMA (electronic supplementary material, **Suppl. Tab. 3**) utilized co-expression together with tissue-specific eQTLs in brain regions to prioritize 99 candidate causal genes for major depressive disorder [94]. These co-expression modules were enriched in brain regions but not in whole-blood, highlighting the tissue-specific nature of the co-expression networks [94].

Population-based co-expression networks describe the relationships between genes through both genetics and environment. Consequently, based on the co-expression alone, it is not possible to separate which part of the co-expression is due to genetics. Therefore, these networks have limited use for fine-mapping causal variants and are mainly used to identify genes and pathways affected by GWAS loci after gene prioritizations have been made. In addition, co-expression networks are not directed [108]. Genetic information of the individuals used to generate the co-expression network would solve this issue, as the genetic and environmental components could be separated and directionality could be added into the network [108], although this is not a trivial task. Fine-mapping would be of great value in modelling the genetic component of the network by facilitating the selection of true causal variants.

3.5 Fine-mapping under the omnigenic model

As discussed throughout this review, it is becoming increasingly clear that complex traits are highly polygenic and that many variants can deregulate *cis*- and *trans*-acting factors in a variety of ways (**Fig. 2A**). In light of this, Boyle et al. [87] proposed an omnigenic model for complex traits in which each gene that is expressed in the cell will have an effect on the trait or disease in some way (**Fig. 1C**) [87,88]. For example, height is so polygenic that most 100kb genomic windows seem to contribute to explaining its variance. Given that the effect sizes of the individual variant are getting so small, it begs the question: What does the causality of the individual variant mean in a complex trait [87,88,114]? If the omnigenic model is true, it presents a major challenge for fine-mapping GWAS loci, particularly for the interpretation of the downstream consequences as the complexity of genetic effects on traits will only increase. In addition, current functional assays may not be suited to model the small and subtle variant effects and gene–gene or gene–environment interactions observed in population studies using millions of individuals.

Instead, the complete GWAS signal from all loci associated with a trait can be used to estimate a polygenic score (PGS) that describes an individual's genetic predisposition for the given trait. In its most basic form, a polygenic score constitutes the linear combination of all independent risk genotypes weighted by the GWAS effect size, but many more sophisticated methods exist (**Fig. 3D**) [115–117]. The PGS for a trait can be associated to the expression level of genes (and proteins) in a population [72,118]. If there are strong correlations, GWAS loci together, as represented by the PGS, are jointly influencing these genes. These genes likely represent core genes in a disease-associated co-expression network. Although PGS have issues when it comes to broad applicability across populations [119], they can be a useful abstraction layer to make sense of a polygenic trait.

Given we are becoming aware of the likely poly- and even omni-genic nature of traits, fine-mapping the individual GWAS locus seems like an impossible task. However, with current approaches the stronger, and arguably more important, genetic effects associated with traits and diseases can be elucidated [70,72,73]. Moreover, by using abstraction layers such as PGS, inferences can be made about the joint consequences of these effects [72]. Indeed, the genes and pathways associated with stronger or joint genetic effects are more likely candidates for drug interventions [120] (electronic supplementary material, **Suppl. Tab. 1**). Although we might never fully comprehend all the tiny effects and interactions underlying a trait, we will likely see an increase in clever ways to arrive at the interpretable biological mechanisms behind traits.

Future perspectives

We have reviewed recent high-throughput GWAS fine-mapping approaches that can identify variants and genes causal for a trait or disease. The complexity and uncertainty present in aspects of these approaches illustrates that a single approach does not suffice to grasp the full cause and effect of candidate variants and genes. In addition, while large datasets, mostly in blood, have identified many potentially causal variants and genes associated with traits, these candidates need to be refined and validated using tissue- and cell-type-specific resources in combination with trait-specific environmental factors to recapitulate the true biological state of each trait as closely as possible. An additional challenge lies in translating these disease genes into clinical practice, as prioritized genes might not be existing, nor practical, drug targets.

Despite these challenges, we believe that combining the use of patient-derived material, with methods that find regulatory regions and their downstream genes will aid drug target identification for complex diseases. In addition, this knowledge could be used to generate prediction models that aid in the fast and non-invasive identification of trait-specific variants and genes in the general population. This will form the foundation of our understanding of complex traits, aid drug development and will allow tailored precision medicine in the near future.

Acknowledgements

We acknowledge Kate McIntyre for editorial assistance and critically reading the manuscript.

Funding

O.B. is supported by an NWO VIDI grant (016.171.047) and an NWO VENI grant (NWO 863.13.011). I.J. and R.B. are supported by a Rosalind Franklin Fellowship from the University of Groningen and an NWO VIDI grant (016.171.047).

Data accessibility

This article has no additional data.

Supplementary material

Supplemenatary material are provided at: https://doi.org/10.1098/rsob.190221

Suppl. Tab. 1-3 can be found in the additional Excel file.

Competing interests

None of the authors have competing interests.

Authors' contributions

R.B. and O.B. conceived and wrote the manuscript. I.J. wrote and critically edited.

Media summary

Complex diseases like type 1 diabetes or cardiovascular disease all have genetic underpinnings that can increase risk or exacerbate symptoms. However, how genetics contributes to disease is not very well understood. This review describes the ways in which science has tried to elucidate the effects of genetics on the molecular mechanisms, genes and pathways that are important for diseases. Moreover, as it is becoming clear that the role of genetics in disease may be more complex than initially thought, we speculate on how the field should progress to ensure that genetics can truly contribute to the understanding of complex diseases.

References

1. Morahan, G. et al. Tests for genetic interactions in type 1 diabetes linkage and stratification analyses of 4,422 affected sib-pairs. Diabetes 60, 1030–1040 (2011).
2. Trynka, G. et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in coeliac disease. Nature Genetics (2011) doi:10.1038/ng.998.
3. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700 000 individuals of European ancestry. Human Molecular Genetics 27, 3641–3649 (2018).
4. Slatkin, M. Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 9, 477–485 (2008).
5. Ozaki, K. et al. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nature Genetics 32, 650–654 (2002).
6. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research (2019) doi:10.1093/nar/gky1120.
7. Kumar, V., Wijmenga, C. & Withoff, S. From genome-wide association studies to disease mechanisms: Coeliac disease as a model for autoimmune diseases. Seminars in Immunopathology 34, 567–580 (2012).
8. Belmont, J. W. et al. A haplotype map of the human genome. Nature (2005) doi:10.1038/nature04226.
9. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. Nature (2014) doi:10.1038/nature12787.
10. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. Nature Reviews Molecular Cell Biology 19, 621–637 (2018).
11. Lambert, S. A. et al. The Human Transcription Factors. Cell 172, 650–665 (2018).
12. Farh, K. K. H. et al. Genetic and epigenetic fine-mapping of causal autoimmune disease variants. Nature 518, 337–343 (2015).
13. Corradin, O. & Scacheri, P. C. Enhancer variants: Evaluating functions in common disease. Genome Medicine 6, 1–14 (2014).
14. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. Human Molecular Genetics 24, R111–R119 (2015).
15. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics 19, 491–504 (2018).
16. Weissenkampen, J. D. et al. Methods for the Analysis and Interpretation for Rare Variants Associated with Complex Traits. Current Protocols in Human Genetics (2019) doi:10.1002/cphg.83.
17. Tak, Y. G. & Farnham, P. J. Making sense of GWAS: Using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics and Chromatin 8, 1–18 (2015).
18. Maurano, M. T. et al. Transcription Factor Occupancy in Vivo. 47, 1393–1401 (2016).
19. Javierre, B. M. et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell 167, 1369-1384.e19 (2016).
20. Bernstein, B. E. et al. The NIH roadmap epigenomics mapping consortium. Nature Biotechnology (2010) doi:10.1038/nbt1010-1045.
21. Yen, A. et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015).
22. Onengut-gumuscu, S. et al. Nihms665498. 47, 381–386 (2015).
23. Chen, L. et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell vol. 167 (2016).
24. Trynka, G. et al. Chromatin marks identify critical cell types for fine-mapping complex trait variants. Nature Genetics (2013) doi:10.1038/ng.2504.
25. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nature Genetics 48, 206–213 (2016).
26. Matthew T. Maurano, Richard Humbert, Eric Rynes1, Robert E. Thurman, E. et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science 337, 1190–1195 (2012).
27. Core, L. J. et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature Genetics (2014) doi:10.1038/ng.3142.
28. Jonkers, I., Kwak, H. & Lis, J. T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. eLife 2014, 1–25 (2014).
29. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science (2008) doi:10.1126/science.1162228.
30. Mahat, D. B. et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nature Protocols 11, 1455–1476 (2016).
31. Shiraki, T. et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proceedings of the National Academy of Sciences of the United States of America (2003) doi:10.1073/pnas.2136655100.
32. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
33. Forrest, A. R. R. et al. A promoter-level mammalian expression atlas. Nature (2014) doi:10.1038/nature13182.
34. Stunnenberg, H. G. et al. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell (2016) doi:10.1016/j.cell.2016.11.007.
35. Benton, M. L., Talipineni, S. C., Kostka, D. & Capra, J. A. Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. BMC Genomics 20, 1–22 (2019).
36. Qu, K. et al. Individuality and Variation of Personal Regulomes in Primary Human T Cells. Cell Systems 1, 51–61 (2015).
37. Hsiao, Y. H. E. et al. Alternative splicing modulated by genetic variants demonstrates accelerated evolution regulated by highly conserved proteins. Genome Research 26, 440–450 (2016).

38. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. American Journal of Human Genetics 102, 11–26 (2018).
39. Paraboschi, E. M. et al. Functional variations modulating PRKCA expression and alternative splicing predispose to multiple sclerosis. Human Molecular Genetics 23, 6746–6761 (2014).
40. Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. Science 352, 600–604 (2016).
41. Fiszbein, A., Krick, K. S. & Burge, C. B. Exon-mediated activation of transcription starts. bioRxiv 565184 (2019) doi:10.1101/565184.
42. Zhang, C. et al. A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. Nucleic Acids Research 34, 2238–2246 (2006).
43. Degner, J. F. et al. DNase-I sensitivity QTLs are a major determinant of human expression variation. Nature 482, 390–394 (2012).
44. Rastogi, C. et al. Accurate and sensitive quantification of protein-DNA binding affinity. Proceedings of the National Academy of Sciences of the United States of America 115, E3692–E3701 (2018).
45. Khan, A. et al. JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Research 46, D260–D266 (2018).
46. Jolma, A. et al. DNA-binding specificities of human transcription factors. Cell 152, 327–339 (2013).
47. Alasoo, K. et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nature Genetics 50, 424–431 (2018).
48. Mansour, M. R. et al. Mutation of a Noncoding Intergenic Element. Science 346, 1373–1377 (2016).
49. Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA Binding Variation. Cell 166, 538–554 (2016).
50. Jolma, A. et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature 527, 384–388 (2015).
51. Harley, J. B. et al. Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity. Nature Genetics 50, 699–707 (2018).
52. Bouziat, R. et al. Reovirus infection triggers inflammatory responses to dietary antigens and development of coeliac disease. Science 356, 44–50 (2017).
53. Tarlinton, R. E. et al. The interaction between viral and environmental risk factors in the pathogenesis of multiple sclerosis. International Journal of Molecular Sciences 20, 1–16 (2019).
54. Garieri, M. et al. The effect of genetic variation on promoter usage and enhancer activity. Nature Communications 8, 1–7 (2017).
55. Claussnitzer, M. et al. FTO obesity variant circuitry and adipocyte browning in humans. New England Journal of Medicine 373, 895–907 (2015).
56. Tewhey, R. et al. Variants Using a Multiplexed Reporter Assay. 165, 1519–1529 (2017).
57. Liu, S. et al. Systematic identification of regulatory variants associated with cancer risk. Genome Biology 18, 1–14 (2017).
58. Van Arensbergen, J. et al. Genome-wide mapping of autonomous promoter activity in human cells. Nature Biotechnology (2017) doi:10.1038/nbt.3754.
59. van Arensbergen, J. et al. High-throughput identification of human SNPs affecting regulatory element activity. Nature Genetics 51, (2019).
60. Jonkers, I. H. & Wijmenga, C. Context-specific effects of genetic variants associated with autoimmune disease. Human Molecular Genetics 26, R185–R192 (2017).
61. Ben-David, U. et al. Genetic and transcriptional evolution alters cancer cell line drug response. Nature 560, 325–330 (2018).
62. Chun, S. et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nature Genetics 49, 600–605 (2017).
63. Mumbach, M. R. et al. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nature Genetics 49, 1602–1612 (2017).
64. Kumasaka, N., Knights, A. J. & Gaffney, D. J. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. Nature Genetics 51, 128–137 (2019).
65. Smemo, S. et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature 507, 371–375 (2014).
66. Ulirsch, J. C. et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. Nature Genetics 51, 683–693 (2019).
67. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nature Genetics 47, 598–606 (2015).
68. Schmitt, A. D. et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Reports 17, 2042–2059 (2016).
69. Gao, T. & Qian, J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Research (2019) doi:10.1093/nar/gkz980.
70. Westra, H. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature Genetics 6, 247–253 (2014).
71. Zhernakova, D. V. et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nature Genetics 49, 139–145 (2017).
72. Võsa, U. et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv 447367 (2018) doi:10.1101/447367.
73. Piasecka, B. et al. Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. Proceedings of the National Academy of Sciences of the United States of America 115, E488–E497 (2018).
74. Lappalainen, T., et al. Transcriptome and genome sequencing uncovers functional variation in humans HHS Public Access Introduction and data set. Nature 501, 506–511 (2013).

75. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. American Journal of Human Genetics 88, 76–82 (2011).
76. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. Nature Communications 8, 1–10 (2017).
77. Giambartolomei, C. et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genetics 10, (2014).
78. Smith, G. D. & Ebrahim, S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? International Journal of Epidemiology 32, 1–22 (2003).
79. Porcu, E. et al. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. Nature Communications 10, 3300 (2019).
80. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature Genetics 48, 481–487 (2016).
81. Graaf, A. van der et al. A novel Mendelian randomization method identifies causal relationships between gene expression and low-density lipoprotein cholesterol levels. bioRxiv 671537 (2019) doi:10.1101/671537.
82. Morrison, J., Knoblauch, N., Marcus, J., Stephens, M. & He, X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. bioRxiv 682237 (2019) doi:10.1101/682237.
83. Hemani, G. et al. The MR-base platform supports systematic causal inference across the human phenome. eLife (2018) doi:10.7554/eLife.34408.
84. Verbanck, M., Chen, C. Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nature Genetics 50, 693–698 (2018).
85. Musunuru, K. et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 466, 714–719 (2010).
86. Morloy, M. et al. Genetic analysis of genome-wide variation in human gene expression. Nature 430, 743–747 (2004).
87. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169, 1177–1186 (2017).
88. Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell 177, 1022-1034.e6 (2019).
89. Wills, Q. F. et al. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. Nature Biotechnology 31, 748–752 (2013).
90. Wijst, M. G. P. Van Der, Brugge, H., Vries, D. H. De, Deelen, P. & Morris, A. Europe PMC Funders Group Single-cell RNA sequencing identifies cell type-specific cis - eQTLs and co-expression QTLs. 50, 493–497 (2018).
91. Watanabe, K., Umićević Mirkov, M., de Leeuw, C. A., van den Heuvel, M. P. & Posthuma, D. Genetic mapping of cell type specificity for complex traits. Nature Communications 10, (2019).
92. Carithers, L. J. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project. Biopreservation and Biobanking 13, 307–308 (2015).
93. Hernandez, D. G. et al. Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. Neurobiology of Disease (2012) doi:10.1016/j.nbd.2012.03.020.
94. Gerring, Z. F., Gamazon, E. R. & Derks, E. M. A gene co-expression network-based analysis of multiple brain tissues reveals novel genes and molecular pathways underlying major depression. PLOS Genetics 15, e1008245 (2019).
95. Human Cell Atlas. https://www.humancellatlas.org/.
96. sc-eQTLGen. https://eqtlgen.org/single-cell.html.
97. The LifeTime Initiative - LifeTime FET Flagship. https://lifetime-fetflagship.eu/.
98. Li, Y. et al. Inter-individual variability and genetic influences on cytokine responses against bacterial and fungal pathogens HHS Public Access Author manuscript. 22, 952–960 (2017).
99. Li, Y. et al. A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans. Cell 167, 1099-1110.e14 (2016).
100. Kettunen, J. et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. Nature Communications 7, 1–9 (2016).
101. Bonder, M. J. et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nature Genetics 49, 131–138 (2017).
102. Wang, J. et al. Meta-analysis of human genome-microbiome association studies: The MiBioGen consortium initiative. Microbiome 6, 1–7 (2018).
103. Orrù, V. et al. Genetic variants regulating immune cell levels in health and disease. Cell (2013) doi:10.1016/j.cell.2013.08.041.
104. Aguirre-Gamboa, R. et al. Differential Effects of Environmental and Genetic Factors on T and B Cell Immune Traits. Cell Reports (2016) doi:10.1016/j.celrep.2016.10.053.
105. Shalem, O., Sanjana, N. E. & Zhang, F. High-throughput functional genomics using CRISPR-Cas9. Nature Reviews Genetics (2015) doi:10.1038/nrg3899.
106. Horlbeck, M. A. et al. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. eLife 5, 1–20 (2016).
107. Gasperini, M. et al. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell 176, 377-390.e19 (2019).
108. Van Der Wijst, M. G. P., De Vries, D. H., Brugge, H., Westra, H. J. & Franke, L. An integrative approach for building personalized gene regulatory networks for precision medicine. Genome Medicine 10, 1–15 (2018).
109. Deelen, P. et al. Improving the diagnostic yield of exome- sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. Nature Communications 10, 1–13 (2019).
110. Rubin, A. J. et al. Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. Cell 176, 361-376.e17 (2019).

111. Shifrut, E. et al. Genome-wide CRISPR Screens in Primary Human T Cells Reveal Key Regulators of Immune Function. Cell (2018) doi:10.1016/j.cell.2018.10.024.
112. Gate, R. E. et al. Mapping gene regulatory networks of primary CD4+ T cells using single-cell genomics and genome engineering. bioRxiv (2019) doi:10.1101/678060.
113. Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. Nature Communications 6, (2015).
114. Visscher, P. M. et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. American Journal of Human Genetics 101, 5–22 (2017).
115. Wray, N. R. et al. Research Review: Polygenic methods and their application to psychiatric traits. Journal of Child Psychology and Psychiatry and Allied Disciplines 55, 1068–1087 (2014).
116. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nature Reviews Genetics 17, 392–406 (2016).
117. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics 50, 1219–1224 (2018).
118. Bakker, O. B. et al. Integration of multi-omics data and deep phenotyping enables prediction of cytokine responses. Nature Immunology 19, 776–786 (2018).
119. Martin, A. R. et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. American Journal of Human Genetics 100, 635–649 (2017).
120. Wishart, D. S. et al. DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Research 46, D1074–D1082 (2018).
121. Barrett T et al. 2013 NCBI GEO: Archive for functional genomics data sets - Update. Nucleic Acids Res. 41, 991–995. (doi:10.1093/nar/gks1193)
122. Papatheodorou I et al. 2018 Expression Atlas: Gene and protein expression across multiple studies and organisms. Nucleic Acids Res. 46, D246–D251. (doi:10.1093/nar/gkx1158)
123. Auton A et al. 2015 A global reference for human genetic variation. Nature 526, 68–74. (doi:10.1038/nature15393)
124. Francioli LC et al. 2014 Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. 46, 818–825. (doi:10.1038/ng.3021)
125. Ward LD, Kellis M. 2012 HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 40, 930–934. (doi:10.1093/nar/gkr917)
126. Bycroft C et al. 2018 The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209. (doi:10.1038/s41586-018-0579-z)
127. Fabregat A et al. 2018 The Reactome Pathway Knowledgebase. Nucleic Acids Res. 46, D649–D655. (doi:10.1093/nar/gkx1132)
128. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999 KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. (doi:10.1093/nar/27.1.29)
129. Ashburner M et al. 2000 Gene ontology: Tool for the unification of biology. Nat. Genet. (doi:10.1038/75556)
130. Carbon S et al. 2019 The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 47, D330–D338. (doi:10.1093/nar/gky1055)
131. Köhler S et al. 2019 Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. 47, D1018–D1027. (doi:10.1093/nar/gky1105)
132. Avsec Ž et al. 2019 The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. Nat. Biotechnol. 37, 592–600. (doi:10.1038/s41587-019-0140-0)
133. Wang K, Li M, Hakonarson H. 2010 ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, 1–7. (doi:10.1093/nar/gkq603)
134. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. 2015 Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience 4, 1–16. (doi:10.1186/s13742-015-0047-8)
135. Gamazon ER et al. 2015 A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. (doi:10.1038/ng.3367)
136. Lotfollahi M, Wolf FA, Theis FJ. 2019 scGen predicts single-cell perturbation responses. Nat. Methods 16, 715–721. (doi:10.1038/s41592-019-0494-8)
137. Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. 2016 Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. PLoS Comput. Biol. 12, 1–20. (doi:10.1371/journal.pcbi.1004714)

Chapter 3

Linking common and rare disease genetics through gene regulatory networks

Olivier B. Bakker [1], Annique Claringbould [2*], Harm-Jan Westra [1], Henry Wiersma [1], Floranne Boulogne [1], Urmo Võsa [3], Sophie Mulcahy Symmons [1], Iris H. Jonkers [1], Lude Franke [1,4#] and Patrick Deelen [1,4#]*

medRxiv (pre-print) 2021

1 University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands
2 Structural and Computational Biology Unit, EMBL, Heidelberg, Germany
3 Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia
4 Oncode Institute, Utrecht, The Netherlands
** These authors contributed equally*
# These authors contributed equally

Abstract

Genetic variants identified through genome-wide association studies (GWAS) are typically non-coding and exert small regulatory effects on downstream genes, but which downstream genes are ultimately impacted and how they confer risk remains mostly unclear. Conversely, variants that cause rare Mendelian diseases are often coding and have a more direct impact on disease development. We demonstrate that common and rare genetic diseases can be linked by studying the gene regulatory networks impacted by common disease-associated variants. We implemented this in the ‘Downstreamer’ method and applied it to 44 GWAS traits and find that predicted downstream “key genes” are enriched with Mendelian disease genes, e.g. key genes for height are enriched for genes that cause skeletal abnormalities and Ehlers-Danlos syndromes. We find that 82% of these key genes are located outside of GWAS loci, suggesting that they result from complex *trans* regulation rather than being impacted by disease-associated variants in cis. Finally, we discuss the challenges in reconstructing gene regulatory networks and provide a roadmap to improve identification of these highly connected genes for common traits and diseases.

Introduction

Genetic variation plays a major role in the development of both common and rare diseases, yet the genetic architectures of these disease types are usually considered quite different. Rare genetic disorders are thought to primarily be caused by a single, mostly protein-coding genetic variant that has a large effect on disease risk. As a consequence, the causal genes for a rare disorder can often be identified by sequencing individual patients or families. In contrast, the genetic risks for common diseases are modulated by a large number of mostly non-coding variants that individually exert small effects. These variants are typically identified through genome-wide association studies (GWASs). However, identification of the causal variants and genes affected by GWAS loci remains challenging, in part due to linkage disequilibrium (LD) and small effect-sizes [1,2].

Despite the differences between rare and complex diseases, it has been shown that GWAS loci for multiple traits are enriched for genes that can cause related rare diseases when damaged [3,4]. For instance, common variants associated to PR interval, a measurement of heart function, have been found within the *MYH6* gene [5], which is known to harbour mutations in individuals with familial dilated cardiomyopathy [6]. Moreover, eQTL studies have found examples of rare disease genes that are affected by distal common variants in trans, such as the immunodeficiency gene *ISG15*, which is affected by multiple systemic lupus erythematosus–associated variants [7]. These results indicate that common and rare diseases can result from damage to or altered regulation of the same genes, suggesting that the same biological pathways underlie these conditions [4]. However, it is not fully known to what extent specific genes and pathways are shared between rare and common diseases.

Over the years, many pathway-enrichment methods have been developed that can identify which biological pathways are enriched for common diseases [8-10] as well as highlighting their most likely cellular context(s) [11,12]. In addition, several methods can prioritize individual genes within GWAS susceptibility loci by studying how they are functionally related to genes in other susceptibility loci [8,13-16]. However, these methods confine themselves to genes in GWAS loci, potentially missing relevant *trans*-regulated up- or downstream effects. In blood, expression quantitative trait locus (eQTL) mapping has been successful in identifying the downstream *trans* regulatory consequences of GWAS-associated variants (i.e. *trans*-eQTLs and eQTSs, where polygenic scores are linked to expression levels) [7]. Unfortunately, large eQTL sample sizes are required to detect such effects, and such datasets are not yet available for most tissues.

Here we build upon the 'omnigenic model' hypothesis, which states that the genes that are most important in complex diseases are those that are modulated by many different common variants through gene regulatory networks [17,18]. The omnigenic model postulates that a limited number of core genes exist that drive diseases, but that many peripheral genes, which contain associated variants, contribute indirectly to disease development by modulating the activity of the core genes. Since the omnigenic model predicts that many core genes map outside GWAS loci, these genes will be missed by methods that prioritize genes within GWAS loci. The omnigenic model hypothesis is supported by recent works assessing RNA levels of blood cells [19] and molecular traits [20] and a large-scale *in vitro* knockdown experiment [21]. However, these studies were performed in blood, limiting their conclusions to GWAS studies on blood-related traits and immunological disorders.

To take this work further, we integrated (mRNA level) gene regulatory networks with GWAS summary statistics to prioritize key genes that we suspect are more likely to directly contribute to disease predisposition than genes in GWAS loci. We have implemented this strategy in a software package called ‘Downstreamer’ that uses GWAS summary statistics and gene co-regulation based on 31,499 multi-tissue RNA-seq samples in order to prioritize these key genes. We also provide pathway, rare disease phenotype (coded by HPO terms) and tissue enrichments to aid in the comprehensive interpretation of GWAS results.

We applied Downstreamer to 44 GWASs for a wide variety of traits (**Suppl. Tab. 1**) and show that the identified key genes are enriched for intolerance to loss-of-function (LoF) and missense (MiS) mutations and for Mendelian disease genes that lead to similar phenotypic outcomes as the GWAS trait. Specifically, we find that key genes for height are strongly enriched for severe growth defects and skeletal abnormalities in humans and mice. Additionally, key genes for auto-immune diseases point to lymphocyte checkpoints and regulators and those for glomerular filtration rate (GFR; a measure of kidney function) are transporters for several metabolites.

Key genes that cause Mendelian disease can therefore highlight the molecular pathways driving the complex disease. Conversely, predicted key genes may aid in identifying new Mendelian disease genes.

Results

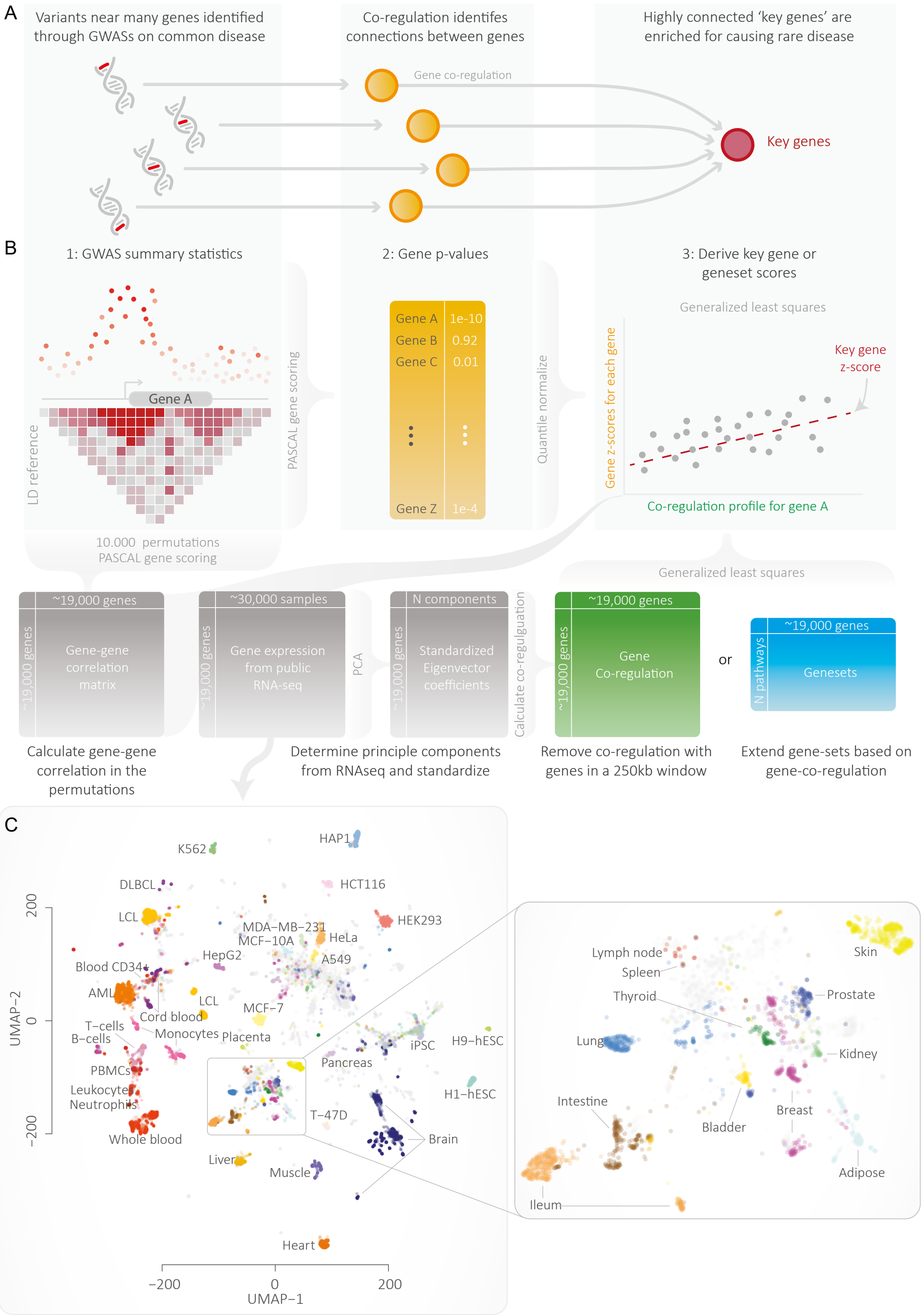
To enable identification of GWAS key genes, we developed Downstreamer (Methods), a tool that integrates GWAS summary statistics with gene expression–based co-regulation networks. Downstreamer first converts individual variant associations to gene p-values by aggregating associations within a 25kb window around the gene body for all protein-coding genes while correcting for LD between variants [9] (**Fig. 1A**). These gene p-values are then converted to z-scores. We calculated gene z-scores for 44 GWAS summary statistics reflecting a wide variety of disorders and complex traits (**Suppl. Tab. 1**).

Association signals for polygenic traits cluster around transcription factors

We observed that the z-scores of individual GWASs were often weakly positively correlated (**Suppl. Fig. 1A**), especially for traits for which many loci have been identified (**Suppl. Fig. 1B**). For instance, the gene-level z-scores for height correlated positively with the gene-level z-scores of all other traits. To investigate the source of this shared signal, we calculated the average gene-level significance across all 44 traits while correcting for bias that might be introduced for traits that are strongly correlated (see Methods).

We observed that 30% of the variation in this 'average GWAS' signal could be explained by both the extent of LD around a gene and by the local gene density (**Suppl. Fig. 2**). The more extensive the LD around a gene, the higher the chance that genetic variants within the gene are associated, especially for highly polygenic traits [22,23]. Consequently, the gene-level z-score for these genes increases. Hence, when collapsing GWAS summary statistics into gene z-scores, some amount of correlation between well-powered GWAS studies is to be expected. However, this is unwanted when using gene z-scores in a pathway-enrichment analysis. We next evaluated if the remaining 70% of the average signal was enriched for any biological processes. After correcting for LD and gene density, we observed that 79 of the top 500 genes are transcription factors (OR: 2.22, p-value: 4.25×10^{-9}). We also saw enrichment among the top 500 genes for pathways related to DNA binding and transcription, for example, transcription regulator activity (OR: 1.98, p-value: 2.24×10^{-11}) (**Suppl. Tab. 2**). Additionally, genes with a higher average gene z-score were enriched for intolerance to LoF (**Suppl. Fig. 3**). These enrichments suggest that there is a set of genes, enriched for GWAS hits, that confer risk to many different types of traits. This is consistent with previous observations that broad functional categories tend to be enriched for many traits [17].

However, as these often-associated genes obscure the specific pathways and key genes for a trait, we corrected the individual gene z-scores for the average gene z-score in order to get disease-specific gene-level significance scores that were as specific as possible. Downstreamer then correlates these corrected gene z-scores to gene expression patterns, pathway memberships and tissue expression through a generalized least squares (GLS) model that accounts for gene–gene correlations resulting from the relationship between LD and sharing of biological functionality (**Fig. 1B**, Methods). This results in a z-score that represents the significance of the association of a gene, pathway or tissue


A
Variants near many genes identified through GWASs on common disease
Co-regulation identifes connections between genes
Highly connected 'key genes' are enriched for causing rare disease
Gene co-regulation
Key genes
B
1: GWAS summary statistics
2: Gene p-values
3: Derive key gene or geneset scores
Gene A
LD reference
PASCAL gene scoring
Gene A 1e-10
Gene B 0.92
Gene C 0.01
Gene Z 1e-4
Quantile normalize
Generalized least squares
Gene z-scores for each gene
Key gene z-score
Co-regulation profile for gene A
10.000 permutations
PASCAL gene scoring
Generalized least squares
~19,000 genes
~19,000 genes
Gene-gene correlation matrix
~30,000 samples
~19,000 genes
Gene expression from public RNA-seq
PCA
N components
~19,000 genes
Standardized Eigenvector coefficients
Calculate co-regulguation
~19,000 genes
~19,000 genes
Gene Co-regulation
or
~19,000 genes
N pathways
Genesets
Calculate gene-gene correlation in the permutations
Determine principle components from RNAseq and standardize
Remove co-regulation with genes in a 250kb window
Extend gene-sets based on gene-co-regulation
C
K562
HAP1
DLBCL
HCT116
LCL
HEK293
MDA−MB−231
MCF−10A
HeLa
HepG2
A549
Blood CD34+
AML
LCL
Cord blood
MCF−7
T−cells
B−cells
Monocytes
Placenta
iPSC
H9−hESC
PBMCs
Leukocytes
Neutrophils
Pancreas
H1−hESC
T−47D
Whole blood
Brain
Liver
Muscle
Heart
UMAP−2
200
0
−200
−200
0
200
UMAP−1
Lymph node
Spleen
Thyroid
Skin
Prostate
Lung
Kidney
Intestine
Bladder
Breast
Adipose
Ileum

Identification of key genes using gene co-expression

To identify key genes, we searched for genes that are co-regulated with the genes within a given trait's GWAS loci. We used a gene expression database containing 31,499 tissue and cell-line RNA-seq samples [24] to calculate gene co-regulation (a measure of the similarity of expression) for each pair of protein-coding genes (**Fig. 1**, Methods). Co-regulation is defined as the correlation between standardized eigenvector coefficients derived from the expression data (**Fig. 1B**). Since each component is given equal weight, co-regulation is less sensitive to the major tissue effects that can confound co-expression correlations calculated using expression data from a set of heterogeneous tissue samples. To ensure that no bias was introduced by GWAS loci located in highly co-regulated gene clusters, co-regulation relationships between genes within 250kb were removed to further compensate for these gene clusters resulting from genomic organisation. We then associated the gene z-scores to gene co-regulation using a GLS model. We use permutations to determine the significance of the association. The resulting association z-score reflects the overall connectivity of that gene to the GWAS genes in the network (**Fig. 1B**, Methods). We call this z-score the 'key gene score' throughout the manuscript, and we call the genes that pass Bonferroni significance and have a positive association 'key genes'. Besides detecting key genes, Downstreamer is also able to identify pathway and tissue enrichments for GWAS traits, using reconstituted gene sets that provide increased statistical power to identify significant pathway enrichment (**Note S1, Note S2**). Pathway and tissue enrichments results yielded plausible results consistent with previous findings, indicating that correction for the average GWAS signal is a useful addition.

In total, we identified 3,648 key genes over the 44 tested traits, with most key genes arising from GWASs for white blood cell composition and other haematological factors (**Fig. 2A**). The number of samples and independent loci for a GWAS is positively correlated to the number of detected key genes (Pearson R: 0.38 and 0.33, p-values: 1×10^{-2} and 2.8×10^{-2}, respectively; **Suppl. Fig. 4**), which is to be expected as larger GWASs typically contain more signal. To determine how similar the key gene predictions are, we correlated the key gene scores of the 44 different traits to each other and observed that traits of the same class cluster together (**Fig. 2B**). For example, the immune diseases (inflammatory bowel disease (IBD), coeliac disease (CeD), type 1 diabetes (T1D), rheumatoid arthritis (RA), asthma and multiple sclerosis (MS)) clustered together, neighboured by traits representing white blood cell composition. Other distinct clusters were found for psychological traits (educational attainment, schizophrenia, major depressive disorder, body mass index (BMI)) and cardiovascular traits (pulse pressure, diastolic and systolic blood pressure, coronary artery disease), further showing that the gene regulatory networks downstream of GWAS signals are partially shared between related traits. To some extent, this sharing is expected, given known co-morbidities between, for example, CeD and T1D [25] and the widespread genetic correlations of related complex traits [26].

Left: Fig. 1 A) Downstreamer works on the idea that many genes identified through GWAS jointly affect a set of key genes that strongly impact disease development. **B)** Schematic overview of Downstreamer methodology. **C)** The 31,499 RNA-seq samples used for the study visualized using Uniform Manifold Approximation and Projection (UMAP). The zoom-in on the right shows a detailed view of the various primary tissues in the dataset.

On average, 82% of predicted key genes are located outside GWAS loci (≥ 250kb from the lead variant) (**Fig. 2C**). This indicates that the key genes may be under *trans* regulation by the genes within the GWAS loci, rather than being impacted by a GWAS variants directly in cis, as is the case for most genes in GWAS loci. The other 18% of key genes are located within GWAS loci, suggesting that there is both a *cis* effect by a genetic variant that directly perturbs the function of these genes and a *trans* effect where the other GWAS loci modulate these genes.

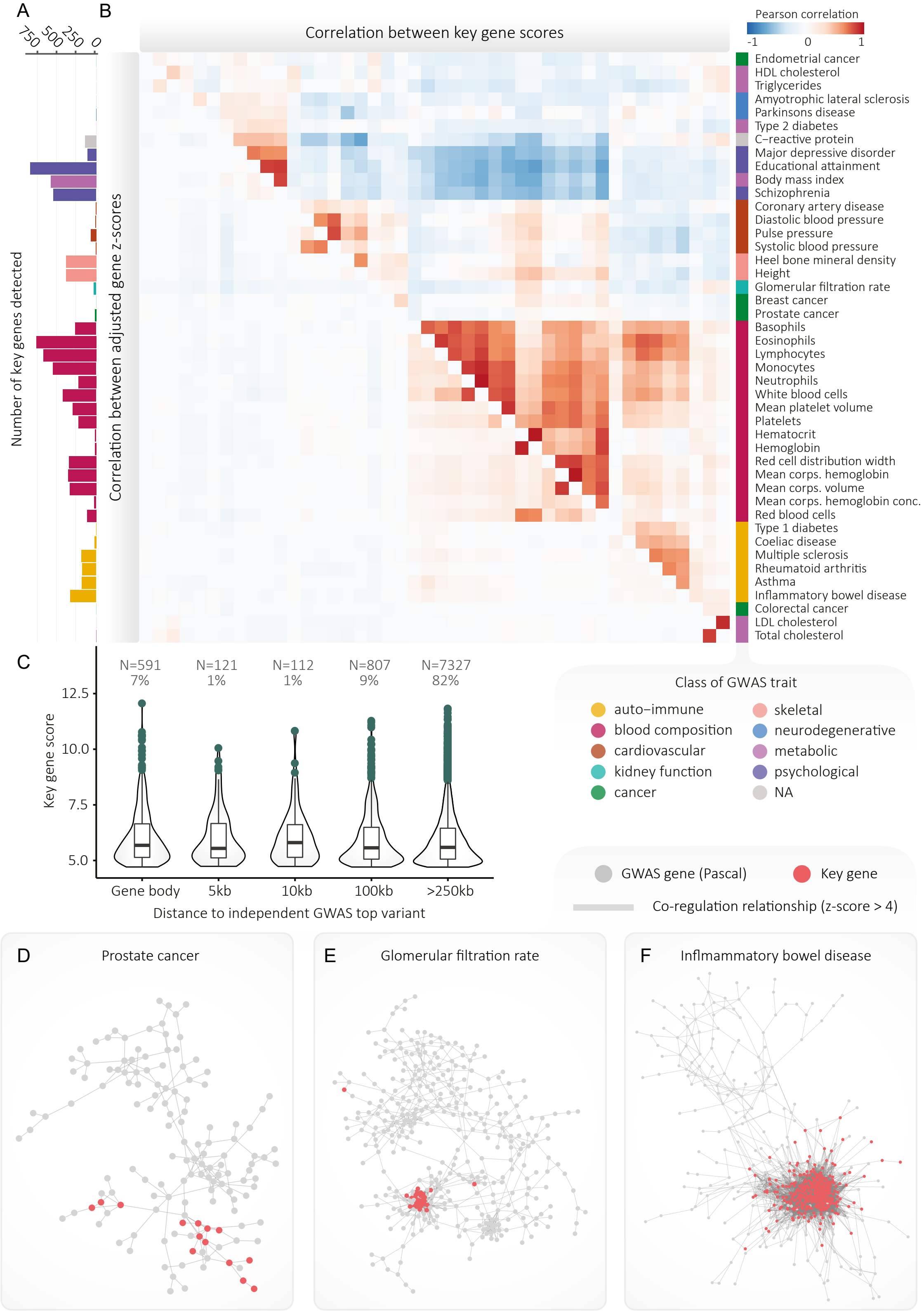
Of note, out of the 3,648 key genes detected, 2,036 (55%) were detected in multiple GWAS traits. However, this number is largely driven by the genes we detected for highly correlated traits such as blood cell composition (**Fig. 2A**, **Suppl. Fig. 5**). To better determine if key genes are trait-specific or shared among different diseases, we assigned each of the 44 traits to 10 broader classes such as auto-immune disease or blood cell composition. We then observed that 1,032 (28%) of the key genes are shared between at least two different classes (**Suppl. Fig. 5A**). This is largely driven by the overlapping key genes of blood cell composition and auto-immune disease, which account for 413 of the 1,032 overlapping key genes.

Below we highlight key genes for a few traits. For prostate cancer, we prioritized 14 key genes, 10 of these are outside the GWAS loci (**Fig. 2D**). The most notable are KLK3, which encodes for PSA (prostate-specific antigen), the marker that is used to screen for prostate cancer [27], and *KLK2*, which is known to activate *KLK3* [28]. Additionally, many other key genes we identified have either been implicated in prostate cancer [29–32] (*TMC5, MLPH, OVOL2* and *CHD1*) or in other types of cancer (*TFAP2C, BAIAP2L1* and *PLEKHN1*)[33–35].

The GWAS for GFR, a measurement of kidney function, revealed 32 key genes (**Fig. 2E**), of which 6 genes are solute carriers (a group of membrane transporters). Two of these, *SLC22A12* and *SLC17A1*, are known to be urate transporters, fitting the known relationship between urate levels and GFR [36]. Other notable GFR key genes include four glucuronosyltransferases (*UGT1A9, UGT2B7, UGT2A3* and *UGT1A6*) that are important in drug metabolism and clearance of drugs by the kidneys [37-39] and four genes (*UMOD, SLC22A12, SLC36A2* and *NPHS2*) known to cause rare forms of kidney disease [40].

The auto-immune diseases shared several key genes, such as *IL-2RA, ICOS* and *CD48*, indicating an adaptive immune signature. Recently, a large-scale CRISPR assay assessing the regulators of *IL-2, IL-2RA* and *CTLA4* systematically knocked down thousands of immune genes in primary CD4+ T cells in order to assess how these genes are co-regulated [21]. The genes regulating *IL-2RA* formed a highly inter-connected network, with the members of this network being significantly enriched for harbouring GWAS signals for MS. We prioritize *IL-2RA* as one of the most significant key genes for MS, but *IL-2RA* is

Right Fig. 2 A) Number of key genes detected for each GWAS tested. **B)** Pearson correlations between the gene z-scores after correction for the mean signal (lower triangle) and the key gene scores (upper triangle). Correlations were calculated using all protein-coding genes (including non-significant ones). **C)** Boxplots showing key gene scores in relation to the independent significant top SNPs from the GWAS. **D)** The gene regulatory network for prostate cancer. The network shows how prostate cancer GWAS genes and key genes are interconnected. Grey nodes represent GWAS genes. Red nodes indicate key genes. A key gene may also be located in a GWAS locus. Only positive co-regulation relationships with a z-score >4 are drawn as an edge in the network. **E)** As D, but showing the network for the GWAS and key genes for glomerular filtration rate. **F)** As D, but showing the network for the GWAS and key genes for inflammatory bowel disease.


A
B
750 500 250 0
Number of key genes detected
Correlation between key gene scores
Correlation between adjusted gene z-scores
Pearson correlation
-1 0 1
Endometrial cancer
HDL cholesterol
Triglycerides
Amyotrophic lateral sclerosis
Parkinsons disease
Type 2 diabetes
C−reactive protein
Major depressive disorder
Educational attainment
Body mass index
Schizophrenia
Coronary artery disease
Diastolic blood pressure
Pulse pressure
Systolic blood pressure
Heel bone mineral density
Height
Glomerular filtration rate
Breast cancer
Prostate cancer
Basophils
Eosinophils
Lymphocytes
Monocytes
Neutrophils
White blood cells
Mean platelet volume
Platelets
Hematocrit
Hemoglobin
Red cell distribution width
Mean corps. hemoglobin
Mean corps. volume
Mean corps. hemoglobin conc.
Red blood cells
Type 1 diabetes
Coeliac disease
Multiple sclerosis
Rheumatoid arthritis
Asthma
Inflammatory bowel disease
Colorectal cancer
LDL cholesterol
Total cholesterol
C
N=591 7%
N=121 1%
N=112 1%
N=807 9%
N=7327 82%
12.5
10.0
7.5
5.0
Key gene score
Gene body
5kb
10kb
100kb
>250kb
Distance to independent GWAS top variant
Class of GWAS trait
auto−immune
blood composition
cardiovascular
kidney function
cancer
skeletal
neurodegenerative
metabolic
psychological
NA
GWAS gene (Pascal)
Key gene
Co-regulation relationship (z-score > 4)
D
Prostate cancer
E
Glomerular filtration rate
F
Inflmammatory bowel disease

also a key gene for IBD, asthma, CeD, RA and white blood cells, consistent with the role of T cells in these diseases. For IBD, low-dose IL-2 has been shown to alleviate symptoms of DSS-induced colitis in mice, highlighting this pathway as a potentially viable therapeutic target [41]. Additionally, a duplication found in the *IL-2RA* locus, causing excessive IL-2 signalling, may predispose carriers to early-onset colitis [42].

Among the key genes identified for IBD (**Fig. 2F**), there are several known drug targets. Some targets were already identified through GWAS (e.g. *TNF, JAK2* and *PRKCB*). Others are located outside the GWAS loci, including *ITGB7*, which is one of the targets of Vedolizumab [43];*JAK3*, which is targeted by JAK inhibitors [44] and *S1PR4* which is targeted by Amiselimod [45]. Additionally, $^{\text{RGS1}}$ has been proposed to be a druggable target that protects against colitis when downregulated [46–48]. While $^{\text{RGS1}}$ has been associated to CeD [49] and MS [50], its loci were not identified by the IBD GWAS.

Similarly, for schizophrenia, we identify key genes within GWAS loci that are targeted by schizophrenia drugs (*RM3, GRIA1* and *GRIN2A*), as well as key genes located outside of the GWAS loci that are established drug targets or being tested as drug targets. These include *HTR1A*, which is target of aripiprazole [51]; *HTR5A* and *HTR1E*, which are both targeted by amisulpride [52] and *GRIA2* and *GRIA3*, which are both targets of topiramate [53].

Key genes can be depleted or enriched for cis-eQTL effects

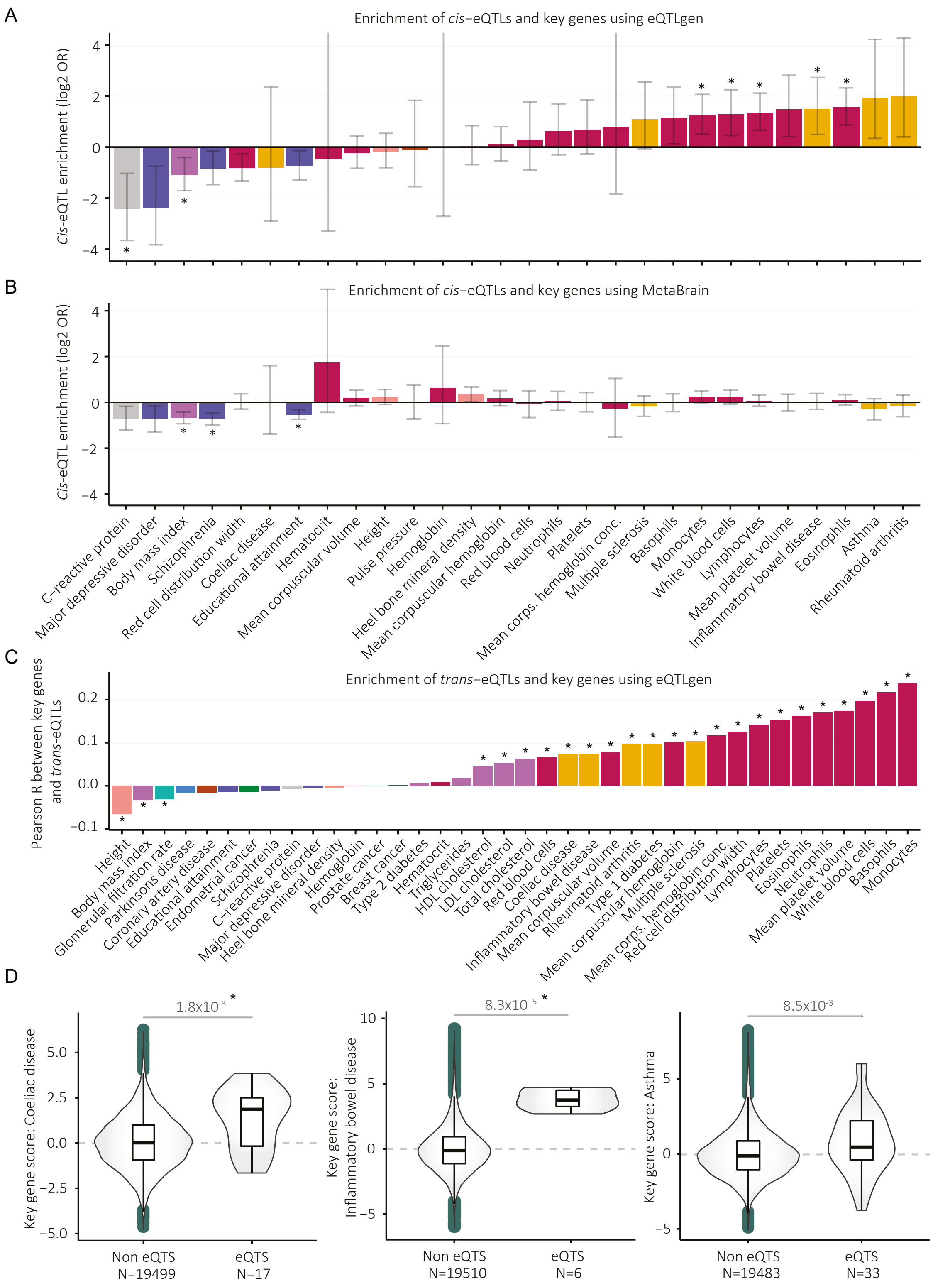
It has been observed in blood that genes without a detectable *cis*-eQTL effect are more intolerant to loss-of-function mutations [7]. This is possibly explained by more extensive buffering of regulatory effects on these important genes [54]. This has implications for the use of *cis*-eQTLs to identify disease-relevant genes. Here, we assessed whether key genes have fewer *cis*-eQTL effects than expected by chance. We did not observe consistent depletion of *cis*-eQTL. When testing 28 traits for which Downstreamer predicted at least 10 key genes, using blood-based *cis*-eQTLs from the eQTLgen consortium (**Fig. 3A**), we found Bonferroni significant enrichments for five traits (IBD and four different white blood cell count measurements, ORs: 2.36 – 2.93, p-values: 1.31×10^{-3} – 2.45×10^{-7}). Two traits were significantly depleted for blood *cis*-eQTLs: C-reactive protein levels (OR: 0.19, p-value: 4.68×10^{-4}) and BMI (OR: 0.47, p-value: 1.19×10^{-3}) (**Fig. 3A**). When using brain-based *cis*-QTLs from the MetaBrain project 55, we found three traits for which the key genes are significantly depleted for being *cis*-eQTLs: educational attainment (OR: 0.69, p-value 7.40×10^{-7}), BMI (OR: 0.63, p-value 1.17×10^{-7}), and schizophrenia (OR: 0.61, p-value 4.54×10^{-8}) (**Fig. 3B**), and observed no significant enrichment.

Overlap between key genes and trans regulatory targets

To investigate if the key genes result from *trans* regulation originating in the GWAS loci, we assessed if there was a correlation between the key gene scores and *trans*-

---

Right: Fig. 3 A) Enrichment of cis-eQTL genes and key genes. Cis eQTL genes are genes that have a significant cis eQTL effect in eQTLgen. * Indicates Bonferroni-adjusted p-values < 5×10^{-2} corrected for 28 tests. **B)** As A, but using MetaBrain cis eQTLs. **C)** Pearson correlation coefficients between key gene scores and the sum chi square statistics (see Methods) of the trans eQTL effects from significant independent GWAS SNPs. * Indicates Bonferroni-adjusted p-values < 5×10^{-2} corrected for 44 tests. **D)** Key gene scores for genes found to be in eQTS in the eQTLgen consortium for CeD, IBD and asthma. Nominal p-values of a t-test between the eQTS and non-eQTS groups are indicated. * Indicates Bonferroni-adjusted p-values < 5×10^{-2} corrected for 11 tests.


A
Enrichment of cis–eQTLs and key genes using eQTLgen
Cis-eQTL enrichment (log2 OR)
4
2
0
−2
−4
*
B
Enrichment of cis–eQTLs and key genes using MetaBrain
Cis-eQTL enrichment (log2 OR)
C–reactive protein
Major depressive disorder
Body mass index
Schizophrenia
Red cell distribution width
Coeliac disease
Educational attainment
Hematocrit
Mean corpuscular volume
Height
Pulse pressure
Hemoglobin
Heel bone mineral density
Mean corpuscular hemoglobin
Red blood cells
Neutrophils
Platelets
Mean corps. hemoglobin conc.
Multiple sclerosis
Basophils
Monocytes
White blood cells
Lymphocytes
Mean platelet volume
Inflammatory bowel disease
Eosinophils
Asthma
Rheumatoid arthritis
C
Enrichment of trans–eQTLs and key genes using eQTLgen
Pearson R between key genes and trans-eQTLs
0.2
0.1
0.0
−0.1
Height
Body mass index
Glomerular filtration rate
Parkinsons disease
Coronary artery disease
Educational attainment
Endometrial cancer
Schizophrenia
C–reactive protein
Major depressive disorder
Heel bone mineral density
Hemoglobin
Prostate cancer
Breast cancer
Type 2 diabetes
Hematocrit
Triglycerides
HDL cholesterol
LDL cholesterol
Total cholesterol
Red blood cells
Coeliac disease
Inflammatory bowel disease
Mean corpuscular volume
Rheumatoid arthritis
Type 1 diabetes
Mean corpuscular hemoglobin
Multiple sclerosis
Mean corps. hemoglobin conc.
Red cell distribution width
Lymphocytes
Platelets
Eosinophils
Neutrophils
Mean platelet volume
White blood cells
Basophils
Monocytes
D
$1.8x10^{-3}$ *
Key gene score: Coeliac disease
5.0
2.5
0.0
−2.5
−5.0
Non eQTS N=19499
eQTS N=17
$8.3x10^{-5}$ *
Key gene score: Inflammatory bowel disease
10
5
0
−5
Non eQTS N=19510
eQTS N=6
$8.5x10^{-3}$
Key gene score: Asthma
5
0
−5
Non eQTS N=19483
eQTS N=33

eQTL effects from the eQTLgen consortium [7]. To do so, for each gene, we summed the squared z-scores of *trans*-eQTL effects from the independent top hits for each GWAS. This results in a chi-square score for each gene that depicts to what extent the top GWAS variants of a trait affect the expression of those genes. We then correlated these scores with key gene scores. We found a significant association between the Downstreamer key gene scores and chi-square statistics of the eQTL effects for 24 of the 44 traits (Bonferroni-adjusted $P \leq 0.05$, **Fig. 3C**). Not surprisingly, the strongest associations were for the GWASs representing blood cell traits and auto-immune diseases. Interestingly, three non-blood traits – height, BMI and GFR – displayed significant negative correlations, which suggests the unsuitability of blood *trans*-eQTLs for interpreting non-blood traits, likely due to low expression of the relevant genes in blood [7].

Since *trans*-eQTL effects are typically small and current datasets are only powered to detect a fraction of these effects [7], the eQTLGen consortium correlated the polygenic scores for a diverse set of traits to gene expression levels. The genes with significantly lower or higher expression depending on the polygenic scores of the individuals (so-called eQTS genes) were prioritized as relevant for the trait. We observed that eQTS genes had higher key gene scores for three traits (t-test p-values, IBD: 1.8×10^{-3}, CeD: 8.3×10^{-5} and asthma: 8.5×10^{-3}), suggesting that key genes are more likely to be influenced by converging *trans*-eQTL effects (**Fig. 3D**).

We identified two genes that were both key genes and eQTS genes for asthma: *RELB* and *CST7*. *RELB* is a member of the NF-κB family of transcription factors that activate the non-canonical NF-κB pathway [56] and has been linked defects in T and B cell maturation, leading to combined immunodeficiency and auto-immune responses [57]. *CST7* has been described as a critical factor in maintaining eosinophile function [58], and eosinophiles are known to be one of the key cell types in asthma [59].

Key genes tend to be highly expressed in relevant tissues and cell types

As GWASs in the same class tended to show enrichment in the same cell types (**Data S1**, **Suppl. Fig. 7**) and shared key genes, we next tested if key genes were highly expressed in the cell types relevant for the corresponding trait. To determine the tissue specificity of each gene for a given tissue, we calculated a statistic for how highly a gene is expressed in that tissue by subtracting the mean expression of the samples of that tissue from the mean of all other samples in our dataset. This revealed significant association between the key gene scores and the expression level of genes in seemingly relevant tissues (**Fig. 4**, **Suppl. Fig. 8**-**10**), highlighting that the key genes tend to be highly expressed in the cell types where the GWAS is most enriched.

For example, several key genes for prostate cancer, such as *KLK3* (coding for the prostate-specific antigen), are highly expressed in prostate (**Fig. 4**). However, we also identified key genes for prostate cancer, such as *GPRS158*, that showed average expression in prostate but were much more highly expressed in other tissues such as muscle. Additionally, we also observed genes that were highly expressed in the prostate that were not key genes, such as *NKX3-1*. We saw similar examples for GFR and IBD (**Suppl. Fig. 9**, **Suppl. Fig. 10**). This indicates that the key gene prioritizations are in all likelihood not purely driven by tissue specific expression.

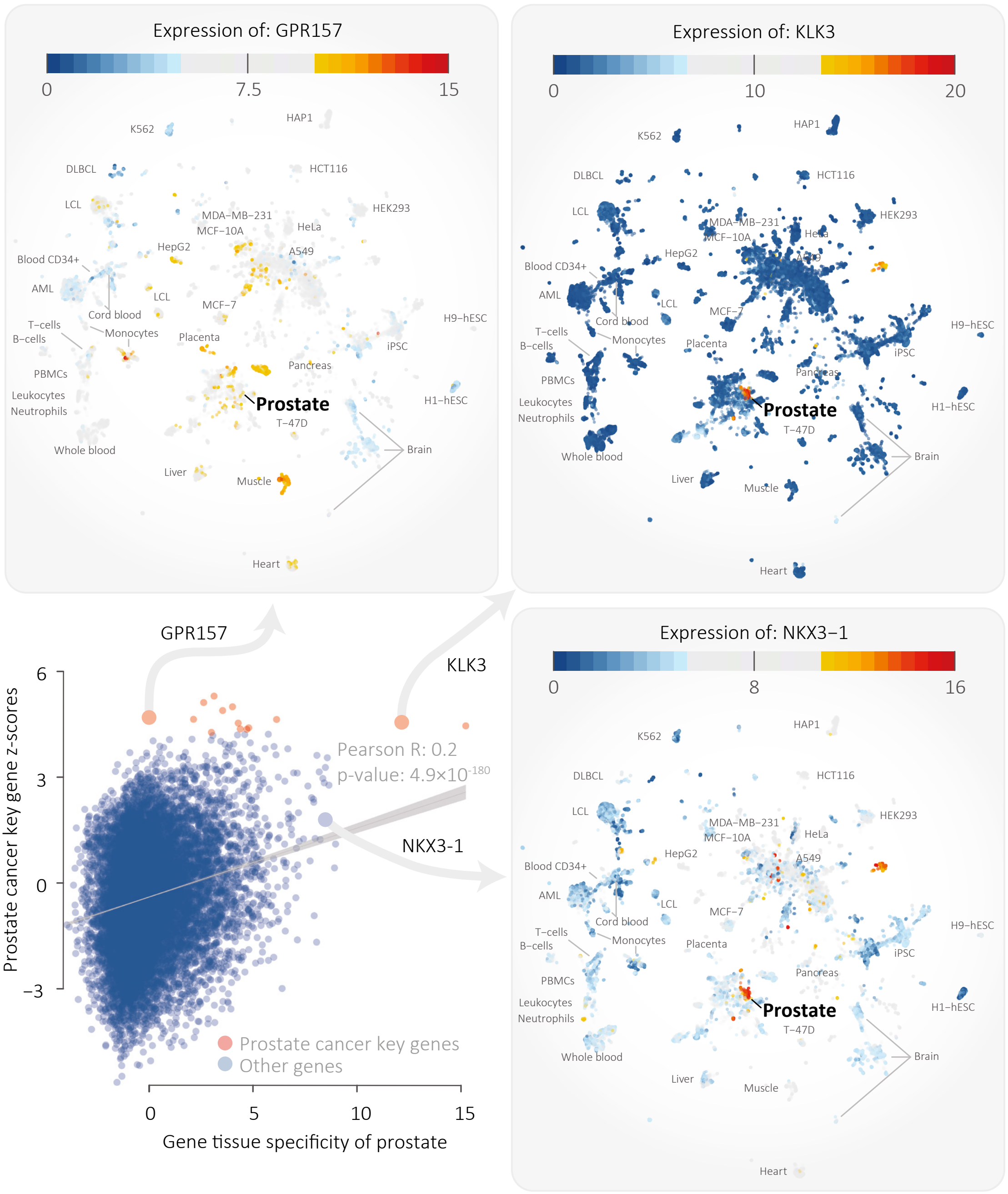


Fig. 4 Scatterplot showing the specificity of expression of a gene in prostate (x-axis) versus the key gene z-score of prostate cancer (y-axis). The Pearson correlation is 0.2 (p-value: 4.9×10^{-180}). Specificity of expression was determined by taking the mean in prostate samples and subtracting the mean from all other annotated samples in the dataset. The three panels show the expression of the highlighted genes, revealing that some key genes are specifically expressed in prostate but there are also key genes that are not prostate-specific and, vice-versa, there are prostate-specific genes that are not predicted to be key genes.

Key genes are more often constrained

We next reasoned that if key genes are essential to fundamental biological processes, they would be more likely to be more evolutionarily constrained. We therefore compared the key gene scores to the MiS and LoF z-scores from gnomAD [60]. These z-scores describe the tolerance level of a gene to MiS or LoF variants. The higher the score, the less tolerant the gene is to these types of variants. We observed significant association between the key gene scores and the MiS (Pearson R: 0.12, p-value: 3.7×10^{-57}) and LoF (Pearson R: 0.13, p-value: 1.4×10^{-7}) z-scores (**Suppl. Fig. 11**). Compared to the key genes, genes that map within GWAS loci had a lower LoF association (Pearson R: 0.07, p-value: 1.2×10^{-05}), but similar association with MiS (Pearson R: 0.12, p-value: 1.3×10^{-12}). Next, we evaluated if this association was driven by genes that have more connections in the gene network (i.e. whether a gene is a 'hub' gene or not), as we observed that the number of connections a gene has in the network is associated to the key gene score (**Suppl. Fig. 13**). After correcting for the number of connections a gene has, the associations for MiS and LoF remained, but were reduced (Pearson R: 0.07, 0.08, p-value 6.31×10^{-21} and 3.9×10^{-24}, respectively). Together, these results suggest that key genes tend to be evolutionarily constrained and are especially intolerant to LoF variants compared to the PASCAL gene p-values.

Key genes are enriched for Mendelian genes for related phenotypes

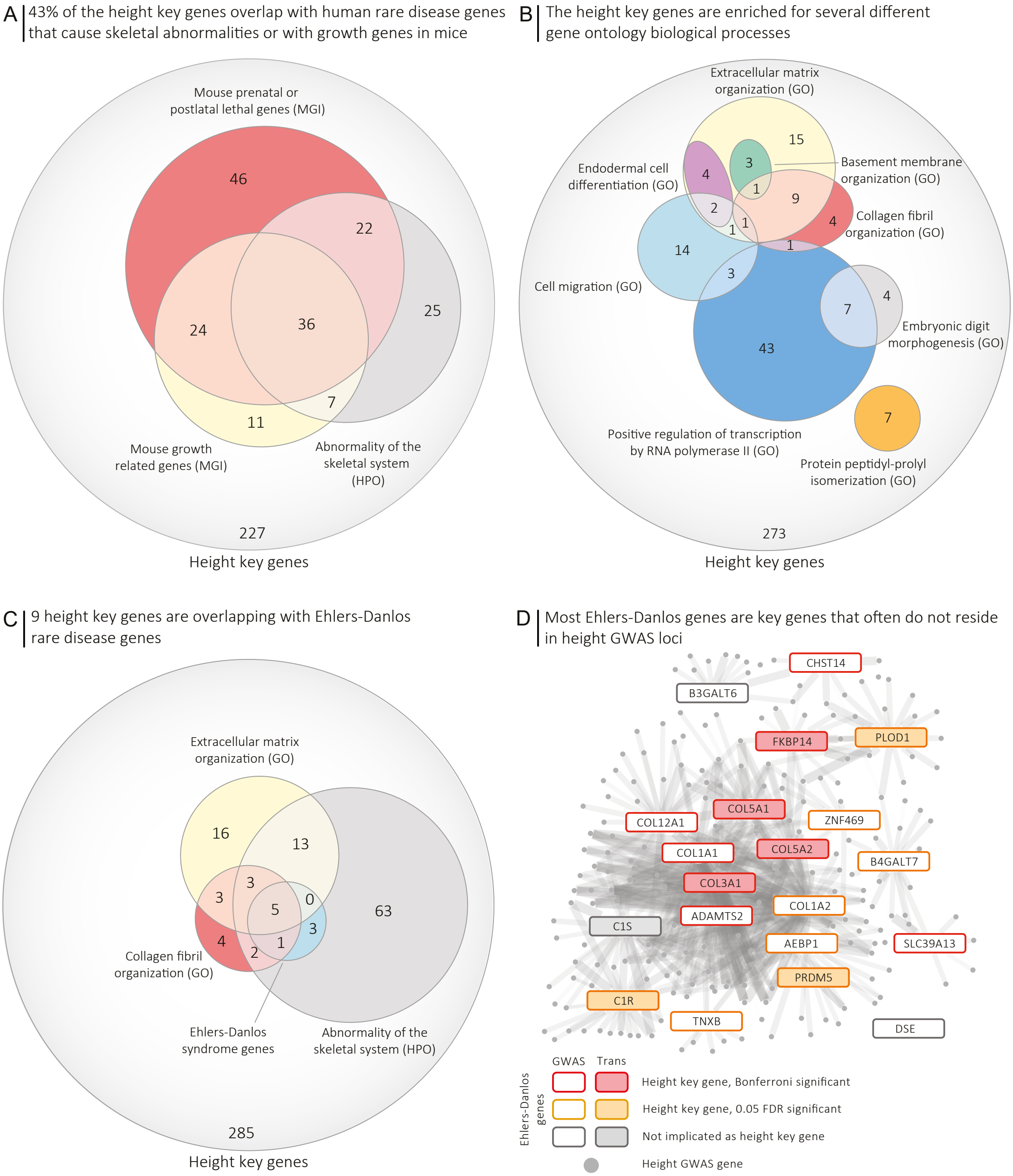
Genes in GWAS loci are known to be enriched for causing Mendelian diseases [3]. We observe that these enrichments of Mendelian disease genes are even stronger for the key genes that we prioritize. For example, we identified 398 Bonferroni-significant key genes for height and 90 (22.6%) of those are Mendelian disease genes causing "Abnormality of the skeletal system" (p-value: 5.18×10^{-9}, OR: 2.13, **Data S4**). This enrichment is stronger than for genes in the GWAS loci, where 319 of the 1,951 (16.4%) genes are annotated to cause "Abnormality of the skeletal system" (p-value: 7.86×10^{-9}, OR: 1.47, Data S2). Even when only considering the closest gene near the lead height GWAS hits, the enrichment of key-genes remains stronger (p-value: 7.02×10^{-12}, OR: 1.85, **Data S3**)

The most significant enrichment for height key genes is for "Abnormal lower limb bone morphology": 43 key genes are annotated to this HPO term (p-value: 6.86×10^{-15}, OR: 4.71). When also considering phenotypes based on mouse orthologs, we found 78 that are associated to growth and 128 that are pre- or post-natal lethal. In total, we can hereby explain 171 of the 398 (43%) associated height genes. These key genes are enriched for various pathways (**Fig. 5A**) including 'Collagen fibril organisation', 'Embryonic digit morphogenesis' and 'Extracellular matrix organisation'.

Among the Bonferroni-significant key genes for height are 9 of the 21 known Ehlers-Danlos genes (**Fig. 5A**). When using a less stringent FDR of 5%, we predict 17 out of the 21 Ehlers-Danlos genes to be key genes for height. Ehlers-Danlos syndromes are disorders of the connective tissues that often result in skeletal malformities [61]. These syndromes are caused by defects in, or related to, the collagen formation needed for the extracellular matrix. This is in line with the pathway enrichments of the height key genes and earlier findings [62].

Three genes with a Bonferroni significant gene p-value for IBD (*SKIV2L, NOD2, RTEL1*) overlapped with the 36 HPO-annotated colitis genes (OR: 4.87, p-value: 2.8×10^{-2}). The enrichment for colitis genes improved when assessing the key genes, which increased the overlap to six genes (*IL10RA, RASGRP1, NCF4, TNFAIP3, FASLG, ZAP70*; OR: 11.46, p-value: 3.3×10^{-5}). *IL10RA, RASGRP1, NCF4* and ZAP70 were all located further than 250kb from an independent *GWAS* hit in the GWAS used, meaning that these genes would not have been identified by overlapping IBD GWAS loci with known Mendelian genes. Other phenotypes that were significantly enriched among the key genes for IBD included those related to recurrent (fungal) infections and various phenotypes relating to immune function. Enriched mouse phenotypes included many related to T and B cell function and abundance (**Data S1**).

We matched each of the 44 GWAS traits to a best-fitting HPO term based on the phenotypic descriptions (**Suppl. Tab. 3**). We observed that 22% of the identified key genes are linked to rare diseases that cause related phenotypes (**Fig. 6A**, **Suppl. Tab. 4**). We found that the key genes for 18 of the 44 traits are significantly enriched (adjusted for 44 tests) for related rare disease genes (**Fig. 6B**; p-values: 2.98×10^{-4} to 4.91×10^{-29}, OR: 1.79 to 71.83). Another eight traits showed nominally significant overlap between key genes and related rare disease genes. For the 18 traits without significant overlap, we found between 0 and 20 key genes, indicating that our power for these traits was limited (**Fig. 6D**). The only exception to this was the GWAS for c-reactive protein levels, for which we found 148 key genes but only 6 genes were linked to its HPO term. Despite the limited power for these 18 traits, 9 traits had significantly larger key gene scores for the HPO-associated genes (U-test p-values: 8.73×10^{-12} to 8.28×10^{-4}), indicating that the key gene scores still have some predictive power for detecting rare disease genes (**Fig. 6C**).


A | 43% of the height key genes overlap with human rare disease genes that cause skeletal abnormalities or with growth genes in mice

B | The height key genes are enriched for several different gene ontology biological processes

C | 9 height key genes are overlapping with Ehlers-Danlos rare disease genes

D | Most Ehlers-Danlos genes are key genes that often do not reside in height GWAS loci

Mouse prenatal or postlatal lethal genes (MGI)
46
22
25
24
36
7
11
Mouse growth related genes (MGI)
Abnormality of the skeletal system (HPO)
227
Height key genes
Extracellular matrix organization (GO)
15
3
4
Basement membrane organization (GO)
Endodermal cell differentiation (GO)
1
9
2
1
1
4
Collagen fibril organization (GO)
14
1
3
Cell migration (GO)
4
7
Embryonic digit morphogenesis (GO)
43
7
Positive regulation of transcription by RNA polymerase II (GO)
Protein peptidyl-prolyl isomerization (GO)
273
Height key genes
Extracellular matrix organization (GO)
16
13
3
3
5
0
63
3
4
2
1
Collagen fibril organization (GO)
Ehlers-Danlos syndrome genes
Abnormality of the skeletal system (HPO)
285
Height key genes
CHST14
B3GALT6
FKBP14
PLOD1
COL5A1
ZNF469
COL12A1
COL1A1
COL5A2
B4GALT7
COL3A1
COL1A2
ADAMTS2
C1S
AEBP1
SLC39A13
PRDM5
C1R
TNXB
DSE
GWAS
Trans
Ehlers-Danlos genes
Height key gene, Bonferroni significant
Height key gene, 0.05 FDR significant
Not implicated as height key gene
Height GWAS gene

Fig. 5 A) 43% of height key genes are known to result in growth abnormalities in either humans or mice, indicating that these key genes are important genes for height. **B)** The height key genes are enriched with different, only partly overlapping, pathways, indicating that the key genes are part of multiple biological processes. **C)** Nine of the height key genes are known to cause Ehlers-Danlos syndromes, which involve abnormalities of the skeletal system. Most of these are annotated to the GO pathways for “Extracellular matrix organization” and “Collagen fibril organization”. It may be that the key genes that we now link to height and that are part of the collagen or extracellular matrix pathways also contribute to Ehlers-Danlos syndromes. **D)** The Ehlers-Danlos genes are co-expressed with many height GWAS loci genes.

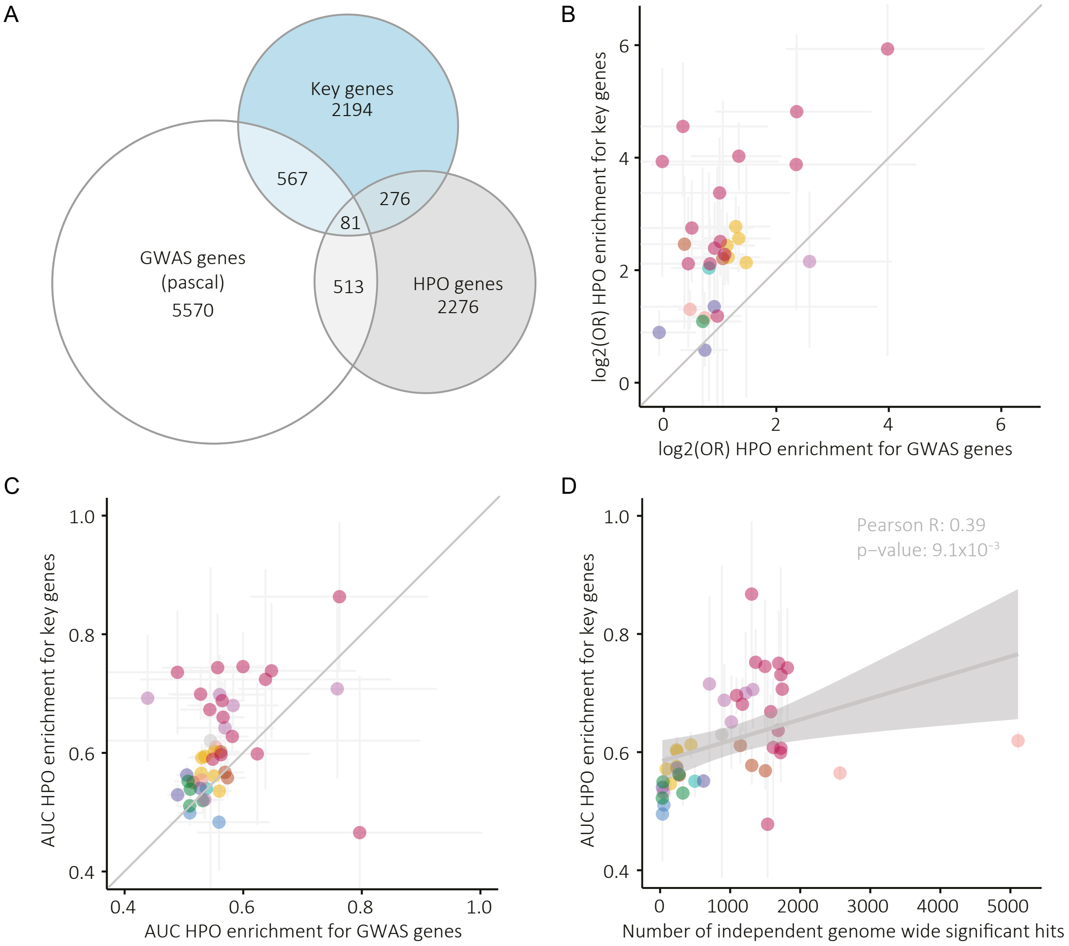


Fig. 6 A) Overview of the overlap between key genes, the genes for the 44 HPO terms we matched to their respective GWAS and the Bonferroni-significant genes identified by PASCAL. **B)** Comparison between the odds ratios of the HPO enrichment done using key genes and Bonferroni-significant GWAS genes identified by PASCAL. Each dot represents the HPO term matched to that GWAS. 95% confidence intervals of the odds ratios are represented. **C)** As in B but showing the AUC values calculated using the entire key gene z-score or GWAS gene p-value vector for all protein-coding genes. **D)** Association between the AUC values and the number of genome-wide significant hits for each GWAS.

Discussion

In this work, we present Downstreamer, a method that integrates gene co-regulation with GWAS summary statistics to prioritize genes central in the respective trait's network. We applied Downstreamer to 44 GWAS studies and prioritized genes that are not directly implicated by GWAS, yet are good candidates based on pathway annotation and their involvement in Mendelian diseases. Some of the genes showed evidence of being directly regulated by *trans*-acting genetic factors. The key genes are enriched for being evolutionarily constrained, indicating they more often have crucial biological functions. These findings suggest that the small effects of GWAS-associated variants ultimately converge on key disease genes.

We observed that the gene prioritization scores of related traits are often correlated (**Fig. 2B**). To some extent this is expected given the known shared genetic signature of, for instance, auto-immune disorders. It could also potentially indicate that the gene prioritizations are confounded by cell type–specific expression levels. Indeed, we found the genes prioritized for a trait to be more abundantly expressed in the samples that best match that trait (**Fig. 4**, **Suppl. Fig. 8**). However, this was not the sole driver of our key gene prioritization. We also found several examples of key genes that are not specifically expressed in the relevant tissues, as well as genes with very low key genes scores that show similar tissue-specific expression to the key genes. For instance, *GPR157* is predicted to be a key gene for prostate cancer, but it is highly expressed in many tissues (**Fig. 4**, **Suppl. Fig. 9**, **Suppl. Fig. 10**). Additionally, high expression of key genes in the disease-relevant tissue is to be expected because rare disease genes are also highly expressed in the tissue relevant to those diseases [63].

We recently also applied a pre-release version of Downstreamer to several neurodegenerative diseases, while using a comprehensive brain-specific gene co-regulation network of the MetaBrain project. This revealed that the signal of underrepresented cell types and tissues can be overshadowed by more abundant tissues in our expression data [55]. This might be especially relevant for diseases in which uncommon or rare cell types are instrumental to disease pathophysiology, such as gluten-specific T cells in CeD [64]. We therefore expect that key gene prioritizations can benefit from creating tissue- or even cell type–specific gene regulatory networks, should enough samples be available for the relevant tissue or cell type to accurately calculate co-expression. The future generation and inclusion of single-cell RNA-seq data should also be able to solve the issues regarding confounding by cell-type composition.

We observe that genes with *cis*-eQTL effects in blood are enriched for being key genes for blood traits. For instance, for IBD the key genes are more likely to be blood eQTLs than is expected by chance (**Fig. 3A**). A similar enrichment is seen at nominal significance levels for rheumatoid arthritis and asthma. Different results were found when using brain-derived *cis*-eQTLs. Using MetaBrain eQTLs, we found a depletion of *cis*-eQTL genes among the key genes of several brain related traits (**Fig. 3B**). This might indicate that genes that are important for the brain are more tightly controlled and are therefore not as easily affected by eQTL effects compared to important immune genes.

When comparing our prioritized genes to the results of a large blood-based *trans*-eQTL and eQTS analysis, we found an overlap in identified genes (**Fig. 3C**). As expected, this primarily holds for traits manifesting in blood, such as immune disorders and blood cell proportions. This confirms that a portion of the genes identified by Downstreamer are modulated by disease-associated variants in trans. We suspect four main causes for why there are key genes for which we cannot confirm *trans* regulation using blood-based *trans*-eQTLs or eQTSs: 1) the gene is not expressed in blood or the regulation is not present in blood, 2) the effects of genetic variants that only act in rarer blood cell types that are diluted by the expression levels in the more common cell types [65], 3) some trait-associated eQTLs depend on specific environmental stimuli [66], which can hinder the ability of population cohort studies to identify the regulatory consequences of disease-associated variants, and 4) Downstreamer works by integrating the many small effects originating from many different loci. Individually these effects might be too weak to currently be detected as *trans*-eQTL effects. We therefore conclude that co-expression-based methods such as Downstreamer are complementary to existing studies that link disease-associated variants to gene profiles.

One assumption we make when calculating the gene p-values is that the genes within 25kb of a GWAS signal are affected by the GWAS variants, but this might not the case for all genes. However, recent work has suggested that except for integrating epigenetic and HiC contact data, the next best predictor of causality is the closest gene to the top of the GWAS signal and this approach outperforms eQTL-based approaches [67]. We therefore decided not to integrate any prior eQTL information when calculating gene p-values, as this would often lead to incorrect prioritizations. In addition, the genes affected by GWAS variants are also likely to be tissue-specific, further complicating the prioritizations, and we would need extensive prior information to select the correct eQTLs or epigenetic information. This presents an area where major improvements could be made in future, when more accurate and systematic predictions can be made about which genes are regulated by GWAS variants in *cis*.

Our findings are in line with the infinitesimal model [68] that postulates that a quantitative trait or complex genetic disease can result from an infinite number of variants, each exerting an infinitely small effect size. An extension of the infinitesimal model is the omnigenic model [17], which predicts that all genes that are expressed in the relevant tissue or cell type will have a non-zero effect on disease outcome. The omnigenic model also postulates the existence of core genes that are pivotal in the development of a disease or trait. These core genes are expected to be enriched for genes that are involved in rare Mendelian diseases. The fact that key genes tend to be highly expressed in the relevant tissues for a trait, together with the enrichments of rare disease genes among the key genes, fits the regulatory pattern hypothesized in the omnigenic model. Hence, (some proportion of) the key genes we predict using Downstreamer could be the core genes described in the omnigenic model.

There is an important implication of the enrichment of key genes among known Mendelian disease genes for rare disease diagnostics. On average, a genetic cause is currently identified for only 30% of the patients with a suspected rare disease [69]. One of the reasons for this low diagnostic yield is that if a rare variant is found in a gene with an unknown function, it is difficult to determine if this variant could be causative for a patient's phenotype. We expect that in the future approaches like ours could be used to leverage the key genes of common diseases and traits to prioritize candidate rare disease genes in a manner similar to what we did previously using GADO [24].

In summary, we present Downstreamer, a method that integrates multi-tissue gene regulatory networks with GWAS summary statistics to prioritize key genes central in the gene network. These key genes were enriched for Mendelian variants that cause related phenotypes, highlighting that GWAS signals partially converge on Mendelian disease genes. While gaps remain in our understanding of the *trans* regulatory architecture of GWAS traits and diseases, assessing the genes most central in their respective regulatory network presents a promising way forward for interpreting both complex and rare disease genetics.

Methods

GWAS summary statistics

We downloaded the publicly available summary statistics from either the GWAS catalogue [70] or supplementary data files. A full list of the summary statistics used is available as **Suppl. Tab. 1**. Downstreamer requires rs identifiers (rsId) of the variants as well as the p-values. These were extracted from the summary statistic files, and any duplicate variants or variants without a rsId were removed. Where needed, summary statistics were lifted to build 37 and the rsIds matched on position and allele to 1000 Genomes phase 3 EUR for all variants with a minor allele frequency (MAF) > 0.05 [71].

Pathway databases

We used the following pathway and gene-set databases: Reactome [72], KEGG [73] and GO [74] (downloaded July 18, 2020), HPO [75] (filtered version as in [76]) and MGI (downloaded October 20, 2020) [77].

For the pathway enrichments below in step 2.2, we first expanded the known pathway annotations using the pathway predictions algorithm described in [76]. We expanded the pathway annotations with all genes with a Bonferroni-significant prediction of a pathway. Using the DEPICT algorithm, we have already shown that using predicted pathway annotations improves pathway enrichments [8]. Therefore, we used these expanded pathways when associating pathways to traits using Downstreamer.

Overview of Downstreamer methodology

In short, Downstreamer associates a gene-level prioritization score (GWAS gene z-scores) to a gene–gene co-regulation matrix to find genes that have many connections (at the expression level) to genes inside GWAS loci (core genes). In addition, Downstreamer can identify pathway enrichments by switching the co-regulation matrix for pathway annotations. Downstreamer implements a strategy that can perform these associations while accounting for LD structure and chromosomal organization. Downstreamer operates in two steps. In the first step, the GWAS gene z-scores are calculated for the GWAS trait and a null distribution. In the second step, the GWAS gene z-scores are associated with either the co-regulation matrix or the pathway annotations. Details of these steps are outlined in the sections below.

Downstreamer step 1.1. Calculation of GWAS gene z-scores

The first step in Downstreamer is to convert GWAS summary statistics from p-values per variant to an aggregate p-value per gene while accounting for local LD structure (1000 Genomes phase 3 EUR). This p-value is then converted to a gene z-score. This aggregate gene-level z-score represents the GWAS signal potentially attributable to that gene.

This was done as follows. First, we applied genomic control to correct for inflation in the GWAS signal. We then integrated the procedure from the PASCAL method into Downstreamer so that we can aggregate variant p-values into a gene p-value while accounting for the LD structure [9]. We aggregated all variants within a 25kb window around the start and end of a gene using the non-Finnish European samples of the

1000 Genomes (1000G) project, Phase 3 to calculate LD [71]. We calculated GWAS gene p-values for all 20,327 protein-coding genes (Ensembl release v75).

Downstreamer step 1.2. Null GWAS to account for chromosomal organization of genes and empirical p-value calculations.

To account for the longer-range effects of haplotype structure, which result in genes having a similar GWAS gene z-score, we use a GLS regression model for all regressions done in Downstreamer. The GLS model takes a correlation matrix that models this gene–gene correlation.

To calculate this correlation matrix, we first simulated 10,000 random phenotypes by drawing phenotypes from a normal distribution and then associating them to the genotypes of the 1000G Phase 3 non-Finnish European samples. Here, we only used the overlapping variants between the real traits and the permuted GWASs to avoid biases introduced by genotyping platforms or imputation. We then calculated the GWAS gene z-scores for each of the 10,000 simulated GWAS signals, as described above. Next, we calculated the Pearson correlations between the GWAS gene z-scores. As simulated GWAS signals are random and independent of each other, any remaining correlation between GWAS gene z-scores reflects the underlying LD patterns and chromosomal organization of genes.

We simulated an additional 10,000 GWASs as described above to empirically determine enrichment p-values. Finally, we used an additional 100 simulations to estimate the false discovery rate (FDR) of Downstreamer associations.

Downstreamer step 1.3. Correction for additional variables and mean gene p-value calculation

To facilitate the correction of additional parameters, variables can be provided which are used to correct the GWAS gene p-values before fitting the GLS (step 2.2). These are fit using a (multivariate) OLS model of which the residuals are taken and used as input for the subsequent steps. We used this option to additionally correct for gene length as well as the mean gene p-value over the 44 traits. The mean gene p-value was calculated by first calculating the mean of the traits in each of the 10 classes of GWAS traits for each gene. Then, for each gene, the mean over these 10 means was calculated in order to avoid having the overrepresented classes (blood cell composition) overshadow the calculation of the means.

Downstreamer step 2.1. Pre-processing GWAS gene z-scores and pruning highly correlated genes

For each GWAS, both real and simulated, we force-normalized the GWAS z-scores into a normal distribution to ensure that outliers will not have disproportionate weights. Due to limitations in the PASCAL methodology that result in ties at a minimum significance level of 1×10^{-12} for highly significant genes, we use the minimum SNP p-value from the GWAS to identify the most significant gene and resolve the tie. We then use the linear model (step 1.3) to correct for gene length, as longer genes will typically harbour more SNPs.

Sometimes, two (or more) genes will be so close to one another that their GWAS gene z-scores are highly correlated, violating the assumptions of the linear model. Thus, genes with a Pearson correlation $r \geq 0.8$ in the 10,000 GWAS permutations were collapsed into 'meta-genes' and treated as one gene. Meta-gene z-scores were averaged across the input z-scores. Lastly, the GWAS z-scores of the meta genes were scaled (mean = 0, standard deviation = 1).

Downstreamer step 2.2. GLS to calculate key gene scores and pathway enrichments

We used a GLS regression to associate the GWAS gene z-scores with the gene co-regulation z-scores or with the expanded pathway annotations. These two analyses result in the key gene prioritizations and pathway enrichments, respectively. We used the gene–gene correlation matrix derived from the 10,000 permutations as a measure of the conditional covariance of the error term (Ω) in the GLS to account for the relationships between genes due to LD and proximity. The pseudo-inverse of Ω is used as a substitute for Ω^{-1}

The formula of the GLS is as follows:

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

Where β is the estimated effect size of the pathway, term or gene from the co-regulation matrix; Ω is the gene-gene correlation matrix; X is the design matrix of real GWAS z-scores and y is the vector of gene z-scores per pathway, term or gene from the co-regulation matrix. As we standardized the predictors, we did not include an intercept in the design matrix and X only contains one column with the real GWAS gene z-scores. We estimated the betas for the 10,000 random GWASs in the same way and subsequently used them to estimate the empirical p-value for β.

Downstreamer step 2.3. Pathway and gene set gene enrichments

To identify pathway and disease enrichments, we used the following databases: HPO, KEGG, Reactome, MGI and GO Biological Process, Cellular Component and Molecular Function. We have previously predicted how much each gene contributes to these gene sets, resulting in a z-score per pathway or term per gene [24]. We then used Bonferroni correction to determine if the gene should be added to the extended pathway membership.

Next, we collapsed genes into meta-genes, in parallel with the GWAS step, to ensure compatibility with the GWAS gene z-scores, following the same procedure as in the GWAS pre-processing. The pathway memberships for a meta-gene were calculated as the sum of the membership divided by the square root of the number of genes. So, a meta-gene containing five genes, of which two are in a pathway, would get a value of $2 / \sqrt{5} = 0.89$ for that pathway. Finally, the pathway memberships of the meta-genes were scaled and centred (mean = 0, standard deviation = 1).

Downstreamer step 2.4: Co-regulation matrix

To calculate key gene scores, we used a previously generated co-regulation matrix based on a large multi-tissue gene network [24]. In short, publicly available RNA-seq samples were downloaded from the European Nucleotide Archive (https://www.ebi.ac.uk/ena). After quality control, 56,435 genes and 31,499 samples covering a wide range of human cell-types and tissues remained. We performed a principal component analysis on this dataset and selected the 165 principal components representing 50% of the variation that offered the best prediction of gene function [76]. We then selected the protein-coding genes and centred and scaled the eigenvectors for these 165 components (mean = 0, standard deviation = 1) such that each component was given equal weight. The first components mostly describe tissue differences [24], so this normalization ensures that tissue-specific patterns do not disproportionately drive the co-regulation matrix. The co-regulation matrix is defined as the Pearson correlation between the genes from the scaled eigenvector matrix. The diagonal of the co-regulation matrix was set to zero to avoid correlation with itself having a disproportionate effect on the association to the GWAS gene z-scores. Finally, we converted the Pearson r to z-scores. To associate the co-regulation to the gene z-scores, the same meta-gene procedure was applied as outlined for the pathway enrichments.

Enrichment of key genes

Enrichments of key genes among HPO/MGI/GO terms and KEGG gene sets was done by Fisher's exact test, taking all key genes at Bonferroni or FDR significance and comparing their overlap to all other genes. AUCs were calculated by dividing the Mann-Whitney U statistic of the key gene z-scores and gene set membership by the product of sample sizes. The gene-pathway/term definitions we used were those provided by the respective databases, thus they were not the extended versions used for the GWAS gene set enrichments. This is implemented in Downstreamer using –T PRIO_GENE_ENRICH.

Enrichment of average gene z-scores and association with LD and gene density

Enrichments of the top 500 average gene z-scores were done by first correcting the mean gene z-score vector (see step 1.3 for details on calculating this) for the extent of the LD around a gene as well as the gene density. To quantify the extent of the LD block, we took the mean of the LD scores of all SNPs in a 25kb window around the gene. Pre-computed European LD scores were downloaded from https://github.com/bulik/ldsc. Gene density was calculated by counting the number of genes in a 500kb window around the start end of the gene. Both these factors were then fit in a linear model with the mean gene z-score as the outcome. The residuals were taken and ranked to arrive at the top 500 genes. We then carried out overrepresentation analysis using https://toppgene.cchmc.org/enrichment.jsp with the default background set.

Association with LoF and MiS intolerance

MiS and LoF intolerance z-scores were downloaded from the gnomAD consortium (https://gnomad.broadinstitute.org/downloads > pLoF Metrics by Gene TSV v2.1.1). As an overall measure of the "keyness" of a gene, we calculated the maximum key gene z-score observed over the 44 traits for each gene. We then associated this to the MiS and LoF z-scores from the gnomAD consortium by Pearson correlation.

Enrichment of cis-eQTL and key genes

Enrichments of *cis*-eQTL and key genes were calculated by fisher exact test, taking all the genes tested in eQTLgen or MetaBrian respectively as the background set. A gene was considered to be a *cis*-eQTL gene if it had a significant association in eQTLgen or MetaBrain analyses respectively.

Overlap with trans-eQTLs and eQTS genes

To investigate their overlap with the key genes identified by Downstreamer, we downloaded the *trans*-eQTL and eQTS results from the eQTLGen Consortium (www.eqtlgen.org). For each GWAS, we selected all *trans*-eQTLs that emanate from independent top SNPs (1000 Genomes phase 3 EUR, R^2 0.2, 500kb window) and calculated the sum of *trans*-eQTL squared z-scores for each gene. We then log-transformed this and associated it to the key gene z-score for the GWAS using Pearson correlation.

For the overlap with eQTS genes, we selected eQTSs for which we had overlapping GWAS traits. We then evaluated if the eQTS genes had a higher key gene z-score compared to all other genes using a Student's t-test.

Code and data availability

Software and scripts are available for download at: https://github.com/molgenis/systemsgenetics/tree/master/Downstreamer

A manual for Downstreamer is available at: https://github.com/molgenis/systemsgenetics/wiki/Downstreamer

All RNA-seq data used in the main analysis are publicly available in the European Nucleotide Archive, for details please see [24].

Acknowledgements

We thank Kate McIntyre for the editorial assistance.

We thank the UMCG Genomics Coordination Center, the UG Center for Information Technology and their sponsors BBMRI-NL & TarGet for storage and compute infrastructure.

Funding

O.B.B. is supported by an NWO VIDI grant awarded to I.H.J (no. 016.171.047).

I.H.J. is supported by a Rosalind Franklin Fellowship from the University of Groningen and an NWO VIDI grant (no. 016.171.047).

L.F. is supported by grants from the Dutch Research Council (ZonMW-VIDI 917.14.374 and ZonMW-VICI to L.F.) and by an ERC Starting Grant, grant agreement 637640 (ImmRisk) and through a Senior Investigator Grant from the Oncode Institute.

P.D. is supported by an NWO, ZonMW-VENI grant (no. 9150161910057).

Competing interests

The authors declare no competing interest

Materials & Correspondence

Lude Franke; l.h.franke@umcg.nl, Patrick Deelen; p.deelen@umcg.nl

Author contributions

Conceptualization: O.B.B., A.C., L.F., P.D.
Data curation: O.B.B., A.C., P.D.
Formal Analysis: O.B.B., A.C., P.D.
Funding acquisition: L.F., P.D.
Investigation: A.C., O.B.B., H.J.W., H.W., F.B., U.V., S.M.S.
Methodology: O.B.B., L.F., P.D.
Software: O.B.B., H.W., P.D.
Supervision: O.B.B., I.H.J., L.F.
Visualization: O.B.B., A.C., L.F., P.D.
Writing – original draft: O.B.B., A.C., L.F., P.D.
Writing – review & editing: I.H.J, H.J.W., U.V.,
Roles as defined by: CRediT (Contributor Roles Taxonomy)

Supplementary material

Supplemenatary material are provided at: https://doi.org/10.1101/2021.10.21.21265342

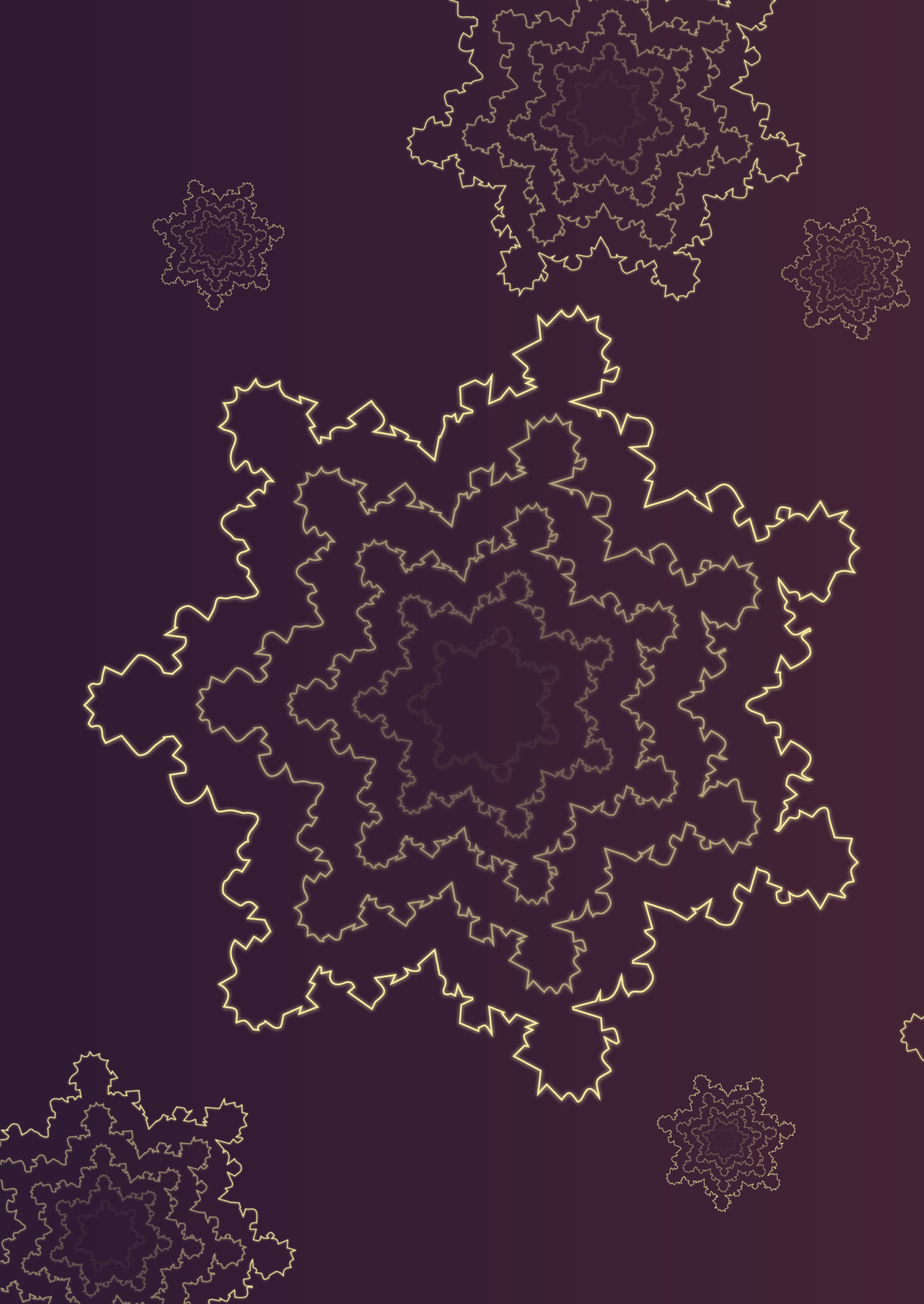
Suppl. Fig. 1. Association between gene z-score profiles of different traits
Suppl. Fig. 2. Association between average gene z-score and LD and gene density
Suppl. Fig. 3. Enrichment of missense and loss of function intolerance in the average gene z-score
Suppl. Fig. 4. Relationship between number of samples used for a GWAS and the power to detect key genes
Suppl. Fig. 5. Sharing among the key genes
Suppl. Fig. 6. Comparison of pathway enrichment results before and after extending pathway definitions using a gene regulatory network
Suppl. Fig. 7. Sample enrichment per plots per traits
Suppl. Fig. 8. Correlations between tissue expression and key gene z-scores
Suppl. Fig. 9. Relation between glomerular filtration rate key gene scores and expression levels
Suppl. Fig. 10. Relation between inflammatory bowel disease key gene scores and expression levels
Suppl. Fig. 11. Association between LoF and MiS scores and key gene z-scores
Suppl. Fig. 12. Association between LoF and MiS scores and GWAS gene p-value
Suppl. Fig. 13. Association between degree and key gene z-scores
Suppl. Fig. 14. Association between LoF and MiS scores and key gene z-scores after correction for the degree
Suppl. Fig. 15.Sharing between GWASs among pathway enrichments for GO biolo gical process
Suppl. Tab. 1. List of the 44 complex traits and diseases to which we applied Downstreamer
Suppl. Tab. 2. Enrichment of top500 genes from average GWAS signal
Suppl. Tab. 3. Mapping of GWAS traits to HPO-terms
Suppl. Tab. 4. Enrichment of GWAS genes and key genes per trait
Data S1. Key gene prediction and pathway enrichments of the 44 tested traits and diseases
Note S1. Pathway enrichments using co-regulation networks
Note S2. Tissue- and cell type–enrichments

References

1. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. Hum. Mol. Genet. 24, R111-119 (2015).
2. Visscher, P. M. et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. 101, 5–22 (2017).
3. Freund, M. K. et al. Phenotype-Specific Enrichment of Mendelian Disorder Genes near GWAS Regions across 62 Complex Traits. Am. J. Hum. Genet. 103, 535–552 (2018).
4. Genetics of Infectious and Inflammatory Diseases: Overlapping Discoveries from Association and Exome-Sequencing Studies - PubMed. https://pubmed.ncbi.nlm.nih.gov/27912315/.
5. Holm, H. et al. Several common variants modulate heart rate, PR interval and QRS duration. Nat. Genet. 42, 117–122 (2010).
6. Carniel, E. et al. Alpha-myosin heavy chain: a sarcomeric gene associated with dilated and hypertrophic phenotypes of cardiomyopathy. Circulation 112, 54–59 (2005).
7. Võsa, U. et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. 1–11 (2021) doi:10.1038/s41588-021-00913-z.
8. Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. Nat Commun 6, (2015).
9. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. PLOS Comput. Biol. 12, e1004714 (2016).
10. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLOS Comput. Biol. 11, e1004219 (2015).
11. Meuleman, W. et al. Index and biological spectrum of human DNase I hypersensitive sites. Nature 584, 244–251 (2020).
12. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. 47, 1228–1235 (2015).
13. Gerring, Z. F., Mina-Vargas, A., Gamazon, E. R. & Derks, E. M. E-MAGMA: an eQTL-informed method to identify risk genes using genome-wide association study summary statistics. Bioinforma. Oxf. Engl. btab115 (2021) doi:10.1093/bioinformatics/btab115.
14. Gerring, Z. F., Mina-Vargas, A. & Derks, E. M. eMAGMA: An eQTL-informed method to identify risk genes using genome-wide association study summary statistics. 854315 https://www.biorxiv.org/content/10.1101/854315v1 (2019) doi:10.1101/854315.
15. Sobczyk, M. K., Gaunt, T. R. & Paternoster, L. MendelVar: gene prioritization at GWAS loci using phenotypic enrichment of Mendelian disease genes. Bioinformatics 37, 1–8 (2021).
16. Weeks, E. M. et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. 2020.09.08.20190561 https://www.medrxiv.org/content/10.1101/2020.09.08.20190561v1 (2020) doi:10.1101/2020.09.08.20190561.
17. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169, 1177–1186 (2017).
18. Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell 177, 1022-1034.e6 (2019).
19. Vuckovic, D. et al. The Polygenic and Monogenic Basis of Blood Traits and Diseases. Cell 182, 1214-1231.e11 (2020).
20. Sinnott-Armstrong, N., Naqvi, S., Rivas, M. & Pritchard, J. K. GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. eLife 10, e58615 (2021).
21. Freimer, J. W. et al. Systematic discovery and perturbation of regulatory genes in human T cells reveals the architecture of immune networks. bioRxiv 2021.04.18.440363 (2021) doi:10.1101/2021.04.18.440363.
22. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. 47, 291–295 (2015).
23. Jordan, D. M., Verbanck, M. & Do, R. HOPS: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. Genome Biol. 20, 222 (2019).
24. Deelen, P. et al. Improving the diagnostic yield of exome- sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. Nat. Commun. 10, 2837 (2019).
25. Cohn, A., Sofia, M. A. & Kupfer, S. S. Type 1 Diabetes and Coeliac Disease: Clinical Overlap and New Insights into Disease Pathogenesis. Curr. Diab. Rep. 14, 517 (2014).
26. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. Nat. Genet. 47, 1236–1241 (2015).
27. Balk, S. P., Ko, Y.-J. & Bubley, G. J. Biology of Prostate-Specific Antigen. J. Clin. Oncol. 21, 383–391 (2003).
28. Williams, S. A., Xu, Y., De Marzo, A. M., Isaacs, J. T. & Denmeade, S. R. Prostate-Specific Antigen (PSA) Is Activated by KLK2 in Prostate Cancer Ex Vivo Models and in Prostate-Targeted PSA/KLK2 Double Transgenic Mice. The Prostate 70, 788–796 (2010).
29. Zhang, W. et al. Transmembrane Channel-Like 5 (TMC5) promotes prostate cancer cell proliferation through cell cycle regulation. Biochimie 165, 115–122 (2019).
30. Bu, H. et al. Putative Prostate Cancer Risk SNP in an Androgen Receptor-Binding Site of the Melanophilin Gene Illustrates Enrichment of Risk SNPs in Androgen Receptor Target Sites. Hum. Mutat. 37, 52–64 (2016).
31. Roca, H. et al. Transcription factors OVOL1 and OVOL2 induce the mesenchymal to epithelial transition in human cancer. PloS One 8, e76773 (2013).

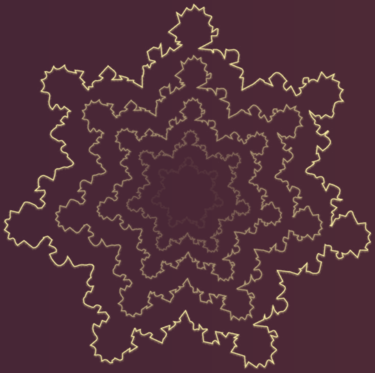
32. Ikonen, T. et al. Association of E-cadherin germ-line alterations with prostate cancer. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 7, 3465–3471 (2001).
33. Wang, L. et al. Downregulation of POTEG predicts poor prognosis in esophageal squamous cell carcinoma patients. Mol. Carcinog. 57, 886–895 (2018).
34. Chao, A. et al. BAI1-Associated Protein 2-Like 1 (BAIAP2L1) Is a Potential Biomarker in Ovarian Cancer. PLoS ONE 10, e0133081 (2015).
35. Kuriyama, S., Tsuji, T., Sakuma, T., Yamamoto, T. & Tanaka, M. PLEKHN1 promotes apoptosis by enhancing Bax-Bak hetro-oligomerization through interaction with Bid in human colon cancer. Cell Death Discov. 4, 11 (2018).
36. Satirapoj, B. et al. High levels of uric acid correlate with decline of glomerular filtration rate in chronic kidney disease. J. Med. Assoc. Thail. Chotmaihet Thangphaet 93 Suppl 6, S65-70 (2010).
37. Liston, H. L., Markowitz, J. S. & DeVane, C. L. Drug glucuronidation in clinical psychopharmacology. J. Clin. Psychopharmacol. 21, 500–515 (2001).
38. Guillemette, C., Lévesque, É. & Rouleau, M. Pharmacogenomics of Human Uridine Diphospho-Glucuronosyltransferases and Clinical Implications. Clin. Pharmacol. Ther. 96, 324–339 (2014).
39. Margaillan, G. et al. Quantitative Profiling of Human Renal UDP-glucuronosyltransferases and Glucuronidation Activity: A Comparison of Normal and Tumoral Kidney Tissues. Drug Metab. Dispos. 43, 611–619 (2015).
40. OMIM - Online Mendelian Inheritance in Man. https://omim.org/.
41. Lee, H. et al. Low-dose interleukin-2 alleviates dextran sodium sulfate-induced colitis in mice by recovering intestinal integrity and inhibiting AKT-dependent pathways. Theranostics 10, 5048–5063 (2020).
42. Joosse, M. E. et al. Duplication of the IL2RA locus causes excessive IL-2 signaling and may predispose to very early onset colitis. Mucosal Immunol. 1–11 (2021) doi:10.1038/s41385-021-00423-5.
43. McLean, L. P., Shea-Donohue, T. & Cross, R. K. Vedolizumab for the treatment of ulcerative colitis and Crohn's disease. Immunotherapy 4, 883–898 (2012).
44. Lefevre, P. L. C. & Vande Casteele, N. Clinical Pharmacology of Janus Kinase Inhibitors in Inflammatory Bowel Disease. J. Crohns Colitis 14, S725–S736 (2020).
45. Sugahara, K. et al. Amiselimod, a novel sphingosine 1-phosphate receptor-1 modulator, has potent therapeutic efficacy for autoimmune diseases, with low bradycardia risk. Br. J. Pharmacol. 174, 15–27 (2017).
46. Caballero-Franco, C. & Kissler, S. The autoimmunity-associated gene RGS1 affects the frequency of T follicular helper cells. Genes Immun. 17, 228–238 (2016).
47. Gibbons, D. L. et al. Cutting Edge: Regulator of G Protein Signaling-1 Selectively Regulates Gut T Cell Trafficking and Colitic Potential. J. Immunol. 187, 2067–2071 (2011).
48. Salaga, M., Storr, M., Martemyanov, K. A. & Fichna, J. RGS proteins as targets in the treatment of intestinal inflammation and visceral pain - new insights and future perspectives. BioEssays News Rev. Mol. Cell. Dev. Biol. 38, 344–354 (2016).
49. Hunt, K. A. et al. Newly identified genetic risk variants for coeliac disease related to the immune response. Nat. Genet. 40, 395–402 (2008).
50. International Multiple Sclerosis Genetics Conssortium (IMSGC). IL12A, MPHOSPH9/CDK2AP1 and RGS1 are novel multiple sclerosis susceptibility loci. Genes Immun. 11, 397–405 (2010).
51. Swainston Harrison, T. & Perry, C. M. Aripiprazole: a review of its use in schizophrenia and schizoaffective disorder. Drugs 64, 1715–1736 (2004).
52. Mortimer, A. M. Update on the management of symptoms in schizophrenia: focus on amisulpride. Neuropsychiatr. Dis. Treat. 5, 267–277 (2009).
53. Hahn, M. K., Cohn, T., Teo, C. & Remington, G. Topiramate in schizophrenia: a review of effects on psychopathology and metabolic parameters. Clin. Schizophr. Relat. Psychoses 6, 186–196 (2013).
54. Wang, X. & Goldstein, D. B. Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. Am. J. Hum. Genet. 106, 215–233 (2020).
55. Klein, N. de et al. Brain expression quantitative trait locus and network analysis reveals downstream effects and putative drivers for brain-related diseases. 2021.03.01.433439 https://www.biorxiv.org/content/10.1101/2021.03.01. 433439v2 (2021) doi:10.1101/2021.03.01.433439.
56. Sun, S.-C. Non-canonical NF-κB signaling pathway. Cell Res. 21, 71–85 (2011).
57. Sharfe, N. et al. The effects of RelB deficiency on lymphocyte development and function. J. Autoimmun. 65, 90–100 (2015).
58. Matthews, S. P., McMillan, S. J., Colbert, J. D., Lawrence, R. A. & Watts, C. Cystatin F Ensures Eosinophil Survival by Regulating Granule Biogenesis. Immunity 44, 795–806 (2016).
59. Lambrecht, B. N. & Hammad, H. The immunology of asthma. Nat. Immunol. 16, 45–56 (2015).
60. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020).
61. Miller, E. & Grosel, J. M. A review of Ehlers-Danlos syndrome. J. Am. Acad. PAs 33, 23–28 (2020).
62. Weedon, M. N. et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nat. Genet. 40, 575–583 (2008).
63. 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).
64. Molberg, O. et al. Gliadin specific, HLA DQ2-restricted T cells are commonly found in small intestinal biopsies from coeliac disease patients, but not from controls. Scand. J. Immunol. 46, 103–109 (1997).
65. van der Wijst, M. G. P. et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat. Genet. 50, 493–497 (2018).
66. Fairfax, B. P. et al. Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. Science 343, 1246949–1246949 (2014).
67. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. Nature 593, 238–243 (2021).

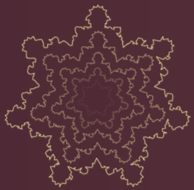
68. Fisher, R. A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Trans. R. Soc. Edinb. 52, 399–433 (1919).
69. Retterer, K. et al. Clinical application of whole-exome sequencing across clinical indications. Genet. Med. Off. J. Am. Coll. Med. Genet. 18, 696–704 (2016).
70. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 47, D1005–D1012 (2019).
71. Gibbs, R. A. et al. A global reference for human genetic variation. Nature 526, 68–74 (2015).
72. Jassal, B. et al. The reactome pathway knowledgebase. Nucleic Acids Res. 48, D498–D503 (2020).
73. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30 (2000).
74. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29 (2000).
75. Köhler, S. et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. 47, D1018–D1027 (2019).
76. KidneyNetwork: Using kidney-derived gene expression data to predict and prioritize novel genes involved in kidney disease | medRxiv. https://www.medrxiv.org/content/10.1101/2021.03.10. 21253054v1.
77. Bult, C. J. et al. Mouse Genome Database (MGD) 2019. Nucleic Acids Res. 47, D801–D806 (2019).
78. Bakker, O. B. et al. Potential impact of coeliac disease genetic risk factors on T cell receptor signaling in gluten-specific CD4+ T cells. Sci. Rep. 11, 9252 (2021).
79. van der Graaf, A. et al. Systematic Prioritization of Candidate Genes in Disease Loci Identifies TRAFD1 as a Master Regulator of IFNγ Signaling in Coeliac Disease. Front. Genet. 11, 1780 (2021).
80. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet. 50, 621–629 (2018).

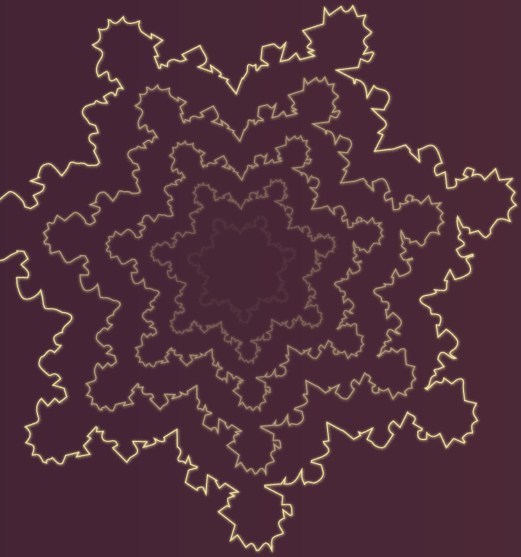


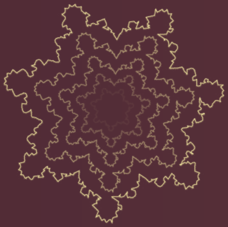
Part II

Genetics and human immune variation









Chapter 4

Deconvolution of bulk blood eQTL effects into immune cell subpopulations

Raúl Aguirre-Gamboa [1], Niek de Klein [2*], Jennifer di Tommaso [1*], Annique Claringbould [2], Monique GP van der Wijst [2], Dylan de Vries [2], Harm Brugge [2], Roy Oelen [2], Urmo Võsa [1,3], Maria M. Zorro [1], Xiaojin Chu [1,4], Olivier B. Bakker [1], Zuzanna Borek [1], Isis Ricaño-Ponce [1], Patrick Deelen [2,5], Cheng-Jiang Xu [4,7], Morris Swertz [1,5], Iris Jonkers [1], Sebo Withoff [1], Irma Joosten [6], Serena Sanna [1], Vinod Kumar [1,7], Hans J. P. M. Koenen [6], Leo A. B. Joosten [7], Mihai G. Netea [7,8], Cisca Wijmenga [1], BIOS Consortium, Lude Franke [1] and Yang Li [1,4,7]*

BMC Bioinformatics (2020)

1 Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands.
2 Department of Genetics, Oncode Institute, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands.
3 Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia.
4 Centre for Individualised Infection Medicine (CiiM) & TWINCORE, joint ventures between the Helmholtz-Centre for Infection Research (HZI) and the Hannover Medical School (MHH), Feodor-Lynen-Str. 7, 30625 Hannover, Germany.
5 University of Groningen and University Medical Center Groningen, Genomics Coordination Center, Groningen, the Netherlands.
6 Department of Laboratory Medicine, Laboratory for aMedical Immunology, Radboud University Medical Centre, Nijmegen, the Netherlands.
7 Department of Internal Medicine and Radboud Center for Infectious Diseases, Radboud University Medical Center, Nijmegen, the Netherlands.
8 Department of Genomics & Immunoregulation, Life and Medical Sciences Institute (LIMES), University of Bonn, Bonn, Germany.
** These authors contributed equally*

Abstract

Background: Expression quantitative trait loci (eQTL) studies are used to interpret the function of disease-associated genetic risk factors. To date, most eQTL analyses have been conducted in bulk tissues, such as whole blood and tissue biopsies, which are likely to mask the cell type-context of the eQTL regulatory effects. Although this context can be investigated by generating transcriptional profiles from purified cell subpopulations, current methods to do this are labor-intensive and expensive. We introduce a new method, Decon2, as a framework for estimating cell proportions using expression profiles from bulk blood samples (Decon-cell) followed by deconvolution of cell type eQTLs (Decon-eQTL).

Results: The estimated cell proportions from Decon-cell agree with experimental measurements across cohorts (R ≥ 0.77). Using Decon-cell, we could predict the proportions of 34 circulating cell types for 3194 samples from a population-based cohort. Next, we identified 16,362 whole-blood eQTLs and deconvoluted cell type interaction (CTi) eQTLs using the predicted cell proportions from Decon-cell. CTi eQTLs show excellent allelic directional concordance with eQTL (≥ 96–100%) and chromatin mark QTL (≥87–92%) studies that used either purified cell subpopulations or single-cell RNA-seq, outperforming the conventional interaction effect.

Conclusions: Decon2 provides a method to detect cell type interaction effects from bulk blood eQTLs that is useful for pinpointing the most relevant cell type for a given complex disease. Decon2 is available as an R package and Java application (https://github.com/molgenis/systemsgenetics/tree/master/Decon2) and as a web tool (www.molgenis.org/deconvolution).

Keywords

eQTL, Deconvolution, Cell types, Immune cells

Background

For many of the genetic risk factors that have been associated to immune diseases by genome-wide association studies (GWAS), the molecular mechanism leading to disease remains unknown [1]. Most of these genetic risk variants are located in the non-coding regions of the genome, implying that they play a role in gene regulation [2, 3]. Expression quantitative trait locus (eQTL) analysis provides a way to characterize the regulatory effect of these risk factors in humans, and many eQTL studies have now been carried out using bulk tissues, for example, whole blood [4, 5]. However, bulk tissues comprise many different cell types, and gene regulation is known to vary across cell types [6-8]. In recent years, efforts to describe eQTL effects in purified cell subpopulations have been carried out in specific cell types [9]. Unfortunately, the length and cost of the study protocols have limited these studies to small sample sizes and only a few cell types. Current developments on single cell (sc) RNASeq technologies have given rise to sc-eQTLs, an approach that, although promising, is still bound to a limited number of individuals, which thereby limits the number of detectable cell type interaction (CTi) eQTLs. Nevertheless, the ability to pinpoint the CT in which a risk factor exerts an eQTL effect could help us to understand its role in disease.

Statistical approaches to detect CT effects using tissue expression profiles have mainly been developed to evaluate gene by environment interaction (GxE) terms, for example to detect CT eQTLs for myeloid and lymphoid lineages using only whole blood gene expression and by evaluating the interaction between genotype and cell proportions for neutrophils and lymphocytes in whole blood [10]. A second study linked eQTL genes to proxy genes through correlation; these proxy genes were then associated with intrinsic or extrinsic factors such as cell proportions or inflammation markers [11]. However, these efforts focused on exploiting only one GxE term, or on indirectly linking the CT proportions to given eQTL, rather than directly ascertaining the interaction between all the main cell proportions comprising the bulk tissue and genotype. Unfortunately, quantifying cell proportions, in particular rare subpopulations (total abundance ≤3% in circulating white blood cells), is expensive and time-consuming. Hence, quantifying immune cell proportions in large functional genomics cohorts is not common practice.

Here we present and validate Decon2, a computational and statistical framework that can (1) predict the proportions of known circulating immune cell subpopulations (Decon-cell), and (2) combine these predicted proportions with whole blood gene expression and genotype information to assign bulk eQTL effects into CTi eQTLs (Decon-eQTL). Our two-step framework provides an improvement over previously published methods. Unlike earlier methods [12], Decon-cell does not rely on any prior information about transcriptome profiles from purified cell subpopulations. It only requires quantification of the cell proportions comprising the bulk tissue, in this case whole blood. Decon-cell identifies signature genes that correlate with cell proportions in a bulk tissue. Secondly, Decon-eQTL is the first approach in which all major cell proportions (the major cell types for which the sum of proportions per sample is approximately 100%) of bulk blood tissue are incorporated into an eQTL model simultaneously. Decon-eQTL can then be used to systematically test for any significant interaction between each CT and genotype, while also controlling for the effect on expression of the other cell types.

We generated the Decon-cell predictive models using data from the 500FG cohort [13], where quantification of immune cell types was carried out using FACS [14] and RNA-Seq-based bulk whole blood transcriptome profiles were available for 89 samples [15]. By using a cross-validation approach, we were able to accurately predict 34 out of 73 cell subtypes using only whole blood gene expression. For validation, we applied Decon-cell to three independent cohorts (Lifelines Deep [16], n = 627; Leiden Longevity cohort [17], n = 660 and the Rotterdam Study [18], n = 773) for which both blood RNA-seq and measured cell proportion data are available (neutrophils, lymphocytes and CD14+ monocytes and granulocytes). Additionally, we benchmarked Decon-cell prediction performance against two other existing methods that quantify immune cell composition using gene expression profiles from whole blood on these three independent cohorts. After showing that we can accurately predict circulating immune cell proportions, we applied Decon-cell to estimate cell proportions in 3194 individuals from the BIOS cohort [16, 19-21] for whom both whole blood RNA-seq and genotypes were available. The BIOS cohort is a valuable resource for functional genomics studies where extensive characterization of the genetic component on gene expression [11] and epigenetics [22] have been performed. We integrated whole blood expression and genotype information and predicted cell proportion with Decon-eQTL to deconvolute [16], 362 significant whole blood *cis*-eQTLs top effects into CT interacting eQTLs (CTi eQTLs). These deconvoluted CTi eQTL results were comprehensively validated using transcriptome profiles from purified cell subpopulations [23], eQTLs and chromatin mark QTLs from purified cell types [9] and eQTLs from single-cell experiments [24]. We also systematically compared the performance of Decon-eQTL against the most used method [10] that detect cell type eQTL effects using whole blood expression profiles.

Results

Decon-cell accurately predicts the proportions of known immune cell types

In order to assign the cell types from which an overall eQTL effect from a bulk tissue sample (e.g. whole blood) arise, we need three types of information: genotype data, tissue expression data and cell type proportions (**Fig. 1**). Here we propose a computational method that predicts the cell proportions of known immune cell types using gene signatures in whole blood expression data using a machine-learning approach. Decon-cell employs the regularized regression method elastic net [26] to define sets of signature genes for each cell type. In other words, these signatures were selected as having the best prediction power for individual cell proportions.

There are 89 samples in the 500FG cohort with both whole blood RNA-seq and quantification of 73 immune cell subpopulations by FACS. This data was used to build the prediction models for estimating cell subpopulations by Decon-cell. First, we determined which of the 73 cell subpopulations could be reliably predicted by Decon-cell. A within-cohort cross-validation strategy was employed by randomly dividing 89 samples (**Fig. 1**) into training and test sets (70 and 30% of the samples, respectively). After generating a model using each training set, we applied the prediction models of each cell type to the samples in the test sets. We compared the predicted and measured cell proportion for each cell type using Spearman correlation coefficients to evaluate prediction performance. We repeated this process 100 times and then used the mean of the correlation coefficient in all 100 iterations to evaluate the prediction performance.

We were able to predict 34 out of 73 cell subpopulations using whole blood gene expression data at a threshold of mean $R \geq 0.5$ across all 100 iterations (**Fig. 2A**, **Suppl. Fig. 1**, **Suppl. Tab. 1**). The number of signature genes selected in our models for predicting cell proportions varied across the cell types, ranging from 2 to 217 signature genes (**Suppl. Fig. 2A**, **Suppl. Tab. 1**), and they were independent of the average abundance of these cell types in whole blood ($R = 0.02$, Spearman correlation coefficient, **Suppl. Fig. 2A**). In particular, cell types that are abundant in whole blood (granulocytes-neutrophils, CD4+ T cells and CD14+ monocytes) were predicted with high confidence (correlation between predicted and measured values, $R \geq 0.73$).

Remarkably, we were also able to predict a number of less abundant cell subpopulations, including NK cells, CD8+ T cells, non-NK T cells (CD3CD56-), CD4+ central memory, CD4+ effector memory T cells and regulatory T cells (**Suppl. Fig. 2A**), as determined by FACS. Cell types with a low prediction performance ($R < 0.5$) are those that have few signature genes with expression levels that correlate sufficiently (i.e. absolute $R < 0.3$) with the measured cell proportions in whole blood (**Suppl. Fig. 2B-C**). For each of the 34 predictable cell types, we used Decon-cell to build models for predicting their cell counts using all 89 samples from the 500FG cohort. These models were applied to 3194 samples in an independent cohort (BIOS cohort) to predict cell proportions of circulating immune cell types for the subsequent deconvolution of eQTL effects.

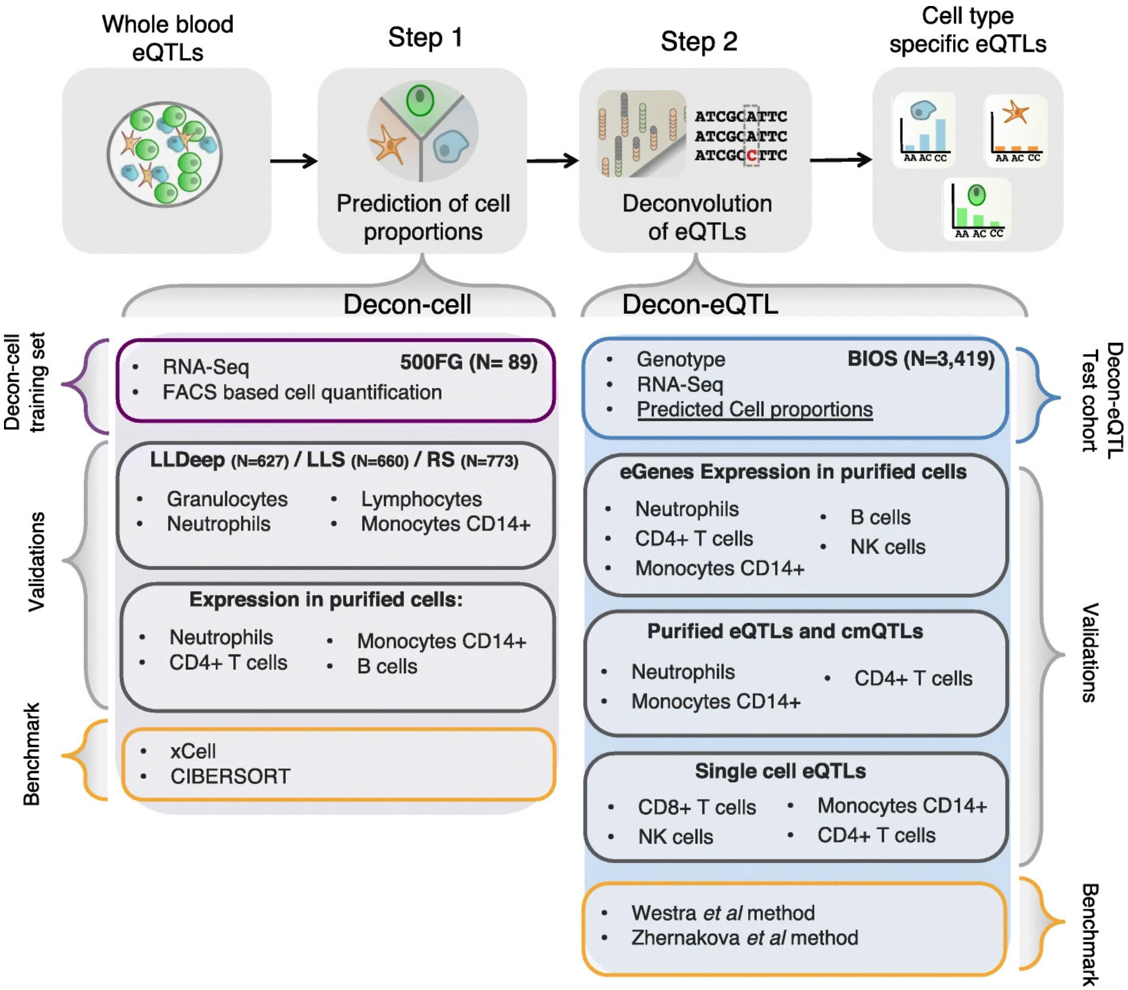


Fig. 1 Workflow of application of Decon2 to predict cell counts followed by deconvolution of whole blood eQTLs. Using whole blood expression and FACS data of 500FG samples, Decon-cell predicts cell proportions with selected marker genes of circulating immune cell subpopulations. Validations of Decon-cell were carried out on three independent cohorts for which measurements of neutrophils/granulocytes, lymphocytes and monocytes CD14+ were available along with expression profiles of whole blood. Benchmarking of Decon-cell was performed against CIBERSORT [25] and xCell [12]. Decon-cell was applied to an independent cohort (BIOS) to predict cell counts using whole blood RNA-seq. Decon-eQTL subsequently integrates genotype and tissue expression data together with predicted cell proportions for samples in BIOS to detect cell type eQTLs. We validated Decon-eQTL using multiple independent sources, including expression profiles of purified cell subpopulations, eQTLs and chromatin mark QTLs (cmQTLs) from purified neutrophils, monocytes CD14+ and CD4+ T cells [9], and single-cell eQTL results [24]. Benchmarking of Decon-eQTL was carried out for comparison with a previously reported methods that detected cell type–eQTL effects using whole blood expression data, i.e. the Westra et al. [10]

In addition to within-cohort validation, we tested our cell proportion models using three independent cohorts (LLDeep, n = 627; LLS, n = 660; RS, n = 773) in which cell type abundances were quantified using a Coulter counter for neutrophils (granulocytes for RS), lymphocytes and CD14+ monocytes (**Fig. 2B**, **Suppl. Fig. 3A-B**). In LLDeep, we were able to accurately predict these three cell types with Spearman correlation coefficients of R = 0.73, R = 0.89 and R = 0.73, respectively. For LLS and RS, the prediction performance was similarly accurate for neutrophils and lymphocytes (R = 0.76 for neutrophils, R = 0.84 for lymphocytes), but less so for monocytes (R = 0.50 for CD14+ monocytes and proportions in LLS and R = 0.74 for granulocytes, R = 0.83 for lymphocytes and R = 0.28 for CD14+ monocytes in RS).

Next, in order to benchmark Decon-cell, we compared its prediction performance against two other existing tools that quantify the abundance of known immune cell types using bulk whole blood expression profiles: CIBERSORT [25] and xCell [12]. We obtained the predicted proportions by CIBERSORT and enrichment scores of circulating immune cells by xCell for the samples in three different cohorts: LLDeep, LLS and RS (**Suppl. Fig. 4A-B**). For each cell type, Decon-cell outperforms CIBERSORT and xCell (**Suppl. Fig. 3B**). The scatterplots of predicted vs measured values (**Suppl. Fig. 3A**, **Suppl. Fig. 4A-B**) further demonstrate that the better performance of Decon-cell is not due to cell proportion outliers.

Finally, we evaluated whether the signature genes showed CT expression in their relevant purified cell types using BLUEPRINT [23] RNA-seq data from the purified cell subpopulations. Here we focused on cell types with more than three samples measured, which included neutrophils, CD14+ monocytes, CD4+ T cells and B cells. The signature genes showed overall higher expression in their relevant cell subpopulations compared to other cell subpopulations. Interestingly, the signature genes were also able to cluster the samples of the relevant CT using unsupervised hierarchical clustering (**Suppl. Fig. 5A-D**). Together, our results demonstrate that the gene signatures identified by Decon-cell using only whole blood gene expression data are predictive for the proportions of circulating immune cell subpopulations.

To facilitate the cell proportion prediction of new samples using whole blood RNAseq, we have made the Decon-cell prediction models and gene signatures available in an R package (Decon-cell) and as a web tool (www.molgenis.org/deconvolution). These two implementations allow users to pre-process their RNA-seq expression counts and estimate cell proportions using the pre-established models for [34] cell types in whole blood. In addition, the Decon-cell R package allows users to generate Decon-cell-like gene signatures to predict their own cell proportions, which requires the input of bulk expression profiles and cell proportions to generate new Decon-cell predictive models.

Decon-eQTL identifies which cell types contribute to the whole blood eQTL effect

As we know, eQTL analysis using whole blood bulk expression data fails to distinguish between a general eQTL present in all cell types and an effect mainly found in a subset of the cell types. We therefore propose a new approach, called Decon-eQTL, that assigns the overall bulk eQTL into CT effects. Using the cell proportions in whole blood, it is possible to formally test if the genetic effect is interacting with the cell proportions. More explicitly, we include both the genotype and all major CT proportions of interest in a linear model, and systematically test if there is a significant interaction effect between

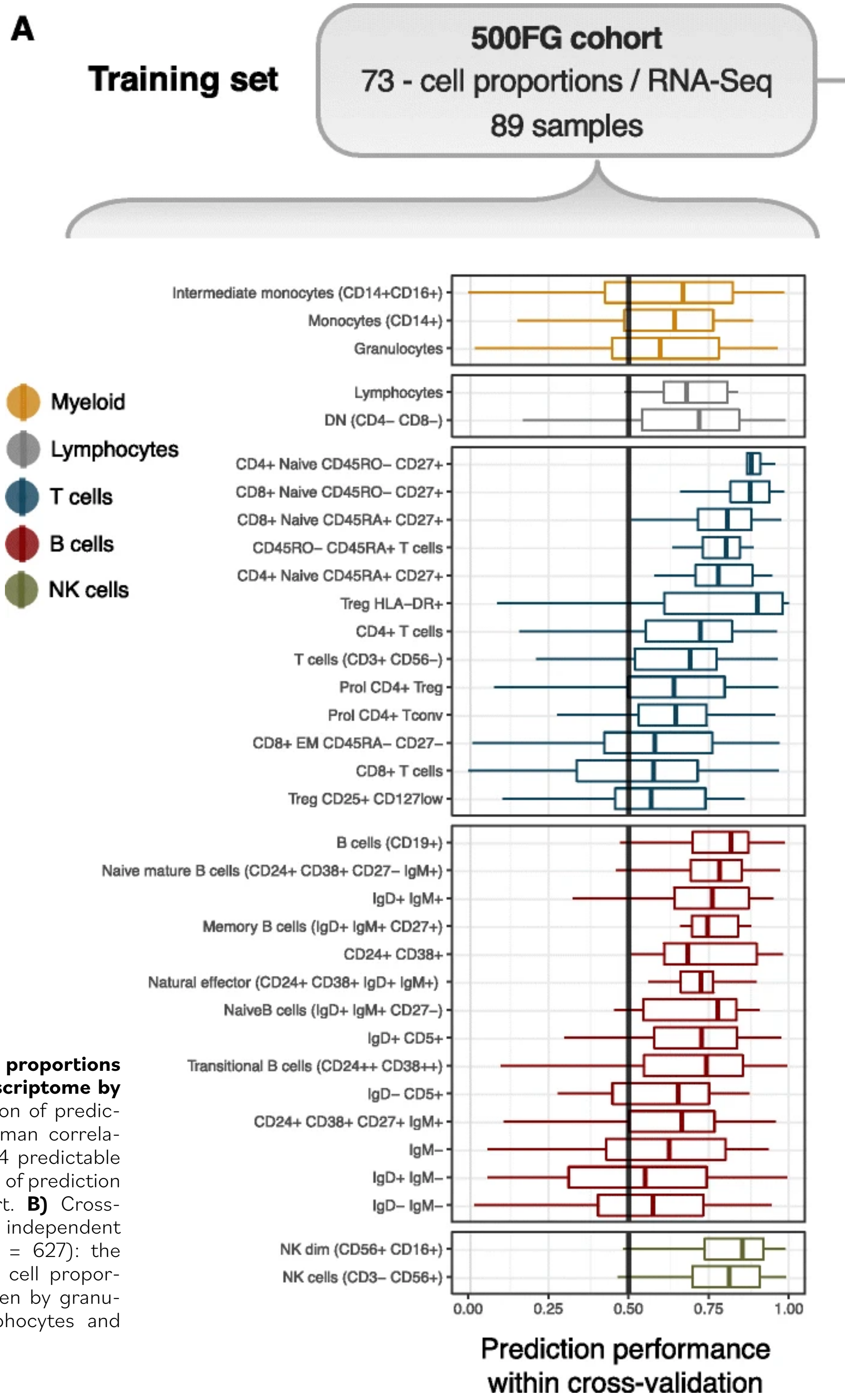
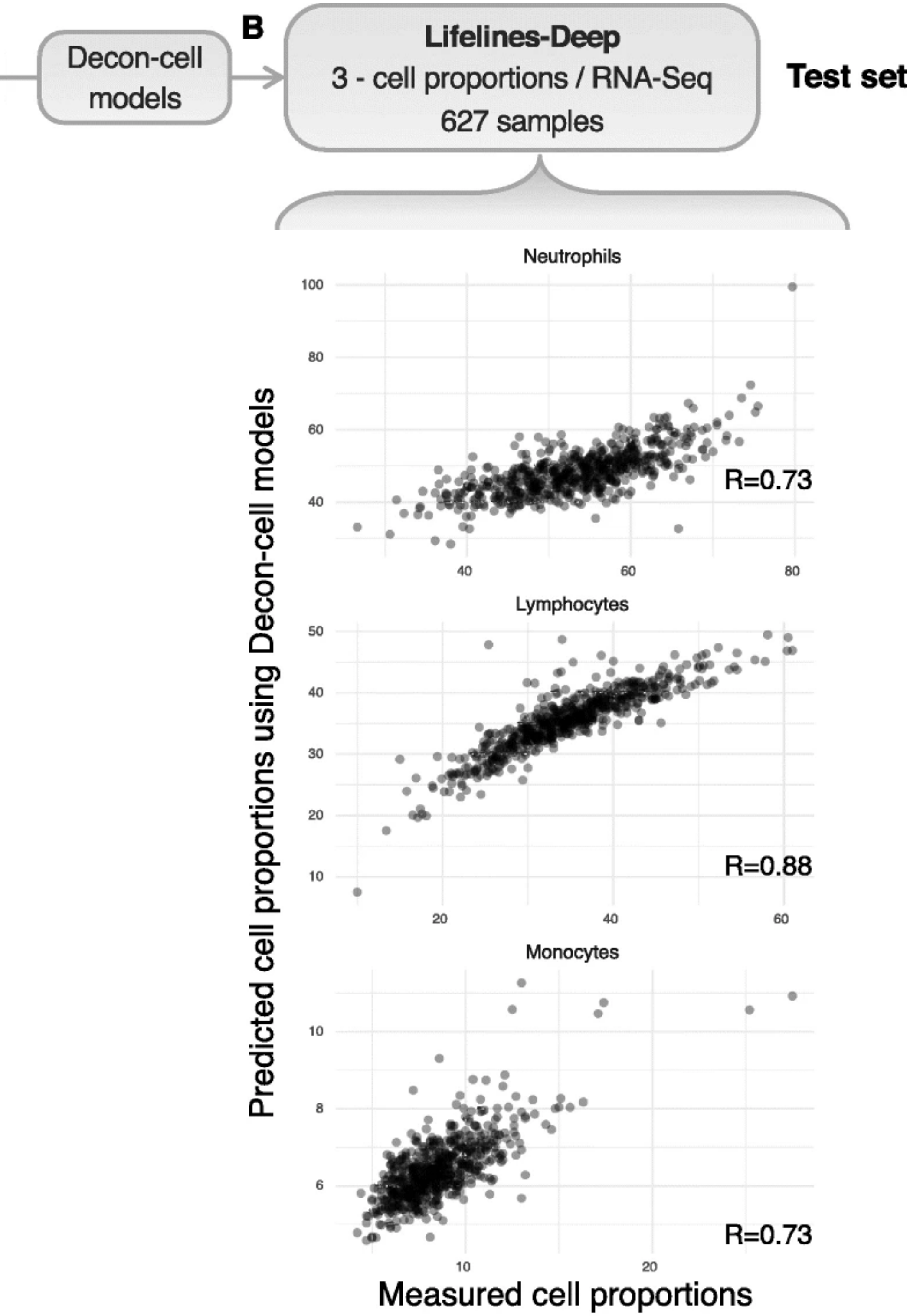


Fig. 2 Prediction of cell proportions using whole blood transcriptome by Decon-cell. A) Distribution of prediction performance (Spearman correlation coefficient) of the 34 predictable cell types in 100 iterations of prediction within the 500FG cohort. **B)** Cross-cohort validation in an independent Lifelines-Deep cohort (n = 627): the measured and predicted cell proportions for neutrophils (given by granulocytes in 500FG), lymphocytes and monocytes are compared


B
Decon-cell models
Lifelines-Deep
3 - cell proportions / RNA-Seq
627 samples
Test set
Neutrophils
100
80
60
40
40
60
80
R=0.73
Lymphocytes
50
40
30
20
10
20
40
60
R=0.88
Monocytes
10
8
6
10
20
R=0.73
Predicted cell proportions using Decon-cell models
Measured cell proportions

genotype and each of the cell proportions in the variation of gene expression in whole blood. At the same time, the model used by Decon-eQTL controls for the effects of the remaining cell types on gene expression. In this way, whole blood expression data, genotypes and (predicted) cell proportions can be integrated to assign a CTi effect from a bulk eQTL (**Fig. 1**).

We applied Decon-eQTL to 3198 samples (BIOS cohort) with transcriptome levels (RNA-seq), genotype information and cell proportions predicted by Decon-cell. Whole blood *cis*-eQTL mapping yielded 16,362 whole blood eQTLs (false discovery rate (FDR) ≤ 0.05). For each of these whole blood *cis*-eQTLs, we applied Decon-eQTL with a focus on 6 major cell subpopulations: granulocytes, CD14+ monocytes, CD4+ T cells, CD8+ T cells, B cells and NK cells. These cell types were selected because the sum of their relative percentages was close to 100% and none of these cell type pairs had an absolute correlation coefficient R ≥ 0.75. Decon-eQTL computationally assigned 4139 CTi eQTLs from these subpopulations, reflecting 3812 genes and 3650 SNPs. 25% of the whole blood eQTLs have a significant (FDR ≤ 0.05) CTi eQTL effect given DeconeQTL. The majority (31%) of the total CTi eQTL effects detected were found to be associated to granulocyte proportions, possibly because granulocytes comprise ~ 70% of circulating white blood cells (**Fig. 3A**). The majority (74%) of CTi eQTLs detected by our method were assigned to a single cell type (**Suppl. Fig. 6A**). Similarly, we find almost no sharing between cell types in single-cell eQTLs from 112 individuals. However, it should be noted that these eQTLs are likely not exclusively present for this particular cell type in biology, but that the statistical power given our sample size was sufficient to detect the interaction effects that we describe as CTi eQTL in this particular cell type. Decon-eQTL was only able to find a few cases of sharing of CTi eQTLs between cell types, likely due to a lack of power of the interaction model. An example of such a shared CTi eQTLs can be seen for the *NOD2* gene, where Decon-eQTL detected a strong granulocyte-eQTL effect alongside a smaller opposite effect in CD14+ monocytes. This opposite effect has also been previously described in eQTL studies on purified CD14+ monocytes and neutrophils [8]. These results demonstrate that the effects of cell proportions on gene expression should be taken into account when interpreting eQTLs derived from bulk tissues.

Decon-eQTL prioritizes genes to relevant cell types

CTi eQTL genes are expected to have higher expression levels in their relevant cell types, and their expression in whole blood should therefore be correlated with the proportions of these relevant cell types. To test this, we evaluated if the expression levels of the CTi eQTL genes detected in the BIOS cohort were correlated with their relevant cell proportions, and compared this to the correlation with nonrelevant cell types. We calculated the Spearman correlation coefficients between the expression of the identified CTi eQTL genes and the measured cell proportions in the 500FG cohort (n = 89). We then compared the correlation coefficients we obtained here with those between expression and the remaining cell proportions. For each of the six cell subpopulations we evaluated in Decon-eQTL, their CTi eQTL genes had a significantly higher correlation with their relevant cell subpopulation than with other cell types (t-test, p-value < 0.05) (**Fig. 3B**). As such, this result shows a significant association between CTi eQTL genes and the proportion of their relevant CT in an independent cohort.

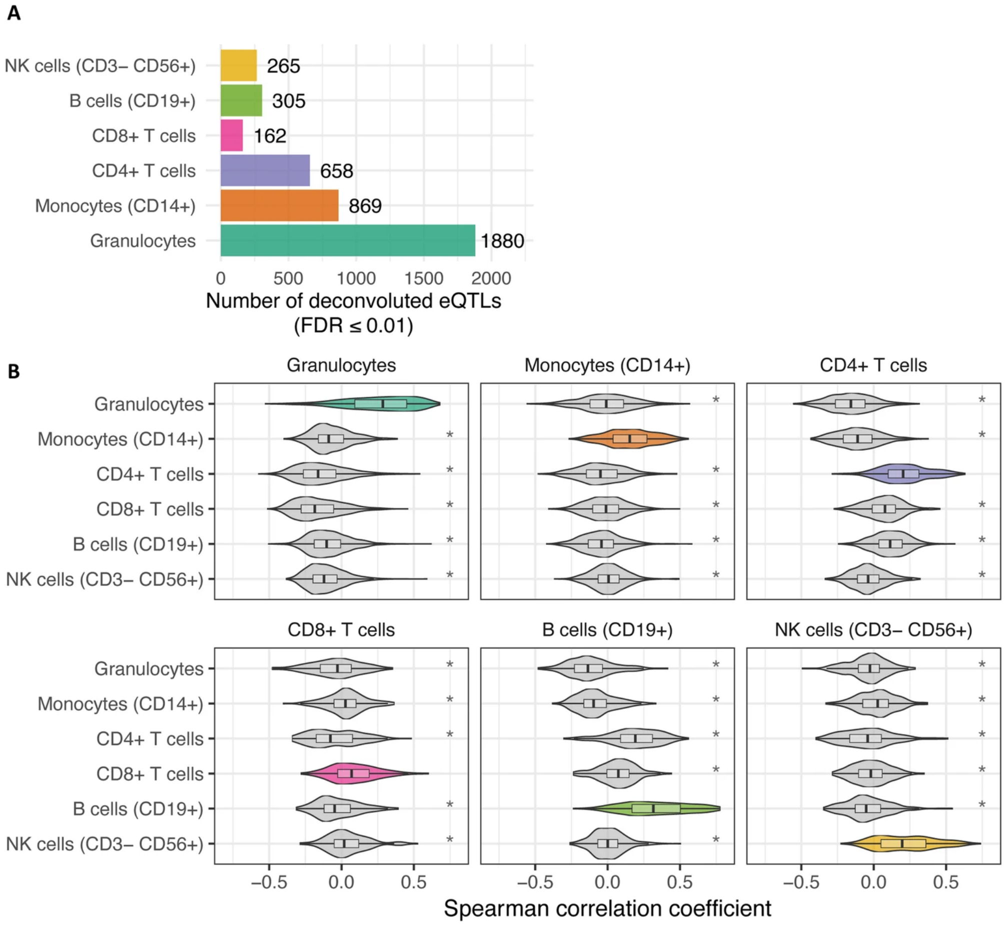
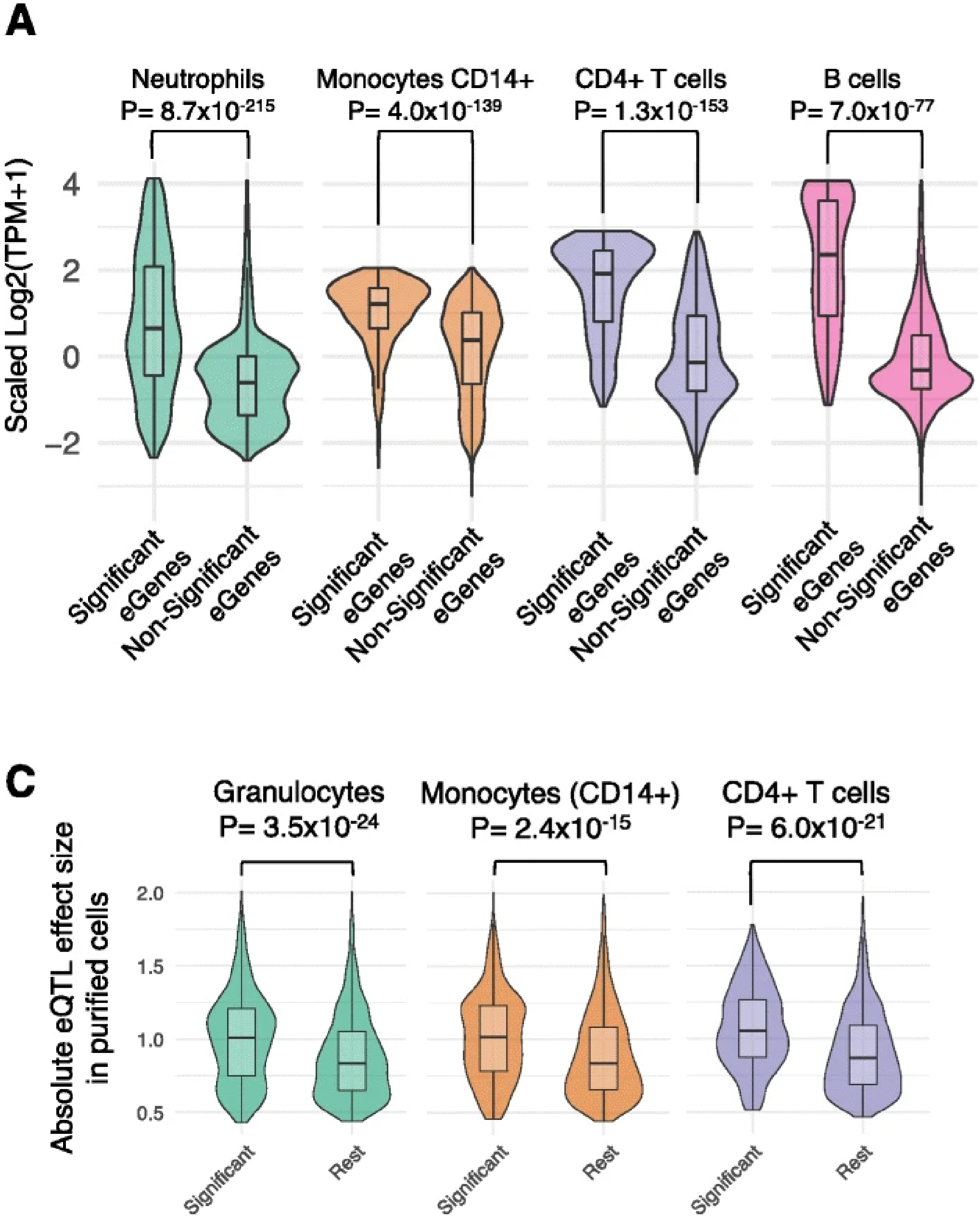


Fig. 3 Deconvolution of whole blood eQTLs into CTi eQTLs. Decon-eQTL detects CTi eQTLs by integrating proportions of cell subpopulations (predicted by Decon-cell), gene expression and genotype information. **A)** Number of deconvoluted CTi eQTLs in each cell type using whole blood RNA-seq data of 3189 samples in BIOS cohort. **B)** Distribution of Spearman correlation coefficients between expression levels of CTi eQTL genes and cell counts for each cell subpopulation. The CTi eQTL genes show positive and statistically higher correlation (Spearman) with the relevant cell type proportions as compared to the rest (t-test pvalue < 0.05) in an independent cohort (500FG)

Next, we evaluated whether the significant CTi eQTL genes were over-expressed in their relevant cell subpopulation compared to eQTL genes that were found to be nonsignificant CTi eQTLs for the same cell type. For this purpose, we made use of the purified neutrophil, CD14+ monocyte, CD4+ T cell and B cell RNA-seq data from the BLUEPRINT dataset. We include these cell types because they were the only ones with more than three samples measured. For each of the four cell types, we observed that the expression of CT eQTL genes detected by Decon-eQTL was significantly higher (t-test, p-value ≤0.05) than the expression of non-significant Decon-eQTL genes (**Fig. 4A**). We also observed that the deconvoluted eQTL genes from granulocytes showed a relatively wider range of variation than the CT eQTL genes from the other three subpopulations. We hypothesized that this could be explained by the fact that granulocytes comprise ~ 70% of the cell composition in whole blood, thus giving us the power to detect eQTL for lowly expressed genes in granulocytes. This is partly supported by the observation that the variation of expression in whole blood of granulocyte CTi eQTL genes was


A
Neutrophils
P= 8.7x10-215
Monocytes CD14+
P= 4.0x10-139
CD4+ T cells
P= 1.3x10-153
B cells
P= 7.0x10-77
Scaled Log2(TPM+1)
4
2
0
−2
Significant eGenes
Non-Significant eGenes
Significant eGenes
Non-Significant eGenes
Significant eGenes
Non-Significant eGenes
Significant eGenes
Non-Significant eGenes
C
Granulocytes
P= 3.5x10-24
Monocytes (CD14+)
P= 2.4x10-15
CD4+ T cells
P= 6.0x10-21
Absolute eQTL effect size in purified cells
2.0
1.5
1.0
0.5
Significant
Rest
Significant
Rest
Significant
Rest

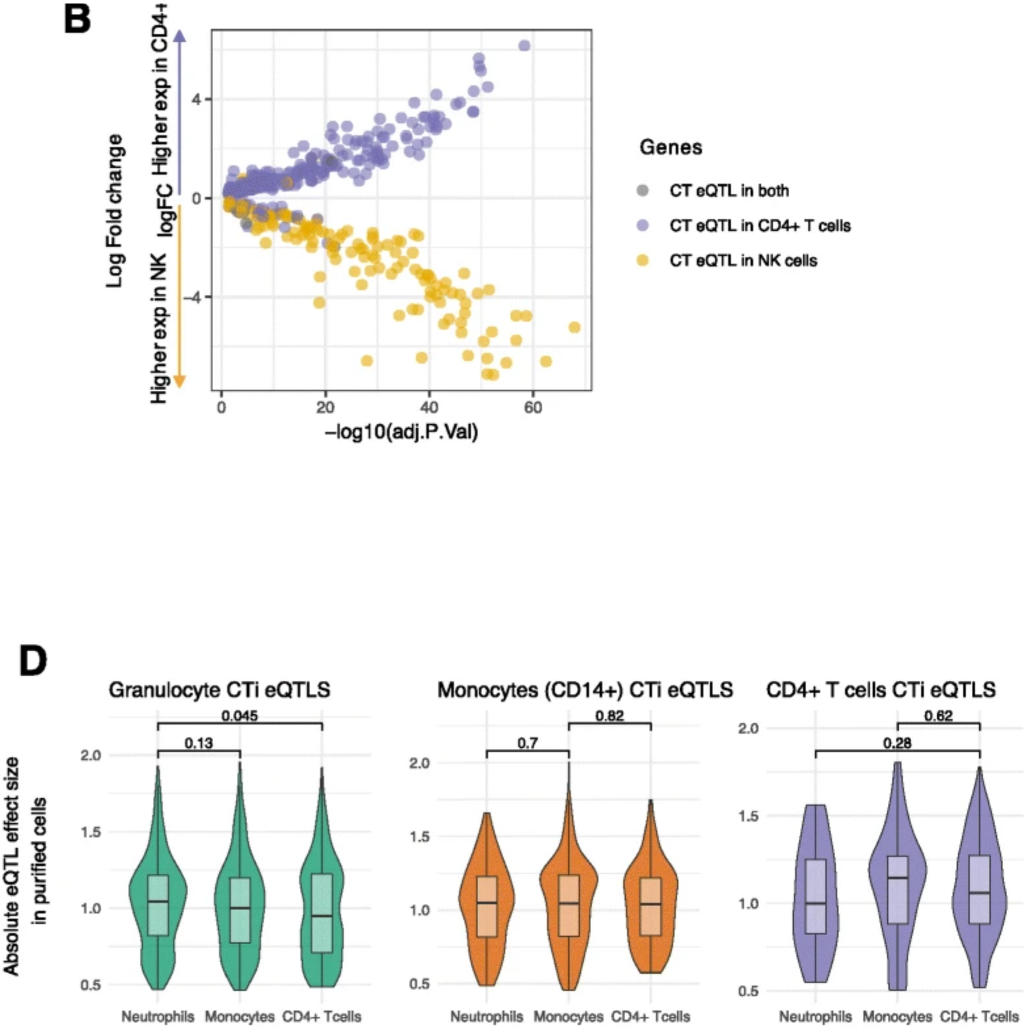


Fig. 4 Validation of CTi eQTLs. A) The expression of CTi eQTL genes in purified cell subpopulations from BLUEPRINT [23] are significantly higher in the relevant cell subpopulation when compared to other available cell subtypes (green for granulocyte eQTL genes showing expression for purified neutrophils; orange for monocytes; purple for CD4+ T cells; pink for B cells). **B)** Genes differentially expressed (Adjusted p-value ≤0.5) between CD4+ T cells and NK cells are significantly enriched for CT eQTLs effects on CD4+ T cells (dots in purple, Fisher exact p=1.8x10^{-17}) and NK Cells (dots in yellow, Fisher exact p=2.3x10^{-18}), respectively. **C)** CTi-eQTLs (FDR ≤ 0.05) show significantly larger effect sizes in the purified cell eQTL data [9] compared to the rest of the whole blood eQTLs for which we do not detect a cell type effect, as shown for deconvoluted granulocyte eQTLs in neutrophil-derived eQTLs (green),-monocytes (orange) and CD4+ T cells (purple). **D)** As C, but showing only the CTi-eQTLs on the X-axis.

significantly greater than for those CTi eQTL genes deconvoluted to the other five cell subpopulations (F-test, p-value ≤0.05, **Suppl. Fig. 7**).

Furthermore, by using publicly available transcriptome profiles (GSE78840 [27]) of purified NK cells and CD4+ T cells, we assessed if the differentially expressed genes across the two cell types were enriched for eGenes of deconvoluted CT eQTLs. Here we observed that the CD4+ differentially expressed genes (Adjusted p-value ≤0.05) were significantly enriched for CD4+ T cell eQTLs (Fisher exact $p=1.8x10^{-17}$), whereas NK cell differential genes (Adjusted p-value ≤0.05) were significantly enriched for NK cell eQTLs (Fisher exact $p=2.3x10^{-18}$) as shown in **Fig. 4B**.

In summary, we were able to show that the eQTL genes detected by Decon-eQTL are transcriptionally active in their relevant cell type because that is where they are more highly expressed.

CT eQTLs identified by Decon-eQTL in whole blood are replicated in purified cell eQTL datasets

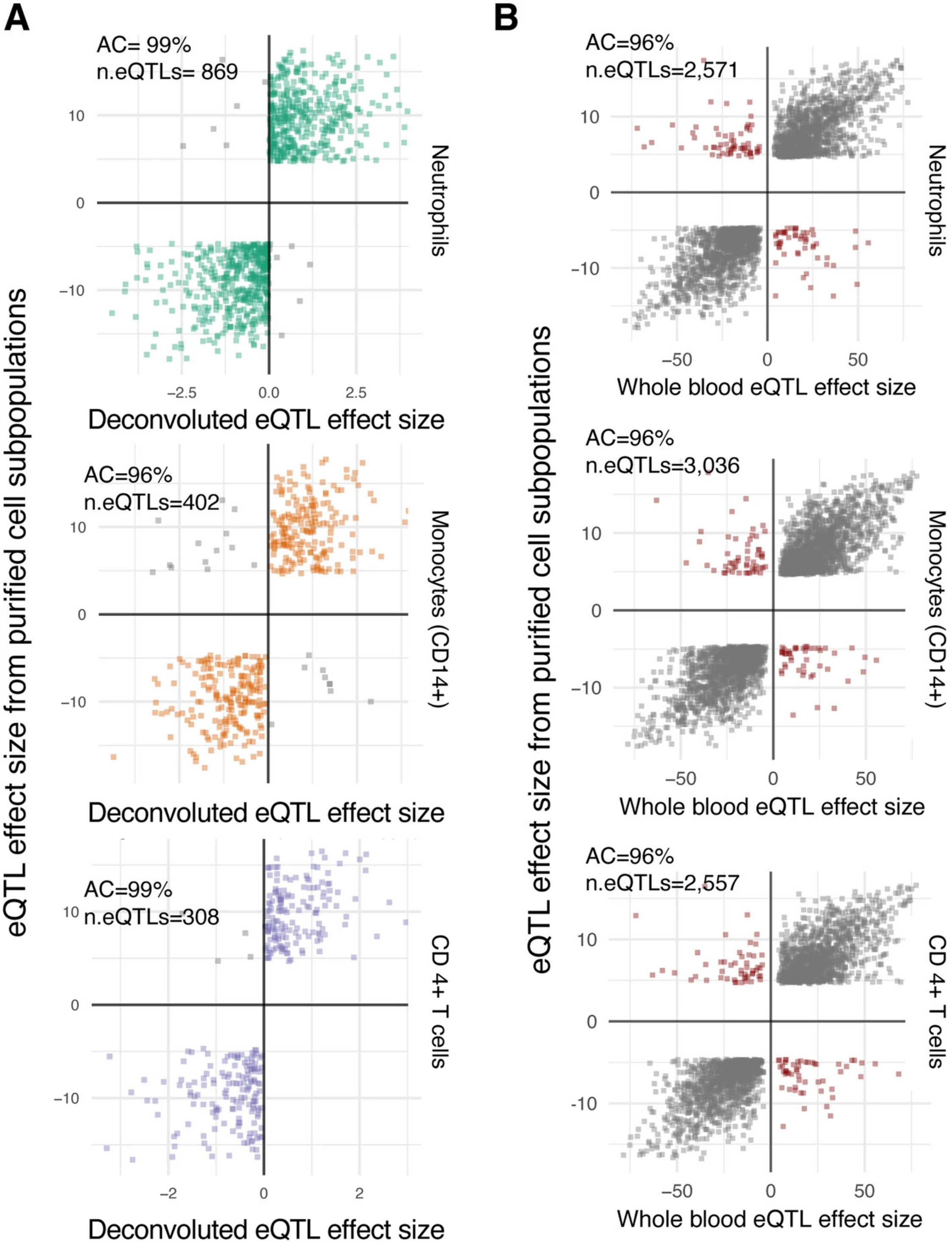
To validate the CT eQTLs defined by Decon-eQTL, we utilized eQTLs identified from purified neutrophils, CD4+ T cells and CD14+ monocytes [9]. We first compared the absolute effect sizes of eQTLs from purified cells that are also significantly deconvoluted CTi eQTLs to the effect sizes of eQTLs from purified cells that are also nonsignificant deconvoluted CTi eQTLs for this cell type. For all three cell populations, effect sizes in our deconvoluted CTi eQTLs were significantly higher than the effect sizes of eQTLs without a significant CTi eQTL (Wilcoxon test, p-value ≤0.05, **Fig. 4C**). Next, we assessed the specificity of our deconvoluted CTi eQTLs by evaluating CTi eQTL effect sizes in non-relevant cell subpopulations. For example, we compared the effect sizes of deconvoluted granulocyte CTi eQTLs against those with non-significant deconvoluted granulocyte CTi eQTLs using the effect sizes of purified CD4+ T cell eQTLs. Notably, we observed no statistically significant differences using effect sizes from nonrelevant cell subpopulations (see off-diagonal comparisons in **Suppl. Fig. 8**), which further supports the biological relevance of our deconvoluted CTi eQTLs. However, when comparing the effect sizes in the purified eQTLs of only the CTi eQTLs that were significant across all three available cell subpopulations, we were not able to find significant differences (**Fig. 4D**). For example, the effect size of neutrophil CTi eQTLs is the same across neutrophils, monocytes CD14+ and CD4+ T cells.

To further demonstrate that the CTi eQTLs assigned by Decon-eQTL are biologically meaningful, we made use of the K27AC and K4ME1 epigenetic QTLs characterized in purified neutrophils, CD4+ T cells and monocytes CD14+ [9]. In a similar fashion to the above comparison of effect sizes with purified eQTLs, we compared the absolute effect sizes from both K27AC and K4ME1 QTLs from eQTLs for which Decon-eQTL detects a significant CTi effect to the effect sizes of the other whole blood eQTLs. Here we observed that for corresponding cell types, e.g. evaluating granulocyte CT eQTLs in K27AC QTLs from purified neutrophils, the distribution of the absolute effect sizes is significantly higher for the chromatin mark QTLs (cmQTLs) than for non-significant CT eQTLs, which provides epigenetic evidence that our method is able to correctly assign cell type eQTL effects, as shown in the diagonal comparisons for both K27AC QTLS (**Suppl. Fig. 9**) and K4ME1 QTLs (**Suppl. Fig. 10**). Notably, for the non-relevant cell subpopulations, we observed that only one comparison (granulocytes vs. CD14+

monocytes) shows statistically significant higher effect sizes for K27AC QTLs and K4ME1 QTLs. For the rest of the non-relevant comparisons (shown in the off-diagonal of both **Suppl. Fig. 9** and **Suppl. Fig. 10**), there are no statistically significant differences. Comparing the eQTL effect sizes in purified KC27AC and K4ME1 QTLs of only the significant CTi eQTLs across all three available cell subpopulations shows that the effect sizes from the relevant cell type are significantly stronger for all pairings except those between granulocytes and CD14+ monocytes (**Suppl. Fig. 11**).

In addition to the comparison of effect sizes, we compared the allelic concordance between deconvoluted eQTLs and eQTLs from purified cell subtypes [9]. For each available cell type (neutrophils, CD14+ monocytes, and CD4+ T cells), we evaluated whether the direction of the eQTL effect on deconvoluted CT eQTLs was the same as the one observed from purified cell subpopulations. The allelic concordance between the deconvoluted eQTLs and purified eQTLs was high across cell types: 99% for granulocyte eQTLs (compared to neutrophil eQTLs), 96% for CD14+ monocytes eQTLs and 99% for CD4+ T cells (**Fig. 5A**). These rates of allelic concordance are significantly higher for granulocyte and CD4+ T cell CTi eQTLs compared to those between whole blood eQTLs and eQTLs from purified cell subpopulations (**Fig. 5B**; Neutrophils, Fisher exact p-value=3.91×10^{-6}; CD4+ T cells Fisher exact p-value=0.005), whereas the allelic concordance for deconvoluted CD14+ monocyte eQTLs is the same as for whole blood eQTLs and purified CD14+ monocyte eQTLs (**Fig. 5B**). We also compared the allelic concordance of deconvoluted CTi eQTLs of specific cell types against the eQTLs of non-relevant purified subpopulations. Interestingly, the allelic concordance across non-relevant cell subtypes is consistently lower (off-diagonal **Suppl. Fig. 12**, Bonferroni-corrected Fisher exact p-value < 0.0001 for all comparisons). Higher allelic concordance across cell types was seen between deconvoluted granulocyte eQTLs and CD14+ monocyte eQTLs with a 95% allelic concordance, which shows that the direction of effect is often shared between related cell types.

Finally, we evaluated the allelic concordance rates for CTi eQTLs assigned by Decon-eQTL and K27AC QTLs from purified cell subpopulations. Here we observed a consistently high allelic concordance rate: 92% for granulocyte eQTLs (in purified Neutrophils), 87% for CD14+ monocytes and 92% for CD4+ T cells (boxed diagonal comparisons in **Suppl. Fig. 13**). These concordance rates are significantly higher than the ones between the whole blood eQTLs and K27AC QTLs from purified cell subpopulations (**Suppl. Fig. 14**) for neutrophils (Fisher exact test p-value=9.06×10^{-14}), CD14+ monocytes (Fisher exact test p-value=3.33×10^{-4}), CD4+ T cells (Fisher exact test p-value=8.64×10^{-9}). Moreover, we noticed a consistent decrease in allelic concordance rates when assessing the concordance of CT eQTLs in K27AC QTLs of non-relevant cell subpopulations (off-diagonal comparisons, **Suppl. Fig. 13**). Taken together, the results from allelic concordance rates between deconvoluted CTi eQTLs and eQTLs/K27AC QTLs from purified cell subpopulations add a further layer of evidence to support the biological relevance of deconvoluted CT eQTLs.


A
AC= 99%
n.eQTLs= 869
10
0
−10
−2.5
0.0
2.5
Deconvoluted eQTL effect size
Neutrophils
AC=96%
n.eQTLs=402
10
0
−10
Deconvoluted eQTL effect size
Monocytes (CD14+)
AC=99%
n.eQTLs=308
10
0
−10
−2
0
2
Deconvoluted eQTL effect size
CD 4+ T cells
eQTL effect size from purified cell subpopulations
B
AC=96%
n.eQTLs=2,571
10
0
−10
−50
0
50
Whole blood eQTL effect size
Neutrophils
AC=96%
n.eQTLs=3,036
10
0
−10
−50
0
50
Whole blood eQTL effect size
Monocytes (CD14+)
AC=96%
n.eQTLs=2,557
10
0
-10
−50
0
50
Whole blood eQTL effect size
CD 4+ T cells
eQTL effect size from purified cell subpopulations

CTi eQTLs identified by Decon-eQTL in whole blood show high allelic concordance with single-cell RNA-seq eQTLs

To replicate the deconvoluted CT eQTLs in the cell subtypes that were not available in Chen et al. [9] purified cell eQTLs, we utilized the recent single-cell RNA-seq eQTLs (sc-eQTLs) identified in CD14+ monocytes, NK cells, CD4+ T cells, CD8+ T cells and B cells [24], as well as new single cell eQTL data that was processed in the same way. In total, we used sc-eQTLs from 112 individuals. We selected all significant eQTLs for each of the cell types (non-classical and classical monocytes were combined) and compared them to the direction of the eQTL effect given by Decon-eQTL, hereby observing an allelic concordance of 96.42% (**Fig. 6A**).

Decon-QTL outperforms conventional interaction method

To our knowledge, our approach is the first to model the effect of multiple components of bulk blood RNA-seq simultaneously in an attempt to fully deconvolute gene expression levels into more precise cell type x genotype effects. Previous studies have used an interaction effect between genotype and cell proportions of one specific cell type to detect cell type eQTL effects using whole blood gene expression [10], [11], or used the correlation of the eQTL effect with cell type proxy genes [10, 11].

The Westra et al. method has often been used to detect cell type eQTL effects using bulk expression data and cell proportions [28-31]. In brief, it focuses on the effect of the GxE interaction (where E represents cell proportions) to explain the variation in gene expression and only incorporates one cell type at a time. To properly compare Decon-eQTL with the 'Westra method', we applied both methods to the BIOS cohort and detected CT eQTLs for the six cell subpopulations. Replication of CT eQTLs identified by the Westra method was done as described above for Decon-eQTL. Here we observed that the eGenes (i.e. genes with eQTLs) detected by the Westra method show significantly higher expression for granulocytes ($p=3.0x10^{-12}$, observed in purified neutrophils) and CD4+ T cells ($p=5.0x10^{-13}$) and B cells ($p=5.1x10^{-11}$), but not for CD14+ monocytes ($p=1$, see **Suppl. Fig. 15A**). Next, we found that the distribution of effect sizes in eQTLs from purified cells is significantly higher for the CT eQTLs detected using the Westra method when compared to the rest of the whole blood eQTLs ($p=2.2x10^{-47}$, $p=9.6x10^{-8}$ and $p =1x 10^{-47}$ for neutrophils, CD14+ monocytes and CD4+ T cells, respectively; boxed-diagonal comparisons in **Suppl. Fig. 15B**), showing similar results to the ones from Decon-eQTL (**Suppl. Fig. 8**).

When we compared the allelic concordance rates between the direction of effects given by the interaction term from the Westra method to those found in eQTLs from purified cell subpopulations, we observed that the allelic concordances for granulocytes eQTLs (99%, evaluated in neutrophils, $p > 0.05$) and CD4+ T cells 93% ($p > 0.05$) (**Suppl. Fig. 16**) are comparable to those observed for Decon-eQTL (**Fig. 4A**). Conversely, the allelic

Left: Fig. 5 Allelic concordance of CTi eQTLs with eQTLs from purified cells. CTi eQTLs show high allelic concordance compared to eQTLs from purified cell subpopulations [9]. **A)** for granulocyte eQTLs (green), CTi eQTLs achieved an allelic concordance of 99% compared to eQTLs from purified neutrophils. Similarly, the allelic concordances were 96 and 99% for CD14+ monocytes and CD4+ T cells, respectively. Except for monocytes, these values are higher than those observed for whole blood eQTLs when comparing to eQTLs from purified subpopulations, as shown in panel **B**

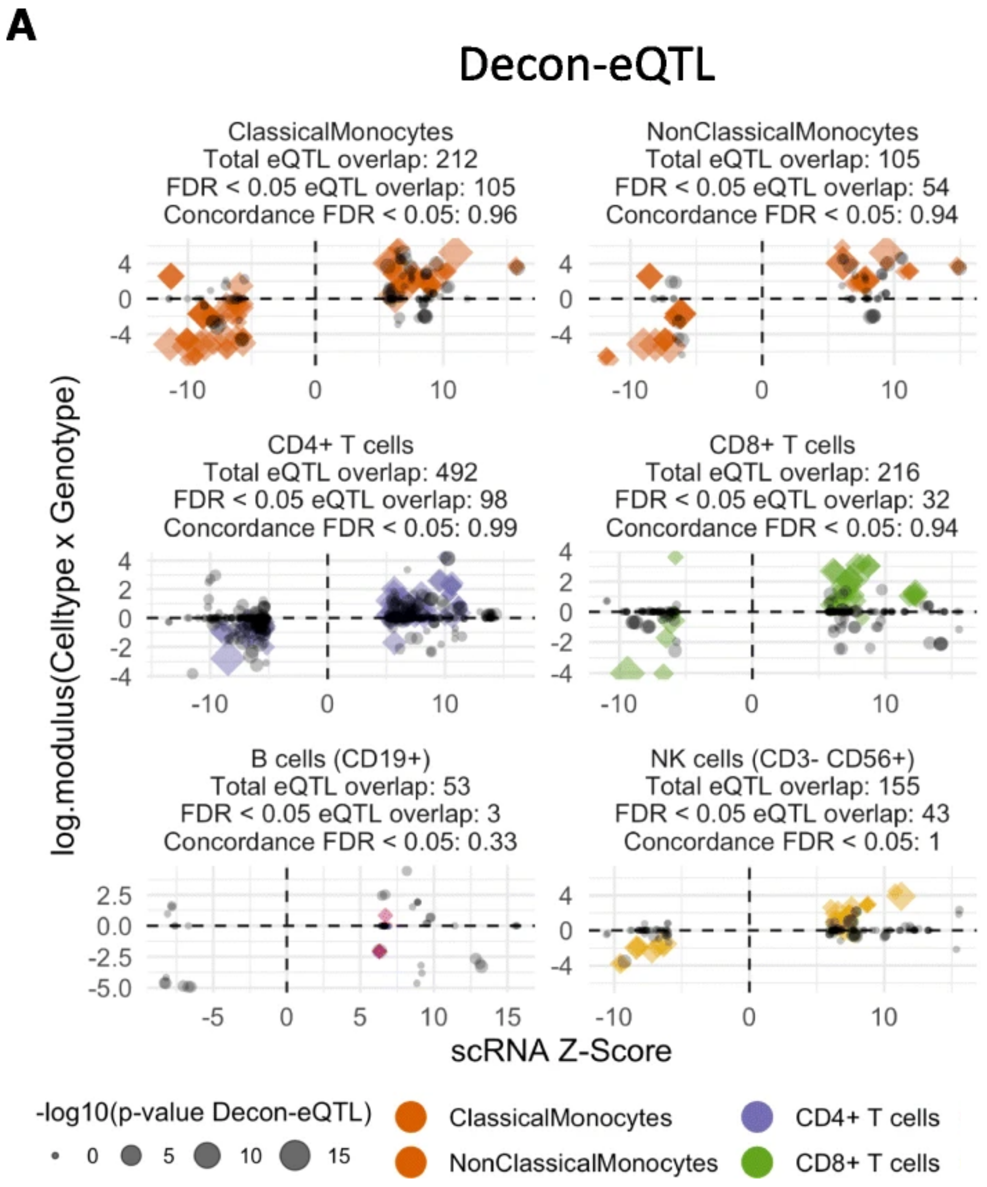
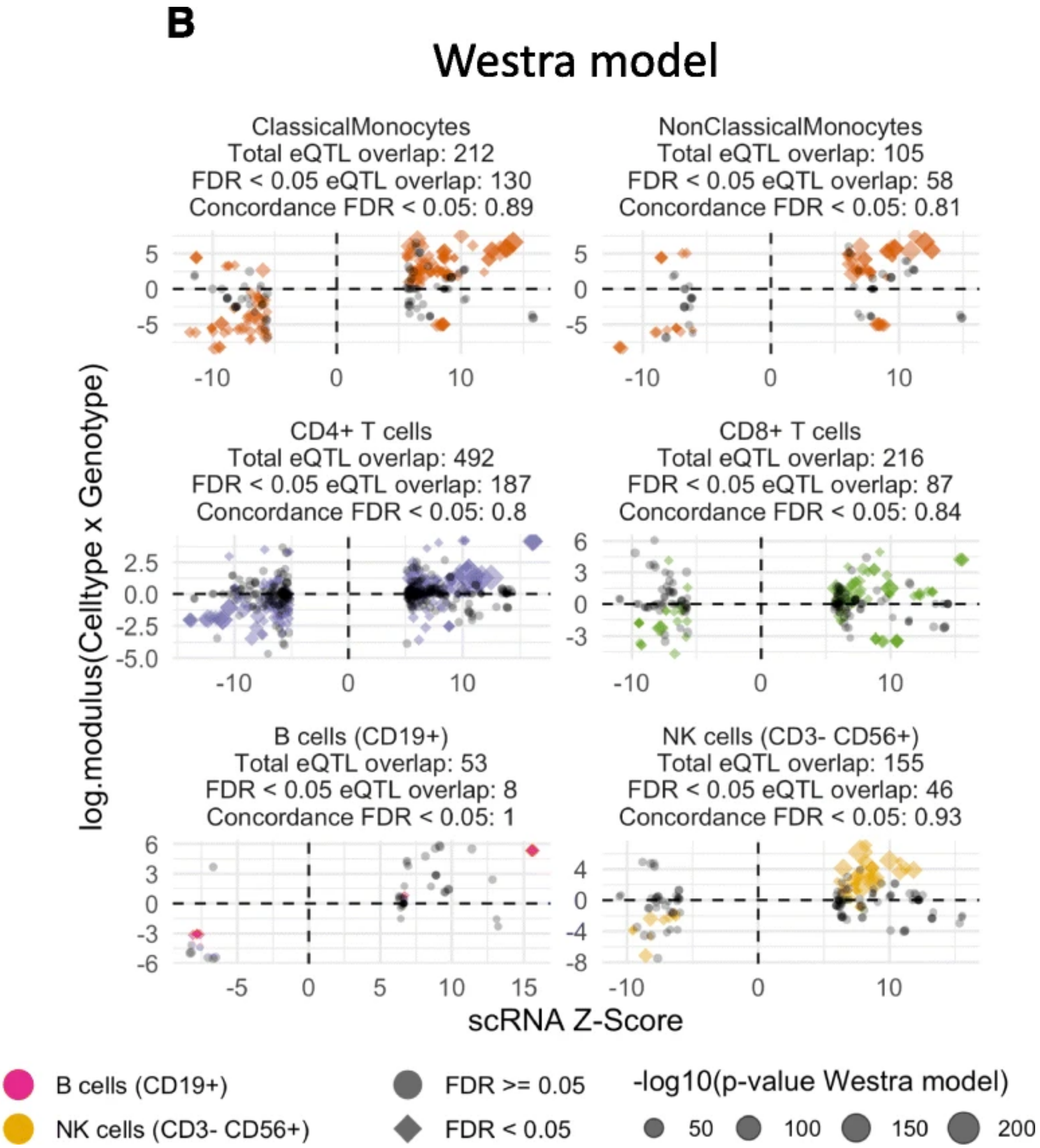


Fig. 6 Allelic concordance of CTi eQTLs with eQTLs from single cell RNAseq. A) Comparison in allelic direction between CTi eQTLs and eQTLs from single cell RNAseq experiments in 6 cell types. **B)** Comparison in allelic direction between Westra model eQTLs and single cell eQTLs. In both panels coloured diamonds are FDR < 0.05, grey circles are FDR > = 0.0 in the single cell data, and the size is the -log10(p-value) of the predicted cell type interacting eQTLs


B
Westra model
ClassicalMonocytes
Total eQTL overlap: 212
FDR < 0.05 eQTL overlap: 130
Concordance FDR < 0.05: 0.89
NonClassicalMonocytes
Total eQTL overlap: 105
FDR < 0.05 eQTL overlap: 58
Concordance FDR < 0.05: 0.81
CD4+ T cells
Total eQTL overlap: 492
FDR < 0.05 eQTL overlap: 187
Concordance FDR < 0.05: 0.8
CD8+ T cells
Total eQTL overlap: 216
FDR < 0.05 eQTL overlap: 87
Concordance FDR < 0.05: 0.84
B cells (CD19+)
Total eQTL overlap: 53
FDR < 0.05 eQTL overlap: 8
Concordance FDR < 0.05: 1
NK cells (CD3- CD56+)
Total eQTL overlap: 155
FDR < 0.05 eQTL overlap: 46
Concordance FDR < 0.05: 0.93
log.modulus(Celltype x Genotype)
scRNA Z-Score
5
0
-5
-10
0
10
2.5
0.0
-2.5
-5.0
6
3
0
-3
-6
4
-4
-8
5
15
B cells (CD19+)
NK cells (CD3- CD56+)
FDR >= 0.05
FDR < 0.05
-log10(p-value Westra model)
50
100
150
200

concordance rate for CD14+ monocytes is only 62%, significantly lower than the results from Decon-eQTL (96%, p=0.001). Finally, for granulocytes, CD4+ T cell eQTLs and monocytes, we overlapped the results from Westra method and Decon-eQTL with the eQTLs from purified cell types (Chen et al. 9) (**Suppl. Fig. 17**). For all three cell types, we found that Decon-eQTL is able to detect a larger number of eQTLs. For neutrophils, the Westra method has a higher replication rate (Fisher p-value=0.002). For CD14+ monocytes, both methods had the same replication rate (Fisher p-value=0.737). For CD4+ T cells, Decon-eQTL had a better replication rate (p-value=7.47×10^{-12}).

Finally, we compared the difference in allelic concordance with sc-eQTLs. The overall allelic concordance of Decon-eQTL CTi QTLs with sc-eQTLs (96.42%, **Fig. 6A**) is higher than that achieved by the Westra model (p=1.24×10^{-8}), where we observed an overall allelic concordance of 84.67% (**Fig. 6B**). For both non-classical monocytes (Fisher p-value=0.045) and CD4+ T cells (Fisher p-value=7.89×10^{-7}), Decon-eQTL has a significantly better allelic concordance. For CD8+ T cells (Fisher p-value=0.230), classical monocytes (Fisher p-value=0.0513), B cells (Fisher p-value=0.055) and NK cells (Fisher p-value=0.242), there is no significant difference. Nevertheless, Decon-eQTL shows a higher allelic concordance for NK cells, classical monocytes, and CD8+ T cells (93.8% vs 83.9, 96.2% vs 89.2, and 100% vs 93.5% respectively), while for B cells it has lower concordance (33% vs 100%).

Overall, these results demonstrate that Decon-eQTL is able to detect more CTi eQTLs that can be replicated in purified eQTL dataset than previously reported methods, especially in less abundant cell types such as CD14+ monocytes. However, the detection of interaction effects between genotype and cell proportions in order to dissect bulk (in this case whole blood) expression data and CTi eQTLs remains an area of great opportunity that could still be explored, particularly given the constantly increasing number of samples present in functional genomic cohorts and the growing numbers of purified and sc-eQTL datasets that can be used for validation.

Discussion

We have developed a novel statistical framework, Decon2, that predicts the proportions of known immune cell subtypes using gene expression levels from whole blood (Decon-cell). These predicted cell proportions can then be used together with genotype information and expression data to deconvolute a whole-blood eQTL effect into cell type interacting effects (Decon-eQTL). Using a set of samples with both whole blood RNA-seq data and cell frequencies of 73 cell subpopulations, we demonstrated that Decon-cell was able to predict 34 independent cell subpopulations. The performance of Decon-cell has been validated in multiple independent cohorts and benchmarked with existing methods. The Decon-cell models were then applied to a cohort of 3189 samples for which whole blood RNA-seq data was available, resulting in predicted cell counts for these samples. By integrating bulk expression data, genotype and predicted cell counts of the BIOS cohort, Decon-eQTL was able to dissect whole blood eQTL effect into CTi eQTLs without purifying immune cell subpopulations. The results of Decon-eQTL were then validated again using several independent data types: 1) eQTLs from purified cell subpopulations, 2) chromatin QTLs of purified cells, 3) gene expression from purified cell types and 4) eQTLs derived from single-cell protocols. Compared with existing methods, Decon-eQTL consistently shows superior performance. To sum up, the proposed framework is useful for (re)-analyzing both existing and new bulk blood tissue datasets in order to detect CTi eQTL effects and can be applied and tested on other tissues once cell count proportions become available. Cataloging and further interpreting the role of CTi eQTLs will improve our understanding of the functional role of the SNPs associated with complex diseases at the level of specific cell subtypes.

The main advantage of our Decon-cell method for predicting cell proportions is that it does not rely on the gene expression measured in purified cell subtypes when defining signature gene sets. Moreover, our method does not require the definition of marker genes based on their differential expression compared to other cell subpopulations, unlike previously reported methods [12]. The signature genes defined by Decon-cell are determined using a completely unsupervised approach that applies regularized regression to select the optimal combination of genes to accurately predict a certain circulating cell proportion. The majority of these marker genes are differentially expressed across purified cell subpopulations, but not all. Nevertheless, these signature gene sets are still correlated to the cell proportions in whole blood. In summary, we have shown that Decon-cell can accurately predict the proportions of circulating immune cell subpopulations in three independent cohorts and that it out-performs previously reported methods within these cohorts.

Our Decon-eQTL method for detecting a CTi eQTL effect with bulk blood tissue expression data is, to our knowledge, the first attempt to simultaneously model whole blood gene expression profiles into its major components. In contrast to a previous method where single cell type (G x E) effects were evaluated one at a time [10, 31], Decon-eQTL incorporates all the major cell proportions simultaneously to better dissect the overall genetic effect of gene expression signal into cell subpopulation effects. We have shown that CTi eQTL genes found with Decon-eQTL have higher expression and higher effect sizes in purified neutrophils, CD14+ monocytes and CD4+ T cells than do non-CTi genes, and we find significantly higher allelic concordance for two out of four tested cell types with sc-eQTLs than with a conventional interaction model (**Fig. 6A, B**). Moreover, we have also shown the biological relevance of the deconvoluted CTi eQTLs by validating our results on cmQTLs where CTi eQTLs have significantly higher effect sizes and allelic concordance rates are significantly higher than those of whole blood eQTLs. Finally, we have also demonstrated that Decon-eQTL can replicate sc-eQTLs derived from scRNA-seq data, showing a higher allelic concordance with sc-eQTLs than when using only whole-blood eQTL effects.

There are also limitations to our method. The CTi eQTLs detected by Decon-eQTL tend to be eQTL exclusive for the specific CT, suggesting that the CT with the strongest eQTL effect was selected by Decon-eQTL. This is likely due to the partial collinearity present between the CT proportions included in the model (as shown by their correlation structure in **Suppl. Fig. 18A-B**). Thus, the genetic effect of one cell type might be masked by another CT with a correlated cell proportion. The highest correlation coefficient among cell types included in the model was 0.75 (between granulocytes and B cells). Therefore, deconvoluting CTi eQTLs for partially correlated cell proportions could lead to false negative results for cell types with relatively weaker eQTL effects.

In our model, we included the six major blood cell types, but there are many more cell types available for which our method is not able to detect a CTi eQTL estimate. Furthermore, we only tested Decon-eQTL using genome-wide whole blood *cis*-eQTLs main effects. Such eQTL effects are very likely shared across multiple cell types, however we are only able to detect its interaction with only one cell type due to statistical power and co-linearity (**Suppl. Fig. 6A**), which is also seen in the sc-eQTLs with limited (112) samples (**Suppl. Fig. 6B**). Nevertheless, this does not imply that the CTi eQTL are exclusive for, or only present in, that specific cell type, as we observe in **Fig. 4D**, where the effect sizes of the significant CTi eQTLs in purified subpopulations are not significantly different across all three purified cell subpopulations. Yet this difference in the effect-size of CTi eQTLs between relevant and non-relevant cell types can be seen in histone modification QTLs (as shown in **Suppl. Fig. 11**), likely due to the cell type-specificity of epigenetic marks. Lastly, Decon2 has only been tested in whole blood, where large numbers of samples are available, and therefore it is not known how it will perform in other tissues.

The proposed framework of Decon2 is generic for predicting cell subpopulations in bulk tissues (Decon-cell) and re-distributing the overall eQTL effect into cell types (Decon-eQTL). Both methods have been implemented in freely available software. In both the R package and the user interface-based webtool, we provide the models for predicting cell subpopulation in whole blood that were constructed and validated in this work so that interested users can estimate immune cell subpopulations in whole blood in healthy people of western European ethnicity, as our models were built using a Dutch cohort (500FG).

Conclusion

In summary, Decon2 is a computational method that can accurately assign CT effects in whole blood functional genomic cohorts. It can be applied to any dataset for which genotypes and expression data is available and could potentially aid in understanding the molecular effects of genetic risk factors associated with complex diseases at the cell-subpopulation level. Our method makes it possible to create CT gene regulatory networks that could explain the different effects that each CT has on a complex disease in a cost-efficient way. Since Decon2 only requires gene expression and genotype information to deconvolute bulk blood eQTLs into CTi eQTLs, it is possible to re-analyze existing bulk blood RNA-seq data for which genotypes are also available. In this scenario, we would use Decon-cell to predict cell proportions in whole blood and obtain CT information on many more eQTLs through an increase in sample size. In addition to whole blood, the methods behind Decon2 can potentially be generalized to use transcriptional profiles derived from any other type of bulk tissue, such as biopsies from tumors or other solid tissues implicated in complex disease etiology. However, the method has not yet been tested in other tissues. Our methods can hence aid in the detection of genetic effects on gene expression in rare cell subpopulations in bulk tissues.

Methods

RNA-seq data collection in 500FG cohort

We selected a representative subset of 89 samples from the 500 participants of the 500FG cohort, which is part of the Human Functional Genomics Project (HFGP). Our subset was balanced for age and sex based on the original distribution in the cohort. RNA was isolated from whole blood and globin transcripts were subsequently filtered by applying the Ambion GLOBINclear kit. The samples were then processed for sequencing using the Illumina TruSeq 2.0 library preparation kit. Paired-end sequencing of 2x50-bp reads was performed on the Illumina HiSeq 2000 platform. The quality of the raw reads was checked using FastQC (http://www.bioinformatics.babraham.ac.uk/ projects/fastqc/). Read alignment was performed with STAR 2.3.0 [32, 33] using the human Ensembl GRCh37.75 as reference, and the aligned reads were sorted using SAMTools [34]. Lastly, gene-level quantification of the reads was done using HTSeq [35].

RNA-seq preparation and data processing in the BIOS cohort

RNA was isolated from whole blood and globin transcripts were subsequently filtered by applying the Ambion GLOBINclear kit. Library preparation was performed using the Illumina TruSeq v2 library preparation kit. Next, Illumina HiSeq 2000 was used for paired-end sequencing of 2x50 bp reads while pooling 10 samples per lane and expecting > 15 million read pairs per sample. Read sets were generated using CASAVA, retaining only those reads that passed Illumina Chastity Filter.

Quality control of the reads was evaluated using FastQC (http://www.bioinformatics. babraham.ac.uk/projects/fastqc/). Adaptor sequences were trimmed out using cutadapt (v1.1) with default settings. Low quality ends of reads were removed using Sickle (v1.200) (https://github.com/najoshi/sickle).

Reads were then aligned using STAR 2.3.0E [33]. All SNPs present in the Genome of the Netherlands (GoNL) with MAF ≥ 0.01 were masked from the reads to avoid reference mapping bias. Read pairs with at most eight mismatches and mapping to at most five positions were used. Quantification of counts per genes was done using Ensembl v.71 annotation (which corresponds to GENCODE v.16).

Genotype data of the BIOS cohort

Genotype information was independently generated for each of the cohorts, further details on data collection and methods used for genotyping can be found in their papers (CODAM [36], LLDeep [16], LLS [17], RS [18] and NTR [21]).

Genotypes were harmonized to GoNL with Genotype Harmonizer [37] and imputed with IMPUTE2 [38] using GoNL as reference panel. SNPs with an imputation score below 0.5, a Hardy-Weinberg equilibrium p-value smaller than $1x10^{-4}$, a call rate below 95%, or a MAF smaller than 0.05 were filtered out. For further analysis, only eSNPs from whole blood *cis*-eQTL top effects were subsequently used in Decon-eQTL.

Quantification of cell proportions in 500FG cohort

Inclusion criteria and further description of the participants of the 500FG cohort can be found at http://www.humanfunctionalgenomics.org. A total of 73 manually annotated immune cell subpopulations were quantified using 10-color flow cytometry. To minimize biological variability, cells were processed immediately after blood sampling and typically analyzed within 2–3 h. Cell populations were gated manually as previously described [14].

Cis-eQTLs in the BIOS cohort

For *cis*-QTL mapping, we tested for association between genes and SNPs located within 250 kb of a gene center. SNPs with MAF ≥ 0.01, call rate = 1 and Hardy-Weinberg equilibrium p-value ≥0.0001 were included. eQTLs were declared to be significant at FDR < 0.05. Pre-processing of RNA-seq and QTL mapping was performed using a custom eQTL pipeline that has been described previously [11].

Normalization and correction of gene expression data for deconvolution of eQTL effects Total read counts from HTSeq were first normalized using the trimmed means of M (TMM) values. TMM expression values were then log2 transformed. For predicting cell proportions, we used scaled expression data in both the 500FG and BIOS cohorts.

For the deconvolution of eQTLs, the expression was log2 transformed and corrected for the effects of cohort, age, sex, GC content, RNA degradation rates, library size and number of detected genes per sample using a linear model. The corrected expression data was then exponentiated to maintain the original linear relationship across read counts (gene expression) and cell proportions.

General description of Decon2

Decon2 is a statistical framework for estimating cell counts using molecular profiling such as expression data from heterogeneous samples (Decon-cell) and consecutive deconvolution of expression quantitative trait loci (Decon-eQTL) into each cell subpopulation. To predict cell proportion levels using Decon-cell built in models, it's only input is a matrix As input Decon-cell takes a table of normalized gene expression counts, with samples as columns and genes as rows, and outputs a table of predicted cell count proportions for cell types that were included in the training model. Decon-cell also enables the user to generate its own custom models, for which it requires a matrix of gene expression to train the model and a matrix of measured cell proportions; this will output a list with one specific model for each of the cell types included. A matrixtable of normalized gene expression levelscounts, a matrixtable of predicted or measured cell count proportions, and a matrixtable of genotype dosages (0 for homozygous reference, 1 for heterozygous, and 2 of homozygous alternative), lastlyand a table with the SNP + gene combinations to test, are used as input for Decon-eQTL, and this outputs for each SNP + gene combination the beta and p-value of the cell-type dependent eQTL effect. See **Suppl. Fig. 20** for a graphical overview.

Prediction of cell proportions using gene expression levels from bulk tissue (Decon-cell)

For cell count prediction, expression data is TMM normalized, log2(expression+ 1) transformed and z-transformed (scaled). We proposed that the abundance of molecular markers such as gene expression could be used as proxies to predict cell proportions. This can be represented as:

$$C_{kj} = \beta_{ki}\, Y_{ij} + e_{kj} \tag{1}$$

where expression data is Yij for genes i = 1, 2, ..., G and samples j = 1, 2, ..., N and cell count data is C_{kj} for sample j in cell type k (k = 1, 2, ..., K). β_{ki} represents the coefficients of gene i in determining cell counts of cell type k of a complex tissue. ekj is the error term.

In order to select only the most informative genes for predicting cell counts, we implemented a feature selection scheme by applying an elastic net (EN) regularized regression [26]. In the EN algorithm, the βk Y are estimated by minimizing:

$$\| C_k - \beta_k Y \|^2 \; subject\ to\ (1-\alpha) \| \beta_k \|^2 + \alpha \|\beta_k\|_1 \leq s \tag{2}$$

s is a tuning parameter that limits the number of features that will be included in the final predictor model. We estimate the best s per cell type by applying a 10-fold crossvalidation approach, where the most optimal penalty parameter (α) was obtained.

Deconvolution of eQTL effects (Decon-eQTL)

Decon-eQTL models the expression level in the bulk tissue by considering the genetic contribution of multiple cell types present in the system. For identifying the CT eQTL effect, the interaction term between a particular cell type and genotype was tested for statistically significant contribution to the explained variance on the expression levels of particular genes, while accounting for the remaining cell proportions. If we consider a generic eQTL linear model for whole blood it can be described as:

$$y = a + \beta . g + e \tag{3}$$

where y is the measured gene expression, a the modeled non-genetic dependent expression, g the genotype coded as 0, 1 or 2, β. g the genotype-dependent expression and e the error, e.g. unknown environmental effects. Here, all three terms are modeling the effect of the mixture of different cell types present in blood. In an RNA-seq-based gene expression quantification of a bulk tissue, one could express gene expression levels (y) as the sum of counts (ψ) per K cell types:

$$y = \sum_{k=1}^{K} \psi_k \tag{4}$$

For every cell type, the expression level can be written as a generic eQTL model (eq. 3) weighted by the cell proportions. ψ_k is a combination of the genetic and non-genetic contribution of the cell type to y. The non-genetic contribution per cell type is β. c, wherec is the cell count proportions. The genetic contribution is β_k. g : c_k. For k cell types the expression is then:

$$y = \sum_{k=1}^{K} \psi_k = \Sigma_k . \left(\beta_k . c_k \right) + \Sigma_k . \left(\gamma_{k.}\, g \times c_k \right) + e \tag{5}$$

where y is the measured expression levels, k is the total number of cell types, ck is the cell count proportions of cell type k, g is the genotype and e is the error term. Since we are assuming a linear relationship between total gene expression and the levels of expression generated by each of the cell types composing a bulk tissue, the cell proportions are scaled to sum to 100% such that the sum of the effect of the cell types equals the effect in whole blood. Here we assume that the true sum of the cell counts should be very close to 100% of the total PBMC count, which is why we include the 6 cell types that together form the top hierarchy given the gating strategy used to quantify the cell subpopulations [14]. The genotype main effect is not included in the model because the sum of the genotype effect per cell type should approximate the main effect.

Because the contribution of each of the cell types to expression level y cannot be negative, we constrain the terms of the model to be positive using Non-Negative Least Squares [39, 40] to fit the parameters to the measured expression levels. However, if the allele that has a negative effect on gene expression is coded as 2, the best fit would have a negative interaction term, which would be set to 0. To address this, we want the allele that causes a positive effect on gene expression to always be coded as 2. However, the effect of an allele can be different per cell type, therefore the coding of the SNP should also be different per cell type. We therefore run the model multiple times, swapping the genotype encoding for one of the interaction terms each time. The encoding that gives the lowest R-squared is then chosen as the optimal genotype encoding. For the encoding, we limit the number of genotypes that have an opposite genotypic encoding to a maximum of one interaction term, as we have observed that this leads to no significant difference when compared to using all possible configurations and limits the number of models that have to be run from k2 to (2*k) + 2.

To test if there is a CT interaction effect, we run the linear model of eq. 5 and, for each CT, run the same model with the cell proportion:genotype interaction term removed. For example, when testing two cell types the full model is:

$$y = \beta_1 . c_1 + \beta_2 . c_2 + \gamma_1 . g \times c_1 + \gamma_2 . g \times c_2 + e \tag{6}$$

and the two models with the interaction terms removed are:

$$\begin{aligned} y &= \beta_1 . c_1 + \beta_2 . c_2 + \gamma_1 . g \times c_1 + e \\ y &= \beta_1 . c_1 + \beta_2 . c_2 + \gamma_2 . g \times c_2 + e \end{aligned} \tag{7}$$

For both the full model and the CT models, we calculated the sum of squares using the different genotype configurations detailed above. For both the full and the CT models, we then selected the genotype configuration with lowest sum of squares. Then, for each CT, we tested if the full model could significantly explain more variance than the CT model using an ANOVA.

We then applied our strategy to 16,362 significant whole blood *cis*-eQTL top effects detected using the BIOS cohort. We then correct the p-values for multiple testing using FDR for each of the cell types, i.e. granulocyte eQTL p-values were corrected for 16, 362 tests in the same way as CD4+ T cells eQTL p-values were corrected for the exact same number of tests.

Westra et al. interaction model

In the Westra et al. model, expression data is normalized in the same way as in Decon eQTL. The effect of the cell type is predicted using a genotype * cell count interaction term:

$$y = I + \beta_1.G + \beta_2.c + \beta_3.c\,x\,G + e \tag{8}$$

where y is expression, I the intercept, G the genotype, c the cell count and c x G the cell count x genotype interaction term. Additional restrictions are set on the p-values. For neutrophils, if (the β of the neutrophil x G interaction term) * (the β of the G interaction term) < 0, the p-value is set to 1. For CD4+ and monocytes, if (the β of the neutrophil x G interaction term) * (the β of the G interaction term) > 0, the p-value is set to 1.

Comparison between allelic concordance

For the comparison between allelic concordances, we counted the concordant and discordant eQTLs for each of the cell type comparisons and did a Fisher exact test between each of the groups. The p-values are Bonferroni-corrected.

Single-cell eQTLs

The sc-eQTLs were obtained for 112 individuals in the same way as described in Van der Wijst et al. [24] For the allelic direction comparison, we used all significant eQTLs. Classical monocyte and non-classical monocyte eQTLs were combined and jointly compared to Decon-eQTL Monocytes.

Supplementary information

Supplemenatary material are provided at:

https://doi.org/10.1186/s12859-020-03576-5

Additional file 1 :

Suppl. Fig. 1. Prediction performance of Decon-cell within 500FG: The Y-axis represents the 73 immune cell types quantified by FACS in the 500FG cohort. The bar plot on the left panel shows the mean Prediction Performance (Spearman correlation coefficient between predicted and measured cells across 100-fold cross validations). On the right panel, box plots represent the distribution of the Prediction Performance within 100 iterations of the cross validations. A cutoff of mean Prediction Performance ≥0.5 was applied to define predictable cell types (green).

Suppl. Fig. 2. Signature genes selected for prediction of cell proportions by Decon-cell: (A) Total number of marker genes (genes selected in ≥80% of all models in the 100 iterations) per predictable cell type. Different colors indicate different subpopulations. (B) The number of genes significantly correlated with cell counts (Spearman correlation, adjusted p ≤ 0.05) (y-axis) shows the total number of significantly correlated genes, while the x-axis shows the prediction performance (x-axis). (C) Distributions of the total number of “strongly” correlated genes (absolute Spearman correlation ≥0.3) between predictable and unpredictable cell subpopulations.

Suppl. Fig. 3. Comparison of prediction performance between Decon-cell and other existing methods. (A) Performance of Decon-cell: the measured (x axis) and predicted cell proportions (y-axis) were compared for neutrophils (given by granulocytes in 500FG), lymphocytes and monocytes CD14+ and granulocytes in three independent cohorts (shown by row, from top to bottom: LLDeep (n = 627); LLS (n = 660); RS (n = 773)). (B) Comparison of prediction performance for Decon-cell, CIBERSORT and xCell in three independent cohorts for a total of 4 major immune subpopulations.

Suppl. Fig. 4. Prediction performance of xCell and CIBERSORT in three independent Dutch populations (LLDeep, n = 627; LLS, n = 660; RS, n = 773). (A) Scatter plots showing the measured cell proportions of circulating immune cells on the x-axis and the xCell enrichment score on the y-axis. (B) Scatter plots showing the measured cell proportions of circulating immune cells on the x-axis and the predicted cell proportions given by CIBERSORT) on the y-axis. **Suppl. Fig. 5**. Expression of marker genes selected by Decon-cell. Expression levels (scaled, log2(TPM + 1)) of signature genes in the data in three purified cell subpopulations: CD4+ T cells (A), neutrophils/granulocytes (B) and monocytes (C) in the data from BLUEPRINT. Cell subpopulations are indicated in different colors by columns. Correlation of each of the signature genes and the cell subpopulation percentage in the 500FG cohort is shown on by the green bar at the left-hand side of heatmap figure, i.e. darker green corresponds to higher correlations.

Suppl. Fig. 6. Many of the CTi eQTL are cell type exclusive. Colored bar plot on the left shows the total number of significant CTi eQTLs in whole blood eQTLs (as also shown in **Fig. 2A**). Gray bar plot shows the total number of eQTLs shared across the possible combinations of the six cell subpopulations under study. **Suppl. Fig. 7**. Variation of gene expression across samples for deconvoluted cell-type eQTLs genes in whole blood. Granulocyte eQTL genes show significantly higher variance across the BIOS samples (F test p-value ≤0.05) compared to those from monocytes, CD4+ T cells, CD8+ T cells, B cells and NK cells.

Suppl. Fig. 8. Validation of CTi eQTLs using effect sizes of eQTLs from purified cells. CTi eQTLs (FDR ≤ 0.05) from the BIOS cohort show a significantly bigger effect size in purified cell eQTLs 9 from their relevant cell subtype as compared to other whole blood eQTLs (diagonal boxed comparisons). The off-diagonal comparisons show that these eQTL genes are specific to a cell subpopulation because the differences in effect sizes are non-significant in all but one case (CD4+ T cell eQTL genes in monocyte-derived eQTLs). **Suppl. Fig. 9**. Validation of CTi eQTLs using effect sizes of K27AC QTLs from purified cells. CTi eQTLs (FDR ≤ 0.05) show a significantly bigger effect size for K27AC QTLs that have peaks located in the promoter region of the eGenes from their relevant cell subtype compared to the rest of the significant whole blood eQTLs (diagonal boxed comparisons). The off-diagonal comparisons show that these eQTL genes are specific to a cell subtype because the differences in effect sizes are non-significant in all but the comparisons across Neutrophils and Monocytes (CD14+).

Suppl. Fig. 10. Validation of CTi eQTLs using effect sizes of K4ME1 QTLs from purified cells. CTi eQTLs (FDR ≤ 0.05) show a significantly bigger effect size for K4ME1 QTLs (where the eGenes is the closest gene tagging the K4ME1 QTLs peak) from their relevant cell subtype compared to the rest of the significant whole blood eQTLs (diagonal boxed comparisons). The off-diagonal comparisons show that these eQTL genes are specific to a cell subtype because the differences in effect sizes are non-significant in all but the comparisons between neutrophils and monocytes (CD14+).

Suppl. Fig. 11. Validation of CTi eQTLs using allelic concordance with eQTLs results from purified cells. CTi eQTLs (FDR ≤ 0.05) show high allelic concordance with their respective purified cell eQTLs. Top row shows allelic concordance of deconvoluted granulocyte eQTLs (all in green) against neutrophils, monocytes and CD4+ T cells. Second row shows deconvoluted monocyte eQTLs against purified cell eQTLs in the same order as the top row. Bottom row shows the same comparisons as for deconvoluted CD4+ eQTLs. Allelic concordance of the off-diagonal (comparing CTi eQLTs with non-relevant cell types) show a consistent decrease in allelic concordance. p-values are Bonferroni-corrected Fisher exact tests between groups.

Suppl. Fig. 12. Validation of CTi eQTLs using allelic concordance with K27AC results from purified cells. CTi eQTLs (FDR ≤ 0.05) show a high allelic concordance in their respective purified cell K27AC QTLs. Top row shows allelic concordance of deconvoluted granulocyte eQTLs (all in green) against neutrophils, monocytes and CD4+ T cells derived from K27AC QTLs. Second row shows deconvoluted monocyte eQTLs (all in orange) against purified cell K27AC QTLs in the same order as top row. Bottom row shows the same comparisons as for deconvoluted CD4+ eQTLs (all in purple). Allelic concordance of the off-diagonal (comparing deconvoluted eQTLs with non-relevant cell types) show a consistent decrease in allelic concordance when compared to the relevant cell type comparisons. p-values are Bonferroni-corrected Fisher exact tests between groups. **Suppl. Fig. 13**. Allelic concordance between whole blood eQTLs and K27AC QTLs for purified neutrophils, CD14+ monocytes and CD4+ T cells.

Suppl. Fig. 14. Comparison of whole blood eQTLs with eQTLs from single cell RNA-seq Whole blood eQTLs show 89% allelic concordance for significant eQTLs derived from scRNA-seq data, comprising monocytes CD14+, B cells, CD4+ T cells, CD8+ T cells and NK cells. **Suppl. Fig. 15** Validation of cell type eQTLs detected in the BIOS cohort using the Westra et al. method: (A) Expression of eGenes in purified cell subpopulations from BLUEPRINT (green for granulocyte eQTL genes showing expression for purified neutrophils; orange for monocytes; purple for CD4+ T cells; pink for B cells). (B) CT eQTLs detected by the Westra method show a significantly larger effect size in purified cell eQTLs [11] as compared to the rest of the whole blood eQTLs. Boxed-diagonal shows the comparisons with relevant cell types where the effect differences are stronger.

Suppl. Fig. 16. Allelic concordance rates of cell type eQTLs detected using the Westra et al. method and eQTLs from purified cells. Top row shows allelic concordance of granulocyte CT eQTLs against neutrophils, monocytes and CD4+ T cells. Second row shows CT monocyte eQTLs against purified cell eQTLs in the same order as top row. Bottom row shows the same comparisons for CT CD4+ eQTLs. **Suppl. Fig. 17** Comparison of Decon-eQTL with Westra et al. method. Overlap of CT eQTLs detected with Decon-eQTL and the Westra et al. method and those found to be significant in purified cell subpopulations for granulocyte QTLs (A), CD4+ T cells (B), and monocytes (C).

Suppl. Fig. 18. Distribution and correlation among circulating cell proportions. Scatter plots show the correlations between different cell subpopulations in 89 samples from 500FG. Blue line indicates a fitted linear model. Diagonal plots depict the overall density distribution per cell type. Upper right triangle shows the Pearson correlation coefficient for each pairwise comparison. (B) Correlations between different cell subpopulations in the BIOS cohort obtained by prediction using Decon-cell. **Suppl. Fig. 19**.General overview of the Decon2 method. (A) Gene expression can be used to predict cell count percentages of cell counts that are already trained in the Decon-Cell model. Additionally, the model can be trained on different cell types if expression data and cell count proportions are available. (B) Decon-eQTL models the cell type dependent eQTL effect using expression, genotype, and measured cell count proportions or, if unavailable, predicted cell count proportions.

Additional file 2 :

Suppl. Tab. 1: Ensembl IDs and symbol names of the marker genes selected by Decon-cell for the 34 predictable circulating immune cell proportions.

Additional file 3 :

Suppl. Tab. 2: Summary statistics from Decon-eQTLs for the 16,362 whole blood eQTLs.

Abbreviations

eQTL: expression quantitative trait loci; CT: Cell type; CTi: Cell type interaction; GWAS: Genome-wide association studies; sc: single cell; GxE: Gene by environment interaction; TMM: Trimmed means of M

Acknowledgements

We thank K Mc Intyre and J Senior for editing the final text. We thank T. Spenkelink for the DeconCell web tool design. We thank the UMCG Genomics Coordination center, the UG Center for Information Technology and their sponsors BBMRI-NL & TarGet for storage and compute infrastructure, and the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

Authors’ contributions

C.W., L.F. and YL initialized the study.

Y.L. and L.F. directed and supervised the project.

Y.L. developed the statistical framework, together with L.F.. R.A-G, N.K., L.F., and Y.L., performed data analysis and interpretation.

J.D.T. was involved in the initial analysis.

N.K. and R.A-G. made the software and webtool.

C, U.V., M. Z, X.C., O.B.B., Z.B., I.R.P., P.D., C.J.X., M.S., I.J. 1, S.W., I.J. 2, S.S., V.K., H.J.P.M.K., L.A.B.J., M.G.N., M.W., D.V., H.B., R.O. and C.W. contributed to data collection, data analysis and interpretation.

R.A-G, N.K., L.F., and Y.L. draft and revise the manuscript.

All authors have read and approved the manuscript.

Funding

L.F. is supported by grants from the Dutch Research Council (ZonMW-VIDI 917.164.455 to M.S. and ZonMW-VIDI

917.14.374 to L.F.), and by an ERC Starting Grant, grant agreement 637640 (ImmRisk). Y.L. was supported by an ZonMW-OffRoad grant (91215206). The HFGP is supported by a European Research Council (ERC) Consolidator grant (ERC 310372). This study was further supported by an IN-CONTROL CVON grant (CVON2012–03) and a Netherlands Organization for Scientific Research (NWO) Spinoza prize (NWO SPI 94–212) to M.G.N.; an ERC advanced grant (FP/ 2007–2013/ERC grant 2012–322698) and an NWO Spinoza prize (NWO SPI 92–266) to C.W.; a European Union Seventh Framework Programme grant (EU FP7) TANDEM project (HEALTH-F3–2012-305279) to C.W. and V.K.;. A CONACYT-I2T2 scholarship (382117) to R.A-G. The Biobank-Based Integrative Omics Studies (BIOS) Consortium is funded by BBMRI-NL, a research infrastructure financed by the Dutch government (NWO 184.021.007). This work was supported by Radboud University Medical Centre Hypatia Tenure Track Grant (2018) to Y.L.

Availability of data and materials

The deconvolution summary statistics are made available as supplementary table. Information on how to request the genotype and RNAseq data used for the eQTL calculation can be found here: https://www.bbmri.nl/acquisition-use- analyze/bios. A subset of the single cell eQTLs is preliminary data for which a manuscript is in preparation, and will be made available after publication of that manuscript. Contact Lude Franke (l.h.franke@umcg.nl) to request access to this data. The GEO accession code for the expression data of 500FG is GSE134080.

Ethics approval and consent to participate

We have used existing and already published data only. Therefore, we did not get prior ethics approval or consent to participate.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106:9362–7.
2. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016;48:214–20.
3. Javierre BM, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016;167:1369–1384.e19.
4. Westra H-J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013;45:1238–43.
5. Joehanes R, et al. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. Genome Biol. 2017;18:16.
6. Raj T, et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science. 2014; 344:519–23.
7. Peters JE, et al. Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease. PLoS Genet. 2016;12:e1005908.

8. Naranbhai V, et al. Genomic modulators of gene expression in human neutrophils. Nat Commun. 2015;6:7545.
9. Chen L, et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell. 2016;167:1398– 1414.e24.
10. Westra H-J, et al. Cell Specific eQTL Analysis without Sorting Cells. PLoS Genet. 2015;11:e1005223.
11. Zhernakova DV, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2017;49:139–45.
12. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18:220.
13. Netea MG, et al. Understanding human immune function using the resources from the Human Functional Genomics Project. Nat Med. 2016;22:831–3.
14. Aguirre-Gamboa R, et al. Differential Effects of Environmental and Genetic Factors on T and B Cell Immune Traits. Cell Rep. 2016;17:2474–87.
15. Bakker OB, et al. Integration of multi-omics data and deep phenotyping enables prediction of cytokine responses. Nat Immunol. 2018;19:776–86.
16. Tigchelaar EF, et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open. 2015;5:e006772.
17. 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. 2014;23:4420–32.
18. Hofman A, et al. The Rotterdam Study: 2014 objectives and design update. Eur J Epidemiol. 2013;28:889–926.
19. van Greevenbroek MMJ, 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 Investig. 2011;41:372–9.
20. Schoenmaker M, et al. Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. Eur J Hum Genet. 2006;14:79–84.
21. Willemsen G, et al. The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. Twin Res Hum Genet. 2010;13:231–45.
22. Bonder MJ, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017; 49:131–8.
23. Adams D, et al. BLUEPRINT to decode the epigenetic signature written in blood. Nat Biotechnol. 2012;30:224–6.
24. van der Wijst MGP, et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat Genet. 2018;50:493–7.
25. Newman AM, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453–7.
26. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33:1–22.
27. Gruden K, et al. A ‘crossomics’ study analysing variability of different components in peripheral blood of healthy caucasoid individuals. PLoS One. 2012;7:e28761.
28. Davenport EE, et al. Discovering in vivo cytokine eQTL interactions from a lupus clinical trial; 2017. https://doi.org/10. 1101/118703.
29. Wilson DR, Sun W, Ibrahim JG. Mapping Tumor-Specific Expression QTLs In Impure Tumor Samples; 2017. https://doi. org/10.1101/136614.
30. Geeleher P, et al. Cancer expression quantitative trait loci (eQTLs) can be determined from heterogeneous tumor gene expression data by modeling variation in tumor purity. Genome Biol. 2018;19:130.
31. Glastonbury CA, Alves AC, Moustafa JE-S, Small KS. Cell-type heterogeneity in adipose tissue is associated with complex traits and reveals disease-relevant cell-specific eQTLs; 2018. https://doi.org/10.1101/283929.
32. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
33. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
34. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
35. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31:166–9.
36. van Dam RM, Boer JM, Feskens EJ, Seidell JC. Parental history of diabetes modifies the association between abdominal adiposity and hyperglycemia. Diabetes Care. 2001;24:1454–9.
37. Deelen P, et al. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. BMC Res Notes. 2014;7:901.
38. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5:e1000529.
39. Zhou X, Song Y, Wang L, Liu Q. Preconditioned GAOR methods for solving weighted linear least squares problems. J Comput Appl Math. (2009;224:242–9.
40. Lawson CL, Hanson RJ. Solving Least Squares Problem; 1995.

Chapter 5

Shared genetics of cytokine production upon diverse stimulations

O. B. Bakker [1,#], R. Aguirre-Gamboa [1,#], M. Jaeger [2], M. Oosting [2], S. P. Smeekens [2,3], R. T. Netea-Maier [4], R. J. Xavier [5,6], L. A.B. Joosten [2], E. Patin [7], L. Quintana-Murci [7], C. Wijmenga1[,8,*], M.G. Netea [2,9,*] and Y. Li [1,2*]

1 Department of Genetics, University of Groningen, University Medical Center Groningen, 9700 RB Groningen, the Netherlands
2 Department of Internal Medicine and Radboud Center for Infectious Diseases, Radboud University Medical Center, 6525 HP Nijmegen, the Netherlands
3 Department of Laboratory Medicine, Laboratory for Medical Immunology, Radboud University Medical Center, 6525 GA Nijmegen, the Netherlands
4 Department of Internal Medicine, Division of Endocrinology, Radboud University Medical Center, 6525 HP Nijmegen, the Netherlands
5 Broad Institute of MIT and Harvard University, Cambridge, MA 02142, USA
6 Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
7 Unit of Human Evolutionary Genetics, Institut Pasteur, CNRS UMR2000, 75015 Paris, France
8 Department of Immunology, University of Oslo, Oslo University Hospital, Rikshospitalet, 0372 Oslo, Norway
9 Department for Genomics & Immunoregulation, Life and Medical Sciences Institute (LIMES), University of Bonn, 53115 Bonn, Germany
# These authors contributed equally
* Corresponding author

Abstract

The human cytokine response shows a remarkable inter-individual heterogeneity in the general population. Previous studies have identified that genetic and non-genetic host factors play an important role in driving this variation. However, these studies focused mostly on single genotype-phenotype pairs and did not take the correlation structure and potential co-regulation between cytokine response phenotypes into consideration. In this study, we aimed to identify the genetic effects that are shared between cytokine response profiles across a wide range of stimuli. Using a multivariate QTL mapping approach, we were able to identify four previously unidentified loci regulating the *ex vivo* cytokine response to pathogens in a population of 500 healthy individuals. Among the identified cytokine QTLs where *TLR1* and *FCGR* loci which have been shown to be strong *trans* regulators of expression levels upon stimulation. We further show that the cytokine QTL signal in the *FCGR* locus co-localizes with inflammatory bowel disease (IBD) GWAS suggesting a shared mechanism between regulation of cytokine response and IBD. Altogether, we highlight the complex nature of the genetic regulation underlying cytokine responses by identifying shared genetic effects between cytokine response phenotypes. Future studies into immune signaling, in particular the cytokine response, should account for the complexity of immune responses in their study design and asses immune responses as a complex interacting network rather than a set of pairwise correlations.

Introduction

Over the last years, studies in systems and population immunology have made it increasingly clear that there is substantial inter-individual heterogeneity in the ways the immune system functions, at both baseline and stimulation conditions [1-5]. Such heterogeneity is apparent at both protein [1,2,6] and gene expression levels [3,5] and is likely to result from a complex interplay between genetics, intrinsic and environmental factors. One area where such heterogeneity is particularly apparent is the cytokine response to stimulation.

We previously evaluated whether cytokine response profiles showed significant inter-correlation, and thus have the potential to share (genetic) regulatory mechanisms [6]. We observed strong correlation patterns between cytokine response profiles that clustered based on the type of cellular assay that was used (PBMC, whole-blood, mo-derived macrophage), the broad response phenotype (innate after 24H; IL-6, TNF-α, IL-1β vs. Adaptive after seven days; IL-17, IL-22 and IFN-γ) and the broad stimulation class (fungal, bacterial, ligand-based). Such correlation clusters suggest that there is a potential sharing of regulatory mechanisms during these correlated cytokine responses. Recent efforts to map the extent of genetic control of immune response variation at the expression level have revealed several strong context-dependent genetic regulators of immune responses [5]. In addition, a plethora of genetic variants that regulate cytokine protein concentrations at baseline and after stimulation [1,3,6-8] have been identified. While a few studies have started to search for shared (pleiotropic) genetic effects on immune function [8], one aspect that has not been yet been assessed is the extent with which genetic regulators are simultaneously influencing multiple cytokine response phenotypes.

In this study, we hypothesize that the observed correlation between cytokine response profiles is partially due to shared genetic effects. To this end, we firstly defined groups of cytokine response phenotypes based on the type of cellular assay, the cytokine and the stimulation class following the patterns observed in previous clustering analysis [6]. We then identified shared associations between cytokine responses and genetic factors by using a multivariate approach. This allowed us to determine whether a set of cytokine responses are jointly regulated by a single genetic factor (mv-cQTL) and in doing so increase the power to detect such associations [9-13].

Results

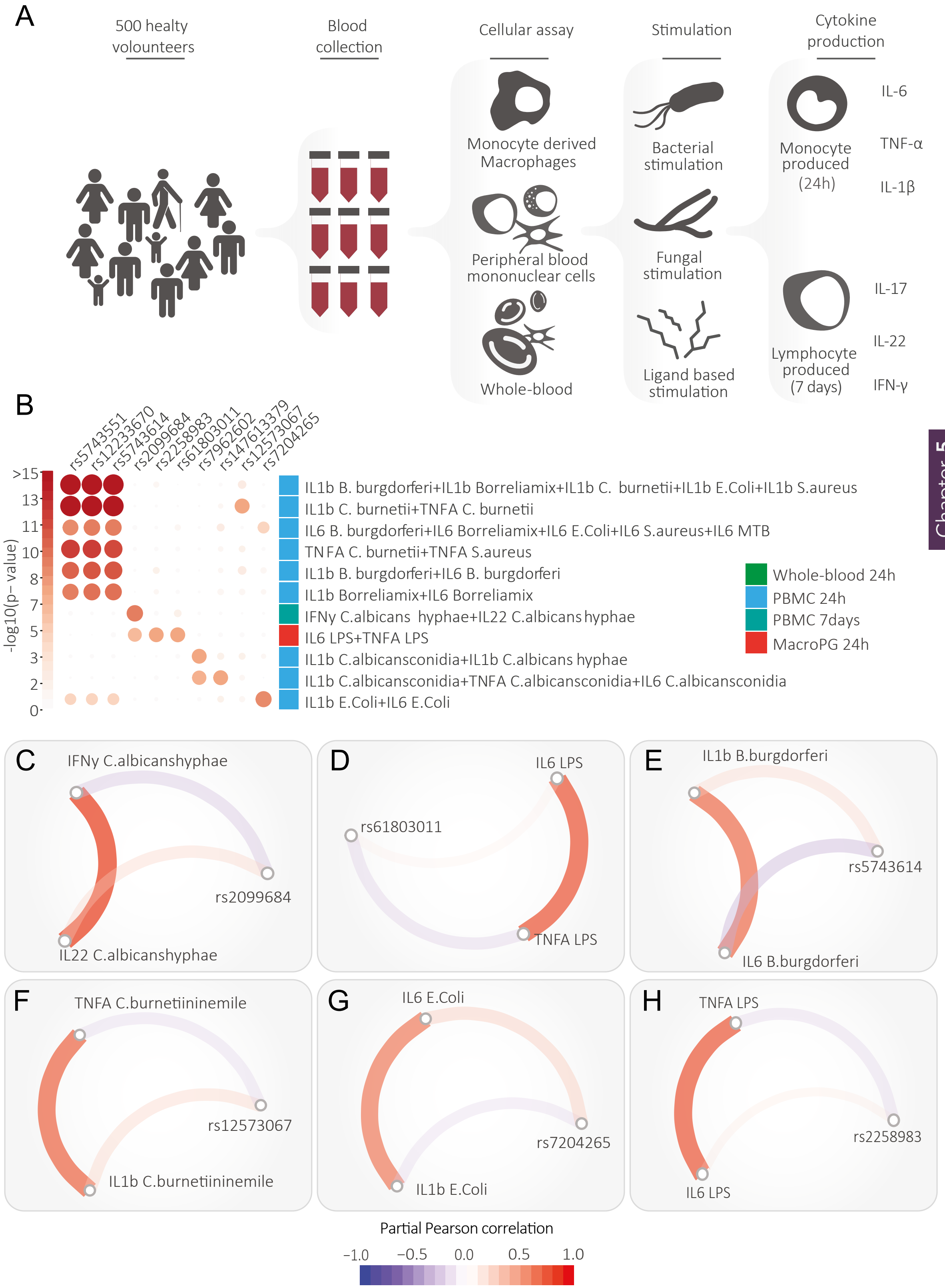
Multivariate genetic analysis reveals seven loci with distinct effects on cytokine production

The main objective of this study was to detect and characterize the shared genetic factors influencing cytokine production upon infectious or immune stimulation. To do so we made use of the database of cytokine responses to multiple stimulations from the 500FG cohort [1,6,14–16] together with the genetic profiles of the same individuals. We firstly defined combinations of relevant traits based on the type of cellular assay (PBMC, whole-blood and mo-derived macrophage), cytokines (monocyte; IL-6, TNF-α, IL-1β or lymphocyte; IL-17, IL-22 and IFN-γ produced) and class of stimulation (bacterial, fungal, ligand based) (**Fig. 1A**). Within these groups combinations of cytokine response phenotypes were then made in both a stimulation and a cytokine based grouping. The stimulation based grouping evaluated if one stimulation has shared genetic factors with downstream effects on multiple cytokines. The cytokine based grouping evaluated if multiple stimulations have shared genetic factors in influencing one cytokine. In total, we defined 26 distinct combinations of cytokine response phenotypes (**Suppl. Tab. 1**).

We then jointly associated each of the cytokine groups with each of the ~ four million available genetic variants genome-wide by using a multivariate mixed linear model (MVLMM) [9,10]. This analysis yielded a total of seven genome-wide significant (at $p<5x10^{-8}$) multivariate cytokine QTL (mv-cQTL), with 13 distinct top variant context pairs (**Fig. 1B**, **Table 1**). We then evaluated if the observed loci were unique to the multivariate condition, or if they were also present in single trait analysis [6]. We found that six (rs2099684, rs61803011, rs5743614, rs12573067, rs7204265 and rs2258983) out of the 13 genome wide significant mv-cQTL were unique to the multivariate analysis, and had not been identified in previous single trait analysis (**Fig 1C-H**, **Table 1**). We then evaluated if there was sharing between association signals for mv-cQTLs or if these associations were unique to their respective conditions. We found substantial overlap between association signals, at a suggestive threshold ($p<5x10^{-5}$), between mv-cQTL phenotype pairs (**Fig. 1B**) suggesting that mv-cQTL might have pleiotropic effects.

Next, we evaluated if the observed associations were likely to be either regulated distinctly (pleiotropy or undecided) or mediated through each other (**Fig. 1C-D**) using partial correlation [17] and mediation analysis [18]. In total, we found that eight of the 13 genome-wide associations showed distinct regulation (no mediation at FDR < 0.05) while five showed at least one mediation relationship (at FDR < 0.05) (**Suppl. Tab. 2**). For all six mv-cQTL that did not show genome-wide significance in single trait analysis (**Table 1**), the partial correlation analysis revealed distinct regulation (**Fig. 1C-D**, **Suppl. Tab. 2-3**). Additionally, we observed that for these six associations the genetic variant had an opposite direction of effect on the cytokine level, suggestive of it facilitating an

Right: Fig. 1. Multivariate QTL mapping strategy reveals 13 shared associations regulating cytokine response. A) Schematic overview of the defined cytokine groups. **B)** Heatmap of -log10 association p-values of the multivariate cytokine QTLs. -log10 p-values have been capped at 15 to ensure readability of the color scale. Colored boxes indicate the cellular assay the models was derived from. **C-H)** Partial Pearson correlation networks between the cytokines and genetic variants for the hits that were not identified in previous univariate analysis. Partial correlations indicate the correlation between two traits (genetic variant and cytokine) after correcting for the other traits (cytokines) in the network.


A
500 healty volounteers
Blood collection
Cellular assay
Stimulation
Cytokine production
Monocyte derived Macrophages
Peripheral blood mononuclear cells
Whole-blood
Bacterial stimulation
Fungal stimulation
Ligand based stimulation
Monocyte produced (24h)
Lymphocyte produced (7 days)
IL-6
TNF-α
IL-1β
IL-17
IL-22
IFN-γ
B
rs5743551
rs12233670
rs5743614
rs2099684
rs2258983
rs61803011
rs7962602
rs147613379
rs12573067
rs7204265
-log10(p- value)
>15
13
11
10
8
7
5
3
2
0
IL1b B. burgdorferi+IL1b Borreliamix+IL1b C. burnetii+IL1b E.Coli+IL1b S.aureus
IL1b C. burnetii+TNFA C. burnetii
IL6 B. burgdorferi+IL6 Borreliamix+IL6 E.Coli+IL6 S.aureus+IL6 MTB
TNFA C. burnetii+TNFA S.aureus
IL1b B. burgdorferi+IL6 B. burgdorferi
IL1b Borreliamix+IL6 Borreliamix
IFNy C.albicans hyphae+IL22 C.albicans hyphae
IL6 LPS+TNFA LPS
IL1b C.albicansconidia+IL1b C.albicans hyphae
IL1b C.albicansconidia+TNFA C.albicansconidia+IL6 C.albicansconidia
IL1b E.Coli+IL6 E.Coli
Whole-blood 24h
PBMC 24h
PBMC 7days
MacroPG 24h
C
IFNy C.albicanshyphae
rs2099684
IL22 C.albicanshyphae
D
IL6 LPS
rs61803011
TNFA LPS
E
IL1b B.burgdorferi
rs5743614
IL6 B.burgdorferi
F
TNFA C.burnetiininemile
rs12573067
IL1b C.burnetiininemile
G
IL6 E.Coli
rs7204265
IL1b E.Coli
H
TNFA LPS
rs2258983
IL6 LPS
Partial Pearson correlation
−1.0
−0.5
0.0
0.5
1.0

interaction between cytokine response phenotypes (**Fig. 1C-D**, **Suppl. Tab. 3**). This suggests that the genetic control of cytokine production upon stimulation is partially controlled by shared genetic factors having distinct effects on cytokine phenotypes.

The TLR1/10/6 gene cluster is a strong (trans) regulator of cytokine responses and co-expression after stimulation

One of the strongest mv-cQTL effects we observed was a shared genetic regulation on IL-1β, TNF-α and IL-6 concentrations, which was detected upon stimulation with *B. burgdorferi*, *E. coli*, MTB, *S. aureus* and *C. burnetii nine mile*. These concentrations

locus	single_trait	rsId	chr	pos	eff	alt	effAF	MAF	p_wald	beta	se	ntrait	N	model
1	**no**	rs2099684	1	161500130	A	G	0.63	0.37	4.48E-10			2	390	combined
									1.06E-03	-0.42	0.13		390	IFNy_C.albicanshyphae_PBMC_7days
									3.47E-01	0.12	0.12		390	IL22_C.albicanshyphae_PBMC_7days
	no	rs61803011	1	161594526	G	T	0.75	0.25	2.68E-08			2	402	combined
									1.00E-05	-0.21	0.10		402	IL6_LPS_macroPG_24h
									3.18E-07	-0.70	0.14		402	TNFA_LPS_macroPG_24h
2	**yes**	rs12233670	4	38787216	C	T	0.74	0.26	9.25E-26			5	380	combined
									8.77E-01	-0.01	0.10		380	IL1b_B.burgdorferi_PBMC_24h
									7.33E-06	-0.41	0.09		380	IL1b_Borreliamix_PBMC_24h
									8.88E-19	**-0.87**	**0.09**		**380**	**IL1b_C.burnetiininemileSerum_PBMC_24h**
									2.24E-05	-0.42	0.10		380	IL1b_E.Coli_PBMC_24h
									6.02E-01	0.03	0.06		380	IL1b_S.aureus_PBMC_24h
	yes	rs12233670	4	38787216	C	T	0.75	0.25	9.83E-11			2	336	combined
									7.33E-06	-0.41	0.09		380	IL1b_Borreliamix_PBMC_24h
									6.70E-11	**-0.68**	**0.10**		**336**	**IL6_Borreliamix_PBMC_24h**
	yes	rs12233670	4	38787216	C	T	0.74	0.26	4.62E-20			2	380	combined
									8.88E-19	**-0.87**	**0.09**		**380**	**IL1b_C.burnetiininemileSerum_PBMC_24h**
									1.52E-11	**-0.47**	**0.07**		**388**	**TNFA_C.burnetiininemileSerum_PBMC_24h**
	yes	rs12233670	4	38787216	C	T	0.75	0.25	3.31E-10			5	335	combined
									1.10E-06	-0.57	0.11		336	IL6_B.burgdorferi_PBMC_24h
									6.70E-11	**-0.68**	**0.10**		**336**	**IL6_Borreliamix_PBMC_24h**
									5.07E-06	-0.44	0.10		336	IL6_E.Coli_PBMC_24h
									1.44E-01	-0.12	0.08		336	IL6_S.aureus_PBMC_24h
									1.92E-04	-0.44	0.12		335	IL6_MTB_PBMC_24h
	no	rs5743614	4	38798935	C	T	0.75	0.25	5.27E-12			2	336	combined
									9.60E-01	0.00	0.10		380	IL1b_B.burgdorferi_PBMC_24h
									1.11E-06	-0.57	0.11		336	IL6_B.burgdorferi_PBMC_24h
	yes	rs5743551	4	38807654	T	C	0.74	0.26	5.62E-13			2	388	combined
									9.61E-12	**-0.48**	**0.07**		**388**	**TNFA_C.burnetiininemileSerum_PBMC_24h**
									2.02E-01	0.12	0.09		388	TNFA_S.aureus_PBMC_24h
3	**no**	rs12573067	10	101360754	G	A	0.77	0.23	2.18E-08			2	380	combined
									1.91E-03	0.33	0.11		380	IL1b_C.burnetiininemileSerum_PBMC_24h
									9.58E-02	-0.12	0.07		388	TNFA_C.burnetiininemileSerum_PBMC_24h
4	**yes**	rs7962602	12	10267996	T	A	0.84	0.16	3.12E-08			2	380	combined
									9.50E-09	**0.29**	**0.05**		**380**	**IL1b_C.albicansconidia_PBMC_24h**
									8.02E-02	0.21	0.12		380	IL1b_C.albicanshyphae_PBMC_24h
5	**yes**	rs147613379	12	133270573	T	G	0.68	0.32	2.92E-08			3	336	combined
									1.08E-02	0.12	0.05		380	IL1b_C.albicansconidia_PBMC_24h
									1.14E-01	0.14	0.09		388	TNFA_C.albicansconidia_PBMC_24h
									1.73E-08	**0.47**	**0.08**		**336**	**IL6_C.albicansconidia_PBMC_24h**
6	**no**	rs7204265	16	29270304	T	C	0.83	0.17	1.19E-09			2	336	combined
									6.78E-01	-0.05	0.12		380	IL1b_E.Coli_PBMC_24h
									9.62E-06	0.50	0.11		336	IL6_E.Coli_PBMC_24h
7	**no**	rs2258983	19	51630482	C	A	0.43	0.57	4.17E-08			2	402	combined
									1.20E-02	-0.19	0.08		402	IL6_LPS_macroPG_24h
									1.03E-07	-0.54	0.10		402	TNFA_LPS_macroPG_24h

Table 1. Summary statistics for the 13 genome wide significant distinct variant context pairs. Models were generated using GEMMA version 0.98. Exact sample sizes after removing NA values are reported for each model in the N column. Genomic coordinates have been standardized to b37.

were strongly associated to rs12233670 (chr4:38787216, cohort MAF=0.26) and its proxies. This variant is located in the *TLR1/TLR10/TLR6* region (**Fig. 2A**), which was also identified in previous single trait analysis at protein [6,7] and expression levels [5,19]. Toll like receptors (TLR) are one of the main family of pattern-recognition innate receptors that activate the innate responses to pathogen associated molecular patterns (PAMPs) [20,21]. This locus has been shown to have effects on individual cytokine response phenotypes, however, our analysis revealed that this locus jointly regulates predominantly bacteria induced (*S. aureus*, MTB, *E. coli*, *B. burgdorferi*, *C. burnetii*) cytokine responses (**Fig. 1B**), consistent with the role of TLRs in the innate immune response. To evaluate if these signals are likely to be distinct, we performed a mediation and partial correlation analysis and found that two out of six TLR-associated traits showed evidence of distinct regulation (no mediation at FDR < 0.05) (**Fig. 2A-B**, **Suppl. Tab. 2**). The remaining four TLR traits did show evidence of mediation (at FDR < 0.05) suggesting that the correlation between these traits is a result of a shared pathway rather than distinct effects of the genetic marker (**Fig. 2C-F**).

To validate these findings, we utilized summary statistics on stimulation-specific expression QTL (eQTL) from healthy individuals from the Milieu Intérieur consortium [3,5]. We observed that the strongest master regulator of gene expression, in *trans*, in response to *E. coli* (97 genes), BCG (80 genes), *S. aureus* (7 genes) and SEB (13 genes) stimulation, was the *TLR1/TLR10/TLR6* locus, the same we observed to be a strong regulator of cytokine response phenotypes. In the *cis*-eQTL analysis, the only gene that was associated with rs12233670 was *TLR1* (β=0.095 p=1.20×10^{-9}, beta standardized to C allele) at the baseline level and upon *E. coli* stimulation (β=0.35 p=2.23×10^{-48}, beta standardized to C allele). This suggests that the downstream immune response to *E. coli* is regulated by rs12233670 through TLR1.

To obtain more insights into the mechanism through which this locus influences cytokine production after stimulation, we reconstructed co-expression networks using the gene expression data on 560 immune genes from Milieu Intérieur Consortium's dataset [5]. We selected all *cis* and *trans* eQTL genes for rs12233670 after *E. coli* stimulation and used them to construct a co-expression network at baseline and after *E. coli* (**Fig. 3A**). We observed widespread positive correlation relationships at baseline (**Fig. 3A**). After stimulation with *E. coli* we observed a striking increase in negative gene-gene correlations (**Fig. 3A**).

Enrichment analysis on the *cis*- and *trans*-eQTLs revealed no significant enrichment after correction for multiple testing, as expected, given that the gene panel analyzed is already enriched for immune responses (560 immune related genes). However, out of the 98 *cis* and *trans* eQTL genes regulated by rs12233670 the most significantly enriched pathways included "Interleukin-10 signalling" (15 / 37 genes, p=7.52×10^{-4}) "diseases associated with the TLR signaling cascade" (9 / 21 genes, p=6.38×10^{-3}) "signalling by interleukins" (38 / 153 genes, p=7.56×10^{-3}) and "MyD88 deficiency (TLR2/4)" (5 / 9 genes, p=1.21×10^{-2}), consistent with the hypothesis that rs12233670 is disrupting the TLR mediated immune response network.

Among the negative associations identified in the co-expression analysis (**Fig. 3A**) were *TLR1* ~ *IL6* (r=-0.54, p=4.20×10^{-62}) and *TLR1* ~ *IL1B* (r=-0.31, p=3.08×10^{-19}), an effect consistent with the reduction in IL-6 and IL-1β released for carriers of the C/T or C/C genotype for rs12233670 in our data and the reported *cis* eQTL effect in the

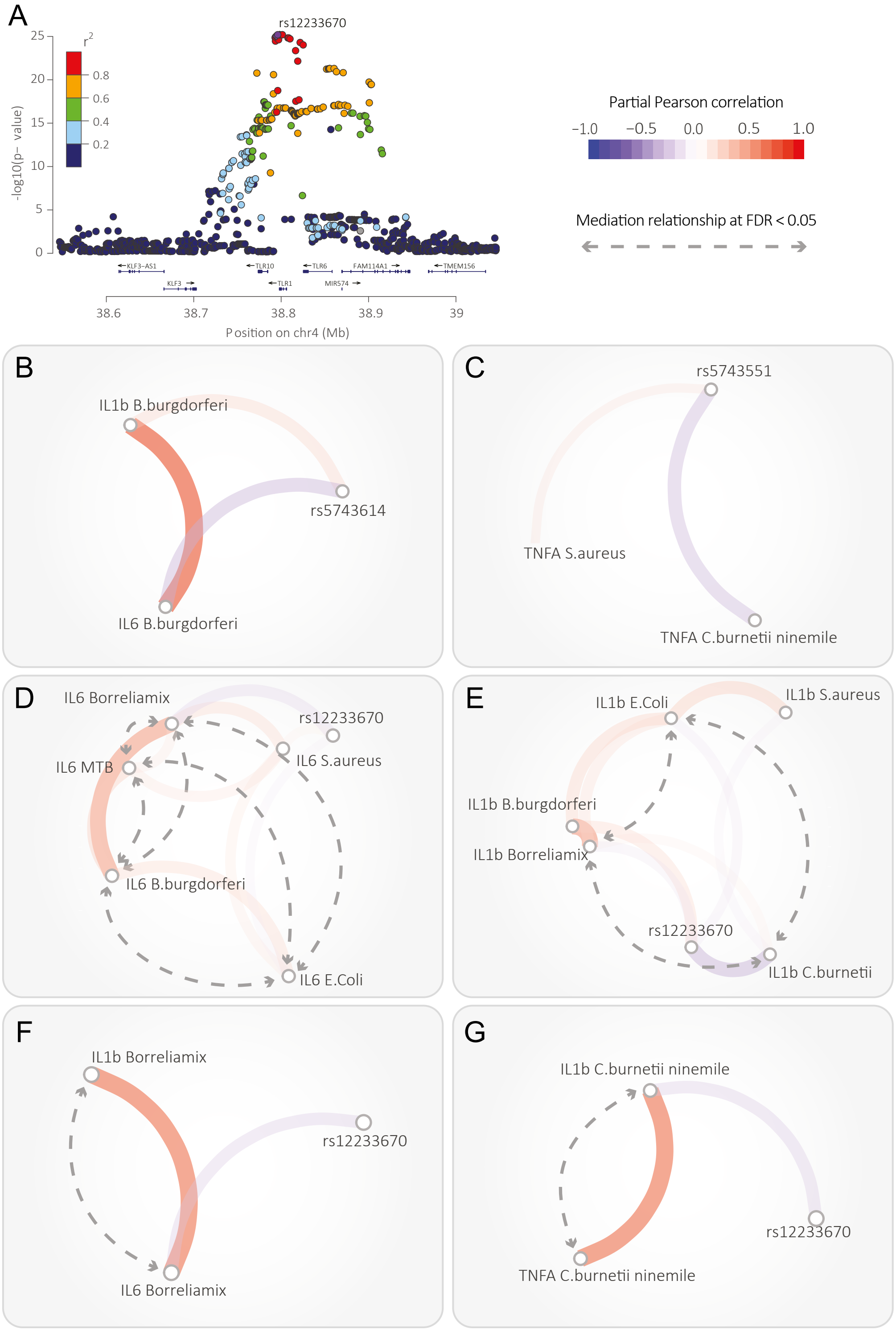


Fig. 2. Associations in TLR1 locus show both distinct and mediation effects. A) Locuszoom plot showing the most significant mv-cQTL effect in the *TLR1/TLR6/TLR10* locus identified as a joint regaulation on bacterial induced IL-1β levels. Y-axis indicates the -log10 p-value of the association. X-axis indicates the genomic position on chromosome 4. **B-G)** Partial Pearson correlations and mediation relationships between the identified top mv-cQTL effects and the cytokine response phenotypes for the corresponding model. Partial correlations indicate the Pearson correlation between two traits after adjusting for all the other traits in the network. Mediation relationships have been identified based on the Causal Inference Test as described by Schadt et al. [18]. Exact statistics for the mediation analysis and partial correlations are reported in **Suppl. Tab. 2 and 3**.

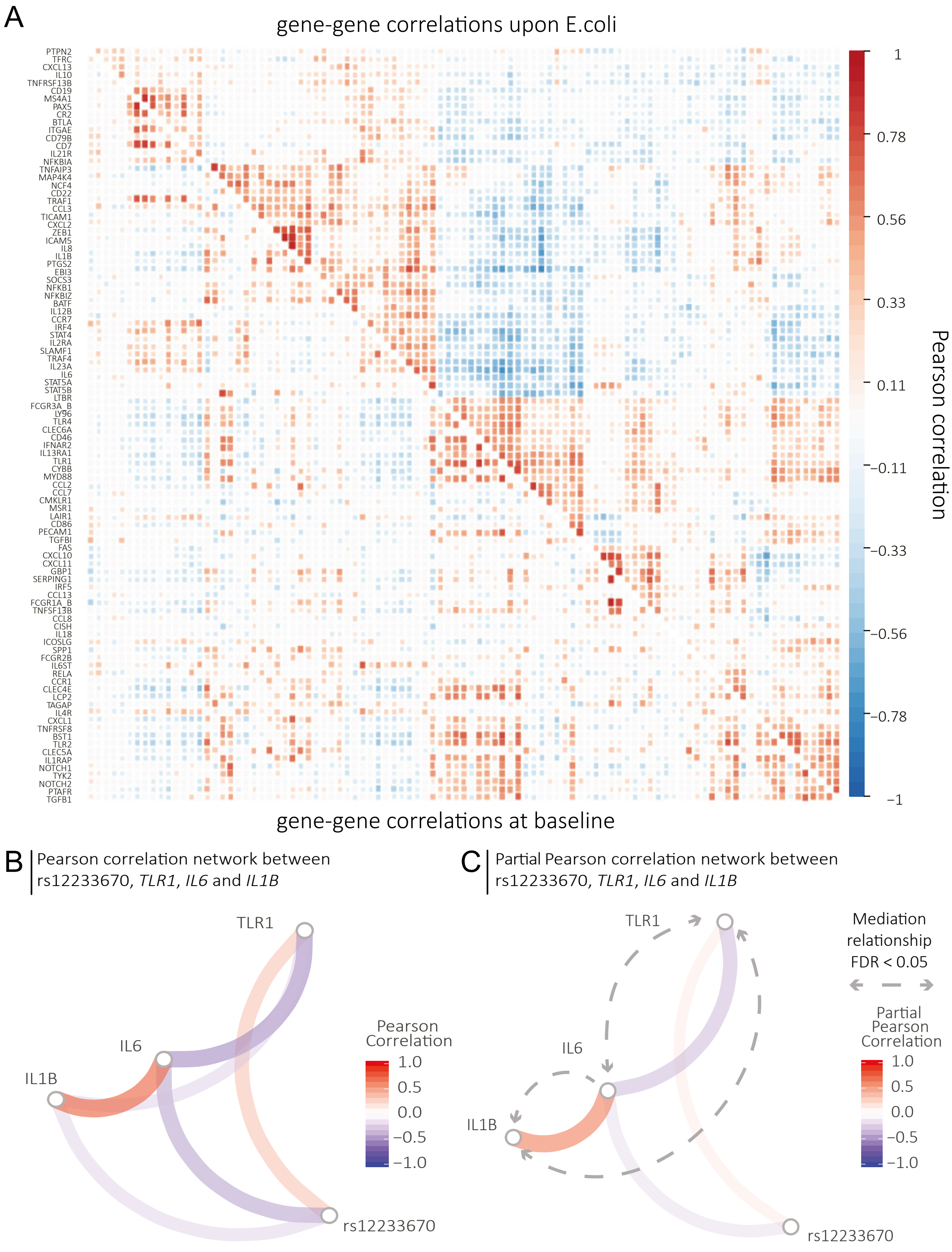


Fig. 3. Associations in TLR1 locus are mirrored at the expression level. A) Pearson correlations between all *cis* and *trans* eQTL genes of rs12233670 in the Milieu Intérieur dataset (804 individuals) at baseline (lower triangle) and upon *E. coli* stimulation (upper triangle). **B)** Pearson correlation network of the mv-cQTL SNP rs12233670 and *TLR1*, *IL6* and *IL1B* gene expression upon *E. coli* stimulation. Strong associations are shown between the SNP and *TLR1*, *IL6* and *IL1B* expression levels. **C)** Partial Pearson correlation network showing the associations between the SNP and *TLR1*, *IL6* and *IL1B* expression upon *E. coli* stimulation after correcting for the other traits in the network. Grey arrows indicate a mediation relationship at FDR < 0.05 based on the Causal Inference Test as described by Schadt et al. [18].

Milieu Intérieur data (**Fig. 3B**). Next we evaluated if there was any evidence that the association between rs12233670 and IL-6 and IL-1β was mediated through TLR1 (**Fig. 3C**, **Suppl. Tab. 4**). We observed significant (FDR=3.50×10^{-4}) mediation relationships between *TLR1*, *IL6* and *IL1B* suggesting that there is a mediation relationship between these traits. Given the role of TLRs in the innate immune response, and the fact that *TLR1* was the only observed *cis*-eQTL gene upon stimulation it might be inferred that *IL6* and *IL1B* expression are regulated by rs12233670 through *TLR1* expression levels.

Together, these results suggest the SNP rs12233670 as a strong regulator of *TLR1* expression upon bacterial stimulation, which has downstream effects on both TLR1-mediated signaling and cytokine responses at a protein level.

The FCGR locus is a strong trans-regulator of IFN-γ and IL-22 responses

The next most significant multi-trait unique effect we observed was located in the *FCGR* locus. This locus is characterized for encoding Fc gamma receptors that bind the fragment crystallizable (Fc) region of IgG. We observed a genome wide significant association between the variant rs2099684 (chr1:161500130, cohort MAF=0.37) and IFN-γ and IL-22 concentrations after stimulation with *C. albicans* hyphae (p=4.48×10^{-10}) (**Fig 4A**). The partial correlation analysis revealed strong associations for both IFN-γ (r=-0.32, p=2.07×10^{-10}) and IL-22 (r=0.23, p=2.80×10^{-8}) with rs2099684 (effects standardized to A allele) and no significant mediation, suggesting that rs2099684 might be a pleiotropic regulator (**Fig. 1C**, **Suppl. Tab. 2**).

We looked into the eQTL effects for this locus and found that the SNP rs2099684 was a *trans* regulator for 5 genes (*GBP1, GBP5, IRF1, STAT1, FCGR1A_B*) at both baseline and stimulated conditions. IRF1 and STAT1 are both TFs that play key roles in type-2 interferon signaling [22,23] and *GBP1* has been shown to be induced by IFN-γ [24]. While there was a reported increase in effect size after stimulation, the *trans* eQTLs of rs2099684 were not unique to the stimulation condition [5]. We then applied an overrepresentation analysis on these 5 genes. The only significant enrichment was found to be the Reactome pathways Interferon Signaling (OR: 3.7, FDR=0.0058) and Interferon Gamma Signaling (OR: 4.3, FDR=0.0022), which fits with the observed effect of rs2099684 on IFN-γ concentrations and known roles of these genes in type-2 interferon signaling. This suggests that rs2099684 is disrupting the IFN-γ signaling leading to a differential in cytokine response for both IFN-γ and IL-22.

Given the key role played by the immune system in a wide range of diseases, we investigated cQTLs as a mechanism underlying genetic associations to complex diseases identified by genome-wide association studies (GWAS). To this end, we identified individual GWAS loci that are likely to share a causal variant with a cQTL in the same locus using the COLOC method [25]. In total we searched for colocalization with the cQTL and 11 immune diseases (**Suppl. Tab. 5**). We observed one significant colocalization in the *FCGR* locus between our cytokine signal for *C. albicans* hyphae induced IFN-γ and IL-22 and inflammatory bowel disease (IBD) (**Fig. 4B**) with the probability of the traits sharing a causal variant exceeding 0.95. Although inferring causal direction between these traits remains challenging, one could interpret that rs2099684 is involved in mediating IFN-γ and IL-22 concentrations after stimulation. Therefore, the de-regulation created by rs2099684 in the production of these pro-inflammatory cytokines could potentially impact the risk of developing IBD.

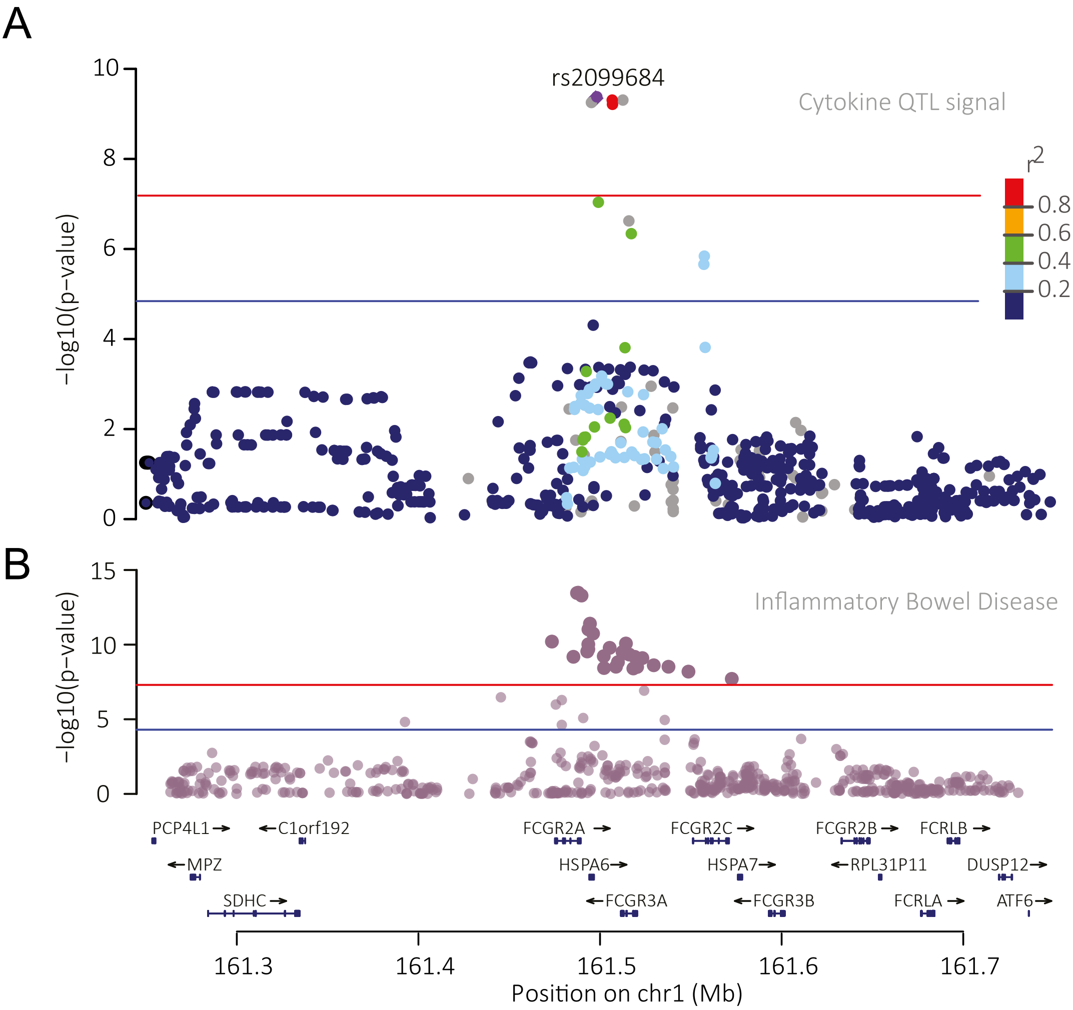


Fig. 4. Associations in FCGR locus. A) Locuszoom plot showing the second most significant mv-cQTL effect. This effect is located in the FCGR locus and identifies a joint regulation on IL-22 and IFN-γ levels in response to *C.albicans* hyphae stimulation. Y-axis indicates the -log10 p-value of the association. X-axis indicates the genomic position on chromosome 1. **B)** Locuszoom plot showing co-localization in the *FCGR* locus between the mv-CQTL effect (top plot) and Inflammatory Bowel Disease GWAS (lower plots). Y-axis indicates the -log10 p-value of the association for the respective GWAS. X-axis indicates the position of chromosome 1.

Discussion

Immune response networks are highly complex and interwoven, here we show that the genetic regulation underlying cytokine responses shows a similar pattern. Using multivariate analysis, we show that accounting for the effect of another response phenotype can help to reveal shared genetic regulation between immune responses. These associations can be missed in single trait pairwise analysis if they are hidden due to there being a (pleiotropic) regulator with opposite directions of effect or due to power [9,11-13]. We observed this pattern for the six hits that could not be observed in single trait analysis.

Here we show that the genetic regulation underlying the *ex vivo* cytokine response to pathogens shows signs of pleiotropy. Others have recently shown a similar pattern for baseline cytokine concentrations circulating in blood [8]. In this study it is suggested that pleiotropic cytokine loci are important for the regulation of hematopoietic and metabolic functions and that they have a particular relevance to cardiovascular disease. Although our overarching conclusion is shared, we do not directly observe an impact on cardiovascular disease. This might be explained by the fact that we study a different array of cytokines. In addition, we are studying the *ex vivo* response to pathogens, rather than the baseline circulating concentrations in serum, a radically different phenotype with likely different regulation. A large difference in sample size (N=~400, vs N=9263) might also limit our power to detect the effects important for metabolism and hematopoietic function. Increasing the power to detect genetic variation underlying cytokine responses might yield many more genetic associations of smaller effect size. Additionally, performing *ex vivo* stimulation experiments in disease specific cohorts might give insights into the nature of response networks in these diseases, rather than evaluating such effects in the general population.

One striking observation is the lack of *cis*-regulation on cytokine responses at a protein level. It has been shown that at the expression level such effects do exist both at baseline [26,27] and after stimulation [5]. The absence of *cis* effects might be due to the fact that the *cis*-mediated response of the six cytokines assessed (IL-6, IL-1β, TNF-α, IL-17, IL-22, IFN-γ) is very early in the stimulation. Taking this into account, it still seems that the *cis* regulation is not important for the total release of cytokines at a protein level and that it might be more important to control the speed of the response. Again our power is relatively limited (~400 individuals) and increased power might help to reveal such effects. Alternatively, the impact of genetic variants on these cytokine genes might be buffered by evolution. Indeed, dosage sensitive genes are known to be depleted for *cis*-eQTL signals [28]. Since we were able to identify strong *trans* regulators of cytokine responses it seems *trans* effects are more important in regulating the total release of cytokines by impacting the pathways controlling the release, rather than *cis*-eQTL effects controlling the expression of the cytokine genes directly.

It should be noted that the associations we identify do not prove a causal mechanism. To truly identify which genes and variants are really (pleiotropically) regulating the cytokine response causally a carefully considered experimental setup would be required. We did evaluate if there were mediation relationships between the different cytokines, however, absence of mediation does not guarantee that the resulting associations are then due to pleiotropy. The presented work does present a credible set of regulatory pathways, such as TLR1 or FCGR mediated signaling, with effects which were independently validated at an expression level. To prove the causality of these types of effects we need to move towards reconstruction directional cytokine signaling networks by performing knockdowns of target genes on a genetic level. However, performing genetic manipulation studies in mixtures of primary cells that are relevant for *in vivo* situations remains a major hurdle.

In conclusion, the work presented represents a small step towards accounting for the complexity in the regulation of immune responses and shows that by jointly modeling related *ex vivo* cytokine responses shared genetic relationships can be identified that are missed with single trait approaches. Future studies into immune signaling, in particular the cytokine response, should account for the complexity of immune responses in their study design and asses immune responses as a complex interacting network rather than a set of pairwise correlations.

Methods

Study cohort

The main analyses were performed in the 500FG cohort, which is part of the Human Functional Genomics Project and has been previously described [6]. This cohort consists of 534 healthy individuals (237 males and 296 females) of Caucasian origin. Volunteers ranged from 18 to 75 years of age, and most (421 individuals) were 30 years or younger. The volunteers had BMI was within normal limits (15 to 35), and most (380 individuals) had a BMI between 20 and 25. Of these 534 original volunteers, 45 were excluded because of genetic background and questionnaire results (medication usage and chronic disease), thus leaving 489 individuals.

Cytokine stimulation experiments

Cytokine stimulation experiments have been previously described [6]. In short, PBMCs were extracted from blood in ~500 volunteers and subsequently stimulated using 7 pathogens (*Candida albicans* conidia*, Candida albicans* hyphae*, Staphylococcus Aureus, Mycobacterium Tuberculosis, Estridia Coli, Borrelia Burgdorferi, Coxiella Burnetii*) and 3 ligand based simulations (lipopolysaccharide (LPS), Pam3Cys and Phytohemagglutinin (PHA)) after which the response of three monocyte produced cytokines (IL-6, IL-1β and TNF-α) was measured by enzyme linked immunosorbent assays (ELISA) 24H after stimulation. The response of the three lymphocyte produced cytokines (IFN-γ, IL-17 and IL-22) was quantified using ELISA after 7 days.

Genotyping and quality control

Exact genotyping procedures have been described previously [6]. In short, Genotyping of individuals was performed using a SNP chip (Illumina Human Omniexpress Exome-8). Opticall 0.70 [29] was used to perform genotype calling with samples with a call rate smaller than 0.99 being removed. This call set was subsequently imputed using IMPUTE2 [30] using GoNL [31] as a reference panel. After imputation, variants with Hardy-Weinberg equilibrium (HWE) $< 5x10^{-7}$, call rate < 0.99, MACH $R^2 < 0.3$ and a minor allele frequency (MAF) < 0.1 as well as multi-allelic sites were excluded yielding a dataset containing 4,358,038 SNPs and 489 individuals. This resulting set was then pruned using a 1Mb window and 0.2 LD threshold using the PLINK 1.9 option --indep-pairwise after having the HLA region (chr6:25- 36Mb) removed. Related individuals where then identified using the option --genome and any relationships with a pi hat score of >0.1 removed. Genetic outliers were also filtered on the same set. This yielded a dataset of 441 individuals used for downstream analysis.

Statistical Methods

Data pre-treatment

Raw cytokine data (ELISA) was log2 normalized prior to performing the QTL mapping. NA values were removed on a case by case basis to maximize the available samples for each trait. Exact sample sizes are reported in **Suppl. Tab. 1**. Only traits with a limited number ties were included in the final analysis. See Supplemental **Fig. 2** for histograms. Cytokine data was then adjusted for several main cell types (lymphocytes, monocytes, Tc, Bc, NK in PBMC and additionally granulocytes cells for WB) as well as age and gender using a linear model. The residuals plus the mean of the original data where taken and used for mapping and partial correlation analysis.

Cytokine correlations and co-expression networks

Correlations between cytokines were generated in R version 3.4.4 using ‘cor’ and ‘cor.test’ from the stats package. Correlations were calculated on the log2 normalized cytokine counts using Pearson’s correlation. The cytokine correlation heatmap was visualized using the R package ‘corrplot’ version 0.84.

Co-expression in the Milieu Interieur dataset was calculated on the normalized expression data as reported in the paper [5] using the Pearson’s correlation implemented in ‘cor.test’. Networks were visualized using the function ‘network_plot’ from the R package ‘corr’ version 0.3.0.9000 and ‘corrplot’

Univariate mixed linear model

To provide an appropriate comparison with the multi trait modeling, we first mapping single trait cytokine QTLs using genome wide efficient mixed model analysis (GEMMA) version 0.98 [9,10] under the univariate model using Wald test (-lm 1) and a kinship matrix generated on the pruned non-HLA 500FG data using GEMMA’s option -gk 1.

Multivariate mixed linear model

QTL mapping was facilitated by GEMMA version 0.98 9,10 under the multivariate linear mixed model using Wald test (-lmm 1) and a kinship matrix generated on the pruned non-HLA 500FG data using GEMMA’s option -gk 1. Combinations of traits to jointly model were defined based on the tissue, the broad function of the cytokine, the stimulation type or the cytokine. Exact combinations used are reported in **Suppl. Tab. 1**.

Partial correlations

Partial correlation were calculated on the same dataset as used for mapping using the R version 3.4.4 and the package ‘ppcor’ version 1.1 [17] And the function ‘pcor’. Input was a matrix containing the cytokines for the relevant model and the genetic marker of interest. Correlations and the corresponding t-statistics were calculated using Pearson’s correlation. Exact correlations and statistics are reported in **Suppl. Tab. 3**. Partial correlation networks were visualized using the function ‘network_plot’ from the R package ‘corrr’.

Mediation analysis

Mediation FDR and p-values values were calculated using Causal Inference Test ‘CIT’ as described by Schadt et al. [18] and implemented in the R package ‘cit’ version 2.2 . FDR was calculated based on 1000 random permutations of the data using the function ‘fdr.cit’. Exact FDR values and their confidence intervals for the CIT test are reported in **Suppl. Tab. 2** and **4**.

Overrepresentation analysis

Gene set enrichments were performed using http://webgestalt.org/ on Reactome pathways using an overrepresentation analysis [32].The full set of 562 genes assessed in the Milieu Interieur study was used as a reference set for enrichments to avoid bias introduced by the platform used for quantifying expression.

Colocalization analysis

Colocalization analysis was facilitated by the R package 'coloc' version 2.3-1 using the function 'coloc.abf' [25]. Full association summary statistics used for co-localization were downloaded from several publicly available resources indicated in **Suppl. Tab. 5** [33-43]. The wrapper used is available at https://bitbucket.org/immunogengroup/gwas-colocolization. Colocalization was considered if the PPH4 was > 0.95.

Data availability

The data that support the findings of this study are available at https://hfgp.bbmri.nl/; the data have been meticulously catalogued and archived at BBMRI-NL, aiming for maximum reuse, by following the findability, accessibility, interoperability and reusability (FAIR) principles. Individual-level genetic data as well as other privacy sensitive datasets are available upon request at http://www.humanfunctionalgenomics.org/site/?page_id=16/. These datasets are not publicly available because they contain information that could compromise the research participants' privacy. The central data stewardship and access have been implemented with the MOLGENIS open-source platform for scientific data, which enables flexible data upload, management and querying, including sufficiently rich metadata and interfaces for machine processing and custom (R statistics) visualization for human processing (http://molgenis.org/).

Acknowledgements

We thank all of the volunteers in the 500FG and Milieu Interieur cohorts for their participation. We thank the EAGLE eczema consortium for making their summary statistics publicly available. The HFGP is supported by a European Research Council (ERC) Consolidator grant (ERC 310372). This study was further supported by an IN-CONTROL CVON grant (CVON2012-03) and a Netherlands Organization for Scientific Research (NWO) Spinoza prize (NWO SPI 94-212) to M.G.N.; an ERC advanced grant (FP/2007-2013/ERC grant 2012-322698) and an NWO Spinoza prize (NWO SPI 92-266) to C.W.; a European Union Seventh Framework Programme grant (EU FP7) TANDEM project (HEALTH-F3-2012-305279) to C.W. and V.K.; and an NWO VENI grant (NWO 863.13.011) and an ZonMw-OffRoad-91215206 grant to Y.L. M.O. was supported by an NWO VENI grant (016.176.006). R.J.X. was supported by National Institutes of Health (NIH) grants DK43351, AT009708 and AI137325. MGN was supported by a Spinoza grant of the Netherlands Organization for Scientific Research.

Author contributions

Y.L., C.W. and M.G.N. designed the study. M.O., S.P.S., M.J., R.T.N.-M., R.J.X. and L.A.B.J. performed the experiments and processed the 500FG data. E.P. and L.Q-M. provided assistance with the Milieu Intérieur dataset. O.B.B. performed statistical analysis with assistance from R.A.-G.; O.B.B., R.A.G., Y.L., M.G.N., interpreted the data the and wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethics statement

The HFGP study was approved by the ethical committee of Radboud University Nijmegen (no. 42561.091.12). Experiments were conducted according to the principles expressed in the Declaration of Helsinki. Samples of venous blood were drawn after informed consent was obtained.

Supplementary material

Supplementary material has been provided to the the University of Groningen Library which hosts the digital version of this thesis.

Suppl. Fig. 1. Cytokine-cytokine correlations show distinct clusters

Pearson correlations (upper triangle) between cytokine stimulation pairs show clustering based on the type of cellular assay, the stimulation class and broad cytokine response type. Lower triangle indicates the -log10 p-value of the association. p-values have been limited to 15 to keep the color scale visible. Any associations not passing bonferroni significance have been removed in the lower triangle.

Suppl. Fig. 2. Histograms of cytokine stimulation pairs included in the study

Y-axis indicates the counts in each bin. Y-axis indicates the log2 normalized cytokine expression after stimulation. Log2 transformed cytokine expression has not been adjusted for age, gender and cell type composition.

Suppl. Tab. 1. Overview of all models evaluated in this study. The model column indicates the combination of cytokine stimulation pairs assessed. Exact sample sizes for each model are reported in the SampleSize column.

Suppl. Tab. 2. Test statistics of the casual mediation test in the cytokine data. Causal mediation test was performed using the R package 'cit'. A pairing in a model was considered subject to a mediation effect if the FDR was < 0.05. Exact sample sizes for each test are reported in the n column. The column 'q.cit' indicates the permutation based FDR estimate. 1000 permutations were performed.

Suppl. Tab. 3. Test statistics of the partial correlation analysis. Partial correlation was calculated using the R package 'ppcor'. Exact sample sizes for each test are reported in the n column.

Suppl. Tab. 4. Test statistics of the casual mediation test in the stimulated gene expression data in the TLR locus. Causal mediation test was performed using the R package 'cit'. A pairing in a model was considered subject to a mediation effect if the FDR was < 0.05. Exact sample sizes for each test are reported in the n column. The column 'q.cit' indicates the permutation based FDR estimate. 1000 permutations were performed.

Suppl. Tab. 5. Overview of the summary statistics used for the colocalization test.

References

1. Li, Y. et al. Inter-individual variability and genetic influences on cytokine responses to bacteria and fungi. Nat. Med. 22, 952–960 (2016).
2. de Craen, A. J. M. et al. Heritability estimates of innate immunity: an extended twin study. Genes Immun. 6, 167–170 (2005).
3. Thomas, S. et al. The Milieu Intérieur study — An integrative approach for study of human immunological variance. Clin. Immunol. 157, 277–293 (2015).
4. Brodin, P. & Davis, M. M. Human immune system variation. Nat. Rev. Immunol. 17, 21–29 (2016).
5. Piasecka, B. et al. Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. Proc. Natl. Acad. Sci. 201714765 (2017) doi:10.1073/pnas.1714765115.
6. Li, Y. et al. A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans. Cell 167, 1099-1110.e14 (2016).
7. Mikacenic, C., Reiner, A. P., Holden, T. D., Nickerson, D. A. & Wurfel, M. M. Variation in the TLR10/TLR1/TLR6 locus is the major genetic determinant of interindividual difference in TLR1/2-mediated responses. Genes Immun. 14, 52–57 (2013).
8. Nath, A. P. et al. Multivariate genome-wide association analysis of a cytokine network reveals variants with widespread immune, haematological and cardiometabolic pleiotropy. bioRxiv 544445 (2019) doi:10.1101/544445.
9. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat. Methods 11, 407–409 (2014).
10. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44, 821–824 (2012).
11. Ferreira, M. A. R. & Purcell, S. M. A multivariate test of association. Bioinformatics 25, 132–133 (2009).
12. Kim, S. & Xing, E. P. Statistical estimation of correlated genome associations to a quantitative trait network. PLoS Genet. 5, e1000587 (2009).
13. O'Reilly, P. F. et al. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PloS One 7, e34861 (2012).
14. ter Horst, R. et al. Host and Environmental Factors Influencing Individual Human Cytokine Responses. Cell 167, 1111-1124.e13 (2016).
15. Schirmer, M. et al. Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. Cell 167, 1125-1136.e8 (2016).
16. Aguirre-Gamboa, R. et al. Differential Effects of Environmental and Genetic Factors on T and B Cell Immune Traits. Cell Rep. 17, 2474–2487 (2016).
17. Kim, S. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. Commun. Stat. Appl. Methods 22, 665–674 (2015).
18. Schadt, E. E. et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nat. Genet. 37, 710–717 (2005).
19. Quach, H. et al. Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. Cell 167, 643-656.e17 (2016).
20. Kawai, T. & Akira, S. Toll-like Receptors and Their Crosstalk with Other Innate Receptors in Infection and Immunity. Immunity 34, 637–650 (2011).
21. Netea, M. G., Wijmenga, C. & O'Neill, L. A. J. Genetic variation in Toll-like receptors and disease susceptibility. Nat. Immunol. 13, 535–542 (2012).
22. Platanias, L. C. Mechanisms of type-I- and type-II-interferon-mediated signalling. Nat. Rev. Immunol. 5, 375–386 (2005).
23. Ramana, C. V., Chatterjee-Kishore, M., Nguyen, H. & Stark, G. R. Complex roles of Stat1 in regulating gene expression. Oncogene 19, 2619–2627 (2000).
24. Honkala, A. T., Tailor, D. & Malhotra, S. V. Guanylate-Binding Protein 1: An Emerging Target in Inflammation and Cancer. Front. Immunol. 10, (2020).
25. Giambartolomei, C. et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLOS Genet. 10, e1004383 (2014).
26. Võsa, U. et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. bioRxiv 447367 (2018) doi:10.1101/447367.
27. Westra, H.-J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet. 45, 1238–1243 (2013).
28. Wang, X. & Goldstein, D. B. Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. Am. J. Hum. Genet. 106, 215–233 (2020).
29. Shah, T. S. et al. optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. Bioinformatics 28, 1598–1603 (2012).
30. Howie, B., Marchini, J. & Stephens, M. Genotype Imputation with Thousands of Genomes. G3 Genes Genomes Genet. 1, 457–470 (2011).
31. 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).
32. Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res. 45, W130–W137 (2017).
33. Morris, A. P. et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. 44, 981–990 (2012).
34. Paternoster, L. et al. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. Nat. Genet. 47, 1449–1456 (2015).

35. Moffatt, M. F. et al. A large-scale, consortium-based genomewide association study of asthma. N. Engl. J. Med. 363, 1211–1221 (2010).
36. Köttgen, A. et al. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. Nat. Genet. 45, 145–154 (2013).
37. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. 47, 979–986 (2015).
38. Trynka, G. et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in coeliac disease. Nat. Genet. 43, 1193–1201 (2011).
39. International Multiple Sclerosis Genetics Consortium et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476, 214–219 (2011).
40. Tsoi, L. C. et al. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. Nat. Genet. 44, 1341–1348 (2012).
41. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506, 376–381 (2014).
42. Bentham, J. et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat. Genet. 47, 1457–1464 (2015).
43. Onengut-Gumuscu, S. et al. Fine-mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nat. Genet. 47, 381–386 (2015).

Chapter 6

Integration of multi-omics data and deep phenotyping enables prediction of cytokine responses

Olivier B. Bakker[1], Raul Aguirre-Gamboa[1], Serena Sanna[1], Marije Oosting[2], Sanne P. Smeekens[2], Martin Jaeger[2], Maria Zorro[1], Urmo Võsa[1], Sebo Withoff[1], Romana T. Netea-Maier[4], Hans J.P.M. Koenen[3], Irma Joosten[3], Ramnik J. Xavier[5,6], Lude Franke[1], Leo A.B. Joosten[2], Vinod Kumar[1,2], Cisca Wijmenga[1,7,], Mihai G. Netea[2,8,*] and Yang Li[1,*]*

Nature Immunology (2018)

1 Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
2 Department of Internal Medicine and Radboud Center for Infectious Diseases, Radboud University Medical Center, Nijmegen, the Netherlands
3 Department of Laboratory Medicine, Laboratory for Medical Immunology, Radboud University Medical Center, Nijmegen, the Netherlands
4 Department of Internal Medicine, Division of Endocrinology, Radboud University Medical Center, Nijmegen, the Netherlands5 Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114, USA
6 Broad Institute of MIT and Harvard University, Cambridge, MA 02142, USA
7 Department of Immunology, University of Oslo, Oslo University Hospital, Rikshospitalet, Oslo, Norway
8 Department for Genomics & Immunoregulation, Life and Medical Sciences Institute (LIMES), University of Bonn, Bonn, Germany
**Corresponding author*

Abstract

The immune response to pathogens varies substantially among people. While both genetic and non-genetic factors contribute to inter-person variation, their relative contributions and potential predictive power have remained largely unknown. By systematically correlating host factors in 534 healthy volunteers, including baseline immunological parameters and molecular profiles (genome, metabolome and gut microbiome), with cytokine-production capacity after stimulation with 20 pathogens, we identified distinct patterns of co-regulation. Among the 91 different cytokine–stimulus pairs, 11 categories of host factors together explained up to 67% of inter-individual variation in cytokine production induced by stimulation. A computational model based on genetic data predicted the genetic component of stimulus-induced cytokine-production (correlation 0.28-0.89), while non-genetic factors influenced cytokine production as well.

Background

Variability in baseline immune response influences an individual's susceptibility to immune-mediated diseases such as infection, autoimmune and inflammatory diseases, as well as their severity [1-5]. Both environmental and host factors are responsible for this variation in immune response [6-9], which makes deciphering their interaction crucial for understanding their influence on susceptibility and instrumental for building quantitative predictors of disease. The Human Functional Genomics Project (HFGP) aims to identify the factors responsible for variability in immune response in the general population and upon perturbations, such as disease state. Within the HFGP, the 500 Human Functional Genomics (500FG) consortium has collected extensive molecular and phenotypic measurements from approximately 500 healthy volunteers of Western-European descent. Earlier 500FG studies assessed the separate impacts of host-related factors, genetic variation or microbiome on cytokine-production capacity [7-9]. However, an integrated understanding of the effect of these factors and of additional host-related factors, such as endocrine hormones, circulating metabolites, platelet-mediated effects or transcriptional profiles of immune cells on stimulus induced cytokine levels has been lacking.

Here, we used a comprehensive systems biology approach to integrate the large-scale genomic, metagenomic and metabolomic data available within the 500FG consortium with the immune cell composition, hormone levels and platelet activation profiles of each person analyzed. This allowed us to describe the baseline heterogeneity of immunological parameters, identify inter-correlated immune components, infer functional connections within the immune system and build predictive models of cytokine-production capacity upon stimulation. Using transcriptome data from a subset of samples, we showed that expression of genes after stimulation explained the variation in cytokine-production better than baseline expression. By integrating multi-omics layers, we showed that cytokine production was regulated by multiple genetic and non-genetic host factors, that production of cytokines after stimulation could be moderately predicted using multiple baseline profiles and that inter-individual variation in immune responses correlated with an individual's genetic risk for (auto) immune disease.

Results

Baseline immune parameters are inter-correlated

To understand inter-individual variation in human immune response, we previously generated a database of immunological measurements, multi-omics data (cytokine response profiles, genetics, gene expression, immune cell frequencies, immune modulators, immunoglobulins, hormone levels, blood platelets, circulating metabolites, gut microbiome composition) and classical phenotypes (age, gender and BMI) from volunteers in the 500FG cohort (**Suppl. Fig. 1A,B** and **Suppl. Tab. 1**). Cytokine production capacity of individuals was assessed using previously generated ELISA profiles on the production of 6 cytokines (IL-1β, IL-17, IL-22, IL-6, TNF-α and IFN-γ), by peripheral blood mono-nuclear cells (PBMC), whole blood and PBMC derived macrophages derived from blood after stimulation with 20 pathogens (**Suppl. Tab. 2**) [7-9]. IL-1β, IL-6 and TNF-α levels were measured 24 hours after stimulation and IL-22, IL-17 and IFN-γ seven days after stimulation in PBMC and PBMC derived macrophages. In whole blood IL-1β, IL-6 and TNF-α levels were measured 48 hours after stimulation.

To map the relationships between these different molecular and immune parameters, we first performed clustering analysis of all immunological measurements besides cytokine production. To reduce the dimensionality of the dataset, the first ten principal components (PCs), covering >75% of variance in each dataset, were individually extracted from the cell count, metabolite and microbiome datasets. These PCs were then combined with the measurements of immune modulators (IL-18, IL-18BP, resistin, leptin, adiponectin, α-1 antitripsyn), immunoglobulins (IgG1-4, IgA, IgM), platelet activation profiles (p-selectin expression, fibrinogen binding, coagulation markers, β-Thromboglobulin) and hormone levels (androsteendion, cortisol, 11 deoxy cortisol, 17 hydroxy progesterone, progesterone, testosterone, 25 hydroxy vitamin D3, TSH, T4) (**Suppl. Tab. 1**). Subsequent unsupervised clustering analysis revealed several clusters (**Fig. 1**) that were consistent with previous observations, validating the current correlations. As such, we observed a negative correlation between the amount of the hormone leptin and the levels of progesterone and testosterone in peripheral blood (**Fig. 1**), consistent with an inhibitory effect of leptin on progesterone and on testosterone in humans[10-13]. We also observed a negative correlation of expression of p-selectin (whole blood flow cytometry) and fibrinogen activation profiles in peripheral blood (**Fig. 1**), consistent with evidence that they are under shared control [14,15]. Similarly, the hormone levels of 17 hydroxy-progesterone and testosterone were positively correlated with progesterone, androsteendion and 11 deoxy cortisol levels in peripheral blood (**Fig. 1**), consistent with these molecules having a common synthesis pathways. Finally, we observed the cluster of α1-antitrypsin with adiponectin and the association of 2 immune cell frequency PC's with total platelet count, as well as a negative association between IL-18 and IgM abundance (**Fig. 1**). These results show that baseline immune parameters in healthy individuals are correlated and likely to be influenced by co-regulatory pathways.

Baseline molecular profiles show substantial variation

Next, we examined the baseline (unstimulated) inter-individual variation in the immunological and molecular profiles described above and found a wide range of variation for the majority of immunological parameters analyzed (**Suppl. Fig. 1C-E**).

Because some variation is known to result from differences in age, gender and season [9,16-19], we corrected for these effects, when applicable. Among the immune-cell populations with high variability, effector T cell subpopulations showed the largest inter-individual variation compared to the other immune cell subpopulations (**Suppl. Fig. 1C**), in agreement with previous observations [6]. Baseline transcript abundance in whole blood also showed substantial inter-individual variation (**Suppl. Fig. 1D**). The top 75 most-variable transcripts were significantly enriched in 23 innate immune gene ontology (GO) terms ($p<0.05$ using an online tool [20]) (**Suppl. Tab. 3**), suggesting that the innate immune response was a major contributor to variations in transcript abundance. This analysis demonstrates that the baseline molecular profiles vary substantially between healthy individuals.

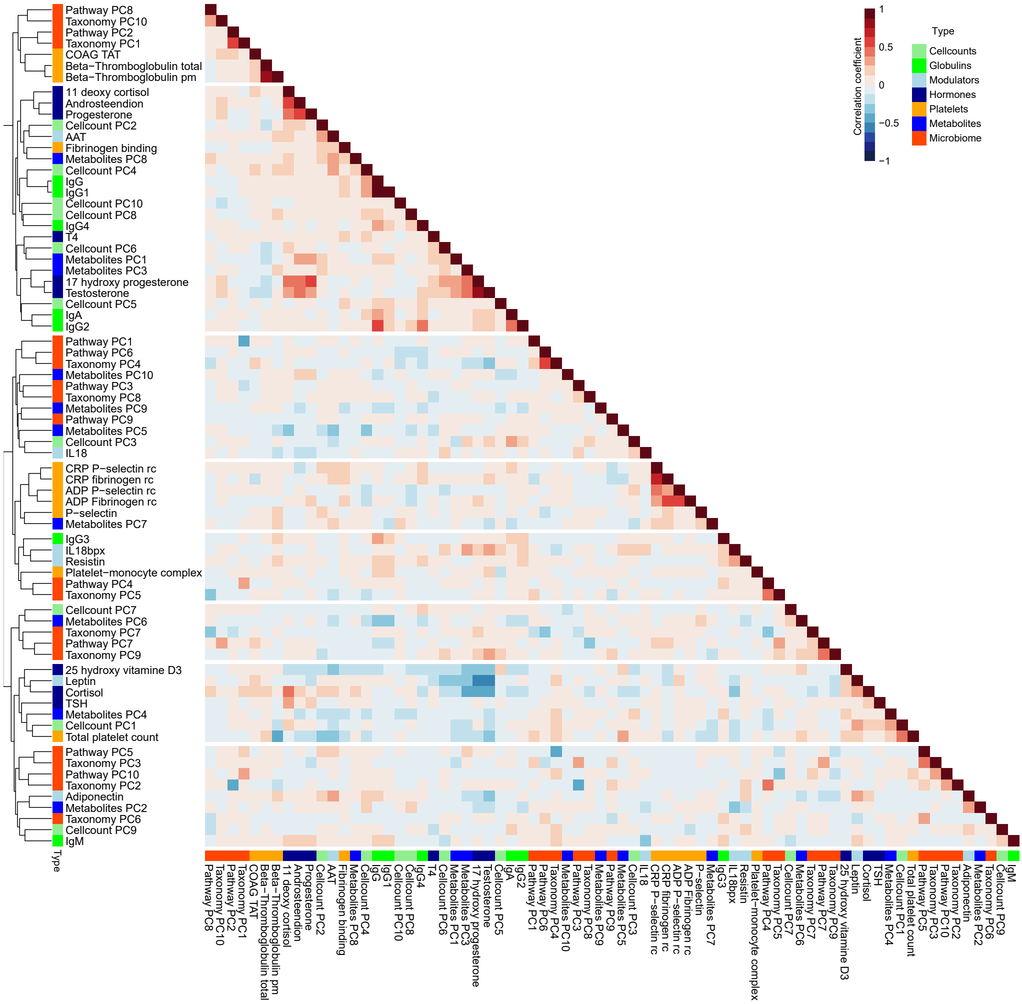


Fig. 1. Analysis of baseline immune parameters and molecular profiling shows baseline parameters are inter-correlated. Spearman's Rank correlations between both immune traits and baseline molecular profiles show that they are inter-correlated (n = 282). For the cell count and omics datasets, the first 10 principal components were extracted and used for calculating the correlation. Colors beside the cluster dendrogram indicate the type of measurements. Every sample represents an individual.

Genetics contributes the most to immune variation

To address to what extent responses to a perturbation were affected by the pre-existing immune status, we first assessed the effect of host factors at baseline on cytokine production. Using a multivariate linear model (MVLM) to examine the percent of variance explained by these factors [21], we found that genetic variation, as measured by single nucleotide polymorphisms (SNP), collectively explained most of the variation in stimulated cytokine production (avg. adj. $R^2 = 0.18$) (**Fig. 2A**). In contrast, the gut microbiome, immune-cell counts, circulating metabolites and seasons displayed only moderate effects (avg. adj. $R^2 = 0.061$, 0.057, 0.047 and 0.041, respectively) on most cytokine-stimulation pairs (**Fig. 2A**), while the concentration of circulating immunoglobulins, inflammatory mediators or hormones, and platelet activation (whole blood flow cytometry) generally had negligible effects (**Fig. 2A**,**B**). To evaluate the significance of the estimates of variation explained by genetics (VG), we performed 1000 permutations of sample labels in the cytokine data and applied the analysis pipeline on the permuted data to obtain the empirical distribution of the estimates of VG (null distribution). We subsequently compared the estimate of VG from the 500FG data with the estimate of VG from the permuted data. In total the estimates of VG in the 500FG were significant in 59 of 91 cases ($p<0.05$, **Suppl. Tab. 4**). For example, we found that the cytokine stimulation pairs explained the best by genetics (Poly I:C and *C. Burnetti* induced IL-6 levels in PBMC) showed significance.

Furthermore, we assessed several specific baseline categories that show cytokine- or pathogen-specificity in explaining the inter-individual variation (**Fig. 2B**). We observed that the abundance of circulating metabolites, including acetate and HDL cholesterol, showed a moderate negative effect on influenza-stimulated cytokine production by PBMC (avg. adjusted $R^2 = 0.19$) (**Fig. 2B**), suggesting that these factors modulate susceptibility to viral infections. The production of the lymphocyte-derived cytokines IL-17, IL-22 and IFN-γ by PBMC in response to *Aspergillus fumigatus* (*A. fumigatus)* conidia was driven more by non-genetic host factors (cell counts, platelet amounts, circulating metabolites, gut microbiome composition and season) than by genetic factors (**Fig. 2B**), which was in contrast to the genetic-component-driven cytokine production in response to all other stimulations used (**Fig. 2B**). More specifically, individuals with high concentration of HDL cholesterol or α1- antitrypsin in the circulation showed lower cytokine production in response to *A. fumigatus*. To validate the link between HDL cholesterol and cytokine production, we cultured PBMCs collected from 6 healthy volunteers in medium containing lipoprotein-deficient plasma (LPDP) and LPDP+HDL cholesterol and measured cytokine production for TNF-α, IL-1B and IL-6 in response to *A. fumigatus* conidia after 24 hours. We observed lower production of all the cytokines assessed in PBMCs cultured with HDL compared to the LPDP control (**Suppl. Fig. 2A**), indicating that HDL cholesterol modulates immune responses to *A. fumigatus* conidia.

Next, we compared the stimulus-dependent cytokine production data from the three different types of stimulation assays (PBMC, whole blood and PBMC derived macrophages) from the same individuals. We found that season, platelet-activation profiles, concentration of immune modulators, and age had a higher impact on stimulus-dependent cytokine production in PBMCs than in macrophages (**Fig. 2A**,**B**). In contrast, stimulus-dependent cytokine production correlated less with baseline metabolite levels in PBMC and whole blood then it did in macrophages (**Fig. 2A**,**B**).

This analysis shows that genetics contribute substantially to the observed inter-individual variation in cytokine level upon stimulation, and the non-genetic molecular profiles and immune parameters contribute as well.

Baseline molecules associate differentially to cytokine response

We next assessed which baseline immune and molecular components contribute most to variation in stimulus-induced cytokine production. We extracted the top five immune modulators (i.e. α1-antitrypsin, IL-18BP, adiponecting, resistin and leptin) and metabolites (i.e. the total cholesterol level in HDL3, glutamine, free cholesterol and α-1 acid glycoprotein) in the analysis of explained variance. They are the molecules that show strong association with most of the cytokine measurements in the analysis of explained variance (**Fig. 3**, **Suppl. Fig. 3**). For example, circulating IL-18BP concentrations negatively correlate with lymphocyte-derived cytokine production (IL-17, IL-22, and IFN-γ) by PBMC after stimulation, but this pattern is not observed for the monocyte-derived cytokine production (IL-1β, IL-6, and TNF-α) by PBMC after stimulation (**Fig. 3**). IL-18BP is an inhibitor of IL-18 [22] and IL-18 induces cytokine production in natural killer (NK) cells and T helper cells [23]. The known function of IL-18BP *in vitro* and the observed correlations suggested IL-18BP could potentially be a biomarker for reduced T cell activity *in vivo*. To validate the divergent effect between IL-18BP concentrations and cytokine production by lymphocytes, we tested for this association in an independent cohort of 300 volunteers of Western-European descent with BMI >25 (300OB), for which we have obtained cytokine production profiles (ELISA) after stimulation of PBMC using the same pathogens and protocols as used in 500FG. In addition, circulating baseline (unstimulated) measurements for IL-18BP were determined. Because this cohort is comprised of mainly obese (BMI >25) and older (age >55) individuals, we limited the analysis to a subset of (n=51) 300-OB volunteers with BMI <28, to bring this distribution more in line with the 500FG cohort. We tested for association (Spearman correlation) between the cytokine production profiles after stimulation and circulating IL-18BP levels (**Suppl. Fig. 2B**). We could replicate the negative effect of IL-18BP on lymphocyte cytokines.

The short chain fatty acid (SCFA) acetate showed the strongest correlation (negative correlation between -0.25 and -0.20) with influenza-induced monocyte–derived IL-1β, IL-6 and TNF-α cytokine production capacity (**Fig. 3**). Cytokine response to bacterial and fungal stimulations showed either positive or negative effects for monocyte-derived cytokine production capacity. In contrast, lymphocyte-derived IL-17, IL-22 and IFN-γ cytokine production showed consistently positive effects in response to most of the bacterial and fungal stimulations. This agrees with previous findings that SCFAs, including acetate, influence cytokine production capacity [24-26]. The negative correlation between acetate and stimulus-induced production of IL-1β, IL-6 and TNF-α was also observed when assessed in PBMC derived macrophages, but not in whole blood (**Fig. 3**). To further investigate the association between acetate and stimulus-induced cytokine production, we cultured PBMC derived macrophages obtained from whole blood of 6 healthy Dutch volunteers *in vitro* in the presence of acetate in the medium, stimulated them with MTB, *C. albicans*, *S. aureus* and *E. coli*, and assessed the cytokine production of TNF-α and IL-6 after 24 hours. We observed an association between acetate and

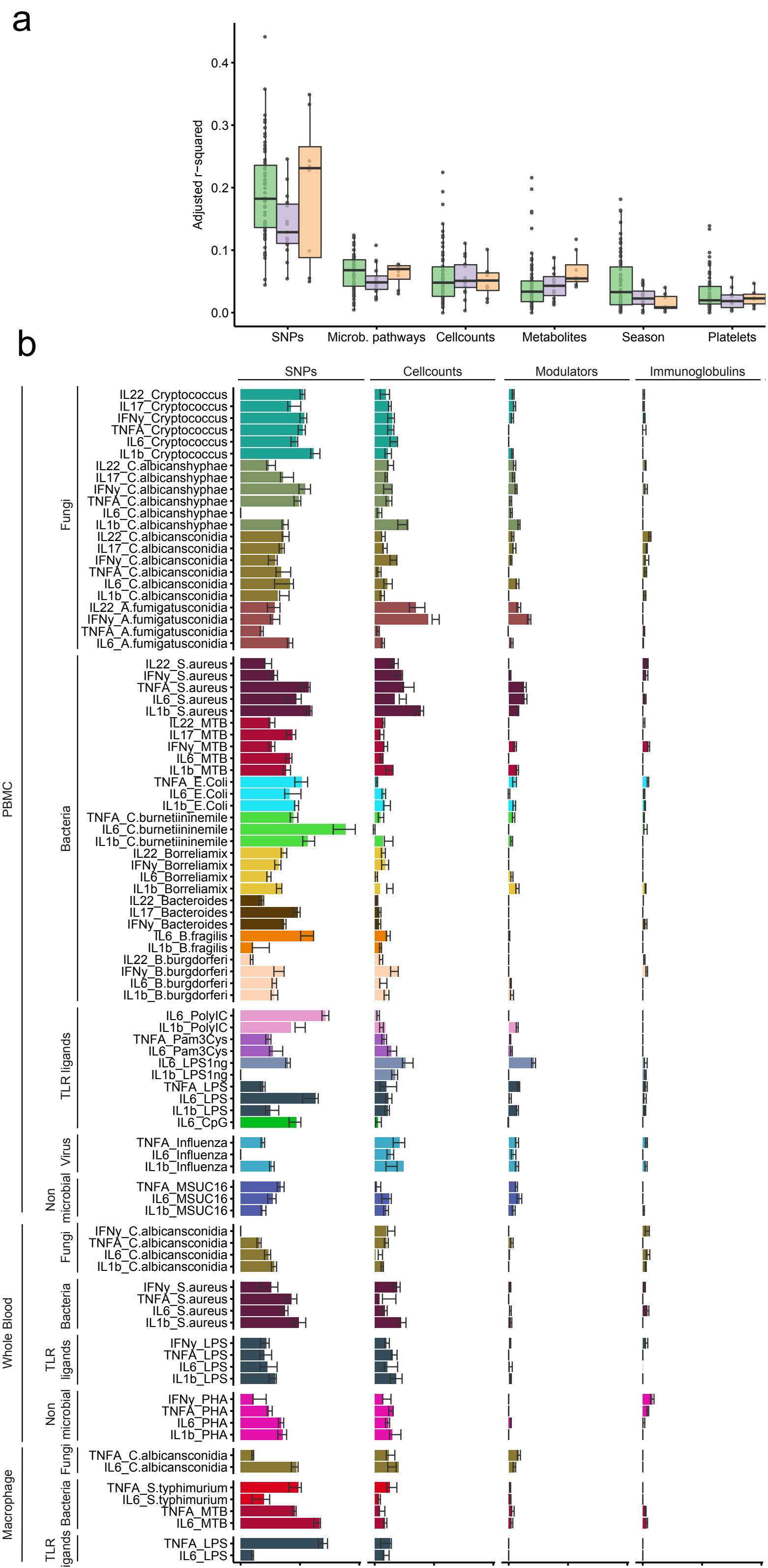
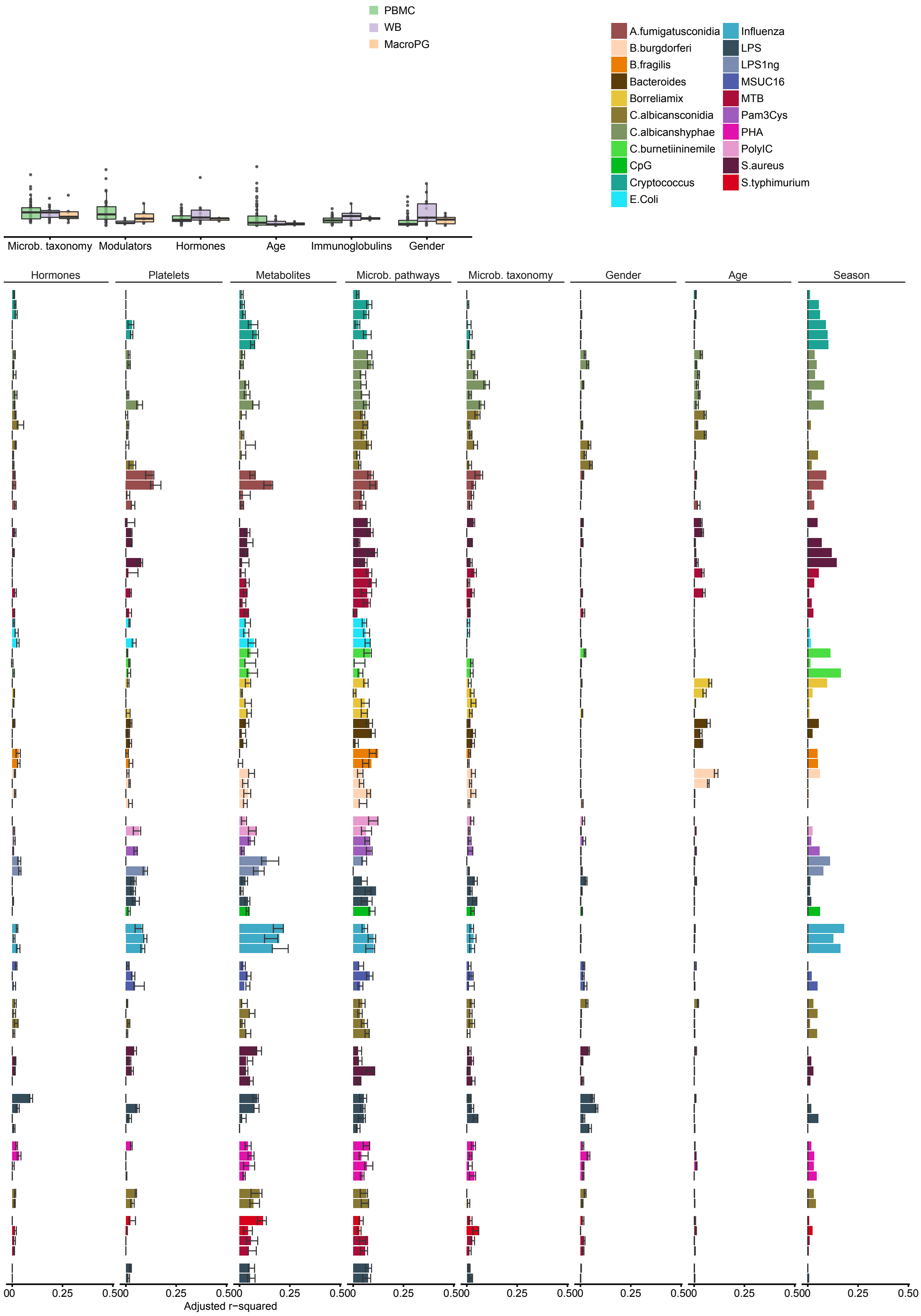


Fig. 2. Contribution of baseline immune parameters and multi-omics to cytokine variation. A) Percentage of variation in stimulated cytokine production explained by each category of measurements. The distribution indicates the adjusted R^2 of a set multivariate linear models (MVLM) representing cytokine stimulation pairs from PBMC (n=67 models), whole blood (n=16 models) and PBMC derived macrophages (n=8 models). Each dot represents the adjusted R^2 of a MVLM for a specific cytokine stimulation pair. **B)** Contribution of each category to inter-individual cytokine variation. X-axis denotes the adjusted R^2 values for the MVLMs. Bars indicate the adjusted R^2 estimated on the full dataset. Error bars indicate the standard deviation in adjusted R^2 of 10 MVLMs trained on a random subset of samples from the full data (90% of all samples). Y-axis denotes the cytokine-stimulation pairs. Colors indicate different stimulations applied in the experiments. Sample sizes differ between the different categories with the platelet, immune modulator, immunoglobulin and classical phenotypes having n = 489, the immune cell counts n = 472, the metabolites n = 377, microbial pathways n = 384, microbial taxonomy n = 411, hormones n = 486 and SNPs n = 392 samples. Every sample represents an individual.


PBMC
WB
MacroPG
A.fumigatusconidia
B.burgdorferi
B.fragilis
Bacteroides
Borreliamix
C.albicansconidia
C.albicanshyphae
C.burnetiininemile
CpG
Cryptococcus
E.Coli
Influenza
LPS
LPS1ng
MSUC16
MTB
Pam3Cys
PHA
PolyIC
S.aureus
S.typhimurium
Microb. taxonomy
Modulators
Hormones
Age
Immunoglobulins
Gender
Hormones
Platelets
Metabolites
Microb. pathways
Microb. taxonomy
Gender
Age
Season
00
0.25
0.500
0.25
0.500
0.25
0.500
0.25
0.500
0.25
0.500
0.25
0.500
0.25
0.500
0.25
0.50
Adjusted r−squared

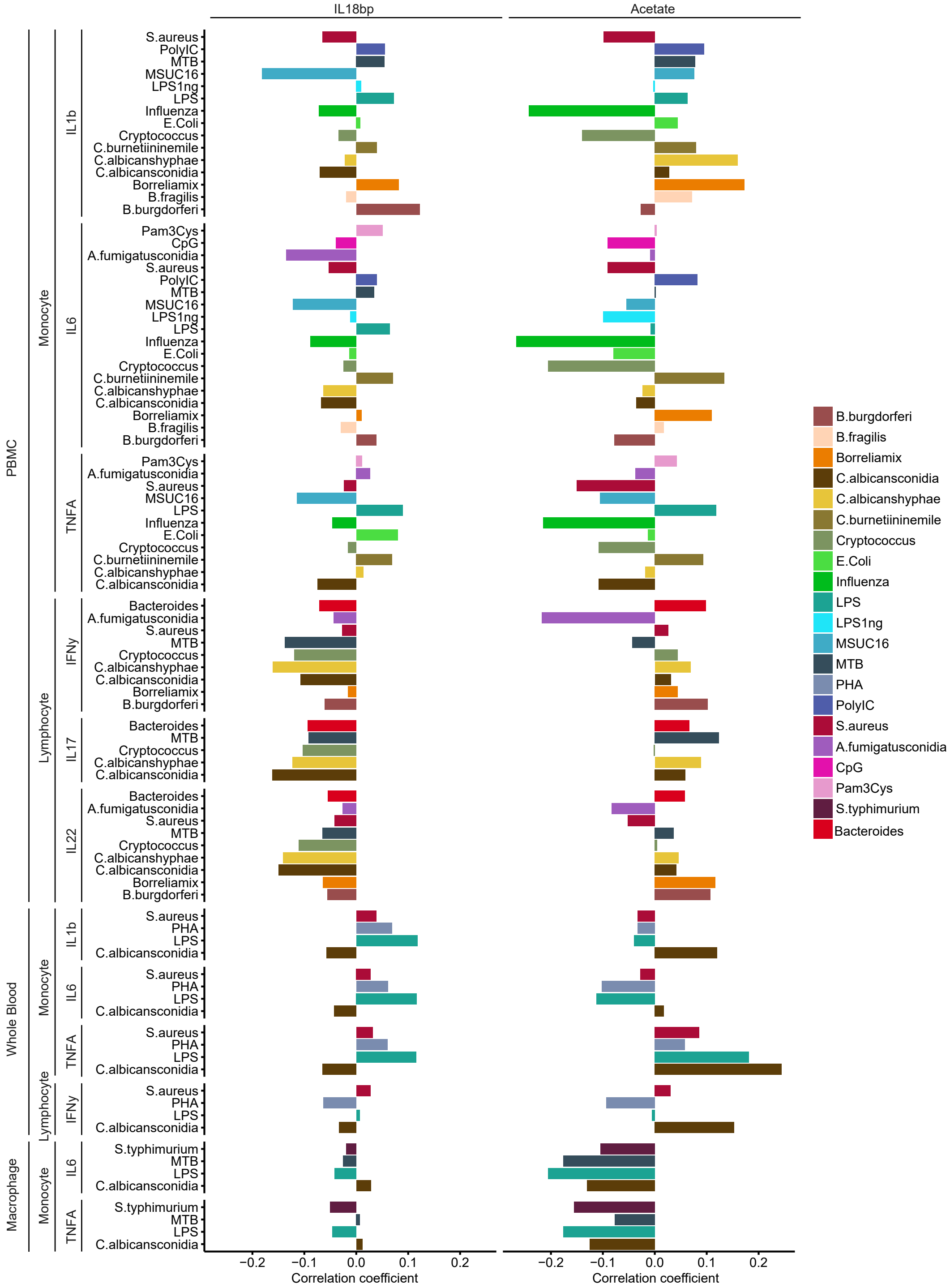


Fig. 3. Examples of baseline molecules which associate differentially to cytokine responses. IL-18BP, a circulating inhibitor of IL-18, displays negative Spearman correlations with general cytokine production capacity of lymphocytes after correcting for age and gender effects (n=489). The metabolite acetate positively correlates with stimulated cytokine production in response to influenza and displays a mostly positive effect on lymphocyte-derived cytokines after correcting for age and gender effects (n=377). Each sample represents an individual.

cytokine production in macrophages where the production of TNF-α and IL-6 in PBMC derived macrophages upon two of the stimuli (*E. coli* and *S. aureus*) were lower in the presence of acetate, but this effect was not observed for *C. albicans* (**Suppl. Fig. 2C**).

Glutamine is known to negatively regulate IL-6 production in human intestinal mucosa [27] and decreases IL-6, TNF-α and IL-1β production in biopsies from Crohn's disease patients [28]. We observed that glutamine, consistently correlated negatively with all monocyte- and lymphocyte-derived cytokines assessed after stimulation (**Suppl. Fig. 3**), suggesting it could be used as an anti-inflammatory biomarker. These results show that baseline molecules are differentially associated with cytokine production between stimuli, as well as between cell types.

Host factors explain up to 67% variation in cytokine level

To determine the collective contribution of genetic variation and immune components at baseline to cytokine production in response to pathogens, a multivariate linear model was used. We constructed a MVLM for each cytokine stimulation pair where we added relevant features from each category of dataset sequentially and subsequently evaluated the increase in variance explained by each added dataset. This integrated approach indicated that a combination of genetic, baseline molecular profiles and immune parameters can explain up to 67% of the inter-individual variation in cytokine production capacity (**Fig. 4**). Because cytokine production is a highly complex phenotype, and many factors that influence it are associated to each other, we tested if changing the order in which specific datasets were added into the models generated different results. When we compared MVLM containing all datasets, to the partial MVLMs, in which each of the 10 datasets were omitted once, we found similar estimates of explained variation as in the sequential analysis (**Suppl. Fig. 4**). For example, regardless of the order the factors were added, genetics remained the largest individual contributor to explaining inter-individual variation (**Suppl. Fig. 4**). This indicated that the order in which various factors were added into the model did not influence the results to a large extent.

Gene expression correlates with cytokine response

Next we integrated baseline transcript abundance with stimulus-induced cytokine expression. We made use of whole genome gene expression profiles obtained using RNA-Seq both before and after stimulation of peripheral blood with *C. albicans* conidia from a subset of volunteers (n = 64) from an independent Dutch cohort (Genome of The Netherlands cohort [29]). We used measurements of the production of TNF-α, IL-6 and IL-1β by PBMC upon stimulation with *C. albicans* conidia after 24 hours in the same individuals. We then applied the same MVLM based analysis approach used earlier to obtain estimates of how much inter-individual variation in cytokine production capacity could be explained by gene expression. We observed that baseline gene expression could explain a substantial portion of the inter-individual variation in production of TNF-α, IL-6 and IL-1β (**Fig. 5**). Production of TNF-α, IL-6 and IL-1β by PBMC stimulated with *C. albicans* conidia showed significantly higher correlations with gene expression induced by stimulation (adj. R^2 reaching up to 0.75) than with baseline gene expression (Wilcox test, $p=1.08 \times 10^{-5}$, $p=8.93 \times 10^{-3}$, $p=1.08 \times 10^{-5}$, for TNF-α, IL-6 and IL-1β respectively). Using GO enrichment (online tool[20]), we found that the genes selected during modelling (**Suppl. Tab. 5**) showed enrichment for several GO terms related to immune responses. For example the genes associated to *C.albicans* induced TNF-α levels were

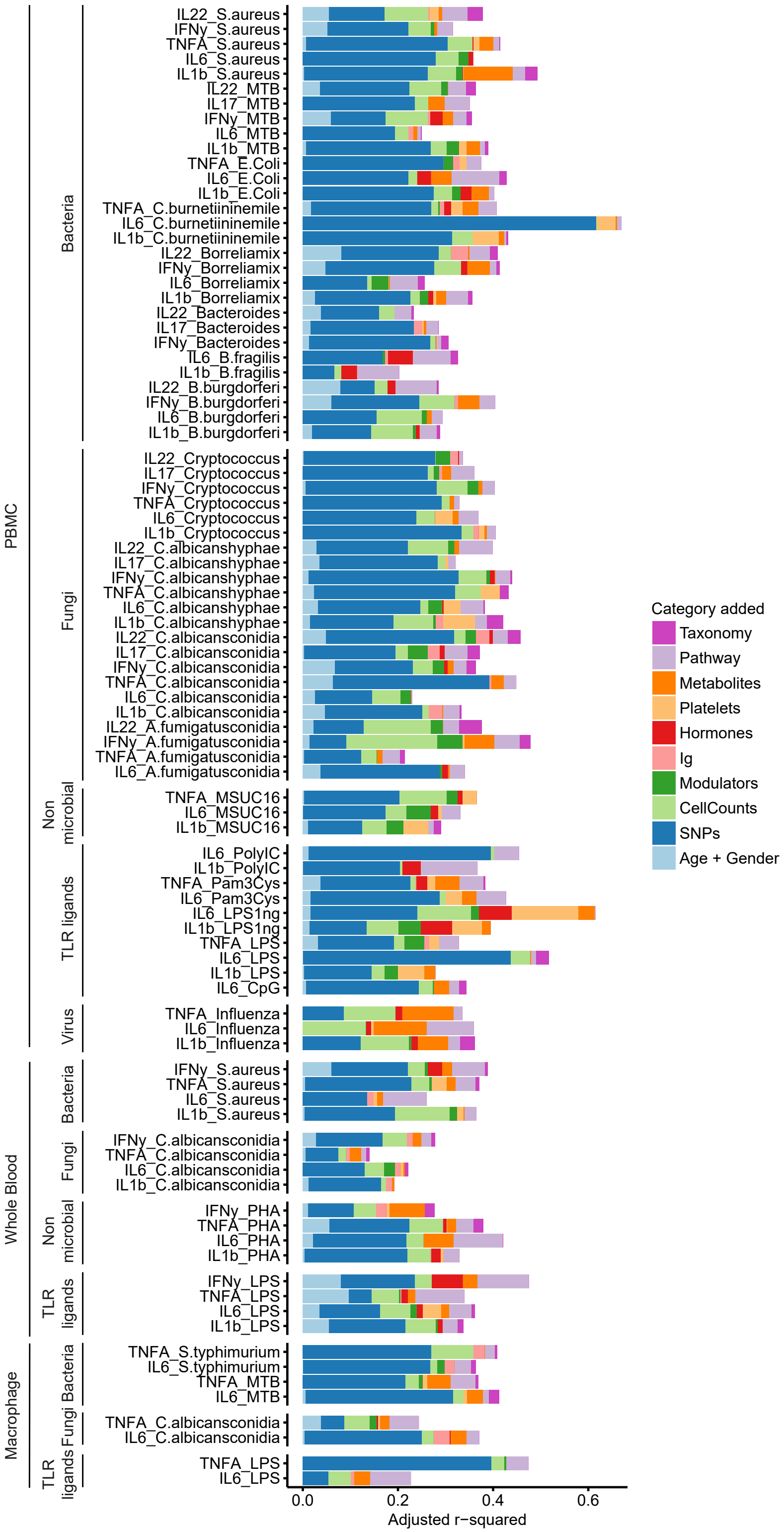


Fig. 4. Cumulative contribution of multiple baseline traits to the variation in stimulated cytokine production. Adjusted R^2 values (x-axis) obtained from multivariate linear models (MVLM) increase when measurements from 10 categories are added sequentially. Each colored bar represents how much additional variation (on top of the preceding colors) the MVLM for that category explains. The order in which features from a dataset were added is from left to right. The combined dataset consisted of 266 samples. Each sample represents an individual. Gene expression was not included in this analysis because of the relatively small sample size of the RNA-seq experiment after overlapping with the other datasets (n = 69). X-axis denotes adjusted R^2 values. Y-axis denotes different cytokine-stimulation pairs.

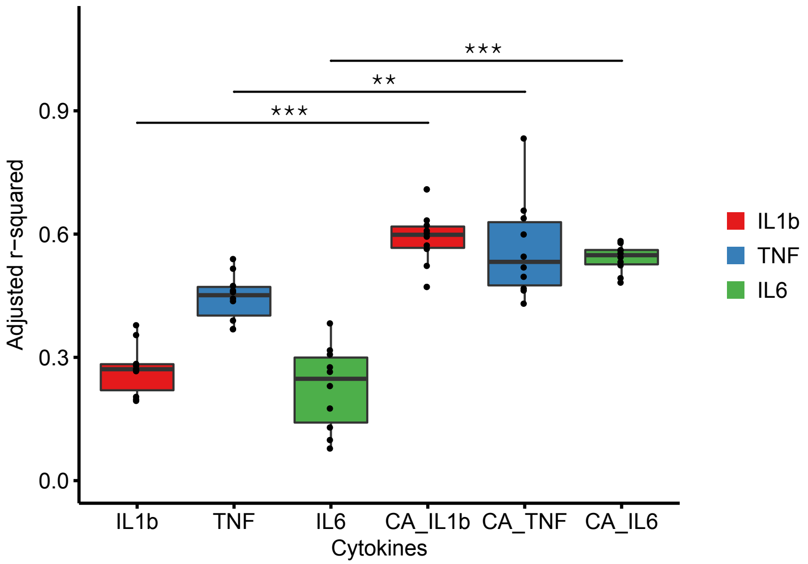


Fig. 5. Integrating gene expression profiles and cytokine production in response to *C. albicans*. Percentage of inter-individual variation (y-axis, adjusted R^2) in stimulated cytokine level of TNF-α, IL-6 and IL-1β explained by gene expression measured at baseline and upon C. albicans stimulation (denoted by CA) is significantly (Wilcox rank sum test, * $p<0.05$, ** $p<0.01$, *** $p<0.001$) higher in the multivariate linear models (MVLM) fitted on stimulated gene expression data. Exact p-values of the Wilcox rank sum test are as follows: IL-1β ($p=1.08x10^{-5}$), TNF-α ($p=8.93x10^{-4}$) and IL-6 ($p=1.08x10^{-5}$). The distribution shows adjusted R^2 (y-axis) of 10 MVLMs fitted after re-sampling using a random subset of samples (90% of all samples each time). Each dot represents the adjusted R^2 of a MVLM. The dataset consisted of 64 samples from the GoNL cohort. Each sample represents an individual.

nominally enriched for negative regulation of mast cell cytokine production ($p=1.28x10^{-3}$), negative regulation of isotype switching to IgE isotypes ($p=1.71x10^{-3}$) and negative regulation of T helper 2 cell differentiation ($p=2.15x10^{-3}$). These results imply a strong correlation between gene expression and functional responses upon stimulation by pathogens, and thus they present gene expression as a target for future studies into the prediction of immune responses.

Immune disease risk is associated with stimulated cytokine level

Many complex diseases appear to result from multiple genetic variants exerting small effects on disease risk [30], which implies that complex diseases conform closely to a classical polygenic model. Using publicly available summary statistics from GWAS we calculated polygenic risk scores (PRS) for 15 immune mediated diseases (**Suppl. Tab. 6**) for all the volunteers in the 500FG cohort as a measure of relative disease risk between individuals. We then tested whether volunteers with a higher risk for an immune mediated disease displayed higher or lower stimulus-induced cytokine production compared to the lower risk individuals. For this analysis, we focused those immune mediated diseases that showed both a significant change (two tailed, two sample t-test, Bonferroni $p<0.05$, **Suppl. Tab. 7**) compared to a permutation-based null distribution, and a consistent pattern at different thresholds used for PRS calculation (**Fig. 6A-C**, **Suppl. Fig. 5A**,**B**). We found that volunteers with higher risk for inflammatory bowel disease, multiple sclerosis, psoriasis and ulcerative colitis had significantly higher ($p<0.05$) stimulus-induced production of lymphocyte-derived (IL-17, IL-22 and IFN-γ) compared to monocyte-derived (TNF-α, IL-6 and IL-1β cytokines) (**Fig. 6**). In contrast, higher risk for type 1 diabetes (T1D) and rheumatoid arthritis was associated with increased stimulus-induced production of monocyte-derived (TNF-α, IL-6 and IL-1β) compared to lymphocyte-derived cytokines (**Fig. 6C**). Higher risk for Crohn's disease, eczema and type 2 diabetes was associated with a significant increase (compared to their respective null distributions, $p<0.05$) in both monocyte- and lymphocyte-derived cytokines compared to the permutation-based null distribution, with no significant differences between the monocyte and lymphocyte derived groups (**Fig. 6B**). These observations suggest that the genetic basis for immune-mediated diseases could influence the functionality of the immune system even in otherwise healthy individuals.

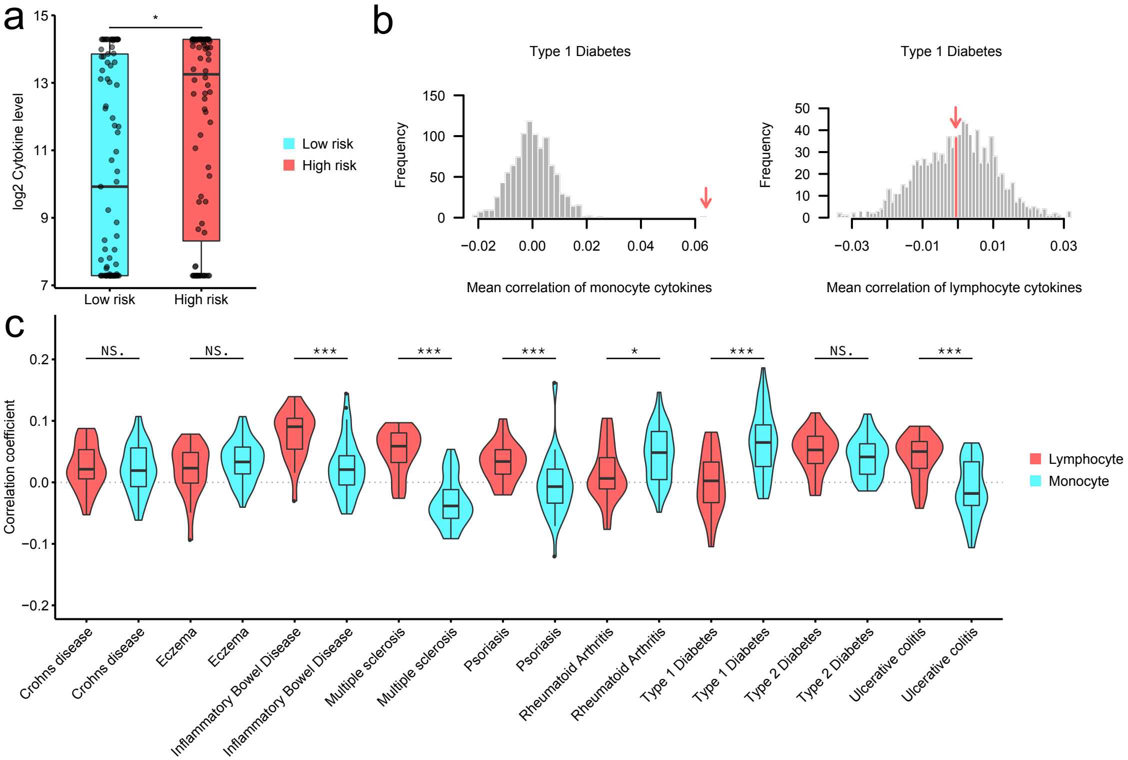


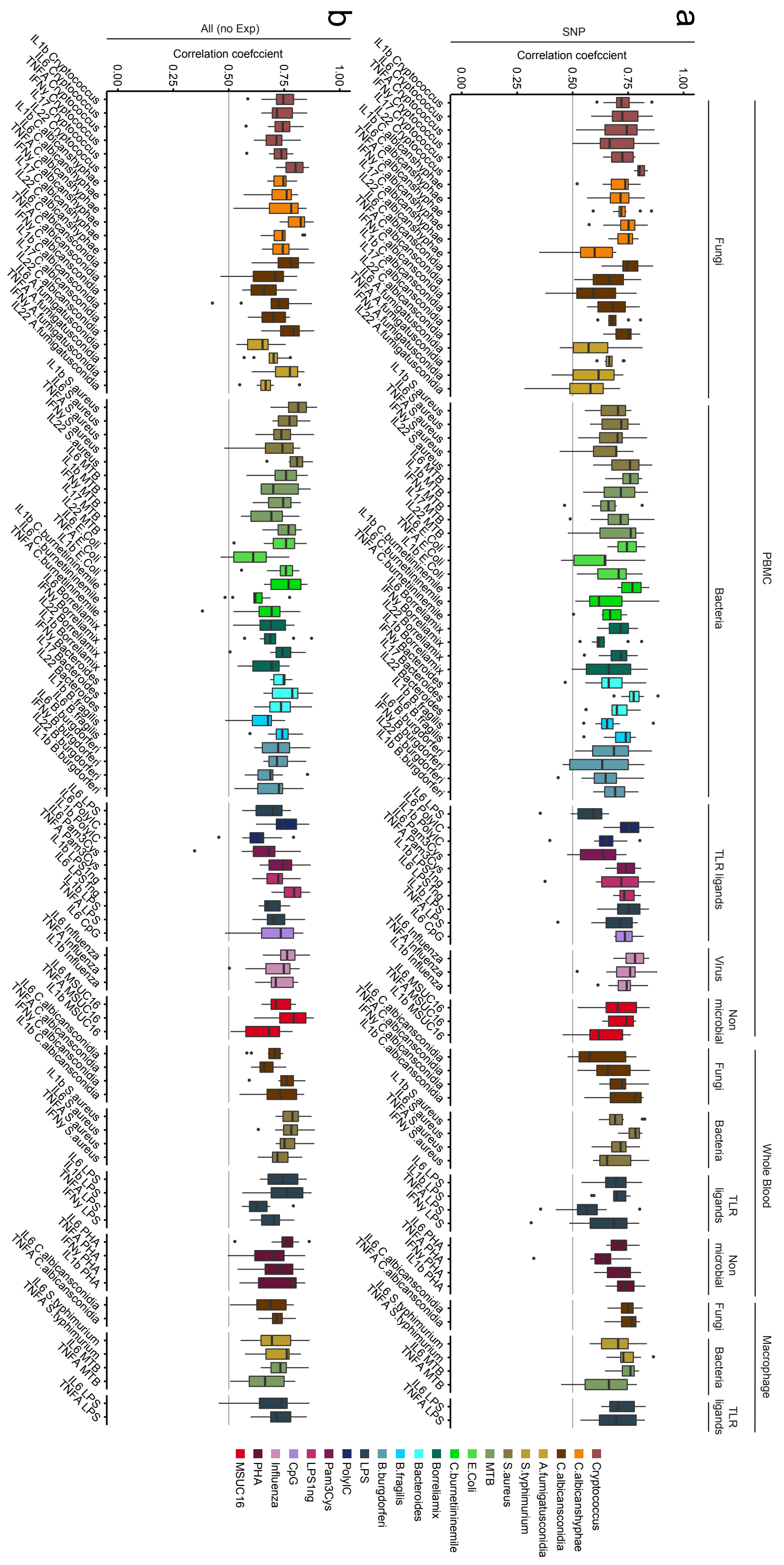
Fig. 6. Stimulated cytokine production correlates with genetic risk score for autoimmune diseases. A) Example individuals with high genetic risk for (auto)immune disease tend to be high producers of cytokines in response to pathogens. * indicates the significance of the Wilcox rank sum test between low- and high-risk groups for T1D ($p=0.011$). Low- and high-risk groups (x-axis) were selected by taking the top and bottom quantile of the PRS for T1D. Y-axis indicates the IL-6 level after stimulation of PBMCs with influenza. **B)** Distribution mean correlations between T1D risk in monocyte-derived cytokines (left panel) and lymphocyte cytokines (right panel) for 1000 permutations. The measured estimate is indicated by the red arrow. T1D shows significance for monocyte derived cytokines (left) but not for the lymphocyte derived cytokines (right). **C)** Distribution of Spearman correlation coefficients between stimulated cytokine production and genetic risk score for immune disease in 430 individuals, shown for PBMC. Genetic risk scores calculated based on genome-wide association studies for different diseases. Significant differences in mean correlation between the lymphocyte- and monocyte-derived cytokines are shown by Wilcox rank sum test (* $p<0.05$, ** $p<0.01$, *** $p<0.001$). Exact p-values are as follows Crohn's disease $p=7.28\times10^{-1}$, eczema $p=2.55\times10^{-1}$, inflammatory bowel disease $p=9.34\times10^{-6}$, multiple sclerosis $p=4.85\times10^{-11}$, psoriasis $p=1.40\times10^{-4}$, rheumatoid arthritis $p=1.41\times10^{-2}$, type 1 diabetes $p=1.00\times10^{-5}$, type 2 diabetes $p=1.65\times10^{-1}$, ulcerative colitis $p=1.34\times10^{-5}$.

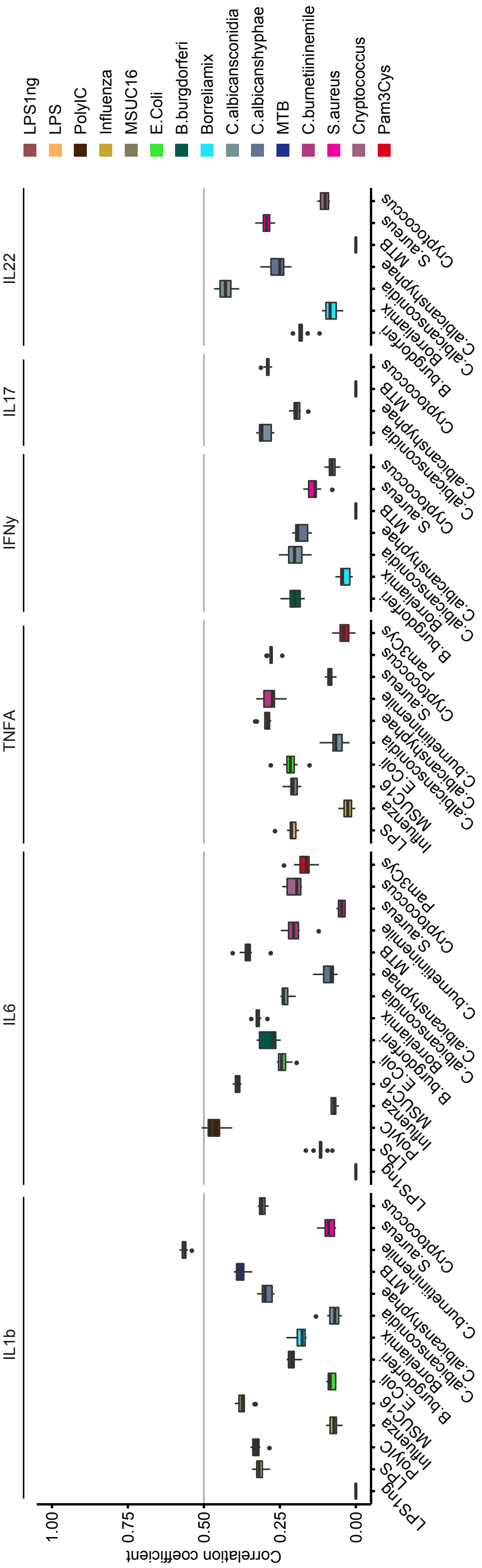
Stimulated cytokine level predicted by genetics

Finally, we integrated both genetics and other molecular features to construct MVLMs to predict each cytokine stimulation pair in PBMC, whole blood and macrophages. To achieve the best prediction of *ex vivo* stimulus-induced cytokine production, we tested several linear prediction methods (Elastic Net, RR-BLUP and PLS) and compared them using both genetic and non-genetic factors to train the MVLMs for each cytokine stimulation pair. Predictive performance was quantified by Spearman's correlation between the measured and the predicted stimulus-induced cytokine production in multiple randomly selected subsets of the volunteers from 500FG. While the prediction performances of the different methods are similar (**Suppl. Fig. 6A-C**), Elastic Net marginally outperformed the others, so we used it for subsequent analyses.

We first tested if SNP data could predict cytokine production. Among the 91 stimulation-cytokine pairs, the correlations between predicted and measured stimulus-induced cytokine production were, on average, 0.69 (range 0.28-0.89) (**Fig. 7A**). Inclusion of the baseline immune parameters and multi-omics data significantly increased the predictive power and stability of the model (two tailed student t-test, $p=1.36x10^{-9}$, t-statistic=6.09, degrees of freedom = 1792) and most predictions for cytokine production increased to, on average, 0.72 (range 0.35-0.90) (**Fig. 7B**). Additional inclusion of the gene expression data from the RNA-seq analysis decreased the predictive power (avg. 0.60, range 0-1) (**Suppl. Fig. 6D**), most likely due to the reduced number of samples for which both RNA-seq and the other factors were available (n = 69).

We then tested the predictive capabilities of the Elastic net trained MVLMs using only SNPs as input and applying it to independent subset of 500FG individuals were new cytokine stimulation experiments were performed (50FG). We found prediction accuracies up to 0.56 for some cytokine stimulation pairs (**Fig. 8**), although the MVLMs performed poorly for most stimulations. Among the best-performing stimulus-cytokine pairs, *C .burnetti* stimulated IL-1β and Poly I:C-stimulated IL-6 gave prediction accuracies of on average 0.56 and 0.46 respectively (**Fig. 8**). Because both pathways are known to have a large genetic component [31] this indicated that the MVLMs could predict cytokine production for stimulus-induced cytokines whose mechanism of induction are primarily driven by genetics. By applying MVLMs to genetics data, we were able to predict the cytokine production upon stimulation, with varying degrees of accuracy.


a
SNP
Correlation coefcient
b
All (no Exp)
Correlation coefcient
0.00
0.25
0.50
0.75
1.00
PBMC
Whole Blood
Macrophage
Fungi
Bacteria
TLR ligands
Virus
Non microbial
Cryptococcus
C.albicanshyphae
C.albicansconidia
A.fumigatusconidia
S.typhimurium
S.aureus
MTB
E.Coli
C.burnetiininemile
Borreliamix
Bacteroides
B.fragilis
B.burgdorferi
LPS
PolyIC
Pam3Cys
LPS1ng
CpG
Influenza
PHA
MSUC16



Left: Fig. 7. Cytokine production in response to pathogens can be predicted using genetics and baseline immune profiles. Spearman correlation between predicted and measured cytokine levels (y-axis) are shown for each of the 10 multivariate linear models from cross validation for all available cytokine stimulation pairs. Cytokine production in response to pathogens can be predicted using SNPs (n = 392 individuals). Prediction accuracy increases when baseline immune parameters and molecular profiles (immune cell frequencies, immune modulators, immunoglobulins, hormone levels, blood platelets, circulating metabolites, gut microbiome composition) are added to the model (n = 353 individuals).

Right: Fig. 8. Prediction using the genetic model in an independent dataset shows some cytokine stimulation pairs can be predicted successfully. Spearman correlations between predicted cytokine level by the multivariate linear models (MVLM) built using genetics (n = 336) and the measured values in an independent set of stimulation experiments (n = 56). The boxplots show the variation in Spearman correlations from each of the 10 MVLMs predictions from the cross validation strategy.

Discussion

In this study we assessed the combined contribution of genetic and non-genetic factors to the inter-individual variation in cytokine production in response to pathogens by examining the cytokine production of immune cells following stimulation with 20 different pathogens or TLR ligands *ex vivo* in PBMC, whole blood and PBMC derived macrophages. This analysis identified new modulators of cytokine production, including circulating inflammatory mediators and metabolites. We found that volunteers with increased genetic risk for immune mediated diseases were more likely to be high responders in terms of stimulus-induced cytokine production. Finally, we trained MVLMs that could predict human stimulus-induced cytokine production for Poly I:C induced IL-6 and *C. burnetti* IL-1β levels in PBMC using only the genetic profiles or a combination of genetic and other molecular profiles.

A recent study on the heritability of immune phenotypes in 210 twins suggested that variations in circulating cytokine concentrations are mostly driven by non-heritable influences [32]. Although we observed here that genetics was the largest single contributor to inter-individual variation (avg. adj. $R^2 = 0.18$), this still leaves room for the majority of the variation to be explained by non-genetic influences. Any differences we observed in estimates of heritability are likely due to differences in the experimental design between the two studies. As such, we assessed cytokine profiles upon stimulation *ex vivo*, whereas the above study [32] measured baseline circulating concentrations *in vivo*. This strongly suggests that it is the response to pathogens during infection that is under stronger genetic pressure rather than the background level of mediators in the circulation. Our study thus agrees with the idea that infections have a strong selective impact on the genetic control of immune responses [33-40].

The present study has potentially important implications for our understanding of the human immune response. We found out that acetate, a circulating metabolite, was associated with changes in stimulus-induced cytokine production and especially in the modulation of Th1 and Th17 responses. SCFA such as acetate, propionate and succinate are released by the gut microbiome and current literature suggests that SCFA have important immunomodulatory properties [24-27]. We show here that acetate has similar effects in humans *in vivo*. It appears important to further investigate the broader impact of SCFA and identify which microbiome profiles modify their concentration in the circulation. We found a strong inhibitory effect of acetate on influenza-stimulated cytokine production, a phenomenon that deserves further scrutiny. Another important metabolic pathway that strongly influenced cytokine responses was the cholesterol and lipoprotein synthesis pathway. Cholesterol pathways have been described to have important immune-modulating effects, with the levels of cholesterol sulfate, a derivative of membrane cholesterol, shown to influence immune processes such as TCR signalling and thymic selection [41]. Here we showed that HDL cholesterol negatively impacted influenza and *Aspergillus*-stimulated cytokine production, possibly with important effects on the pathophysiology of these infections.

The ability to calculate prediction scores for specific immune mediated diseases and to link them to cytokine production shows that certain stimulus-induced cytokine profiles may contribute to particular diseases, e.g. the capacity to release high amounts of monocyte-derived cytokines in T1D. Although we acknowledge that our power to detect these smaller associations is relatively limited, our approach can be used to link any given phenotype to disease scores when individual-level data is available. This offers the opportunity to identify immune pathways important in disease, which may represent new therapeutic targets.

A second limitation of the 500FG cohort is that it contains a higher proportion of young people than the general population [9], which could introduce age bias into the MVLM's predictions. While we acknowledge that the performance of the MVLMs prediction may vary in a population with a different range in age, BMI or ancestry, our study represents a proof-of-concept that stimulus-induced cytokine production can be moderately predicted. Future studies in larger general population cohorts with greater ranges of age and ethnicity will contribute to the generation of models with improved predictive potential for a general population. Future studies should also aim to extend the current analysis, which was limited to common SNP polymorphisms (MAF >0.1), to include rare variants and mutations, a broadening of scope likely to further increase the observed impact of genetics on cytokine production upon stimulation.

In conclusion, we present the most comprehensive assessment to date of the host factors that influence cytokine production. We show that genetics was a major contributor to the inter-individual variation in cytokine production upon pathogen stimulation. However, other non-genetic factors also influenced cytokine production in response to most stimuli, including gut microbiome composition, immune cell numbers in circulation and circulating metabolite concentrations. Individuals with increased genetic risk for a given immune disease tended to have increased cytokine production, and stimulus-induced cytokine production could be predicted for Poly I:C induced IL-6 and *C. burnetti* IL-1β levels. This study provides the fundamentals for predicting components of cytokine production based on genetics and baseline host factor profiles, paving the way towards personalized immune-based therapies.

Methods

Study cohort.

The main analyses were performed in the 500FG cohort, which is part of the Human Functional Genomics Project. This cohort consists of 534 healthy individuals (237 males and 296 females) of Caucasian origin. Volunteers range from 18 to 75 years of age, with the majority (421 individuals) being 30 years or younger (**Suppl. Fig. 1A**). BMI is within normal limits (15 to 35) with the majority (380 individuals) having a BMI between 20 and 25 (**Suppl. Fig. 1B**). Of these 534 original volunteers, 45 were excluded based on genetic background and questionnaire results (medication usage, chronic disease) leaving 489 individuals.

Replication cohort.

Validation experiments were performed in the 300-OB cohort. This cohort consists of ~300 Dutch individuals. All individuals had a BMI >25, with an average BMI of 31, and range in age from 55 to 80 years, with an average age of 67 years. Validations were performed in a subset of the 300-OB cohort with an BMI <28 (N=55). Circulating metabolites and mediators as well as stimulated cytokine levels were measured in the same way as in 500FG.

Experimental procedures.

The experimental procedures used to measure levels of cytokines, modulators, immunoglobulins and hormones have been described previously [9]. Genotyping, metagenomic sequencing of the gut microbiome, FACS sorting of PBMCs and determination of platelet activation profiles have also been described previously [7,8,42]. We selected a representative subset of 89 samples from the 500FG cohort for RNASeq (balanced for age and sex to match the original distribution in the cohort). These samples were processed for sequencing using the Illumina TruSeq version 2 library preparation kit. Paired-end sequencing of 2×50-bp reads was performed using the Illumina HiSeq 2000 platform. The quality of the raw reads was checked using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Read alignment was performed using STAR 2.3.0 [43], and aligned reads sorted using SAMTools. Gene level quantification of reads was done using HTSeq[44]. Circulating metabolites were measured and analysed using the BrainShake Biomarker Analysis Platform that is based on nuclear magnetic resonance (NMR) spectroscopy (BrainShake, Finland).

Statistical methods

Data pre-filtering.

After pre-processing, the gene expression, SNP, metabolite and microbiome datasets were filtered to remove any non-significantly-associated features. This was done to increase the efficiency of downstream analysis. The gene expression metabolite and microbiome datasets were correlated to all of the cytokine measurements, and all features showing a Spearman correlation with a Benjamini-Hochberg adjusted p <0.05 to at least one cytokine were kept. This resulted in a dataset of 4,499 genes, 205 metabolites, 509 microbial pathways and 162 microbial taxonomies. The genetic variants were filtered using previously generated cytokine QTL profiles [7] by setting the p-value cut-off at various thresholds depending on the application. To calculate the variance explained by genetics, a p-value threshold of $p < 5\times10^{-6}$ was chosen. For prediction using the Elastic Net model, various thresholds were evaluated after which all SNPs with a $p < 5\times10^{-5}$ were included in the analysis.

Estimation of explained variance.

The estimation of variance explained by each of the data levels on the different stimulated cytokine production profiles was performed by applying a correlation-based feature selection approach. In this approach, we built a model for each stimulated cytokine measurement in which only features associated to this measurement are included in the model. We select these features by first regressing out the effects of age and gender, then associating the features in a data level to the current cytokine stimulation pair. If a feature showed a significant association (Spearman p-value <0.05), the feature was included in the set of potential predictors. Once all the associations had been computed, the set of potential predictors was correlated to itself to identify collinearity among this predictor set. If features within this predictor set showed an association (Spearman correlation >0.4), the feature which showed the least association (based on the correlation p-values) to the cytokine stimulation pair is removed. This yielded a unique set of predictors for every cytokine stimulation pair, which was then used to fit a multivariate linear model to estimate the variance explained by these features for that cytokine stimulation pair. To account for the inflation that adding predictors has on the explained variation, the adjusted R^2 was taken as the measure of explained variance.

Permutation of cytokine GWAS.

The baseline cytokine GWAS was performed as described previously [7]. We randomly permuted the cytokine and covariate datasets 1000 times then ran the GWAS using these datasets to obtain 1000 random profiles for each cytokine stimulation pair. For each run we obtained the QTL profile and estimated the explained variance using the permuted cytokine and covariate dataset and the pipeline described above. This yielded a distribution of 1000 estimates of explained variance for each cytokine stimulation pair. A measured estimate was considered significant if it was in the top 5% of the permuted distribution of estimates for that cytokine stimulation pair.

Estimation of age and gender effects.

Age and gender effects on cytokine production were assessed by fitting univariate linear models for each cytokine stimulation pair with age and gender as the independent variables, respectively. The R^2 was taken as the measure of explained variation of these models.

Estimation of seasonal effect.

The effect of season on stimulated cytokine production was assessed using a linear combination of sine and cosine terms with the same period (equation 1) as described by ter Horst et al.[9]:

$$y = \beta + \alpha_1 \sin\left(\frac{2\pi x}{365}\right) + \alpha_2 \cos\left(\frac{2\pi x}{365}\right) + \epsilon \ (1)$$

Where y represents the response (cytokine level), β the estimated intercept, α the estimated predictor effect, x the day of the year the sample was taken in, and e the residual effect.

Estimation of cumulative explained variance.

To assess the proportion of variance that can be explained by all levels cumulatively, individual levels were added to a multivariate linear model one by one, and the total model adjusted R^2 calculated for each step. If adding a level showed an increase in the total adjusted R^2 of the model, this value was extracted. To assess the contribution of each level conditional upon the others, the full model was fit first. Subsequently several reduced models were fit where one data level was missing. The adjusted R^2 for this full model was then compared against the model with the missing level. The difference between the reduced model and the full model was taken as a measure of the variance explained by that level when accounting for the effects of the other levels.

Cytokine level prediction.

Our objectives were to investigate whether genetic variants can reveal predictive insights into the cytokine production upon stimulation and whether baseline immune parameters, which are treated as quantitative phenotypes that are continuously distributed over a population, can improve predictive power for cytokine production upon stimulation. Using our population-based study, we searched for those subsets of genetic variants and immune components that are most predictive of the various stimulated cytokine production profiles, rather than using exclusively those variants meeting a stringent level of statistical significance.

We assessed the validity of this approach by applying multiple methods, each of which is discussed in detail below. In total three datasets were evaluated: one for predicting stimulated cytokine production using only SNPs, one containing all levels except gene expression, and one with all levels including gene expression. Firstly, features with little association with cytokine production levels (Spearman p >0.05) were removed for building the prediction models. For the SNP dataset, all SNPs with an association to a cytokine stimulation pair with $p<5\times10^{-5}$ were used as input for feature selection. No filtering for collinearity was applied because Elastic Net accounts for potential collinearity among predictors [45].

Elastic Net.

Prediction of the cytokine levels was facilitated by training an Elastic Net model. A 2×10-fold cross-validation approach was used, where the data was first split up into 10 random training and test sets to validate the prediction, and the training set was then split up once more for feature selection. Prediction accuracy was evaluated by calculating Spearman correlations between the measured cytokine levels and the predictions of the Elastic Net model on the test sets.

RR BLUP.

To show that the prediction results are not influenced to a large extent by the methodology, a mixed linear model (equation 2) , as implemented in the package rrBLUP [46], was applied:

$$y = 1\mu + Zu + \epsilon \; (2)$$

Where y represents the response (cytokine level), 1 a vector of 1S, u the overall mean of the training set, Z the matrix of predictors (traits), u the random effect of the predictors, and e a vector of residual effects. Predictions were made using 10-fold cross-validation. Spearman correlation was then calculated between predicted and measured values. We applied this model as was described previously [47].

Partial least squares regression.

In addition to the Elastic Net and rrBLUP a partial least squares model was applied. Models were validated using 10-fold cross-validation. Prediction of cytokine levels on the test set was done using a linear model (equation 3):

$$y = \beta + \alpha X + \epsilon \; (3)$$

Where y represents the response (cytokine level), β the intercept, α a vector containing the coefficients from the model, X the matrix of predictors (immune traits), and e the residual error.

Polygenic risk scores.

We carried out polygenic scoring of disease risk using publically available GWAS results. Quantitative scores were computed for each trait in this study based on the set of SNPs with p-values lower than predefined p-value thresholds (pT) in the GWAS. Multiple pT were evaluated ($pT < 5x10^{-8}$, $1x10^{-5}$, $1x10^{-4}$, $1x10^{-3}$, and $pT < 1x10^{-2}$). Throughout this work, we refer to the scores defined at $pT < 1x10^{-5}$ as Polygenic Risk Scores (PRS). Full association summary statistics were downloaded from several publicly available resources indicated in **Suppl. Tab. 6** [48-60], . Studies done exclusively in non-European cohorts were omitted. Filters applied to the separate data sources are indicated below. All the dbSNP rs numbers were standardized to match GIANT 1KG p 1V3 and the directions of the effects were standardized to correspond to the GIANT 1KG p 1V3 minor allele. SNPs with different opposite-strand alleles compared to GIANT alleles were flipped. SNPs with A/T and C/G SNPs and SNPs with different alleles GIANT 1KG p 1V3 (tri-allelic SNPs, indels, unknown alleles) were removed from the analysis. Genomic control was applied to all p-values for the datasets not genotyped by Immunochip or Metabochip. We calculated PRS by first clumping variants based on the threshold pT, linkage-disequilibrium ($R^2 < 0.2$) and a 250kb window using the PLINK 1.9 option "clump" and exclusively European samples from 1000 genomes data as a reference for linkage disequilibrium calculation. PRS were subsequently obtained for each threshold pT by calculating them using the linkage-disequilibrium-clumped subset of SNPs using the PLINK 1.9 option “score”.

Association between polygenic risk scores and cytokine production.

The association between the PRS and cytokine production capacity upon stimulation was determined by calculating the Spearman correlation between each of the PRS profiles and each of the stimulated cytokine profiles. To evaluate the statistical significance of association, a permutation method was used. The cytokine data was permuted 1000 times and the correlation was calculated for each of these permuted datasets. Both the measured and permuted distributions were separated into the lymphocyte and monocyte groups, and a student t-test was applied between the measured distribution and the permuted distribution. When either the monocyte or lymphocyte group showed a significant deviation from the permuted distribution (Bonferroni adjusted two sample t-test $p < 0.05$) the disease was selected for interpretation.

Data availability

The data that support the findings of this study are available at https://hfgp.bbmri.nl/ were it has been meticulously catalogued and archived at BBMRI-NL aiming for maximum reuse following the FAIR principles, i.e., Findability, Accessibility, Interoperability, and Reusability. Individual level genetic data as well as other privacy sensitive datasets are available upon request at http://www.humanfunctionalgenomics.org/site/?page_id=16. These datasets are not publicly available because they contain information that could compromise the research participants privacy. The central data stewardship and access has been implemented using MOLGENIS open source platform for scientific data that enables flexible data upload, management and querying, including sufficiently rich metadata and interfaces for machine processing and custom (R statistics) visualization for human processing (see http://molgenis.org). Also summaries of the study have been submitted to BBMRI central catalogues https://catalogue.bbmri.nl (Netherlands) and http://www. bbmri-eric.eu/news-events/bbmri-eric-directory-2-0/ (EU).

Acknowledgements

We thank all of the volunteers in the 500FG and GoNL cohorts for their participation. We thank T.A. Wassenaar and L. Steenhuis of the Hanze University of Applied Science for providing input on the project and helping with the analyses. We thank K. Mc Intyre for editing the text. We thank the EAGLE eczema consortium for making their summary statistics publicly available. The HFGP is supported by a European Research Council (ERC) Consolidator grant (ERC 310372). This study was further supported by an IN-CONTROL CVON grant (CVON2012-03) and a Netherlands Organization for Scientific Research (NWO) Spinoza prize (NWO SPI 94-212) to M.G.N.; an ERC advanced grant (FP/2007-2013/ERC grant 2012-322698) and an NWO Spinoza prize (NWO SPI 92-266) to C.W.; a European Union Seventh Framework Programme grant (EU FP7) TANDEM project (HEALTH-F3-2012-305279) to C.W. and V.K.; and an NWO VENI grant (NWO 863.13.011) and an ZonMw-OffRoad-91215206 grant to Y.L. M.O. was supported by an NWO VENI grant (016.176.006). RJX was supported by Nation Institutes of Health (NIH) grants - DK43351, AT009708, AI137325.

Author contributions

Y.L., C.W. and M.G.N. designed this study. M.O., S.P.S., M.J., R.T.N.-M., H.J.P.M.K., I.J., R.J.X., and L.A.B.J. performed the experiments and processed the data. U.V. collected and pre-processed public summary statistics. O.B.B. performed statistical analysis assisted by R.A.-G., S.S. ,U.V. and L.F.. O.B.B., M.Z., Y.L., S.W., V.K., M.G.N, and C.W interpreted the data. Y.L., C.W., M.G.N. and O.B.B. wrote the manuscript with input from all authors.

Competing Financial Interests

The authors declare no competing interests.

Ethics statement

The HFGP study was approved by the ethical committee of Radboud University Nijmegen (no. 42561.091.12). Experiments were conducted according to the principles expressed in the Declaration of Helsinki. Samples of venous blood were drawn after informed consent was obtained.

Supplementary material

Supplemenatary material are provided at: https://doi.org/10.1038/s41590-018-0121-3

References

1. Fairfax, B. P. et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science 343, 1246949 (2014).
2. Kumar, V., Wijmenga, C. & Xavier, R. J. Genetics of immune-mediated disorders: from genome-wide association to molecular mechanism. Current Opinion in Immunology 31, 51–57 (2014).
3. Lee, M. N. et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science 343, 1246980 (2014).
4. Netea, M. G., Wijmenga, C. & O'Neill, L. A. J. Genetic variation in Toll-like receptors and disease susceptibility. Nat. Immunol. 13, 535–542 (2012).
5. Ye, C. J. et al. Intersection of population variation and autoimmunity genetics in human T cell activation. Science 345, 1254665 (2014).
6. Brodin, P. & Davis, M. M. Human immune system variation. Nat Rev Immunol 17, 21–29 (2017).
7. Li, Y. et al. A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans. Cell 167, 1099-1110.e14 (2016).
8. Schirmer, M. et al. Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. Cell 167, 1125-1136.e8 (2016).
9. ter Horst, R. et al. Host and Environmental Factors Influencing Individual Human Cytokine Responses. Cell 167, 1111-1124.e13 (2016).
10. Brannian, J. D., Zhao, Y. & McElroy, M. Leptin inhibits gonadotrophin-stimulated granulosa cell progesterone production by antagonizing insulin action. Hum Reprod 14, 1445–1448 (1999).
11. Härle, P. et al. Possible role of leptin in hypoandrogenicity in patients with systemic lupus erythematosus and rheumatoid arthritis. Annals of the Rheumatic Diseases 63, 809–816 (2004).
12. Blum, W. F. et al. Plasma Leptin Levels in Healthy Children and Adolescents: Dependence on Body Mass Index, Body Fat Mass, Gender, Pubertal Stage, and Testosterone. J Clin Endocrinol Metab 82, 2904–2910 (1997).
13. Behre, H. M., Simoni, M. & Nieschlag, E. Strong association between serum levels of leptin and testosterone in men. Clinical Endocrinology 47, 237–240 (1997).
14. Xu, T. et al. P-Selectin Cross-Links PSGL-1 and Enhances Neutrophil Adhesion to Fibrinogen and ICAM-1 in a Src Kinase-Dependent, but GPCR-Independent Mechanism. Cell Adh Migr 1, 115–123 (2007).
15. Gawaz, M., Langer, H. & May, A. E. Platelets in inflammation and atherogenesis. J Clin Invest 115, 3378–3384 (2005).
16. Furman, D. et al. Apoptosis and other immune biomarkers predict influenza vaccine responsiveness. Mol Syst Biol 9, 659 (2013).
17. Furman, D. et al. Cytomegalovirus infection improves immune responses to influenza. Sci Transl Med 7, 281ra43 (2015).
18. Furman, D. et al. Systems analysis of sex differences reveals an immunosuppressive role for testosterone in the response to influenza vaccination. PNAS 111, 869–874 (2014).
19. Davis, M. M., Tato, C. M. & Furman, D. Systems immunology: just getting started. Nat Immunol 18, 725–732 (2017).

20. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. Nat Genet 25, 25–29 (2000).
21. Shah, S. et al. Improving Phenotypic Prediction by Combining Genetic and Epigenetic Associations. Am. J. Hum. Genet. 97, 75–85 (2015).
22. Novick, D. et al. Interleukin-18 Binding Protein. Immunity 10, 127–136 (1999).
23. Okamura, H. et al. Cloning of a new cytokine that induces IFN-gamma production by T cells. Nature 378, 88–91 (1995).
24. Vinolo, M. A. R., Rodrigues, H. G., Nachbar, R. T. & Curi, R. Regulation of Inflammation by Short Chain Fatty Acids. Nutrients 3, 858–876 (2011).
25. Tedelind, S., Westberg, F., Kjerrulf, M. & Vidal, A. Anti-inflammatory properties of the short-chain fatty acids acetate and propionate: a study with relevance to inflammatory bowel disease. World Journal of Gastroenterology 13, 2826 (2007).
26. Cavaglieri, C. R. et al. Differential effects of short-chain fatty acids on proliferation and production of pro- and anti-inflammatory cytokines by cultured lymphocytes. Life Sciences 73, 1683–1690 (2003).
27. Coëffier, M., Marion, R., Ducrotté, P. & Déchelotte, P. Modulating effect of glutamine on IL-1β-induced cytokine production by human gut. Clinical Nutrition 22, 407–413 (2003).
28. Lecleire, S. et al. Combined Glutamine and Arginine Decrease Proinflammatory Cytokine Production by Biopsies from Crohn’s Patients in Association with Changes in Nuclear Factor-κB and p38 Mitogen-Activated Protein Kinase Pathways. J. Nutr. 138, 2481–2486 (2008).
29. 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).
30. Lvovs, D., Favorova, O. O. & Favorov, A. V. A Polygenic Approach to the Study of Polygenic Diseases. Acta Naturae 4, 59–71 (2012).
31. Li, Y. et al. Inter-individual variability and genetic influences on cytokine responses to bacteria and fungi. Nature Medicine 22, 952–960 (2016).
32. Brodin, P. et al. Variation in the human immune system is largely driven by non-heritable influences. Cell 160, 37–47 (2015).
33. Barreiro, L. B. & Quintana-Murci, L. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet 11, 17–30 (2010).
34. Casals, F. et al. Genetic adaptation of the antibacterial human innate immunity network. BMC Evolutionary Biology 11, 202 (2011).
35. Andrés, A. M. et al. Targets of Balancing Selection in the Human Genome. Mol Biol Evol 26, 2755–2764 (2009).
36. Pickrell, J. K. et al. Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 19, 826–837 (2009).
37. Tang, K., Thornton, K. R. & Stoneking, M. A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. PLOS Biology 5, e171 (2007).
38. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A Map of Recent Positive Selection in the Human Genome. PLOS Biology 4, e72 (2006).
39. Wang, E. T., Kodama, G., Baldi, P. & Moyzis, R. K. Global landscape of recent inferred Darwinian selection for Homo sapiens. PNAS 103, 135–140 (2006).
40. Williamson, S. H. et al. Localizing Recent Adaptive Evolution in the Human Genome. PLOS Genetics 3, e90 (2007).
41. Wang, F., Beck-García, K., Zorzin, C., Schamel, W. W. A. & Davis, M. M. Inhibition of T cell receptor signaling by cholesterol sulfate, a naturally occurring derivative of membrane cholesterol. Nat Immunol 17, 844–850 (2016).
42. Aguirre-Gamboa, R. et al. Differential Effects of Environmental and Genetic Factors on T and B Cell Immune Traits. Cell Reports 17, 2474–2487 (2016).
43. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
44. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169 (2015).
45. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320 (2005).
46. Endelman, J. B. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome Journal 4, 250 (2011).
47. Riedelsheimer, C. et al. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet 44, 217–220 (2012).
48. Morris, A. P. et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. 44, 981–990 (2012).
49. Paternoster, L. et al. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. Nat. Genet. 47, 1449–1456 (2015).
50. Putov, N. V., Bulatov, P. K., Gorovenko, G. G., Fedoseev, G. B. & Brusilovskiĭ, B. M. [Classification of unspecific diseases of the bronchopulmonary system]. Vrach Delo 52–56 (1977).
51. Köttgen, A. et al. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. Nat. Genet. 45, 145–154 (2013).
52. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. 47, 979–986 (2015).
53. Trynka, G. et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in coeliac disease. Nat. Genet. 43, 1193–1201 (2011).
54. Hinks, A. et al. Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. Nat. Genet. 45, 664–669 (2013).
55. International Multiple Sclerosis Genetics Consortium et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476, 214–219 (2011).

56. Cordell, H. J. et al. International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. Nat Commun 6, 8019 (2015).
57. Tsoi, L. C. et al. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. Nat. Genet. 44, 1341–1348 (2012).
58. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506, 376–381 (2014).
59. Bentham, J. et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat. Genet. 47, 1457–1464 (2015).
60. Onengut-Gumuscu, S. et al. Fine-mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nat. Genet. 47, 381–386 (2015).

Chapter 7

Evolution of cytokine production capacity in ancient and modern European populations

Jorge Domınguez-Andre´s[1,2], Yunus Kuijpers[3,4*], Olivier B. Bakker[5], Martin Jaeger[1,2], Cheng-Jian Xu[1,3,4], Jos W.M. Van der Meer[1], Mattias Jakobsson[6,7], Jaume Bertranpetit[8], Leo A.B. Joosten[1,2], Yang Li[1,2,3,4#] and Mihai G. Netea[1,2,9,#]*

Elife (2021)

1 Department of Internal Medicine and Radboud Center for Infectious diseases (RCI), Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands;
2 Radboud Institute for Molecular Life Sciences (RIMLS), RadboudUniversity Medical Center, Nijmegen, Netherlands;
3 Department of Computational Biology for Individualised Infection Medicine, Centre for Individualised Infection Medicine (CiiM), a joint venture between Helmholtz-Centre for Infection Research (HZI) and the Hannover Medical School (MHH), Hannover, Germany;
4 TWINCORE, Centre for Experimental and Clinical Infection Research, a joint venture between Helmholtz-Centre for Infection Research (HZI) and the Hannover Medical School (MHH), Hannover, Germany;
5 Department of Genetics, University Medical Centre Groningen, Nijmegen, Netherlands; 6Human Evolution, Department of Organismal Biology, Uppsala University, Uppsala, Sweden;
7 Centre for Anthropological Research, Department of Anthropology and Development Studies, University of Johannesburg, Auckland Park, South Africa;
8 Institut de Biologia Evolutiva (UPF- CSIC), Universitat Pompeu Fabra, Barcelona, Spain;
9 Department for Genomics & Immunoregulation, Life and Medical Sciences Institute (LIMES), University of Bonn, Bonn, Germany
** These authors contributed equally*
# These authors contributed equally

Abstract

As our ancestors migrated throughout different continents, natural selection increased the presence of alleles advantageous in the new environments. Heritable variations that alter the susceptibility to diseases vary with the historical period, the virulence of the infections, and their geographical spread. In this study we built polygenic scores for heritable traits that influence the genetic adaptation in the production of cytokines and immune-mediated disorders, including infectious, inflammatory, and autoimmune diseases, and applied them to the genomes of several ancient European populations. We observed that the advent of the Neolithic was a turning point for immune-mediated traits in Europeans, favouring those alleles linked with the development of tolerance against intracellular pathogens and promoting inflammatory responses against extracellular microbes. These evolutionary patterns are also associated with an increased presence of traits related to inflammatory and auto-immune diseases.

Introduction

Human history has been shaped by infectious diseases. Human genes, especially host defence genes, have been constantly influenced by the pathogens encountered [1-3]. Pathogens drive the selection of genetic variants affecting resistance or tolerance to the infection, and heritable variations that increase survival to diseases with high morbidity and mortality will be naturally selected in people before reproductive age [2]. These selection signatures vary with historical period, virulence of the pathogen, and the geographical spread.

Here we investigated the historical evolutionary patterns leading to genetic adaptation in cytokine production and immune-mediated diseases, including infectious, inflammatory, and autoimmune diseases. Cytokine production capacity is a key component of the host defence mechanisms: it induces inflammation, activates phagocytes to eliminate the pathogens and present antigens, and controls induction of T helper (Th) adaptive immune responses. We have therefore chosen to investigate the evolutionary trajectories of cytokine production capacity in modern human populations during history. To determine the difference in polygenic regulation of diseases and cytokine production capacity, we used data derived from the 500 Functional Genomics (500FG) cohort of the Human Functional Genomics Project (HFGP; http://www.humanfunctionalgenomics.org). The HFGP is an international collaboration aiming to identify the host and environmental factors responsible for the variability of human immune responses in health and disease [4]. Within the HFGP project, the 500FG study generated a large database of immunological, phenotypic, and multi-omics data from a cohort of 534 individuals of Western-European ancestry, which has been used to integrate the impact of genetic and environmental factors on cytokine production and immune parameters. We subsequently deciphered the factors that influence inter-individual variations in the immune responses against different stimuli [5-8.]

Results and discussion

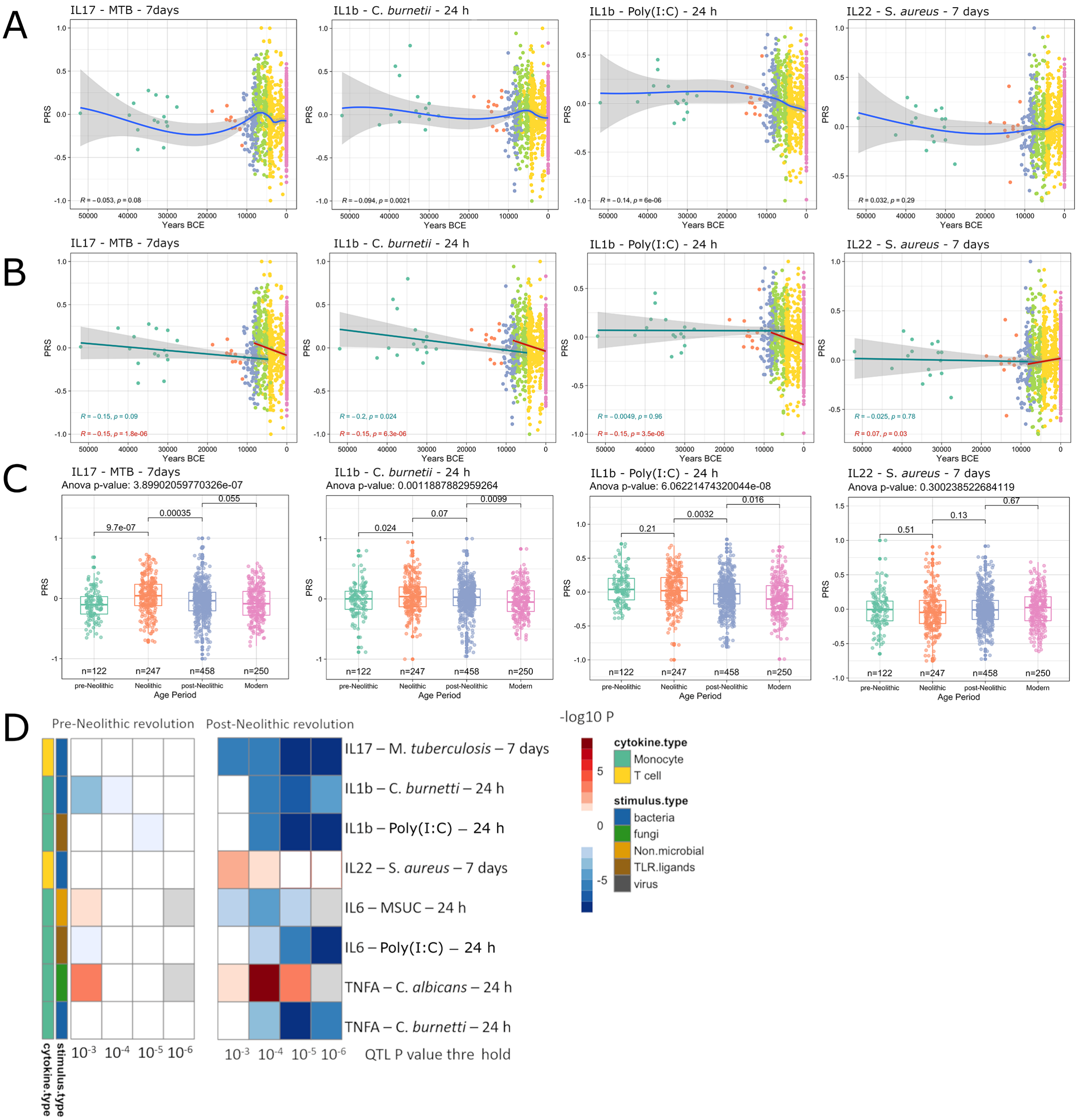
Peripheral blood mononuclear cells from these individuals were challenged with bacterial, fungal, viral, and non-microbial stimuli, and six cytokines (tumor necrosis factor (TNF)-α, interleukin (IL)- 1β, IL-6, IL-17, IL-22, and interferon (IFN)-γ were measured at 24 hr or 7 days after stimulation, generating 105 cytokine-stimulation pairs (**Fig. 1—Suppl. Fig. 1** and **Suppl. file 1A**). The stimulation time intervals were chosen based on extensive studies that showed that the time points used were best suited for assessing monocyte-derived and lymphocyte-derived cytokines per stimulus. Not all the stimuli induce the production of all cytokines; so the selection of the cytokine-stimulus pairs was performed for those pairs for which cytokine production was measurable [6-9]. We correlated cytokine production with genetic variant data to obtain cytokine quantitative trait loci (QTLs), which were employed to compute and compare the polygenic risk score (PRS) of the genomes of 827 individuals from different human historical eras (early upper Palaeolithic, late upper Palaeolithic, Mesolithic, Neolithic, post-Neolithic), which were downloaded from version 37.2 of the compiled dataset containing unimputed published ancient genotypes (https://reich.hms.harvard.edu/downloadable-genotypes-present-day-and-ancient-dna-data-compiled-published-papers) and 250 modern Europeans randomly selected from the European 1000G cohort. The individuals in this cohort present a similar genetic background (Western-European ancestry) and balanced characteristics in terms of age, sex, body mass index (BMI), and habits [8], allowing us to represent the natural variability in the immune responses and minimizing the impact of the interindividual variability in our results. In line with this, cytokines were measured in batches in which a certain cytokine was measured the same day for all the samples, in order to decrease the potential influence of technical variations. Intra-individual variation of cytokine production capacity was found to be limited 10 while inter-individual variability was largely dependent on genetic variants [5,6].

We investigated the PRS changes over time through constructing linear models and correlation analysis. In order to account for the ancient DNA (aDNA) samples being pseudo-haploid, ambiguous single-nucleotide polymorphisms (SNPs) (A/T and C/G) were excluded when computing PRS to prevent errors due to strand flips. PRS was computed using the most significant QTLs that had a p-value lower than our predetermined threshold for each given trait and removing all variants within a 250kb window around these variants. Additional PRS models were calculated at varying window sizes (250, 500, 1000 kb) in order to show consistency in the direction of the trends (**Fig. 1—Suppl. Fig. 2**). The dosage of these variants was multiplied by their effect size while the dosage of missing variants in a sample was supplemented with the average dosage. Although replacing missing values with the population average does not skew the observed trends over time, it does affect the scale of the data. We opted for this approach, however, due to the scarcity of variants shared by all samples; we focused instead on consistent observations after excluding samples with higher missing genotype rates. Finally, we scaled the PRS to a range of -1 and 1 and correlated the scores of the samples with their respective carbon-dated ages. In order to verify the robustness of our results, we repeated the analysis at multiple threshold combinations for variant missingness and QTL thresholds.

We plotted locally estimated scatterplot smoothing (LOESS) regression models to show the variations in PRS across time for each trait and highlight the time periods around which they drastically change. After plotting the changes before and after the Neolithic revolution at various thresholds for both the missing genotype rate in samples as well as QTL inclusion threshold, only traits that showed consistently significant trends in the same direction were chosen. Additional boxplots and t-tests show the variations in PRS between different pairings besides pre and post-Neolithic samples. A schematic representation of the steps performed is shown in **Fig. 1—Suppl. Fig. 3**. Lastly, we used a separate test to determine whether traits were under selective pressure between any two broad time periods using the trait-associated Wright's fixation index (Fst). This test shows an increase in the number of traits under selective pressure after the start of the Neolithic and highlights the traits and points in time for which the observed trends in PRS can be partially attributed to selective pressure as opposed to drift (**Suppl. file 1B**).

Applying the methodology described above, several patterns were apparent (**Fig. 1**). The first overall observation is that the estimation of cytokine production capacity based on PRS shows significant differences between populations in various historical periods, and the strength of evolutionary pressure on cytokine responses was different before and after the Neolithic revolution. We did not observe significant changes in cytokine production capacity between individuals who lived in different historical periods before the Neolithic, whereas strong pressure is apparent after adoption of agriculture and animal domestication in Europe. This different pattern may have resulted from the more limited number of samples available for the older time periods, resulting in lower statistical power, but the presence of some evolutionary pressure also before the Neolithic argues that this is most likely not the full explanation. The development of agriculture and domestication of animals in the Neolithic increased population densities on the one hand and the contact between humans and domesticated animals as a source of pathogens on the other hand. The number of zoonoses increased dramatically (examples being tuberculosis, brucellosis, Q-fever, and influenza), which strongly increased the selective pressure and caused significant adaptations of immunity at the genetic level [11]. Most of the genetic adaptations to pathogens took place in the period since modern humans abandoned their hunting-gathering lifestyle and developed agriculture [12]. In this respect, the strongest changes leading to tolerance (decreased cytokine production) were exerted in the cytokine responses to intracellular zoonotic infections (*tuberculosis* and *Coxiella*). In contrast, responses to the extracellular pathogens *Staphylococcus aureus* and *Candida albicans* indicate increased resistance, with high production of IL-22 and TNF-α, respectively. The increased response to the important fungal pathogen *C. albicans* after the Neolithic period is validated also at the transcriptional level. Overall, these patterns are reminiscent of the studies showing that human immune responses need to adapt to a new landscape of infectious agents depending on the geographical location and types of microbe encountered [13]. Such different patterns were most likely encountered also through history.

Importantly, our results also show significant patterns in changes in the production of specific cytokines during history. The resistance against intracellular pathogens increased after Neolithic with higher IFN-γ responses (see **Fig. 1—Suppl. Fig. 1**): indeed, it is known that Th1-IFN-γ responses are crucial for the host defence against intracellular pathogens such as mycobacteria or *Coxiella* [14]. In addition, the resistance to the extracellular pathogens *C. albicans* and *S. aureus* is also increased after this Neolithic era, with TNF-α and IFN-γ production increasing steadily after. These two


A
IL17 - MTB - 7days
R = −0.053, p = 0.08
IL1b - C. burnetii - 24 h
R = −0.094, p = 0.0021
IL1b - Poly(I:C) - 24 h
R = −0.14, p = 6e-06
IL22 - S. aureus - 7 days
R = 0.032, p = 0.29
PRS
Years BCE
B
IL17 - MTB - 7days
R = −0.15, p = 0.09
R = −0.15, p = 1.8e-06
IL1b - C. burnetii - 24 h
R = −0.2, p = 0.024
R = −0.15, p = 6.3e-06
IL1b - Poly(I:C) - 24 h
R = −0.0049, p = 0.96
R = −0.15, p = 3.5e-06
IL22 - S. aureus - 7 days
R = −0.025, p = 0.78
R = 0.07, p = 0.03
C
IL17 - MTB - 7days
Anova p-value: 3.89902059770326e-07
9.7e-07
0.00035
0.055
IL1b - C. burnetii - 24 h
Anova p-value: 0.0011887882959264
0.024
0.07
0.0099
IL1b - Poly(I:C) - 24 h
Anova p-value: 6.06221474320044e-08
0.21
0.0032
0.016
IL22 - S. aureus - 7 days
Anova p-value: 0.300238522684119
0.51
0.13
0.67
n=122
n=247
n=458
n=250
pre-Neolithic
Neolithic
post-Neolithic
Modern
Age Period
D
Pre-Neolithic revolution
Post-Neolithic revolution
-log10 P
IL17 – M. tuberculosis – 7 days
IL1b – C. burnetti – 24 h
IL1b – Poly(I:C) – 24 h
IL22 – S. aureus – 7 days
IL6 – MSUC – 24 h
IL6 – Poly(I:C) – 24 h
TNFA – C. albicans – 24 h
TNFA – C. burnetti – 24 h
cytokine.type
Monocyte
T cell
stimulus.type
bacteria
fungi
Non.microbial
TLR.ligands
virus
5
0
-5
10-3 10-4 10-5 10-6
QTL P value thre hold
cytokine.type
stimulus.type

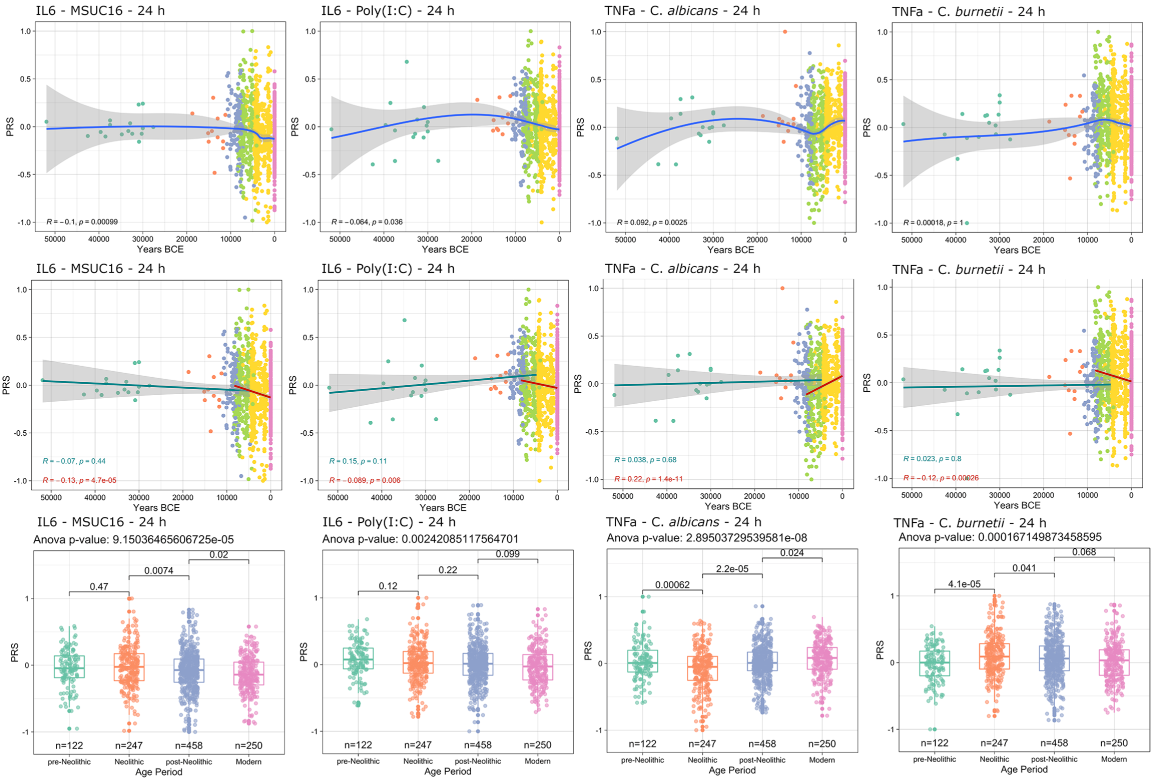


Fig. 1. Distribution of PRS of cytokine response QTLs across time. A–C) Polygenic risk score (PRS) models based on quantitative trait loci (QTLs) at a p-value threshold of 1x10^{-5}. Gray area around regression lines represents the 95% confidence interval. **A**) LOESS regression models showing the changing differences in PRS across time. **B)** Dual linear models showing the difference in trends before and after the Neolithic revolution. **C)** Boxplots showing the difference in mean PRS between adjacent broad time periods using a t-test and the overall difference in means using analysis of variance (ANOVA). **D)** Heatmap showing the consistency in regression trends before and after the Neolithic revolution using different QTL inclusion thresholds (1x10^{-3} to 1x10^{-6}). Intracellular organisms that can cause zoonotic disease: *Mycobacterium tuberculosis, Coxiella burnetti*. Extracellular organisms: *Staphylococcus aureus*, *Candida albicans*. MSUC: monosodium urate crystals. PolyIC: polyinosinic:polycytidylic acid.

cytokines are very well known to be important for anti-Candida and anti- Staphylococcus host defence [15,16]. On the other hand, a different pattern emerges in relation with the IL-1β/IL-6/IL-17 axis: the production of these cytokines is seen decreasing after Neolithic (see **Fig. 1A and B**). In this context, the decrease through time of poly I:C induction of cytokines, as a model of viral stimulation, is intriguing but potentially very important: many important viruses such as influenza and coronaviruses (severe acute respiratory syndrome (SARS), Middle East respiratory syndrome (MERS), and SARS coronavirus 2 (SARS-CoV-2)) exert life-threatening effects through induction of cytokine mediated hyperinflammation (also termed 'cytokine storm') [17]; evolutionary processes to curtail these exaggerated responses are thus likely to be protective, and tolerance against viruses becomes a host defence mechanism [18].

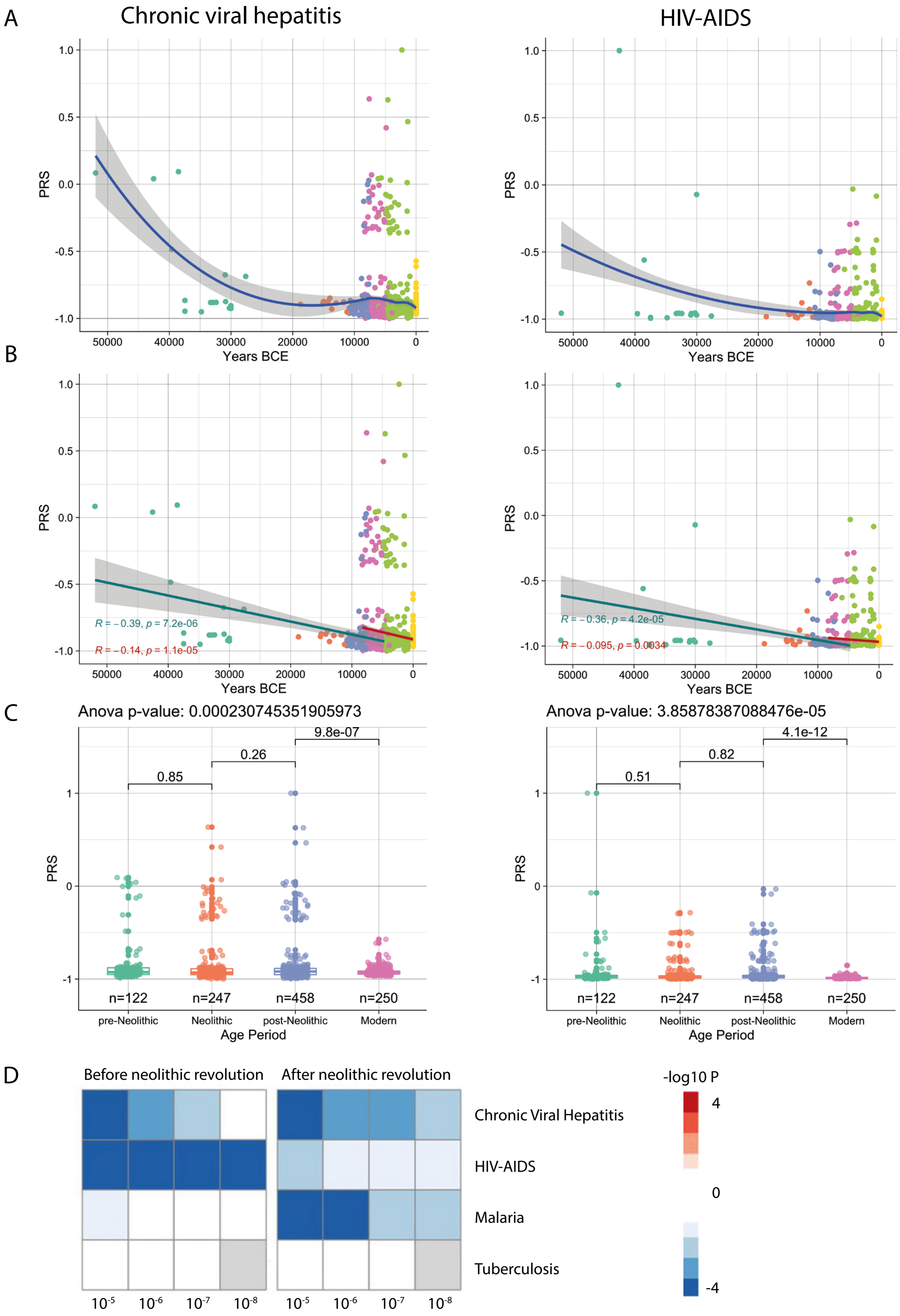
These evolutionary genetic adaptations to pathogens throughout human history greatly influence the way we respond to multiple diseases in modern times as well. To assess these effects, we calculated the PRS associated with the risk of several highly prevalent immune-mediated diseases. The first focus was on common infectious diseases such as malaria, human immunodeficiency virus-acquired immunodeficiency syndrome (HIV-AIDS), tuberculosis, and chronic viral hepatitis; we calculated the changes in susceptibility to these diseases in the last 50,000 years of human history, based on summary statistics from genome-wide association studies (GWAS) databases available from the literature (**Fig. 2**—**Suppl. Fig. 1**). Our results show that humans are becoming more resistant to these diseases, with the notable exception of tuberculosis, whose risk score remained stable along the period studied (**Fig. 2**). Of note, the QTLs that passed our thresholds were scarce, resulting in a PRS model using a limited number of SNPs, making them sensitive to changes in the p-value cut-off, especially in traits related with infectious diseases. Our results suggest that humans have built up a genetic makeup that made them more resistant to a variety of microbes. The pattern of this adaptation is very interesting as well, with a suggested decrease of susceptibility to malaria especially in the last 10,000 years. The reason for this accelerated resistance after Neolithic might be linked to a higher disease prevalence due to increased population density, as otherwise Plasmodium parasites are known to have circulated in Africa since at least the Paleogene [30] million years ago [19], and we have likely inherited it from gorillas [20]. Intriguingly, we also observed a strong decrease in susceptibility to HIV: this is a contemporary pathogen, therefore this signal could be due to common genetic and immune pathways with other infections that were present in human populations. The increased resistance to HIV in Europeans may be derived from selective pressures induced by other pathogens such as Yersinia pestis [21]. Our data suggest, on the other hand, that the source of this increased resistance is even older.

In contrast, the lack of genetic adaptation in susceptibility to tuberculosis is intriguing. This surprising finding may be explained by a concept in which *Mycobacterium tuberculosis* is at the same time a pathogen and a symbiont, in which latent infection enhances the resistance against other pathogens, and this is why our immune system tolerates mycobacterial presence [22]. In this regard, individuals with latent tuberculosis exhibit enhanced macrophage functions that may protect against other pathogens through the induction of trained immunity [23]. In this context, humanity may not be adapting to tuberculosis because increased resistance against mycobacteria is not evolutionarily advantageous. All in all, these results suggest that the risk of suffering infectious diseases has steadily decreased at least for the last 50,000 years as a result of the selection of genetic variants that confer resistance to infections.

It has been proposed that the increased prevalence of inflammatory and autoimmune diseases is associated with the immune-related alleles that have been positively selected through evolutionary processes to protect against infection; hence, the contrasting differences in the prevalence of autoimmune diseases between populations result from diverse selective pressures [24]. In line with this, it has been hypothesized that genetic variants associated with protection against infectious agents are behind the increased prevalence of autoimmune diseases in populations with low pathogen exposure, such as Europeans [25,26]. To study the changing patterns of susceptibility to autoimmune and inflammatory diseases during history, we used publicly available summary statistics from GWAS of digestive tract-related autoimmune and inflammatory diseases and arthritis-related diseases (see **Fig. 2**—**Suppl. Fig. 1**) and calculated the PRS for each of the samples under study. Interestingly, we observed a robust increase of the genetic variants related with the development of inflammatory diseases in the digestive tract after the Neolithic revolution (**Fig. 3**). PRS scores associated with coeliac disease, Crohn's disease, ulcerative colitis, and inflammatory bowel disease were strongly associated with the age of the samples, regardless of the p-value thresholds or the missing genotype rates used for PRS calculation, showing the robustness of these results (see **Fig. 1**—**Suppl. Fig. 2**). The fact that especially intestinal inflammatory pathology is increased after a historical event that fundamentally modified human diet is unlikely to be an accident. Our results are in line with earlier research demonstrating that variants in genes important for immune responses and involved in coeliac disease pathophysiology (such as *IL12, IL18RAP, SH2B3*) are under strong positive selection [27]. The reasons for the selection pressure on these genes are not completely understood, but an advantage for host defence has been suggested [27].

In contrast to intestinal inflammation, the PRS of traits linked with juvenile-idiopathic arthritis, rheumatoid arthritis, and multiple sclerosis shows a decrease in genetic susceptibility with the age of the sample after the Neolithic revolution. For pre-Neolithic periods, these patterns had little impact with decreasing PRS for digestive tract diseases and increasing PRS for ankylosing spondylitis and juvenile idiopathic arthritis. A strong decrease in susceptibility to juvenile idiopathic arthritis, rheumatoid arthritis, and multiple sclerosis is seen after the Neolithic period (see **Fig. 3**). This is likely linked to the decreased production of the IL-1/IL-6/IL-17 axis described in **Fig. 2**, which is particularly important in the pathophysiology of these disorders [28,29].

The significant changes in cytokine production and disease susceptibility in European populations after the Neolithic can be due to selective processes on the one hand (as described above), but also due to important demographic changes due to migrations of human communities such as the Anatolians (in Neolithic) or the Yamnaya populations from the Pontic steppe (during the Bronze Age) [30]. In this regard, several loci associated with inflammatory diseases displayed a group of alleles linked with Crohn's disease, coeliac disease, and ulcerative colitis in Neolithic Aegeans, the community that spread farming across Europe [31], with several of these alleles showing signs of positive selection in modern Europeans [26]. In addition, the gene expression PRS of several cytokines based on the *cis* and *trans* [gene] expression quantitative trait locus (eQTLs) from the


A
Chronic viral hepatitis
HIV-AIDS
PRS
1.0
0.5
0.0
-0.5
-1.0
50000
40000
30000
20000
10000
0
Years BCE
B
R = −0.39, p = 7.2e-06
R = −0.14, p = 1.1e-05
R = −0.36, p = 4.2e-05
R = −0.095, p = 0.0034
C
Anova p-value: 0.000230745351905973
Anova p-value: 3.85878387088476e-05
0.85
0.26
9.8e-07
0.51
0.82
4.1e-12
1
0
-1
n=122
n=247
n=458
n=250
pre-Neolithic
Neolithic
post-Neolithic
Modern
Age Period
D
Before neolithic revolution
After neolithic revolution
Chronic Viral Hepatitis
HIV-AIDS
Malaria
Tuberculosis
10^{-5}
10^{-6}
10^{-7}
10^{-8}
-log10 P
4
0
-4

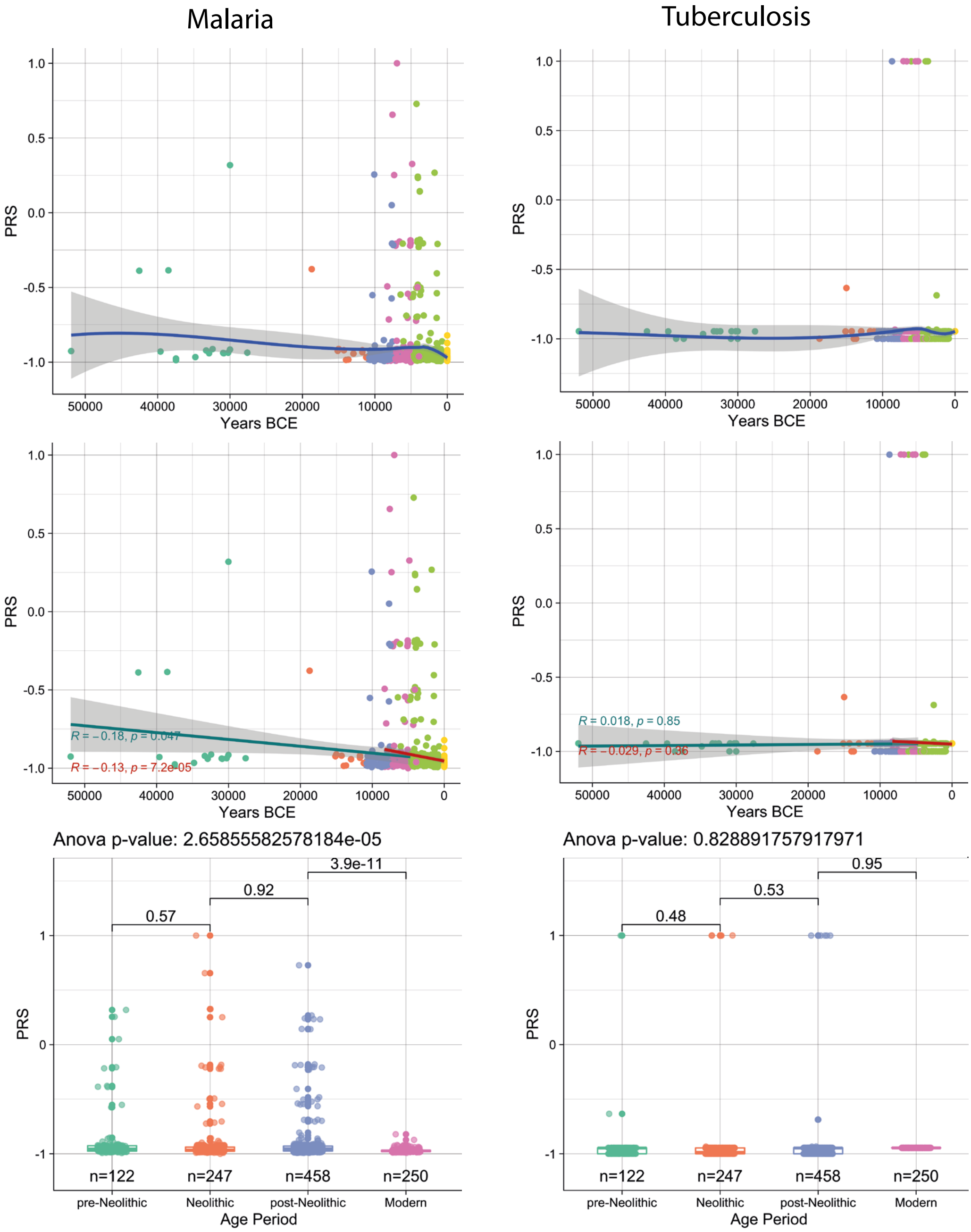
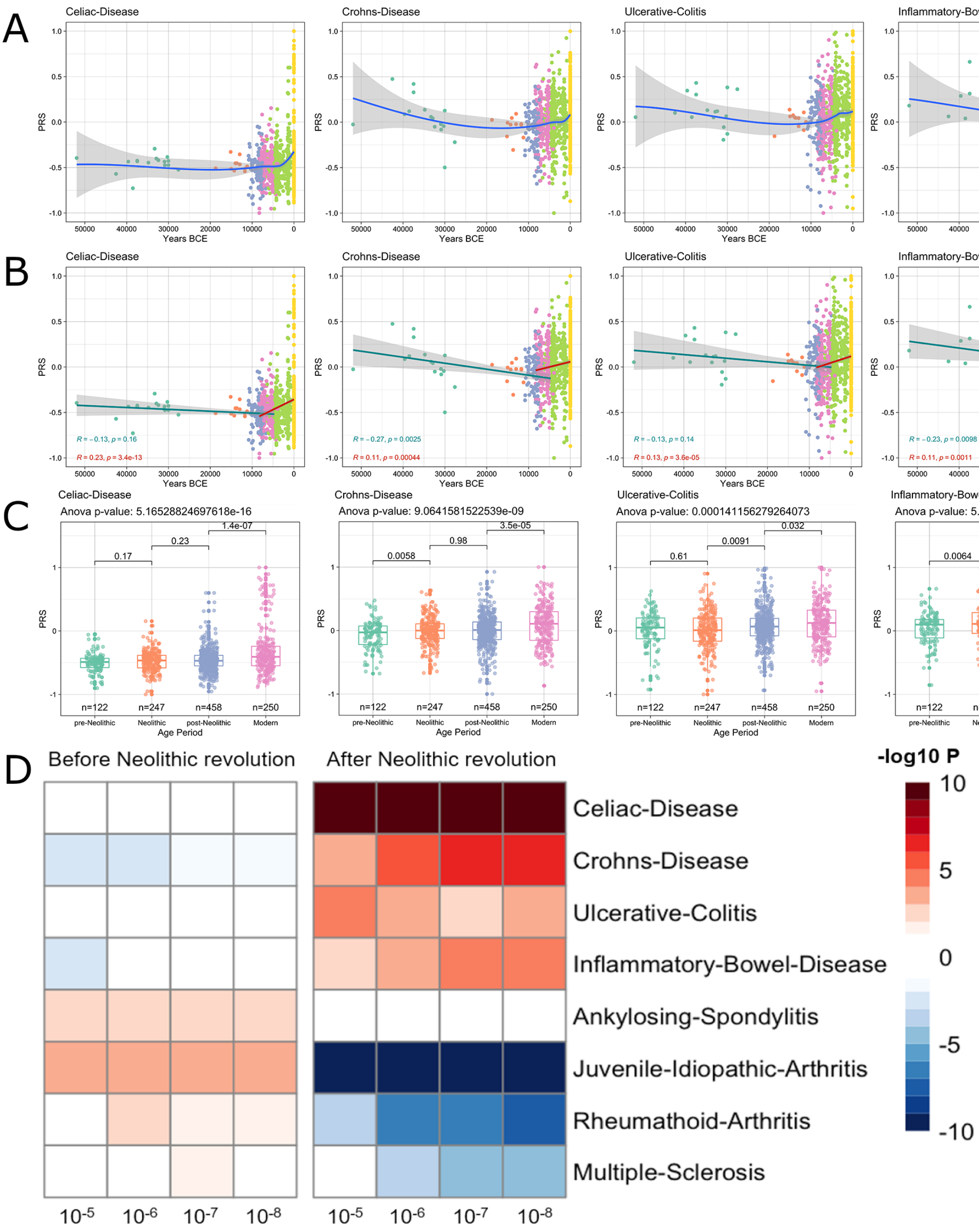


Fig. 2. Distribution of PRS of infectious diseases across time. A–C) Polygenic risk score (PRS) models based on a p-value threshold of 1x10^{-5}. Gray area around regression lines represents the 95% confidence interval. **A)** LOESS regression models showing the changing differences in PRS across time. **B)** Dual linear models showing the difference in trends before and after the Neolithic revolution. **C)** Boxplots showing the difference in mean PRS between adjacent broad time periods using a t-test and the overall difference in means using analysis of variance (ANOVA). **D)** Heatmap showing the consistency in regression trends before and after the Neolithic revolution using different quantitative trait locus (QTL) inclusion thresholds (1x10^{-5} to 1x10^{-8}).


A
Celiac-Disease
Crohns-Disease
Ulcerative-Colitis
Inflammatory-Bowel-Disease
PRS
Years BCE
50000
40000
30000
20000
10000
0
1.0
0.5
0.0
-0.5
-1.0
B
R = −0.13, p = 0.16
R = 0.23, p = 3.4e-13
R = −0.27, p = 0.0025
R = 0.11, p = 0.00044
R = −0.13, p = 0.14
R = 0.13, p = 3.6e-05
R = −0.23, p = 0.0098
R = 0.11, p = 0.0011
C
Anova p-value: 5.16528824697618e-16
0.17
0.23
1.4e-07
Anova p-value: 9.0641581522539e-09
0.0058
0.98
3.5e-05
Anova p-value: 0.000141156279264073
0.61
0.0091
0.032
Anova p-value: 5.89416386837156e-08
0.0064
0.8
2.2e-05
n=122
n=247
n=458
n=250
pre-Neolithic
Neolithic
post-Neolithic
Modern
Age Period
D
Before Neolithic revolution
After Neolithic revolution
-log10 P
10
5
0
-5
-10
Celiac-Disease
Crohns-Disease
Ulcerative-Colitis
Inflammatory-Bowel-Disease
Ankylosing-Spondylitis
Juvenile-Idiopathic-Arthritis
Rheumathoid-Arthritis
Multiple-Sclerosis
10-5
10-6
10-7
10-8

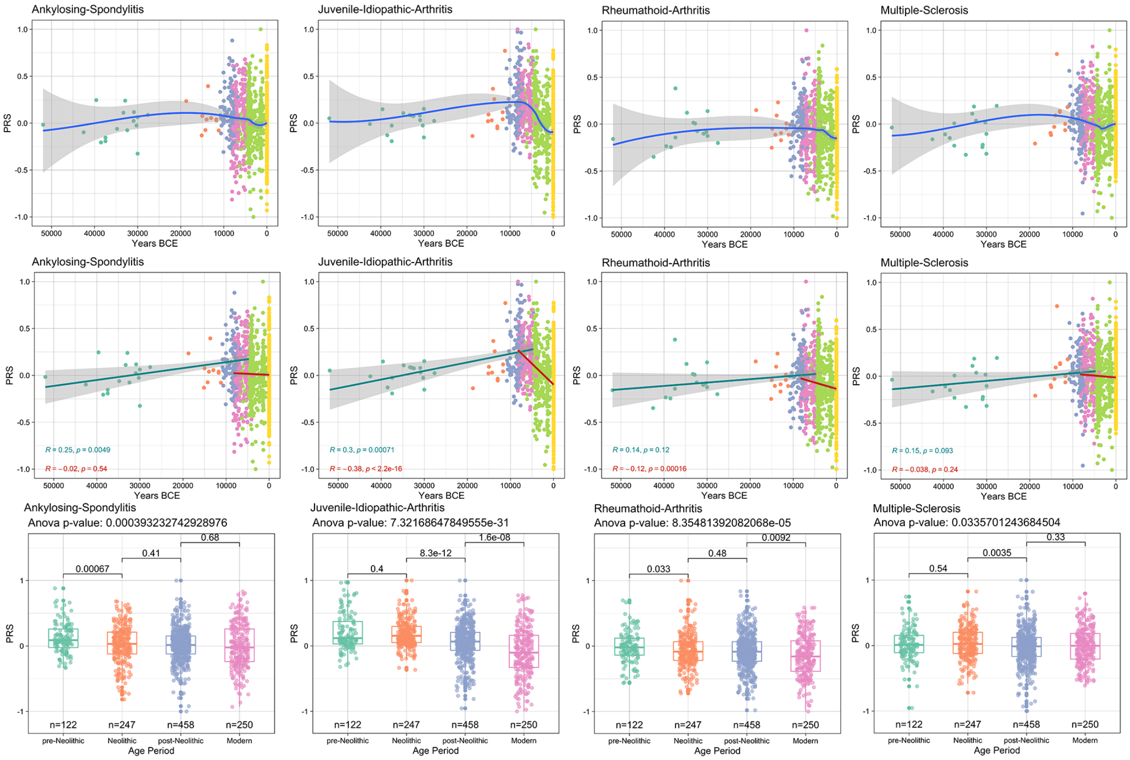


Fig. 3. Distribution of PRS of inflammatory diseases across time. A–C) Polygenic risk score (PRS) models based on a p-value threshold of 1x10^{-5}. Gray area around regression lines represents the 95% confidence interval. **A)** LOESS regression models showing the changing differences in PRS across time. **B)** Dual linear models showing the difference in trends before and after the Neolithic revolution. **C)** Boxplots showing the difference in mean PRS between adjacent broad time periods using a t-test and the overall difference in means using analysis of variance (ANOVA). **D)** Heatmap showing the consistency in regression trends before and after the Neolithic revolution using different quantitative trait locus (QTL) inclusion thresholds (1x10^{-5} to 1x10^{-8}).

eQTLGen Consortium (https://www.eqtlgen.org/) displayed a very strong association with time for TNF-α after the Neolithic revolution (**Fig. 4**).

Of note, the availability of samples in the pre-Neolithic dataset is limited compared to the post-Neolithic era. To test the robustness of the trends observed, we recalculated the correlation coefficients and p-values to show the consistency of the results from the down-sampled data of the post-Neolithic samples to match the sample size of the pre-Neolithic samples. This analysis showed that the trajectories observed in our results are consistent regardless of the sample size (**Fig. 4**— **Suppl. Fig. 1**).

In this study we focus on the effects of different pathogens on cytokine production and, through that, on the response to different infections. Human evolution has been influenced by multiple factors such as migration, urbanization, climate, and diet, which also influence the responses to pathogens. The advent of the Neolithic lifestyle was a milestone for human societies: after the Neolithic, the density of human populations and the rate of contact with domesticated animals increased significantly, which increased the emergence of pathogens and the ease with which those could spread. Our data show that while various events and conditions throughout human history have influenced our immune responses to pathogens, the advent of the Neolithic lifestyle was an important turning point for our capacity to respond to pathogens.

Collectively, our results show that the advent of the Neolithic era was a turning point for the evolution of immune-mediated traits in European populations, driving the expansion of alleles that favour the development of tolerance against intracellular pathogens and promote inflammatory responses against extracellular microbes. This is associated with a higher presence of genetic traits related with inflammatory and auto-immune diseases of the digestive tract and a lower number of alleles linked with the development of arthritis. It is important to underline that our results are obtained using European aDNA samples and GWAS summary statistics computed using European cohorts. Therefore, our results would be specific to European populations and should be interpreted with caution for populations from other geographical locations. Further research should compare the trends in different populations that have been exposed to different environments across the planet and clarify the influence of ancestry, time, and rural vs urban lifestyle to shed light on the influence of the infectious environment on genetics and human evolution.

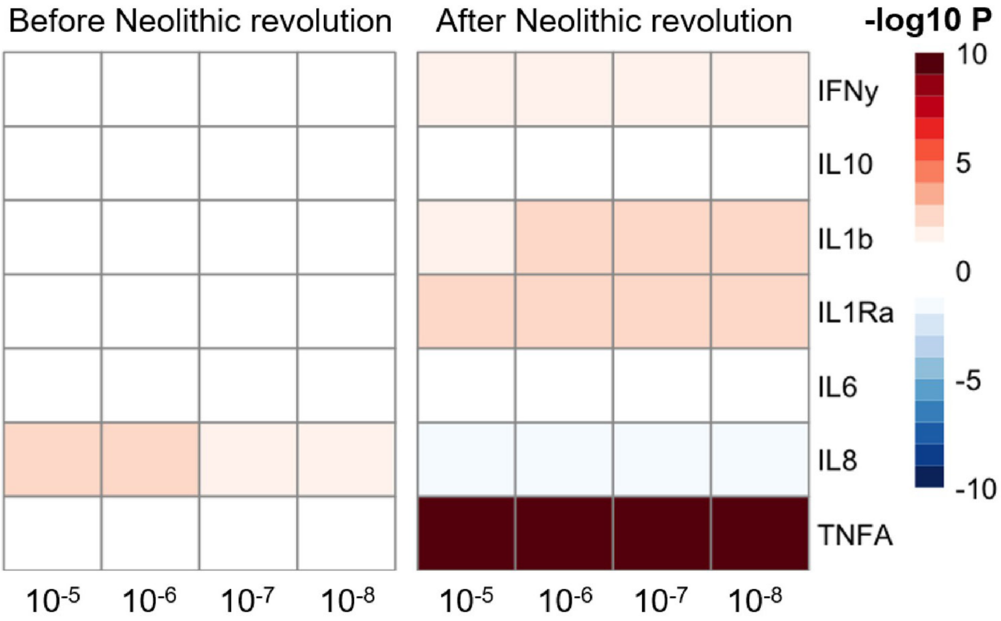


Fig. 4. Changes in cytokine gene expression PRS across time before and after the Neolithic revolution using different QTL thresholds ($1x10^{-5}$ to $1x10^{-8}$). Missing genotype rates ranged through 0.96, 0.9, 0.8, and 0.7. Quantitative trait locus (QTL) p-values for variants included in our polygenic risk score (PRS) models ranged through 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, and 10^{-8}. The color key indicates the range of -log10 p-values of the Pearson correlation between PRS and time. Red and blue indicate positive and negative association, respectively.

Materials and methods

Cohort selection

Ancient DNA genotype data were downloaded from version 37.2 of the published aDNA genotype database, compiled by and available on the David Reich Lab website (https://reich.hms.harvard.edu/ downloadable-genotypes-present-day-and-ancient-dna-data-compiled-published-papers). The aDNA samples consisted of pseudo-haploid genotype data. This was due to the low genotyping coverage. Samples with variant missingness above 96% were filtered out using Plink [32]. This was done in order to remove outliers with extremely low coverage. Only samples within Europe were used for this study, and these samples were selected based on their geographic location, that is latitude (within 35 and 70 degrees north) and longitude (within 10 degrees west and 40 degrees east). Samples without a carbon-dated age were also filtered out. We also selected 250 European samples from the 1000 genomes project phase 3. Only variants present in both the ancient samples and the modern samples were retained. This resulted in a dataset of 827 ancient samples and 250 modern samples containing 1,233,013 variants.

Carbon-dated sample origin and geographical location

Both carbon-dated age of origin and latitudinal and longitudinal data were available for these 827 ancient European samples. Broad time periods were assigned to these samples with the early upper Paleolithic era for all samples originating from before 25,000 years before the common era standardized to 1950 (BCE). The late upper Paleolithic era follows until 11,000 BCE. The Mesolithic era ranges from 11,000 to 5500 BCE. The Neolithic era ranges from 8500 to 3900 BCE, and the postNeolithic era ranges from 5000 BCE and more recent ages. Using the geographical data in combination with archeological clues and the genetic data, the broad time period of origin was also available for samples that were dated to a point in time with overlapping broad time periods. This allowed the samples to be classified as either early upper Paleolithic, late upper Paleolithic, Mesolithic, Neolithic, or post-Neolithic. The sample age of the 250 modern European samples was set to 0.

Summary statistics of GWAS and cytokine QTLs

Summary statistics for complex traits were obtained from the UK Biobank [33] and the GWAS catalog [34] last accessed on March 29, 2020. The stimulated cytokine response summary statistics from the 500FG cohort of the HFGP were used [6]. Some complex traits had multiple different sets of summary statistics available. In these cases, the data, which were more recent and used bigger cohorts that were either of European or mixed (European and Asian) ancestry, were selected. The variants of these summary statistics were then filtered by only keeping bi-allelic variants. Most aDNA genotypes available are pseudo-haploid as a consequence of their lower sample quality. We excluded ambiguous SNPs (A/T and C/G) in order to prevent errors due to strand flips present in these pseudo-haploid samples.

Polygenic risk score calculation

Polygenic risk scores were then calculated by first intersecting the filtered variants from the summary statistics with the variants present in the DNA samples. Starting at the most significant variant, all variants within a 250kb window around that variant were excluded until no variants remained. We then multiplied the dosage of these variants with the effect size and these values were summed. If a variant is missing in a sample, the dosage is substituted with the average genotyped dosage for that variant within the entire dataset. This way the PRS is not skewed in any specific direction. The formula for this is described below with the score S being the weighted sum of a variant's dosage X_n multiplied by its associated weight or β_n calculated using m variants.

Relation between PRS and carbon-dated sample age

We constructed piecewise linear models for each trait by separating the samples into two groups. These two groups consisted of all samples preceding the Neolithic era and those of the Neolithic era and later, respectively. We correlated PRS with the carbon-dated age of our samples. We then multiplied the -log10 of the Pearson correlation p-values with the sign of the correlation coefficients. In addition, we plotted LOESS regression models to highlight the change in PRS at each point in time independent of any predefined breakpoint between historical periods. We also performed a group-based comparison using Student's t-test. We compared pre-Neolithic, Neolithic, post-Neolithic, and Mesolithic samples with their respective adjacent historical periods to show the difference between other historical transitions besides the pre- and post-Neolithic.

Selection test

We tested whether traits were under selection or if observed changes were due to the genetic drift between adjacent time periods. We performed a two-tailed test using the mean Fst calculated with trait-specific SNPs between two adjacent periods and a reference distribution of 10,000 random linkage disequilibrium (LD) and minor allele frequency (MAF)-matched mean Fst scores calculated using an equal amount of SNPs . Bonferroni correction was performed to account for multiple testing.

Robustness of results

In order to test the robustness of our results we calculated PRS using multiple different p-value thresholds for QTL inclusion. We used p-value thresholds from 10^{-3} to 10^{-8} for the complex traits obtained through GWAS catalog and the UK Biobank. The thresholds used for the stimulated cytokine responses ranged from 10^{-3} to 10^{-6}. We also calculated PRS using different variant missingness thresholds. This means we removed samples with a variant missingness rate higher than 96, 90, 80, or 70%. All of the results from the piecewise linear models were then used to create a heatmap depicting the consistency and robustness of our observed correlations.

Additionally, various window sizes were used for clumping the QTLs, and LD-based clumping was also performed excluding variants with an LD greater than 0.2 compared to our lead SNP within a window. In order to see whether our observations were due to sample imbalances between the pre Neolithic period and the later periods, samples originating from the Neolithic period and later were randomly down-sampled to the same number of samples as the pre-Neolithic samples. Correlation coefficients between PRS and sample age were then recalculated for the Neolithic and younger samples and compared to the coefficients obtained using all Neolithic and younger samples.

Acknowledgements

MGN was supported by an ERC Advanced Grant (833247) and a Spinoza Grant of the Netherlands Organization for Scientific Research. YL was supported by an ERC Starting Grant (948207) and the Radboud University Medical Centre Hypatia Grant (2018) for Scientific Research. JD-A is supported by The Netherlands Organization for Scientific Research (VENI grant 09150161910024). JB was supported by PID2019-110933GB-I00 (AEI/FEDER, UE) MINECO, Spain.

Competing interests

The other authors declare that no competing interests exist.

Author contributions

Jorge Domınguez-Andre´s, Conceptualization, Investigation, Writing original draft;

Yunus Kuijpers, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing original draft;

Olivier B Bakker, Resources, Formal analysis, Investigation;

Martin Jaeger, Resources, Writing review and editing;

Cheng-Jian Xu, Resources, Methodology, Writing review and editing;

Jos WM Van der Meer, Supervision, Writing review and editing;

Mattias Jakobsson, Resources, Methodology;

Jaume Bertranpetit, Resources, Supervision, Writing review and editing;

Leo AB Joosten, Resources, Supervision;

Yang Li, Mihai G Netea, Conceptualization, Resources, Supervision, Funding acquisition, Project administration, Writing review and editing

Data availability

All the data employed in this manuscript have been obtained from publicly available databases.

Supplementary material

Supplemenatary material are provided at: https://doi.org/10.7554/eLife.64971

References

1. Fumagalli, M. & Sironi, M. Human genome variability, natural selection and infectious diseases. Curr. Opin. Immunol. 30, 9–16 (2014).
2. Karlsson, E. K., Kwiatkowski, D. P. & Sabeti, P. C. Natural selection and infectious disease in human populations. Nat. Rev. Genet. 15, 379–393 (2014).
3. Quintana-Murci, L. & Clark, A. G. Population genetic tools for dissecting innate immunity in humans. Nat. Rev. Immunol. 13, 280–293 (2013).
4. Netea, M. G. et al. Understanding human immune function using the resources from the Human Functional Genomics Project. Nat. Med. 22, 831–833 (2016).
5. Bakker, O. B. et al. Integration of multi-omics data and deep phenotyping enables prediction of cytokine responses. Nat. Immunol. 19, 776–786 (2018).
6. Li, Y. et al. A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans. Cell 167, 1099-1110.e14 (2016).
7. Schirmer, M. et al. Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. Cell 167, 1125-1136.e8 (2016).
8. ter Horst, R. et al. Host and Environmental Factors Influencing Individual Human Cytokine Responses. Cell 167, 1111-1124.e13 (2016).
9. van de Veerdonk, F. L. et al. The Macrophage Mannose Receptor Induces IL-17 in Response to Candida albicans. Cell Host Microbe 5, 329–340 (2009).
10. Horst, R. ter et al. Seasonal and Nonseasonal Longitudinal Variation of Immune Function. J. Immunol. 207, 696–708 (2021).
11. Flandroy, L. et al. The impact of human activities and lifestyles on the interlinked microbiota and health of humans and of ecosystems. Sci. Total Environ. 627, 1018–1038 (2018).
12. Deschamps, M. et al. Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. Am. J. Hum. Genet. 98, 5–21 (2016).
13. Ferwerda, B. et al. TLR4 polymorphisms, infectious diseases, and evolutionary pressure during migration of modern humans. Proc. Natl. Acad. Sci. 104, 16645–16650 (2007).
14. Thakur, A., Mikkelsen, H. & Jungersen, G. Intracellular Pathogens: Host Immunity and Microbial Persistence Strategies. J. Immunol. Res. 2019, e1356540 (2019).
15. Chan, L. C. et al. Protective immunity in recurrent Staphylococcus aureus infection reflects localized immune signatures and macrophage-conferred memory. Proc. Natl. Acad. Sci. 115, E11111–E11119 (2018).
16. Domínguez-Andrés, J. et al. Inflammatory Ly6Chigh Monocytes Protect against Candidiasis through IL-15-Driven NK Cell/Neutrophil Activation. Immunity 46, 1059-1072.e4 (2017).
17. Tay, M. Z., Poh, C. M., Rénia, L., MacAry, P. A. & Ng, L. F. P. The trinity of COVID-19: immunity, inflammation and intervention. Nat. Rev. Immunol. 20, 363–374 (2020).
18. Diard, M. & Hardt, W.-D. Evolution of bacterial virulence. FEMS Microbiol. Rev. 41, 679–697 (2017).
19. Poinar, G. Plasmodium dominicana n. sp. (Plasmodiidae: Haemospororida) from Tertiary Dominican amber. Syst. Parasitol. 61, 47–52 (2005).
20. Liu, W. et al. Origin of the human malaria parasite Plasmodium falciparum in gorillas. Nature 467, 420–425 (2010).
21. Duncan, S. R., Scott, S. & Duncan, C. J. Reappraisal of the historical selective pressures for the CCR5-Δ32 mutation. J. Med. Genet. 42, 205–208 (2005).
22. Pai, M. et al. Tuberculosis. Nat. Rev. Dis. Primer 2, 1–23 (2016).
23. Joosten, S. A. et al. Mycobacterial growth inhibition is associated with trained innate immunity. J. Clin. Invest. 128, 1837–1851 (2018).
24. Ramos, P. S., Shedlock, A. M. & Langefeld, C. D. Genetics of autoimmune diseases: insights from population genetics. J. Hum. Genet. 60, 657–664 (2015).
25. Fumagalli, M. et al. Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. PLOS Genet. 7, e1002355 (2011).
26. Raj, T. et al. Common Risk Alleles for Inflammatory Diseases Are Targets of Recent Positive Selection. Am. J. Hum. Genet. 92, 517–529 (2013).
27. Zhernakova, A. et al. Evolutionary and Functional Analysis of Coeliac Risk Loci Reveals SH2B3 as a Protective Factor against Bacterial Infection. Am. J. Hum. Genet. 86, 970–977 (2010).
28. Akioka, S. Interleukin-6 in juvenile idiopathic arthritis. Mod. Rheumatol. 29, 275–286 (2019).
29. Mei, Y. et al. Increased serum IL-17 and IL-23 in the patient with ankylosing spondylitis. Clin. Rheumatol. 30, 269–273 (2011).
30. Racimo, F. et al. The spatiotemporal spread of human migrations during the European Holocene. Proc. Natl. Acad. Sci. 117, 8989–9000 (2020).
31. Hofmanová, Z. et al. Early farmers from across Europe directly descended from Neolithic Aegeans. Proc. Natl. Acad. Sci. 113, 6886–6891 (2016).
32. 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).
33. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209 (2018).
34. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 45, D896–D901 (2017).

ATCG
UMCG
TƆЯ
HLA
ごめんなさい

Part III

The genetics of CeD in different contexts

Chapter 8

Potential impact of coeliac disease genetic risk factors on T cell receptor signaling in gluten-specific CD4+ T cells

Olivier B. Bakker [1], Aarón D. Ramírez-Sánchez [1*], Zuzanna A. Borek [1*], Yang Li [1], Niek de Klein [1], Rutger Modderman [1], Yvonne Kooy-Winkelaar [2], Marie K. Johannesen [3,4], Filomena Matarese [5], Joost H.A. Martens [5], Vinod Kumar [1], Jeroen van Bergen [2], Shuo-Wang Qiao [3,4], Knut E.A. Lundin [3,6], Ludvig M. Sollid [3,4], Frits Koning [2], Cisca Wijmenga [1,3], Sebo Withoff [1] and Iris H. Jonkers [1]*

Scientific Reports (2021)

1 University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands
2 Department of Immunohematology and Blood Transfusion, Leiden University Medical Center, the Netherlands
3 K.G. Jebsen Coeliac Disease Research Centre, Institute of Clinical Medicine, University of Oslo, Oslo, Norway
4 Department of Immunology, Oslo University Hospital, Rikshospitalet, Oslo, Norway
5 Department of Molecular Biology, Nijmegen Centre for Molecular Life Sciences, Radboud University, Nijmegen, the Netherlands
6 Department of Gastroenterology, Oslo University Hospital, Rikshospitalet, Oslo, Norway
** These authors contributed equally*

Abstract

Coeliac disease is an auto-immune disease in which an immune response to dietary gluten leads to inflammation and subsequent atrophy of small intestinal villi, causing severe bowel discomfort and malabsorption of nutrients. The major instigating factor for the immune response in coeliac disease is the activation of gluten-specific CD4+ T cells expressing T cell receptors that recognize gluten peptides presented in the context of HLA-DQ2 and DQ8. Here we provide an in-depth characterization of 28 gluten-specific T cell clones. We assess their transcriptional and epigenetic response to T cell receptor stimulation and link this to genetic factors associated with coeliac disease. Gluten-specific T cells have a distinct transcriptional profile that mostly resembles that of Th1 cells but also express cytokines characteristic of other types of T helper cells. This transcriptional response appears not to be regulated by changes in chromatin state, but rather by early upregulation of transcription factors and non-coding RNAs that likely orchestrate the subsequent activation of genes that play a role in immune pathways. Finally, the lack of changes in chromatin profile and the dynamic transcription factor expression profiles suggest that genes activated by T cell receptor stimulation of gluten-specific T cells may be impacted by genetic variation at several genetic loci associated with coeliac disease.

Introduction

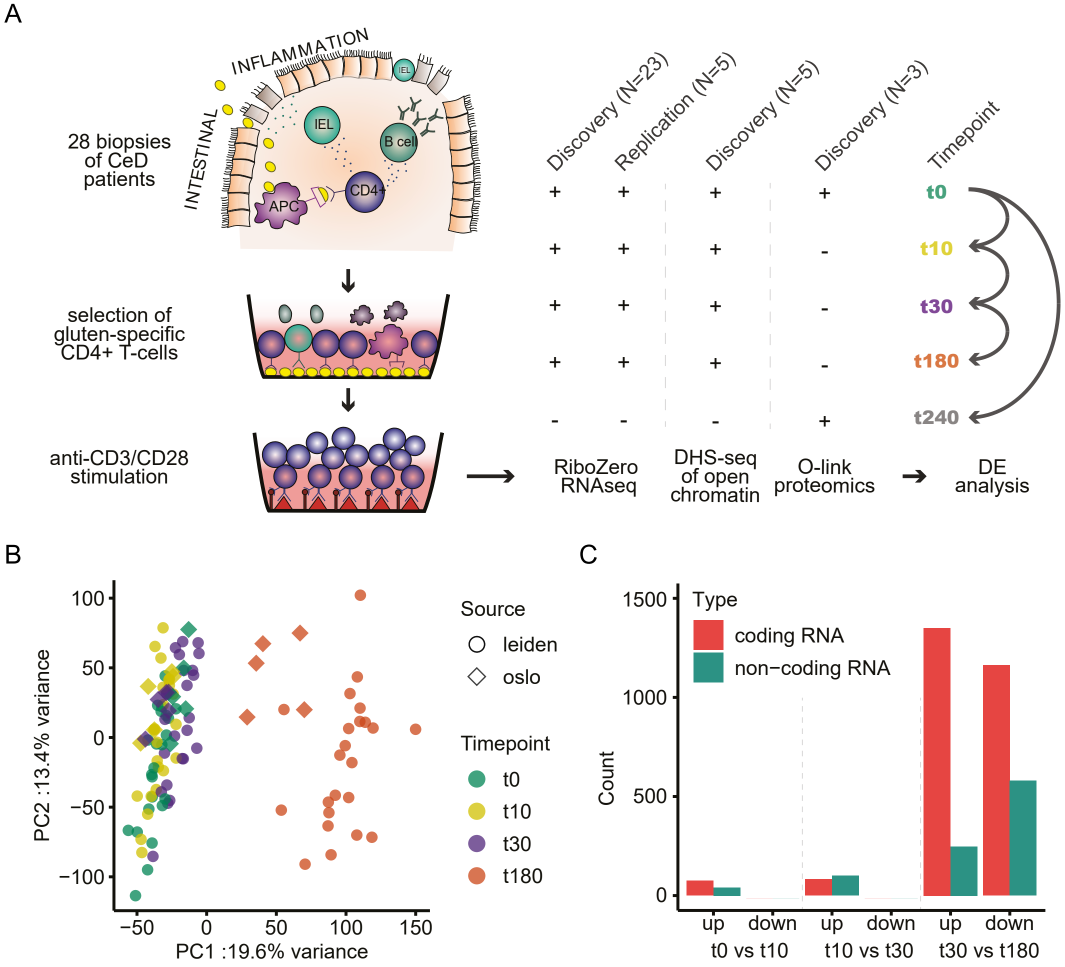
In coeliac disease (CeD), cereal-derived gluten peptides penetrate the small intestinal barrier, are subsequently modified by tissue *trans*-glutaminase 2 (TG2), then presented by HLA-DQ2- or HLA-DQ8-positive antigen-presenting cells to gluten-specific CD4+ T helper cells (gsTcells) [1]. This leads to robust activation of gsTcells that subsequently stimulate B cells to start producing auto-antibodies to TG2 and deamidated gluten peptides [1,2] and activate CD8+ intraepithelial lymphocytes (IELs) to attack intestinal epithelial cells, leading to the villous atrophy that is characteristic of CeD. GsTcells are only found persistently in CeD patients [3,4] and can induce villous atrophy in patients upon gluten ingestion even after these individuals have been on a gluten-free diet for years [5]. Activation of gsTcells is thus central to CeD onset and pathology.

GsTcells have been shown to secrete many signaling molecules upon stimulation, including interleukin (IL)-2, IL-4, IL-6, IL-8, IL-10, IL-21, CD40LG, IFN-γ and TNF, and are often classified to be of type 1 helper class [6-12]. GsTcells uniquely express IL-21 and CXCL13, as well as several other markers characteristic of follicular and regulatory T cells [13]. IL-21 and CXCL13, together with CD40LG and IL-4, play an important role in the interaction, differentiation and activation of T cells and plasma B cells [13-16]. Cytokines secreted by gsTcells are also important for activation and proliferation of IELs, in combination with IL-15, a cytokine important in CeD etiology that is produced by IELs[17-20]. GsTcells are thus central in the response to gluten peptides that leads to inflammation, anti-TG2 antibody production and villous atrophy in CeD.

To date, 43 genetic risk factors have been associated with CeD [21-23], the most important being the HLA haplotypes HLA-DQ2 and -DQ8. While the role of HLA-DQ2 and -DQ8 in CeD is well defined [24-26], the contribution of the non-HLA CeD risk-loci is mostly unclear. More than 95% of the single nucleotide polymorphisms (SNPs) associated with CeD are located in the non-coding genome and presumably deregulate genes important for CeD etiology [27]. Enrichment analysis of the CeD SNPs in regulatory regions suggests that CD4+ T cells are the major cell type affected by genetic risk factors [28-30]. Moreover, pathway and *cis*-eQTL analyses of genes in CeD loci suggest that they affect T cell receptor (TCR) signaling via alteration of expression of genes such as *UBASH3A*, *CD28* and *CSK* [30-32]. Overall, these observations confirm the importance of CD4+ T cell activation in CeD but do not delineate how CeD-associated SNPs affect gsTcells upon activation.

This knowledge gap is partially due to an incomplete understanding of the regulation of the response to stimulation in gsTcells. Recently, it was shown that the genetic risk loci associated with CeD are enriched with binding sites of specific transcription factors (TFs), including STAT4, STAT5A, STAT5B, T-BET, AP-1 subunit FOS and TFs from the NFκB signaling pathway [29]. Indeed, many of these TFs have been implicated in regulation of CD4+ T cell activation or in CeD [12,33-36]. However, the role of these TFs, as well as the dynamic transcriptional and epigenetic response in the activation of gsTcells, has not been described. Nor has the role of CeD-associated genetic variants in these dynamic transcriptional processes been explored in gsTcells.

Here, we set out to profile the transcriptomic and epigenetic response of gsTcells derived from CeD patients upon TCR-stimulation with anti-CD3 and anti-CD28 (aCD3/aCD28). This allowed us to identify the regulatory steps essential for the rapid and robust activation of cytokines important for CeD etiology and to prioritize which CeD-associated risk loci are related to the activation of gsTcells. Overall, we elucidate the dynamic events in gsTcells that can be induced by gluten peptides in CeD patients.



	Discovery (N=23)	Replication (N=5)	Discovery (N=5)	Discovery (N=3)	Timepoint
	+	+	+	+	t0
	+	+	+	-	t10
	+	+	+	-	t30
	+	+	+	-	t180
	-	-	-	+	t240

Fig. 1: Stimulation of gluten-specific CD4+ T cells with aCD3/aCD28. A) Experimental scheme of the discovery and replication cohort. The '+' symbol indicates that a measurement is available at that timepoint. 28 gsTcell clones were isolated from CeD biopsies, 23 clones were used in discovery analysis and 5 were used for replication and DHS-sequencing. A final 3 clones from the discovery set were used for proteomic analysis. **B)** PCA of the complete expression data of the discovery (circles) and replication (diamonds) cohorts. Each time point is indicated in a different color. **C)** Differentially expressed genes identified by differential expression (DE) analysis between consecutive timepoints, plotted per comparison. Biotypes and direction of each DE gene are indicated.

Results

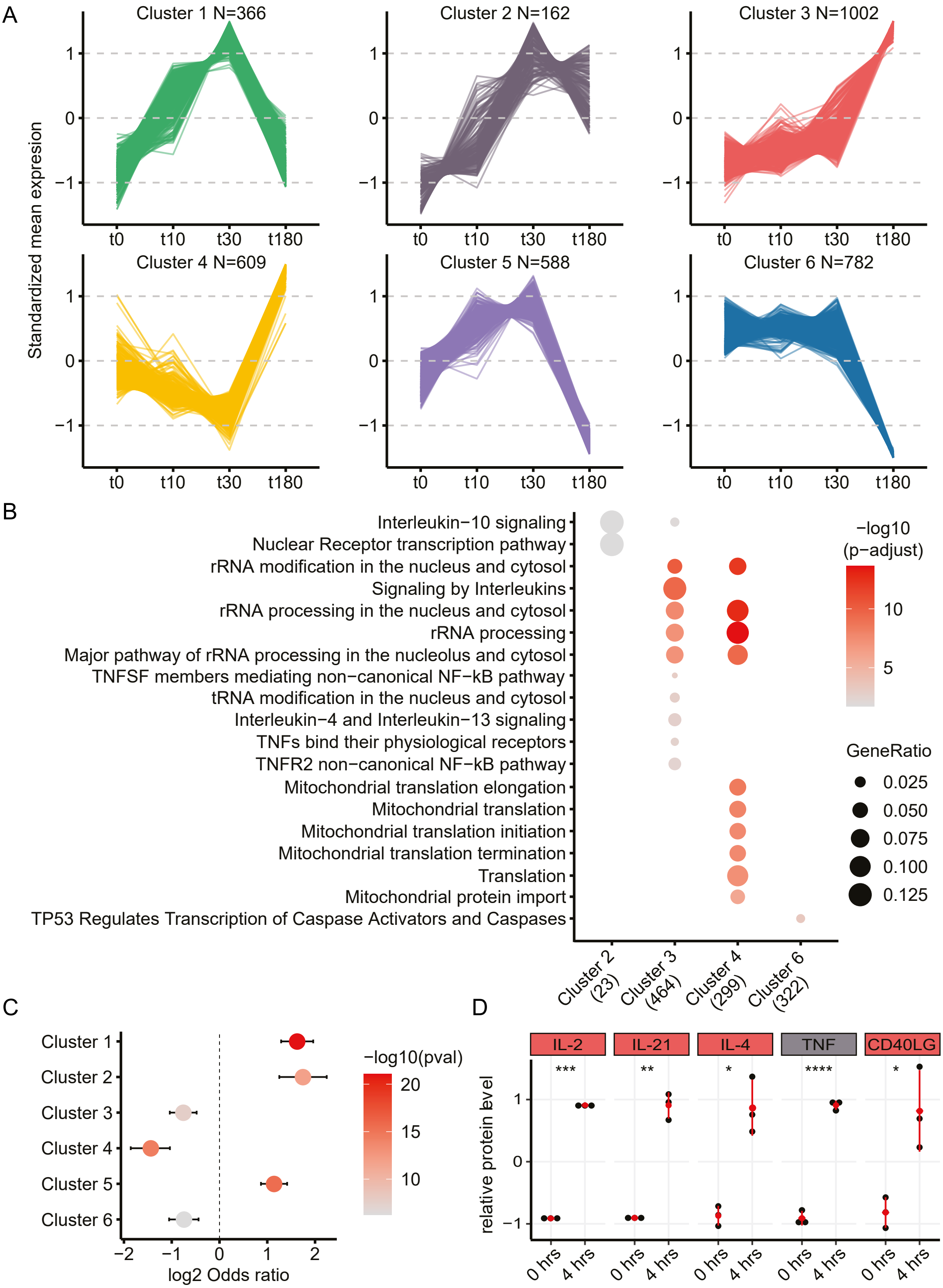
Dynamic transcriptome changes in stimulated gsTcells

To study the TCR-induced response of gsTcells, we opted for aCD3/aCD28 stimulation as a proxy for the interaction of gsTcell-TCR with gluten peptides presented by antigen-presenting cells in the context of HLA-DQ2 or -DQ8. Twenty-three CD4+ gsTcell clones isolated from biopsies from patients with active CeD were cultured and stimulated *in vitro* and used to perform transcriptomic (RNA-Seq; n=23) and targeted proteomics analysis (n=3) (discovery cohort, **Suppl. Tab. 1**). An additional five gsTcell clones were used for replication of the transcriptomic data and detection of open chromatin through DNase-Hyper-Sensitivity sequencing (DHS-seq) (**Fig. 1A**, **Suppl. Tab. 1**). The transcriptomic response of gsTcells to stimulation showed strong and consistent effects after 180 minutes relative to the earlier timepoints. Although there was considerable inter-clonal variation, the replication and discovery cohorts behaved very similarly (**Fig. 1B**). We used the discovery cohort to determine the dynamic transcriptional response of genes during the course of the stimulation, which showed clear distinctions between each time point (**Suppl. Fig. 1**). We performed differential expression (DE) analysis between consecutive timepoints to reveal the changes in gene expression over time. Between 0-10 min, 10-30 min and 30-180 min, 115, 182 and 3339 DE genes were identified, respectively (**Fig. 1C**, **Suppl. Tab. 2**). Finally, non-coding genes were found to be differentially expressed at all timepoints, but at 180 mins the downregulated set was roughly twice as large as the upregulated set, which was not the case for the coding genes (**Fig. 1C**). GsTcell clones thus rapidly displayed strong and dynamic transcriptional changes after stimulation.

DE genes cluster into response patterns with distinct functions

To understand and categorize the activation of gsTcells, we clustered the DE genes using k-means clustering to identify temporal response patterns. This identified six distinct clusters that each represent a specific response (**Fig. 2A**, **Suppl. Fig. 2**, **Suppl. Tab. 2**). Genes in Clusters 1 (n=366) and 2 (n=162) were upregulated early (at 10 min). In contrast, cluster 3 genes (n=1002) were upregulated after 30 and 180 min. Cluster 4 (n=609) genes displayed an early decrease but recovered after 180 min. Cluster 5 (n=588) genes responded similarly to cluster 1 genes, but their gene expression levels after 180 min were decreased compared to unstimulated expression levels. Finally, cluster 6 genes (n=782) show a consistent late decrease in expression.

Gene set enrichment analysis using the Reactome database pinpointed enriched gene sets (false discovery rate (FDR) < 0.05) for clusters 2, 3, 4 and 6 (**Fig. 2B**) but found no enrichment in clusters 1 and 5. Cluster 2 is enriched for Nuclear Receptor transcription pathway genes (*NR4A1, NR4A2* and *NR4A3*) and IL-10 signaling genes (*TNF* and *ICAM1*). Cluster 3 is predominantly associated with immune function, showing an enrichment of cytokine signaling pathway genes. Moreover, genes associated with transcriptional and translational processes are also enriched in cluster 3, consistent with an immune response that requires the production and secretion of cytokines and other signaling proteins. Cluster 4 is enriched in translational and mitochondrial response genes that are associated with a shift towards protein production. Finally, the downregulated genes in cluster 6 are enriched in genes involved p53-mediated regulation of caspases. Downregulation of cluster 6 genes may thus decrease apoptosis and cell death and


A
Standardized mean expresion
Cluster 1 N=366
Cluster 2 N=162
Cluster 3 N=1002
Cluster 4 N=609
Cluster 5 N=588
Cluster 6 N=782
1
0
−1
t0
t10
t30
t180
B
Interleukin−10 signaling
Nuclear Receptor transcription pathway
rRNA modification in the nucleus and cytosol
Signaling by Interleukins
rRNA processing in the nucleus and cytosol
rRNA processing
Major pathway of rRNA processing in the nucleolus and cytosol
TNFSF members mediating non−canonical NF−kB pathway
tRNA modification in the nucleus and cytosol
Interleukin−4 and Interleukin−13 signaling
TNFs bind their physiological receptors
TNFR2 non−canonical NF−kB pathway
Mitochondrial translation elongation
Mitochondrial translation
Mitochondrial translation initiation
Mitochondrial translation termination
Translation
Mitochondrial protein import
TP53 Regulates Transcription of Caspase Activators and Caspases
Cluster 2 (23)
Cluster 3 (464)
Cluster 4 (299)
Cluster 6 (322)
−log10 (p−adjust)
10
5
GeneRatio
0.025
0.050
0.075
0.100
0.125
C
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
−2
−1
0
1
2
log2 Odds ratio
−log10(pval)
20
15
10
D
IL-2
IL-21
IL-4
TNF
CD40LG
***
**
*
****
*
relative protein level
1
0
−1
0 hrs
4 hrs

favor proliferation, paving the way for a robust immune response by gsTcells that is mediated by the genes in cluster 3.

To ascertain how these pathways might be regulated, we investigated the relative enrichment of non-coding RNAs (ncRNAs) per cluster. The early-responding clusters 1, 2 and 5 are enriched for ncRNAs, implying that these RNAs play a role in regulating expression of genes at 180 min, when protein-coding genes are enriched (clusters 3, 4 and 6) (**Fig. 2C**). Several genes encoding TFs that mediate early immune and stress responses are also found in clusters 1 and 2. These TFs include all EGR TFs, NR4A1, NR4A2, NR4A3, ATF3, FOS and FOSB, of which the latter two are subunits of AP-1 (**Suppl. Fig. 3A**). Additionally, cluster 3 encompasses *REL* (encoding NFκB subunit c-REL), *NFKB1* (encoding NFκB subunit p50) and the NFκB inhibitory genes *NFKBIA*, *NFkBID* and *NFKIZ*, which suggests an early activation of the NFκB pathway and a subsequent feedback loop after 180 min (**Suppl. Fig. 3B**)[37,38]. Thus, the action of several TFs that are either activated or transcribed soon after stimulation, possibly in conjunction with ncRNAs, seems to mediate the strong response of genes after 180 min.

To confirm that transcriptional changes lead to secretion of cytokines, we measured a panel of 92 proteins in the gsTcell culture medium after 4 hours of stimulation. Thirty of these proteins are encoded by genes differentially expressed during the stimulation of gsTcells (**Suppl. Fig. 4A**, **Suppl. Tab. 3**). We found that levels of IL-21, CD40LG, IL-2, IL-4 and TNF were significantly increased (nominal p-value < 0.05) (**Fig. 2D**) in concert with increased expression of their corresponding genes (**Suppl. Fig. 4B**). These cytokines all have pro-inflammatory roles and contribute to activation and proliferation of other cell types, including B cells (IL-21, IL-4 and CD40LG) and other T cells like IELs (IL-2 and TNF)[39,40].

In summary, the distinct and dynamic transcriptional changes we observe represent a robust translational immune response to TCR activation that leads to the secretion of several cytokines within 4 hours.

Transcriptional changes identified in gsTcells upon activation are similar to those in other T cells

To examine the specificity of the gsTcell response, we compared the DE genes of gsTcells (between 0 and 180 min) to those of naïve CD4+ T cells stimulated with aCD3/aCD28 (DICE consortium (n=90))[41]. A large proportion (57%) of the DE effects we observed in gsTcells were also found in the DICE consortium data (96% directional concordance; **Fig. 3A**, **Suppl. Tab. 4-5**). Directionally concordant genes (Q1 and Q3 in **Fig. 3B**) showed an enrichment for genes involved in interleukin signaling and rRNA processing (**Suppl. Fig. 5A**), as we had observed in cluster 3 (**Fig. 2B**), suggesting this to be a

Left: Fig. 2: Differentially expressed genes cluster into response patterns showing distinct functions. A) Cluster analysis identified 6 robust clusters encompassing the response profiles of all 3509 differentially expressed (DE) genes. Numbers of genes per cluster are shown. Y-axis shows the mean expression of the genes at each respective time point, centered to mean zero and scaled to standard deviation one. **B)** Reactome gene set enrichment analysis shows enriched pathways for 4 out of the 6 clusters. **C)** Enrichment analysis by Fisher's exact test of non-coding RNAs per cluster. Significance is shown in shades of red. Log2 odds ratios are plotted on the x-axis and indicate enrichment or depletion. Error bars indicate 0.95 confidence intervals. **D)** Scaled relative protein levels (Olink) in the unstimulated condition and after 4 hrs of stimulation for the three independent experiments (black dots). Box colors above the dotplots indicate the cluster in which the DE genes are found. Mean, minimum and maximum relative protein levels are indicated in red. Nominal significance is indicated by asterisks (*p-value<0.05, **p-value< 0.01, ***p-value < 0.001 and ****p-value<0.0001).

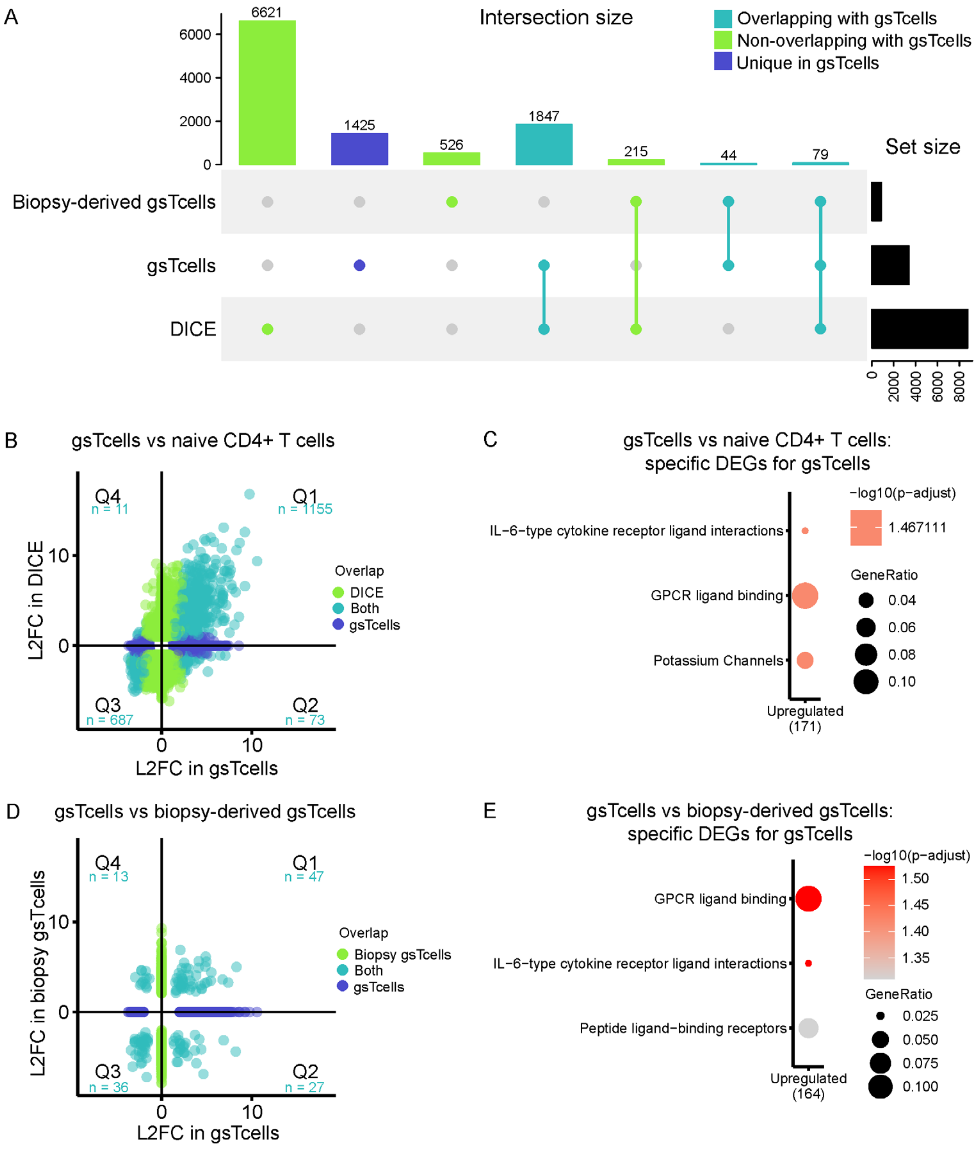


Fig. 3: The transcriptional response of gsTcells shows differences and similarities with other T cells. A) Upset plot comparing significant DE genes between stimulated naïve CD4+ T cells (DICE) (t0 vs t240, FDR < 0.05 and absolute log2FC > 1)[41], gsTcells (t0 vs t180, FDR < 0.05 and absolute log2FC > 1) and biopsy-derived gsTcells (gsTcells vs CD4+ T cells in cases/controls, FDR < 0.05 and absolute log2FC > 2)[13]. At the top, the size of the intersecting sets with gsTcells are indicated in light blue, non-overlapping genes with gsTcells are shown in light green and genes unique to gsTcells are shown in dark blue. The right barplot shows the total number of DE genes per dataset. **B)** Scatterplot of log2FC of DE genes between DICE (y-axis) and gsTcells (x-axis). Numbers in light blue indicate the number of genes in each quadrant that are significant in both analyses. **C)** Gene set enrichment analysis done using Reactome of DE genes unique for gsTcells as compared to DICE naïve CD4+ T cells (adjusted p-value < 0.05, absolute log2FC > 2). At the bottom is the direction of expression of the DE genes in gsTcells. Numbers in brackets indicate the number of DE genes present in all enriched pathways. Dot size indicates the ratio of the number of genes present in the gene set and the total gene set used in each pathway. **D)** Comparison as in (B) for DE effects in CeD biopsy-derived gsTcells (y-axis). **E)** Gene set enrichment analysis as in (C) for genes unique to gsTcells compared to CeD biopsy-derived gsTcells and DICE naïve CD4+ T cells (adjusted p-value < 0.05, absolute log2FC > 2).

general response of CD4+ T cells to stimulation. Genes uniquely activated in gsTcells include chemokines such as *CCL1*, *CXCL1* and *CCL4L1*, which encode for peptides that bind the receptors CCR8, CXCR2 and CCR5 and that can mediate recruitment of immune cells, including Type 2 innate lymphocyte cells, neutrophils and activated CD8+ T cells, respectively (**Fig. 3C**, **Suppl. Fig. 6A**)[42]. In addition, the cytokine-encoding genes *IL5, IL9, IL19, IL17F* and *IL26*, as well as *RORC* (encoding for TF RORγt), are uniquely differentially expressed in gsTcells, albeit at low levels, and each points to a different subset classification of T helper cells for gsTcells [43].

To determine which T helper subset the gsTcells most resemble, we compared the expression patterns of the main cytokines associated with Th1 (*IFNG, TNF*), Th2 (*IL4, IL5, IL13*), Th17 (*IL17A*), Treg (*IL10, TGFB1*), T follicular helper (*IL21*) and Th9 (*IL9*) cells (**Suppl. Fig. 6B**)[6-13,43,44]. We observed that cytokines characteristic for several subsets are strongly upregulated upon stimulation, with Th1 cytokines most strongly expressed. GsTcells are therefore most similar to Th1 cells but also share characteristics with multiple Th subsets.

Overall, the strong concordance between the TCR-induced response profile of naïve CD4+ T cells and gsTcells (**Fig. 3A,B**) suggests that the TCR response is largely shared in CD4+ T cells. However, we also observed 1371 unique DE genes that were not observed in the naïve CD4+ T cells or biopsy-derived gsTcells. This set of genes was mainly upregulated (1047/1371) and was enriched for G protein-coupled receptor ligand binding, IL-6–type cytokine receptor ligand interaction and peptide ligand-binding receptors (**Fig. 3E**). Nonetheless, when compared to naïve CD4+ T cells, *in vitro*-cultured gsTcells show a distinct response profile on top of the shared TCR response profile that includes a diverse set of cytokines and chemokines that are comparable to that of Th1, Tfh and Th2 cells, which is consistent with previous evidence [6-13,43,44].

The expression profile of in vitro-cultured gsTcells is similar but not identical to the expression profile of gsTcells isolated directly from biopsies

Next, we compared the transcriptomic response of our cultured gsTcells with gsTcells taken directly from CeD biopsies six days after an *in vivo* gluten challenge[13]. These biopsy-derived gsTcells were obtained from Christophersen et al. [13] and contained tetramer+ CD4+ T cells and tetramer- CD4+ T cells from CeD patients (n=5) and gut CD4+ T cells from healthy controls (n=4). We found that 123 out of 3395 DE genes in gsTcells are shared with the 865 DE genes from biopsy-derived gsTcells (p-value=0.0056; two-sided Fisher's exact test) (**Fig. 3D**, **Suppl. Tab. 4,6**). These overlapping genes were enriched for the 'TNFR2 non-canonical NFκB pathway' in Reactome and included immune genes such as *CD200, MAP4K3 & PDCD1, IL21* (a key regulator in CeD) and *IL22* (a regulator of intestinal epithelial homeostasis [45]) (**Suppl. Fig. 5B**, **Suppl. Tab. 7**). The differences between the *in vitro*-cultured gsTcells and biopsy-derived gsTcells could partly be attributable to differences in sample size. Moreover, the *in vitro*-cultured gsTcells were stimulated for 3 hrs, and the DE genes therefore reflect early transcriptional changes upon stimulation, which may not reflect the activation state of *in vivo* biopsy-derived gsTcells. Finally, continuous *in vitro* culturing of the gsTcell clones in the presence of IL-2 and IL-15 is likely to introduce changes in the expression profile of the gsTcell clones compared to the biopsy-derived gsTcells. Thus, gsTcells show disease-relevant characteristics but are not identical to 'fresh' biopsy-derived gsTcells.

Chromatin of gsTcells remains largely stable

Next, we assessed if changes in chromatin state agreed with changes in gene expression upon stimulation of gsTcells. We observed that the chromatin state measured by DHS-Seq remained mostly stable at all four timepoints when assessed genome-wide, and most of the variation in open chromatin was explained by which gsTcell clone the data was derived from, and not by stimulation effects (**Fig. 4A**, **Suppl. Tab. 8**). Nonetheless, when investigating the open chromatin sites in a 5kb window around the 3509 DE genes, we observed some variation between t180 and the other timepoints (**Fig. 4B**).

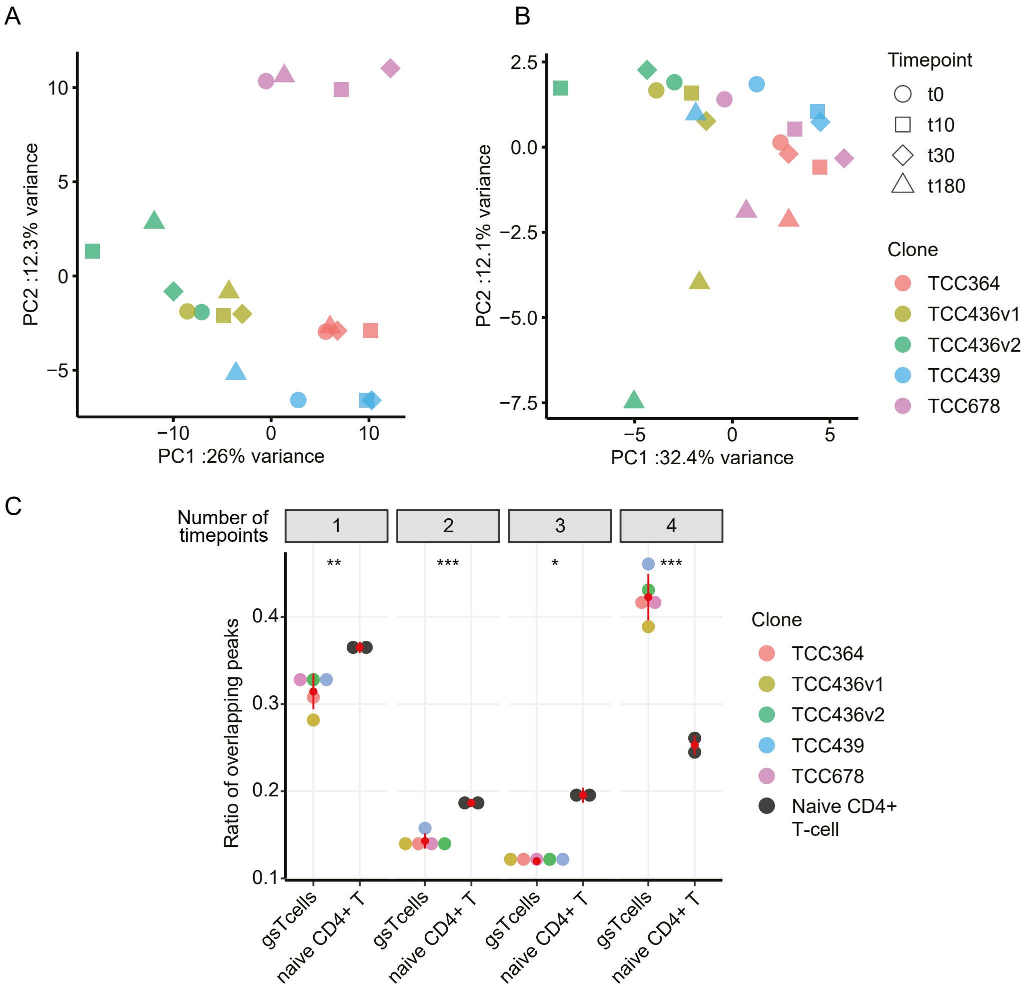


Fig. 4: Open chromatin changes are minimal in gsTcells upon stimulation. A) PC analysis of all peaks in each individual gsTcell clone in the discovery cohort. Timepoints indicated by shapes, clones by colors. **B)** As in (A), but only using peaks within a 5kb window around the transcription start sites of the 3509 DE genes in the PCA. **C)** Comparison of the relative changes in the individual gsTcells and naïve CD4+ T cells of two individuals between all timepoints. Overlaps between all timepoints were calculated using the multi-inter Bedtools function, and the relative number of overlaps is plotted for each. Peaks that are present at all timepoints are represented in the '4' category. Peaks unique to one time point are represented in the '1' category. Clones are indicated with the same colors as in (A) and (B). Nominal significance indicated with asterisks (*p-value < 0.05, **p-value < 0.01 and ***p-value < 0.001).

However, differential peak-calling analysis did not find any sites with a log2 fold change > 1. These results are probably due to the large variation between the clones (**Fig. 4A,B**) and the modest sample size. We also investigated the overlap between peaks in each clone and compared that to the overlap between ATAC-seq peaks in naïve CD4+ T cells in a similar stimulation experiment [46]. Overall, the naïve CD4+ T cells showed more stimulation-specific effects, with 25% of peaks being shared at all four timepoints, as opposed to gsTcells, where 45% of the peaks are shared (**Fig. 4C**). Moreover, TF binding site enrichment analysis on peaks uniquely present in unstimulated gsTcells, but not seen in naïve CD4+ T cells under unstimulated conditions, showed that AP-1 binding sites are enriched in gsTcells even prior to stimulation (HOMER, p-value= $1x10^{-16}$). Altogether, this suggests that the early upregulation of gene expression upon activation in gsTcells is mostly independent of chromatin state and instead driven by the interplay of TFs and ncRNAs.

DE genes show enrichment for CeD loci and CeD-associated genes

To ascertain if the DE genes we found are affected by the genetic background associated with CeD, we integrated our transcriptional data with 118 *cis*-genes for CeD identified by a previous gene prioritization effort [30] (**Fig. 5A**). We observed that 26 of the prioritized genes are DE in the gsTcells. Of these 26 genes, 18 belong to cluster 3, in which immune response pathways are overrepresented, consistent with the mostly T cell–based GWAS signal of CeD [23]. Of particular interest are *IL21* and other cytokine-encoding genes (*IL2, TNFSF11* and *FASLG*), several cytokine receptor genes (*IL21R, IL1RL1, TNFRSF9* and *IL2RA*) and genes for TFs that can play a role in immune response (*REL, BACH2* and *IRF4*). Interestingly, *UBE2E3, CSK* and *SLC22A4* are DE in gsTcells but not in naïve CD4+ T cells (**Fig. 3B**), which implies that these genes may have a relatively specialized function in gsTcells.

We subsequently checked if the DE genes were overrepresented in CeD loci [23] using genomic region enrichment analysis (GREA)[19] (**Fig. 5 B,C**; **Suppl. Fig. 8**). GREA operates by comparing the overlap of DE genes in a window of ± 125kb around a CeD locus to the overlap found randomly in a permutation-based null distribution. We found no significant enrichment for genes located in CeD loci when assessing all 3509 DE genes (**Fig. 5B**). However, when we assessed the enrichment per cluster (**Suppl. Fig. 8**), cluster 3 showed a nominally significant enrichment for genes located in CeD loci (p-value=0.042; **Fig. 5C**), with around 18% of cluster 3 genes being located near a CeD locus. Thus, the genes in cluster 3 may be affected by the genetic predisposition for CeD.

Next, we assessed whether any of the TFs that were DE in gsTcells were enriched to bind in CeD loci using REgulatory trait Locus Intersection (RELI) in conjunction with the provided database of ChIP-seq data for 389 TF-cell type pairs [29]. In total, 98 TF–cell type pairs showed significant enrichment for binding in CeD loci (FDR <0.05) (**Suppl. Tab. 9**). Of these TFs, FOS, STAT5A and TBX21 (also called T-BET) were of particular interest as their genes showed a DE effect in gsTcells and the ChIP-seq data used in RELI was derived from CD4+ T cells. *FOS* falls in cluster 1 and showed an early but transient response to stimulation (**Suppl. Fig. 9A**), whereas *STAT5A* and *TBX21* showed a later response corresponding to the cluster 3 profile (**Suppl. Fig. 9B and C**, respectively). FOS is a well-known early response immune and stress response TF that can heterodimerize with the JUN or ATF TF families. Indeed, *ATF3* is also DE and has a very similar transcriptional profile to *FOS* (**Suppl. Fig. 3A**). STAT5A is required for IL2 signaling in CD4+ T cells [47], and *TBX21*, which encodes T-BET, is the major TF for Th1 differentiation, expression of IFN-γ and other Th1-specific cytokines [48]. Thus, these TFs all have prominent roles in CD4+ T cell differentiation and activation.

To explore if any of the SNPs in the 43 CeD loci from the latest CeD meta-analysis [23] could directly affect gene regulation in gsTcells, we overlapped the TF binding sites of FOS, STAT5A and TBX21 with the open chromatin regions, CeD GWAS summary statistics [23], DE information, ATAC-seq of naïve CD4+ T cells [46] and capture Hi-C data of activated CD4+ T cells (**Fig. 6**). We found one locus of interest near the *IL18RAP/IL1RL1* genes. In this locus, one SNP (rs1420106, GWAS p-value= 8.7×10^{-14}) located in the promoter of *IL18RAP* overlapped with all three of the enriched TF binding sites, a DHS peak from gsTcells and an ATAC-seq peak from naïve CD4+T cells (**Fig. 6A**). We assessed if this locus could act as an enhancer using publicly available capture Hi-C data from activated CD4+ T cells [49] and found interactions with the promoters of *IL1RL1* and *IL18R1* (**Fig. 6A**). Moreover, rs1420106 strongly affected the expression of *IL18RAP*, *IL1RL1* and *IL18R1* (eQTL p-values 3.27×10^{-310}, 1.95×10^{-144} and 1.63×10^{-185}, respectively) in

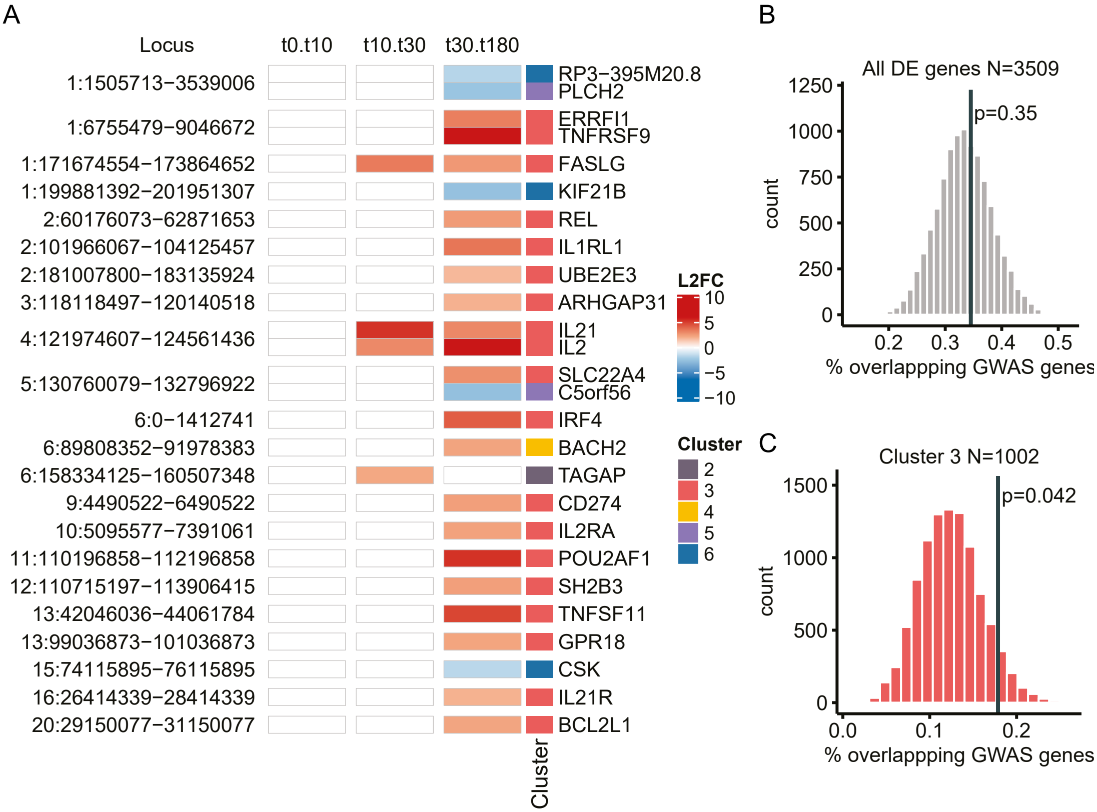


Fig. 5: DE genes show enrichment for CeD loci. A) Genes prioritized from CeD loci [30] were overlapped with DE genes and plotted per CeD locus. Log2 fold change between timepoints is indicated, and clusters are depicted with colored boxes (right). **B)** Enrichment of DE genes over the null distribution (histogram) in CeD loci (±125kb window around start and end of gene) using GREA[19] for all DE genes (n=3509). X-axis indicates the number of genes that overlap with CeD loci as a percentage. The histogram shows the null distribution based on 10,000 permuted gene-sets. The black line indicates the value of the true gene-set. Nominal p-values are indicated. **C)** As in (B), but only for genes in cluster 3 (n=1002).

whole blood (eQTLgen database [50]). Finally, both *IL1RL1* and *IL18RAP* were significantly upregulated in gsTcells at t180 compared to t0 (**Fig. 6B**). Together, this suggests that rs1420106 might have a role in the activation response of gsTcells by modifying the expression of *IL18RAP*, *IL1RL1* and *IL18R1*. Similarly, we identified three more CeD loci (containing genes *BACH2*, *IL21-IL2* and *TAGAP*) that showed overlap of SNPs associated with CeD with DHS sites found in gsTcells that also bind the TFs FOS, STAT5A and/or TBX21 (**Suppl. Fig. 10-12**).

Thus, CeD-associated genetics may play a complex role in gsTcells during activation of these important cells in CeD pathology.

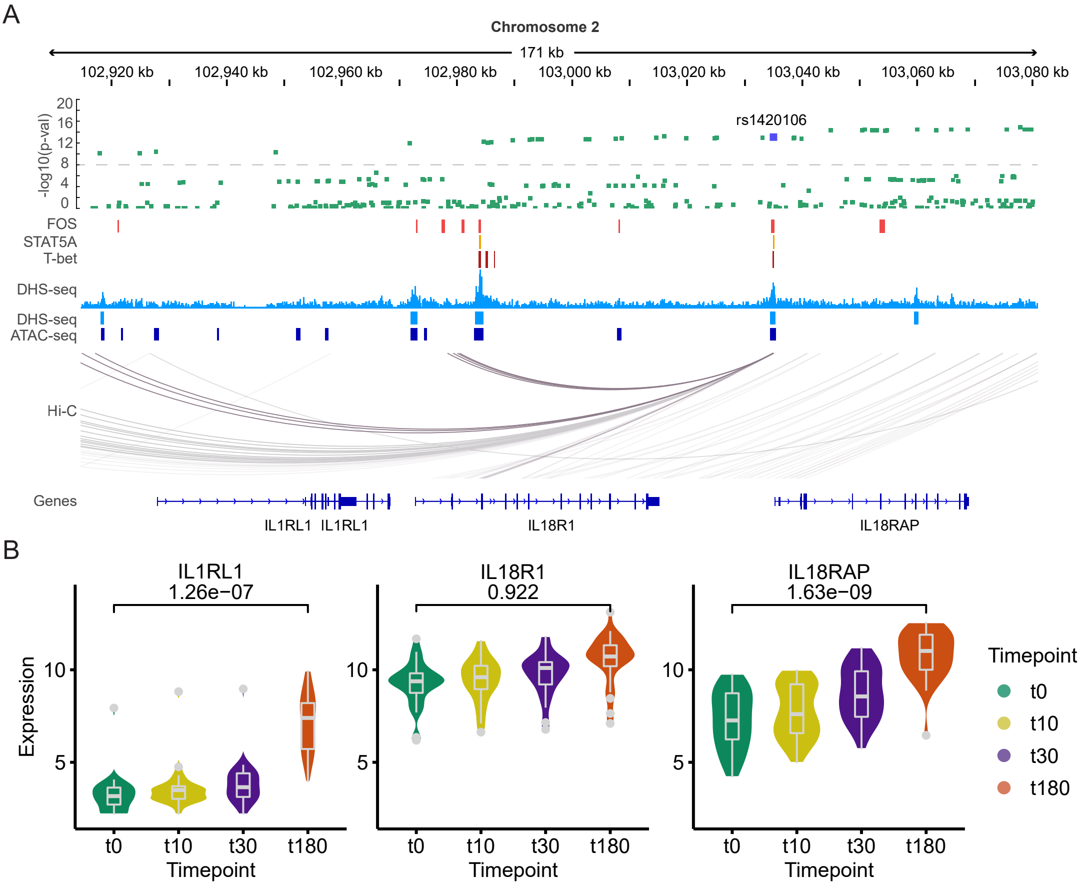


Fig. 6: Rs1420106 in the *IL1RL1/IL18R1/IL18RAP* locus overlaps with open chromatin and TF binding sites in gsTcells. A) Overview of the *IL1RL1/IL18R1/IL18RAP* locus. From top to bottom: summary statistics of the CeD GWAS meta-analysis [23] (green); TF binding sites based on Chip-seq of FOS (red), STAT5A (orange) and T-BET/TBX21 [29] (brown); DHS profile of unstimulated gsTcells with peak calls depicted below (light blue); peaks from naïve CD4+ T cell ATAC-seq [46] (dark blue); capture Hi-C data of activated CD4+ T cells depicting the 3D interactions between the highlighted region and other locations in the locus [82] (grey) and the gene annotations. The prioritized SNP rs1420106 is indicated in dark blue. **B)** Gene expression pattern of genes in the locus. Y-axis represents the VST-normalized expression data. Black line and number indicate the adjusted p-value (DeSeq2) of the DE effect between the t0 and t180 timepoints. Blood eQTL p-values of SNP rs1420106 for the indicated genes are: *IL18RAP*, 3.27×10^{-310}; *IL1RL1*, 1.95×10^{-144} and *IL18R1*, 1.63×10^{-185} (eQTLgen.org[50]).

Discussion

In this study, we characterized gsTcells, one of the key players in CeD pathogenesis, by profiling transcriptomic and epigenetic changes during the early response to aCD3/aCD28 activation. We pinpointed pathways and TFs that may regulate these cells and confirmed that gsTcells are not restricted to any specific class of Th cells, but rather express cytokines, chemokines and TFs that are characteristic of Th1 but also of Tfh and Th2 subsets. We also identified *CCL1*, *CXCL1* and *CCL4L1* as unique DE genes that had not previously been shown to be expressed by CD4+ T cells in the context of CeD. Finally, we showed that the early response of gsTcells to stimulation is regulated by rapid upregulation of many TFs and ncRNAs in combination with the activation of JNK/AP-1 and NFκB TFs, whereas changes in chromatin were minor. Overall, our results provide an in-depth analysis of the molecular pathways that are activated in gsTcells upon TCR activation.

Our study illustrates that gsTcells express and secrete cytokines that can be associated to various subsets of Th cells, most prominently with the Th1 (IFN-γ and TNF), Th2 (IL4, IL5, IL13) and Tfh (IL21) subsets, which is in agreement with previous studies [6-13,43,44]. Based on their unique cytokine profile, gsTcells may exert multiple functions in CeD pathogenesis. Firstly, IFN-γ is important for eliciting a strong response to foreign antigens such as the gluten peptides in CeD [51]. Moreover, IFN-γ can also directly affect the integrity of the intestinal barrier [52]. Secondly, Th2-associated cytokines and IL-21 produced by Tfh cells are important for plasma cell differentiation, B cell activation and autoantibody production [43]. Finally, IL-21, IL-2 and TNF have been shown to activate CD8+ T cells and IELs in the gut, which thereby become “licensed to kill” epithelial cells, leading to the villus atrophy in CeD [12,19,20,53]. Thus, cytokines derived from gsTcells may play a role in several distinct disease mechanisms.

Our comprehensive analysis of the regulatory mechanisms that drive gene expression in activation of gsTcells agrees in part with the results of a study where the authors analyzed biopsy-derived gsTcells (Christophersen et al. [13]) but also uncovered differences. These differences might be due to differences in sample size and experimental design between the two studies but may also reflect the fact that the gsTcells used in our study have been expanded *in vitro* in the presence of IL-2 and IL-15. Nonetheless, the unique expression of genes that are not DE in activated naïve CD4+ T cells and the enrichment of genes that overlap with genes specific to biopsy-derived gsTcells suggest that *in vitro*-cultured gsTcells are a unique and appropriate model to delineate the dynamic transcription and regulation of gsTcells.

Similarly, comparison with other datasets such as the DICE data (**Fig. 3A,B**) and CeD-patient derived CD4+ T cells from PBMCs described by Quinn et al. (data not shown), validate our findings that the gsTcells have a similar but distinct expression profile and that important cytokines and other genes are associated to CeD etiology, such as *IFNG, IL21, IL17A* and *IL4* [54], are differentially expressed. Still, the unique expression profile of gsTcells relative to naïve or blood-derived CD4+ T cells may be lost if the comparison with other antigen-specific T cells or memory CD4+ T cells activated with similar stimuli and timepoints.

Based on the expression pattern of the DE genes, we identified six major gene clusters with distinct dynamic responses and functions. Early responding genes were represented by clusters 1, 2 and 5 and contained multiple TFs, some cytokines and a disproportionate number of ncRNAs, implying that these genes regulate the response at later time points. An earlier study also observed a ncRNA response during the activation of lymphocytes, which supports the idea that ncRNAs are key in the development and activation of CD4+ T cell [55]. Some ncRNAs that are DE in gsTcells have also been implicated in immune activation, inflammation and proliferation in other studies, including *LINC00174*, *AF131217.1*, *TINCR* and *LINC00342*, which were all found in cluster 1 [56-60]. Thus, ncRNAs seem to play a pivotal role in the immune response by changing the expression of specific genes in gsTcells.

The gsTcells we studied showed a stable open chromatin profile genome-wide, with only minor changes upon stimulation near the transcription start sites of DE genes. However, we cannot exclude the possibility that the stability we observe is a consequence of culturing gsTcells in the presence of cytokines to induce expansion[61]. In contrast to the stable open chromatin profiles in gsTcells, the transient expression of specific TFs and ncRNAs, in concert with the activation of common signaling pathways like JNK/AP-1 and NFκB, may be the source of the unique expression profile observed in the gsTcells.

Several DE genes are located in CeD-associated loci, and we found a subtle enrichment for genes of immune cluster 3 in CeD loci. However, the largest CeD association study to date was performed using the Immunochip platform, which is enriched for known immune regions, and thus we may have missed enrichment in other functional clusters. Nonetheless, 18% of the genes in cluster 3 are located in CeD loci, highlighting that TCR-mediated T cell activation, particularly in gsTcells, may be affected by CeD-associated SNPs.

To ascertain the role of CeD-associated SNPs in regulating gene expression in gsTcells, we integrated multiple publicly available functional data layers with the gene expression and DHS regions of the gsTcells and CeD-associated SNPs. While this integration provides suggestive evidence that the prioritized SNPs have regulatory potential in gsTcells, we cannot directly confirm that the genetic effect has a regulatory role. This would require an eQTL analysis with primary gsTcells derived from biopsies or functional validation by targeting these candidate regulatory elements, both of which are beyond the scope of this study. Despite these challenges, we provide evidence for potential genetic interference of CeD-associated SNPs in gsTcells in several loci and pinpoint several SNPs and regulatory regions within the genome that are the most likely candidates to cause this interference.

In summary, we present an in-depth characterization of early transcriptional dynamics of gsTcells in response to TCR activation. We highlight that this transcriptional response is most likely regulated by TFs and ncRNAs rather than large changes in chromatin state. Finally, we prioritize several CeD-associated genetic loci that may impact the TCR-activation in gsTcells directly.

Methods

Obtaining T cell clones from biopsies of CeD patients

Gluten-specific CD4+ T cells were isolated from CeD patient small intestinal biopsies, as described previously [25,62-65]. Patients were diagnosed with CeD with small bowel biopsy confirmation and included at Leiden University Medical Center (LUMC), the Netherlands (n=18), and the Riks Hospital in Oslo (RHO), Norway (n=4), from which 23 and 5 gluten-specific T cell lines were isolated, respectively (**Suppl. Tab. 1**). Briefly, written informed consent was given by all patients and the small intestine biopsies from CeD patients were cultured with a mixture of gluten and TG2-treated gluten (deamidation) for 5 days. To expand the T cells, IL-2 (20 Cetus units/ml; Novartis, Arnhem, the Netherlands) and IL-15 (10 ng/ml; R&D systems, Abingdon, UK) were added. Subsequently, irradiated allogeneic peripheral blood mononuclear cells in the presence of phytohemagglutinin (1 µg/ml; Remel Inc. Lenexa, USA), IL-2 (20 Cetus units/ml) and IL-15 (10 ng/ml) were mixed with T cells for re-stimulation[25]. The resulting T cell clones were tested for reactivity against gluten digested by pepsin and trypsin and TG2-treated in proliferation assay. The pepsin/trypsin digest of gluten was prepared as described by Van de Wal et al.[63]. For deamidation, the pepsin/trypsin-digested gluten (500 mg/ml) was incubated with 100 mg/ml of guinea pig tTG (T-5398; Sigma, St. Louis, MO) at 37°C for 2 h in PBS with 1 mM CaCl2 and subsequently used in T cell proliferation assays. Proliferation assays were conducted as described by Van de Wal et al.[63], for samples from LUMC, and Molberg et al.[65], for samples from RHO. Gluten-specific lines were cloned by limiting dilution and expanded again by re-stimulation at 1- to 3-week intervals [25]. Clones were stored in liquid nitrogen. All methods were performed in accordance with relevant guidelines and regulations.

Stimulation of T cell clones

In all, 28 gluten-specific T cell clones were stimulated in 6-well plates coated overnight with anti-CD3 (2.5 µg/ml; Biolegend, San Diego, CA, USA) and anti-CD28 (2.5 µg/ml; Biolegend) or PBS (negative control) for 0, 10, 30 and 180 minutes. At each timepoint, cells were harvested for RNA isolation. Cell culture medium was harvested after 240 minutes for proteomic analysis.

RNA isolation and library preparation of stimulated gsTcells

GsTcells were harvested at each time point, washed with PBS and resuspended in a lysis buffer (Ambion, Life Technologies, Carlsbad, CA, USA). RNA was extracted with the mirVana RNA isolation kit (Ambion) according to the manufacturer's instructions. The quantity and quality of RNA was determined by Bioanalyzer (Agilent technologies, Santa Clara, CA, USA). The sequencing libraries were prepared from 1 µg of total RNA using the TruSeq Stranded Total RNA with Ribo-Zero Globin kit (Illumina, San Diego, CA, USA) according to the manufacturer's instructions. Sequencing was done with the Illumina HiSeq 2500 (Illumina).

DNase I hypersensitivity sequencing and analysis

Standard protocols for nuclei isolation, DNase I (Roche #04716728001) treatment and library preparation for DNase I hypersensitivity sequencing generated within the Blueprint consortium were followed. Protocol details [66] can be found at:

http://www.blueprint-epigenome.eu/UserFiles/file/Protocols/Blueprint_DNase1_Protocol.pdf. All samples were sequenced to a sequencing depth of approximately 50-60 million 50 bp single-end reads.

Protein analysis

Supernatants from unstimulated and 4 hour-stimulated gsTcell cell cultures were taken and analyzed with the Immuno-Oncology panel of Olink (http://www.olink.com/products/immuno-oncology). Data was analyzed by subtracting relative log2 protein levels in a blank medium control from the relative log2 protein levels in the supernatants, followed by a two-sided t-test to measure significant change between supernatants of unstimulated and 4 hour-stimulated gsTcells.

Statistical methods

Statistical analyses were performed in R (version 3.6.3) [67] unless otherwise specified. Visualization of results was done using the R package ggplot2 (version 3.3.0)[68].

RNA-seq quantification

Before alignment, the reverse complement of the fastQ sequences were taken using the FASTX-Toolkit [69]. Alignment was done using Hisat2 (version 2.0.4) [70] against the forward strand, with default alignment parameters. The reference genome index was made using the Hisat2-build indexer and 1000 genomes reference genome version GRCh37 v75 with default parameters. For the samples that had paired-end data, only the first mate file was used for alignment. Reads mapping to multiple positions were removed. The genes were quantified using HTSeq (version 0.6.1.p1)[71] with options -m union, -t exon, --stranded yes and other options on default.

DE analysis

The raw count matrix, containing 63,682 genes and 112 samples (92 samples from the Leiden cohort and 20 from the Oslo cohort), was first filtered to remove any non-expressed genes by selecting only genes that had at least 1 read in 20 samples. This resulted in 29,772 genes to be tested for DE effects. Samples from the Leiden and Oslo cohorts were then split, and the DE effects assessed separately.

DE effects were quantified using the R package DEseq2 (version 1.26.0) [72], including RNAseq batch and sex as covariates for the Leiden samples. No covariates were included for the Oslo samples because no sex information was available and all samples had been sequenced in the same batch. DE effects were then mapped between the t0 and t10, t10 and t30, and t30 and t180 timepoints. DE effects in the Leiden cohort were filtered on having an absolute log2 fold change (log2FC) of at least 1 and an FDR < 0.05. Oslo samples were used as the replication cohort, and comparisons between the two were made using unfiltered Oslo data. In total, we identified 3509 unique DE genes in the Leiden cohort. These genes were used for interpretation and downstream analysis. PC analyses were performed on the variance-stabilized count data.

Clustering of DE genes into distinct response patterns

DE genes were clustered into time patterns as follows. The gene expression matrix was VST-normalized using DESeq2 (version 1.26.0) [72], after which the mean expression level for each gene was determined at each of the four time points. Each row was then centered to mean 0 and scaled to standard deviation of 1. The data was clustered using k-means clustering (k=6) on a Euclidean distance matrix using the R package TCseq (version 1.10.0) [73]. Cluster number was determined by assessing the stability of the clustering in terms of within-cluster sums of squares over 100 iterations of the clustering. We then determined that the optimal tradeoff between stability and informativeness of each cluster occurred with a cluster number of k=6. To verify the stability, we ran another 100 random k-means clustering runs using different parameters (nstart=100, k=6 and iter.max=1000). This yielded 100 fully stable clusters that matched very well with the clustering definition maintained in the manuscript (98.3% of genes matched their cluster).

Comparisons with DICE and biopsy-derived gsTcells data

DE genes from primary naïve CD4+ T cells were retrieved from DICE (https://dice-database.org/)[41]. Briefly, naïve CD4+ T cells from healthy donors (n=91) were obtained from blood by FACS and stimulated using aCD3/aCD28 for 4 hours. Biopsy-derived gsTcells were obtained from Christophersen et al. [13] and contained tetramer+ CD4+ T cells and tetramer- CD4+ T cells from CeD patients (n=5) and gut CD4+ T cells from healthy controls (n=4). First, we obtained all DE genes that were significant (adjusted p-value < 0.05) and showed an absolute log2FC > 1. Next, we intersected all DE genes from each dataset to obtain those that were unique per dataset and those that were shared with gsTcells. We then obtained the overlapped DE genes of gsTcells between DICE (n=1926) and CeD biopsies (n=144) to evaluate the concordance of those genes using the log2FC. The overlapped genes were divided in four quadrants (Q1-Q4). Q1 and Q3 included concordant DE genes that were upregulated and downregulated, respectively. Q2 and Q4 consisted of non-concordant DE genes, with Q2 being upregulated in gsTcells but not in reference dataset and Q4 being vice-versa.

Gene set enrichment analysis

Reactome pathways [74] were used to identify the pathways or biological processes that were enriched for each set of genes. This analysis was performed using the R package clusterProfiler (version 3.14.3) [75]. p-values were adjusted using the Benjamini-Hochberg procedure to account for multiple testing.

Quantification and peak calling of DHS sequencing

DHS reads were aligned to hg19 reference genome using bwa (version 0.6.1-r104)76 with default settings, after which duplicates were marked using bamUtil (version: 1.0.2) [77]. Alignments were then filtered to have a mapping quality of at least 30 and a primary alignment and to not be duplicated using Samtools (version 1.9) [78]. Peaks were then called using macs2 (version 2.2.6) [79] enabling --broad --nomodel –shift -125 --extsize 250. Peaks were considered at an FDR threshold < 0.05.

Differential peak calling of DHS sequencing

To identify differentially accessible sites between timepoints, consensus peaks were first defined using the R package DiffBind (version 2.14.0) [80], after which raw read counts were determined for each consensus peak. Differentially accessible peaks were then quantified between t0 and t10, t10 and t30, and t30 and t180 using DEseq2 (version 1.26.0) [72]. PC analyses of DHS data were performed on the RPKM-normalized log10-transformed read counts for the consensus peaks. Overlap of peaks between the gsTcells or public datasets was determined with Bedtools multiinter –cluster. TF binding site enrichment was performed with Homer findMotifsGenome.pl with the merged regions of untreated gsTcells as background [81].

Enrichment of differentially expressed genes in coeliac disease loci

To test for enrichment of DE genes in CeD loci, we used the R package GREA (https://github.com/raguirreg/GREA, version 0.1.0) [19]. We defined CeD genes as genes within a 125kb window of a CeD GWAS top SNP. We then generated 10,000 random gene-sets that matched the CeD gene-sets in size. The 10,000 random gene-sets were used to generate an empirical null distribution of the overlap between our DE gene-set per cluster and the random gene-sets. We then estimated the one-sided empirical p-value of the enrichment for each cluster of DE genes.

Acknowledgements

We thank Kate Mc Intyre for editing the manuscript. We would like to acknowledge Morris Swertz for facilitating data processing and storage. O.B.B. is supported by a Netherlands Organization for Scientific Research (NWO) VIDI grant (no. 016.171.047) and an NWO VENI grant (no. 863.13.011). A.D.R.S. is supported by CONACYT-I2T2 scholarship (no. 459192). S.W. and C.W. were supported by The Netherlands Organ-on-Chip Initiative, an NWO Gravitation project (024.003.001) funded by the Ministry of Education, Culture and Science of the government of the Netherlands. K.E.A.L., L.M.S., C.W., S.W.Q. and M.K.J. are supported by Stiftelsen KG Jebsen (project SKGJ-MED-017). I.J. is supported by a Rosalind Franklin Fellowship from the University of Groningen and an NWO VIDI grant (no. 016.171.047).

Author contributions

Z.A.B., R.M., Y.K.W., M.K.J., S.W.Q., F.M., J.v.B. and I.H.J. performed wet-lab experiments. K.E.A.L., L.M.S. and F.K. provided samples. Z.A.B., I.H.J, A.D.R.S. and O.B.B. conceived and wrote the manuscript. A.D.R.S. and O.B.B. performed the statistical analyses. Z.A.B. and N.K. assisted with quantifying the RNAseq and DHS data. I.H.J, Y.L., J.H.A.M., V.K., K.E.A.L., L.M.S., F.K. S.W. and C.W. supervised and edited the manuscript.

Data availability

All data (RNA-Seq count tables, DHS counts and peaks) required to reproduce this study have been provided as supplementary files. The raw RNA-Seq and DNAse reads supporting this study are available upon request to the authors as this is privacy sensitive. All code and scripts used to generate the results and figures are available on Github (https://github.com/OlivierBakker/gluten_specific_tcells).

Conflict of interest statement

The authors declare no conflict of interest.

Ethics statement

Biological material was obtained from coeliac disease patients according to protocols approved by the regional ethics committees (Medical Ethical Committees of VU University Medical Center, Leiden University Medical Center and University of Oslo), and the individuals donating material gave their written informed consent.

Supplementary material

Supplemenatary material are provided at:

https://doi.org/10.1038/s41598-021-86612-5

Suppl. Fig. 1: PCA based on all DE genes in the discovery cohort. Timepoints indicated with colors.

Suppl. Fig. 2: Evaluation of temporal clustering of gene expression. A) Log2 fold changes between consecutive timepoints of all DE genes plotted in a heatmap. On the left, a dendrogram based on hierarchical clustering of the gene's log2 fold changes, followed by a bar indicating the biotype of the gene, with protein-coding genes in red and non-coding genes in green. On the right, a bar indicating which cluster the gene belongs to, as depicted in **Fig. 2A**. Overall concordance with the k-means-based clustering of the expression data is good. B) Heatmap of another 100 random k-means clustering runs using different parameters (nstart=100, k=6 and iter.max=1000). Rows indicate genes. Columns indicate the random cluster runs. Red-blue colors indicate the size of the cluster. This yielded 100 fully stable clusters that matched very well with the clustering definition used in the manuscript (98.3% of genes matched their cluster). C) Elbow plot showing the within-cluster sums of squares (y-axis) for different cluster numbers (x-axis). Boxplots show the overall stability of the clustering over 100 repeats of the same clustering. Cluster numbers 2, 3, 4 and 6 showed most stability over the 100 runs.

Suppl. Fig. 3: Expression profiles of specific transcription factors. A) TFs that respond early to stimulation. All these TFs are in cluster 1 or 2. Mean profile of VST-normalized expression is shown with 0.95 confidence interval in the shaded area. B) As in (A), but for NFκB-associated factors. NFκB inhibitory proteins encoded by *NFKBIA*, *NFKBID* and *NFKBIZ* are in cluster 1 and 2, whereas *REL* and *NFKB1* are in cluster 3.

Suppl. Fig. 4: Proteins secreted by gsTcells. A) Scaled relative protein levels as detected in the medium of unstimulated gsTcells and after 4 hrs of stimulation from three independent experiments (black dots). Only protein levels of genes that are DE are shown (30 out of 92 proteins measured). Box colors above the dotplots indicate the cluster in which the corresponding DE genes are found. Cluster 2 is in grey, cluster 3 in red and cluster 6 in blue. Mean, minimum and maximum relative protein levels are indicated in red. Nominal significance is indicated with asterisks (*p-value < 0.05, **p-value < 0.01, ***p-value < 0.001 and ****p-value < 0.0001). B) Expression level of the genes corresponding to the proteins significantly changed after 4 hrs in (A). Mean expression profile is shown with 0.95 confidence interval in shaded area.

Suppl. Fig. 5: Reactome gene set enrichment analysis of shared DE genes with gsTcells. Significant enrichments in Reactome pathways found in: A) overlapped DE genes of gsTcells and DICE consortium, only concordant quadrants (Q1 and Q3) and B) shared DE genes of gsTcells with biopsy-derived gsTcells. Only for overlapping upregulated gsTcell DE genes an enrichment was found.

Suppl. Fig. 6: Expression level of cytokines and chemokines in gsTcells. A) Expression profile of chemokine genes uniquely expressed in gsTcells. Mean profile of VST-normalized expression is shown with 0.95 confidence interval in the shaded area. B) As in (A), but for cytokines expressed by specific T helper cell subsets.

Suppl. Fig. 7: Reactome gene enrichment analysis of DE genes specific for external datasets. Pathway enrichments of specific DE genes in external references compared to gsTcells in this study. Results of comparison with DICE consortium (adjusted p-value < 0.05, absolute log2FC > 2). No enrichment was found in DE genes specific for biopsy-derived gsTcells.

Suppl. Fig. 8: Enrichment of DE genes in CeD loci per cluster. Enrichment of DE genes over the null distribution in CeD loci (±125kb window around start and end of gene) for each individual cluster using GREA[19]. Nominal p-values and number of genes per cluster are indicated. X-axis indicates number of genes that overlap with CeD loci, expressed as a percentage.

Suppl. Fig. 9: Gene expression profile of prioritized TFs in gsTcells. Mean profile of VST-normalized expression is shown with 0.95 confidence interval in shaded area. Fig. shows A) *FOS*, a subunit of AP-1, located in cluster 1, B) *STAT5A*, located in cluster 3, and C) *TBX21*, located in cluster 3.

Suppl. Fig. 10: Rs1738074 in the *TAGAP* locus overlaps with open chromatin and TF binding sites in gsTcells. A) Overview of the *TAGAP* locus. From top to bottom: summary statistics of the CeD GWAS meta-analysis [22] (green); TF binding sites based on Chip-seq of FOS (red), STAT5A (red) and T-BET/TBX2129 (red); DHS profile of unstimulated gsTcells with peak calls depicted below (light blue); peaks from naïve CD4+ T cell ATAC-seq [46] (dark blue); capture Hi-C data of activated CD4+ T cells depicting the 3D interactions between the highlighted region and other locations in the locus [82] (purple) and the gene annotations. Rs1738074 is indicated with the light blue box. B) Gene expression pattern of genes in the locus. Y-axis represents the VST-normalized expression data. Black line and number indicate the adjusted p-value (DeSeq2) of the DE effect between the t0 and t180 timepoints. Blood eQTL p-values of SNP rs1738074 on the indicated genes are: *TAGAP*, 2.05×10^{-19} and *RSPH3*, 7.37×10^{-13} (eQTLgen.org[50]).

Because *TAGAP* is DE in gsTcells and because rs1738074 is directly located in the promoter of *TAGAP*, has an eQTL and contains binding sites for FOS and T-BET, the expression of *TAGAP* is likely affected by CeD-associated genetics in gsTcells. Conversely, *RSPH3* is not DE, nor is there a 3D interaction between the location of rs1738074 and the promoter of *RSPH3*, which suggests that the eQTL effect observed in whole blood in eQTLgen data is not present in gsTcells.

Suppl. Fig. 11: Rs13140464 in the *IL2/IL21* locus overlaps with open chromatin and TF binding sites in gsTcells. A) Overview of the *IL2/IL21* locus. From top to bottom: summary statistics of the CeD GWAS meta-analysis [22] (green); TF binding sites based on Chip-seq of FOS (red), STAT5A (red) and T-BET/TBX2129 (red); DHS profile of unstimulated gsTcells with peak calls depicted below (light blue); peaks from naïve CD4+ T cell

ATAC-seq [46] (dark blue); capture Hi-C data of activated CD4+ T cells depicting the 3D interactions between the highlighted region and other locations in the locus [82] (purple) and the gene annotations. Rs13140464 is indicated with the light blue box. B) Gene expression pattern of the genes in the locus. Y-axis represents the VST-normalized expression data. Black line and number indicate the adjusted p-value (DeSeq2) of the DE effect between the t0 and t180 timepoints. Blood eQTL p-values of SNP rs13140464 on the indicated genes: *FGF2*; 1.63×10^{-5} (eQTLgen.org[50]).

This locus is highly complex and only has one eQTL effect, through rs13140464 on *FGF2*, based on whole blood data (eQTLgen.org50). However, the interaction of the DHS site that overlaps with rs13140464 does not interact directly with this gene in activated CD4+ T cells, indicating that this gene may not be affected by the SNP in gsTcells in this context. Other genes for which no eQTL effects by SNP rs13140464 have been identified do have direct interactions, including *IL2*, *CETN4P*, *NUDT6* and *SPATA5*. Out of these, only *IL2* is differentially expressed, which indicates that it may be affected by rs131140464 under stimulated conditions in gsTcells.

Suppl. Fig. 12: rs905671 and rs943689 in the *BACH2* locus overlap with open chromatin and TF binding sites in gsTcells. A) Overview of the *BACH2* locus. From top to bottom: summary statistics of the CeD GWAS meta-analysis [22] (green); transcription factor binding sites based on Chip-seq of FOS (red), STAT5A (red) and T-BET/TBX2129 (red); DHS profile of unstimulated gsTcells with peak calls depicted below (light blue); peaks from naïve CD4+ T cell ATAC-seq [46] (dark blue); capture Hi-C data of activated CD4+ T cells depicting the 3D interactions between the highlighted region and other locations in the locus [29] (purple) and the gene annotations. rs905671 and rs943689, located 185bp apart, are indicated by the light blue box. B) Gene expression pattern of genes in the locus. Y-axis represents the VST-normalized expression data. Black line and number indicate the adjusted p-value (DeSeq2) of the DE effect between the t0 and t180 timepoints. Blood eQTL p-values of SNP rs943689 on the indicated genes are: *BACH2*, 7.59×10^{-53} (eQTLgen.org[50]).

In whole blood, *BACH2* is strongly affected by rs905671 and rs943689. Moreover, this TF is differentially expressed in activated gsTcells, and the locus is strongly enriched for 3D interactions between the *BACH2* promoter and many other intergenic locations in *BACH2*, including with the DHS site that overlaps with rs905671 and rs943689. Thus, *BACH2* is likely affected by CeD-associated genetics, with the most likely candidate SNPs being rs905671 and rs943689.

Supplemental Tables

Suppl. Tab. 1: Description of the gsTcell clones used in this study. LUMC denote clones in the discovery cohort, and RHO are the clones used for the replication and for DHSseq analysis.

Suppl. Tab. 2: DE genes for discovery and replication cohorts. Results from replication cohort prefixed by ‘o’.

Suppl. Tab. 3: DE results for all the proteins assayed using the O-link panel. Summary statistics reported for a two-tailed t-test between baseline and stimulated condition after 4 hours.

Suppl. Tab. 4: Combinatory matrix of DE genes used for the upset plot in **Fig. 3A**. Columns indicate the dataset, and rows name the genes. The number 1 indicates that the gene is DE.

Suppl. Tab. 5: Comparison of DE genes of gsTcells with the DICE dataset.

Suppl. Tab. 6: Comparison of DE genes of gsTcells with the biopsy-derived gsTcell dataset.

Suppl. Tab. 7: Gene Set Enrichment Analysis results using Reactome.

Suppl. Tab. 8: DHS accessibility profiles per timepoint and normalized DHS counts for consensus sites.

Suppl. Tab. 9: Results from TF binding enrichment analysis for CeD.

References

1. Lindfors, K. et al. Coeliac disease. Nat. Rev. Dis. Prim. 5, 1–18 (2019).
2. Jabri, B. & Sollid, L. M. T Cells in Coeliac Disease. J. Immunol. 198, 3005–3014 (2017).
3. Molberg et al. Gliadin specific, HLA DQ2-restricted T cells are commonly found in small intestinal biopsies from coeliac disease patients, but not from controls. Scand. J. Immunol. 46, 103–108 (1997).
4. Christophersen, A. et al. Healthy HLA-DQ2.5+ Subjects Lack Regulatory and Memory T Cells Specific for Immunodominant Gluten Epitopes of Coeliac Disease. J. Immunol. 196, 2819–2826 (2016).
5. Risnes, L. F. et al. Disease-driving CD4+ T cell clonotypes persist for decades in coeliac disease. J. Clin. Invest. 128, 2642–2650 (2018).
6. Gianfrani, C. et al. Gliadin-Specific Type 1 Regulatory T Cells from the Intestinal Mucosa of Treated Coeliac Patients Inhibit Pathogenic T Cells. J. Immunol. 177, 4178–4186 (2006).
7. Bodd, M. et al. HLA-DQ2-restricted gluten-reactive T cells produce IL-21 but not IL-17 or IL-22. Mucosal Immunol. 3, 594–601 (2010).
8. Nilsen, E. M. et al. Gluten specific, HLA-DQ restricted T cells from coeliac mucosa produce cytokines with Th1 or Th0 profile dominated by interferon γ. Gut 37, 766–776 (1995).
9. Brottveit, M. et al. Mucosal cytokine response after short-term gluten challenge in coeliac disease and non-coeliac gluten sensitivity. Am. J. Gastroenterol. 108, 842–850 (2013).
10. Lahat, N. et al. Cytokine profile in coeliac disease. Scand. J. Immunol. 49, 441–447 (1999).
11. Sjöberg, V. et al. Intestinal T cell Responses in Coeliac Disease – Impact of Coeliac Disease Associated Bacteria. PLoS One 8, e53414 (2013).
12. Kooy-Winkelaar, Y. M. C. et al. CD4 T cell cytokines synergize to induce proliferation of malignant & nonmalignant innate intraepithelial lymphocytes. Proc. Natl. Acad. Sci. U. S. A. 114, E980–E989 (2017).
13. Christophersen, A. et al. Distinct phenotype of CD4+ T cells driving coeliac disease identified in multiple autoimmune conditions. Nat. Med. 25, 734–737 (2019).
14. Moens, L. & Tangye, S. G. Cytokine-Mediated Regulation of Plasma Cell Generation: IL-21 Takes Center Stage. Front. Immunol. 5, 1–13 (2014).
15. Vazquez, M. I., Catalan-Dibene, J. & Zlotnik, A. B cells responses and cytokine production are regulated by their immune microenvironment. Cytokine 74, 318–326 (2015).
16. Mesin, L., Sollid, L. M. & Niro, R. Di. The intestinal B cell response in coeliac disease. Front. Immunol. 3, 1–12 (2012).
17. Meresse, B., Korneychuk, N., Malamut, G. & Cerf-Bensussan, N. Interleukin-15, a master piece in the immunological Jigsaw of coeliac disease. Dig. Dis. 33, 122–130 (2015).
18. Korneychuk, N. et al. Interleukin 15 and CD4+ T cells cooperate to promote small intestinal enteropathy in response to dietary antigen. Gastroenterology 146, 1017–1027 (2014).
19. Zorro, M. M. et al. Tissue alarmins and adaptive cytokine induce dynamic and distinct transcriptional responses in tissue-resident intraepithelial cytotoxic T lymphocytes. J. Autoimmun. 108, 102422 (2020).
20. Ciszewski, C. et al. Identification of a γc Receptor Antagonist That Prevents Reprogramming of Human Tissue-resident Cytotoxic T Cells by IL15 and IL21. Gastroenterology 158, 625–637 (2020).
21. Dubois, P. C. A. et al. Multiple common variants for coeliac disease influencing immune gene expression. Nat. Genet. 42, 295–302 (2010).
22. Trynka, G. et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in coeliac disease. Nat. Genet. 43, 1193–1201 (2011).
23. Ricaño-Ponce, I. et al. Immunochip meta-analysis in European and Argentinian populations identifies two novel genetic loci associated with coeliac disease. Eur. J. Hum. Genet. 28, 313–323 (2020).
24. Kim, C. Y., Quarsten, H., Bergseng, E., Khosla, C. & Sollid, L. M. Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in coeliac disease. Proc. Natl. Acad. Sci. U. S. A. 101, 4175–4179 (2004).
25. Petersen, J. et al. T cell receptor recognition of HLA-DQ2-gliadin complexes associated with coeliac disease. Nat. Struct. Mol. Biol. 21, 480–488 (2014).

26. Sollid, L. M., Qiao, S. W., Anderson, R. P., Gianfrani, C. & Koning, F. Nomenclature and listing of coeliac disease relevant gluten T cell epitopes restricted by HLA-DQ molecules. Immunogenetics 64, 455–460 (2012).
27. Withoff, S., Li, Y., Jonkers, I. & Wijmenga, C. Understanding Coeliac Disease by Genomics. Trends Genet. 32, 295–308 (2016).
28. Farh, K. K.-H. et al. Genetic and epigenetic fine-mapping of causal autoimmune disease variants. Nature 518, 337–43 (2015).
29. Harley, J. B. et al. Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity. Nat. Genet. 50, 699–707 (2018).
30. van der Graaf, A. et al. Systematic Prioritization of Candidate Genes in Disease Loci Identifies TRAFD1 as a Master Regulator of IFN-γ Signaling in Coeliac Disease. Front. Genet. 11, 1–16 (2021).
31. Kumar, V., Wijmenga, C. & Xavier, R. J. Genetics of immune-mediated disorders: from genome-wide association to molecular mechanism. Curr. Opin. Immunol. 31, 51–57 (2014).
32. 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, 62–74 (2016).
33. Monteleone, G. et al. A Failure of Transforming Growth Factor-β1 Negative Regulation Maintains Sustained NF-κB Activation in Gut Inflammation. J. Biol. Chem. 279, 3925–3932 (2004).
34. Fernandez-jimenez, N. et al. Coregulation and modulation of NFκB-related genes in coeliac disease: Uncovered aspects of gut mucosal inflammation. Hum. Mol. Genet. 23, 1298–1310 (2014).
35. Maiuri, M. C. et al. Nuclear factor κB is activated in small intestinal mucosa of coeliac patients. J. Mol. Med. 81, 373–379 (2003).
36. Li, P. et al. BATF-JUN is critical for IRF4-mediated transcription in T cells. Nature 490, 543–546 (2012).
37. Renner, F. & Schmitz, M. L. Autoregulatory feedback loops terminating the NF-κB response. Trends Biochem. Sci. 34, 128–135 (2009).
38. Shih, V. F. S. et al. Kinetic control of negative feedback regulators of NF-κB/RelA determines their pathogen- and cytokine-receptor signaling specificity. Proc. Natl. Acad. Sci. U. S. A. 106, 9619–9624 (2009).
39. Crotty, S. A brief history of T cell help to B cells. Nat. Rev. Immunol. 15, 185–189 (2015).
40. Zhu, J. & Paul, W. E. Peripheral CD4+ T cell differentiation regulated by networks of cytokines and transcription factors. Immunol. Rev. 238, 247–262 (2010).
41. Schmiedel, B. J. et al. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell 175, 1701–1715 (2018).
42. Hughes, C. E. & Nibbs, R. J. B. A guide to chemokines and their receptors. FEBS J. 285, 2944–2971 (2018).
43. Raphael, I., Nalawade, S., Eagar, T. N. & Forsthuber, T. G. T cell subsets and their signature cytokines in autoimmune and inflammatory diseases. Cytokine 74, 5–17 (2015).
44. Spolski, R. & Leonard, W. J. IL-21 and T follicular helper cells. Int. Immunol. 22, 7–12 (2009).
45. Keir, M. E., Yi, T., Lu, T. T. & Ghilardi, N. The role of IL-22 in intestinal health and disease. J. Exp. Med. 217, 1–9 (2020).
46. Qu, K. et al. Individuality and Variation of Personal Regulomes in Primary Human T Cells. Cell Syst. 1, 51–61 (2015).
47. Farrar, M. A. & Owen, D. L. STAT5 and CD4+ T Cell Immunity. F1000Research 6, 1–10 (2017).
48. Zhang, Y., Zhang, Y., Gu, W. & Sun, B. Th1/Th2 Cell Differentiation and Molecular Signals. in T Helper Cell Differentiation and Their Function (ed. Sun, B.) 15–44 (Springer Netherlands, 2014). doi:10.1007/978-94-017-9487-9_2
49. Burren, O. S. et al. Chromosome contacts in activated T cells identify autoimmune disease candidate genes. Genome Biol. 18, 1–19 (2017).
50. Võsa, U. et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv 1–57 (2018). doi:10.1101/447367
51. Schroder, K., Hertzog, P. J., Ravasi, T. & Hume, D. A. Interferon-γ: an overview of signals, mechanisms and functions. J. Leukoc. Biol. 75, 163–189 (2004).
52. Eriguchi, Y. et al. Essential role of IFN-γ in T cell–associated intestinal inflammation. JCI Insight 3, (2018).
53. Setty, M. et al. Distinct and Synergistic Contributions of Epithelial Stress and Adaptive Immunity to Functions of Intraepithelial Killer Cells and Active Coeliac Disease. Gastroenterology 149, 681–691 (2015).
54. Quinn, E. M. et al. Transcriptome analysis of CD4+ T cells in coeliac disease reveals imprint of BACH2 and IFNI regulation. PLoS One 10, 1–25 (2015).
55. Xia, F. et al. Dynamic transcription of long non-coding RNA genes during CD4+ T cell development and activation. PLoS One 9, 1–11 (2014).
56. Lu, Q., Meng, Q., Qi, M., Li, F. & Liu, B. Shear-sensitive lncRNA AF131217.1 inhibits inflammation in HUVECs via Regulation of KLF4. Hypertension 73, E25–E34 (2019).
57. Liu, J., Yang, T., Zhang, Y. & Wang, S. Promotion of BZW2 by LINC00174 through miR-4500 inhibition enhances proliferation and apoptosis evasion in laryngeal papilloma. Cancer Cell Int. 20, 1–10 (2020).
58. Guo, J. et al. Role of linc00174/miR-138- 5p (miR-150- 5p)/FOSL2 Feedback Loop on Regulating the Blood-Tumor Barrier Permeability. Mol. Ther. - Nucleic Acids 18, 1072–1090 (2019).
59. Liu, C., Xu, Y., Wu, X. & Zou, Q. Clinical significance of linc00342 expression in the peripheral blood lymphocytes of patients with chronic kidney disease. Int. J. Nephrol. Renovasc. Dis. 12, 251–256 (2019).
60. Xu, S., Kong, D., Chen, Q., Ping, Y. & Pang, D. Oncogenic long noncoding RNA landscape in breast cancer. Mol. Cancer 16, 1–15 (2017).
61. Hedfors, I. A. & Brinchmann, J. E. Long-Term Proliferation and Survival of In vitro-Activated T Cells is Dependent on Interleukin-2 Receptor Signalling but not on the High-Affinity IL-2R. Scand. J. Immunol. 58, 522–532 (2003).
62. Vader, W. et al. The Gluten response in children with coeliac disease is directed toward multiple gliadin and glutenin peptides. Gastroenterology 122, 1729–1737 (2002).

63. Van De Wal, Y. et al. Small intestinal T cells of coeliac disease patients recognize a natural pepsin fragment of gliadin. Proc. Natl. Acad. Sci. U. S. A. 95, 10050–10054 (1998).
64. Kooy-Winkelaar, Y. & Koning, F. Isolation and Cloning of Gluten-Specific T Cells in Coeliac Disease. in Methods in Molecular Biology 53–59 (2015). doi:10.1007/978-1-4939-2839-2_6
65. Molberg, Ø., McAdam, S. N., Lundin, K. E. A. & Sollid, L. M. Studies of Gliadin-Specific T cells in Coeliac Disease. in Coeliac Disease 105–124 (Humana Press, 2000). doi:10.1385/1-59259-082-9:105
66. Yi, G. et al. Chromatin-Based Classification of Genetically Heterogeneous AMLs into Two Distinct Subtypes with Diverse Stemness Phenotypes. Cell Rep. 26, 1059-1069.e6 (2019).
67. R Core Team. R: A Language and Environment for Statistical Computing. (2019).
68. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (Springer-Verlag New York, 2016).
69. Gordon, A., Hannon, G. J. & others. Fastx-toolkit. FASTQ/A short-reads preprocessing tools (unpublished) http://hannonlab. cshl. edu/fastx_toolkit 5, (2010).
70. Kim, D., Langmead, B. & Salzberg, S. L. hisat2. Nat. Methods (2015).
71. Anders, S., Pyl, P. T. & Huber, W. HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169 (2015).
72. Love, A. M., Anders, S., Huber, W. & Love, M. M. Package ‘ DESeq2 ’. (2017).
73. Wu, M. & Gu, L. TCseq: Time course sequencing data analysis. R Packag. version 1, (2018).
74. Fabregat, A. et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 46, D649–D655 (2018).
75. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. ClusterProfiler: An R package for comparing biological themes among gene clusters. Omi. A J. Integr. Biol. 16, 284–287 (2012).
76. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
77. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Res. 25, 918–925 (2015).
78. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
79. Gaspar, J. Improved peak-calling with MACS2. bioRxiv 496521 (2018). doi:10.1101/496521
80. Stark, R., Brown, G. & others. DiffBind: differential binding analysis of ChIP-Seq peak data. R Packag. version 100, 3–4 (2011).
81. 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).
82. Wang, Y. et al. The 3D Genome Browser: A web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol. 19, 1–12 (2018).

Chapter 9

Fine-mapping of Coeliac disease associated genetic variants using SuRE-SNP implicates role for genetics in epithelial cells

R.V. Broekema [1], O.B. Bakker [1*], A. v.d. Graaf [1], J. Gelderloos-Arends [1], R. Modderman [1], L. Paige [2], B. v. Steensel [2], C. Wijmenga [1], S. Sanna 1, S. Withoff [1], J. v. Arensbergen [2] and I.H. Jonkers [1#]*

1 Department of Genetics, University of Groningen, University Medical Center Groningen, 9700 RB Groningen, the Netherlands
2 Netherlands Cancer Institute, 1066CX Amsterdam, the Netherlands

** These authors contributed equally*
# Corresponding author

Abstract

Over 40 genetic loci have been associated to coeliac disease (CeD) risk in the past years in several genome-wide association studies. It is not well understood how these loci exert their effects as most of the variants therein are non-coding and there is a lack of information on which of these variants are causal. Furthermore, recent studies have shown that the effects of non-coding genetic factors are highly cell-type and context dependent.

To get insight into the regulatory role of genetic variants in the pathophysiological context of CeD, we applied the Survey of Regulatory Elements & SNPs (SuRE-SNP), a linkage disequilibrium independent reporter assay that measures the intrinsic regulatory activity of genomic regions and the variants located therein. Using SuRE-SNP, we enrich CeD loci from patient DNA, allowing for a more targeted yet still high-throughput fine-mapping approach encompassing all disease associated SNPs. We applied SuRE-SNP in two representative cell-lines with and without stimulation: epithelial barrier cells of the small intestine (Caco-2) and CD4+ T cells representing the T cell mediated immune response (Jurkat). The intrinsic regulatory activity was dependent on the cell-type and context, which was reflected by the cell-type-specific transcription factor binding sites located in the active regions, and the mRNA levels of downstream genes. Moreover, we identified, and replicated by luciferase assay, three genetic variants that assert an allele-specific effect on the activity of their SuRE-SNP element. Two of these variants, rs2888524 and rs71327063, exhibit allele-specific effects in epithelial cells in the *CCR3/CCR5* locus thought to be linked to immune cells. Our findings suggest a role for CeD genetics in mediating both immune and epithelial barrier function, which ultimately impacts CeD development.

Key words

Coeliac Disease, SuRE-SNP, Fine-mapping, GWAS, ASE, Epithelial

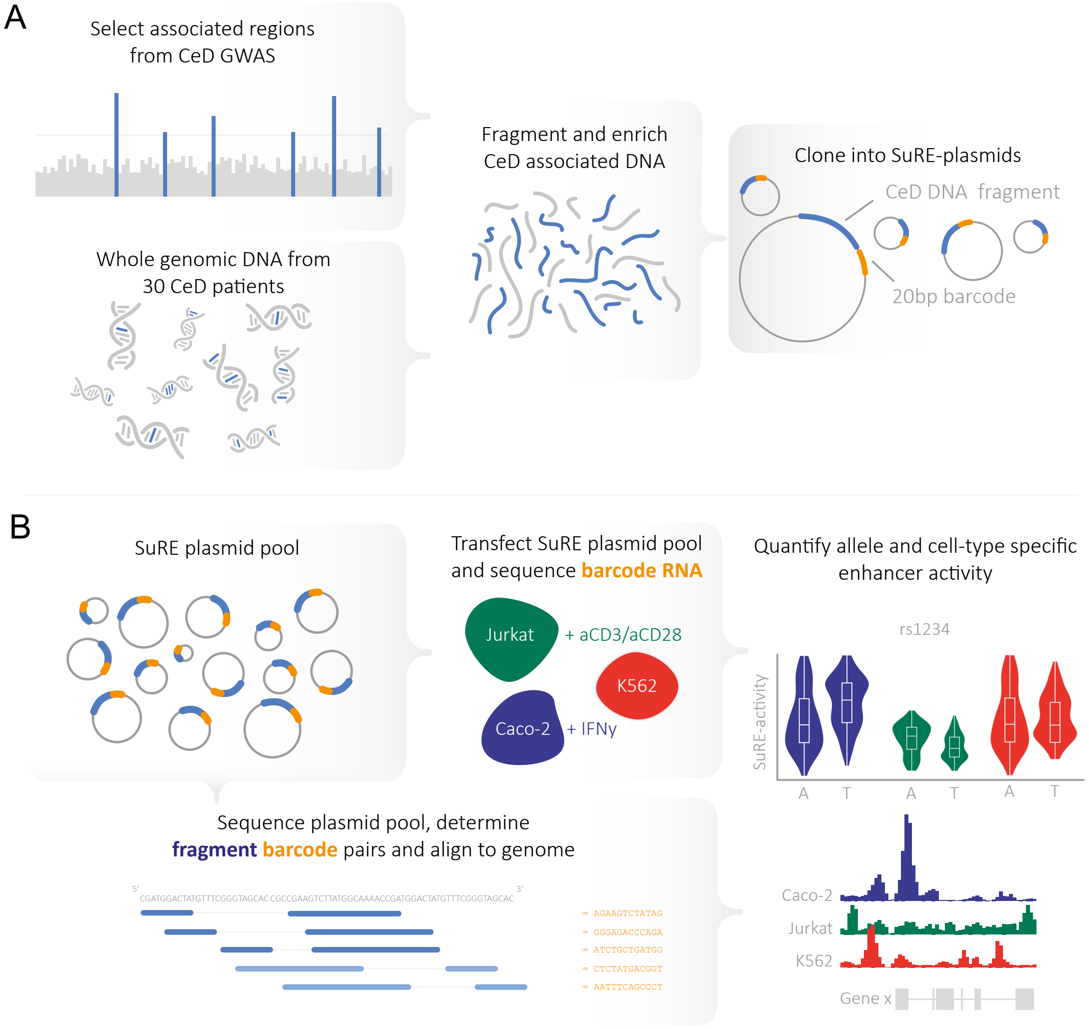
Introduction

Most complex diseases have a genetic component that partly determines susceptibility and outcome. This component has been well studied for a wide range of diseases in genome wide association studies (GWASs), revealing hundreds-of-thousands of disease-associated loci [1]. However, due to patterns of linkage between genetic variants resulting from recombination and evolution, it is still challenging to identify which exact variants within each GWAS locus are causal in affecting the disease susceptibility and outcome [2]. Most variants identified in GWASs are located in non-coding regions where they can affect genes in various direct and indirect ways. For example, non-coding variants can affect the expression of genes by altering transcription factor (TF) binding sites in regulatory elements (promoters and enhancers), or by affecting insulators which in turn change topologically associated domains. The major challenges in post-GWAS interpretation are fine-mapping which variants causally affect which genes, the mechanism by which those variants do this, and in what disease-relevant cell-type and stimulation-context this occurs [2].

A common approach to post-GWAS interpretation is to use expression quantitative trait loci (eQTLs) to link GWAS loci to genes. Such eQTLs were expected to be the answer to link GWAS loci to relevant target genes, and many efforts have used them to prioritize genes relevant for disease [3-6]. While eQTLs can help pinpoint the correct gene for a GWAS trait in certain contexts, they do not help to disentangle the linkage disequilibrium (LD) between variants, which is needed to pinpoint the causal variants. Moreover, recent work [7-9] has shown that eQTL prioritizations using current eQTL resources do not always give the correct link between variant and gene, possibly due to the lack of context specificity and lack of knowledge on regulatory element structure. Therefore, acquiring a good understanding of the causal genetic variants through fine-mapping is essential.

There are many approaches for fine-mapping causal genetic variants in genomic loci. For example, overlap of genetic variants with chromatin marks, functional elements, and TF binding sites [10,11] ,as well as computational models [12-14]. These methods are however limited by the fact that they do not explicitly test for a regulatory effect, but rather rely on association or the assumption that overlap constitutes change in activity of a functional element. *In vitro* assays such as massively parallel reporter assays (MPRAs) [15] are a high-throughput fine-mapping methods that assess the effects of genetic variants on regulatory activity. MPRAs test genetic variants in a cell-type specific and quantitative manner, and they rely on artificially generated DNA fragments. This makes variant screening static (one fragment length and sequence) and limited by short pre-defined genomic target locations. The MPRA-based method ‘survey of regulatory elements’ (SuRE) overcomes these limitations by assessing fragmented whole-genomic DNA to assess genome-wide regulatory activity) [15]. In the present study, we focus on coeliac disease (CeD) as it is an auto-immune disease of which the causative agent (gluten) and most relevant cell types associated with pathology are known (i.e., immune and epithelial cells). Briefly, in CeD the ingestion of dietary gluten instigates the immune system to damage the small intestinal epithelial barrier [16]. Genetic studies on CeD [17-19] have established that many of the associated genes are implicated in T cell receptor signaling and other immune processes. Less is known about the role of these genes in the affected epithelial cells. However, several genes associated with epithelial barrier function have been genetically implicated in CeD [20-22]. Thus, the mechanism and context in which barrier genes may play a role in the susceptibility and mechanism of CeD is of interest.

To study the role of CeD genetics in epithelial cells we applied SuRE-SNP, a modified version of SuRE with the aim of improved fine-mapping in GWAS loci. Instead of assaying the whole genome, SuRE-SNP enriches for the CeD loci from patient DNA, allowing for a more targeted yet still high-throughput fine-mapping approach of all CeD-associated SNPs carried by the patients at a MAF of 0.05. We assessed the intrinsic transcriptional activity within each entire locus, and the effect of genetic variants therein, using hundreds of millions of SuRE-SNP plasmid constructs. In the context of CeD we applied SuRE-SNP in three cell-lines with and without stimulation: Epithelial barrier cells of the small intestine (Caco-2) and CD4+ T cells representing the immune response (Jurkat), and K562 cells (erythroblasts), as a methodological control. We identified cell-type and stimulation specific SuRE peaks, enrichment of cell-type-related TF binding-motifs, and allele-specific expression (ASE) variants that were replicated by luciferase assay. Our results suggest a genetic role of epithelial cells in the development of CeD.


A
Select associated regions from CeD GWAS
Whole genomic DNA from 30 CeD patients
Fragment and enrich CeD associated DNA
Clone into SuRE-plasmids
CeD DNA fragment
20bp barcode
B
SuRE plasmid pool
Transfect SuRE plasmid pool and sequence barcode RNA
Jurkat
+ aCD3/aCD28
K562
Caco-2
+ IFNy
Quantify allele and cell-type specific enhancer activity
rs1234
SuRE-activity
A
T
A
T
A
T
Sequence plasmid pool, determine fragment barcode pairs and align to genome
5'
CGATGGACTATGTTTCGGGTAGCAC CGCCGAAGTCTTATGGCAAAACCGATGGACTATGTTTCGGGTAGCAC
3'
= AGAAGTCTATAG
= GGGAGACCCAGA
= ATCTGCTGATGG
= CTCTATGACGGT
= AATTTCAGCCCT
Caco-2
Jurkat
K562
Gene x

Results

Generation of a high-throughput CeD-specific SuRE-SNP plasmid library

To identify the candidate causal genetic variants in the CeD-associated GWAS loci, we applied the survey of regulatory elements and single-nucleotide polymorphisms (SuRE-SNP) method [15], a next generation reporter assay that can assay the intrinsic regulatory activity of specifically enriched DNA loci from patient material. We first selected all genome wide ($p<5x10^{-8}$) and suggestively significant loci ($p<5x10^{-5}$) from the latest CeD meta-analyses [17,18]. We then determined the regions of interest by extending the window around independent GWAS top variants from the most downstream to the most upstream variants with LD R^2 0.8, including 50kb padding on both sides of each locus. These thresholds were determined based on simulated GWAS data where the causal variants were known (**Fig. 1A**, **Suppl. Fig. 1**). We included regions within the HLA locus that have been associated with CeD independently from the HLA genes [23] but excluded most of the HLA region as the total amount of associated DNA was too large to include due to the strong LD in that locus. In total, this yielded 14 megabases (Mb) of DNA that contains associations to CeD (**Suppl. Tab. 1**). Whole-genomic DNA derived from 30 CeD patients was then selected from a large DNA sample cohort of CeD patients [17] by optimizing for the minor alleles of GWAS top-variants. The 30 DNA samples were fragmented to approximately 300bp, modified to include a custom designed patient-indexed adapter of 37bp on both ends of each fragment and used to generate a library of CeD enriched DNA by means of RNA-probe hybridization (**Fig. 1A**, Methods). These DNA fragments were then cloned into the custom promoterless SuRE plasmids, each tagged with a unique 20bp barcode (Methods) [15]. Fragments are randomly cloned into a plasmid in either forward or reverse orientation. This allows for testing bidirectionally similar to how enhancers and promoters can have bidirectional transcription [24]. Thus, we generated a CeD-specific SuRE-SNP library of which we sequenced around 331 million plasmids that we can use to couple the fragments to the unique 20bp barcodes in the plasmids (**Fig. 1A**).

Subsequently, this SuRE-SNP plasmid library was transfected into K562 (control), Caco-2 (epithelial cells) and Jurkat (CD4+ T cells) cell lines. Caco-2 and Jurkat cells were also stimulated for 3 hours with IFN-γ and aCD3/aCD28 respectively, to mimic conditions of active CeD. For each cell line 120 million cells were transfected in biological duplicates and split in two additional technical replicates to account for bias in library preparation (**Fig. 1B**). RNA containing the barcodes expressed from the DNA

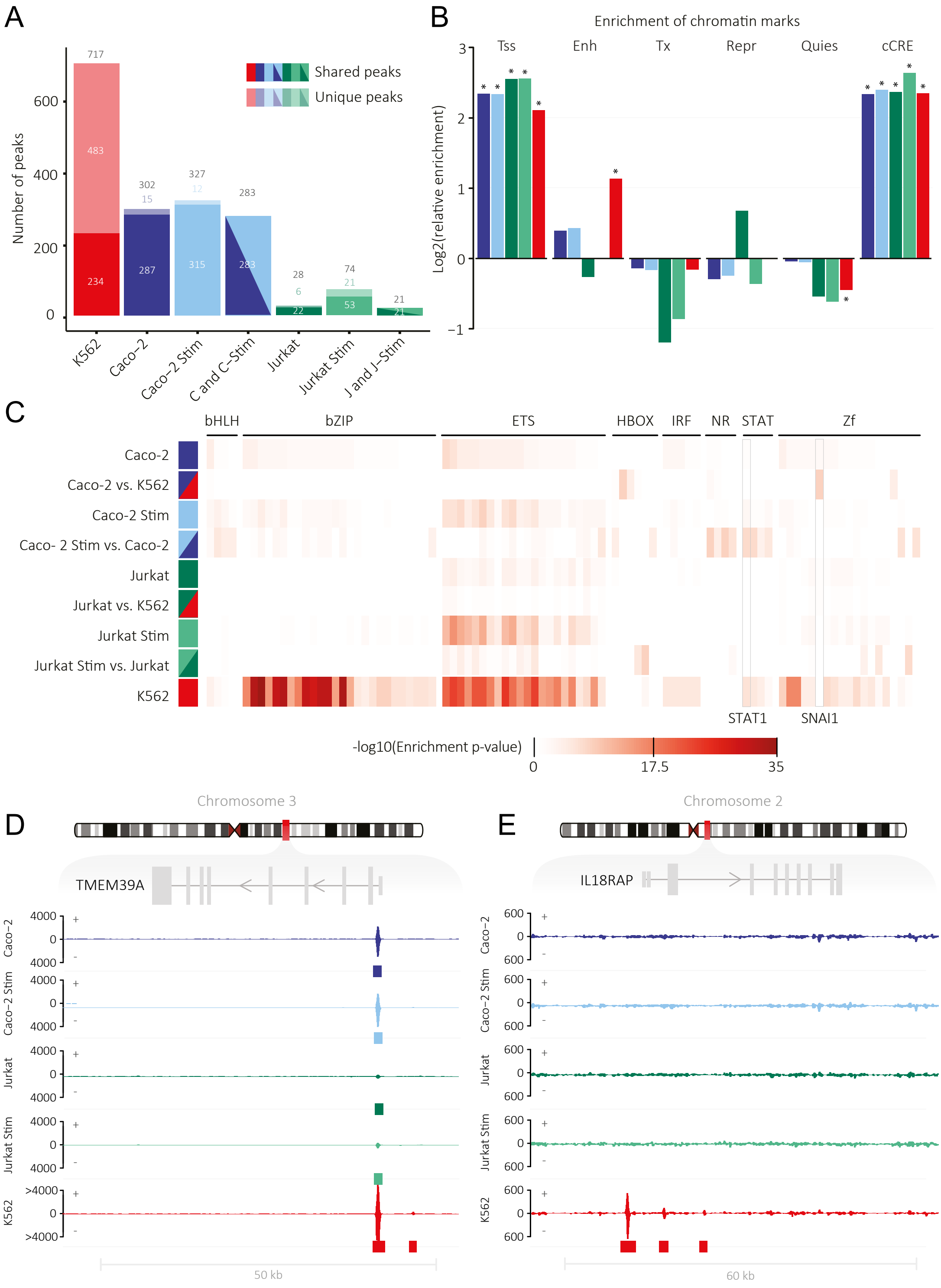
Left: Fig. 1. Schematic representation of SuRE-SNP study design. A) In the first step of the SuRE-SNP protocol, CeD GWAS loci were defined by selecting LD blocks around top variants (Methods). In conjunction, 30 CeD patients of European ancestry were selected, enriching the population for the CeD associated minor alleles with a MAF <= 10%. The DNA of these patients was extracted, fragmented, and enriched for the identified CeD associated LD blocks. This enriched library of DNA fragments was cloned into SuRE-plasmids which carry a unique 20bp reporter barcode. **B)** In the second step, the generated pool of plasmids is first sequenced to identify fragment - barcode pairs. The fragments are then aligned to the genome to identify their origin. Once fragment-barcode pairs were established, the SuRE plasmid pool was transfected into Jurkat (T cell line), Caco-2 (epithelial cell line) and K562 (Erythroblast cell line) cells. Jurkat cells were stimulated with aCD3/aCD28 to model the TCR activation that occurs in CeD. Caco-2 were stimulated with IFN-γ to mimic the inflammatory state of CeD. RNA-seq was performed to identify the expression of the 20bp barcodes, which informs on the intrinsic regulatory activity of the fragment associated with that barcode. Using the barcode expression and the barcode-fragment links, regions in the genome that have intrinsic regulatory activity were identified, as well as the genetic variants that disrupt this activity fragments in the plasmids of the SuRE-SNP library was isolated and sequenced. Overall, this strategy allowed us to determine intrinsic regulatory activity of all DNA in CeD loci with greater coverage compared to the regular SuRE protocol (**Suppl. Fig. 2**), resulting in better resolution for detecting the effects of most genetic variants associated with CeD on the functional elements.

Identifying intrinsic regulatory activity of CeD-associated DNA fragments

We identified genomic regions that showed intrinsic regulatory activity by calling peaks on the barcode expression using MACS2. In total we identified 717 peaks in K562, 302 in Caco-2, 327 in the stimulated Caco-2, 28 in Jurkat and 74 in stimulated Jurkat cells that were significant (FDR 0.05), showed at least a threefold enrichment over background and could be replicated in both biological replicates (**Fig. 2A**, **Suppl. Fig. 3**, **Suppl. Tab. 2**). Importantly, we observed substantial differences in the sequencing saturation and signal to noise ratio between the three cell-types we assayed, with the Jurkat, THP-1 and GM12878 cell-lines showing the lowest signal to noise ratios and THP-1 and GM12878 showing low complexity (**Suppl. Fig. 4**). While differences in transfection efficiency between each cell type may explain part of this problem, the majority of the noise seems to be derived from a-specific background expression of the SuRE-SNP plasmid library in CD4+ T cells and other immune cells. We expect the cause to be a possible interaction between specifically expressed transcription factors and the backbone of the plasmid (discussion).

To investigate if our SuRE peaks locate in known functional elements, we overlapped them with the genomic element definitions by the Epigenome Roadmap [11] and Encode [10] consortia. Depending on cell-type, the peaks located mostly (27-60% of peaks) in transcription start sites (TSS) and enhancers (Enh) as defined by Epigenome Roadmap (**Suppl. Fig. 5**). The other 40-73% overlap with quiescent regions, active transcription (Tx) and repressed regions (Repr), likely attributable to the fact that SuRE-SNP measures the intrinsic activity of a fragment independently of the chromatin state [15]. Similarly, 20-47% of our SuRE peaks locate in known candidate *cis*-regulatory elements (cCRE) from the Encode consortium (**Suppl. Fig. 5**). Given that there is a discrepancy between the number of elements annotated for each functional class (i.e., there are more quiescent elements in the genome than there are enhancer elements), we next

Right Fig. 2. Characterization of regions that show intrinsic regulatory activity. A) Bar plot showing the number of peaks that were identified using MACS2 in each cell type. Peaks were defined by a 3-fold enrichment over the plasmid coverage and were shared between biological replicates. Colours indicate the different cell types. **B)** Enrichment analysis of the peaks and the regulatory elements identified by Epigenome Roadmap and the candidate cis regulatory elements (cCRE) from the ENCODE consortium. Enrichment statistics were calculated compared to 10.000 randomly sampled regions (methods). * Indicates a significant enrichment, adjusting for 28 tests. Epigenome roadmap's ChromHMM elements were grouped as transcription start site (Tss; 1_TssA + 2_TssAFlnk + 10_TssBiv + 11_BivFlnk), enhancer (Enh; 6_EnhG + 7_Enh + 12_EnhBiv), transcription (Tx; 3_TxFlnk + 4_Tx + 5_TxWk), repressive (Repr; 8_ZNF/Rpts + 9_Het + 13_ReprPC + 14_ReprPCWk), and Quiescent (Quies; 15_Quies) states. **C)** Transcription factor binding enrichment analysis (HOMER) for the peaks. Enrichment was calculated for a genome wide background for Caco-2, Jurkat and K562. For Caco-2 Stim and Jurkat Stim we calculated enrichment against a genome wide background but added a second condition comparing the enrichment with respect to the baseline state. We added two more conditions, comparing Caco-2 and Jurkat vs. K562. **D)** Example of an active region present in all cell-types assayed located in the promoter of *TMEM39A*. Each track represents the SuRE-activity in a different cell-line. + indicates the activity stemming from reads on the forward orientation, - indicates the activity in the reverse orientation. **E)** As D but showing three peaks in the promoter of *IL18RAP* that are only present in K562 data.


A
Number of peaks
0
200
400
600
717
483
234
302
15
287
327
12
315
283
283
28
6
22
74
21
53
21
21
K562
Caco−2
Caco−2 Stim
C and C−Stim
Jurkat
Jurkat Stim
J and J−Stim
Shared peaks
Unique peaks
B
Enrichment of chromatin marks
Log2(relative enrichment)
3
2
1
0
−1
Tss
Enh
Tx
Repr
Quies
cCRE
*
C
bHLH
bZIP
ETS
HBOX
IRF
NR
STAT
Zf
Caco-2
Caco-2 vs. K562
Caco-2 Stim
Caco- 2 Stim vs. Caco-2
Jurkat
Jurkat vs. K562
Jurkat Stim
Jurkat Stim vs. Jurkat
K562
STAT1
SNAI1
-log10(Enrichment p-value)
0
17.5
35
D
Chromosome 3
TMEM39A
Caco−2
Caco−2 Stim
Jurkat
Jurkat Stim
K562
4000
0
4000
>4000
0
>4000
+
-
50 kb
E
Chromosome 2
IL18RAP
Caco−2
Caco−2 Stim
Jurkat
Jurkat Stim
K562
600
0
600
+
-
60 kb

evaluated if the functional element overlap with SuRE peaks represented an enrichment (**Fig. 2B**). We found that the peaks were most strongly enriched in TSSs and cCREs for all cell-types. Regions of active transcription (Tx) and quiescent (Quies) regions were slightly depleted of SuRE peaks. Peaks from K562 were also significantly enriched for enhancer marks (Enh), while Caco-2 and Jurkat samples did not show an enrichment for enhancer marks. This is likely due to the relatively modest number of peaks detected in these cell-types (**Fig. 2A**).

We next assessed if the peaks were enriched for particular transcription factor binding motifs. To do so we applied HOMER to screen for enrichment of 440 known vertebrate TF motifs [25]. We tested genome-wide backgrounds for each cell-type, and additionally evaluated the enrichment compared to the other cell-types and conditions, to assess any cell-type specific TF binding in our assay. We observed that binding sites of basic leucine zipper (bZIP) and especially E26 transformation-specific (ETS) family TFs were enriched in all cell-types (**Fig. 2C**). The TFs in these families cover a multitude of different ubiquitous processes, as such it is expected that they are enriched in all cell-types [26,27]. When we assessed cell-type specific TF binding enrichment in Caco-2 (by comparing to K562) we observed that SNAI1 binding sites were most significantly enriched. SNAI1 is a TF that is involved in the regulation of E-cadherin (an epithelial cell adhesion molecule) and has a role in modulating the epithelial to mesenchymal transition of epithelial cells [28]. This highlights that the SuRE peaks we identify are representative of cell-type specific processes mediated through specific TF binding.

Between the cell-types strong correlation (Pearson r ~0.8) could be observed, showing that a large portion of peaks are shared across cell-types (**Suppl. Fig. 6**). For example, a peak, active in all assayed cell-types, locates to the promoter region of *TMEM39A* (**Fig. 2D**). *TMEM39A* was in the > 65Th percentile of expression in all cell-types. In primary tissues, *TMEM39A* is most abundantly expressed in fibroblasts, but also active in most other cell-types [29].

Despite substantial correlation between cell-types, the activity in some of the peaks showed cell-type specific patterns (**Suppl. Fig. 6A-C**). For example, three peaks around the promoter of *IL18RAP* which are uniquely active in K562 cells (**Fig. 2E**). *IL18RAP* was moderately (40Th percentile) expressed in K562 cells but not in Caco-2 and Jurkats (< 10Th percentile), highlighting that the SuRE-activity (in promoters) tends to be indicative of gene expression levels. The correlation patterns were much stronger between peaks in promoter regions (< 1kb from TSS) and proximal enhancers (< 5kb from TSS) compared to distal enhancers (> 5kb from TSS) which showed lower correlation between cell-types and lower activity overall compared to promoter and proximal enhancer peaks (**Suppl. Fig. 6A-D**).

Regulatory activity impacts gene expression in a stimulation specific manner

To ascertain how SuRE-activity relates to gene expression we looked for genes impacted by the stimulation of the Caco-2 and Jurkat cells and correlated these to the activity of the regulatory elements identified. The pattern of cell-type specificity observed in the SuRE-activity was mirrored at the RNA level (**Fig. 3A**). We then looked for changes in SuRE-activity within 1kb of the TSSs of the differentially expressed (DE) genes in either IFN-γ stimulated Caco-2 or aCD3/aCD28 stimulated Jurkat (**Suppl. Tab. 3**) and found a significant increase in SuRE-activity near DE genes in Caco-2 but not in Jurkat (**Fig.

3B). This suggests that these *cis*-acting SuRE peaks are involved in regulation of the DE genes, at least in Caco-2. Notable examples of this are *IRF1* and *STAT1* for Caco-2 (**Fig. 3C,D**) and *BACH2* and *ICOS* for Jurkats (**Fig. 3E,F**), as we observed both increased SuRE-activity in their promoter regions as well as them being DE at the mRNA level. The lack of correlation between SuRE activity and DE gene expression in Jurkats is likely due to low signal to noise ratio in the Jurkat SuRE assay.

IRF1 was not expressed in baseline Caco-2 cells but saw a significant upregulation upon IFN-γ stimulation (**Fig. 3C**, log2 fold change: 5.72, p-value: $p<2.2x10^{-308}$). *STAT1* was already expressed in Caco-2 prior to stimulation, but its expression was significantly increased upon stimulation (**Fig. 3D**, log2 fold change: 1.79, p-value: $6.82x10^{-11}$). *IRF1* and *STAT1* encoding TFs of the same name are both known to be activated upon IFN-γ

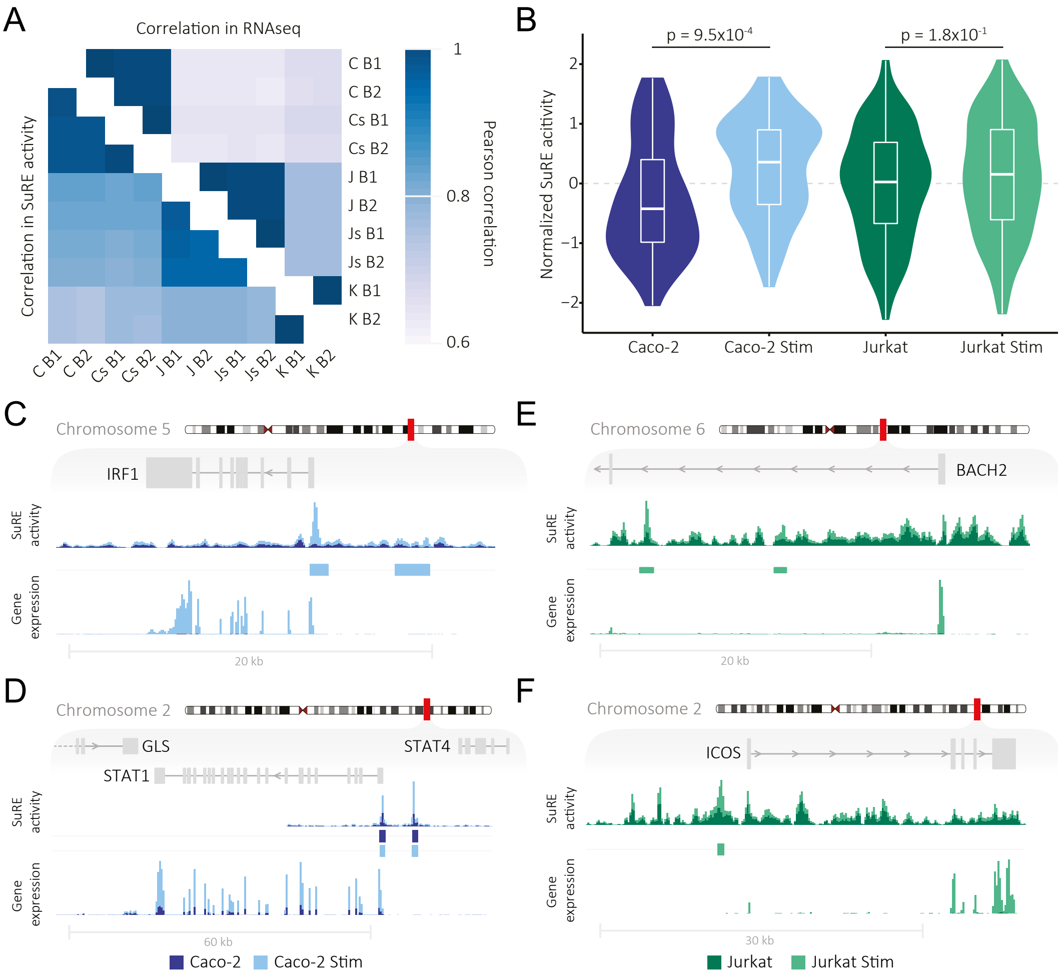


Fig. 3. Stimulation specific effects on SuRE-activity coincide with expression of downstream genes. A) Correlation heatmap between the replicates in RNA-seq data (upper triangle) and SuRE-activity (lower triangle). Samples cluster in a similar fashion in both data. **B)** Activity of SuRE peaks located in promoter regions (< 1kb upstream or overlapping TSS) of differentially expressed genes. Activity of the peaks in the promoters of differentially expressed genes is increased in Caco-2 stim but not in Jurkat (discussion). **C-F)** Examples showing the regions around four differentially expressed genes, *IRF1* and *STAT1* in Caco-2 and *BACH2* and *ICOS* in Jurkats. The top track shows the SuRE-activity in the regions, the bottom tracks the RNA-seq expression.

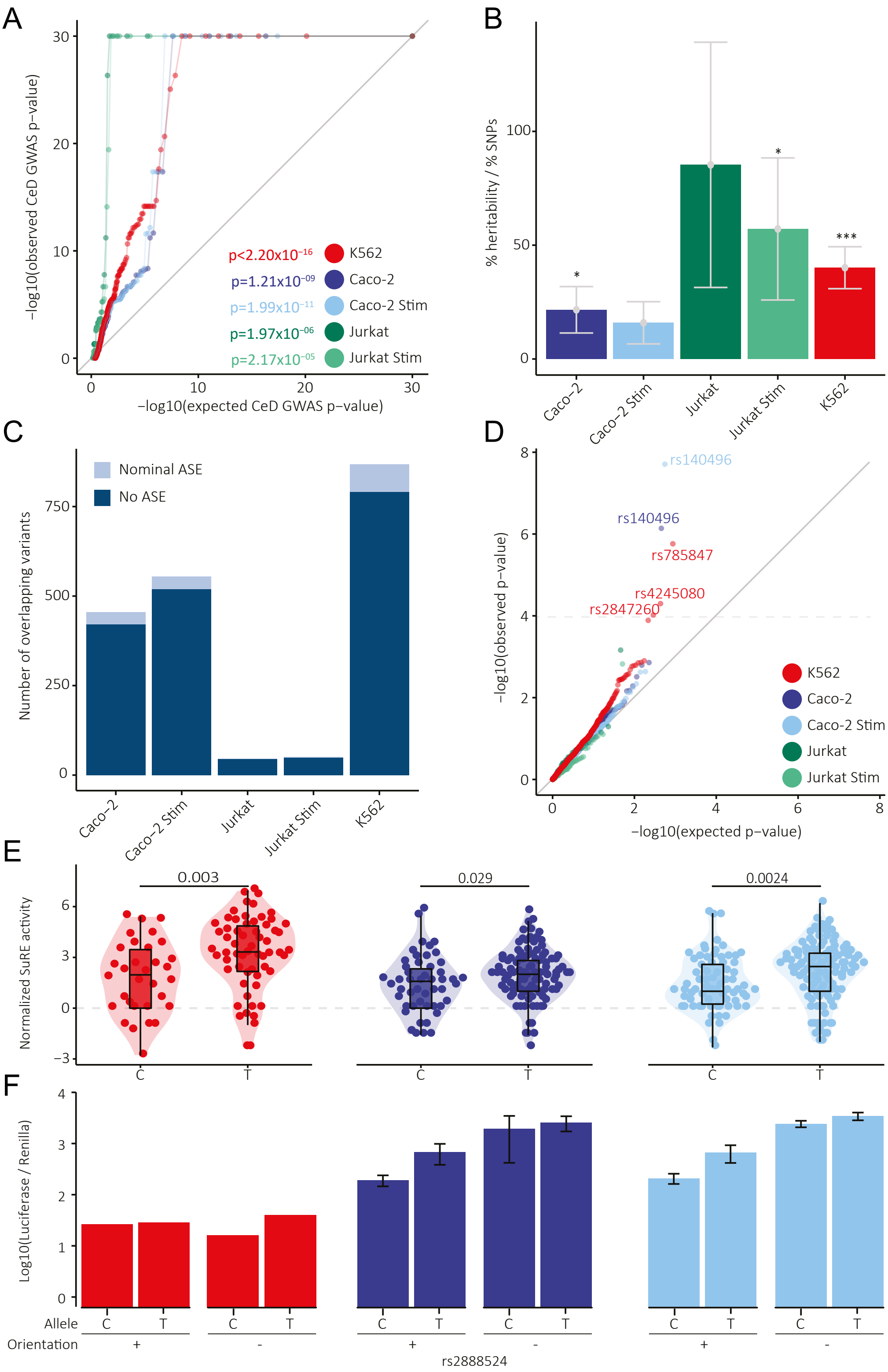
stimulation and to regulate the expression of hundreds of downstream targets [30,31]. IRF1 is known to be mostly expressed by immune cells but is also present in the small intestinal epithelium [29] and has been shown to regulate the expression of cytokines by epithelial cells [32]. In CeD, IFN-γ is produced when gluten-specific T cells are activated [33,34], which likely activates *IRF1* and *STAT1* in the cells of the epithelial barrier. Furthermore, when assaying which TF binding sites were enriched under SuRE peaks in stimulated Caco-2 compared to unstimulated, we found that STAT1 sites were significantly enriched (**Fig. 2C**). Together this indicates a robust activation of downstream IFN-γ signaling in epithelial cells, mediated through STAT1 and IRF1. Dysregulation of the promoter regions of *STAT1* and *IRF1* regions by genetic factors may therefore have wide reaching effects on the response to IFN-γ by epithelial cells. Indeed, rs2549005, a SNP suggestively associated with CeD ($p<5x10^{-5}$), located in this SuRE peak shows a significant eQTL effect on *IRF1* in blood [4]. The SNP rs41430444, also located in a SuRE peak and suggestively associated with CeD, is an eQTL for *STAT1* [4]. This suggest that these SNPs, located in SuRE peaks, have the potential to regulate the gene expression of *IRF1* and *STAT1*.

BACH2 and *ICOS*, both showing increased regulatory activity and expression in stimulated Jurkats (**Fig. 3E,F**, expression log-2 fold changes: 3.49 and 3.66, p-values: $2.42x10^{-240}$ and $1.78x10^{-9}$ respectively), are key factors for TCR signaling and T cell function. BACH2 is an essential TF with a key role in both T and B cell function and Mendelian variants in BACH2 lead to primary immune deficiencies [35]. ICOS, expressed by gluten-specific T cells, acts as a co-stimulatory checkpoint and can induce the proliferation of several Th subsets [36]. Hence, genetic factors altering the expression of *ICOS* and *BACH2* may impact the severity of the inflammatory response in CeD [37]. With regards to genetic effects that may impact the regulatory elements, for *BACH2*, the SNP rs905671, located in an intronic SuRE peak (**Fig. 3E**), is significantly associated with CeD ($p<5x10^{-8}$) and has an eQTL on BACH2 in blood [4]. The SNP rs11571306 located in the SuRE peak in the *ICOS* promoter (**Fig. 3F**) is associated with CeD ($p<5x10^{-8}$) and has a significant eQTL effect for *ICOS* in blood [4]. Thus, stimulation specific regulatory regions may be affected by genetic variants associated with CeD.

Enrichment of CeD heritability in regions with regulatory activity

Next, we looked for genetic variants overlapping the SuRE peaks to assess their effects on the regulatory potential of the peak, and subsequently their involvement in CeD (**Suppl. Tab. 4**). We observed that genetic variants located in SuRE peaks were significantly enriched for being associated to CeD in a cell-type specific manner (**Fig. 4A,B**). The strongest effect was observed for Jurkat cells, followed by K562 and finally Caco-2.

Right: Fig. 4. CeD heritability enrichment in SuRE peaks and allele-specific expression of SuRE plasmids. A) QQplot showing the distribution of CeD associated p-values for SNPs overlapping with SuRE-peaks versus the distribution of CeD p-values of non-overlapping SNPs. p-values show the significance of the Kolmogorov–Smirnov test between the two distributions. This test was done for each cell-type separately. **B)** Enrichment of heritability under SuRE-peaks by stratified LD score regression. p-values from LD score regression analysis are indicated by * $p<5x10^{-2}$, ** $p<5x10^{-3}$, *** $p<5x10^{-4}$. **C)** Number of genetic variants (1000 genomes MAF 5%) that overlap with SuRE-peaks in CeD loci. Light blue indicates the number of variants for which a nominal (un-adjusted $p<0.05$) ASE effect was found. **D)** QQplot of the ASE analysis p-values. The dotted line marks Bonferroni significance ($1x10^{-4}$), all SNPs passing this threshold are written out in full. **E)** ASE effect of rs2888524 in the SuRE-data for K562, Caco-2 and Caco-2 stim. **F)** ASE effect of rs2888524 based on the luciferase assay. The luciferase assay tests the effects in both orientations which are indicated by + and – respectively.


A
−log10(observed CeD GWAS p−value)
−log10(expected CeD GWAS p−value)
0
10
20
30
p<2.20x10⁻¹⁶ K562
p=1.21x10⁻⁰⁹ Caco-2
p=1.99x10⁻¹¹ Caco-2 Stim
p=1.97x10⁻⁰⁶ Jurkat
p=2.17x10⁻⁰⁵ Jurkat Stim
B
% heritability / % SNPs
0
50
100
*
*
***
Caco−2
Caco−2 Stim
Jurkat
Jurkat Stim
K562
C
Number of overlapping variants
Nominal ASE
No ASE
0
250
500
750
Caco−2
Caco−2 Stim
Jurkat
Jurkat Stim
K562
D
−log10(observed p−value)
−log10(expected p−value)
0
2
4
6
8
rs140496
rs140496
rs785847
rs4245080
rs2847260
K562
Caco-2
Caco-2 Stim
Jurkat
Jurkat Stim
E
Normalized SuRE activity
−3
0
3
6
0.003
0.029
0.0024
C
T
C
T
C
T
F
Log10(Luciferase / Renilla)
0
1
2
3
4
Allele
C
T
C
T
C
T
C
T
C
T
C
T
Orientation
+
-
+
-
+
-
rs2888524

To see if there was also an enrichment for the heritability of CeD we applied stratified LD score regression [38] using the SuRE peaks and found significant enrichment of heritability in stimulated Jurkats, K562 and Caco-2 cells (**Fig. 4B**). The enrichment in stimulated Jurkats is striking, as we only detected 74 peaks, indicating that the variants located in these peaks have a relatively high effect size on CeD risk. While the T cell signature in CeD genetics is well described [20,22], less is known about the role of CeD genetics on epithelial cells. This is partially due to the CeD GWAS having been performed on the Immunochip platform, so a bias towards immune-gene rich loci exists. Nevertheless, we do observe enrichment in Caco-2 cells which points to a potential role for either immune genes in epithelial cell function or barrier-related genes within immune-gene rich loci.

Allele specific regulatory activity by CeD-associated SNPs

To quantify if specific genetic variants had an impact on the intrinsic transcriptional activity of the DNA fragments in the CeD-specific SuRE-SNP library, we first called SNPs, insertions and deletions (INDELs) in the library of genetic fragments we obtained from the 30 CeD patients. In total we called ~ 64K (bi-allelic) variants at a minor allele frequency (MAF) of 5% genome wide. We then assigned each fragment which overlapped with a variant in a CeD locus to an allele. This gave us two pools of fragments for each variant, representing the two alleles. We then compared the mean activity between each of these pools using a Mann-Whitney U-test to estimate the significance of the allele specific expression (ASE). Additionally, we further split these pools into either forward or reverse orientation fragments to test for orientation specific ASE. We considered only variants that overlapped with a SuRE peak for interpretation. This yielded 868 variants for K562, 455 for Caco-2, 555 for stimulated Caco-2, 46 for Jurkat and 51 for stimulated Jurkats. Overall, the power to detect ASE effects was limited with only a small proportion of the variants overlapping regulatory elements attaining nominal significance (**Fig. 4C,D**, **Suppl. Tab. 4**). We identified four SNPs which attained Bonferroni significance, rs140496 (Caco-2 & Caco-2 stim), rs785847, rs4245080 and rs2847260 (K562, **Fig 4D**).

We selected a total of seven ASE effects and controls to independently validate using a luciferase assay. We included the two Bonferroni significant top effects in K562 and Caco-2 cells, a fine-mapped systemic lupus erythematosus SNP (rs140490) [39] that was not located in a SuRE peak and did not show ASE, as a negative control, and added four additional SNPs. We selected these four SNPs based on the following criteria: the SNP must 1) have a nominally significant ASE effect, 2) be at least suggestively associated to CeD itself, or have an LD proxy (R^2>0.8) that is, and 3) be located in a SuRE peak.

The Bonferroni significant K562 top ASE effect (rs785847) showed consistent directions of effect in the luciferase assay (**Suppl. Fig. 7**). In addition, our negative control SNP rs140490 also consistently does not show luciferase activity nor any allelic effect (**Suppl. Fig. 8**). The fragment with the only Bonferroni significant ASE effect we observed in Caco-2 cells (rs140496, not associated to CeD) was not very active in the luciferase assay (**Suppl. Fig. 9**). While there were clear differences between the alleles with the G-allele seemingly being repressed in the minus orientation, they were opposite to what we observed in the ASE analysis where the G-allele shows increased transcriptional activity in the plus orientation. Two of the four additionally selected SNPs, rs2888524 and rs71327063, that are both part of the *CCR3/CCR5* locus, also replicated the ASE (3-fold expression difference) in Caco-2 cells (**Fig. 4E,F**, **Suppl. Fig. 10**). Unfortunately, luciferase could not replicate the increased C-allele activity in Caco-2 cells of SNP rs11072504 (**Suppl. Fig. 11**). Rs55950816 in the HLA-region could also not be replicated and showed large standard deviations between three luciferase biological replicates leaving some uncertainties (**Suppl. Fig. 12**).

In summary, four of the (nominal) ASE effects identified by SuRE-SNP can be replicated in an independent luciferase assay, however, the assay method may affect the strength and directionality of the ASE effect in some cases.

Disruption of TF binding motifs by genetic variants located in SuRE peaks

Finally, we assessed possible TF-binding motif disruption by each of the seven selected SNPs to understand how the three replicated ASE are caused, why three do not replicate, and why the negative control does indeed not have an ASE in Caco-2 cells. We tested 440 known TF motifs for each SNP allele separately in a 30bp window centered around the SNP using HOMER [25] (**Fig. 5**, **Suppl. Fig. 7-12**).

The ASE effect in Caco-2 from SNP rs2888524, located in the *CCR3/CCR5* locus (**Fig. 5A**), points to a change in the LRF TF-motif, as the T-allele of rs2888524 is located where only a C-nucleotide is tolerated by LRF (**Fig. 5B,C**). *LRF*, also known as *ZBTB7A*, is a zinc-finger gene that was previously shown to repress genes by itself or by maintaining nucleosome occupancy by recruiting the nucleosome and remodeling deacetylase (NuRD) complex to repress gene expression [40,41]. This is in-line with our observation that the C-allele has a lower expression than the T-allele (**Fig. 4E**).

The other CCR-locus ASE SNP we identified is rs71327063, with its G-allele providing stronger transcription in Caco-2 cells (**Suppl. Fig. 10**). We identified a TF-motif for SMAD3 in the A allele, but not the G allele (**Suppl. Fig. 10**). In addition, a small cluster of AP-1 subunit TF-motifs are present up-stream of the SNP and SMAD3 TF-motif. SMAD3 has been shown to interact with AP-1 subunits to regulate promoter activity [42-44]. In our case the lower SuRE-activity and binding site for SMAD3 in the A-allele suggests a mechanism of repression by SMAD3, possibly in combination with AP-1 [45,46]. This effect could also potentially be mediated through SMAD3 and SNAI1 interactions, as this is a known TF interaction that suppresses promoters during epithelial-mesenchymal transition [47,48]. A second TF overlapping the SNP, ZNF711, is a ubiquitously present yet uncharacterized TF [49] and, although it is not as strongly expressed in Caco-2 as SMAD3, its TF-motif has zero tolerance for the A-allele of SNP rs71327063 and thus may affect binding of ZNF711. We further assessed the potential disruption of TF binding as an explanation for the ASE effects in the other five SNPs selected for follow up (**Suppl. Fig. 7-12**) and found potential disruption of TF binding in four cases.

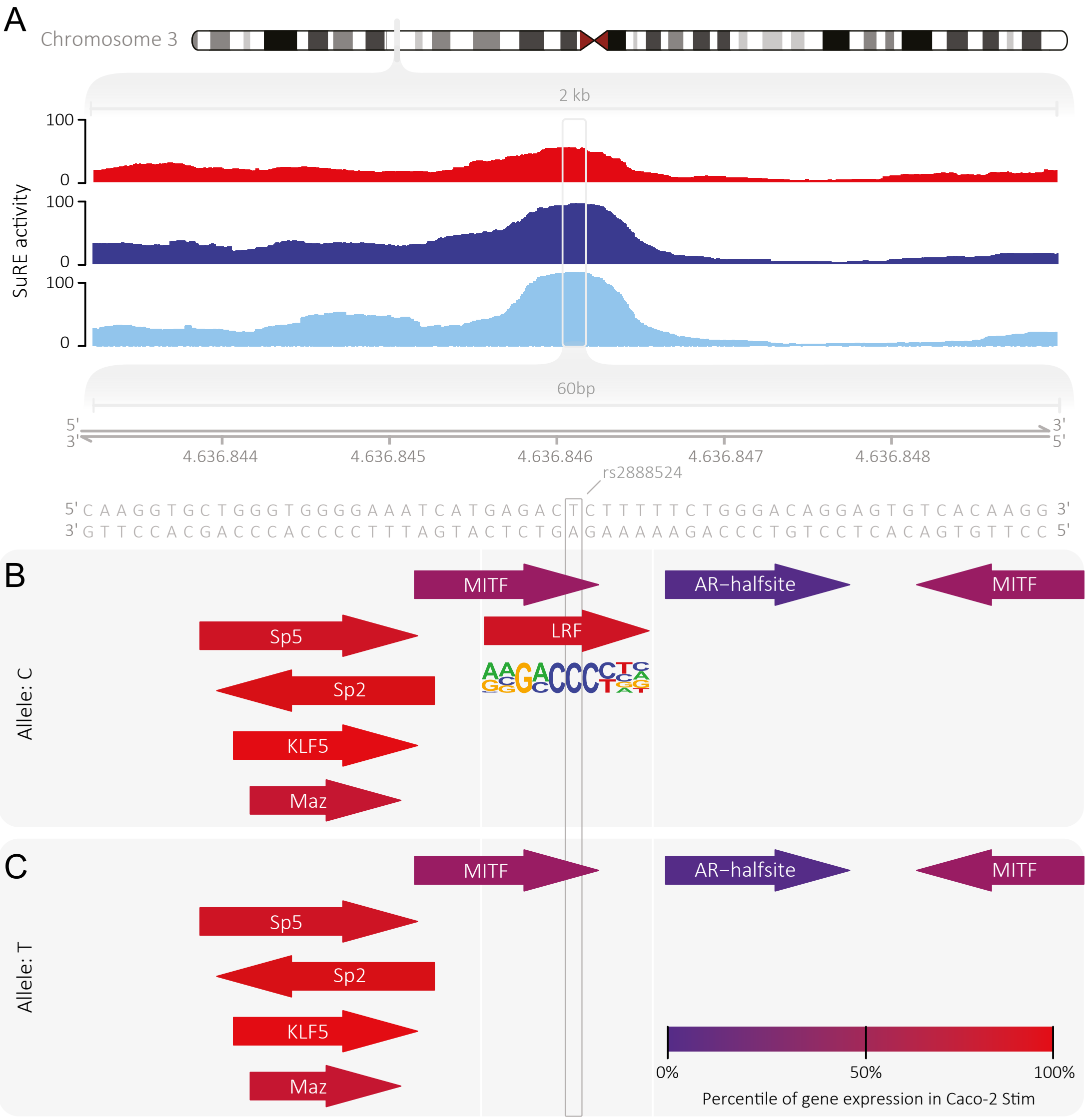


Fig. 5. Disruption of LRF binding motif as a possible mechanism for the regulatory effect of rs2888524. A) The SNP rs2888524, located centrally in the SuRE-peak that is active in K562, Caco-2 and most strongly in Caco-2 Stim. **B-C)** Zoomed in view of 60bp window around rs2888524. Results of TF binding analysis (HOMER) in this 60bp window for the C-allele of rs2888524. The C-allele (panel B) shows a binding site for LRF which is not present in the T-allele (panel C) due to a highly conserved C in the binding motif. Colour of the arrows indicates the expression of the associated gene in Caco-2 Stim as a percentile, with red being highly expressed, and blue lowly expressed.

Discussion

In this work we present SuRE-SNP, a reporter assay with the goal of identifying cell-type specific genetic regulatory effects of SNPs on promoters and enhancers. We applied SuRE-SNP to study CeD and identified several hundred regulatory elements in K562, Jurkat and Caco-2 cells in CeD-associated genetic loci. We show that the intrinsic regulatory activity of CeD-associated DNA is context and cell-type dependent. We highlight that the SuRE-elements are indicative of downstream regulation by showing that they modulate genes involved in the type-2 interferon response in epithelial cells upon IFN-γ stimulation. Furthermore, several SuRE-elements contain genetic variants for which we assessed regulatory potential by ASE analysis. Based on TF-motif analysis we identified the TF LRF/ZBTB7A as a likely candidate to regulate a regulatory element at SNP rs2888524. Additionally, another regulatory element at SNP rs71327063 is possibly regulated by repression at the A-allele through a SMAD3 and AP-1 TF-complex.

Both ASE SNPs rs2888524 and rs71327063 locate to the CCR locus on chromosome-3 (**Suppl. Fig. 13**). Within this locus these SNPs likely regulate one or perhaps multiple local genes such as the immune-related CCR genes or the barrier-related *LZTFL1* and *SACM1L* genes, as also associated by eQTL analysis in blood [4]. While the genes *CCR1, CCR2, CCR3* and *CCR5* are the closest genes to both SNPs, these CCR-genes are lowly expressed in our Caco-2 expression data. C-C chemokine receptor (CCR) genes are involved in immune cell signaling as receptors to chemokines excreted by inflammatory responses and are mostly expressed in T cells and monocyte-macrophages [50]. As SNPs rs2888524 (T allele in perfect LD with proxy variant rs35053103's G allele OR:1.18) and rs71327063 (A allele in LD, R^2 0.84, with proxy variant rs34671664's C allele OR: 1.27) identified in the CCR locus associated with CeD appear to facilitate repression of activity, via the TFs LRF/ZBTB7A and SMAD3 respectively, they may act to repress CCR expression. Alternatively, the affected regulatory element may also promote expression, either directly or indirectly, over longer distances. The genes *LZTFL1* and *SACM1L* are strongly expressed in our Caco-2 cells but are more remotely located from both SNPs (>300- 500Kb). LZTFL1 functions as a barrier protein by colocalizing with e-cadherin and interacting with the epithelial-mesenchymal transition (EMT) pathway [51]. SACM1L is a membrane phosphatase involved in vesicle and membrane regulation at the endoplasmic reticulum (ER), Golgi, and plasma membrane [52]. Thus, the identified SNPs in this locus could transmit downstream effects on several genes that may have a functional consequence in the context of CeD. However, the targets of the affected enhancers need to be identified through additional functional experiments.

SuRE-SNP is an unbiased method to identify and characterize functional elements in genomic loci associated with a specific disease without prior knowledge of the structure and location of enhancers and promoters. Additionally, as DNA fragments are sequenced directly, effects of *de novo* SNPs, deletions and insertions can be assayed. Despite these advantages, we identified challenges with translating the SuRE protocol to different cell-types. One striking observation was that the background expression of the SuRE plasmid appears to be cell-type specific. In K562 cells, the cell-type with which SuRE was originally developed, we observed a relatively clean signal with clear peaks and detectable enhancer effects (**Fig. 2**, **Suppl. Fig. 4**). In Caco-2 we also saw clear signal, however there was noticeably more background expression present as compared to the K562 signals. For Jurkat cells, we detected some signals that could be replicated, but most signal was obscured by strong background expression creating a very poor signal-to-noise ratio in this immune cell type. Similar results were also obtained for GM12878 (B cells) and THP-1 (monocytes) (**Suppl. Fig. 4C**) however, these cells also showed low complexity in the barcode expression, whereas Jurkats did not (**Suppl. Fig. 4A-B**). As the poor signal-to-noise ratio was observed only in immune cell-types, we speculate that there could be initiation of transcription at cell-type specific TF binding sites in the backbone of the SuRE-plasmid. While we extensively screened for this (data not shown) we could not conclusively pinpoint if and where initiation of transcription in the SuRE-plasmid backbone occurred.

Unfortunately, the high background expression in combination with the large complexity of our SuRE-SNP library has a significant impact on the power for detecting ASE effects, especially in Jurkat T cells. While the coverage in the plasmid pool per SNP was good (on average 877 fragments), many plasmid barcodes were not detected after sequencing of cDNA, and final coverage was much lower, with on average 330 fragments per SNP found to be expressed at least once in the cDNA (~37% of the input). Hence, we only managed to detect four Bonferroni significant ASE effects. This presumably leads to our SuRE-SNP being underpowered for the majority of SNPs. Indeed, if we equate the number of fragments covering the variants to samples in an eQTL study, we can see the "sample size" is limited compared to for example the largest blood eQTL study to date [4] which would cover the minor allele of a MAF 5% variant with ~1500 samples as opposed to the 17 independent measurements (the 330 unique fragments times the 0.05 MAF in our study). In this case, the coverage cannot easily be increased by sequencing more of the plasmid pool, as this would also increase noise, leading to a point of diminishing returns.

Another factor that may impact our ability to detect ASE effects is the fact that the strongest SuRE-activity tends to be located in promoter regions (**Suppl. Fig. 6D**). Promoter regions are generally more evolutionally constrained compared to enhancer domains [53]. Therefore, any genetic effects exerted by common variants located in these promoter peaks, are likely to be very small or not even present due to selection. Hence, common genetic factors might have more of an impact on enhancers than promoters. Given that we detected the strongest in or near promoters in our assay, especially in Jurkats, we might be missing genetic regulation by common variants overlapping enhancers.

To verify that the ASE effects we observe are not false positives, we tested seven of these effects (including controls) using a luciferase assay and were able to replicate four of them (**Fig. 4**, **Suppl. Fig. 7-12**). However, the other three showed either low activity in the luciferase assay or inconsistent directions of effect. SNP rs140496 shows very low luciferase activity while we identify it as overlapping with a moderately active SuRE peak (**Suppl. Fig. 9**). Moreover, even with its low activity there appears to be an allelic effect in the opposite allelic direction and on the opposite strand orientation as compared to our measured SuRE-ASE. Similarly, SNP rs55950816 also shows opposite allelic direction and strand orientation, even though activity in the luciferase assay is clearly detected (**Suppl. Fig. 12**). These discrepancies are not new, nor are they limited to just our SuRE and luciferase plasmids. Subtle or vast read-out differences can be observed between similar reporter assay tests when 1) using different reporter genes, in our case the unique barcode versus luciferase [54], 2) using minimal or core promoters, as in the luciferase assay, or none, as in SuRE-SNP [55-57], 3) differences in the assayed DNA sequence length between the randomly inserted SuRE DNA fragments and the static DNA fragment in the luciferase assay [55], 4) scrambling the location of minimal promoter, reporter gene, and inserted element within the plasmid [55]. Another aspect to consider is the role of DNA shape characteristics in regulating enhancer activity [58]. The same assayed DNA-fragment may be differently accessible for regulatory proteins due to folding differences of each plasmid based on their elements and total size differences, thus influencing the shape of the assayed DNA-fragment. Indeed, the plasmids of the luciferase and SuRE assays are very different, which may lead to unexpected and difficult to control results in both assays. Overall, we believe that high-throughput assays as SuRE-SNP are valuable to detect initial ASE effects, but functional validation with other reporter assays and multi-omics approaches is necessary to confirm the SNP effects and especially the downstream consequences of these SNPs.

Finally, we note that the (best powered) CeD GWAS we used, has been performed on the Immunochip platform. Consequently, most CeD-associated loci are by definition immune loci, and those barrier genes that are in these loci are there by chance. Indeed, several studies, including ours, imply that the heritability of auto-immune diseases, including CeD, is best explained by SNPs located in regulatory regions in lymphoid cells (**Fig. 4A,B**) [38,59,60]. Thus, the chances of finding CeD-associated SNPs that affect epithelial gene expression in Caco-2 cells is low. Unfortunately, as we had poor signal-to-noise ratios in Jurkat T cells, we were also unable to find many T cell specific SNPs associated with CeD that caused ASE.

In conclusion, we identified many cell-type and context specific regulatory elements governing CeD heritability and identified and replicated several genetic effects on these regulatory elements. While the power to detect regulatory elements and genetic effects was limited by cell-type dependent technical variation, our data suggests that genetic loci associated to CeD may also play a role in epithelial cells ultimately affecting the susceptibility and development of CeD.

Methods

Selecting CeD associated LD blocks

CeD associated loci were defined based on three CeD meta-analyses by selecting both genome wide ($p<5x10^{-8}$) and suggestively ($p<5x10^{-5}$) associated variants [17,18,23]. These variants were then clumped in a 1 mega base (Mb) window around the top variant to identify possible independent signals (LD $R^2<0.1$) within the locus. Then all variants with an $R^2 > 0.8$ around these independent top variants were identified. The outer boundaries of this set of variants were padded with 50kb to ensure the majority of the possibly causal variants were present. The thresholds for LD R^2 of 0.8 and 50kb of padding were determined based on simulated GWASs where the causal variants are known, allowing evaluation of what cutoffs are best to recover the majority of causal variants (https://shiny.cnsgenomics.com/gwasMP/ , **Suppl. Fig. 1**) [61]. Due to the size of the LD blocks in the HLA region (chr6:25-35Mb) and the fact we could only enrich for a maximum of 14Mb, we opted to exclude these regions, apart from 2 blocks identified through HLA fine-mapping by Gutierrez-Achury et al. [23]. In total, we defined 72 loci comprising 136 independent top variants to be included in the SuRE-SNP assay (**Suppl. Tab. 1**). These analyses were performed using Plink 1.9 with the 1000 genomes phase 3 non-Finnish European samples as the LD reference [62].

Sample selection

To get the highest representation of minor alleles for each CeD associated variant, we selected the optimal combination of CeD-patient samples for which we had material available from the latest CeD GWAS [18]. A maximum of 30 samples could be enriched and thus we ranked all samples by the count of minor alleles they carried for CeD-associated independent top variants with a MAF < 0.1. Any samples with fewer than 2 minor alleles with MAF<0.1 were excluded. The top 30 samples with the highest minor allele count were selected for analysis. For this selection we used the previously generated genotypes, which were made using the Immunochip platform [18] and were confirmed by performing a new GSA SNP-array on the 30 selected DNA samples to prevent sample swaps.

Wetlab methods

Enrichment of CeD associated DNA from patient material

Enrichment of the CeD associated regions (see section: Selecting CeD associated LD blocks) for the selected 30 samples (see section: Sample selection) was done using Agilent's SureSelect XT Target Enrichment (SureSelect custom 12-24Mb, catalog no. 5190-4896) on Agilent's Bravo automated liquid handling platform. SureSelect XT makes use of pre-designed RNA-probes that bind and pull-down target DNA by magnetic bead binding. A total of 230.330 RNA-probes were computationally designed using Agilent's online SureDesign platform. Because of the limited number of probes allowed, the design prioritized genome wide ($p<5x10^{-8}$) loci by a two-fold coverage versus single-fold coverage of suggestive ($p<5x10^{-5}$) loci. The SureDesign tool makes use of two different boosting settings (maximum, and balanced) that ensure a different locus coverage by tiling of probes, as well as three repeat-mask settings (most stringent, least stringent, no masking) to reduce inclusion of known repetitive elements. For both genome wide and suggestive loci we designed probes in three parts, 1) Most stringent repeat masking and maximum boosting, 2) Least stringent repeat masking and balanced boosting, 3) no repeat masking and balanced boosting (**Suppl. Tab. 5**). The design included two-fold coverage of all genome wide loci (46,82%, 25,15%, and 7,69% of total probes per respective setting), one-fold coverage of all suggestive loci (9,21%, 8,70%, and 0,70% of total probes per respective setting), and additional coverage of the IL-21/IL-2 locus to utilize the maximum allowed number of probes (one-fold coverage, least stringent repeat masking, balanced boosting, 1,73% of total probes).

Per sample we started with 400Ng DNA and processed it according to the manufacturer's protocol. In short, whole genomic DNA was sonicated with the Covaris S220 followed by end-repair, a-tailing, and adapter ligation. Originally the SureSelect XT protocol makes use of Illumina's Truseq adapters, however, the SuRE-plasmids also contain the Truseq adapters. The consecutive use of Truseq adapters in both methods would result in likely amplification and sequencing errors. Thus, we created and ligated with custom made adaptors (**Suppl. Tab. 6**). Next an 11-cycle pre-PCR was required followed by Speedvac (Thermo Scientific, Savant DNA120) to concentrate the DNA before the actual hybridization and streptavidin-coated magnetic bead (Invitrogen, Dynabeads MyOne Streptavidin T1, catalog no. 65604D) capture of the hybridized DNA. A 12-cycle post-PCR finalized the hybridized libraries with an average fragment size of 300bp including adapters. All required purifications in between each step were done using Agencourt's AMPureXP beads (catalog no. A63882). DNA shearing quality, pre-PCR libraries, and post-PCR libraries were assessed using Agilent's Tapestation D1000 (D1000 ScreenTape, catalog no. 5067-5582).

Generating the plasmid pool

The enriched DNA library with an average fragment size of 300bp including FlexAdapters was transformed in CloneCatcher DH5G electrocompetent Escherichia coli cells (Genlantis, catalog no. C810111), or in E. cloni 10G cells (Lucigen, catalog no. 60107-1), followed by purification using a GIGA plasmid purification kit (#10091; Qiagen).

Barcode to fragment library preparation

Library preparation and barcode to fragment sequencing was done according to the previously described short DNA insert size.

Cell culturing, plasmid pool transfections, stimulations, and flow cytometry

The finalized SuRE-SNP plasmid pool was transfected into all cell types at 120 million cells per biological replicate with two biological replicates being generated for each unstimulated and stimulated cell type. Cell culturing of Caco-2 cells (ATCC, catalog no. HTB-37) was done in DMEM medium with high glucose and pyruvate (Gibco, catalog no. 41966052), 1% penicillin/streptomycin (Lonza, catalog no. DE17602E), 10% heat-inactivated fetal bovine serum (Gibco, catalog no. 10270), 1% 1M 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES, Gibco, catalog no. 15630080), and 1% MEM non-essential amino acids solution (Gibco, catalog no. 11140050). All other cell lines, Jurkat (ATCC, clone E6-1, catalog no. TIB-152), K562 (ATCC, catalog no. CCL-243), GM12878 (NIGMS, Corriel Institute, GM12878), THP-1 (Sigma Aldrich, ECACC 88081201), were cultured in RPMI 1640 with L-glutamine and HEPES (Gibco, catalog no. 52400-025), and the same 1% penicillin/streptomycin and 10% heat-inactivated fetal bovine serum as used for Caco-2 cells. All cells were cultured up to passage 5-7 after thawing before used for further experiments.

Caco-2 cells were transfected according to the manufacturer in batches of 40Ug plasmid per 20 million cells by lipofection at a 3:1 ratio of lipofectamine and DNA (Thermo Fisher Scientific, Lipofectamine-2000, catalog no. 11668019). All other cell types were transfected by 4D-nucleofection (Lonza Amaxa) in batches of 10Ug plasmid per 10Million cells: K562 (Lonza Amaxa, SG kit, catalog no. V4XC-3024, program FF-120), Jurkat (Lonza Amaxa, SE kit, catalog no. V4XC-1024, program CK-116), THP-1 (Lonza Amaxa, SG kit, catalog no. V4XC-3024, program ED-100), GM12878 (Lonza Amaxa, SF kit, catalog no. V4XC-2024, program CY-100). Transfection method and program choices were based on extensive optimizations with optimal cell numbers and plasmid concentrations (data not shown). However, the bulk nature of these experiments has (severely) reduced efficiencies for some cell lines.

All cell stimulations were performed for 3 hours. Caco-2 cells were stimulated with 60Ng/ml IFN-γ (Biolegend, Recombinant Human IFN-γ (carrier free), catalog no. 570206). Jurkat cells were stimulated with 2.5Ug/ml anti-CD3 (Biolegend, Ultra-LEAF Purified anti-human CD3 Antibody, catalog no. 317326) and 2.5Ug/ml anti-CD28 (Biolegend, Ultra-LEAF™ Purified anti-human CD28 Antibody, catalog no. 302934). THP-1 cells were stimulated with 16.67Ng/ml LPS (Sigma-Aldrich, Lippopolysaccharides from Escherichia coli 026:B6, catalog no. L8274-10MG). GM12878 cells were stimulated with 20Ug/ml Anti-IgM (Jackson Immunoresearch Europe, AffiniPure F(ab')$_2$ Fragment Goat Anti-Human IgM, Fc5µ fragment specific, catalog no. 109-006-129), 2Ug/ml CD40 (R&D Systems, Human CD40/TNFRSF5 Antibody, catalog no. mAB6321-500), and 50Ng/ml IL-21 (Abcam, Recombinant human IL-21 protein (Animal Free), catalog no. Ab179621).

Flow cytometry was used to confirm transfection efficiency by GFP (Lonza Amaxa, nucleofection kits, pmaxGFP), to estimate cell death based on cell size, and to confirm stimulation efficiency by staining for CD69 for Jurkat cells (data not shown, Biolegend, CD69 PE, catalog no. 310906) (**Suppl. Tab. 9**).

SuRE Library preparation

RNA was taken using 1Ml per 10 million cells of Qiazol lysis reagent (Qiagen, catalog no. 79306) 24-hours after nucleofection (K562, Jurkat, THP-1, GM12878) and 48-hours after lipofection (Caco-2) as these times were determined to provide the strongest GFP expression and thus also SuRE-barcode expression. RNA isolations were performed in batches of 12Ml bulk using Corning 15Ml high-speed centrifuge tubes (Sigma Aldrich, catalog no. CLS430791-500EA) according to the Qiagen's Qiazol protocol with the inclusion of an additional acid-phenol-chloroform step (Thermo Fisher Scientific, Acid-Phenol:Chloroform pH 4.5 with IAA 125:24:1, catalog no. AM9720). cDNA was produced in batches of 10-20 reactions per biological replicate with the SuRE-plasmid specific primer 247JvA as previously described [63]. PCR amplification was also done as previously described but with the MyTag HS Red Mix (Bioline Meridian Bioscience, catalog no. BIO-25048) and with an optimal number of PCR cycles that was determined by qPCR for each cell line (data not shown, cycles: Caco-2 17-cycles, Jurkat and K562 16-cycles, THP-1 19-cycles, GM12878 21-cycles). Library clean-up was initially done by AMPureXP beads (Agencourt, catalog no. A63882) and followed by library assessment with Agilent's D1000 tapestation (Agilent, catalog no. 5067-5582). An additional PAGE-purification was performed to remove unwanted PCR artifacts, the protocol was adapted from the methods publication by Lopez-Gomollon et al. [64]. In short, DNA was loaded onto 15% polyacrylamide gels, the 200- 400bp fragment band was cut from the gel, purification of DNA was done overnight by PAGE-elution buffer (0.5M ammonium acetate and 10MM magnesium acetate), followed by ethanol precipitation of the DNA. The final libraries were again assessed by Agilent's D1000 tapestations before pooling and sequencing.

RNA-seq library preparation

A few micrograms of RNA from all biological replicates of Caco-2, Jurkat, and K562 were used to generate stranded polyA RNA-seq libraries with Illumina adapter (NovoGene).

Sequencing

Sequencing of the plasmid pool for barcode to fragment association (iPCR) was done by NovoGene sequencing 150bp PE on a Hiseq X ten. Sequencing was performed in two runs on the same library. The first run sequenced 469.895.653 reads, the second run sequenced 743.846.252 reads for a total of 1.213.741.905.

Sequencing of the barcode expression (cDNA) library sequencing was done by NovoGene at SE 75bp on Illumina NextSeq, or by the UMCG sequencing facility on Illumina NextSeq at SE 75bp in several runs for a total of 4.301.099.097 reads for all samples 32 samples (8 cell-lines / conditions, 4 replicates).

RNA-seq libraries were also sequenced by NovoGene for Caco-2, Caco-2 Stim, Jurkat, Jurkat Stim, K562 each in duplicate. Sequencing was 150bp PE on the Illumina NovaSeq6000 for aiming for 25 million reads per sample.

Luciferase assays to validate ASE effects

Double-stranded DNA (dsDNA) of the seven SNPs selected for luciferase validation were manufactured as gBlocks gene fragments (IDT) of about 300bp with the SNP of interest centered in each fragment (**Suppl. Tab. 7**). Per SNP two gBlocks were ordered, one for each SNP-allele, any other known variants within the fragment were always kept at reference alleles. In some cases, but always for both alleles, the FlexAdapter sequence was included on one-side of the fragment to compensate for too high or low GC-content at the 5' or 3' of the gBlocks which the gBlocks manufacturing process could not handle.

The Promega Dual-Luciferase Reporter Assay System (Promega, catalog no. E1910) was performed according to the manufacturer's instructions. Thus, cloning of the fragments was done in both orientations in the pGL3-Basic vector (Promega, E1751). Plasmid transfections with the cloned pGL3-basic and TK-Renilla (Promega, E2241), were performed in 500K to 1 million Caco-2 or K562 cells as described above. K562 luciferase cell lysates were taken 48-hours post transfection and Caco-2 lysates 72-hours post transfection due to the nucleofection versus lipofection differences. All SNPs and controls were also stimulated in Caco-2, as described above, stimulations were started 3-hours before the 72-hour time-point. A single biological test was performed in K562 and Caco-2 cells, with triplicate technical replicates. Replicated

ASE SNPs were tested in biological triplicates in Caco-2 with and without stimulations. Luminescence was measured with double cell lysate volumes (20Ul) and half LAR-II and Stop&Glo volumes (comparable results to normal volumes, data not shown) using the GloMax 96 Microplate Luminometer (Promega, E6521), other settings were according to the manufacturer's manual.

Statistical Methods

Step 1: Barcode to fragment association

Before the associations between unique barcode activity and genomic regions can be established the barcodes and genomic fragments need to be linked. First, the sequencing data of the plasmid pool is demultiplexed into fastq files per donor using Cutadapt v2.7 [65] enabling -e 0.2 --match-read-wildcards --action=none --no-indels. After which the adapter sequences in both forward (CCTAGCTAACTATAACGGTCCTAAGGTAGCGAACCAGTGAT NNNNNNNNNNNNNNNNNNNNNNNNAGCGTACCGTAGT) and reverse (CCAGTCGT NNNNNNNNNNNNNNNNNNNNNNNNAGCGTACCGTAG) reads are trimmed using Cutadapt using nonstandard options - q 25 -m 20:50 -e 0.075 --max-n 5 --no-indels --discard-untrimmed. The 20 nucleotides preceding the adapter in the forward read represent the barcode. The nucleotides after the adapter represent the genomic sequence of the donor.

Next, the sample DNA fragments are aligned to the human genome (b37, 1KG version) using BWA mem v0.7.15 (http://bio-bwa.sourceforge.net/bwa.shtml) using default options and sorted by readname using samtools v1.9. Reads are sorted by readname so the paired reads can be extracted more efficiently. Next, the aligned reads stemming from different sequencing runs and lanes are merged into one BAM file per donor using Picard v2.18.26 (https://github.com/broadinstitute/picard). Alignments were then filtered removing any with a mapping quality under 30, non-primary alignments and reads that do not map in a proper pair using samtools.

Step 2: Quantifying barcode expression

To quantify the mRNA expression of the barcodes that is mediated by the inserted random DNA fragments of patient DNA, the RNAseq reads were trimmed on the 5' adapter (CCTAGCTAACTATAACGGTCCTAAGGTAGCGAA) using Cutadapt. Only reads with an adapter that had fewer than 3 mismatches were kept (on average 98% of reads). Any reads with more than 3 mismatches with the adapter sequence were discarded. After this, any barcodes which were not 20 nucleotides were also removed. Finally barcodes were counted on how often they were present in the set using iPCR-tools MakeBarcodeCounts (https://github.com/OlivierBakker/CeD-SuRE-SNP/tree/master/pipeline/iPCR-tools).

Step 3: Assigning fragments to barcodes

Finally, the fragments and barcode expression were integrated into a single file using iPCR-tools MakeIpcrFile, removing any fragments that did not have a barcode, were not paired, unmapped, or primary alignments. We hereafter refer to these fragment pairs as iPCR-records. As it is possible for the same fragment to be sequenced multiple times, we collapsed iPCR-records with the same barcode by merging records where both ends mapped within 50bp of each other, that mapped in the same orientation and to the same chromosome. We allowed for this 50bp window to account for slight variations in mapping, although this was a relatively rare occurrence. Out of the set of fragments with the same barcode, the fragment that had the best mapping quality and the longest stretch of aligned bases was used to determine the mapping position of that barcode. For each iPCR-record, we recorded how many times we observed the barcode. This information is later used as a proxy for transfection likelihood, as a plasmid that has been sequenced more is also more likely to have been transfected. This was done using iPCR-tools CollapseIpcr. This set of filtered iPCR-records was used as the basis for all downstream analysis.

Calling SuRE peaks

To call regions that showed SuRE-activity, we merged technical replicates, as these showed consistent activity in the active regions (Pearson r > 0.9, **Suppl. Fig. 3**,**6**). While rare, we then filtered any iPCR-records that had a barcode count >500 as these are likely to be PCR artifacts and could potentially skew the results. We then called peaks using macs2 v2.2.6, supplying the count of iPCR-records as the control track and the barcode count as the treatment. As we can be sure the iPCR-records are unique, we supplied –keep-dup all to keep fragments matching to the same position. As we restricted the analysis to CeD associated regions, we enabled –g 14E6 to represent the 14Mb of DNA we assessed. Further non-standard options were --format BEDPE -nomodel --bdg –SPMR. We filtered the resulting peaks on FDR 0.05 and an enrichment over background of 3. After which we only kept peaks that were present in both biological replicates using iPCR-tools OverlapPeaks. This set of peaks was used for downstream analysis (**Suppl. Tab. 2**). We repeated this procedure to assess if peaks showed orientation specific effects by only using iPCR-records in + and – orientations respectively.

Genotyping of genomic fragments

To enable the analysis of ASE effects on SuRE activity, we genotyped the genomic fragments using GATK HaploytypeCaller v3.8 [66]. First, the alignments were de-duplicated based on the barcode associated with the reads using iPCR-tools SubsetBam, keeping the best aligning reads. Base quality scores were then recalibrated, and variants called following the GATK best practices.

We applied hard filters to filter the genotype calls, as we have an over-representation of reads in the 14Mb of CeD associated DNA, which invalidates the model in the best practices. We applied separate filters for SNPs and INDELs. SNPs were filtered to be bi-allelic, and the following filter string, “QD < 10.0 || FS > 10.0 || MQ < 50.0 || SOR > 3.0 || MQRankSum < -5.0 || ReadPosRankSum < -4.0 || ReadPosRankSum > 4.0 || AF < 0.05”. INDELs were filtered to be bi-allelic, no larger than 40 bases and the filter string “QD < 10.0 || FS > 25.0 || SOR > 10.0 || ReadPosRankSum < -4.0 || ReadPosRankSum > 4.0 || AF < 0.05”. Post-filtering, we identified 64.697 variants of which 59.552 were SNPs and 5.257 were indels genome wide.

Quantifying ASE effects

IPCR-fragments were then assigned an allele if they overlapped a called genetic variant using iPCR-tools AssignVariantAlleles. Fragment were assigned an allele if the sequenced part of the fragment fully overlapped with the variant, the allele matched either of the variant alleles and the allele matched the genotype of the donor as called by GATK for homozygous individuals.

We then summed the barcode counts for the iPCR-fragments for all replicates of the same cell-type and condition and divided it by the number of times we observed the fragment in step 1. The log2 of this ratio was taken as the activity of an iPCR-record in a cell-type and condition. We considered variants if they overlapped a SuRE peak, and the minor allele had at least 12 fragments showing activity. We considered 868 variants for K562, 455 for Caco-2, 555 for stimulated Caco-2, 46 for Jurkat and 51 for stimulated Jurkats (**Suppl. Tab. 4**).

We then applied a Mann-Whitney U test for each variant to test if the mean activity was different between the two alleles for the variant. We further tested if the orientation of the fragment had an effect, by repeating this step for + and – orientations separately.

Overlap of SuRE peaks with known regulatory elements

Overlap of SuRE peaks with epigenome roadmap's ChromHMM 15-state predicted regulatory elements (K562 dataset E123_K562, Caco-2 dataset E109_Smallintestine, Jurkat dataset E037_THelperMemory) [11], and with Encode's cCRE (K562 dataset ENCFF464BRU, Caco-2 dataset ENCFF610OFX) [67] was done using Bedtools intersect 68. UCSC's liftover tool [69] was used to change Encode's cCRE's genomic coordinates from hg38 to hg19, and we filtered out any region marked as 'low DNAse region' for increased stringency. Entire SuRE peaks were overlapped with Encode's cCRE and one overlap per SuRE peak was counted to avoid inflation by multiple cCRE overlap with a single SuRE peak. Epigenome roadmap's ChromHMM elements were grouped as transcription start site (Tss; 1_TssA + 2_TssAFlnk + 10_TssBiv + 11_BivFlnk), enhancer (Enh; 6_EnhG + 7_Enh + 12_EnhBiv), transcription (Tx; 3_TxFlnk + 4_Tx + 5_TxWk), repressive (Repr; 8_ZNF/Rpts + 9_Het + 13_ReprPC + 14_ReprPCWk), and Quiescent (Quies; 15_Quies) states. Each SuRE peak was matched to a single element when multiple states overlapped by prioritization of states (Tss > Enh > Tx > Repr > Quies).

Enrichment of SuRE peaks among known regulatory elements

Enrichment of SuRE peaks among annotated genomic regions was calculated by overlapping SuRE peaks with a given reference dataset and comparing the overlap against an empirical null distribution generated by creating 100.000 sets of random peaks matching the SuRE peaks in length and size. To ensure no bias was introduced by the fact that we focus on known CeD associated regions, which are likely to be enriched for regulatory signal compared to random segments of genome, we restricted the random sampling of regions to the 14Mb of CeD associated genome we assessed. The observed percentage divided by the mean of the permuted distribution was taken as the ratio of enrichment, and two-sided empirical p-values were calculated to assess the significance. This has been implemented in iPCR-tools GenomicRegionEnrichment.

Stratified LD score regression & enrichment of variants in CeD GWAS

We assessed the enrichment of CeD signal in SuRE peaks in two ways. First, we compared the distribution of CeD GWAS p-values of variants located in peaks to the distribution of all CeD GWAS p-values. We applied a Kolmogorov-Smirnov test to assess if the p-values of SNPs located in SuRE peaks deviated significantly from the background distribution.

Secondly, we applied stratified LD score regression (sLDSC) [38] to assess which cell-types were relatively more enriched for containing GWAS signal. As it is generally not recommended to run sLDSC on immunochip GWASs with the provided LD scores, we generated LD scores using only the SNPs which were tested in the CeD GWAS used. We then applied sLDSC using the SuRE peaks for each cell-type to quantify the relative enrichment of heritability in the peaks.

TF binding enrichment and allele specific TF binding

To calculate enrichment of TF binding sites under SuRE peaks we applied HOMER's [25] findMotifsGenome.pl using standard options providing each of the peak files to calculate enrichment and using the genome as background. To calculate cell-type or stimulation specific enrichments we used the respective peak sets as the background set. For instance, to calculate Caco-2 stimulation specific motif enrichments, we used the Caco-2 unstimulated peaks as background. To identify Caco-2 specific motif enrichments, we used the K562 peaks as the background.

To calculate the allele specific TF binding for the seven SNPs we studied more in depth, we scanned for binding sites in the sequence +- 30bp around the SNP. We again used HOMER to scan for known motifs in the sequences of both alleles of each SNP. We applied the by HOMER supplied database of known vertebrate TF motifs. We coupled the HOMER TF-motifs to TF-genes by using the extensive human TF resource by Lambert et al. and manually annotating where needed [70] (**Suppl. Tab. 8**). To order quantify the expression of TFs overlapping each SNP we used the RNA-seq data generated from the cell-lines used for the SuRE-assay.

RNA-seq analysis and differential expression

The trimmed fastQ files where aligned to build human_g 1K_v37 ensemble Release 75 reference genome using hisat/0.1.5-beta-foss- 2015B 71 with default settings. Before gene quantification SAMtools/1.2-foss- 2015B 72 was used to sort the aligned reads. The gene level quantification was performed by HTSeq-count HTSeq/0.6.1P1-foss- 2015B [73] using --mode=union, Ensembl version 75 was used as gene annotation database.

Quality control (QC) metrics are calculated for the raw sequencing data. This is done using the tool FastQC FastQC/0.11.3-Java-1.7.0_80 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). QC metrics are calculated for the aligned reads using Picard-tools picard/1.130-Java-1.7.0_80 (http://broadinstitute.github.io/picard/) CollectRnaSeqMetrics, MarkDuplicates, CollectInsertSize-Metrics and SAMtools/1.2-foss- 2015B flagstat.

Raw read counts were filtered to remove zeroes and lowly expressed genes, by removing genes with a read count < 10 in either duplicate. Differential expression analysis was performed by Deseq2. SuRE elements were linked to differentially expressed genes by taking all differentially expressed genes at FDR < 0.05 (**Suppl. Tab. 3**) and looking for SuRE peaks 1kb upstream of the promoter of these genes. This was done using the function findGenes in the R package bumphunter [74].

Data availability

Full summary statistics for the ASE analysis as well as peak calls and RNA counts have been provided as supplementary tables. The raw sequencing and genotype data is not publicly available due to privacy concerns with regards to the patient genetics but is available upon request. All code and scripts used to generate the results in this study are available at https://github.com/OlivierBakker/CeD-SuRE-SNP.

Author contributions

Conceptualization: R.V.B., O.B.B., J.V.A., B.V.S., I.H.J.
Data curation: O.B.B.
Formal Analysis: O.B.B, R.V.B., A.V.D.G.
Funding acquisition: J.V.A, B.V.S, I.H.J
Investigation: R.V.B., J.G-A., R.M.
Methodology: R.V.B., I.H.J., J.V.A., B.V.S
Project administration: I.H.J.
Software: O.B.B., L.P.
Supervision: I.H.J., S.W., S.S.
Validation: R.V.B., J.G-A., R.M.
Resources: C.W.
Visualization: O.B.B., R.V.B
Writing – original draft: O.B.B., R.V.B.
Writing – review & editing: I.H.J., S.W.

Roles as defined by: CRediT (Contributor Roles Taxonomy)

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We thank the genomics coordination center of the UMCG for their assistance with running the RNA-seq pipeline. We thank the genome diagnostics laboratory of the UMCG for their help, and financial contribution, with the SureSelect hybridization. We thank Kate McIntire for the editorial assistance.

Supplementary material

Supplementary material has been provided to the the University of Groningen Library which hosts the digital version of this thesis.

Suppl. Fig. 1. Recovery of simulated causal variants at different parameters. A) Percentage of recovered causal variants (y-axis) versus genomic distance (x-axis). B) As A, but x-axis show the LD R^2 between the variants. C) As A, but x-axis shows the difference in minor allele frequency. Causal variant simulations determined in [61] using the webtool https://shiny.cnsgenomics.com/gwasMP/.

Suppl. Fig. 2. Barplots per chromosome showing the number of SuRE-fragments. Each chromosome was split up into 1Mb windows, after which the number of SuRE-fragments in that 1Mb window was counted. 1Mb windows overlapping with a CeD associated region (as defined in **Suppl. Tab. 1**) are colored blue.

Suppl. Fig. 3: Scatterplots of SuRE-activity in active regions between technical replicates. VST normalized (DEseq2) counts of SuRE-activity between different technical replicates of each sample. Each dot represents a peak active in at least one of the cell-types.

Suppl. Fig. 4: Sequencing saturation curves and signal to noise ratios. A) Number of unique barcodes detected in barcode expression data (y-axis) at different levels of random down sampling of the total sequenced barcode pool (x-axis). B) As A, but y-axis indicates the proportion of unique barcodes detected. C) The mean ratio of barcode count (cDNA) and plasmid count (iPCR) over all detected peaks. D) Signal to noise ratio's for the different cell-types assayed. Signal to noise ratio was determined by taking the peak ratio in the peaks as signal and dividing this by the mean peak ratio in randomly selected regions (methods). A ratio of 1 indicates no difference, 2 a 2 fold difference etc. * indicates that the signal to noise ratio is significant at a p-value of < 0.05 adjusted for 9 tests. E) Track overview of the different cell-types in the region around the promoter of *PARK7*

Suppl. Fig. 5: Overlap between SuRE-peaks and functional elements identified by the Epigenome Roadmap and Encode consortia. Y-axis shows the proportion of SuRE-elements that overlaps with the respective element in Roadmap or Encode. The numbers above the bar indicate the number of peaks that overlap.

Suppl. Fig. 6: Association between regions of active SuRE-activity and average activity for different peak types. Heatmaps of Pearson correlations between VST normalized (DEseq2) counts of SuRE-activity between different replicates and cell-types. A) Correlation between promoter peaks 1kb upstream of a TSS. B) Correlation between proximal enhancer peaks 5kb upstream of a TSS. C) Correlation of distal enhancers located > 5kb from a TSS. D) Average activity of the SuRE-peaks in each of the three types of peaks.

Suppl. Fig. 7: Region overview of rs785847. A) ASE effect of SNP on SuRE-activity per cell-type. Nominal p-values of Wilcox test between the alleles indicated. B) ASE effect of the SNP per orientation using luciferase assay. C) Trackplot showing the SuRE-activity in a 2kb region around the SNP. D) Zoomed in view showing a region of 60bp around the SNP. Bottom tracks show TF binding sites per SNP allele as determined by HOMER. The TF overlap with our K562 top-ASE rs785847 almost exclusively shows AP-1 TF subunits from the basic leucine-zipper (bZIP) family like FOS, JUN and ATF [75]. The bZIP-family proteins all contain the TGAsTCA binding-motif (s is a C or G nucleotide) with only very subtle differences in binding preference for the middle (n) nucleotide. All bZIP-family TFs as shown in (panel D) have no or little tolerance for a T-nucleotide in the middle position of the TF-motif and are therefore likely not properly binding at the T-allele of SNP rs785847 resulting in its lower transcription compared to the C-allele. Interestingly, in Caco-2 SuRE and luciferase data we do not see much expression of this regulatory region (panel C), possibly because bZIP TF-family members form dimers before binding to DNA [75] while only BATF of all these bZIP TF-family members is expressed in Caco-2 cells which by itself may not be enough to recruit the required transcription machinery.

Gene expression changes in AP-1 subunit genes have previously been associated with many immune diseases [76]. Similarly, abundant overlap of AP-1 subunit proteins with SNPs has been observed within CeD-associated loci and many other auto-immune diseases [77]. Although we only identify this ASE SNP in K562 it may play a role in other CeD-associated cell types with extensive use of AP-1 gene-regulation. For example, enrichment of AP-1 subunit TF-motifs has been observed in the open chromatin regions of gluten-specific T cells [34].

Suppl. Fig. 8: Region overview of rs140490. A) ASE effect of SNP on SuRE-activity per cell-type. Nominal p-values of Wilcox test between the alleles indicated. B) ASE effect of the SNP per orientation using luciferase assay. C) Trackplot showing the SuRE-activity in a 2kb region around the SNP. D) Zoomed in view showing a region of 60bp around the SNP. Bottom tracks show TF binding sites per SNP allele as determined by HOMER. Rs140490 was included as a negative control for Caco-2 cells since it has been previously fine-mapped as an SLE relevant SNP in B cells and monocytes with higher T-allele than G-allele expression [39]. We do not measure SuRE-SNP ASE nor luciferase activity for rs140490 (panel A, B). Nonetheless, our TF-motif analysis does reveal many new TF-motif matches at the T-allele (panel D). Any of the newly formed TF-motifs at the T-allele are likely candidates to induce ASE in the context of B cells, monocytes and perhaps other cell types but not in epithelial cells. The c-Myc TF-motif solely present at the G-allele (and actively expressed in Caco-2 cells) may compensate the otherwise potentially lower transcribed G-allele and explain why this SNP does not have an ASE in all cell types (those with c-Myc expression). Furthermore, similar to rs55950816 this SNP rs140490 does not overlap properly with our identified SuRE peak (**Suppl. Fig. 12**).

Suppl. Fig. 9: Region overview of rs140496. A) ASE effect of SNP on SuRE-activity per cell-type. Nominal p-values of Wilcox test between the alleles indicated. B) ASE effect of the SNP per orientation using luciferase assay. C) Trackplot showing the SuRE-activity in a 2kb region around the SNP. D) Zoomed in view showing a region of 60bp around the SNP. Bottom tracks show TF binding sites per SNP allele as determined by HOMER. The strongest Caco-2 and stimulation ASE rs140496, with an expected higher G-allele expression, could not be replicated by luciferase assay. Surprisingly, it did not pass the threshold for an active enhancer when compared to a negative control synthetic DNA fragment (not shown). Moreover, in contrast to the ASE effect in SuRE-SNP, the G-allele showed suggested repression. Aside from methodological differences between the luciferase assay and our SuRE-SNP method that might explain this discrepancy, the TF-binding motif differences between both alleles could be informative in identifying the enhancer disrupting potential of rs140496. The major difference in TF-binding sites is that the A-allele has binding opportunities for AP-1 subunits that the G-allele does not have, while TF-motifs for E2A (TCF3) and the in Caco-2 enriched TF-motif SNAI1 (**Fig. 2C**) are located just next to the SNP for both alleles (panel D). Possibly there is competition for binding between SNAI1 or E2A and AP-1 at the A-allele reducing its potential expression compared to the G-allele, or an opposite increase of A-allele expression happens due to synergistic binding between SNAI1 or E2A and AP-1. SNAI1 was shown to recruit lysing-specific demethylase-1 (LSD1) which is an epigenetic co-repressor component that can remove the active enhancer mark H3K4Me1 to silence regulatory elements [78,79]. Similarly, the E2A (TCF3) TF can also act as a silencing factor [80].

Suppl. Fig. 10: Region overview of rs71327063. A) ASE effect of SNP on SuRE-activity per cell-type. Nominal p-values of Wilcox test between the alleles indicated. B) ASE effect of the SNP per orientation using luciferase assay. C) Trackplot showing the SuRE-activity in a 2kb region around the SNP. D) Zoomed in view showing a region of 60bp around the SNP. Bottom tracks show TF binding sites per SNP allele as determined by HOMER.

Suppl. Fig. 11: Region overview of rs11072504. A) ASE effect of SNP on SuRE-activity per cell-type. Nominal p-values of Wilcox test between the alleles indicated. B) ASE effect of the SNP per orientation using luciferase assay. C) Trackplot showing the SuRE-activity in a 2kb region around the SNP. D) Zoomed in view showing a region of 60bp around the SNP. Bottom tracks show TF binding sites per SNP allele as determined by HOMER Rs11072504 is a Caco-2 only ASE in our data with stronger C-allele than T-allele expression. However, this does not replicate in our luciferase assay (panel A, B). TF-binding motif analysis only reveals GATA3 binding at the T-allele as a potential TF-binding difference (panel D). Considering the higher C-allele expression as identified with the SuRE-SNP ASE, GATA3 binding at the T-allele would have to induce repression, which GATA3 is indeed capable of [81]. The role of rs11072504 if regulated by GATA3 may be more relevant in T cells [82,83], especially in CD8+ T cells since GATA3 is known to repress functionally similar NK-cell related genes [81], especially in CD8+ T cells since GATA3 is known to repress functionally similar NK-cell related genes. Moreover, rs11072504 is located in between two SuRE peaks and may therefore not be as biologically impactful as the SuRE-SNP ASE analysis suggests (panel C).

Suppl. Fig. 12: Region overview of rs55950816. A) ASE effect of SNP on SuRE-activity per cell-type. B) ASE effect of the SNP per orientation using luciferase assay. C) Trackplot showing the SuRE-activity in a 2kb region around the SNP. D) Zoomed in view showing a region of 60bp around the SNP. Bottom tracks show TF binding sites per SNP allele as determined by HOMER. Rs55950816 is located within the HLA region on chromosome-6 and has a stronger transcription with its G-allele compared to its C-allele in stimulated Caco-2 cells according to our SuRE-SNP ASE analysis. We could not replicate this in our luciferase assay (panel B). The position of this SNP is on the far edge of our identified SuRE peak, which is 690bp in total and not completely overlapping with the 300bp synthetic DNA fragment used in the luciferase assay. This may explain the difficulty in replicating the expected ASE (panel A). Additionally, we do not identify any TFs overlapping the SNP and can therefore not speculate on any possible transcriptional mechanism that causes the potential ASE at this SNP (panel D). However, our motif search does not include every known TF and may therefore exclude a poorly characterized Caco-2 TF candidate for which the TF binding motif is unknown.

Suppl. Fig. 13. CCR locus overview containing the two ASE SNPs rs2888524 and rs71327063. Genes are coloured when they are associated to these two ASE SNPs according to eQTL-gen, and they remained black if there is no eQTL association. Colouring is based on Caco-2 gene expression with green representing highly expressed genes, yellow representing lowly expressed genes, and red representing not expressed genes.

References

1. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 47, D1005–D1012 (2019).
2. Broekema, R. V., Bakker, O. B. & Jonkers, I. H. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. Open Biol. 10, 190221.
3. Kerimov, N. et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat. Genet. 53, 1290–1299 (2021).
4. Võsa, U. et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. 53, 1300–1310 (2021).
5. Giambartolomei, C. et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLOS Genet. 10, e1004383 (2014).
6. Bossini-Castillo, L. et al. Immune disease variants modulate gene expression in regulatory CD4+ T cells and inform drug targets. 654632 (2019) doi:10.1101/654632.
7. Connally, N. et al. The missing link between genetic association and regulatory function. 2021.06.08.21258515 (2021) doi:10.1101/2021.06.08.21258515.
8. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. Nature 593, 238–243 (2021).
9. Barbeira, A. N. et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. 814350 (2020) doi:10.1101/814350.
10. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
11. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015).
12. Taylor, K. E., Ansel, K. M., Marson, A., Criswell, L. A. & Farh, K. K.-H. PICS2: next-generation fine-mapping via probabilistic identification of causal SNPs. Bioinformatics 37, 3004–3007 (2021).
13. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying Causal Variants at Loci with Multiple Signals of Association. Genetics 198, 497–508 (2014).
14. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat. Rev. Genet. 19, 491 (2018).
15. van Arensbergen, J. et al. High-throughput identification of human SNPs affecting regulatory element activity. Nat. Genet. 51, 1160–1169 (2019).
16. Caio, G. et al. Coeliac disease: a comprehensive current review. BMC Med. 17, 1–20 (2019).
17. Trynka, G. et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in coeliac disease. Nat. Genet. 43, 1193–1201 (2011).
18. Ricaño-Ponce, I. et al. Immunochip meta-analysis in European and Argentinian populations identifies two novel genetic loci associated with coeliac disease. Eur. J. Hum. Genet. 28, 313–323 (2020).
19. Dubois, P. C. A. et al. Multiple common variants for coeliac disease influencing immune gene expression. Nat. Genet. 42, 295–302 (2010).
20. Kumar, V. et al. Systematic annotation of coeliac disease loci refines pathological pathways and suggests a genetic explanation for increased interferon-gamma levels. Hum. Mol. Genet. 24, 397–409 (2015).
21. Wapenaar, M. C. et al. Associations with tight junction genes PARD3 and MAGI2 in Dutch patients point to a common barrier defect for coeliac disease and ulcerative colitisAn unusual case of ascites. Gut 57, 463–467 (2008).
22. van der Graaf, A. et al. Systematic Prioritization of Candidate Genes in Disease Loci Identifies TRAFD1 as a Master Regulator of IFNγ Signaling in Coeliac Disease. Front. Genet. 11, 562434 (2020).
23. Gutierrez-Achury, J. et al. Fine-mapping in the MHC region accounts for 18% additional genetic risk for coeliac disease. Nat. Genet. 47, 577–578 (2015).
24. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat. Rev. Mol. Cell Biol. 19, 621–637 (2018).
25. 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).
26. Gutierrez-Hartmann, A., Duval, D. L. & Bradford, A. P. ETS transcription factors in endocrine systems. Trends Endocrinol. Metab. TEM 18, 150–158 (2007).
27. Amoutzias, G. et al. One Billion Years of bZIP Transcription Factor Evolution: Conservation and Change in Dimerization and DNA-Binding Site Specificity. Mol. Biol. Evol. 24, 827–835 (2007).
28. Villarejo, A., Cortés-Cabrera, Á., Molina-Ortíz, P., Portillo, F. & Cano, A. Differential Role of Snail1 and Snail2 Zinc Fingers in E-cadherin Repression and Epithelial to Mesenchymal Transition. J. Biol. Chem. 289, 930–941 (2014).
29. THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318–1330 (2020).
30. Ramana, C. V., Chatterjee-Kishore, M., Nguyen, H. & Stark, G. R. Complex roles of Stat1 in regulating gene expression. Oncogene 19, 2619–2627 (2000).
31. Platanias, L. C. Mechanisms of type-I- and type-II-interferon-mediated signalling. Nat. Rev. Immunol. 5, 375–386 (2005).
32. Oshima, S. et al. Interferon regulatory factor 1 (IRF-1) and IRF-2 distinctively up-regulate gene expression and production of interleukin-7 in human intestinal epithelial cells. Mol. Cell. Biol. 24, 6298–6310 (2004).
33. Jabri, B. & Sollid, L. M. T Cells in Coeliac Disease. J. Immunol. 198, 3005–3014 (2017).
34. Bakker, O. B. et al. Potential impact of coeliac disease genetic risk factors on T cell receptor signaling in gluten-specific CD4+ T cells. Sci. Rep. 11, 9252 (2021).
35. Afzali, B. et al. BACH2 immunodeficiency illustrates an association between super-enhancers and haploinsufficiency. Nat. Immunol. 18, 813–823 (2017).

36. Chen, L. & Flies, D. B. Molecular mechanisms of T cell co-stimulation and co-inhibition. Nat. Rev. Immunol. 13, 227–242 (2013).
37. Christophersen, A., Risnes, L. F., Dahal-Koirala, S. & Sollid, L. M. Therapeutic and Diagnostic Implications of T Cell Scarring in Coeliac Disease and Beyond. Trends Mol. Med. 25, 836–852 (2019).
38. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. 47, 1228–1235 (2015).
39. Lewis, M. J. et al. UBE2L3 Polymorphism Amplifies NF-κB Activation and Promotes Plasma Cell Development, Linking Linear Ubiquitination to Multiple Autoimmune Diseases. Am. J. Hum. Genet. 96, 221–234 (2015).
40. Masuda, T. et al. Transcription factors LRF and BCL11A independently repress expression of fetal hemoglobin. Science 351, 285–289 (2016).
41. Constantinou, C. et al. The multi-faceted functioning portrait of LRF/ZBTB7A. Hum. Genomics 13, 1–14 (2019).
42. Wong, C. et al. Smad3-Smad4 and AP-1 Complexes Synergize in Transcriptional Activation of the c-Jun Promoter by Transforming Growth Factor β. Mol. Cell. Biol. (1999) doi:10.1128/MCB.19.3.1821.
43. Verrecchia, F. et al. Smad3/AP-1 interactions control transcriptional responses to TGF-β in a promoter-specific manner. Oncogene 20, 3332–3340 (2001).
44. Sundqvist, A. et al. Specific interactions between Smad proteins and AP-1 components determine TGFβ-induced breast cancer cell invasion. Oncogene 32, 3606–3615 (2013).
45. Mittelstadt, M. L. & Patel, R. C. AP-1 Mediated Transcriptional Repression of Matrix Metalloproteinase-9 by Recruitment of Histone Deacetylase 1 in Response to Interferon β. PLoS ONE 7, e42152 (2012).
46. Dennler, S., Prunier, C., Ferrand, N., Gauthier, J.-M. & Atfi, A. c-Jun Inhibits Transforming Growth Factor β-mediated Transcription by Repressing Smad3 Transcriptional Activity *. J. Biol. Chem. 275, 28858–28865 (2000).
47. Vincent, T. et al. A SNAIL1–SMAD3/4 transcriptional repressor complex promotes TGF-β mediated epithelial–mesenchymal transition. Nat. Cell Biol. 11, 943–950 (2009).
48. Smad3-mediated recruitment of the methyltransferase SETDB1/ESET controls Snail1 expression and epithelial–mesenchymal transition. EMBO Rep. 19, 135–155 (2018).
49. Uhlén, M. et al. Tissue-based map of the human proteome. Science 347, 1260419 (2015).
50. White, G. E., Iqbal, A. J. & Greaves, D. R. CC Chemokine Receptors and Chronic Inflammation—Therapeutic Opportunities and Pharmacological Challenges. Pharmacol. Rev. 65, 47–89 (2013).
51. Wei, Q. et al. LZTFL1 suppresses lung tumorigenesis by maintaining differentiation of lung epithelial cells. Oncogene 35, 2655–2663 (2016).
52. Wood, C. S. et al. Local control of phosphatidylinositol 4-phosphate signaling in the Golgi apparatus by Vps74 and Sac1 phosphoinositide phosphatase. Mol. Biol. Cell 23, 2527–2536 (2012).
53. Villar, D. et al. Enhancer Evolution across 20 Mammalian Species. Cell 160, 554–566 (2015).
54. Neefjes, M. et al. Reporter gene comparison demonstrates interference of complex body fluids with secreted luciferase activity. Sci. Rep. 11, 1359 (2021).
55. Klein, J. C. et al. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. Nat. Methods 17, 1083–1091 (2020).
56. Wu, G.-Q. et al. Evidence for transcriptional interference in a dual-luciferase reporter system. Sci. Rep. 5, 17675 (2015).
57. Zabidi, M. A. et al. Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. Nature 518, 556–559 (2015).
58. Schöne, S. et al. Synthetic STARR-seq reveals how DNA shape and sequence modulate transcriptional output and noise. PLOS Genet. 14, e1007793 (2018).
59. Meuleman, W. et al. Index and biological spectrum of human DNase I hypersensitive sites. Nature 584, 244–251 (2020).
60. Soskic, B. et al. Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. Nat. Genet. 51, 1486–1493 (2019).
61. Wu, Y., Zheng, Z., Visscher, P. M. & Yang, J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. Genome Biol. 18, 1–10 (2017).
62. Auton, A. et al. A global reference for human genetic variation. Nature 526, 68–74 (2015).
63. van Arensbergen, J. et al. Genome-wide mapping of autonomous promoter activity in human cells. Nat. Biotechnol. 35, 145–153 (2017).
64. Lopez-Gomollon, S. & Nicolas, F. E. Chapter Six - Purification of DNA Oligos by Denaturing Polyacrylamide Gel Electrophoresis (PAGE). in Methods in Enzymology (ed. Lorsch, J.) vol. 529 65–83 (Academic Press, 2013).
65. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12 (2011).
66. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. 201178 (2018) doi:10.1101/201178.
67. Moore, J. E. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583, 699–710 (2020).
68. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
69. Hinrichs, A. S. et al. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 34, D590–D598 (2006).
70. Lambert, S. A. et al. The Human Transcription Factors. Cell 172, 650–665 (2018).
71. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12, 357–360 (2015).
72. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 25, 2078–2079 (2009).

73. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169 (2015).
74. Jaffe, A. E. et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int. J. Epidemiol. 41, 200–209 (2012).
75. Bejjani, F., Evanno, E., Zibara, K., Piechaczyk, M. & Jariel-Encontre, I. The AP-1 transcriptional complex: Local switch or remote command? Biochim. Biophys. Acta BBA - Rev. Cancer 1872, 11–23 (2019).
76. Trop-Steinberg, S. & Azar, Y. AP-1 Expression and its Clinical Relevance in Immune Disorders and Cancer. Am. J. Med. Sci. 353, 474–483 (2017).
77. Harley, J. B. et al. Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity. Nat. Genet. 50, 699–707 (2018).
78. The SNAG domain of Snail1 functions as a molecular hook for recruiting lysine-specific demethylase 1. EMBO J. 29, 1803–1816 (2010).
79. Baulida, J., Díaz, V. M. & García de Herreros, A. Snail1: A Transcriptional Factor Controlled at Multiple Levels. J. Clin. Med. 8, 757 (2019).
80. Nguyen, H., Rendl, M. & Fuchs, E. Tcf3 Governs Stem Cell Features and Represses Cell Fate Determination in Skin. Cell 127, 171–183 (2006).
81. Van de Walle, I. et al. GATA3 induces human T cell commitment by restraining Notch activity and repressing NK-cell fate. Nat. Commun. 7, 11171 (2016).
82. Miettinen, M. et al. GATA3: A Multispecific But Potentially Useful Marker in Surgical Pathology: A Systematic Analysis of 2500 Epithelial and Nonepithelial Tumors. Am. J. Surg. Pathol. 38, 13–22 (2014).
83. Chen, A. F. et al. NEAT-seq: Simultaneous profiling of intra-nuclear proteins, chromatin accessibility, and gene expression in single cells. 2021.07.29.454078 (2021) doi:10.1101/2021.07.29.454078.

Part IV

Reflections

Chapter 10

General discussion

I. In summary

The past two decades have seen an exponential increase in the knowledge we have gained on complex traits and diseases. Previously textbook examples of simple traits, such as eye colour, are now understood to have a highly polygenic basis [1], illustrating that this increase in knowledge has also increased the complexity of the models we use to explain complex traits and diseases. We now understand that genetic variants not only act by eliciting protein-coding changes, but they also mould gene expression in subtle ways that are highly context-sensitive. Given the high polygenicity of complex traits, it is unlikely that we will gain much more understanding by studying single genes in isolation, as the disease-associated variants likely carry out their effects on disease risk through a complex cascade of pathways. As such, there has been a shift in focus away from identifying these genetic factors toward trying to interpret how they operate and what they represent. This post genome-wide association study (GWAS) interpretation has, however, proven challenging, and no model currently exists that can reliably predict the causal effects of GWAS variants.

This thesis evaluated how genetics modulates the development of complex traits and diseases from several different perspectives. In Chapter 2, we evaluated and reflected on current strategies for post-GWAS interpretation. Chapter 3 described Downstreamer, a strategy we developed for quantifying how genetic variants impact the gene regulatory networks underlying disease. We hypothesised that genes central in this network may be important for the disease process. Chapter 4 described a new method to identify eQTL effects that operate in a cell-type-specific manner that does not require generation of single-cell data. In Chapters 5 and 6, we evaluated the genetic and environmental factors that influence cytokine production capacity and linked it back to disease loci and risk scores. In Chapter 7, we looked at the genetics of immune traits in ancient individuals and discussed how evolution may have played a part in shaping them. Finally, in Chapters 8 and 9, we took a deeper look at how the genetic factors associated with coeliac disease may elicit a functional effect in cell types relevant for the disease course.

Overall, these chapters represent a broad evaluation of the current strategies for post-GWAS interpretation. Naturally, they are not without their limitations, and it would be an error to conclude that they provide definitive answers to these highly complex questions. While the specific and technical drawbacks have been extensively discussed in each respective chapter, this final chapter reflects on the work in this thesis and examines the challenges the field currently faces and will face in future. Finally, this chapters reflects on what actionable insight we have gained through work presented in this thesis.

II. Thoughts on interpreting the genetic basis of complex traits and diseases

Fine-mapping and effect sizes of GWAS variants

In Chapter 8, and especially in Chapter 9, we focussed on fine-mapping genetic variants associated with coeliac disease. However, as has become apparent, the effect sizes of individual GWAS variants are generally small, and a single variant is unlikely to make the difference between an individual developing or not developing a complex disease. In addition, in most cases, it is ultimately the protein encoded by the gene that is affected by the variant that exerts the effect. This then begs two questions: *Why do you care about fine-mapping variants? Why not just focus on the genes and proteins?* In principle, if you know the causal gene, then the variants are of less interest. However, in order to identify which gene is causally affected, I believe it is essential to have a comprehensive understanding of the regulatory structure in a locus. Given that GWAS loci generally harbour many variants and genes and linkage disequilibrium (LD) can span across mega-bases, without fine-mapping, you can never be sure which variants are casual. Consequently, you can never identify which enhancers or other regulatory elements are affected by the variants, and you are left with all genes in a locus as the credible set. While this is a workable backbone for subsequent analysis, it would improve specificity to truly know in which context which GWAS variants are affecting which genes in *cis*. Furthermore, if sufficient certainty can be attained about which genes are affected by which variants though fine-mapping, this will increase the understanding we have about the disease process.

Linking GWAS loci to genes

Fine-mapping does not solve the issue of prioritising genes. While it is a first step to identify which regulatory elements are disrupted by the genetic variants, linking these elements to genes remains non-trivial, as elements may be kilo-bases or even mega-bases away from the target gene.

One currently popular approach to link variants to genes (that we also applied in Chapters 5, 8 and 9) is to use eQTLs, as they provide a direct link between genetic variants and genes. However, as noted in Chapters 2 and 4, eQTLs can differ greatly depending on the cell type and the context the cell is present in. Current resources generally lack this specificity, and this is likely one of the reasons eQTLs are currently not very informative for providing variant gene-linking (Chapter 2) [2-5]. For example, the eQTLs from the GTEx project have been able to pinpoint candidate target genes for only 47% of GWAS loci [6], while on average only 11% of trait heritability can be explained by GTEx eQTLs [7]. It should be noted that a large proportion of the candidate genes are likely to be non-causal to the trait. In addition, Nasser et al. evaluated the performance of various models currently used for this task [5] and found that eQTLs, depending on the statistical model used, had either high recall and low precision or low precision and high recall for identifying links between IBD variants and known IBD genes. Moreover, the eQTL-based methods were outperformed by simply taking the closest gene to the variant. Furthermore, recent work applying a Mendelian randomisation–based approach (see *Randomised control trials and Mendelian randomisation* below for a discussion on this technique) only found evidence for causal mediation between transcripts and complex traits through protein levels in ~5% of loci, on average [8]. While these assays are also likely to contain bias and are far from representative of the entire human regulatory repertoire, they do show that there is still a lot that we cannot empirically describe about the *cis*-regulatory architecture of GWAS loci using eQTL, even for diseases where blood-based eQTL should be relatively informative.

While the GTEx project and the eQTL catalogue [4] have made great strides towards cell-type-specificity, and promising results are likely to come from the gut cell atlas [9], human cell atlas and similar single-cell consortia [10], these large-scale projects will still be limited by context-specificity. While I believe studying a gut disorder such as coeliac disease would be much more informative using eQTLs from gut-derived cells, doing this using "healthy" samples will likely not be representative of the disease context. In Chapter 9, for example, we identify regulatory elements in promoters of genes in the type-2 interferon pathway that are not active in the baseline state. In Chapters 5 and 8, a plethora of other examples show massive and dynamic transcriptional changes that only occur once cells are activated. At the risk of stating the obvious, it would be much better, if not essential, to map genetic effects in the truly relevant contexts if we want to understand the disease process and close some of the gaps in our knowledge. I have focussed here on the context-specificity of eQTL and genetic effects, but the same point holds when studying other factors related to disease.

This is, however, no trivial matter. It takes great expense and effort to build up biobanks that house the large number of samples of disease tissue required to do such a study. Additionally, the generation of protocols that can reliably and uniformly dissociate and process patient-derived material is a challenge that should not be dismissed. Furthermore, when studying the response of cells upon stimulation, a time-course study design would be ideal, as genetic variants may exert their effects depending on the time in the stimulation. For example, an eQTL modulating the expression of a transcription factor that is activated very early on after stimulation might only be active at that early timepoint, while this effect might not be observable anymore after several hours.

In the next 10-15 years, we will likely see the generation and expansion of such large scale biobanks, which will grow in size and quality as the technical ascpects are perfected and costs reduced. The prospect of biobanks containing thousands of case and control single-cell RNA-seq samples, perhaps even coupled with protein and chromatin measurements, is tremendously exciting. This will hopefully definitively awnser the question if the context specificity truly is the limiting factor in linking GWAS loci to genes using eQTL. However, we will likely never completely cover all of the GWAS loci through eQTL mapping, as the variation in expression measurements will always outweigh the extremely subtle effect of some of the GWAS variants. Furthermore, the power needed to detect such small and transient eQTL effects will be unfeasable, even in the near future. For instance, most of the variability in standing height can now be explained using common polymorphisms, but that has taken five million individuals to do so.

Other fine-mapping strategies

In Chapter 9, we utilised a modified version of the Survey of Regulatory Elements (SuRE) to identify genetic effects on the activity of regulatory elements. As noted in Chapter 2, however, various other strategies exist. Given the challenges we faced in translating the SuRE to other cell types, I thought it prudent to reflect on how some of these strategies might have played out when applied to studying the genetics of coeliac disease (CeD).

While I still believe that, in principle, the SuRE represents a very elegant way of evaluating genetic effects on enhancer activity, even if it works perfectly, the link between the regulatory element affected by a disease-associated variant and the gene still needs to be proven.

A very common way to fine-map, which we also applied in Chapters 8 and 9, is to overlap GWAS variants with known regulatory elements. The major disadvantage to this is that you have no guarantee that the variants that overlap these elements are affecting gene expression. Hence, the next logical follow-up would be to overlap with eQTLs. However, given the issues outlined above, this still leaves major gaps. A second strategy would involve looking at the transcription factor (TF) binding sites to check if these are disrupted. As these TF binding strategies are generally computational in nature, they still suffer from the issue that they do not prove a functional effect on regulatory activity or gene expression. Additionally, there are currently major differences in the predicted TF binding sites depending on the data and software used to do the predictions. Furthermore not all TF binding motifs are known [11].

In principle, combining these strategies using data from the correct functional context should provide much of the answer and yield a highly credible set of causal variants. But the key point here is that this should be done in the correct context. As most techniques for detecting open chromatin (DNA-seq/ATAC-seq) and the various CHIP-seq protocols for detecting promoters, enhancers and transcription (H3K27Ac, H3K27Me3, H3K 36Me3 etc.) are fairly standardised by efforts like ENCODE [12], it should be feasible to generate such data, just rather costly and time-consuming. However, even if variants are identified that meet all the outlined criteria, this will still not have proven that the variant is affecting the regulatory element. To do so, some form of functional test will remain necessary, as not all variants overlapping a regulatory element are guaranteed to elicit a functional effect.

Additional approaches to fine-map causal variants involve performing a meta-analysis using ethnically diverse populations [13,14] or various statistical models [15]. However, these are either currently limited by the data used or suffer from being observational/statistical in nature and, as such, do not provide explicit evidence that the variant is truly causal. While we have learned a great deal of fundamental knowledge from such efforts, they lack the specificity and functional evidence to make them actionable on their own. Recent efforts to increase the diversity of the genetic data that is available to researchers will likely help to increase the scope and power of such approaches.

The recent explosion of CRISPR-based experimental setups presents a veritable goldmine for designing assays that could assess genetic effects on regulatory elements and gene expression simultaneously. One could now theoretically design a high-throughput assay that could target enhancers or SNPs, and then, through (single-cell) RNA-seq, identify the effects this has on the expression of genes. Given that CRISPR-based techniques are being applied to primary cells, it will become possible to forgo model (cancer) cell lines and utilise actual patient material. However, these assays are not perfect, and substantial technical challenges still exist when it comes to clonal variability [16], guide RNA design [17], single-cell sequencing [18] and cost. In 5–10 years, these may have been solved, just as they have been in genome sequencing, allowing for the robust execution of such protocols.

Trans-eQTL effects, core genes and the omnigenic model

As mentioned throughout this thesis and in the literature, GWAS variants generally have small effect sizes and current *cis*-eQTL resources are imperfect when it comes to explaining their effects. Instead, it has been suggested that the heritability of GWAS traits is mostly modulated through *trans*-eQTL effects [3,19]. Indeed, Võsa & Claringbould showed that *trans*-eQTL effects stemming from systemic lupus erythematosus loci converge on genes related to interferon signalling, showing proof of concept for this idea. In Chapter 3, we evaluated whether the genes we prioritised as central in the gene regulatory network for a disease are enriched for *trans*-eQTL convergence stemming from the disease loci. However, while we observed significant enrichment, we could not explain most of the signal using *trans*-eQTL effects. Several reasons for this were discussed in Chapter 3. Additionally, many of the issues in *cis*-eQTL mapping discussed above also apply to *trans*-eQTLs.

The omnigenic model that was introduced in 2017 describes a similar idea [19]. In this model, GWAS variants are assumed to mostly influence peripheral genes that have no direct effect on the disease process. Instead, these peripheral genes are responsible for regulating core genes that are central to the disease process. In the omnigenic model, it is possible for genes in GWAS loci to be core genes. Since the model's introduction, several studies have come out that evaluate its credibility [20-22], and, in Chapter 3, we describe a method that operates along a highly similar vein and find that the genes we prioritise have properties similar to core genes.

There has been criticism on the omnigenic model. It has been described as an over-simplification of highly polygenic disease processes [23]. Personally, I feel one major challenge with this model is that it is so general that it fits almost everything, and it is a fair question how useful this model is in practice. Furthermore, it is hard to falsify, meaning we can currently only provide positive evidence for it. While the omnigenic model is a helpful abstraction to guide thinking about the genetic effects on gene regulatory networks, it remains and open quesition whether the search for core genes is truly helpful in understanding disease biology.

It would be an interesting challenge to re-define the omnigenic model into something more concrete, and perhaps apply Karl Popper's falsifiability principle to it. However, as with most concepts in complex trait genetics, we are still in an exploratory phase, and it would likely be hard to define a strict hypothesis that is general and still falsifiable and useful. While the hypothesis "*All CeD heritability converges on core genes*" is falsifiable by observing that there is a gene where this is not the case, this is not a very informative hypothesis to answer because it is very general and does not really help further the understanding of the disease process. Instead, one could ask more specific questions like – "*Is Gene X a core gene that significantly affects coeliac disease risk?*" – which is more informative and could be answered for all genes. However, the next question is then how a core gene is defined and whether this is a useful definition. Additionally, the word "significant" is also up to interpretation, a very small effect size of 0.0001% may still be a significant increase in risk, but it is not very actionable. It will take discussion by the scientific community to arrive at sensible and accepted definitions for such terms.

Hopefully, in time, core genes will provide better drug targets for complex disease. While we noted in Chapter 3 that several genes that fulfil the core gene criteria are also existing drug targets, it is an open question if this is coincidence or not. Moreover, not all core genes, should they exist, may be viable drug targets. However, even a marginal enrichment in viable targets would make identification of core genes a worthwhile endeavour. A pre-requisite for achieving this would be the generation of a widely accepted gold standard for core genes and drug targets that could be used to benchmark prioritisation approaches. While services like Open Targets are getting close, to my knowledge, no such standard currently exists. Moreover, generating such a standard is massively challenging given all the unknowns about complex disease genetics. Indeed, the generation of robust gold standard sets has been recognised as a key goal by the International Common Disease Alliance in their 2020 white paper [24].

The use of polygenic scores for interpreting disease genetics

A currently widely used approach for the interpretation of complex traits is the use of polygenic scores (PGSs) (also known as polygenic risk scores when applied to diseases). Generally, a PGS aggregates the thousands of GWAS effects linearly into a single score. This score can then be associated with a variety of traits (such as gene expression [3]) to estimate the collective contribution of the genetic signal to that trait. We applied these scores in Chapter 6 to estimate how the genetic factors associated with immune disease are associated with cytokine production. In Chapter 7, we used PGSs for immune traits to assess how the genetic factors collectively changed during the recent human past (in evolutionary terms).

While PGSs provide a useful abstraction layer, they should be interpreted with caution. For one, the heritability that is explained by the GWAS used to generate PGSs should always be considered when interpreting. For highly powered GWASs such as height, about ~40% of the variation can be accurately explained by SNPs in European ancestries [25]. But for many traits, this number is lower. This impacts interpretation, as you can only draw conclusions on the proportion of heritability you observe. It is very easy to not take this into account and to generalise the conclusions to the entire spectrum of variation for a trait. This is, however, not just an issue for interpreting PGSs, but one that impacts the whole scope of research into complex traits.

Furthermore, the now well-described issues with translating PGSs to non-European populations should also be considered [26,27]. Because of the different scope of variation and LD structure, European GWAS results cannot directly be translated to, for example, African populations. Interestingly, fine-mapping the variants used to generate the PGSs improves the *trans*-ancestry portability [27]. One of the major questions in Chapter 7 is how applicable modern GWAS is to ancient populations. While we did limit analysis to ancient samples discovered in Europe, and the genetic clustering showed that the samples do resemble modern European populations much more closely than, for instance, African samples, we cannot be certain that the variants (and their LD structure) that are associated to complex traits now, were the same 50,000 years ago. Furthermore, it has been suggested that the signals of polygenic adaptation detected using large sets of variants (such as height) are over-estimated due to un-corrected population stratification in the GWAS [28]. The increased *trans*-ancestry portability of fine-mapped variants may in future help improve the predictions we can make on the evolution of complex traits.

Interpreting the genetic basis of complex disease as a whole

As mentioned throughout this section, the genetic basis of complex disease is highly complex and polygenic. Furthermore, we are becoming aware that the effects of genetics are often transient and context-specific. Further study on the fundamental principles of how genes are regulated in *cis* and *trans* by the non-coding genome, genetics and contextual impact will close some of the gaps that cannot currently be explained by eQTL studies. This will require systematically solving the issues for all loci regarding enhancer–gene links, the discrepancies between chromatin-modulating variants and eQTLs and detection of TF binding and disruption. While this is a massive challenge, it also presents exciting opportunities for designing assays and models that take this complexity into account.

III. The limits and opportunities of statistical models and biological model systems

The widespread use of additive linear models in complex-trait genetics

As noted throughout this thesis, GWAS effect sizes are quite small. Hence, strategies have been and are being developed that amalgamate the collective of genetic effects into a more easily interpretable score (Chapter 2). The assumption in such models is often that stronger association equals more important function. In Chapter 3, we developed such a method, Downstreamer, that considers the collective impact of genetic effects rather than focusing on specific loci. The major limitation of most of such approaches is that they usually reduce the complex regulation that happens in disease to a few linear models. Indeed, ours is no exception in this.

This does raise a major question of whether it is a fair assumption to aggregate genetic effects into a single linear score, as done in a PGS (Chapter 7) or in Downstreamer (Chapter 3). Additionally, many approaches [29-31] such as Downstreamer, use p-values and not the effect size of the variant. Consequently, the directions of effect in such models are not interpretable. While a substantial portion of variability can certainly be explained by simple additive linear models, as evidenced by the fact that a good PGS (such as height) can be replicated in independent cohorts [32], other modes of operation for genetic variants exist. This might be one of the reasons why GWASs are currently only partially able to explain their trait's heritability, as noted in Chapter 1 [33]. For example, in GWASs, the dominant/recessive model is rarely assessed, and the alleles are assumed to contribute linearly to the trait. For instance, a recent study on type-2-diabetes found five additional loci by assessing the recessive model, and two of these had substantial effects with an odds ratio > 2 [34]. Furthermore, epistatic interactions (Chapter 2), where two or more alleles need to be jointly present for there to be an effect, are seldom tested due to the huge multiple testing burden and general complexity [35,36].

Additionally, gene x environment interactions might also form an important area where improvements can be made. For instance, for certain auto-immune diseases there is likely a link between viral infection and the triggering of the auto-immune disease [37-40]. Without also phenotyping the infection state of the individuals included in the GWAS, any genetic factors that interact with this process may be missed. However, the scope of possible interactions is practically infinite, so careful consideration is required when assessing such effects. Future methodological developments, improved phenotyping and increases in sample size will likely help to answer these questions.

On the other hand, making a model too complex is also not ideal. As the complexity of a model increases, so does the likelihood of error, leading to issues with reproducibility, as evidenced by the lack of reproducibility by machine learning algorithms in the biological sciences [41-43]. Additionally, more complex models, especially machine learning algorithms, are much harder to interpret because they can be quite opaque in terms of how they arrive at the model. The generation of a set of accessible gold standards for evaluating algorithms and the proper use of validation sets (ideally not just cross-validation) will be key to deliver results that hold up in the long term.

In the end, no perfect model exists, and it boils down to a famous adage attributed to George Box "*all models are wrong, but some are useful.*"

Perspectives on better model systems

This thesis extensively discusses the importance of studying the effects of genetic factors in their correct context, either a complex cellular context and/or the correct micro-environment generated by cytokines and other signalling molecules. However, I have not yet extensively discussed how this may be achieved. Luckily, there are many existing and emerging approaches that will allow for more control over context.

Co-cultures

So far in this thesis, the cell type composition of tissues has been treated primarily as a factor causing confounding in assays. However, cell proportions vary greatly in the body, and this does have a massive impact on phenotypes. A single activated T cell is unlikely to cause inflammation, but thousands of them will have a major impact. From the genetic point of view, it is very appealing to think of the eQTL or regulatory disrupting effects as we describe them theoretically, as static isolated effects that activate or repress the expression of the target gene. However, cells in tissues recruit, interact and influence each other, creating complex, interactive micro-environments that are dependent on disease status. Hence, this static picture of an eQTL in the context of tissues and disease states is not accurate. However, most current assays used to model complex disease genetics do so in single cell-lines or homogeneous cell pools, and thus fail to capture this aspect of biology.

This is where the development of co-cultures could provide an answer. In a co-culture, multiple cell types are cultured simultaneously *in vitro* and allowed to interact with each other (**Fig. 1**). As a simplified example, for CeD, the gluten-specific T cells could be co-cultured with intestinal tissue in the presence of CD8+ T cells and antigen presenting cells that present gluten peptides. If it is then possible to show that intestinal damage occurs in the gluten condition, you have a very nice model on which to perform (genetic) experiments. While such models would still not fully recapitulate the true disease process, they could allow for more accurate recapitulation of partial disease phenotypes and the pathways central to them. This is not limited to intestinal tissue, however, and studying the response of mixed sets of immune cells to different antigens at single-cell resolution at different timepoints would be massively interesting and provide insight into the causal cascade and feedback loops that occur in an immune response. This would be especially true if combined with an assessment of protein levels in conjunction with the mRNA levels. Furthermore, specific immune checkpoints or the key genes identified in Chapter 3 could then be modulated using CRISPR technologies

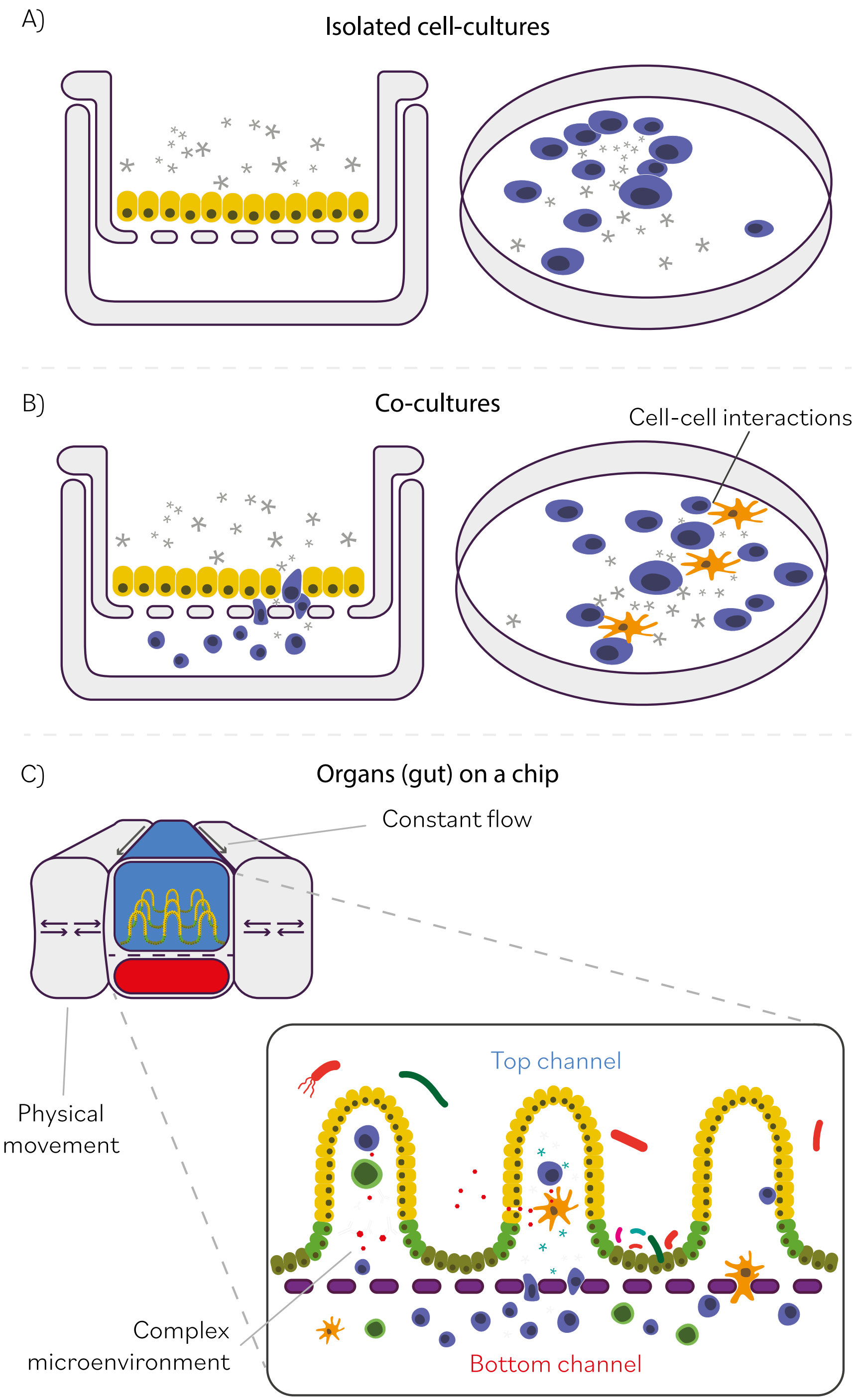


Fig. 1. Schematic overview of different cell-culturing approaches. A) A cell-culture grown in an isolated state. **B)** A co-culture where two cell-types are grown together and allowed to interact. **C)** An example of an organ on a chip, in this case representing gut tissue. These chips are tailored to mimic the environment of a tissue as best as possible. Panel C is based on Moerkens and Mooiweer et al. [47].

to assess their impact on the cellular phenotype, or a cellular pool could be pre-selected based on a known genetic background.

Biopsies

Additionally, future approaches in single-cell RNA sequencing will prove very exciting, once most of the technical wrinkles (such as the high variability between replicates [18]) have been worked out and protocols to process biopsy materials optimised. The generation of large databanks containing both healthy and diseased tissue with uniformly processed (single-cell) data could provide much more accurate insights into genetic regulation, especially if this data is available for different stages of any given disease. Indeed, for CeD, such experiments have been performed with intestinal biopsies taken from inflamed intestinal tissue [44,45]. While the scope of such experiments is currently still too limited to do major genetic studies, it is a very exciting future prospect. A major limiting factor lies in obtaining sufficient samples to perform these studies, especially from healthy subjects as it is not ethical to subject healthy individuals to the invasive medical procedures required. Furthermore, while intestinal biopsies are attainable with relatively low risk to the patient, heart or brain biopsies pose a greater challenge. Hence, studies using these challenging tissues, like GTEx, are generally done on post-mortem samples, which may have a large impact on the cellular phenotypes observed.

Organs on a Chip

To avoid the ethical issues in obtaining patient material, tissues can be grown *in vitro*. One approach that allows this are organs-on-chip (OOCs). OOCs are, as the name implies, small versions of organs grown on a microfluidic chip that can encompass a single cell type or co-cultures of various cell types (**Fig. 1**). The main difference between co-cultures and OOCs is that such chips generally contain channels through which a medium can flow and can be subjected to physical forces that mimic *in vivo* conditions, for instance, mimicking blood flow and peristalsis [46,47], whereas co-cultures are grown under static conditions. This introduces another physiologically relevant layer of complexity to the model. OOCs are grown based on organoids that are either made using patient-derived cells or grown based on (patient-derived) induced pluripotent stem cells (iPSC) [46] that can be obtained relatively easily. While currently there is a substantial cost to growing OOCs on a large scale, the fact that simplified tissues can be grown and kept alive for several weeks [46] presents an exciting opportunity to test and modulate potential drug targets on human-derived material. For example, the key genes identified in Chapter 3 could be targeted using a CRISPR screen on an OOC to assess their functional impact on this system. If a disease-state OOC is available, potential drugs could be added directly to the medium and the effects assessed.

OOCs also present a potential goldmine for answering more fundamental questions about disease processes. For instance, in CeD, it is still an open question how the gluten peptides end up in the intestinal tissue before inflammation occurs and which genes affect this process [48]. Genes that are identified as important for maintaining the intestinal barrier integrity and associated to CeD could be targeted using a CRISPR screen to assess the affects they have on permeability and the ability of gluten to enter the lumen. Furthermore, when the costs are reduced enough and the protocols standardised, it should be possible to grow hundreds of OOCs in parallel, each derived from donors with a different genetic background. This would enable *in vitro* QTL mapping and could allow for the study of (genetic) interaction effects such as the interaction between viral infection and the triggering of auto-immune disease. The ability to get functional readouts of tissue-like material beyond "classical omics" is tremendously exciting and opens up the possibility of assessing the downstream effects of genes on more disease-relevant functional readouts.

While I am definitely over-simplifying with respect to the huge amount of effort and work that goes into optimising such model systems, I believe that in the near future we will be using such systems to model complex disease. These models will likely not be very general but should be tailored towards a specific disease to achieve the best interpretability possible. Still, no model is perfect, and one should carefully consider the pitfalls of using models to understand the complexities of the human body.

IV. Causality

Unfortunately, causal statements are often made based on associations made with observational data [49]. A high-profile example of this would be the purported relationship between vitamin C levels and mortality from cardiovascular disease (CVD) [50], which was later disproven when the causality was assessed by a randomised control trial (RCT) [51]. Alternatively, a recent study applying Mendelian randomisation (MR) (see *Randomised control trials and Mendelian randomisation* below for a discussion), found that differential expression analyses on disease tissue are more likely to identify disease-induced effects than disease-causing ones [52].

In its basic definition, causality indicates a relationship between two events where event A precedes event B and hence is foundational to event B occurring (**Fig. 2A**). This definition is irrespective of the effect size A has on B. Event A may be fully causal, explaining 100% of the variation in B, or A may only be impacting B by 1%.

In association, A and B are linked, but A does not necessarily influence B. Instead, A and B could be linked by a third event C that influences A and B independently. Event C would be a confounder of the relationship between A and B (**Fig. 2A**). For instance, age or sex are common confounders in influencing the relationship between two events.

Furthermore, B could be causal for A, this is known as reverse causation. For example, it has been suggested that the observed protective effect of alcohol consumption on CVD is due to 'sick quitters', a form of reverse causation [53,54]. Associational studies performed on alcohol consumption do so by comparing non-drinkers to drinkers and assessing the effect on CVD. However, a bias arises when non-drinkers stopped drinking because of adverse health outcomes. Hence, the association between alcohol consumption and CVD is caused by the CVD, not the alcohol consumption.

In this thesis, we have mostly applied association-based techniques. This section reflects on what the implications of this are on the findings as well as the next steps that are needed to prove that the observations are due to a causal relationship rather than confounding or reverse causation. While there is nothing inherently wrong with observation-based association studies – they are a necessary first step – great care should be taken to account for biases during interpretation (see *Abstraction and bias in science* for an expanded discussion).

Genetic studies and causality

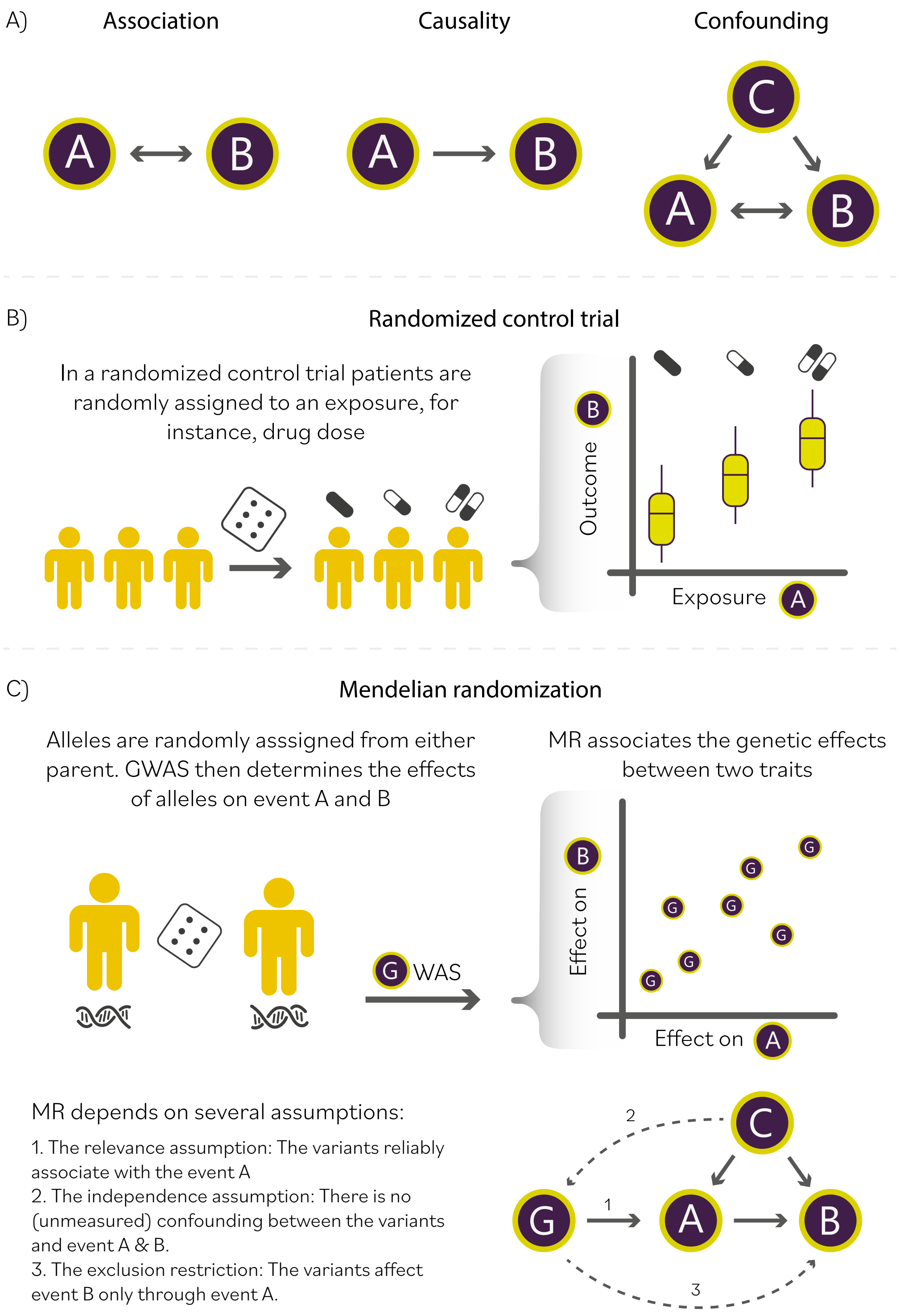
Most genetic studies, such as GWASs, are association-based and observational, and as such they do not provide causal mechanisms. However, genetics has one oddity; it is a highly stable trait that is unlikely to change once associated alleles have been established. While somatic mutations do occur, they are most likely not foundational to complex traits (not considering cancer), and therefore genetic risk for complex traits is basically pre-determined at birth. For example, your smoking status does not impact your genetics (apart from a potential increased risk of somatic mutation), so smoking status cannot confound the observed association between genetics and heart disease. However, genetic factors may impact your 'risk of smoking', which in turn impacts your risk of heart disease. Regardless, it is then still the genetic factor that is causal, just through a more complex pathway.

Because of this, some assumptions can be made about the causal role of genetic variants on complex traits based on GWAS data, assuming population effects and biases are properly accounted for. Of note is that, while the epigenetic state of the genome controls the activity of a genetic variant and the epigenetics may be impacted by the environment, in a GWAS you are measuring the presence of a variant. Therefore, any associations you observe cannot be due to environment because environment cannot change genetic background. What may happen is that the effect of the genetic variant is obscured or enhanced by epigenetics/environment, meaning there is an interaction effect, as has been discussed above (see *The widespread use of additive linear models in complex trait genetics*).

Randomised control trials and Mendelian randomisation

Perhaps the best way to prove a causal relationship between two factors is through an RCT. In an RCT, the study population is divided randomly into two (or more) groups, one of which receives a treatment and the other a placebo (**Fig. 2B**). Because the assignment to the groups is random, the possibility of confounding is reduced so that causal conclusions can be drawn on the results. However, an RCT needs to have a study population sufficiently large to measure the expected effect size and to be unbiased in the selection of the study population to avoid unwanted confounding factors, e.g. when assessing the effectiveness of a treatment for ageing, one should not study a purely paediatric study population.

Inspired by RCTs, MR uses the random segregation of alleles to define the "treatment" and "placebo" groups (**Fig. 2C**) [54-57]. This is done by first selecting genetic variants that are associated with event A, and subsequently evaluating if they impact B in the same way (**Fig. 2C**). As the alleles are randomly combined, they are (in principle) not subject to confounding. However, in MR, causal conclusions can only be drawn when several assumptions about the relationship between the instruments, events A and B, are met [54,58]. There are several approaches to do MR that depend on the data that is available and that deal with the various assumptions in different ways. These will not be discussed in detail here, as these have been well described [54] and this is beyond the scope of this section. Instead, this section discusses the overall concept and the main assumptions impacting interpretability of MR results.



In a randomized control trial patients are randomly assigned to an exposure, for instance, drug dose

Alleles are randomly asssigned from either parent. GWAS then determines the effects of alleles on event A and B

MR associates the genetic effects between two traits

MR depends on several assumptions:

1. The relevance assumption: The variants reliably associate with the event A
2. The independence assumption: There is no (unmeasured) confounding between the variants and event A & B.
3. The exclusion restriction: The variants affect event B only through event A.

Fig. 2. A) Association, confounding and causality. **B)** Schematic of a randomized control trail where the effect of an exposure A on an outcome B is measured. **C)** The basic principles of Mendelian randomization. Figure is inspired by Davey-Smith et al. [61].

These assumptions are:

1. The relevance assumption: the variants reliably associate with event A.
2. The independence assumption: there is no (unmeasured) confounding between the variants and events A and B.
3. The exclusion restriction: the variants affect event B only through event A.

The relevance assumption can be verified by taking replicated associations from good quality GWASs. The independence assumption can be accounted for when using genetics, as explained above (in GWASs and causality), however unaccounted population stratification is a potential violation of this assumption. The exclusion restriction is more difficult to account for. As the same genetic variants can have effects on multiple biological pathways (pleiotropy), it is often impossible to say if this assumption is met. Hence, the current focus of the MR field is in identifying ways to address pleiotropy [59], especially when it comes to the regulation of gene expression by genetic variants [52,60].

When studying the causality of gene expression using MR, the relevance assumption is also under question. The variants that associate with gene expression are eQTL, and, as discussed above (*Linking GWAS loci to genes*), current eQTL resources are far from perfect for explaining how disease heritability is modulated. Furthermore, there are generally only one to three genetic variants that independently associate with the gene. This makes the estimates done by MR less reliable. Hence, great care should be taken when doing such studies to ensure that the eQTL matches the trait on which you want to study the causal effect and that the variants are reliable instruments.

While MR is a wonderful tool to disentangle association from causality, given the underlying assumptions, it is not advisable to apply MR to all genetic datasets without careful consideration. Great care should be taken to verify that the fundamental assumptions are met, or at least addressed with sufficient certainty. Otherwise, any conclusions drawn based on such analyses are just as non-causal as those obtained from observational omics and are potentially more dangerous as causality is implied.

In vitro approaches to model the causal effects of genes

Besides MR and RCTs, the causality of genetic effects and genes can also be evaluated *in vitro* through knockdown and knockout experiments. In such experiments, the genetics of a carefully controlled model system are modulated, and the effects on a phenotype assessed. As this occurs in carefully controlled environments, the possibility of confounding is strongly reduced, and the phenotypic effect is therefore likely due to the knockdown or knockout. The disadvantage is that such systems are more artificial, so there is no guarantee that the knockdown or knockout will be biologically meaningful. As noted in Chapter 2, there are various experimental approaches to perform such studies. Perhaps the most exciting approach to studying complex trait genetics is the CRISPR interference/activation system, as it allows for the assessment of a large number of genes at the same time. Furthermore, it is theoretically possible to control the strength of the knockdown effect. This would allow for experimental setups that could model the large number of small effects observed in GWASs. Secondly, such systems can be used to directly model gene regulatory connections, rather than relying on association as we did in Chapter 3. Naturally, the issues regarding context, cell types and quality also apply to such assays. Moreover, a knockdown, and especially a knockout effect, may not be representative of the true effect of the variant. Indeed, knockout models may better represent Mendelian traits than complex traits. Furthermore, the technical challenges regarding CRISPR assays should not be dismissed (see *Other fine-mapping strategies*). It would be very interesting to see how the field develops assays to modulate the much smaller effects that GWAS variants have.

Time-based designs for inferring causal cascades

Finally, causal relationships between genes can be inferred through experimental setups based on time course. For instance, if you can show that the expression of a TF precedes the expression of one of its target genes, you can infer that the TF is causally affecting the expression of its target, and we applied just such an approach in Chapter 8. While this does provide an indication that the relationship between the TF and gene is causal, the approach is still associational in nature and confounders are likely to be present, so you would still need to verify this through a knockout or knockdown experiment.

Using single-cell RNA sequencing technology, it is now possible to capture the state of single cells. This has as the advantage that cells can be assessed in varying stages of activation within one experiment. This information can be extracted from the RNA data, and the cells can be ordered according to a trajectory through the stages present within a cell population. This is known as pseudo-time analysis. While the time range of such an analysis is relatively limited, it does potentially allow for the estimation of causal relationships between genes. This could be a way to systematically generate directed gene regulatory networks, which would greatly help in interpreting the causal cascade from GWAS variant to phenotype. Due to the limited time range of pseudo-time analysis, it would be optimal to employ a true time-course-based study design, but incorporate pseudo-time analysis to get more resolution for the cellular state at each timepoint.

V. Abstraction and bias in science

The impact of making reality abstract

The ability to make concepts is one of humanity's great tools, allowing us to understand the complex world around us. In science, such abstraction is key for helping us to understand the vast complexity underlying complex traits (for instance, the omnigenic model). However, balancing the reduction of complexity in theoretical models with our ability to understand the true complexity of nature is a massive challenge. In many ways the concept determines what observations can be done, and perhaps moreso, how these oberervations are interpreted. As Ronald Fisher noted on the subject:

> *"Yet any one of the great names of the past, De Moivre or Bayes, Boole or Gauss himself, if by a miracle we can imagine him indoctrinated with the thought of our time, would, I believe, be astonished by the cogency and precision, the directness and accuracy with which problems formerly intolerably encumbered, can, in this age, be recognised, and resolved. In fact they lacked the concepts needed to think clearly about many of their problems.*
>
> *Genetics and Statistics, then, have in common that each in its own field represents a distinctive point of view, which profoundly influences the intellectual processes with which scientific work is approached."*
>
> *– R.A. Fisher, Bateson Lecture at the John Innes Horticultural Institution on, July 6th 1951* [62]

Bias and abstraction in observational omics

The work presented in this thesis depends largely on observational study designs that first produce empirical measurements on a large scale without having a specific hypothesis in mind, instead, basing the design on a concept of how the biology works. One can then test a myriad of hypotheses based on these empirical observations. Therfore, observational omics studies are neither truly inductive, nor deductive. The measurements that are performed and the statistics applied in omics studies are certainly empirical, however the same cannot be said about the hypotheses to which these tools are applied. As the observations that can be made depend on the framework there is for observing them, there is always an inductive component to any study that is done on such data. While there is nothing wrong with this in principle, it is key to be aware of the biases that influence which hypotheses are tested. Such a bias may arise from technical limitations, such as when a technique is difficult to apply in primary cells, so cell-lines are used (Chapter 8 and 9). Or these could be cognitive biases, such as unconsciously wanting to prove a specific idea and being unable to see evidence suggesting differently.

A major source of such cognitive bias lies in the enviroment in which science is performed. Be it the beliefs of a societiy at the time, or by the specific acedemic enviroment. As an example of the latter, in academia there is currently a lot of pressure to perform well, leading to issues of reproducibility, wellbeing and the quality of the scientific output [63-68]. A side effect of this pressure is that there seems to be a tendency in fundamental research to frame projects as actionable while this is, and should, not be the goal of the project [66,68]. While this may be an obvious statement, open and honest reporting is essential for science to progress efficiently. Not to imply, that all of science is dishonest, the precentage of truly fraudulent studies has remained relatively constant, but in recent years there has been a trend heading towards the grey area [63-68].

The observational approaches applied in this thesis are not short of potential biases. These include issues with underrepresented populations, biases in phenotyping, population stratification, inclusion bias, cognitive biases in interpretation, differences in lab protocols and statistical methodology and many more. While bias can never be avoided when doing science, it should not be dismissed. Many obvious pitfalls can be addressed during initial study design, by carefully thinking about biases and pitfalls before any sample is collected and, perhaps most importantly, having a clear hypothesis and plan beforehand. This final point is the major drawback of observational studies, as these are often perfomed on existing data. Consequently, the specific hypotheses are generally generated *ad hoc*. Hence, any biases introduced during study design and data generation that might impact the research question cannot be addressed. Instead, such biases need to be dealt with *in silico*, which is not ideal or not done at all as it is sometimes impossible to do so after the fact.

Bias resulting from the scope of obervational omics

Given the extensive nature of the measurements done in observational omics studies, it is challenging for the individual researcher to become an expert on every single trait, protein, gene or cell type. In fact, many researchers have dedicated their scientific career to studying single genes, indicative of the complexity behind each association. Therefore, in large-scale studies, abstractions are applied, which carry their own limitations, as noted By Ronald Fisher:

> *"So long as the statistician was supposed to concern himself only with vast aggregates of massed data, his acquaintance with the detailed processes by which they came into existence was bound to be vague; each part of the whole contained its own complications and its own enigmas; in the treatment of the mass these were necessarily almost wholly unrecognisable, and the interpretation was harassed by innumerable unanswerable queries.*
>
> …
>
> *Direct contact with what is actually done in experimentation helps the statistician in another very essential way, by leading him to consider variations in procedure, and the reasons why one method is to be preferred to others. The whole wide subject of experimental design is opened out by this consideration."*
>
> *– R.A. Fisher, Bateson Lecture at the John Innes Horticultural Institution on July 6th 1951* [62]

As a consequence of this large scope, there always are some associations to be found. In light of this, critics have described such studies as "fishing expeditions" [69]. An advantage of observational omics research, when properly applied, is that the researcher is less likely to be biased by an a priori model of what they are observing and hence is more likely to observe novel mechanisms. This paradigm has been described as "night science" which contrasts with "day science", the formal testing of structured hypotheses [70]. As such, observational studies lend themselves well to exploring new areas of research, such as complex trait genetics, where the scope of the subject is not yet fully defined [71].

It should be noted that it takes years of careful follow-up study to verify and understand the mechanisms causal to the associations made in observational studies. That is not to say that that observational studies are not useful. On the contrary, association-based studies are an essential first step, but they should be viewed as what they are and should not be interpreted as causal or actionable in a clinical sense.

Managing bias through triangulation

One of the best ways to deal with bias and confounding is to apply triangulation. The term triangulation is derived from surveying, where a point on the map is determined using measurements from two other points in space. When testing hypotheses, you use multiple methods, each with different biases and confounding, to test the hypothesis in question. If all these methods give the same answer, you can be confident that you have the right answer. For example, in Chapter 5, we identify genetic factors in the *TLR1* locus that are associated to the *ex vivo* cytokine response of immune cells. We then show, in an independent dataset, that these factors also regulate gene expression levels. Furthermore, this locus harbours strong signals of natural selection [72] and GWASs have identified associations to several immune diseases. These are all signs indicating that the identified genetic factors do indeed have an important role in the cytokine response of immune cells.

While triangulation is great in theory, it should be noted that as an individual lab, this is unlikely to always be successful because being part of that lab is a bias in itself [73]. As such triangulation is more a way to deal with bias as a scientific community, than it is for the individual project. Triangulation is often difficult to execute in wet lab experiments due to time and cost constraints, but it is often possible to use multiple methods when analysing the data. This does not cover the biases in data generation but does capture issues with various statistical tests. However, the type of statistical tests should be carefully chosen, as many methods rely on the same basic ideas and might be susceptible to the same biases, giving a false sense of security when using them to triangulate a hypothesis. Furthermore, while open data-sharing is fantastic, many studies rely on the same resources, e.g. 1000 genomes, GTEx or the UK biobank. Hence, any bias present in these datasets will impact all the conclusions made in papers using this data.

VI. Epilogue

To summarise this discussion, I present the following statement, which is obvious but nonetheless true: The downstream effects of genetic variants must be studied from multiple perspectives by integrating different data to get validated and robust answers. This is, however, not possible for one individual or even one research group. Therefore, collaboration in the community is of paramount importance, either to generate huge databanks, gather different perspectives or critically reviewing each other's work.

In conclusion, the main take home points of this thesis are:

- Co-expression networks are a promising tool to interrogate the networks that lead genetic variants to impact a phenotype, however given their complexity and the transient nature of cellular states, careful and extensive validation of any findings is required.
- The genes central in such co-expression networks have properties that match the core genes described by the omnigenic model, however, cell-type composition of the network as well as the overall expression level of such genes in the network are confounding factors that should be taken into account
- Genetic effects on the regulatory capacity of enhancers or promoters are most informative for interpreting GWAS variants if they are studied in a cell-type and context that matches the phenotype under study.
- The *ex vivo* cytokine production capacity of immune cells has a substantial genetic component. However, future studies are required to determine if these effects also carry though *in vivo*. Furthermore, the work in this thesis provides a proof of concept, but replication in larger cohorts is required to draw definitive conclusions.
- The genetic basis of cytokine production capacity may have been shaped by polygenic adaptation, but future work is needed to confirm this hypothesis.
- Fine-mapping the exact causal variants underlying GWAS loci is a complex task that requires multiple lines of independent evidence from assays with different fundamental assumptions.

With regards to the question that was posed at the start of this discussion: Have we gained actionable insight from the work in this thesis? My answer would be no. In my view, the goal of fundamental research is not to deliver such actionable answers, but rather to shed light on biological mechanisms that are opaque and complex in the hope that future research will be able to ask such actionable questions. I would like to think that we have made some very small steps into that direction with the work presented in this thesis. In the end it is our task as scientists to try and comprehend the inner workings of the world to the best of our ability, whether that world is "*an ordered universe*" or "*a stew of mixed ingredients*". Given the vast complexity of this task, it may at times feel as though we are like Sisyphus pushing the boulder up the hill. While indeed, like Marcus Aurelius wrote, "*all things, distinct as they are, nevertheless permeate and respond to each other*", we should not let this discourage us, rather we should take pride in the effort, be inspired and wonder at the beauty that can be found in the madness of it all.

References

1. Simcoe, M. et al. Genome-wide association study in almost 195,000 individuals identifies 50 previously unidentified genetic loci for eye color. Sci. Adv. 7, eabd1239.
2. Connally, N. et al. The missing link between genetic association and regulatory function. 2021.06.08.21258515 (2021) doi:10.1101/2021.06.08.21258515.
3. Võsa, U. et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. 53, 1300–1310 (2021).
4. Kerimov, N. et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat. Genet. 53, 1290–1299 (2021).
5. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. Nature 593, 238–243 (2021).
6. Barbeira, A. N. et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. 814350 (2020) doi:10.1101/814350.
7. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. Nat. Genet. 52, 626–633 (2020).
8. Sadler, M. C., Auwerx, C., Porcu, E. & Kutalik, Z. Quantifying mediation between omics layers and complex traits. 2021.09.29.462396 (2021) doi:10.1101/2021.09.29.462396.
9. Elmentaite, R. et al. Cells of the human intestinal tract mapped across space and time. Nature 597, 250–255 (2021).
10. van der Wijst, M. et al. The single-cell eQTLGen consortium. eLife 9, e52155.
11. Jayaram, N., Usvyat, D. & R. Martin, A. C. Evaluating tools for transcription factor binding site prediction. BMC Bioinformatics 17, 1–12 (2016).
12. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
13. Asimit, J. L., Hatzikotoulas, K., McCarthy, M., Morris, A. P. & Zeggini, E. Trans-ethnic study design approaches for fine-mapping. Eur. J. Hum. Genet. 24, 1330 (2016).
14. Kanai, M. et al. Insights from complex trait fine-mapping across diverse populations. 2021.09.03.21262975 (2021) doi:10.1101/2021.09.03.21262975.
15. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat. Rev. Genet. 19, 491 (2018).
16. Tian, R. et al. Pitfalls in Single Clone CRISPR-Cas9 Mutagenesis to Fine-Map Regulatory Intervals. Genes 11, 504 (2020).
17. Fu, R. et al. Systematic decomposition of sequence determinants governing CRISPR/Cas9 specificity. Nat. Commun. 13, 474 (2022).
18. Squair, J. W. et al. Confronting false discoveries in single-cell differential expression. Nat. Commun. 12, 5692 (2021).
19. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169, 1177–1186 (2017).
20. Vuckovic, D. et al. The Polygenic and Monogenic Basis of Blood Traits and Diseases. Cell 182, 1214-1231.e11 (2020).
21. Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell 177, 1022-1034.e6 (2019).
22. Sinnott-Armstrong, N., Naqvi, S., Rivas, M. & Pritchard, J. K. GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. eLife 10, e58615 (2021).
23. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. Cell 173, 1573–1580 (2018).
24. ICDA Recommendations and White Paper.pdf. Google Docs https://drive.google.com/file/d/16SVJ 5lbneN 9hB9E03PZMhpescAN527HO/view?usp=embed_facebook.
25. Yengo, L. et al. A Saturated Map of Common Genetic Variants Associated with Human Height from 5.4 Million Individuals of Diverse Ancestries. 2022.01.07.475305 (2022) doi:10.1101/2022.01.07.475305.
26. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. 51, 584 (2019).
27. Amariuta, T. et al. Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. Nat. Genet. 52, 1346–1354 (2020).
28. Sohail, M. et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. eLife 8, e39702 (2019).
29. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLOS Comput. Biol. 11, e1004219 (2015).
30. Giambartolomei, C. et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLOS Genet. 10, e1004383 (2014).
31. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. PLOS Comput. Biol. 12, e1004714 (2016).
32. Warmerdam, R., Lanting, P., Lifelines Cohort Study, Deelen, P. & Franke, L. Idéfix: identifying accidental sample mix-ups in biobanks using polygenic scores. Bioinformatics 38, 1059–1066 (2022).
33. Manolio, T. A. et al. Finding the missing heritability of complex diseases. Nature 461, 747–753 (2009).
34. O'Connor, M. J. et al. Recessive Genome-wide Meta-analysis Illuminates Genetic Architecture of Type 2 Diabetes. Diabetes db210545 (2021) doi:10.2337/db21-0545.
35. Slim, L., Chatelain, C., Azencott, C.-A. & Vert, J.-P. Novel methods for epistasis detection in genome-wide association studies. PLOS ONE 15, e0242927 (2020).

36. Niel, C., Sinoquet, C., Dina, C. & Rocheleau, G. A survey about methods dedicated to epistasis detection. Front. Genet. 6, (2015).
37. Bouziat, R. et al. Reovirus infection triggers inflammatory responses to dietary antigens and development of coeliac disease. Science 356, 44–50 (2017).
38. Harley, J. B. et al. Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity. Nat. Genet. 50, 699–707 (2018).
39. Taussig, D. & Wine, Y. When a virus lies in wait. eLife 10, e71121 (2021).
40. Bjornevik, K. et al. Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. Science 375, 296–301 (2022).
41. Heil, B. J. et al. Reproducibility standards for machine learning in the life sciences. Nat. Methods 18, 1132–1135 (2021).
42. Kapoor, S. & Narayanan, A. (Ir)Reproducible Machine Learning: A Case Study. 7.
43. Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of applying machine learning in genomics. Nat. Rev. Genet. 23, 169–181 (2022).
44. Christophersen, A. et al. Distinct phenotype of CD4+ T cells driving coeliac disease identified in multiple autoimmune conditions. Nat. Med. 25, 734–737 (2019).
45. Atlasy, N. et al. Single cell transcriptome atlas of immune cells in human small intestine and in coeliac disease. 721258 (2019) doi:10.1101/721258.
46. Low, L. A., Mummery, C., Berridge, B. R., Austin, C. P. & Tagle, D. A. Organs-on-chips: into the next decade. Nat. Rev. Drug Discov. 20, 345–361 (2021).
47. Moerkens, R., Mooiweer, J., Withoff, S. & Wijmenga, C. Coeliac disease-on-chip: Modeling a multifactorial disease in vitro. United Eur. Gastroenterol. J. 7, 467–476 (2019).
48. Caio, G. et al. Coeliac disease: a comprehensive current review. BMC Med. 17, 1–20 (2019).
49. Smith, G. D. & Ebrahim, S. Data dredging, bias, or confounding: They can all get you into the BMJ and the Friday papers. BMJ 325, 1437–1438 (2002).
50. Khaw, K. T. et al. Relation between plasma ascorbic acid and mortality in men and women in EPIC-Norfolk prospective study: a prospective population study. European Prospective Investigation into Cancer and Nutrition. Lancet Lond. Engl. 357, 657–663 (2001).
51. Heart Protection Study Collaborative Group. MRC/BHF Heart Protection Study of antioxidant vitamin supplementation in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet Lond. Engl. 360, 23–33 (2002).
52. Porcu, E. et al. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. Nat. Commun. 12, 5647 (2021).
53. Klatsky, A. L. & Udaltsova, N. Abounding confounding. Addiction 108, 1549–1552 (2013).
54. Sanderson, E. et al. Mendelian randomization. Nat. Rev. Methods Primer 2, 1–21 (2022).
55. Smith, G. D. & Ebrahim, S. ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? Int. J. Epidemiol. 32, 1–22 (2003).
56. Smith, G. D. & Ebrahim, S. Mendelian randomization: prospects, potentials, and limitations. Int. J. Epidemiol. 33, 30–42 (2004).
57. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. 48, 481–487 (2016).
58. Davies, N. M., Holmes, M. V. & Smith, G. D. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ 362, k601 (2018).
59. Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M. & He, X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. Nat. Genet. 52, 740–747 (2020).
60. van der Graaf, A. et al. Mendelian randomization while jointly modeling cis genetics identifies causal relationships between gene expression and lipids. Nat. Commun. 11, 4930 (2020).
61. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum. Mol. Genet. 23, R89-98 (2014).
62. Fisher, R. Statistical methods in genetics1. Int. J. Epidemiol. 39, 329–335 (2010).
63. Gandevia, S. Publication pressure and scientific misconduct: why we need more open governance. Spinal Cord 56, 821–822 (2018).
64. Baker, M. 1,500 scientists lift the lid on reproducibility. Nat. News 533, 452 (2016).
65. Publish or perish. Nature 521, 159–159 (2015).
66. Chiu, K., Grundy, Q. & Bero, L. ‘Spin’ in published biomedical literature: A methodological systematic review. PLOS Biol. 15, e2002173 (2017).
67. Gopalakrishna, G. et al. Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. PLOS ONE 17, e0263023 (2022).
68. Caulfield, T. & Ogbogu, U. The commercialization of university-based research: Balancing risks and benefits. BMC Med. Ethics 16, 1–7 (2015).
69. Felin, T., Koenderink, J., Krueger, J. I., Noble, D. & Ellis, G. F. R. The data-hypothesis relationship. Genome Biol. 22, 1–6 (2021).
70. Yanai, I. & Lercher, M. Night science. Genome Biol 20, 179 (2019).
71. Yanai, I. & Lercher, M. A hypothesis is a liability. Genome Biol. 21, 231 (2020).
72. Heffelfinger, C. et al. Haplotype structure and positive selection at TLR1. Eur. J. Hum. Genet. EJHG 22, 551–557 (2014).
73. Schweinsberg, M. et al. Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. Organ. Behav. Hum. Decis. Process. 165, 228–249 (2021).

Appendices

Summary

Variation is omnipresent, it shapes our world, and it shapes us. A source of such variation that impacts us all is the DNA. The past years have seen extensive profiling of the DNA in relation to a whole host of traits. This has yielded extensive maps of which genetic variants are associated to which traits. These maps are known as genome wide association studies (GWASs). Some GWASs provide clear interpretation for their trait. However, for many traits whose causal cascade is complex (complex traits), GWASs have proven challenging to interpret.

This thesis studies the genetic basis of complex traits and the regulatory mechanisms that are foundational to them, with the aim of providing better interpretation for GWASs on complex traits. Part I presents a broad overview of the mechanisms with which genetic variants can impact complex traits through the regulation of gene expression. In part II, the relationship between genetic factors and immune function is discussed. Part III dives deeply into the role genetic factors have in causing coeliac disease by impacting specific cellular contexts. In part IV, the work is placed into context and future perspectives discussed.

Part I

In Chapter 1, an introduction is given to the work that is foundational to this thesis. Background and context are provided for the topics addressed in this thesis and how they fit together.

Chapter 2 continues this trend by reviewing current approaches to interpreting GWAS effects. Most notably, we discuss *in vitro* and in silico approaches to fine-mapping causal variants within GWAS loci. We further discuss approaches to link the GWAS loci to genes to facilitate interpretation of the downstream consequences of the GWAS variants. We highlight the cell-type and context specific nature of genetic effects and the challenges that are associated with detecting them. Finally, we discuss the relevance of these processes considering the small effects of individual GWAS variants.

Considering the small effects GWAS variants have, we asses how the collective of subtle effects observed in GWAS impact the regulatory relationships between genes in Chapter 3. To do so we built a co-expression map of publicly available RNA-seq data and linked this to GWAS summary statistics. We observe that the gene networks associated with disease are interconnected, and the genes that are central in their respective network more likely to lead to rare diseases when variants occur in these genes.

Part II

As highlighted in Chapter 1, 2 and 3, the effects of GWAS variants can be cell-type specific. In Chapter 4, we attempt to enhance the detection of such cell-type specific genetic effects in immune cells. To do so, we develop a method that can identify cell-type specific expression quantitative trait loci (eQTL) using RNA sequencing data obtained from mixed blood samples. We show that by simultaneously modelling the cell-proportions in blood with the genetic effect on expression, cell-type specific eQTL effects can be identified. We validate these effects by using eQTL data from isolated cell-types and single-cell eQTL.

Chapters 5, 6 and 7 study variation in the immune cells, not solely in terms of gene expression, but also in terms of proteins expressed by the immune cells. In these chapters we study cytokines, a class of immune proteins acting as messengers between immune cells.

In Chapter 5, we study the genetic factors that impact the normal variation in *ex vivo* immune response by measuring cytokine responses at the protein level. To do so, we model genetic effects on correlated cytokines jointly to increase detection power. We assess how the cytokines interact, and if their association with genetic factors is due to pleiotropy or mediation. We link the observations on the protein level to independent data on *ex vivo* stimulated gene expression.

In Chapter 6, we perform a broad screen to identify the relative contribution of different genetic and environmental factors to variation in *ex vivo* cytokine responses. We observe that several host factors, including genetics, metabolite levels, cell-proportions and immune markers are associated with the capacity of immune cells to produce cytokines when stimulated. The magnitude of the associations depended greatly on the cytokine measured and which stimulation was used.

In Chapter 7, we assess if the genetic basis for *ex vivo* cytokine response, as well as immune diseases has been shaped by selective pressures that act polygenically. We applied polygenic risk scores to genetic data obtained from ancient humans to trace how the relative genetic risk for these traits changed over time.

Part III

Every complex trait has its own context and complications. In chapters 8 and 9 we study the role of specific genetic factors by fine-mapping genetic loci associated with coeliac disease, an auto-immune disorder where ingestion of gluten lead to a disproportionate immune response.

In Chapter 8, we study how one of the essential cell-types in coeliac disease, gluten specific CD4+ T cells, get activated by studying gene expression, chromatin state and protein levels. We link this information to the known genetic factors for coeliac disease.

In Chapter 9, we apply an *in vitro* fine-mapping strategy to identify genetic factors that may disrupt enhancer activity in several cell-types relevant for coeliac disease. We identify several enhancer and promoter elements that show suggestive evidence of being impacted by the variants located in them.

Part IV

Finally, in the last chapter of this thesis, several aspects of interpreting the genetic basis for complex traits are discussed in detail. The approaches applied in this thesis are critically reflected on, and their interpretability discussed. The inherent difficulties with observational data and causality are discussed, as are some of the biases present in this work. Furthermore, perspectives are provided as to where the field might be heading next and what exciting prospects are on the horizon.

Samenvatting

We worden omringd door variatie, variatie vormt onze wereld, en variatie vormt ons. Een bron van variatie die een invloed heeft op ons allen is ons DNA. In de afgelopen jaren heeft de wetenschap uitvoerige profielen opgesteld waaraan te zien is hoe de variatie in ons DNA relateert aan onze eigenschappen. Zulke profielen worden genoom wijde associatie studies (GWASs) genoemd.

Voor sommige van deze GWASs is het relatief eenvoudig om te interpreteren hoe de variatie in het DNA de eigenschap beïnvloedt. Echter, voor veel van deze GWASs is het niet duidelijk hoe deze variatie uiteindelijk de eigenschap beïnvloedt. Dit komt deels omdat het verband tussen het DNA en de eigenschap vaak erg complex is en veroorzaakt wordt door meerdere bronnen van variatie naast het DNA.

In dit proefschrift wordt bestudeerd hoe de genetische basis voor zulke complexe eigenschappen haar effect uitvoert, met het doel om een beter begrip te krijgen van het tot stand komen van deze complexe eigenschappen. In deel I van dit proefschrift wordt een overzicht gegeven van de mechanismen waarmee variatie in het DNA de regulatie van genexpressie kan beïnvloeden. In deel II wordt de relatie tussen genetische factoren en immuun functie besproken. In deel III wordt de rol van genetische factoren bestudeerd in het veroorzaken van coeliakie in een aantal specifieke cellulaire contexten. Tot slot wordt in deel IV het werk in context geplaatst en worden toekomstperspectieven besproken.

Deel I

In hoofdstuk 1 wordt een introductie gegeven van het werk dat ten grondslag ligt aan de hoofdstukken in dit proefschrift

Hoofdstuk 2 zet deze trend voort door huidige technieken voor het interpreteren van GWAS-effecten in detail te bespreken. Specifiek gaan wij hier in op *in vitro* en *in sillico* methoden om de causale genetische effecten te bepalen binnen GWAS loci. Verder bespreken we methoden om de genen te vinden die beïnvloed worden door GWAS-varianten. We sluiten af met een discussie over de waarde van de individuele variant in het licht van de vaak kleine effecten die ze hebben.

Gezien deze kleine effecten die GWAS-varianten hebben, bestuderen we in hoofdstuk 3 hoe het geheel van deze effecten samen zou kunnen werken om de regulatoire relaties tussen genen te beïnvloeden. Dit doen we door een co-expressie netwerk op te stellen aan de hand van publiek beschikbare RNA-seq data. We combineren dit netwerk aan GWAS-informatie om zo te kunnen achterhalen of er genen zijn waar de GWAS-informatie samenkomt. We zien dat bepaalde genen in deze netwerken verrijkt zijn om te leiden tot zeldzame ziekten als varianten plaatsvinden.

Deel II

In hoofdstuk 1, 2 en 3 wordt besproken dat de effecten van GWAS-varianten cel-type-specifiek kunnen zijn. In hoofdstuk 4 proberen we de detectie van zulke cel-type-specifieke effecten te verbeteren in immuuncellen. Dit doen we door een methode te verbeteren die aan de hand van bulk RNA-seq en geschatte cel proporties een indicatie kan geven van de celtype specifieke genetische effecten (eQTL). We valideren deze methode met eQTL data uit geïsoleerde celtypen en ‘single-cell’ eQTL data.

Hoofdstukken 5, 6 en 7 bestuderen ook de regulatie van immuuncellen, maar nu niet enkel op het niveau van genexpressie, maar op het niveau van de eiwitten die van belang zijn in de signaalfuncties van het immuunsysteem (cytokines).

In hoofdstuk 5 bestuderen we hoe genetische factoren de variatie in *ex vivo* immuunreacties beïnvloeden door cytokineproductie te meten. We koppelen de gemeten cytokineproductie profielen aan genetische varianten door cytokines die vergelijkbaar reageren samen te modeleren. Dit geeft meer statistische kracht om associaties te vinden tussen het DNA en de cytokine expressie. We kijken verder naar hoe het netwerk eruitziet rond de geïdentificeerde genetische factoren op een eiwit- en genexpressie niveau.

In hoofdstuk 6 voeren wij een brede zoektocht uit naar wat de relatieve bijdrage van verschillende genetische- en omgevingsfactoren is in het bepalen van de *ex vivo* cytokineproductie. We observeren dat meerdere factoren een significante invloed hebben op de cytokineproductie, waaronder genetica, metabolieten, de hoeveelheid immuun cellen en verscheidene immuun eiwitten. Het blijkt dat de bijdrage van deze factoren erg varieert erg afhankelijk van de gemeten cytokine en de stimulatie die gebruikt is.

In hoofdstuk 7 bespreken we hoe de genetische basis van *ex vivo* cytokineproductie gevormd is door selectieve druk. We passen polygene risico scores voor cytokineproductie en immuunziekten toe op genetische data van archeologische opgravingen om zo te bestuderen hoe deze genetische profielen door de tijd heen veranderd zijn.

Deel III

Elke complexe eigenschap heeft zijn eigen context en complicaties. In hoofdstukken 8 en 9 bestuderen we de rol die specifieke genetische factoren hebben bij het veroorzaken van coeliakie, een auto-immuun ziekte waarbij de inname van gluten leidt tot een immuunreactie die uiteindelijk tot darm schade leidt.

In hoofdstuk 8 bestuderen we hoe één van de essentiële celtypen in coeliakie, de glutenspecifieke T cellen, geactiveerd worden. Dit doen we door te kijken naar genexpressie, de openheid van het chromatine en de eiwit niveaus. We koppelen deze informatie aan de bekende genetische factoren die geassocieerd zijn met coeliakie.

In hoofdstuk 9 passen we een *in vitro* methode toe om te bepalen welke specifieke genetische factoren causaal zouden kunnen zijn voor het veroorzaken van coeliakie. We identificeren meerdere ‘enhancer’ en ‘promoter’ elementen in het DNA die beïnvloed worden door genetische varianten in die elementen.

Deel IV

In het laatste hoofdstuk van dit proefschrift worden meerdere aspecten bediscussieerd die invloed hebben op het interpreteren van de genetische basis van complexe eigenschappen. De invloed die de gebruikte methoden hebben op de interpreteerbaarheid van de resultaten wordt besproken evenals de inherente moeilijkheden van het gebruik van observationele data. Ook worden de spannende vooruitzichten voor het veld geschetst.

Curriculum vitae

Olivier Berend Bakker was born on 15 April 1996, in the village of Swifterbant in The Netherlands. He graduated from Ichthus College Dronten in 2014 after which he proceeded to study Bio-informatics at the Hanze University of Applied Science in Groningen. In 2016, he started a research internship at the Department of Genetics of the University Medical Centre Groningen under the supervision of Yang Li where he studied the genetic basis of cytokine production. He obtained his bachelors degree *cum laude* in 2017. His bachelor thesis on the shared genetic basis of cytokine production won second place for both the Hanze Innovation Award as well as the Royal Dutch Chemical Society's 'Golden Spatula' award for best bachelors thesis.

In 2017, Olivier began his PhD studies under the supervision of doctors Yang Li and Iris Jonkers and prof. Cisca Wijmenga at the Department of Genetics of the University Medical Centre Groningen. Here he continued to study the genetic basis of complex traits, with a particular focus on coeliac disease. During this time, Olivier presented his work at the EMBL symposium: From Genomes to Complex Traits in 2019 and at the European Society for Human Genetics conference in 2020.

Olivier is currently continuing his studies into the genetic basis of complex immune disease as a post-doctoral fellow at the Wellcome Sanger Institute in Hinxton, United Kingdom in the lab of dr. Gosia Trynka.

First author publications

- Bakker O.B., Ramirez-Sanchez A & Borek Z.A. (2021), Potential impact of celiac disease genetic risk factors on T cell receptor signalling in gluten-specific CD4+ T cells, Scientific reports, 11,1,1-15
- Bakker O.B & Claringbould A. (2021) Linking common and rare disease genetics through gene regulatory networks, medRxiv
- Bakker O.B. & Broekema R.V. (2020) A practical view of fine-mapping and gene prioritization in the post-genome-wide association era, Open biology, 10,1,190221
- Bakker O.B. (2018), Integration of multi-omics data and deep phenotyping enables prediction of cytokine responses, Nature immunology,19,7,776-786

Co-authored publications

Second author publications papers indicated by ‡

- ‡ Kuijpers Y. et al. (2022) Evolutionary trajectories of complex traits in European populations of modern humans, Frontiers in Genetics
- Sanchez-Maldonado et al. (2022) Type 2 Diabetes-Related Variants Influence the Risk of Developing Prostate Cancer: A Population-Based Case-Control Study and Meta-Analysis, Cancers
- Ruth Katherine S et al. (2021) Genetic insights into biological mechanisms governing human ovarian ageing, Nature. 596
- Van Rheenen W. et al. (2021) Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology, Nature Genetics
- ‡ Domínguez-Andrés J. et al. (2021) Evolution of cytokine production capacity in ancient and modern European populations, Elife
- de Klein N. et al. (2021) Brain expression quantitative trait locus and network analysis reveals downstream effects and putative drivers for brain-related diseases, bioRxiv

- Mc Intyre K et al. (2021) Lifelines COVID-19 cohort: investigating COVID-19 infection and its health and societal impacts in a Dutch population-based cohort, BMJ open, 11

- Chu X. et al. (2021) Integration of metabolomics genomics and immune phenotypes reveals the causal roles of metabolites in disease, Genome biology, 22

- Moorlag SCFM et al. (2021) An integrative genomics approach identifies KDM4 as a modulator of trained immunity, European journal of immunology

- Aguirre-Gamboa R. et al. (2020) Deconvolution of bulk blood eQTL effects into immune cell subpopulations, BMC bioinformatics, 21

- de Vries D. et al. (2020) Integrating GWAS with bulk and single-cell RNA-sequencing reveals a role for LY86 in the anti-Candida host response, PLoS pathogens, 16

- Mc Intyre K. et al. (2020) The Lifelines COVID-19 Cohort: a questionnaire-based study to investigate COVID-19 infection and its health and societal impacts in a Dutch population-based cohort, medRxiv

- Sánchez-Maldonado J. et al. (2019) Steroid hormone-related polymorphisms associate with the development of bone erosions in rheumatoid arthritis and help to predict disease progression:Results from the REPAIR consortium, Scientific reports, 9

Acknowledgements

I have put off writing this part for far too long, probably because I have so enjoyed my time in Groningen and putting this in words truly signifies the end of a chapter. The past four years have been an unexpected journey, which I would not have completed without your fellowship. Before I came to Groningen, I wasn't a man of many words, however as illustrated by this thesis, that has changed. So, without further ado...

Colleagues and collaborators

To Yang, Iris and Cisca I can easily say that without you offering me a PhD position I would be on a very different path. I am very gratefull you saw potential that I did not see myself at that time, and I hope it has been as good as an experience to work with me as it has been to work with you. Yang, I greatly appreciate the opportunity and guidance you have given during those first years which has formed the foundation on which this thesis is built. Cisca, I was inspired right of the bat by your approach to voicing concerns and being open about asking questions, as well as your relentless work ethic and no-nonsense attitude. Even though our paths diverged halfway through, with Yang becoming a full Professor in Hannover and Cisca becoming the Rector Magnificus of our fine university, that has not lessened the impact of your guidance on my scientific thinking.

To Iris I can truly and honestly say you have been a fantastic supervisor, always knowing when to push forward in the face of adversity. While my critical nature must have been a pain to deal with at times, it never deterred you from having an open discussion about problems while keeping a pragmatic attitude towards solving them. This is a quality that I learned a great deal from. Although I do hope that, like a fine wine, my critical nature has mellowed a bit with time and your guidance. Ours has been a great collaboration in that our knowledgebase complemented each other nicely, and I have greatly enjoyed learning from you about coeliac disease, T cells, transcriptional mechanics and getting to know a bit of what black magic happens in the wet-lab. I am also very appreciative of your management style, where I felt there was a trust from both sides that things will be done right. While Fortuna wasn't always on our side, we still produced nice work that I am proud of having done!

To Sebo, I would like to thank you for the guidance and your efforts, when suddenly you had an extra student to supervise. Furthermore, your feedback on the biological aspects, as well as the structure of the chapters in this thesis have helped to improve them greatly. I also greatly enjoyed our converstations about various science fiction media over the years, as well as the great team BBQ's.

To Lude and Patrick and Annique, where one path diverges, another one opens. It has been a pleasure to work with you and expand the scope of my studies into the wacky and wild world of gene networks. I have learned a lot from your points of view in finding the balance between pragmatism and idealism, optimism and realism and have greatly enjoyed our lengthy debates. For instance, I was greatly amused to find that our slack channel on the Downstreamer project was the most active channel in the department by a considerable margin, with around 12,000 messages sent. While we often had different points of view on scientific nitty gritty between the four of us, we always managed to find common ground after a good discussion, this is a way of working that I greatly appreciated and hope to carry forth with me.

To Mihai, your enthusiasm for science is inspiring, and I have greatly enjoyed the work we have done together. As a born sceptic (although I prefer the term realist) I tend to get a bit down with matters. However, a trip to Nijmegen always was a good cure, and I always came back with new sense of inspiration and excitement!

Furthermore, I would like to thank all the members of the assessment committee, professors Gilissen, Duffy and Laman, for their efforts in carefully reading through the many words in this thesis.

To Raul, while it has been a few years now, I had a great time working with and learning from you. You were always calm, happy to explain things I didn't understand and willing to have a good discussion. I still benefit every day from having learned your ggplot skills!

To Roeland, my partner in crime. In many ways our collaboration has been one of yin & yang or wet and dry lab if you will. It has been a pleasure to work together for the past four years and learn from you about what goes on in the wet-lab. We always managed to have a good discussion and to complement each other's expertise, making the work better for it! One day tough, I will beat you in Shards of Infinity ;)

To Adriaan, my fellow critical brother in arms. I have greatly enjoyed our discussions and conversations over the years and have learned a lot from your expertise on statistical genetics. You were never afraid to highlight where I went wrong in my thinking, a rare manifestation of honesty which I greatly appreciate. I do greatly miss our "hobbeltjes" where we could freely vent our frustrations of the day or discuss the tricky problems we ran into.

To Harm, you were there from day one when we both joined the department for an internship in what now seems a lifetime ago! I have greatly enjoyed learning from your Java and web programming skills, skills which I still use constantly. I wish you all the best in your own journey!

To Eva & Joram. Ah, those glorious few months pre-corona. A wierd time when you could just sit next to eachother and talk shop over a coffee, I look back on them with great fondness, as there was always the right balance between debugging science, and complete nonsense. In hindsight the only thing I would have done differently is make more pizzas and drink more coffee & wine together. Also much love to Fred & Berend, it has been amazing to see them grow up ;)

To Dylan, you were the glue that held everybody together, always helping people out, organizing things or being crazy enough to embark on a journey to organize 10 years of research data with me. I think everybody in the department has benefitted enormously from your efforts in bringing people together, and I for one am certainly grateful for that fact!

To Kate and Jacky, having an editor onsite has been a privilege, and you have certainly helped improve my writing a lot. Although I am still hoping the day will come when I properly use 'where' and don't write 'enhancher' first time ;). Your willingness to edit our manuscripts, with deadlines that were far too short (mea culpa) is amazing. I also enjoyed our lunchtime discussions on wildly varying topics from etymology of the various European words for French fries to the intricacies of studying family trees.

To the students I have supervised, Yunus, Twan, Leon and Mathijs, I have enjoyed supervising you a lot, it was always fun, and I wish you all the best of luck for the future!

To all the other PI's, group members and support staff, Nine, Sasha, Ying, Morris, Helene, Janneke Harm-Jan, Shuang, Marije, Niek, Urmo, Freerk, Roy, Robert, Martijn, Monique, Paulien, Sipko, Kieu, Kai, Rutger, Jody, Werna, Arnau, Sergio, Yanni, Miaozhen, Xiaojing, Zenhua, Shixian, Pieter, Gerben, Damiano. You have been very kind, welcoming, and nice bunch to work with, be it in keeping the cluster and wet-lab running smoothly, helping navigate the intricacies of the UMCG paperwork, having a good conversation over lunch, helping to think along on new ideas, nerding out over woodworking tools, eating pho, arepas, shabu-shabu or pizzas, playing boardgames and Factorio or just having a laugh and discussing terrible films and nerding out over science fiction.

And finally, to the people I inevitably forgot to mention, mea culpa, and thank you!

Oud docenten

Aan meneer Moraal, ook aan u heb ik het nodige te danken. Hoewel ik vrees dat er van de wiskunde en scheikunde niet veel terecht is gekomen, kan ik wel zeggen dat als wij al die jaren geleden niet met programmeren bezig geweest waren ik wellicht nooit had ontdekt dat het verschrikkelijk leuk is om te doen, met als gevolg dat ik waarschijnlijk nooit aan mijn studie bio-informatica was begonnen.

Sprekende over bio-informatica, ook dank aan het docenten team op de Hanze, Michiel, Tsjerk, Marcel, Martijn, Arne en Piet. Ik heb been ontzettende leuke tijd gehad tijdens mijn studie op de Hanze, waar uiteindelijk de basis gelegd is voor het werk in dit proefschrift.

Friends

Reinout and Saul, how long have we known each other? By my reconning, it's been 14 years. That number is both scary and amazing at the same time, more so because its more than half our lives. It has been a pleasure to have you as friends. Even though we may have completely opposite ideologies at times, I have always felt at ease, and have enjoyed our discussions where we test the limits of our points of view. It has challenged my preconceptions and made me a better thinker in the process. Not to give the impression that it has been all serious, rather the opposite, there has been plenty of, in fact mostly, nonsense, be it in the production of FG, or in late night game sessions during our student years, or just hanging out and raving about politics, the state of the world, and frustrations in our jobs. I hope the risk of rain will remain severe for the years to come.

My dear Yunus. Where to begin ... You have helped me grow in so many ways that I can truly and honestly say that without your support, I wouldn't be the person I am today. Although upon further consideration, I don't know if that is a compliment or an insult, I'll leave that up to you ;). If ever I got frustrated with the state of science, you were there to listen to me complain. Without your support and advice, this journey would probably have had a different end. Furthermore, the gamenights with Tarik and the rest of the gang have helped to keep me somewhat sane these past years. Now, after this outburst of sentimentality, for the bottom line. I hold the following truths to be self-evident: 1) I was born stupid 2) I will not die hungry 3) Big up Bonsley! And as always, a happy new-year.

My dear Aaron, without a single doubt in my mind, you are the most honest and inherently kind person I have ever had the fortune of meeting. Working together was a pleasure, and always felt natural without either of us ever getting terribly frustrated with one another, I hope ;). You have been a chipper, constant force, who never judged, while always listening. I hope I can return that favour someday. To put it rather sentimentally, you just make the lives of everybody arround you better. I have greatly enjoyed and benefitted from your insights into philosophy, life and science. Furthermore, your willingness to share in your ideas and culture is something I will carry with me always.

My dear Esteban. Your sheer dedication and perseverance towards your goals, especially through tough times, has been an inspiration and something for which I have nothing but the utmost respect. You have inspired me to push myself further, for example in getting me to (almost) complete a 5k run, forcing me to learn how to parallel park or starting me on the road to ‘hablo un poquito Espanol’. I have loved every second of the rooftop Arepa lunches, game nights and milkshakes. I miss them greatly. Long may the sun shine!

Family

Aan mijn familie, Beppe, Pake, Carla, Felix, Martine, Oma en Rob, Jullie hebben mij altijd gesteund zonder iets terug te vragen, een feit dat ik zeer waardeer!

Florian, my dear brother. You have always inspired and stimulated me to get out of my comfort zone, especially in the early years in Groningen. I look back on those years we spent together with great fondness, and I do miss them dearly.

And finally, as I have been putting of the hardest and most important part to write to last, to my parents. Only the lord knows how many hours of my relentless rants you have patiently listened to over the years, a fact that surely deserves a medal ;) In all seriousness, you have always stimulated me to make the best out of myself, while letting me make my own choices. I have always felt supported by you in the choices I have made. Furthermore, I have learned so many valuable life lessons from you both over the years that have kept me going on this path, without which I surely would not have completed this thesis.