Context matters

de Vries, Dylan

DOI:
10.33612/diss.255266474

Context matters

The power of single-cell analyses in identifying context-dependent effects on gene expression in blood immune cells

Dylan H. de Vries

Cover design and lay-out: Publiss | www.publiss.nl
Print: Ridderprint | www.ridderprint.nl

rijksuniversiteit
groningen

Context matters

The power of single-cell analyses in identifying context-dependent effects on gene expression in blood immune cells

PhD thesis

To obtain the degree of PhD at the
University of Groningen
Rector Magnificus Prof. C. Wijmenga
And in accordance with
The decision by the College of Deans.

This thesis will be defended in public on
Wednesday 14 December 2022 at 9.00 hours

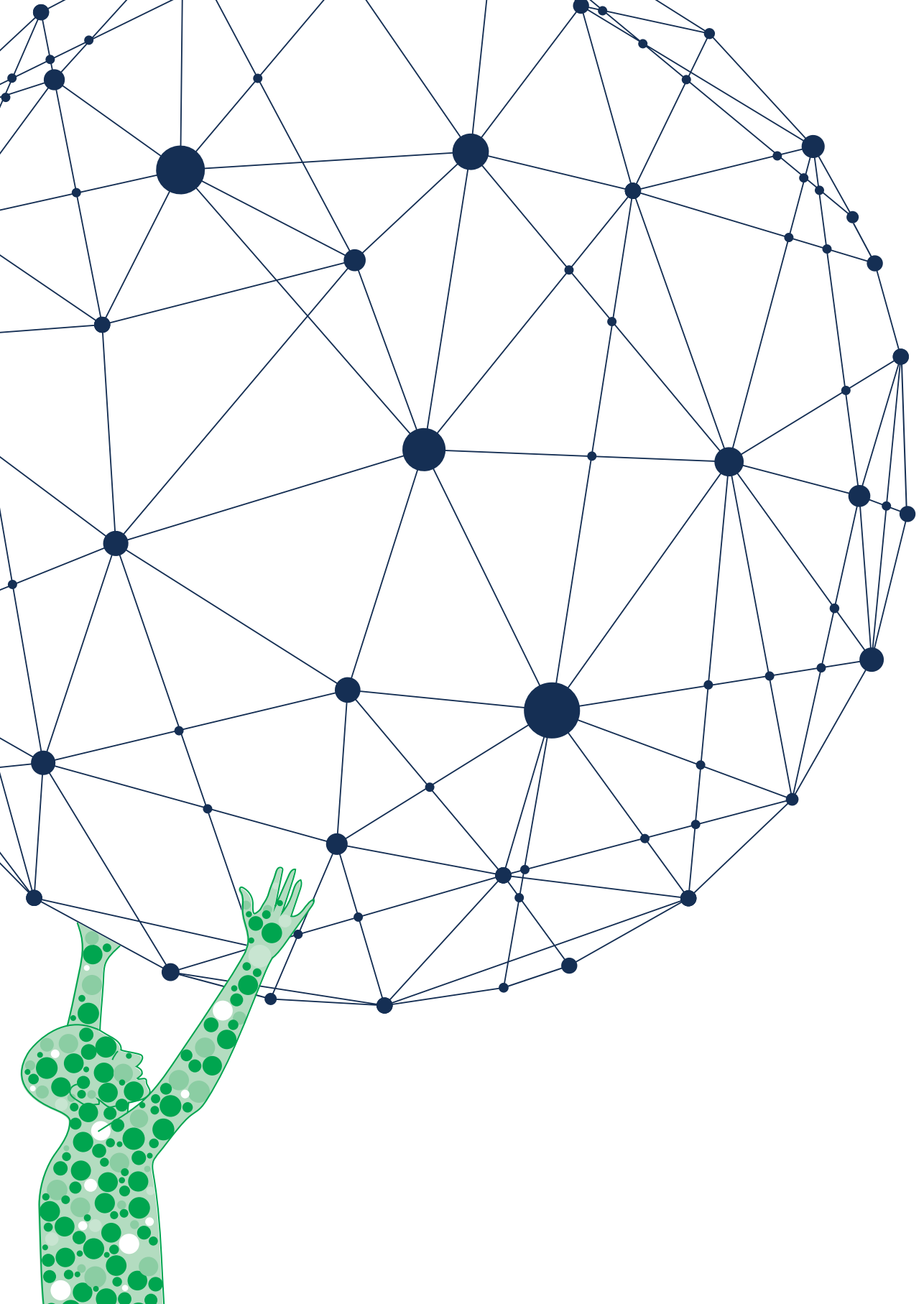
By

Dylan Hanraoi de Vries
Born on 27 May 1994
in
Gouda, the Netherlands

Supervisors
Prof. L.H. Franke
Prof. C. Wijmenga
Dr. M.G.P. van der Wijst

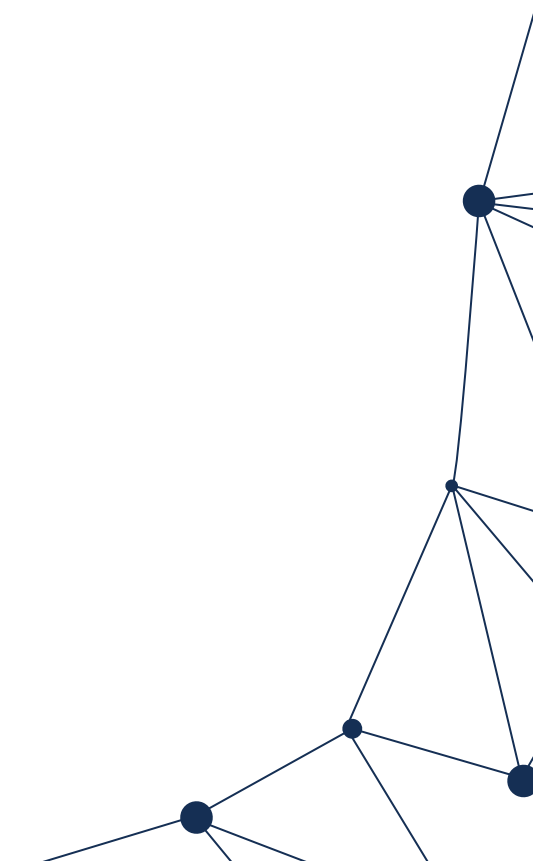
Assessment Committee
Dr. Ir. M.C. Nawijn
Prof. S. Aerts
Prof. B. Wertheim

Table of contents



CHAPTER 1

Introduction



The importance of genetic context

The idea that people would benefit from different medical treatments based on inherent biological differences is not new. In 1900, Karl Landsteiner identified the ABO blood group system and the need to give specific blood transfusions based on individual blood types, work for which he would go on to win a Nobel Prize[1]. But it was not until 1999 that the term ***personalized medicine*** was first introduced, in an article describing the need to target drugs based on genetic profiles[2]. In an article published in 2015, Nicholas Schork and his coauthors exemplified this need by showing that the top-ten highest-grossing drugs in the United States only help between 4%–25% of the people to whom they are prescribed (Fig 1.1)[3]. This situation is even worse for statins, a drug commonly used to lower cholesterol levels, where only 2% of people taking the medication appear to benefit[4]. Learning who will benefit from a given medication, and who will not, will lead to more effective treatment and a reduction in unnecessary side-effects. However, effectively implementing personalized medicine requires a better understanding of the sources of inter-individual heterogeneity in biological processes.

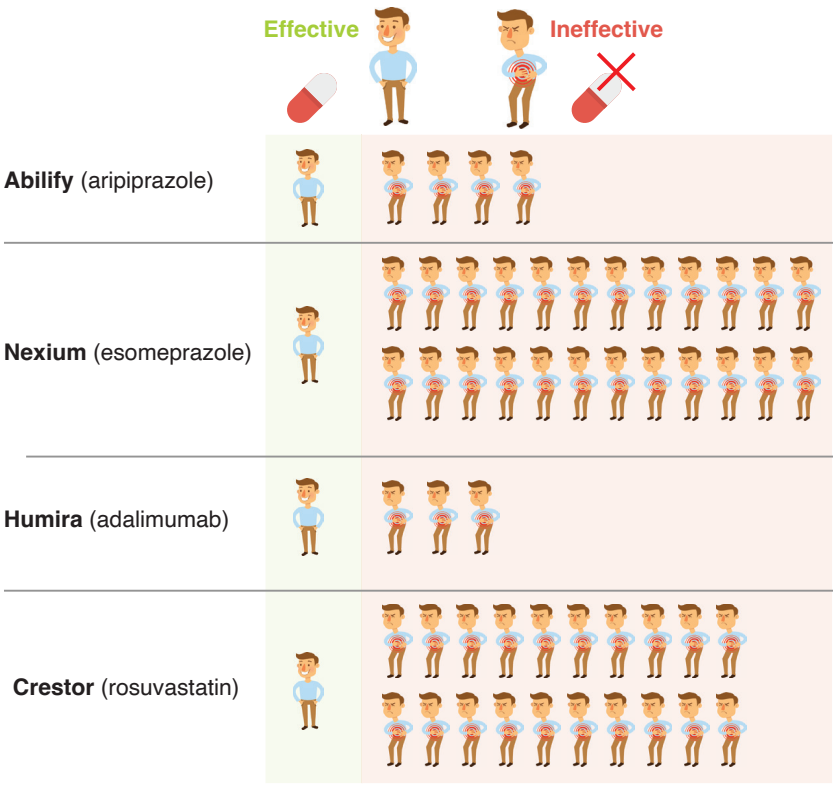


Figure 1.1*. The effectiveness of the top grossing drugs in the USA, showing how many patients are given ineffective drugs for every person with an effective response.**

In recent years, technological advances have made it feasible to study many more contexts to disentangle the factors that drive this inter-individual heterogeneity, with genetic factors playing a central role. Genetic heterogeneity plays a role in many drug responses and thousands of diseases[5]. In **monogenic diseases**, a single mutation can lead to the disease. In **complex diseases**, a combination of multiple genetic and environmental risk factors defines an individual's risk of getting the disease[6]. However, identifying which combinations of genetic risk factors and environmental factors lead to complex disease is incredibly difficult and requires data on many individuals due to the much smaller genetic **effect sizes** for complex diseases compared to monogenic diseases, even when combining the effects of all genetic risk factors for complex disease. Additionally, the ability to capture these genetic risk factors at a large scale has only become available fairly recently. It was not until 2001 that the Human Genome Project released a draft of the first fully sequenced human genome[7]. This draft drove efforts to systematically study common genetic variation and **linkage disequilibrium** (LD), which describes the pattern of inheritance of nearby variants, so that by knowing the genotype of one variant, those in high LD can be inferred. This information made it possible to design oligonucleotide arrays that profile a limited number of genetic variants but cover large parts of the genome and the most common human genetic variation. With the availability of the first beta-versions of these arrays, the first genome-wide association study (**GWAS**) was performed in 2005[8]. The study used genetic information of 96 patients with age-related macular degeneration and 50 healthy controls to identify a single nucleotide polymorphism (**SNP**) that significantly increases the risk for developing this disease. Since then, thousands of GWASs have been performed on a wide variety of traits, and studies now often include hundreds of thousands and even millions of individuals to identify low frequency and small effect-size genetic variants[9]. However, knowing that a SNP is associated to a disease provides little insight into how it may lead to disease. Various strategies exist to unravel this black box between genetic association and disease. One such strategy is to link the SNP genotypes to changes in gene expression, also known as expression quantitative trait locus (**eQTL**) mapping. When a SNP affects the gene expression of a gene in close proximity, the SNP is called a ***cis*-eQTL**, while distal effects are called ***trans*-eQTLs** (Fig. 1.2). Generally, eQTLs have stronger effect sizes the closer they are to transcription start sites, meaning *cis*-eQTLs often have stronger effect sizes than *trans*-eQTLs[10]. Low effect sizes, combined with the much larger number of tests required in *trans*-eQTL analyses, mean that very large sample-sizes are needed in order to yield sufficient statistical power to identify *trans*-eQTLs. This is now becoming possible through **meta-analyses** on multiple eQTL datasets[10,11].

**Figure adapted from Schork et al.[3] using images from Macrovector.*

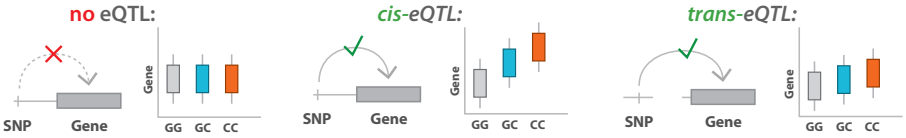


***Figure 1.2**. The effects of SNPs on gene expression when there is no eQTL, a cis-eQTL, or a trans-eQTL. For cis-eQTLs, the SNP is in close genomic proximity to the gene, whereas trans-eQTLs are distal effects where the SNP is far from the gene or even on another chromosome.*

The majority of eQTL studies have been done using **bulk** RNA-seq data, i.e. measuring mRNA expression levels of all the cells in the tissue together without a way to trace expression back to the cell of origin. Generation of bulk RNA-seq data is relatively inexpensive compared to other techniques that allow for such **cell type–specific resolution**, and therefore, often contain much larger cohort sizes[12]. However, such studies often miss **context-specific effects**. Current studies show that as little as 11% ± 2% of disease heritability is mediated by known *cis*-eQTLs[13]. We hypothesize that the true mediation is higher, but that the right *cis*-eQTLs have not yet been identified due to the missing cell type- and context-specific effects. In a large effort by the Genotype-Tissue Expression project, eQTLs were studied across 44 human tissues using 7,051 samples from 449 donors to learn about the importance of tissue-specificity[14]. Despite the wide variety of tissues, only 61.5% of SNPs within significantly associated GWAS regions, or **loci**, had an eQTL effect. However, some of these effects may still be hidden due to other contextual information, such as cell type and environmental factors. As **single-cell technologies** such as single-cell RNA sequencing (**scRNA-seq**) are maturing, they are increasingly being used to map high-resolution context-specific eQTL effects.

Single-cell analyses to disentangle context-specificity

Single-cell sequencing was first introduced by Tang et al in 2009 in a study where they sequenced the RNA of a single mouse cell[15]. Since then, single-cell technology has advanced dramatically and now permit the sequencing of hundreds of thousands of cells per run, although the technique remains more expensive than bulk tissue alternatives[12,16]. Many different single-cell sequencing strategies exist. The most commonly used techniques, which also allows users to sequence most cells, are **droplet-based techniques**[17]**.** In future studies, these will likely be combined with combinatorial indexing to improve the throughput[18]. The droplet-based technique used throughout this thesis uses a **microfluidic chip** with oil, reagents and gel beads with primers and both bead-specific **barcodes** and barcodes specific to the **lane** of the chip (Figure 1.3)[19]. Cells are isolated into a droplet of oil, together with the reagents and the barcoded gel beads. Each mRNA molecule, or **read**, receives a **cell barcode** and a unique molecular identifier (**UMI**) before being transcribed into cDNA, amplified and finally sequenced together with all other reads. The amplification increases the

odds of capturing the reads, which can be correctly counted using the UMIs and assigned to the original cell using the cell barcode[20]. Unlike some other techniques, this technique does not sequence the entire read, but instead sequences from the **3' end** up to a fixed number of nucleotides to still be able to map the read to a gene.

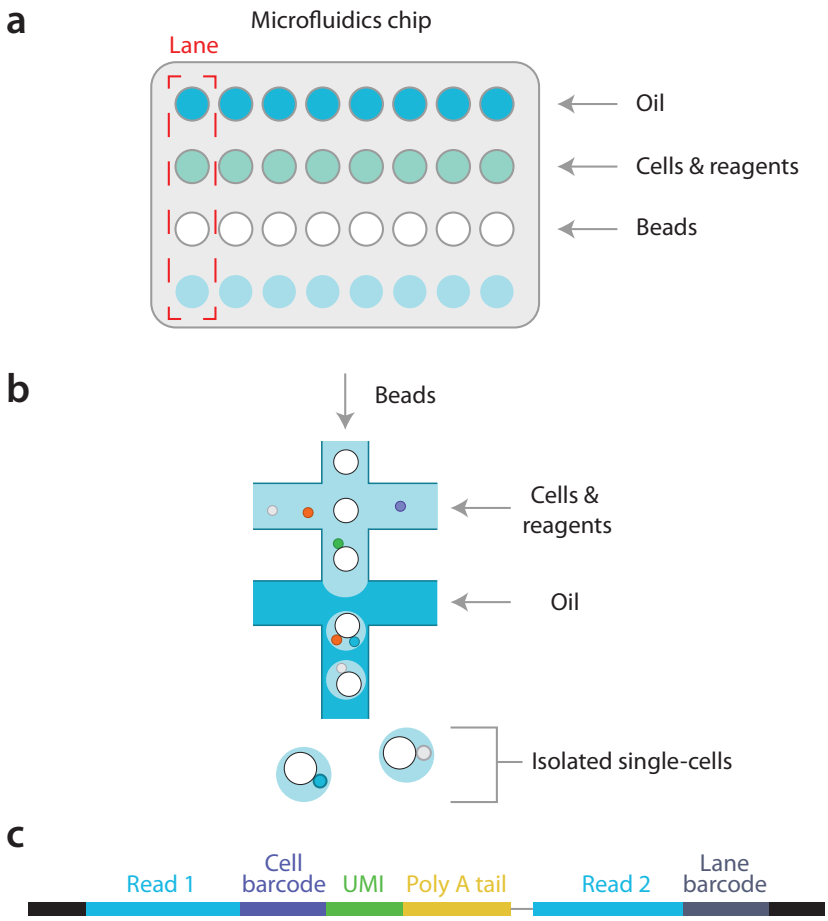


***Figure 1.3**. Overview of how the single-cell RNA-seq method we used works. **a)** The microfluidics chip with eight lanes, each with four wells. **b)** The process to acquire single cells. **c)** The read structure to be sequenced.*

In addition to mapping reads to genes and cells, all the reads of a cell together can also be used to map the cell back to an individual donor using **genetic demultiplexing**[21]. When sequence reads cover a genomic region containing a common variant, this information can be used to link the cell back to an individual. While a single read would not allow us to distinguish between individuals, many such

reads together can. Using genetic demultiplexing, it is possible to pool cells from multiple individuals in a single experiment, greatly reducing the reagent costs per individual while also reducing technical batch effects.

Given the large quantities of data being generated and advances in computational capacity, it is now possible use these massive datasets to map the underlying biological interactions. One way to study how transcriptional regulation works is to reconstruct gene regulatory networks (**GRNs**) that map the transcriptional relationships between genes. Such GRNs typically consist of genes (**nodes**) and co-expression relations between genes (**edges**) and can provide insight into how genes affect specific biological processes within cells. Knowledge of these GRN, in combination with information on how individual genetic variants affect the gene within these networks (e.g. through eQTL mapping), is likely to reveal new insights into how the genetic variants associated in GWAS eventually cause disease, and this knowledge can be used in drug development. Various consortia have been initiated to generate such data in a wide variety of cellular contexts, including the Human Cell Atlas[22], LifeTime[23] consortium and the eQTLGen[10] consortium. This work also opens up many avenues for the development of new tools and strategies to utilize the vast amounts of data coming from these consortia.

Aim of the thesis

The aim of this thesis is to learn more about the impact of genetic and environmental variation on gene expression and their interactions in blood, using the latest single-cell technologies and computational methods, with the ultimate goal of identifying routes for the realization of personalized medicine approaches. We show how scRNA-seq can be used to study how genetic variation and the environment affect gene expression. In later chapters of this thesis, we delve into how genetic and environmental factors interact and affect gene expression and co-expression and eventually describe the single-cell eQTLGen consortium, which will work on studying such interactions in the future.

Outline of the thesis

In **Chapter 2**, we outline how to integrate bulk and single-cell–derived data to reconstruct GRNs and hypothesize how context-specific, personalized GRNs may help to identify key driver genes for disease and for drug targeting.

In **Chapter 3**, we perform one of the first single-cell eQTL studies in which we identified cell type–specific eQTLs for blood immune cell types. We show how cell type–specific effects can be masked in a bulk-like analysis due to small effect sizes or through opposing eQTL effects in other cell types. In addition, we introduce the concept of co-expression QTLs, where the co-expression relationship between pairs of genes is dependent on genetic variation.

In **Chapter 4**, we look at the effects of fungal exposure (*Candida albicans*) on blood immune cells. We used bulk tissue–based approaches to ensure sufficient power to detect pathogen-specific effects, then single-cell data to pinpoint the specific cell types in which these effects occur. Using this approach, we were able to identify the likely mechanism through which the LY86 gene affects the response to Candida infection of the blood.

In **Chapter 5**, we use three different pathogen stimulations at two time points to better capture the cell type–specific and pathogen-dependent responses to infection. We identified the expression changes upon stimulation and saw that the pathogen-specificity was lower than the timepoint-specificity for both differential expression and context-specific eQTLs. Similarly, co-expression QTLs often showed timepoint-specificity. Using a combination of transcription factor binding information and pathway enrichment for the co-expressed genes, we identify the biological mechanism through which pathogen-stimulation affects the co-expression between genes.

In **Chapter 6**, we assess strategies to reconstruct gene regulatory networks derived from scRNA-seq data and quantify the impact of several challenges associated with reconstructing personalized gene regulatory networks. We further develop our co-expression QTL capturing strategy to systematically identify such events and provide guidelines on how co-expression QTL mapping can be performed in other datasets. With this method, we were able to identify a potential working mechanism for how a genetic variant affects the manifestation of autoimmune diseases.

In **Chapter 7**, the plans of the single-cell eQTLGen Consortium are outlined. As a continuation of the eQTLGen consortium's work, this new consortium aims to provide a comprehensive, cell type–specific resource of eQTLs in blood. We present how we want to harmonize cell type classification and perform a federated eQTL analysis and how to use to reconstruct GRNs.

Finally, in **Chapter 8**, I describe how new developments in single-cell technologies can be used to further study context-specificity and how I envision the use of these technologies for precision medicine.

References

1. Farhud, D. D. & Yeganeh, M. Z. A brief history of human blood groups. *Iran. J. Public Health* **42**, 1–6 (2013).
2. Jørgensen, J. T. Twenty Years with Personalized Medicine: Past, Present, and Future of Individualized Pharmacotherapy. *Oncologist* **24**, e432 (2019).
3. Schork, N. J. Personalized medicine: Time for one-person trials. *Nature* **520**, 609–611 (2015).
4. Mukherjee, D. & Topol, E. J. Pharmacogenomics in cardiovascular diseases. *Prog. Cardiovasc. Dis.* **44**, 479–498 (2002).
5. 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).
6. Huang, Q. Genetic Study of Complex Diseases in the Post-GWAS Era. *Journal of Genetics and Genomics* **42**, 87–98 (2015).
7. Craig Venter, J. *et al.* The sequence of the human genome. *Science (80-.).* **291**, 1304–1351 (2001).
8. Edwards, A. O. *et al.* Complement factor H polymorphism and age-related macular degeneration. *Science (80-.).* **308**, 421–424 (2005).
9. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications* **11**, 1–3 (2020).
10. 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. 2021 539* **53**, 1300–1310 (2021).
11. Klein, N. de *et al.* Brain expression quantitative trait locus and network analysis reveals downstream effects and putative drivers for brain-related diseases. *bioRxiv* 2021.03.01.433439 (2021). doi:10.1101/2021.03.01.433439
12. Kuksin, M. *et al.* Applications of single-cell and bulk RNA sequencing in onco-immunology. *Eur. J. Cancer* **149**, 193–210 (2021).
13. 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).
14. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).
15. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
16. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* **13**, 599–604 (2018).
17. Lee, J., Young Hyeon, D. & Hwang, D. Single-cell multiomics: technologies and data analysis methods. *Exp. Mol. Med.* **52**, 1428–1442 (2020).
18. Mulqueen, R. M. *et al.* High-content single-cell combinatorial indexing. *Nat. Biotechnol. 2021* 1–7 (2021). doi:10.1038/s41587-021-00962-z
19. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 1–12 (2017).
20. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
21. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
22. Regev, A. *et al.* The human cell atlas. *Elife* **6**, (2017).
23. Rajewsky, N. *et al.* LifeTime and improving European healthcare through cell-based interceptive medicine. *Nat. 2020 5877834* **587**, 377–386 (2020).

CHAPTER 2

An integrative approach for creating personalized gene regulatory networks for precision medicine

Monique G.P. van der Wijst[#], Dylan H. de Vries[#], Harm Brugge, Harm-Jan Westra, Lude Franke
[#]Both authors contributed equally

Published in Genome Medicine. https://doi.org/10.1186/s13073-018-0608-4

Abstract

Only a small fraction of patients actually respond to the drug prescribed to treat their disease, which means most are risking unnecessary exposure to side-effects through ineffective drugs. This inter-individual variation in drug response is driven by differences in gene interactions caused by each individual's genetic background, environmental exposures, and distribution of specific cell types. These gene interactions can now be captured by building gene networks. In this perspective, we propose an integrative approach that leverages the sensitivity of bulk data with recent advancements in single-cell data to enable the reconstruction of personalized, cell-type- and context-specific networks. We expect this approach will allow us to prioritize key driver genes for specific diseases, knowledge that will open new avenues towards better personalized healthcare.

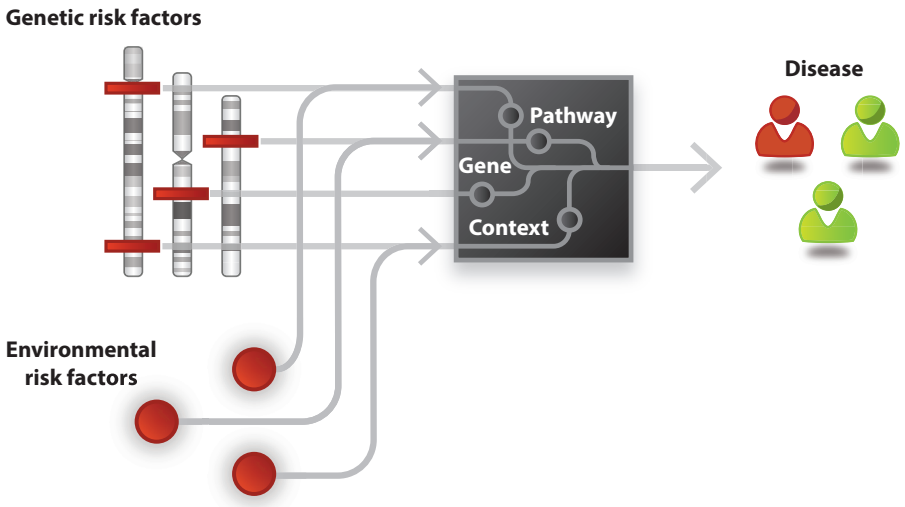


Figure 1 | link between genetic and environmental factors on disease. Revealing the interplay between genetic and environmental risk factors enables identification of the disease-associated context, causal genes and pathways. This leads to a better understanding of why certain individuals become ill, whereas others do not.

1. Unravelling the link between genetic risk factors and disease

In the last decade, genome-wide association studies (GWAS) have identified over 10,000 genetic risk factors, mainly single nucleotide polymorphisms (SNPs), for more than 100 common diseases[1]. Together these GWAS loci can explain up to ~25% of the heritability of complex diseases[2] and up to 56% of disease-related traits[3]. The majority of these genetic risk factors are located in non-coding regions[4] and, as the function of these regions is challenging to decipher, it remains largely unclear how the SNPs are linked to disease. Several studies have shown that the gene nearest to the genetic association may not always be the causal gene[5-7]. In consequence, more sophisticated approaches have been developed to unravel the link between genetic risk factors and disease (e.g. causal genes, pathways and cell types) (Fig. 1). Expression quantitative trait loci (eQTL) studies, for example, have been performed to identify the local (*cis*-eQTL) and distal (*trans*-eQTL) downstream effects of genetic variation on gene expression[8, 9]. These eQTL studies have provided the first clues about how genetic variation is linked to disease (Fig. 2). Other methods to further prioritize putatively causal genes include co-localization analysis, fine-mapping and summary-data-based Mendelian randomization (for detailed discussions of these techniques see [10, 11]). To provide a greater understanding of gene regulatory mechanisms, several large consortia—including the ENCODE project[12], FANTOM[13], Epigenome Roadmap[14] and Blueprint[15]—have systematically classified more than 80% of the genome as non-coding regulatory elements. Genetic variation has now been linked to many of these elements, including epigenetic marks[16, 17], transcription factor binding and chromatin accessibility[18,19] and post-transcriptional regulation[20, 21].

What all the above-mentioned studies make clear is the importance of studying both gene expression[22] and its regulation. However, despite all these advances in our understanding of GWAS variants, a recent study of 7,051 samples from 449 donors across 44 tissues from the Genotype-Tissue Expression (GTEx) project could still only link 61.5% of the SNPs within a GWAS locus to an eQTL effect[23]. The reason that not all GWAS SNPs can be linked to an eQTL effect could be that eQTL studies have been performed in the wrong context for a specific disease. We know now that many genetic risk factors have cell-type-specific effects[22, 24, 25] or are modulated by environmental factors[26, 27] and these are contexts that eQTL studies usually do not completely capture.

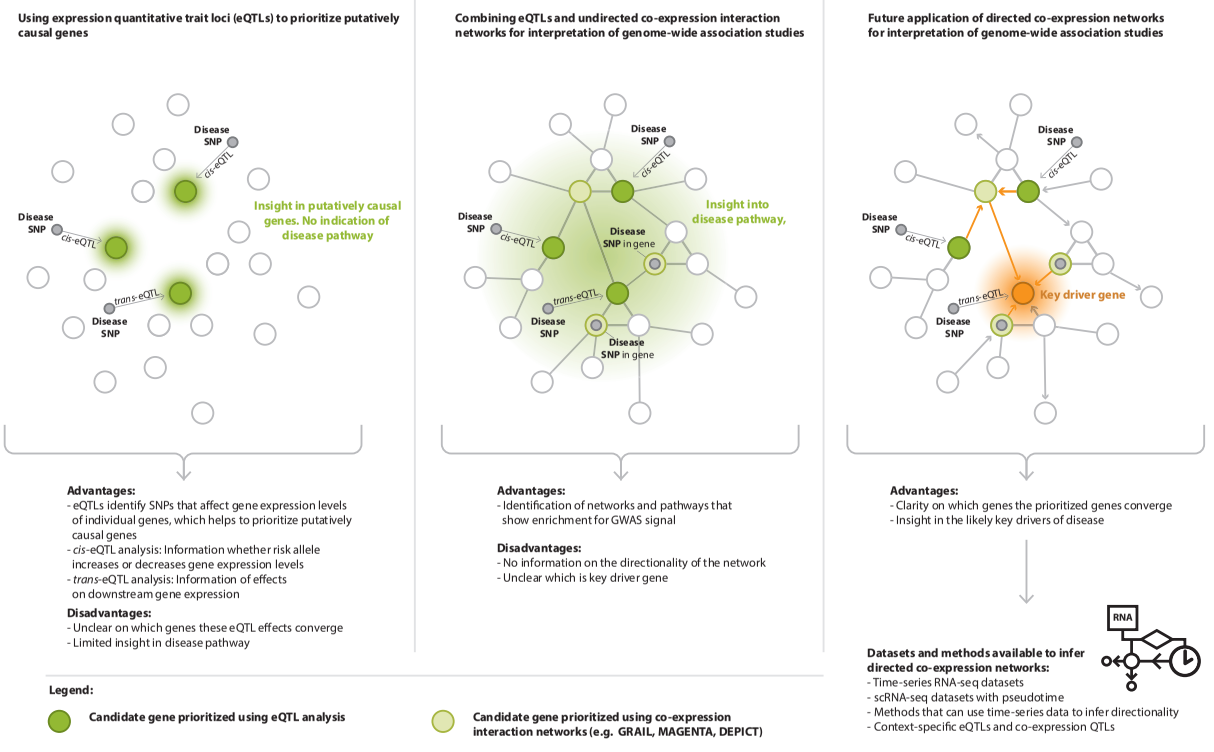


Figure 2 | Current and future approaches shedding light on the link between genetics and disease. To identify putatively causal genes, GWAS SNPs have been linked to gene expression using eQTL analysis (left). To obtain greater understanding of the disease pathogenesis, it is essential to look beyond the disruption of individual genes and identify potential disease-associated pathways. This can be done by identifying the co-expression relations between genes in all loci linked to a specific disease, for example using methods such as GRAIL, MAGENTA and DEPICT (middle). In the future, to pinpoint disease-relevant key driver genes, directional co-expression networks can be generated using a combination of current and novel approaches, including pseudotemporal ordering of scRNA-seq data, and context-dependent eQTL and co-expression QTL analysis (right).

Independent genetic risk factors can converge into key regulatory pathways[24, 28] and may act beyond the disruption of individual genes[29, 30]. Therefore, we expect that a comprehensive overview of the many processes at work will be required to better understand disease pathogenesis. This kind of overview can be acquired by reconstructing gene regulatory networks (GRNs), which are uniquely based on cell type[22, 24, 25], environment[26, 27] and an individual's genetic makeup[29, 30]. Understanding the effect of genetic variation on GRNs is particularly important because this may contribute to the large inter-individual variation in drug responsiveness (Fig. 3); currently some of the most commonly prescribed drugs are only effective for 4% to 25% of the people to whom they are prescribed[31].

In this perspective, we outline our vision for an integrative approach to reconstructing context-specific GRNs. We focus here on gene-expression-based regulatory networks because gene expression data has been extensively collected and because the generation of this type of data at the bulk and single-cell level has advanced the most compared to other single-cell technologies. However, there are other molecular levels, such as metabolites or proteins, that can and should be added to GRNs in the future to capture the full complexity of a disease[32].

We begin this perspective by briefly introducing the concept of a co-expression network and describing the methods to create directional GRNs from co-expression networks using bulk data. We then discuss the limitations of bulk data and how they can be resolved by the unique properties of novel single-cell gene expression approaches to enable the reconstruction of causal GRNs. We do not cover the methods used to reconstruct single-cell GRNs in detail because they have recently been covered in an excellent review by Fiers *et al.*[33]. We conclude by describing how combining bulk and single-cell data permits reconstruction of context-specific, personalized GRNs, and describe their use in advancing personalized healthcare.

2. Gene networks in bulk data

Understanding the pathways affected in disease requires a clear definition of which genes act together in specific cellular processes. To this end, model organisms have been instrumental in defining the most basic pathways present in each cell. By performing knockout experiments, for instance, the relationships between genes can be identified by studying the downstream effects on gene expression or enzymatic function, and these effects are now catalogued in databases such as KEGG[34] and REACTOME[35]. The pathways defined in these databases, however, can be incomplete or biased towards well-studied cellular phenotypes or genes. Co-expression networks and GRNs can therefore be useful for extending the knowledge provided by such databases, and methods for constructing such networks have been reviewed in detail elsewhere[36, 37].

Gene networks can be used to infer the functions of genes by assuming that genes with similar functions are located close together in these networks (Fig. 2).

For example, known pathway annotations can be combined with network properties, such as the number of connections between nodes (which effectively define genes) or distances between individual nodes. This enables the identification of clusters of genes that are enriched in a particular pathway, and unannotated genes can be assigned to the overrepresented pathways in the cluster (i.e. guilt-by-association)[38-42]. A recent comparison of methods for the identification of disease-relevant gene clusters in a co-expression network showed that while a number of approaches were successful, the challenge was best tackled using a consensus of several predictors[43].

Implementing directionality in the gene network

While disease-relevant gene clusters can be identified using the methods discussed above, they do not provide insight into how genetic risk factors affect the network. To identify the downstream consequences of genetic risk factors, directionality must be added to co-expression networks. A directional co-expression network that also has information about the regulators that control gene expression is a GRN. Information obtained from databases such as KEGG[34] can be used to add directionality to specific pathways, but they are limited in their ability to determine whether these pathways are active in specific cell types or if they function similarly in all cells or individuals. Additional approaches are therefore required to generate context-specific GRNs.

Directionality can be added to a co-expression network using a combination of perturbations, time-series data and Bayesian models[44-46] (Fig. 2). Bayesian models can be divided into static and dynamic models. Static Bayesian models cannot implement feedback loops and are thus restricted in the type of models they can generate. Since feedback loops are a common property of GRNs[47], dynamic Bayesian models that incorporate time-series data and allow feedback loops are more suitable[44, 45]. However, generating time-series data is very costly because it requires a high sampling-rate to correctly define directional relationships between genes (follows from the Nyguist-Shanning sampling theorem[48, 49]). Undersampling could result in incorrect correlations between genes or in missing key events in the network[50]. Moreover, correct directional inference requires cells to be synchronized before time-series experiments are started, and even when successful, cells may lose their synchronization over time[51]. Finally, constructing Bayesian networks is computationally intensive[52]. This burden can be reduced by including prior knowledge from other sources (e.g. pathway databases), but the quality of the resulting network will be determined by the reliability of that prior knowledge[53, 54].

Information about the regulators that control gene expression can also be gained by linking GWAS variants to additional molecular layers such as transcription factor binding motifs, and enhancer and promoter elements[55]. This information can be helpful in determining directionality and understanding how genes are being regulated, providing additional support for putatively causal interactions. Similarly, it has been

previously shown that eQTL information can be linked to additional molecular layers to reveal the functional mechanism of how the genotype affects interactions between genes, so-called context-dependent eQTLs[29]. The directionality of these interactions can be exposed by understanding the underlying regulatory mechanism. The *cis*-regulatory effect of SNP rs968567 on *FADS2* expression, for example, is modified by the sterol regulatory element-binding transcription factor *SREBF2* (Fig. 4)[29]. Using ENCODE ChIP-seq data, a binding site of *SREBF2* near rs968567 was identified, which suggested that the binding of *SREBF2* depends on the allele at rs968567 and consequently modulates the downstream expression effect on *FADS2*.

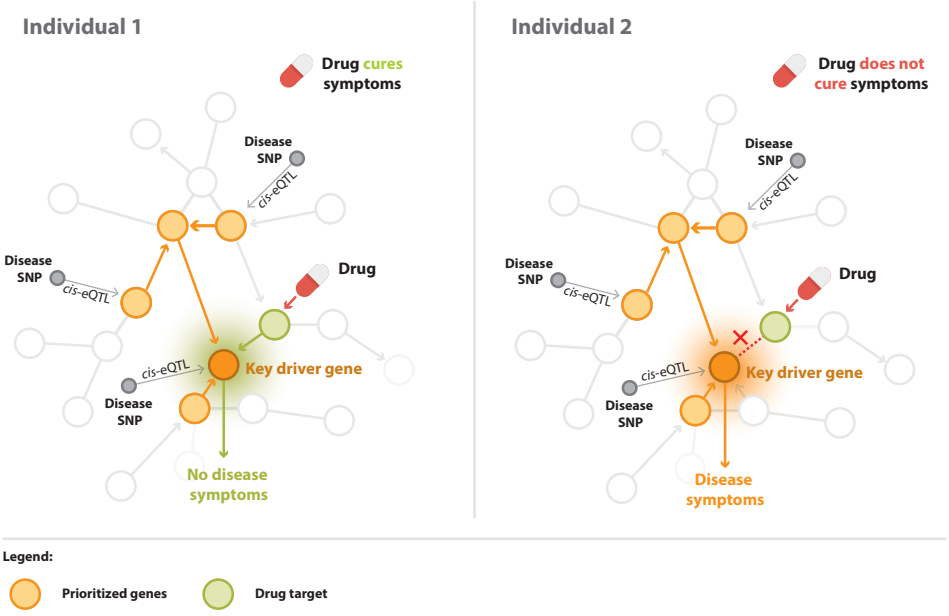


Figure 3 | Implications of personalized GRNs on precision medicine. Depending on an individual's regulatory wiring, specific drugs may or may not be effective. Personalized GRNs will provide guidance for precision medicine in the future. In this particular example, GRNs of two individuals are shown in which the regulatory wiring between the drug target gene and the key driver gene is different: in individual 1, the drug target gene activates the key driver gene, whereas in individual 2, the interaction between both genes is absent. Thus, in individual 1, the drug is effective, whereas in individual 2, the drug is ineffective.

While additional molecular data layers can be very informative for inferring directionality, they are not always available in the disease-relevant context. Recent bulk-based RNA-seq studies have generated tissue-specific co-expression networks for up to 144 different tissues[56, 57]. However, the associated time and cost of implementing directionality and context-specificity have hampered the generation of tissue-specific GRNs in bulk data. In the following sections, we describe how a combination of bulk and single-cell data can be used to resolve these issues and to create GRNs that help us understand the link between genetic risk factors and disease.

3. Improving networks with single-cell data

The first scRNA-seq experiment was performed with a single cell in 2009[58]. Since then, the technique has developed to the point where more than a hundred thousand cells can be processed in parallel[59, 60]. Initial efforts have used single-cell qRT-PCR data to build co-expression networks consisting of a limited number of genes involved in hematopoiesis[61]. More recently, efforts have been made to build gene co-expression networks using scRNA-seq data[62-64]. The reliability of these networks will improve with increasing numbers of cells, reads-per-gene and genes-per-cell, but exact numbers are difficult to provide as they are influenced by many factors[63, 64]. Experimentally-induced perturbations using CRISPR-based approaches coupled to scRNA-seq show that using 100-200 cells with 1000 unique reads per cell is sufficient to detect effects on individual genes[65]. Assuming that such perturbations reflect natural perturbations, comparable numbers of cells and reads are expected to be required to reconstruct a co-expression network.

Most aspects of reconstructing a co-expression network will not differ between single-cell and bulk expression data (see review in [33]). However, single-cell-based co-expression networks will be enhanced when the unique features of single-cell data are taken into account. For example, network inference can be improved by extending the model with transcriptional dynamics revealed by single-cell data[62]. Additionally, the mRNA capture efficiency is lower in a single cell sample compared to a bulk sample[66]. Therefore, especially in single-cell data, being able to confidently impute gene expression levels or increase capture efficiency will increase the power to detect co-expression relationships[63], and we discuss several of these aspects in the following sections.

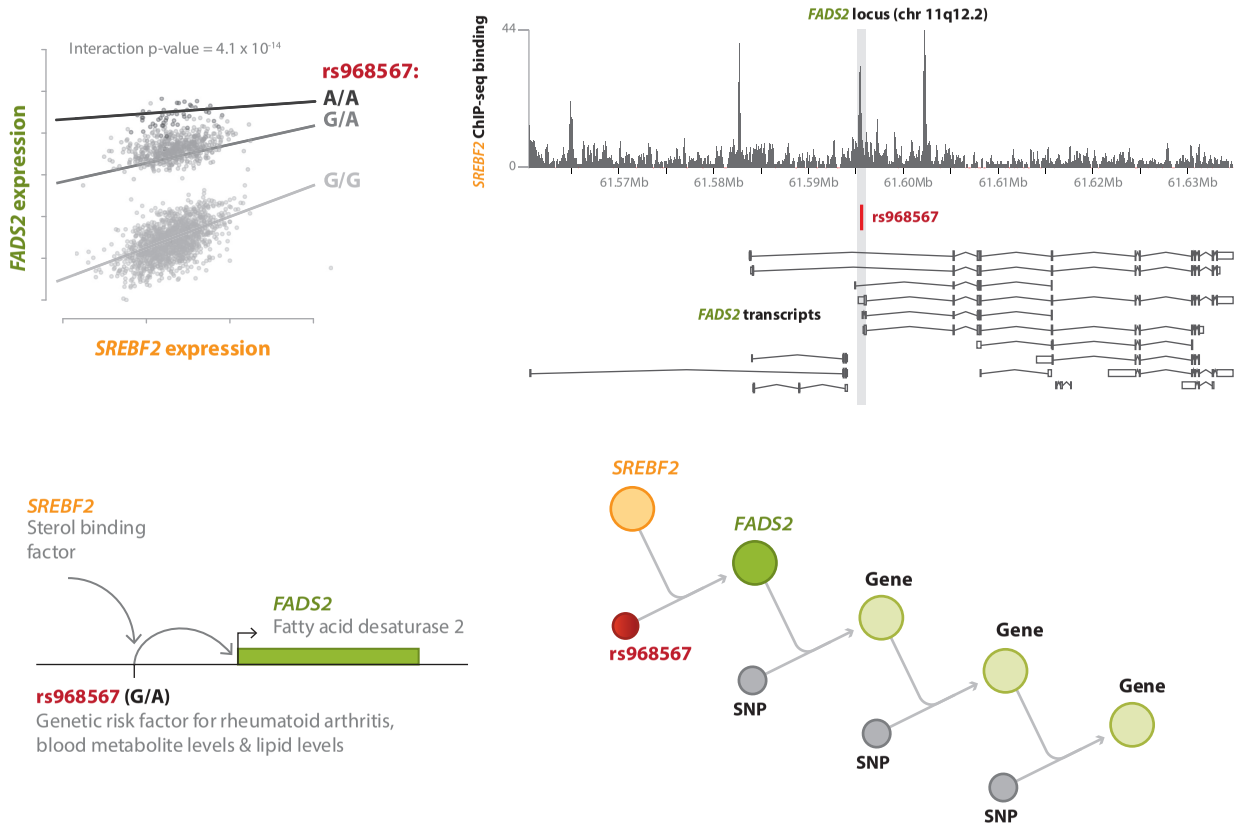


Figure 4 | Reconstruction of a gene regulatory network using eQTLs. SNP rs968567 regulates the interaction between the *cis*-regulated eQTL gene *FADS2* and the sterol binding transcription factor *SREBF2* (context-dependent eQTL) (top left). ENCODE ChIP-seq data shows that this SNP is located within a *SREBF2* binding site, thereby modulating *FADS2* gene expression (top right). Combining *cis*-, *trans*- and context-dependent eQTLs or co-expression QTLs has the potential to allow reconstruction of a branch of a gene regulatory network (bottom).

Specifying the context

Gene expression networks change depending on a number of factors, including cell type[22, 24, 25], environment[26, 27] and genetic signature[29, 30], and the influence of each of these contexts can be determined using scRNA-seq.

The ability of scRNA-seq data to dissect complex tissues and detect cell types/states in an unbiased manner[67-69] is valuable for reconstructing cell-type-specific co-expression networks. This kind of dissection was recently applied to detect single-cell eQTLs with high resolution[30, 70], and this analysis revealed that gene regulation can differ even between subcell types[30]. Unbiased classification has also led to the identification of specific cell states and combinations of transcription factors that drive cell-type-specific gene regulatory programs[71]. That study showed that networks are different between brain cell types and that cell-type classification using networks gives better separation than classification on gene expression levels alone.

Cellular heterogeneity induced by environmental perturbations can also be dissected using single-cell analysis[72]. In the context of co-expression networks, Martins *et al*. used single-cell qRT-PCR to identify the heterogeneous effects of cytokine stimulations on the rewiring of the network in macrophages[73]. Importantly, some of the effects on the co-expression network they identified would have been overlooked if they had pooled the expression of 10 cells, a demonstration of how population-level co-expression networks cannot fully capture gene regulation at the single-cell level.

scRNA-seq can also be used to identify differences induced by genetic variation between individuals, which enables the reconstruction of a person-specific or personalized co-expression network. It has recently been shown that, unlike with bulk RNA-seq, it is feasible to generate many measurements per individual with scRNA-seq, which enables the calculation of correlations between genes per individual[30]. These correlations can be used to identify the relationships between genes within a personal co-expression network. This approach was recently applied to identify relations between genetic variants and the modulation of co-expression in CD4+ T cells. Within a cohort of 45 individuals, genetically modulated co-expression relationships, so-called co-expression QTLs, were identified that could be replicated in a bulk RNA-seq dataset of 2,116 individuals. However, these relationships would not have been detected using a genome-wide approach in bulk data only[30]. Another advantage of scRNA-seq data is that true correlations between genes can be identified that would otherwise be masked by the effects of averaging in bulk RNA-seq data due to Simpson's paradox[74].

However, a disease-specific network is not defined by any of the above-mentioned factors alone, but rather by a combination of them. Celiac disease, as an example, occurs only in individuals who both carry specific HLA genotypes (genetics) and consume foods containing gluten (environment)[75]. Celiac disease is a well-known example of what is called a Genotype by Environment (GxE) interaction, where an

environmental perturbation is modulated by an individual's genetic background. Future scRNA-seq studies should expand our understanding of how GxE interactions modulate the co-expression network. For example, by exposing cells from different individuals to various environmental conditions.

Exploiting scRNA-seq data variability to infer directionality

Measured gene expression levels can vary considerably between different cells even when accounting for cell type, environment and genotype. This is the result of biological and technical variation, with the former affected by processes such as stochastic bursting of transcription[76] and the cell cycle stage, while the latter is inherently associated with the challenges of accurately measuring single-cell gene expression levels [77]. Although large variability in gene expression levels is usually seen as a problem, it can actually provide unique insights that cannot easily be extracted from bulk expression data. During dynamic processes, such as cell differentiation or a response to an environmental stimulus, cells will move towards another stage over time. However, the pace at which cells move into this new stage differs between cells. By exploiting the asynchronous nature of these processes between cells, cells can be computationally ordered in pseudotime based on expression similarity[78, 79]. This pseudotemporal ordering of cells can provide temporal resolution in an experiment that does not explicitly capture cells along a time-series. Insights can therefore be gained using scRNA-seq data that would remain hidden in bulk data, while requiring only one snapshot from a single sample (Fig. 5). Currently, there are more than 50 different methods for pseudotemporal ordering cells (see [80] for a recent comparison of these).

Previously, pseudotime has been used to reconstruct co-expression networks[81, 82] or small directional GRNs[83] from single-cell data[33]. However, the assumptions required for pseudotemporal ordering of cells are often violated in practice, which can result in incorrect assignment of directionality[84, 85]. The sampling frequency inferred by these methods, for instance, depends on sample size, which could be insufficient to recreate the complete underlying process of interest. Further, several different networks may give plausible explanations for the same observed distribution of cell states. Therefore, it is difficult to determine the correct underlying mechanism of gene regulation without prior knowledge.

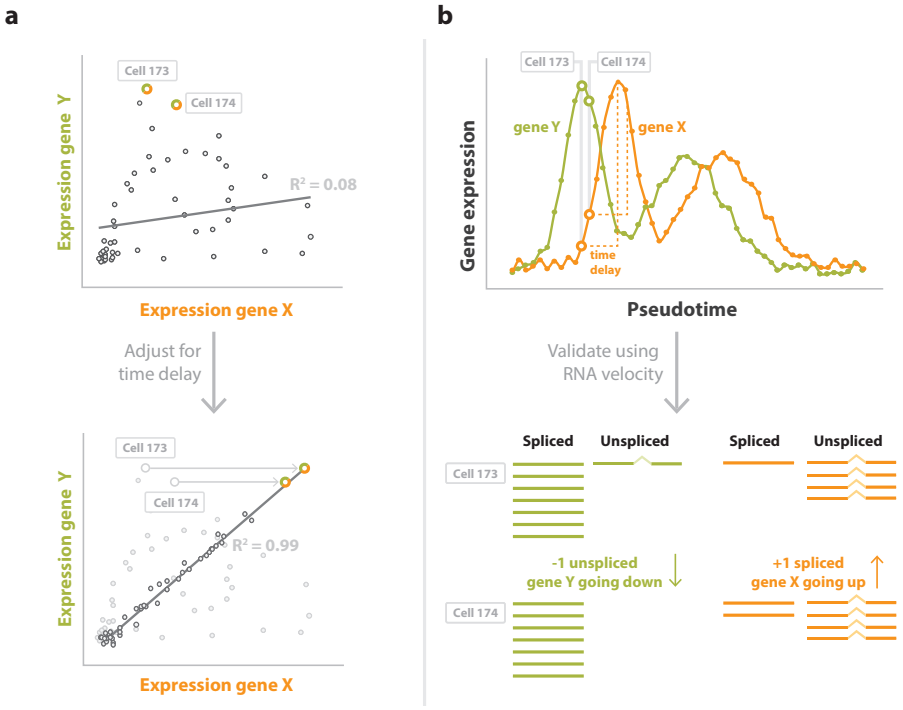


Figure 5 | Inferring causality using pseudotime analysis and RNA velocity. (**a**) In this example, when determining the relation between gene X and gene Y, no correlation would be observed (top). However, the relationship between both genes may be masked by a time delay and correcting for this time delay might reveal a clear correlation between the expression of gene X and gene Y (bottom). (**b**) To identify the length of a time delay, the cells can be ordered along pseudotime, i.e. an inferred timeline based on the variable gene expression states of single cells measured at a single moment in time (top). RNA velocity, a method that exploits the unidirectional character of splicing, allows the prediction of the future state of a cell and its genes (bottom). As such, the correct ordering of cells can be validated using RNA velocity. When plotting gene expression against pseudotime, it will become clear that the expression of gene X is following the expression of gene Y. From this, it can be deduced that gene Y is regulating gene X, and not the other way around.

Both these issues can be resolved using a method called RNA velocity[86] that exploits the unidirectional character of splicing. RNA velocity examines the rate of change of mRNA molecule abundances in the cell by modeling the relationship between unspliced mRNA (an indicator of current transcription) and spliced mRNA (an indicator of transcription in the recent past). By taking the RNA velocity information of all genes together, a cell's future state can be successfully predicted[86]. Moreover, RNA velocity artificially enriches the inferred sampling frequency and prioritizes the pseudotemporal order that explains the observed distribution of cell states. This approach was recently used to independently validate differentiation trajectories predicted using another approach that combines perturbation experiments with clustering and pseudotemporal ordering algorithms[87].

Interestingly, in the context of GRNs, combining the information extracted from RNA abundance and RNA velocity has been shown to improve the ability to predict true targets of transcription factors across a range of species and for experimental settings that mimic the sparseness and noisiness of scRNA-seq data[88]. Moreover, implementation of a time-delay between gene-gene interactions was shown to result in greater accuracy to identify time-delayed interactions and infer network topology[89, 90]. As such, similar to application of time-series bulk data, we reason that causality can be inferred in GRNs using a combination of RNA velocity and pseudotemporal ordering (Fig. 5).

Overcoming the drawbacks of single-cell data

While the unique properties of scRNA-seq data can specify context and add directionality to the co-expression network, scRNA-seq data does have some drawbacks that can reduce its downstream usage for GRN construction. Single cells contain a limited amount of RNA, which means input material requires many rounds of PCR amplification before being sequenced. In consequence, amplification bias and dropouts[91] (genes not detected even though they are expressed, resulting in observed zero-inflated gene expression) are two important factors that can affect scRNA-seq data. In addition to dropouts, zero-inflated gene expression can also be the result of biological heterogeneity[76, 92]. To estimate the contribution of dropouts to zero-inflated gene expression in scRNA-seq data, the abundance of known amounts of synthetic RNA spike-ins and endogenous transcripts have been quantified using both scRNA-seq and single molecule RNA fluorescent in situ hybridization[93, 94]. These comparisons have revealed an average RNA capture efficiency of ~5-25%, depending on the gene expression level and the scRNA-seq protocol used[66, 95, 96].

Amplification bias has been largely addressed by including unique molecular identifiers (UMIs), which tag individual molecules in the 3'-end scRNA-seq protocols, before PCR amplification[97]. However, so far, no general solution has been found for the dropout events. Two major approaches have been proposed to resolve this issue: gene expression imputation and multi-omics imputation. The challenge of zero-inflated expression in scRNA-seq data is one that several groups are trying to tackle by developing gene expression imputation algorithms, such as scImpute, CIDR and MAGIC[98-100]. All these algorithms look for patterns in other cells that can be used to infer the true expression value of the dropout, but they differ in approach and their performance varies depending on the number of dropouts in the data (for a detailed comparison of recent imputation algorithms see [101]). The complex biological relationship between multiple omics layers may also provide additional information that would otherwise not be extracted by investigating a single omics layer. By taking advantage of cross-omics relationships, multi-omics imputation has been shown to outperform single omics imputation approaches in bulk data, especially when the sample size is small or noise is large[102].

Single-cell data may also benefit from multi-omics imputation. Several technologies have been developed to measure multiple omics layers in a single cell, varying from combinations of genomics, epigenomics and transcriptomics to transcriptomics and proteomics (see review in [103]). These studies neatly illustrate the potential of single-cell multi-omics data, however we do not expect direct usage for imputation purposes because the single-cell omics layers are extremely sparse and generating these datasets is very time intensive and costly. Instead we expect that a close exchange between the information obtained in bulk and single-cell data can reveal the subcell type and genetic-context-specific rules to improve multi-omics imputation in single-cell data. The feasibility of this concept was recently assessed by Buenrostro *et al*., who elucidate the relationship between chromatin accessibility and expression levels using a combination of bulk and single-cell data[104]. Another approach was taken by Welch *et al.*, who computationally combined multiple omics layers from different single cells ordered in pseudotime[105]. This enabled them to retrieve the cross-omics relationship within individual cells and accurately infer the cross-omics correlations. While several promising solutions are being developed, none have completely solved the problems surrounding the sparseness of single-cell data, and this will likely remain an area of intense study for the next years to come.

4. Integrative approach for GRN assembly

Considering the unique features and applicability of both bulk and scRNA-seq data for generating GRNs, we propose using an integrative approach to assemble context-specific, personalized GRNs that can help move towards improved precision medicine in the future. This integrative approach combines the richness of bulk data with the finer detail and unique insights obtained from single-cells (Fig. 6). Our proposed approach consists of an interplay alternating between bulk and single-cell data, iteratively updating GRNs with knowledge acquired from both sources of data. This allows us to take full advantage of both technologies and recent collaborative efforts, such as the Human Cell Atlas[106] and the GTEx consortium[22]. In the following sections we describe the three steps of this integrative approach using the example of $CD4^+$ T-cell data illustrated in Figure 6.

Bulk-based reference co-expression network

The first step in assembling a context-specific GRN is establishing a cell-type-specific reference network that can be used as a baseline onto which the specific contexts can be projected. To create this reference network, numerous publically available datasets for specific cell types made with bulk RNA-seq can be used. Public RNA-seq repositories, such as the European Nucleotide Archive and the Sequence Read Archive, already contain hundreds of bulk RNA-seq datasets from purified cell types. Combining these datasets from different resources requires uniform

alignment, quantification and removal of batch effects[107], and multiple recent efforts have worked to combine such uniformly processed bulk RNA-seq datasets in large repositories[108-111].

While single-cell data provides a more detailed context of the network, it does not currently have the sensitivity of bulk data and will create an incomplete network due to dropouts. The bulk reference co-expression network thus serves two purposes. The first is to fill gaps in the network where expression, and therefore any possibility of an interaction, is missing for genes. The second is to provide additional supporting information when evidence from single-cell expression data is insufficient to confidently report the interaction between two genes. In this scenario, comparison between the single-cell and bulk RNA-seq reference can be used to gain additional support for the interaction.

Fine-tuning the reference co-expression network to reflect the context

The second step in assembling the context-specific GRN is to use scRNA-seq data to add context-specific information to the bulk-based reference co-expression network. Sampling can be performed on a full tissue, after which individual cell types can be dissected using the single-cell expression profiles. This allows for the creation of cell-type-specific networks without the need to predefine the studied cell types. Furthermore, for each of the identified cell types, the effect of environmental perturbations can be studied. To illustrate this second step, we provide an example in Figure 6 of a CD4+ T cell-specific and pathogen stimulation-perturbed network. By generating such a network for each individual within a study separately, the higher complexity of the network can be captured.

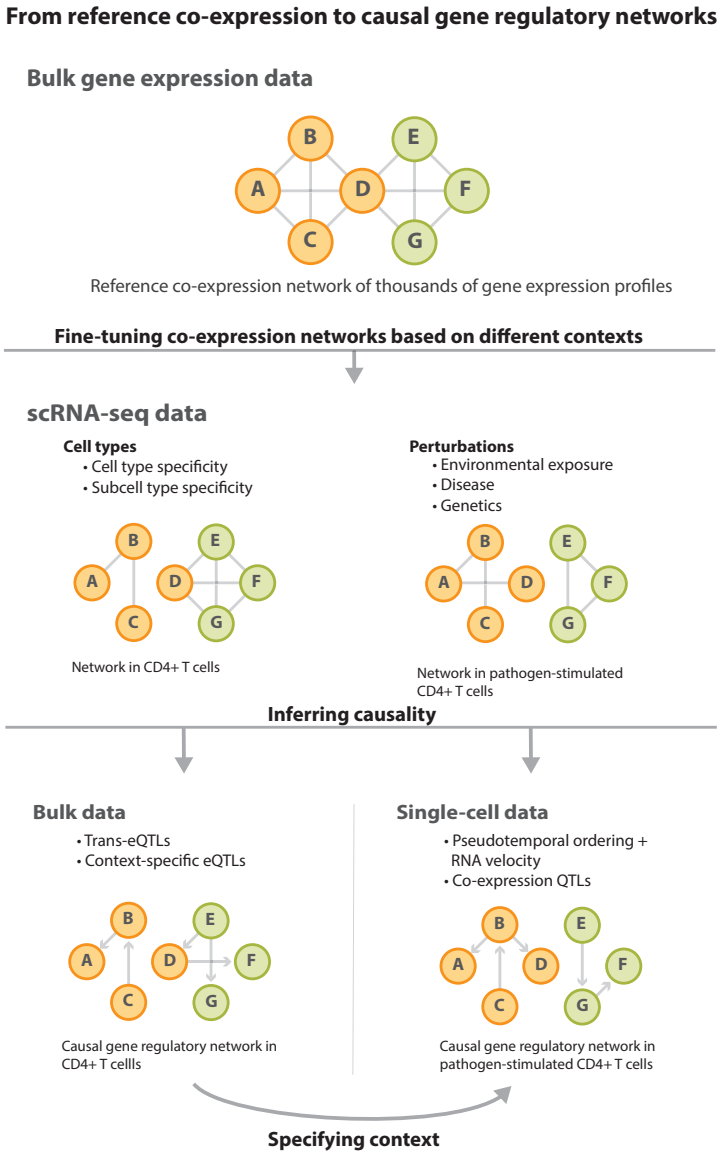


Figure 6 | Reconstruction of personalized, context-specific GRNs through the integration of bulk and single-cell data. We expect the richness of bulk expression data (e.g. number of genes or transcript variants detected, number of datasets available for any given tissue) combined with the context-specificity of scRNA-seq data (e.g. cell type, environmental exposure) to facilitate the generation of context-specific co-expression networks. Finally, integrating additional data layers, such as context-specific eQTLs and co-expression QTLs combined with ChIP-seq data, will allow the direction of effects to be determined. This will enable the reconstruction of personalized, context-specific GRNs.

Transitioning from associations to causal relations

The final step in assembling the GRN is adding directionality to the context-specific network to gain insight into the putatively causal relations between genes and to validate them using experimental approaches. Our proposed method solves the issue of directionality by integrating information from eQTLs or pseudotemporal ordering into the network.

There are several ways in which eQTLs can be used to gain insight into the GRN. First, they can reveal downstream effects of gene regulation. SNPs that have both *cis* and *trans* effects on gene expression can be used to uncover regulatory relationships between genes. For example, Westra *et al*. have shown that such SNPs may affect the expression of a transcription factor in *cis* and consequently affect the expression of many downstream genes in *trans*[24]. For a number of these downstream genes supporting ChIP-seq data was found, suggesting directionality of regulation. Second, context-dependent eQTLs[29] and co-expression QTLs[30] can uncover the upstream interactors of some genes and identify parts of the network where the relationships between genes change with genotype (Fig. 4). Altogether, by combining *cis*-, *trans*- and context-dependent eQTLs or co-expression QTLs, branches of a GRN can be reconstructed and extended with genetic information.

To put the regulatory information obtained from eQTLs in the correct context, the cell types in which the eQTL effects manifest have to be identified[22, 24, 25]. Identification of *trans*-eQTLs and context-dependent eQTLs requires enormous sample sizes in bulk expression data (thousands of samples) to overcome a severe multiple-testing burden[24, 29]. Such massive datasets are currently only available for whole tissues in bulk (such as whole blood), but these do not allow identification of the relevant cell type. While the sample size of single-cell datasets does not permit these analyses at the genome-wide level, single-cell datasets can be used to determine the cell type in which an eQTL effect identified from bulk data manifests. After pinpointing the relevant cell type, the bulk multi-omics data of this specific cell type can be used to identify or verify the regulating mechanism behind the context-dependent interaction. In one example of how this could work, a genetic variant was shown to change enhancer-promoter looping by affecting the affinity of a cell-type-specific transcription factor[112]. By providing this kind of evidence for the regulating mechanism, causality can be integrated into the parts of the cell-type-specific GRN for which eQTLs can be found.

Combining pseudotemporal ordering with RNA velocity allows the identification of directionality between all genes, not just eQTL genes. Identifying which genes share similar expression patterns and the pseudotime at which they are expressed can establish the directional relationship between these genes (Fig. 5). Van Dijk *et al.* recently showed promising results with a comparable approach in which gene expression imputed scRNA-seq data was ordered along pseudotime[100]. Subsequently, kNN-DREMI, an approach to quantify the strength of non-linear and noisy gene-gene relationships, was used to select those transcription factors and potential targets that changed considerably along pseudotime. In this way, they could reconstruct a

large GRN consisting of 719 transcriptional regulators and 11,126 downstream target genes. Of the predicted target genes that were tested, 92% significantly overlapped with target genes as assessed by ATAC-seq. This study showed promising results to identify target genes without experimental perturbation. However, experimental validation is required to transition from associations to causal relations, and pooled CRISPR-screens coupled with a scRNA-seq readout (such as CROP-seq[113], CRISP-seq[114] or PERTURB-seq[65, 115]) are especially suited for this purpose. These methods enable the mapping of downstream consequences of gene perturbations on the whole transcriptome level. Our proposed integrative approach will help provide focus on those regions of the network that are of particular interest and alleviates the need to perform experimental validation on every possible gene. Altogether, we expect that such an integrative approach enables the reconstruction of well-validated context-specific, personalized GRNs.

5. The future of precision medicine

A major challenge in healthcare is that the majority of currently prescribed drugs are only effective in a small subset of patients[31]. This not only leads to money lost on ineffective drugs, but it also unnecessarily exposes patients to adverse drug side-effects. Well-validated, context-specific, personalized GRNs will be essential to move from more traditional medicine towards precision medicine, which will provide treatment or preventive measures that will be effective for specific patients based on their genetic, environmental and lifestyle characteristics (Fig. 3).

Disease-specific GRNs may provide novel insights into disease pathogenesis and have enhanced power to prioritize disease-causing genes[116]. These GRNs provide a bird's-eye view which is required to look beyond the disruption of individual disease genes: while each gene may have a small individual effect, several disease genes together may have a large additive effect when converging into a few disrupted key regulatory pathways[117-119]. Despite different individual disease genes, similar key regulatory pathways are likely to be disturbed in different diseases. Likewise, exposure to specific environmental factors may disturb regulatory pathways in a fashion comparable to specific disease-associated genetic variants. These insights may provide novel links between different diseases or clues to how environmental factors can contribute to one or more diseases, and these new associations should provide novel directions for treatment.

Generation of context-specific GRNs may never fully capture the complexity of multifactorial interactions (genetic background, environmental exposures, disease, etc.) and the intercellular communication that influences the whole organism. Nevertheless, GRNs will be valuable for predicting the outcome of perturbations, and this particular function of GRNs will be useful for predicting potential drug targets for disease. Tumor-specific networks inferred using a combination of gene expression data and cancer-related signaling pathways have already been successfully applied

to identify oncogenes and previously identified targets of cancer treatment[120]. Gene expression and network data of 235 bioactive molecules have also been combined to identify known drug targets with high accuracy (AUC > 0.9)[121].

Not only are GRNs expected to lead to more effective treatment, they may also enable the identification of high risk individuals who would benefit from preventive medicine. Predictions based on genetic information through polygenic risk scores (PRS) have been shown to be valuable in this regard. In one example, individuals given the highest likelihood of developing coronary heart disease by a PRS model were shown to benefit the most from statin treatment[122, 123]. Comparable to genetic prediction models, personalized GRNs are expected to enable identification of high risk individuals, but with the added benefit of understanding the mechanism that underlies the prediction and providing insight into environmental risk factors that might exacerbate genetic risk. We expect that a combination of GRNs and PRS would provide valuable information for disease risk reduction, especially for complex diseases that are less well understood.

An integral component to predict disease in individuals before disease manifestation using both PRS and GRN, is to have their genotype information available. However, the problem is that (seemingly) healthy individuals infrequently come into contact with the healthcare system. Therefore, to implement this in current clinical practice, every individual would ideally be genotyped, for example, directly after birth or during a visit to the general practitioner, as is done with vaccinations. Genotyping using genotype arrays in combination with imputation-based approaches has become very cost-effective, mapping clinically relevant genetic variants with high coverage at low cost (~30 euros/individual). With the steady decrease in price, large-scale genotyping projects have already begun. 500,000 individuals have been genotyped within the UK Biobank population cohort for example[124]. The small investment in genotyping early in life may pay off in the future, as the resulting personalized GRNs are expected to provide informed personal lifestyle advice that could reduce disease risk. The Finnish GeneRISK study is already demonstrating how this could work by showing how providing individuals with a personal disease risk assessment can have a major, long-term effect on their lifestyle[125]. This disease risk assessment was based on 49,000 disease-associated genetic variants and lifestyle-associated risk factors. Overall, 1.5 years after the disease risk assessment, 32.4% of the individuals who were informed they were at >10% risk for cardiovascular diseases showed risk-reducing behavior.

Genotyping individuals may also allow doctors to select effective drugs without adverse effects. However, before this can be implemented in clinical practice, a shift in both public perception and healthcare regulations is required. In 2013 the company 23andMe, which offers commercial genetic testing, was prohibited from providing health risk assessments as part of their products by the US Food and Drug Administration (FDA), as the health information provided by their genetic testing was not accurate enough and not individually actionable. Four years later, the FDA gave

23andMe approval to provide the first at-home DNA test for 10 diseases and clinical conditions, marking a shift in the regulations required to provide genotyping for every individual[126]. Before this DNA test was FDA approved, 23andMe had to assure that users fully understood what the test entails and what the results mean. In addition to this, we believe that trained professionals, such as genetic counselors, are essential for helping individuals interpret genetic results in order to prevent misconceptions and un-needed anxiety. These recent developments are expected to facilitate implementation of genetic testing in the future.

6. Conclusions and future perspectives

In this perspective we have highlighted the importance of using a gene network-based approach rather than a single-gene focused approach to gain the bird's eye view required to understand disease pathogenesis. As diseases arise in highly specific contexts, context-dependent GRNs are needed to understand these diseases. To build these context-dependent GRNs, we have proposed an integrative approach of generating GRNs using both bulk and single-cell data. We have not covered the computational implementation of our approach, as this would go beyond the scope of this perspective. Nevertheless, we expect that our iterative approach is well-suited to implementation using machine learning or deep learning models that learn from large datasets and make predictions on likely outcomes of complex cellular systems such as GRNs[127, 128]. However, massive datasets of this kind are not yet available at the single-cell level. Therefore, large single-cell reference panels, like those being assembled by the Human Cell Atlas[106], will be instrumental for executing our integrated approach using machine learning algorithms. Moreover, platforms like the Human Cell Atlas are expected to provide more uniform guidelines and solutions for generating, processing and handling large-scale scRNA-seq data. This will facilitate the combining of scRNA-seq datasets as part of our integrative approach.

We conclude this perspective with the possibilities that may open up with the integration of person-specific information in GRNs. As initiatives such as 23andMe and the UK Biobank produce ever larger genetic datasets that could be used to reconstruct personalized GRNs, and new initiatives are started, the ability to accurately predict disease risk through a combination of genotype associations and personalized GRNs will improve. However, before these personalized GRNs may be adopted in clinical practice, a number of ethical and legal issues will have to be resolved. Clinical guidelines, for instance, will have to be developed so that the interpretation of the results can be guided by trained professionals and the actionability of individual findings has to become clear[32]. Additionally, updated privacy and data protection regulations, such as the General Data Protection Regulation implemented in the EU in 2018, will be an important aspect to reduce privacy concerns in the general public. Once these issues have been addressed, however, we expect that personalized, context-dependent GRNs will accelerate the progress required to make the next big leap in precision medicine.

Glossary terms

genome-wide association studies (GWAS): genome-wide approach in which genetic variants such as single nucleotide polymorphisms (SNPs) are linked to a molecular trait or disease.

expression quantitative trait loci (eQTL): SNP that explains variation in gene expression levels. When the SNP explains the expression of a gene within a 1 megabase distance, it is called a *cis*-eQTL. When it explains the expression of a gene beyond a 1 megabase distance, it is called a *trans*-eQTL.

co-localization: a method that determines whether the association signals in a locus correspond between two association studies (e.g. between a GWAS and an eQTL study).

fine-mapping: a statistical approach that is used to prioritize the most likely causal genetic variant in a previously identified locus that is linked to a specific phenotype.

Summary-data-based Mendelian randomization (SMR): a summary statistics based variant of Mendelian Randomization that leverages the principle that genetic variation is randomly assigned to a sample with a specific phenotype to infer causality between genetic variation and the phenotype in an observational study.

gene regulatory network (GRN): a directional co-expression network that also contains information about the regulators that control gene expression.

precision medicine: healthcare that is individually tailored on the basis of a person's genetic, environmental and lifestyle characteristics.

co-expression network: an undirected network that describes which genes often behave in a coordinated manner. This network consists of nodes, representing genes, that are connected through edges that represent relationships between nodes. These relationships can be unweighted or weighted, indicating either a binary relationship (on/off) or a more continuous relationship.

Bayesian model: statistical modeling to calculate probabilities for an observation while taking into account the empirical or theoretical expected distribution of these observations or factors expected to influence the observations. Used in co-expression networks to assign probabilities for directionality between genes.

Nyquist–Shannon sampling theorem: describes the sample frequency that is sufficient to capture all the information from a continuous-time signal of a finite bandwidth.

context-dependent eQTLs: eQTLs for which the strength of association depends on a secondary factor. This may be either intrinsic (e.g. expression of another gene, cell type frequency) or extrinsic (e.g. environmental exposure). Gene expression can be used as proxy measurements for both intrinsic and extrinsic factors.

co-expression QTLs: SNPs that modulate the correlation between the co-expression of two genes. To calculate these, many observations (e.g. on multiple cells or tissues) per individual are required.

Simpson's paradox: the situation in which an observed relationship within each sample reverses when the samples are combined.

genotype by environment (GxE) interactions: interactions between an individual's genotype and the environment. Context-dependent eQTLs are a subset of GxE interactions.

stochastic bursting: a fundamental property of genes in which gene transcription occurs in "bursts".

pseudotime: inferred temporal sequences of gene expression states in cells from measurements made at a single moment in time.

RNA velocity: the rate of change of mRNA molecule abundances in the cell determined by modeling the relationship between unspliced mRNA (indicator of current transcription) and spliced mRNA (indicator of transcription in the recent past).

amplification bias: bias that arises due to the sequence-dependent amplification efficiency of PCR.

dropouts: genes that fail to be detected even though they are expressed (resulting in a zero-inflated gene expression distribution).

synthetic RNA spike-in: an RNA transcript of known sequence and quantity used to calibrate measurements in gene expression assays.

unique molecular identifiers (UMIs): barcode sequences tagging individual molecules.

cross-omics relationships: relationships between different measurements of molecular phenotypes (e.g. methylation and transcription).

polygenic risk score: predictive model for a disease or trait based on multiple genetic variants.

machine learning approaches: methods to analyze massive amounts of data to build predictive models from multi-dimensional datasets.

References

1. Visscher, P. M. *et al*. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5-22 (2017).
2. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am. J. Hum. Genet.* **99**, 139-153 (2016).
3. Yang, J. *et al*. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114-1120 (2015).
4. Brodie, A., Azaria, J. R. & Ofran, Y. How far from the SNP may the causative genes be? *Nucleic Acids Res.* **44**, 6046-6054 (2016).
5. McGovern, A. *et al*. Capture Hi-C identifies a novel causal gene, IL20RA, in the pan-autoimmune genetic susceptibility region 6q23. *Genome Biol.* **17**, 212-016-1078-x (2016).
6. Claussnitzer, M. *et al*. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.* **373**, 895-907 (2015).
7. Smemo, S. *et al*. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371-375 (2014).
8. Schadt, E. E. *et al*. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297-302 (2003).
9. Cheung, V. G. *et al*. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* **33**, 422-425 (2003).
10. Zhu, Z. *et al*. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481-487 (2016).
11. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* (2018).
12. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
13. FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al*. A promoter-level mammalian expression atlas. *Nature* **507**, 462-470 (2014).
14. Roadmap Epigenomics Consortium *et al*. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).
15. Adams, D. *et al*. BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotechnol.* **30**, 224-226 (2012).
16. Bell, J. T. *et al*. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12**, R10-2011-12-1-r10. Epub 2011 Jan 20 (2011).
17. McVicker, G. *et al*. Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747-749 (2013).
18. Kasowski, M. *et al*. Variation in transcription factor binding among humans. *Science* **328**, 232-235 (2010).
19. Degner, J. F. *et al*. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390-394 (2012).
20. Pai, A. A. *et al*. The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLoS Genet.* **8**, e1003000 (2012).
21. Pickrell, J. K. *et al*. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768-772 (2010).
22. GTEx Consortium *et al*. Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).

23. Gamazon, E. R. *et al*. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* (2018).

24. Westra, H. J. *et al*. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238-1243 (2013).

25. Battle, A. *et al*. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14-24 (2014).

26. Knowles, D. A. *et al*. Allele-specific expression reveals interactions between genetic variation and environment. *Nat. Methods* **14**, 699-702 (2017).

27. Fave, M. J. *et al*. Gene-by-environment interactions in urban populations modulate risk phenotypes. *Nat. Commun.* **9**, 827-018-03202-2 (2018).

28. Fagny, M. *et al*. Exploring regulation in tissues with eQTL networks. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E7841-E7850 (2017).

29. Zhernakova, D. V. *et al*. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139-145 (2017).

30. van der Wijst, M. G. P. *et al*. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* (2018).

31. Schork, N. J. Personalized medicine: Time for one-person trials. *Nature* **520**, 609-611 (2015).

32. Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nature Reviews Genetics* **19**, 299 (2018).

33. Fiers, M. W. E. J. *et al*. Mapping gene regulatory networks from single-cell omics data. *Brief Funct. Genomics* (2018).

34. Ogata, H. *et al*. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **27**, 29-34 (1999).

35. Fabregat, A. *et al*. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649-D655 (2018).

36. van Dam, S., Vosa, U., van der Graaf, A., Franke, L. & de Magalhaes, J. P. Gene co-expression analysis for functional classification and gene-disease predictions. *Brief Bioinform* (2017).

37. Wang, Y. X. & Huang, H. Review on statistical methods for gene network reconstruction using expression data. *J. Theor. Biol.* **362**, 53-61 (2014).

38. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17-6115.1128. Epub 2005 Aug 12 (2005).

39. Pers, T. H. *et al*. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).

40. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* **9**, 1090-018-03424-4 (2018).

41. Shim, U. *et al*. Pathway Analysis Based on a Genome-Wide Association Study of Polycystic Ovary Syndrome. *PLoS One* **10**, e0136609 (2015).

42. Raychaudhuri, S. *et al*. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5**, e1000534 (2009).

43. Choobdar, S. *et al*. Open Community Challenge Reveals Molecular Network Modules with Key Roles in Diseases. *bioRxiv* (2018).

44. Sanchez-Castillo, M., Blanco, D., Tienda-Luna, I. M., Carrion, M. C. & Huang, Y. A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics* **34**, 964-970 (2018).

45. Sima, C., Hua, J. & Jung, S. Inference of gene regulatory networks using time-series data: a survey. *Curr. Genomics* **10**, 416-429 (2009).

46. Pe'er, D., Regev, A., Elidan, G. & Friedman, N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* **17 Suppl 1**, S215-24 (2001).

47. Freeman, M. Feedback control of intercellular signalling in development. *Nature* **408**, 313-319 (2000).

48. Whittaker, E. T. XVIII.—On the Functions which are represented by the Expansions of the Interpolation-Theory. *Proceedings of the Royal Society of Edinburgh* **35**, 181-194 (1915).

49. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379-423 (1948).

50. Bar-Joseph, Z. Analyzing time series gene expression data. *Bioinformatics* **20**, 2493-2503 (2004).

51. Whitfield, M. L. *et al*. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13**, 1977-2000 (2002).

52. Tasaki, S. *et al*. Bayesian network reconstruction using systems genetics data: comparison of MCMC methods. *Genetics* **199**, 973-989 (2015).

53. Liao, W. & Ji, Q. Learning Bayesian network parameters under incomplete data with domain knowledge. *Pattern Recognit* **42**, 3046-3056 (2009).

54. Feelders, A. J. &Gaag, L. C. v. d. *Learning Bayesian Network Parameters with Prior Knowledge about Context-Specific Qualitative Influences* (UAI, 2005).

55. Marbach, D. *et al*. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods* **13**, 366-370 (2016). **Integration of multiple data layers to reconstruct tissue-specific GRNs.**

56. Saha, A. *et al*. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res.* **27**, 1843-1858 (2017).

57. Greene, C. S. *et al*. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47**, 569-576 (2015).

58. Tang, F. *et al*. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377-382 (2009).

59. Rosenberg, A. B. *et al*. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* (2018).

60. Zheng, G. X. Y. *et al*. Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049 (2017).

61. Pina, C. *et al*. Single-Cell Network Analysis Identifies DDIT3 as a Nodal Lineage Regulator in Hematopoiesis. *Cell. Rep.* **11**, 1503-1510 (2015).

62. Herbach, U., Bonnaffoux, A., Espinasse, T. & Gandrillon, O. Inferring gene regulatory networks from single-cell data: a mechanistic approach. *BMC Syst. Biol.* **11**, 105-017-0487-0 (2017).

63. Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell. Syst.* **5**, 251-267.e3 (2017).

64. Bartlett, T. E., Muller, S. & Diaz, A. Single-cell Co-expression Subnetwork Analysis. *Sci. Rep.* **7**, 15066-017-15525-z (2017).

65. Dixit, A. *et al*. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e17 (2016).

66. Marinov, G. K. *et al*. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.* **24**, 496-510 (2014).

67. Villani, A. C. *et al*. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, 10.1126/science.aah4573 (2017).

68. Jaitin, D. A. *et al*. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776-779 (2014).

69. Buettner, F. *et al*. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155-160 (2015).

70. Kang, H. M. *et al*. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* (2017).

71. Aibar, S. *et al*. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083-1086 (2017).

72. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* (2018).

73. Martins, A. J. *et al*. Environment Tunes Propagation of Cell-to-Cell Variation in the Human Macrophage Gene Network. *Cell. Syst.* **4**, 379-392.e12 (2017).

74. Simpson, E. H. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society.Series B (Methodological)* **13**, 238-241 (1951).

75. Sollid, L. M. & Jabri, B. Triggers and drivers of autoimmunity: lessons from coeliac disease. *Nat. Rev. Immunol.* **13**, 294-302 (2013).

76. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA Synthesis in Mammalian Cells. *PLOS Biology* **4**, e309 (2006).

77. Brennecke, P. *et al*. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093-1095 (2013).

78. Ji, Z. & Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* **44**, e117 (2016).

79. Trapnell, C. *et al*. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381-386 (2014).

80. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *bioRxiv* (2018).

81. Specht, A. T. & Li, J. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics* **33**, 764-766 (2017).

82. Hamey, F. K. *et al*. Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 5822-5829 (2017).

83. Ocone, A., Haghverdi, L., Mueller, N. S. & Theis, F. J. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics* **31**, i89-96 (2015).

84. Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E2467-E2476 (2018).

85. Rashid, S., Kotton, D. N. & Bar-Joseph, Z. TASIC: determining branching models from time series single cell data. *Bioinformatics* **33**, 2504-2512 (2017).

86. La Manno, G. *et al*. RNA velocity in single cells. *bioRxiv* (2017).

87. Plass, M. *et al*. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* (2018).

88. Desai, J. S., Sartor, R. C., Lawas, L. M., Jagadish, S. V. K. & Doherty, C. J. Improving Gene Regulatory Network Inference by Incorporating Rates of Transcriptional Changes. *Sci. Rep.* **7**, 17244-017-17143-1 (2017).

89. Finkle, J. D., Wu, J. J. & Bagheri, N. Windowed Granger causal inference strategy improves discovery of gene regulatory networks. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 2252-2257 (2018).

90. Schmitt, W. A.,Jr, Raab, R. M. & Stephanopoulos, G. Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. *Genome Res.* **14**, 1654-1663 (2004).

91. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740-742 (2014).

92. Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell* **123**, 1025-1036 (2005).

93. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877-879 (2008).

94. Femino, A. M., Fay, F. S., Fogarty, K. & Singer, R. H. Visualization of single RNA transcripts in situ. *Science* **280**, 585-590 (1998).

95. Svensson, V. *et al*. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381-387 (2017).

96. Zeisel, A. *et al*. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138-1142 (2015).

97. Islam, S. *et al*. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163-166 (2014).

98. Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9**, 997-018-03405-7 (2018).

99. Lin, P., Troup, M. & Ho, J. W. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* **18**, 59-017-1188-0 (2017).

100. van Dijk, D. *et al*. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 1-14 (2018).

101. Zhang, L. & Zhang, S. Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1-1 (2018).

102. Lin, D. *et al*. An integrative imputation method based on multi-omics datasets. *BMC Bioinformatics* **17**, 247-016-1122-6 (2016).

103. Hu, Y. *et al*. Single Cell Multi-Omics Technology: Methodology and Application. *Front. Cell. Dev. Biol.* **6**, 28 (2018).

104. Buenrostro, J. D. *et al*. Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).

105. Welch, J. D., Hartemink, A. J. & Prins, J. F. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol.* **18**, 138-017-1269-0 (2017).

106. Regev, A. *et al*. The Human Cell Atlas. *Elife* **6**, 10.7554/eLife.27041 (2017).

107. Wang, Q. *et al*. Unifying cancer and normal RNA sequencing data from different sources. *Sci. Data* **5**, 180061 (2018).

108. Li, J. R. *et al*. Cancer RNA-Seq Nexus: a database of phenotype-specific transcriptome profiling in cancer cells. *Nucleic Acids Res.* **44**, D944-51 (2016).

109. Vivian, J. *et al*. Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.* **35**, 314-316 (2017).

110. Collado-Torres, L. *et al*. Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **35**, 319-321 (2017).

111. Lachmann, A. *et al*. Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **9**, 1366-018-03751-6 (2018).

112. Weintraub, A. S. *et al*. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171**, 1573-1588.e28 (2017).

113. Datlinger, P. *et al*. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297-301 (2017).

114. Jaitin, D. A. *et al*. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* **167**, 1883-1896.e15 (2016).

115. Adamson, B. *et al*. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882.e21 (2016).

116. Gao, L. *et al*. Identifying noncoding risk variants using disease-relevant gene regulatory networks. *Nat. Commun.* **9**, 702-018-03133-y (2018).

117. Menche, J. *et al*. Integrating personalized gene expression profiles into predictive disease-associated gene pools. *NPJ Syst. Biol. Appl.* **3**, 10-017-0009-0. eCollection 2017 (2017).

118. Menche, J. *et al*. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).

119. Chatterjee, S. *et al*. Enhancer Variants Synergistically Drive Dysfunction of a Gene Regulatory Network In Hirschsprung Disease. *Cell* **167**, 355-368.e10 (2016).

120. Peng, Q. & Schork, N. J. Utility of network integrity methods in therapeutic target identification. *Front. Genet.* **5**, 12 (2014).

121. Laenen, G., Thorrez, L., Bornigen, D. & Moreau, Y. Finding the targets of a drug by integration of gene expression data with a protein interaction network. *Mol. Biosyst* **9**, 1676-1685 (2013).

122. Natarajan, P. *et al*. Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting. *Circulation* **135**, 2091-2101 (2017).

123. Mega, J. L. *et al*. Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *Lancet* **385**, 2264-2271 (2015).

124. Bycroft, C. *et al*. Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* (2017).

125. Widen, E. Returning cardiovascular disease risk prediction back to individuals motivates beneficial lifestyle changes : Preliminary results from the GeneRISK study. (Individual access to genomic disease risk has a beneficial impact on lifestyles, European Society of Human Genetics 2018, June 15 2018).

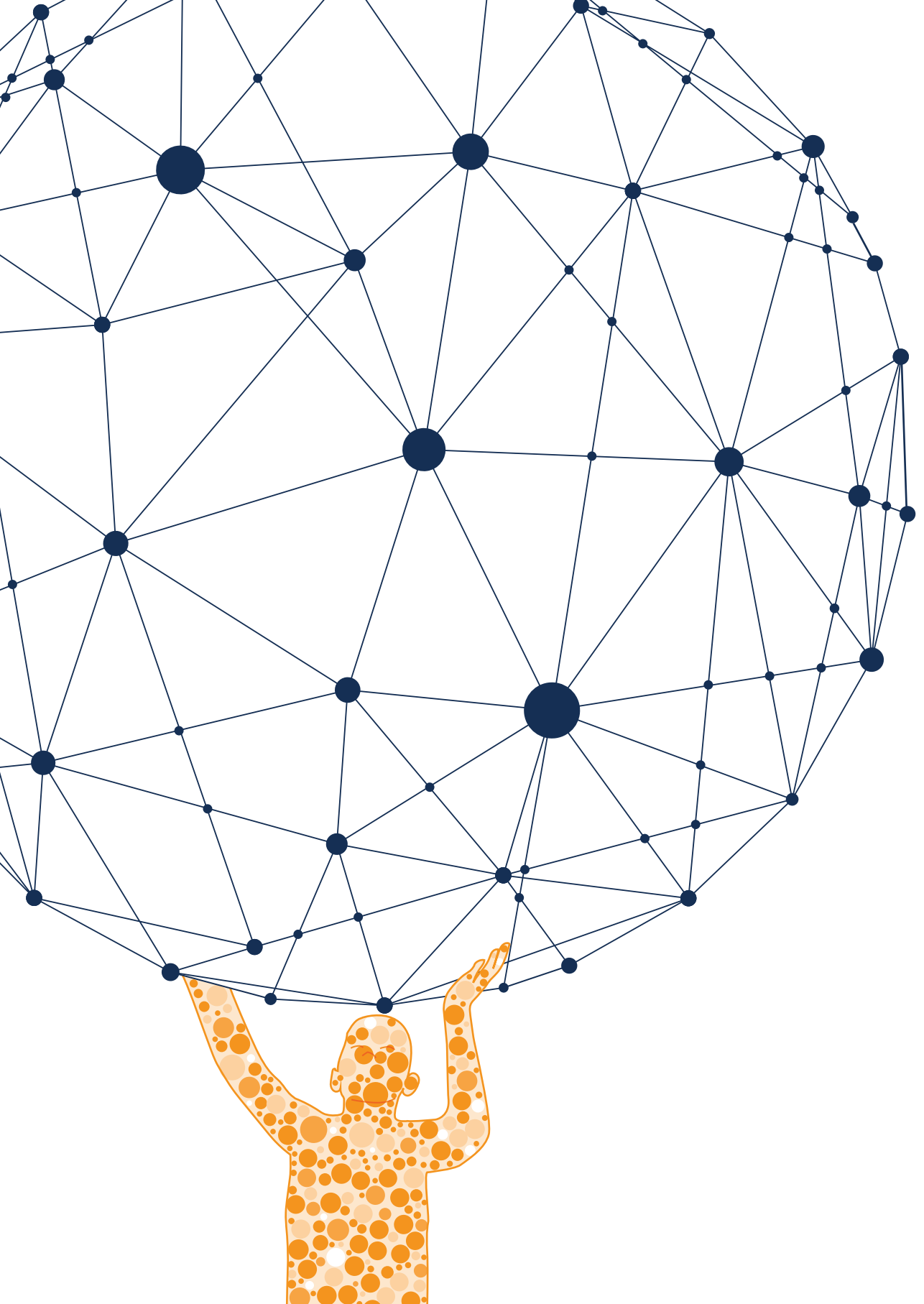
126. US Food and Drug Administration. FDA allows marketing of first direct-to-consumer tests that provide genetic risk information for certain conditions. https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm551185.htm (April 6, 2017).

127. Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. & Collins, J. J. Next-Generation Machine Learning for Biological Networks. *Cell* **173**, 1581-1592 (2018).

128. Ching, T. *et al*. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, 10.1098/rsif.2017.0387 (2018).

Acknowledgements

We thank J. Senior and K. McIntyre for careful editing of the manuscript. L.F. is supported by grants from the Dutch Research Council (ZonMW-VIDI 917.14.374 to L.F.) and a European Research Council Starting Grant, grant agreement 637640 (ImmRisk).



CHAPTER 3

Single-cell RNA sequencing identifies cell type-specific cis-eQTLs and co-expression QTLs

Monique G.P. van der Wijst, Harm Brugge[#], Dylan H. de Vries[#], Patrick Deelen, Morris A. Swertz, LifeLines Cohort Study, BIOS Consortium, Lude Franke[*].

Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.
[#]Both authors contributed equally.
[*] Correspondence should be addressed to L.F.: lude@ludesign.nl

Published in Nature Genetics. https://doi.org/10.1038/s41588-018-0089-9

Main text

Genome-wide association studies have identified thousands of genetic variants that are associated with disease.[1] Most of these variants have small effect sizes, but their downstream expression effects, so-called expression quantitative trait loci (eQTLs), are often large[2] and cell type-specific[3-5]. To identify these cell type-specific eQTLs using an unbiased approach, we used single-cell RNA sequencing (scRNA-seq) to generate expression profiles of ~25,000 peripheral blood mononuclear cells (PBMCs) from 45 donors. We identified previously reported *cis*-eQTLs, but also identified new cell type-specific *cis*-eQTLs. Finally, we generated personalized co-expression networks, and identified genetic variants that significantly alter co-expression relationships (which we termed 'co-expression QTLs'). Single-cell eQTL analysis thus allows for the identification of genetic variants that impact regulatory networks.

Previously, purified cell types[4,6-8] or deconvolution methods[9,10] have been used to identify cell type-specific eQTLs. However, these methods are biased towards specific cell types, or are of limited use for less abundant cell types and dependent on accurately defined marker genes.[11] In contrast, scRNA-seq can be used to investigate rare cell types[12], and thus, enables identification of cell type-specific eQTLs using an unbiased approach. Indeed, proof of concept was previously shown in a study on 15 individuals, where 92 genes were studied in 1,440 cells.[13]

Here, we studied cell type-specific effects of genetic variation on genome-wide gene expression by generating scRNA-seq data of ~25,000 PBMCs from 45 donors of the population-based cohort study Lifelines Deep[14]. After quality control (Online Methods, Suppl. Fig. 1), we first assessed to which extent previously reported *cis*-eQTLs from bulk whole blood, using either 94 DeepSAGE samples [15] (a 3'-end oriented RNA-sequencing strategy similar to our scRNA-seq approach) or 2,116 RNA-seq[11] samples, also show significant effects in the scRNA-seq dataset. For this analysis, we treated the scRNA-seq data as being bulk PBMCs (by averaging expression levels of all cells per gene per sample, referred to as 'bulk-like PBMCs'). We detected 50 and 311 significant *cis*-eQTLs (gene-level false-discovery rate (FDR) of 0.05) that were previously reported in the DeepSAGE[15] and RNA-seq[11] study, respectively (Fig. 1a, Suppl. Table 1). Although only a small proportion (8% and 1%) of previously reported *cis*-eQTLs were significant in our scRNA-seq analysis, 96% and 90.4% had identical allelic directions as in the DeepSAGE[15] and RNA-seq[11] study, respectively, indicating that these *cis*-eQTLs reflect similar regulatory effects. The few discordant eQTLs may reflect the slightly different sample composition of both datasets (PBMCs versus whole blood) and the relatively few sequence reads targeting the 3'-end of genes in the bulk RNA-seq dataset.

We subsequently performed a genome-wide *cis*-eQTL discovery analysis on the bulk-like PBMCs. Separate *cis*-eQTL analyses were conducted on each of the identified major cell types (cell type classification was performed using Seurat[16], Suppl. Fig. 2a, 2b) by averaging the normalized gene expression of all cells per cell

type, gene and donor. In total, 379 unique top *cis*-eQTLs were identified, reflecting 287 unique eQTL genes (gene-level FDR of 0.05) (Table 1), as sometimes in different cell types different SNPs showed the most significant association for an eQTL gene. While 331 (reflecting 249 unique *cis*-eQTL genes) of these 379 *cis*-eQTLs were significant in the bulk-like PBMC eQTL analysis, 48 *cis*-eQTLs (reflecting 38 unique *cis*-eQTL genes) were only detected in specific cell types (i.e. 'cell type-dependent' eQTLs, Suppl. Table 2).

We subsequently attempted to replicate these eQTLs. For the 249 eQTL genes found in the bulk-like PBMC analysis, 233 *cis*-eQTLs were testable and 181 (78%) were associated with the same SNP (90.1% shared allelic direction, Suppl. Table 2) in the whole blood RNA-seq eQTL data set[11]. For the 48 cell type-dependent *cis*-eQTLs, 29 (60%) replicated in the RNA-seq dataset[11]. This lower percentage suggests that in bulk RNA-seq datasets cell type-dependent eQTLs might become too diluted, resulting in low statistical power to recover these. While this most likely happens for rare cell types, we also observed this in common cell types. For instance, in the most abundant cell type ($CD4^+$ T cells), rs2272245 significantly affects the expression of the *TSPAN13* gene in *cis* ($P = 2.21 \times 10^{-6}$). However, this effect was not significant in the bulk-like PBMCs (P = 0.88), because *TSPAN13* is lowly expressed in CD4+ T-cells, whereas it is highly expressed in dendritic cells (DCs) where it did not show a *cis*-eQTL effect (Fig. 1b). *Cis*-eQTLs might also be missed in bulk data, because they might show opposite allelic effects across different cell types. We could not study this in detail, due to lack of power given the sample size and limited number of cells for rare cell types (Suppl. Fig. 2c). Nevertheless, in $CD4^+$ T cells, the A allele of rs4804315 significantly decreased expression of *ZNF414* in *cis* ($P = 6.09 \times 10^{-6}$), whereas in natural killer (NK) cells this allele increased expression of *ZNF414* at nominal significance (P = 0.0339) (Fig. 1b). However, it cannot be excluded that specifically in NK cells, the effect of rs4804315 on *ZNF414* expression is the result of a residual effect on *ZNF414* expression of a second, independent variant.

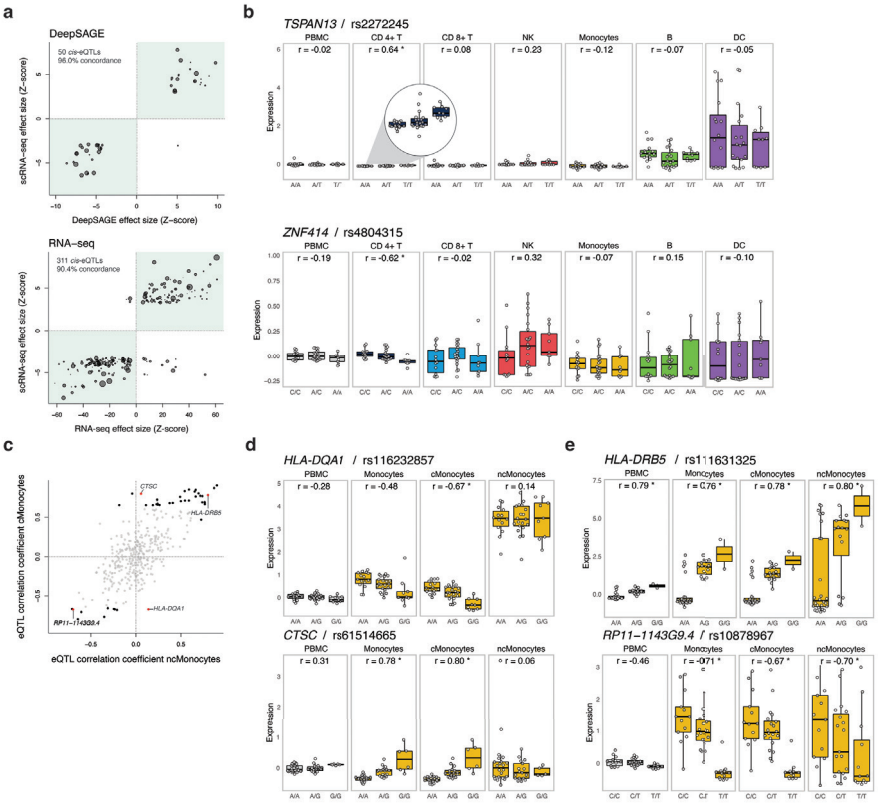


Figure 1. *Cis*-eQTL analysis in single-cell RNA-seq data. (**a**) Effect size of the *cis*-eQTLs detected in the bulk-like PBMC scRNA-seq sample in which the analysis was confined to previously reported *cis*-eQTLs in (top) whole blood DeepSAGE or (bottom) bulk RNA-seq data. The number and percentage represent, respectively, the detected *cis*-eQTLs and their concordance (i.e. same allelic direction – green quadrants) between the bulk-like PBMC population scRNA-seq eQTLs and (top) whole blood DeepSAGE or (bottom) bulk RNA-seq data. The size of each dot represents the mean expression of the *cis*-regulated gene in the total scRNA-seq dataset. (**b**) Examples of undetectable *cis*-eQTLs in the bulk-like PBMC population caused by (top) masking of the *cis*-eQTL present in CD4⁺ T cells but absent in DCs with comparatively high expression of the *cis*-regulated gene or (bottom) opposite allelic effects in CD4⁺ T and NK cells. (**c**) Spearman's rank correlation coefficient for the cMonocytes against the ncMonocytes of all top eQTLs that were identified in the total dataset or at least one (sub)cell cluster (see **Suppl. Table 2**). Significant correlations are shown in black (four red highlighted examples are shown in **d** and **e**), the non-significant in gray. (**d**) *Cis*-eQTLs specifically affecting expression in the cMonocytes, and not the ncMonocytes. (**e**) *Cis*-eQTLs significantly affecting the expression in both the cMonocytes and ncMonocytes. Each dot represents the mean expression of the eQTL gene in a donor. Box plots show the median, the first and third quartiles, and 1.5 times the interquartile range. r, Spearman's rank correlation coefficient; *FDR≤0.05.

Since some *cis*-eQTLs did not replicate in whole blood bulk RNA-seq data, we subsequently investigated eQTL datasets of purified cell types. Indeed, 3 out of 19 remaining cell type-dependent *cis*-eQTLs were detected (each with consistent allelic direction) in purified eQTL datasets of the Blueprint consortium (naïve $CD4^+$ T cells and $CD14^+$ monocytes)[17] or Kasela et al. ($CD4^+$ and $CD8^+$ T cells)[6] (Suppl. Table 3). Hence, only 16 cell type-dependent *cis*-eQTLs were not identified before using bulk eQTL datasets of blood or purified immune cells. Although some *cis*-eQTLs were only significant in specific cell types, this does not prove cell type-specificity; particularly in less abundant cell types power is lacking to detect many *cis*-eQTLs. Ways to partially overcome this, would be to use methods that consider multiple eQTL datasets together, such as eQTL-BMA[18] or Meta-Tissue[19]. However, these methods are currently computationally too demanding for large scRNA-seq data or do not define the cell type in which the eQTL effect occurs.[19,20]

Table 1. *Cis*-eQTL genes identified per cell type

Cell type	Median number of cells/donor	Unique genes with significant *cis*-eQTL effect
PBMC	507	249
$CD4^+$ T	282	145
$CD8^+$ T	74	21
NK	59	14
Monocyte	44	23
B	18	6
DC	11	9
Total (unique)		287

The median number of cells per donor (column 2) correlates fairly well with the number of detected *cis*-eQTL genes (column 3). In total, 379 unique top *cis*-eQTL effects, reflecting 287 unique eQTL genes, have been identified in the total dataset. Within each cell type, the number of unique *cis*-eQTL genes that we identified was equal to the number of unique, top *cis*-eQTL effects.

A major advantage of using scRNA-seq data is the flexibility by which any cell population of interest can be selected for eQTL analysis. In contrast, when using RNA-seq data of purified cell types, one cannot retrieve data from subcell types anymore. Moreover, while finer differences between subcell types may be detectable using gene expression profiles, it is not always recapitulated by different cell membrane markers, complicating cell sorting. Here, we show the added value of performing eQTL analysis on subcell types using two monocyte subsets: classical (cMonocytes) and non-classical monocytes (ncMonocytes). When plotting Spearman's rank correlation of each top eQTL for the cMonocytes against the ncMonocytes, several

examples were revealed that pinpointed the eQTL effect specifically to cMonocytes (Fig. 1c). Two such examples, which were previously identified in RNA-seq data of purified $CD14^+$ monocytes[17], are shown in Figure 1d. The scRNA-seq data now allowed us to specifically assign these effects to cMonocytes (Fig. 1d). Despite having lower power for detecting eQTLs in ncMonocytes due to an almost five times lower abundance compared to cMonocytes (Suppl. Fig. 2b), power in the ncMonocytes remains sufficiently high to detect several other significant ncMonocyte *cis*-eQTLs (Fig. 1e, Suppl. Table 2).

Another opportunity of scRNA-seq data is to use it for determining whether genetic variants can alter gene co-expression. Although recently genes and environmental factors altering the effect size of eQTLs ('context-specific eQTLs') have been identified in bulk RNA-seq eQTL datasets[11,21], a large sample size was required to ensure sufficient power. In contrast, scRNA-seq data enables generation of co-expression networks on an individual donor basis, which vastly reduces the number of samples required to identify SNPs altering co-expression relationships. This enabled us to study whether SNPs showing *cis*-eQTL effects also affect the co-expression relationship of the *cis*-eQTL genes with other genes, which we further define as 'co-expression QTLs'. We confined our analysis to the most abundant cell type ($CD4^+$ T cells), and calculated the co-expression between individual pairs of genes using Spearman's rank correlation. We restricted the analysis to the 145 *cis*-eQTL genes identified in $CD4^+$ T cells (Table 1), thereby increasing the likelihood of finding co-expressed genes that are modulated by the same genetic variant. Out of these, 102 genes showed variance in gene expression within each of the 45 donors and were investigated. For two of these genes we identified significant co-expression QTLs: 93 co-expression QTLs were detected for *RPS26* and one for *HLA-B* (P-value ≤ 1.27×10^{-7}, corresponding to an eQTL-gene level FDR of 0.05). The most significant interaction was found for rs7297175 affecting the co-expression between *RPS26* and *RPL21* ($P = 2.70 \times 10^{-16}$) (Fig. 2a, 2b). When using a more liberal FDR of 0.10 (P-value ≤ 4.72×10^{-7}), we identified significant co-expression QTLs for three eQTL genes (Suppl. Table 4): 13 additional co-expression QTLs were found for *RPS26* and one for *SMDT1.* As a result of co-expression between genes, we cannot rule out that the 106 co-expression QTLs identified for *RPS26* are actually representing just one effect.

To assess the robustness of the identified co-expression QTLs, we tested whether they remained significant after gene expression imputation, which was used to overcome the problem that in scRNA-seq data usually many genes are undetected despite being expressed (i.e. zero-inflated expression). Several computational strategies have been developed to do this.[22-24] However, most current methods are either computationally too demanding for large datasets like ours[23], or cannot sufficiently impute the 94.1% zero values present in our dataset[24]. To overcome this, we used MAGIC[22], a method that imputes gene expression levels for nearly every gene. To prevent that imputation removes effects of genetic differences between donors or cell types, we performed imputation for each donor separately and again

only for $CD4^+$ T cells (see Data availability). In general, imputation worked well, but in some circumstances artifacts were introduced (Suppl. Fig. 3). Therefore, we only used the imputed gene expression data to determine whether the co-expression QTLs, identified prior to imputation, remained significant after imputation (Suppl. Table 4). For the three eQTL genes that were involved in a co-expression QTL, two out of three top co-expression QTLs (rs7297175 affecting the co-expression between *RPS26* and *RPL21,* $P = 3.97 \times 10^{-12}$ (Fig. 2c) and rs4147641 affecting the co-expression between *SMDT1* and *RPS3A,* $P = 2.57 \times 10^{-4}$) remained after imputation (Suppl. Table 4). Subsequently, we were able to replicate both effects in a whole blood bulk RNA-seq eQTL dataset[11] ($P = 1.69 \times 10^{-3}$ for *RPS26-RPL21* (Fig. 2d), $P = 1.59 \times 10^{-4}$ for *SMDT1-RPS3A*) (Suppl. Table 4). Interestingly, SNP rs7297175, affecting the co-expression between *RPS26* and 106 other genes, is in near perfect linkage disequilibrium with the type I diabetes (T1D) SNP rs11171739[25] ($r^2 = 0.98$). Therefore, the numerous co-expression QTLs for *RPS26* may shed new light on *RPS26* and its link with T1D. This interaction effect was also observed in other cell types (Suppl. Fig. 4), indicating it is not cell type-specific. In addition, various analyses were performed to rule out potential technical confounders (see Online Methods).

The co-expression QTL analysis as outlined above highlights another advantage of scRNA-seq data; with PBMCs from only 45 donors, we could identify effects that would otherwise only become apparent in large-scale (2,116 samples) bulk RNA-seq eQTL datasets[11]. Due to Simpson's paradox[26], it may occur that when looking at all individuals together, the interaction between two genes does not show a correlation, while each of the individuals separately do show a correlation. So, even though the effect may be observed in bulk RNA-seq data, the true correlation will only be revealed using scRNA-seq data.

The eQTL and co-expression QTL analyses performed in this study show the benefit of scRNA-seq data for linking genetic variation to gene expression regulation. In addition to these analyses, we expect scRNA-seq data to offer many other opportunities for selecting cells of interest for eQTL and co-expression QTL analysis. For example, one could use the intercellular variation within scRNA-seq data to group cells along the cell cycle[13], along a differentiation path[27] or along a response to an environmental stimulus[28]. By doing so, one might identify eQTLs or co-expression QTLs that are influenced by cell cycle phase, differentiation or environmental status.

In conclusion, this proof of concept study shows the feasibility of using scRNA-seq data for eQTL and gene-gene interaction analysis. The identified eQTLs and co-expression QTLs replicated well with earlier reported whole blood RNA-seq data. Moreover, we extended the list of genes known to be under genetic control or specified the cell type in which the effect is most prominent. Finally, several SNPs were linked to modulation of gene co-expression, implying that gene regulatory networks can be highly personal. We expect that larger single-cell eQTL datasets will enable the identification of many cell type-specific eQTLs and genetic variants that affect regulatory network relationships.

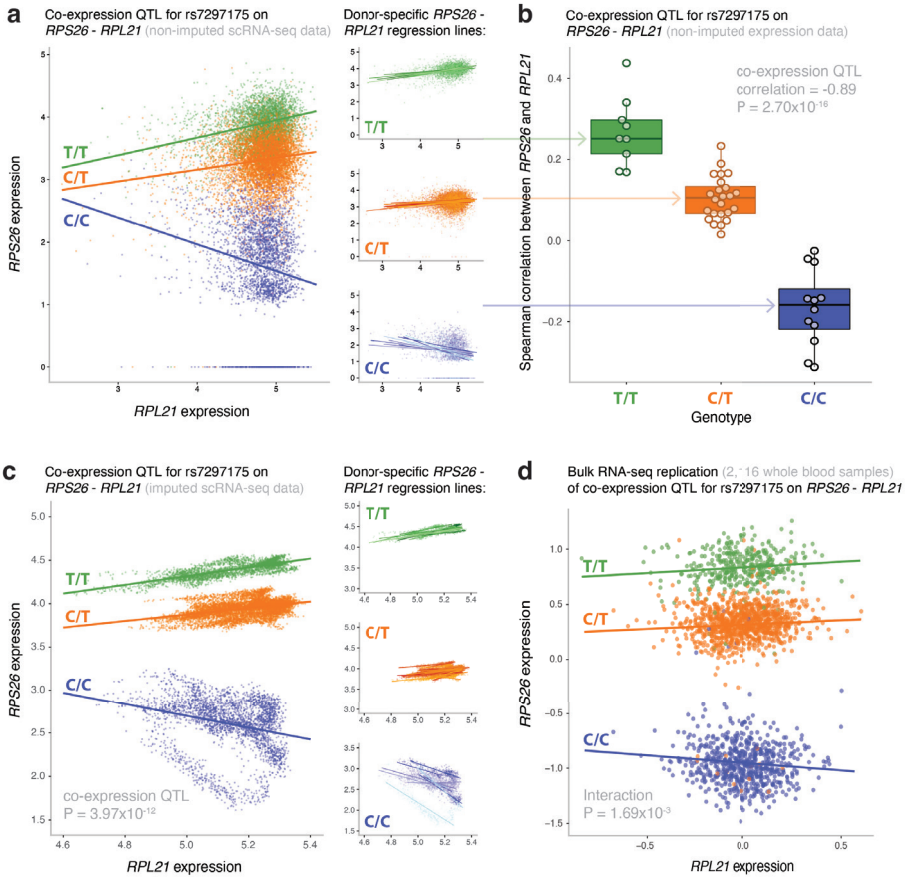


Figure 2. Most significant co-expression QTL in the CD4⁺ T cells. (**a**) The non-imputed expression of *RPS26* and *RPL21* of all individual CD4⁺ T cells colored by genotype (left panel) and stratified per SNP rs7297175 genotype (right panels). Genotype- and donor-specific regression lines are shown in the left and right panel, respectively. Each data point represents a single cell. The nominal P-value is given for the co-expression QTL. (**b**) The Spearman's rank correlation coefficient (r) between *RPS26* and *RPL21* expression stratified by SNP rs7297175 genotype in the CD4⁺ T cells per donor. Each data point represents a single donor. Box plots show the median, the first and third quartiles, and 1.5 times the interquartile range. The nominal P-value is given for the co-expression QTL. (**c**) The imputed expression of *RPS26* and *RPL21* of all individual CD4⁺ T cells colored by genotype (left panel) and stratified per SNP rs7297175 genotype (right panels). Genotype- and donor-specific regression lines are shown in the left and right panel, respectively. Each data point represents a single cell. (**d**) The expression of *RPS26* and *RPL21* of whole blood bulk RNA-seq samples colored by SNP rs7297175 genotype. Genotype-specific regression lines are shown. Each data point represents a single bulk RNA-seq sample. The nominal P-value is given for the interaction effect.

Acknowledgements

We are very grateful to all the volunteers who participated in this study. Moreover, we thank J. Dekens for arranging informed consent and contact with LifeLines. We thank A. Maatman and M. Platteel for their assistance in the lab. M.S and L.F. are 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 L.F. is supported by an ERC Starting Grant, grant agreement 637640 (ImmRisk). 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).

Author contribution

MW generated the scRNA-seq data. MW, HB and DV performed bioinformatics and statistical analyses. PD and BIOS consortium performed replication of co-expression QTLs. MW and LF designed the study and wrote the manuscript. MS and the LLD consortium provided biomaterials, genotype data and computational resources. All authors discussed the results and commented on the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Ethics approval and consent to participate

The LifeLines DEEP study was approved by the ethics committee of the University Medical Centre Groningen, document number METC UMCG LLDEEP: M12.113965. All participants signed an informed consent form prior to study enrollment. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

1. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7-24 (2012).
2. Westra, H. J. *et al*. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238-1243 (2013).
3. Brown, C. D., Mangravite, L. M. & Engelhardt, B. E. Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genet.* **9**, e1003649 (2013).
4. Fairfax, B. P. *et al*. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502-510 (2012).
5. Fu, J. *et al*. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.* **8**, e1002431 (2012).
6. Kasela, S. *et al*. Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+ versus CD8+ T cells. *PLOS Genetics* **13**, e1006643 (2017).
7. Naranbhai, V. *et al*. Genomic modulators of gene expression in human neutrophils. *Nat. Commun.* **6**, 7545 (2015).
8. Ishigaki, K. *et al*. Polygenic burdens on cell-specific pathways underlie the risk of rheumatoid arthritis. *Nat. Genet.* (2017).
9. Westra, H. *et al*. Cell Specific eQTL Analysis without Sorting Cells. *PLOS Genetics* **11**, e1005223 (2015).
10. Venet, D., Pecasse, F., Maenhaut, C. & Bersini, H. Separation of samples into their constituents using gene expression data. *Bioinformatics* **17 Suppl 1**, S279-87 (2001).
11. Zhernakova, D. V. *et al*. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139-145 (2017).
12. Villani, A. C. *et al*. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, 10.1126/science.aah4573 (2017).
13. Wills, Q. F. *et al*. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat. Biotechnol.* **31**, 748-752 (2013).
14. Tigchelaar, E. F. *et al*. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772-2014-006772 (2015).
15. Zhernakova, D. V. *et al*. DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. *PLoS Genet.* **9**, e1003594 (2013).
16. Macosko, E. Z. *et al*. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
17. Chen, L. *et al*. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e24 (2016).
18. Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. *PLOS Genetics* **9**, e1003486 (2013).
19. Sul, J. H., Han, B., Ye, C., Choi, T. & Eskin, E. Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genet.* **9**, e1003491 (2013).
20. Duong, D. *et al*. Applying meta-analysis to genotype-tissue expression data from multiple tissues to identify eQTLs and increase the number of eGenes. *Bioinformatics* **33**, i67-i74 (2017).
21. Knowles, D. A. *et al*. Allele-specific expression reveals interactions between genetic variation and

environment. *Nat. Methods* **14**, 699-702 (2017).

22. van Dijk, D. *et al*. MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *bioRxiv* (2017).
23. Huang, M. *et al*. Gene Expression Recovery For Single Cell RNA Sequencing. *bioRxiv* (2017).
24. Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9**, 997-018-03405-7 (2018).
25. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).
26. Simpson, E. H. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society.Series B (Methodological)* **13**, 238-241 (1951).
27. Trapnell, C. *et al*. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381-386 (2014).
28. Shalek, A. K. *et al*. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363 (2014).
29. Zheng, G. X. Y. *et al*. Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049 (2017).

Online methods

Isolation and preparation of PBMCs

Whole blood of 47 donors from the general population Lifelines Deep (LLD) cohort[14] was drawn into EDTA-vacutainers (BD). Within 2h, peripheral blood mononuclear cells (PBMCs) were isolated using Cell Preparation Tubes with sodium heparin (BD). For all procedures, PBMCs were kept in RPMI1640 supplemented with 50 µg/mL gentamicin, 2 mM L-glutamine and 1 mM pyruvate. Isolated PBMCs were cryopreserved in RPMI1640 containing 40% FCS and 10% DMSO. Within one month, PBMCs were further processed for scRNA-seq. First, cells were thawed in a 37°C water bath until almost completely thawed, after which the cells were slowly washed in warm medium. After washing, cells were resuspended in medium and incubated for 1h in a 5° slant rack at 37°C in a 5% CO_2 incubator. After this 1h resting period, cells were washed twice in medium supplemented with 0.04% bovine serum albumin. Cells were counted using a haemocytometer and cell viability was assessed by Trypan Blue. Eight, sex-balanced sample pools were prepared each containing 1750 cells/donor from 6 (or 5) donors (10,500 cells).

Single-cell library preparation and sequencing

Single cells were captured using the 10X Chromium controller (10X Genomics) according to the manufacturer's instructions (document CG00026), and as previously described.[29] Each sample pool was loaded into a different lane of a 10X chip (Single cell chip kit, 120236). cDNA libraries were generated using the Single Cell 3' Library & Gel Bead kit version 2 (120237) and i7 Multiplex kit (120262) in line with the company's guidelines. These libraries were sequenced using a custom program (27-9-0-138) on 8 lanes of an Illumina HiSeq4000 using a 75bp paired-end kit, per GenomeScan (Leiden, the Netherlands) sequencing guidelines. In total, 28.855 cells were captured and sequenced to an average depth of 74k.

Alignment and initial processing of sequence-data

CellRanger v1.3 software with default settings was used to demultiplex the sequencing data, generate FASTQ files, align the sequencing reads to the hg19 reference genome, filtering of cell and UMI (unique molecular identifier) barcodes, and counting gene expression per cell (see Data availability).

Demuxlet algorithm: demultiplexing samples per lane and doublet detection

Genotypes of the LLD-samples were previously generated[14] and were phased using Eagle v2.3[30] and imputed with the HRC-reference panel[31] using the Michigan Imputation Server[32]. As genotype data of each donor (except 2) was available, we could

use the Demuxlet method[33] that uses variable SNPs between the pooled individuals to determine which cell belongs to which individual and to identify doublets (two cells encapsulated in a single droplet by the 10X Chromium controller).

To determine how well every genotype matches each cell, a likelihood score was calculated by the formula:

$$L_c(s) = \prod_{v=1}^{V} \left[\sum_{g=0}^{2} \left\{ \prod_{i=1}^{d_{cv}} \left(\sum_{e=0}^{1} \Pr(b_{cvi} | g, e) \right) P_{sv}^{(g)} \right\} \right].$$

Here, *c* is the cell, *s* is the individual, *v* are the unique genetic variants (SNPs) found on the reads of the cell, d_{cv} the number of unique reads overlapping with the v^{th} variant from the c^{th} cell. b_{cvi} is the variant-overlapping base call from the i^{th} read, representing reference (R), alternate (A), and other (O) alleles respectively. e_{cvi} is a latent variable indicating whether the base call is correct (0) or not (1) and finally *g* is the true genotype. This likelihood score was calculated by taking into account the genotype probabilities of a sample at all known SNPs, the variant-overlapping base calls with base quality (Phred quality score) > 15, and a probability that the base was not called correctly, which is fixed at 0.001. In this way, for each pool of cells, the genotype within this pool with the highest likelihood was assigned as the most likely person the cell belonged to.

To identify doublets, likelihoods for a 50/50 ratio of all possible combinations of two genotypes were calculated, similarly as for singlets but now considering two genotypes at the same time. To consider a mix of genotypes from two individuals, the following formula was used:

$$L_c(s_1, s_2, \alpha) = \prod_{v=1}^{V} \left[\sum_{g_1, g_2} \left\{ \prod_{i=1}^{d_{cv}} \left(\sum_{e=0}^{1} (1-\alpha) \Pr(b_{cvi} | g_1, e) + \alpha \Pr(b_{cvi} | g_2, e) \right) P_{sv}^{(g_1)} P_{sv}^{(g_2)} \right\} \right]$$

Here, s_1 and s_2 are the two individuals, g_1 and g_2 the corresponding true genotypes and α is the expected proportion of the SNPs in every cell for each of the individuals. An α of 0.5 was consistently used, assuming a 50/50 ratio. The maximum likelihood in the mixed-genotype case was divided by the maximum likelihood in the singlet case to obtain a likelihood ratio. If this ratio was less than 1/t for some number t, the cell was assigned to be a singlet of the sample corresponding to the maximum singlet likelihood. If the ratio was greater than t, the cell was assigned to be a doublet. When the ratio was in between 1/t and t, the cell was called inconclusive: no confident call could be made from which sample(s) the cell originated. The decision boundary factor t was fixed at 2. In theory, if there are n samples in a lane, (n – 1)/n doublets can be identified using the Demuxlet algorithm, because doublets from the same individual (1/n) cannot be identified. Further details of the algorithm can be found in Kang et al.[33]

Using the Demuxlet algorithm, we could confidently assign the majority (99.8%) of cells to one of the individual donors (singlets) or to two different donors (doublets) (Suppl. Fig. 1a, Suppl. Table 5). Remarkably, in two out of eight sample pools, no cells were assigned to one of the six donors within the pool. Moreover, the detected

doublet rate in those sample pools was abnormally high (17.5% and 21.1%, while 3-4% was expected) (Suppl. Table 5). This is most probably due to a sample mix-up in the lab which resulted in an artificially high doublet rate. Since the genotypes of these two mixed-up samples were not available, those samples were excluded from the analysis (marked as “doublet”).

Two additional tests were performed to confirm the correct assignment of cells using Demuxlet. First, we determined what would happen if the cells did not match with their genotypes by taking six random genotypes not present in the sample pool itself. This resulted in 0.02% of the cells being a singlet, 0.03% being inconclusive and 99.95% being a doublet. Second, the number of reads mapping to the Y-chromosome was determined for the singlets of each donor. Cells belonging to a female donor showed (almost) no Y-reads (mismapping reads[34] may explain the few sporadic Y-reads), whereas the majority of cells from male donors did (Suppl. Fig. 1b). So, the correct gender for each of the donors could be confirmed by looking at the number of Y-reads. These tests indicated that the Demuxlet method is correctly assigning cells to their respective donor and is suitable for detecting sample swaps.

Cell type classification

Version 1.4 of the R package Seurat[16] was used to determine the cell types using the raw UMI counts from CellRanger. First, all genes that were not detected in ≥3 cells were removed. Cells in which >5% of the UMIs mapped to the mitochondrial-encoded genes were discarded as this can be a marker of bad quality cells; broken cells will leak cytoplasmic RNA, while the mitochondrial RNA content is retained inside the mitochondria.[35] Also, cells expressing >3,500 genes were considered outliers and discarded (Suppl. Fig. 1c, Suppl. Table 6). Finally, all cells that were marked as doublet or inconclusive by the Demuxlet method were discarded. Supplementary figure 1d shows a t-Distributed Stochastic Neighbor Embedding (t-SNE) plot[36] in which all cells failing the above QCs are visualized. Library size normalization was performed on the UMI-collapsed gene expression for each barcode by scaling the total number of transcripts per cell to 10,000. The data was then log-2 transformed. In total, 25,291 cells and 19,723 genes (average of 1147 detected genes/cell) (see Data availability) were used in the cell type determination.

Linear regression was used to regress out the total number of UMIs and the fraction of mitochondrial transcript content per cell. The variable genes were identified using Seurat’s MeanVarPlot function which places all genes in 20 bins based on their average expression (the mean of non-zero values) and calculates the dispersion (standard deviation of all values) within each bin. Standard parameters were used except the bottom gene expression cut-off (x.low.cutoff) was set to 0 and the bottom dispersion cut-off (y.cutoff) was set to 1.0, resulting in the identification of 1,090 genes. These 1,090 variable genes were used in the principle component analysis (PCA). The first 16 principal components were used for cell clustering using Seurat’s

FindCluster function (default parameters, resolution 1.2) and a t-SNE plot was used to visualize this. Based on known marker genes and differentially expressed genes per cluster (found using Seurat's FindMarkers function), we could assign 11 cell types to the clusters, including some smaller subcell types (Suppl. Fig. 2a, 2b, Suppl. Table 7). The smallest cluster we could detect consisted of plasma cells, making up 0.3% of the total PBMC population.

eQTL analysis

To find the association between genotype and expression per cell type, genome-wide *cis*-eQTL analysis for 18,264 genes (only autosomal genes, gene expressed in at least 3 cells within the total dataset and in at least 1 cell within the cell type queried, within 100 kb distance of the SNP and the gene midpoint, MAF>0.1, call rate >0.95, a Hardy-Weinberg equilibrium P value of >0.001) was performed using our previously described eQTL pipeline, version 1.2.4F (Suppl. Table 2, see Data availability).[11] To assure sufficient power, cell types were merged to a more general classification: $CD4^+$ T cells, $CD8^+$ T cells, NK cells ($CD56^{dim}$ $CD16^+$ and $CD56^{bright}$ $CD16^{+/-}$), monocytes ($CD14^{bright}$ $CD16^-$ classical, cMonocyte, and $CD14^{dim}$ $CD16^+$ non-classical, ncMonocyte), B cells and DCs ($CD1C^+$ myeloid, mDC and plasmacytoid, pDC). The mean expression per gene per cell type per donor was calculated on the normalized (Z-score transformed) expression and used as input for the eQTL analysis. eQTLs were mapped using Spearman's rank correlation coefficient on imputed genotype dosages. eQTLs were considered significant at a gene-level FDR of 0.05. To control the FDR at 0.05 we used the permutation method described previously by us.[2] Here we permute the link between the genotypes and expression data and create an overall null distribution using all genes. We performed in total 10 permutations and use for each gene the total null distribution of all genes to determine a gene-level FDR: during FDR estimation only the most significant SNP per gene is used, both for the real analysis and for each of the permutations.

Concordance and detection

Concordance with previously found independent top eQTLs from a whole blood DeepSAGE (3'-end transcriptomics)[15] and RNA-seq study[11] were computed. For this, the mean expression per gene per individual of all cells was calculated and the *cis*-eQTL mapping was confined to the independent top eQTLs found in the DeepSAGE[15] or RNA-seq study[11]. Subsequently, detection of the same SNP-gene combination and concordance (with same allelic direction) were assessed between the significant top effects (Suppl. Table 1). Vice-versa, we also determined how many of the 379 top eQTLs in our scRNA-seq dataset could be detected and with which allelic direction within the whole blood RNA-seq study[11]. Similarly, we assessed detection rate and concordance with two studies containing RNA-seq data of purified cell types: Kasela et al. performed eQTL analysis on purified $CD4^+$ and $CD8^+$ T cells[6], whereas the data

from the Blueprint consortium contains purified CD14+ monocytes and naïve CD4+ T cells[17] (Suppl. Table 2, 3). Moreover, for the eQTLs that were specifically detected in the cMonocytes and not the ncMonocytes (Fig. 2d), detection rate and concordance were determined using the RNA-seq data of the purified CD14+ monocytes from the Blueprint consortium[17].

Single-cell gene expression imputation

To overcome the zero-inflated expression, the computational method MAGIC[22] was used to impute practically all values of genes with at least some expression. MAGIC imputation (using the following parameters: 20 PCs, t=4, k=9, ka=3, ε=1) was performed per donor separately and only in the CD4+ T cells (see Data availability). The effect of MAGIC imputation was validated by comparing the co-expression of typical cell type-specific marker genes (Suppl. Fig. 3).

Co-expression QTL analysis

For every individual, a Spearman's rank correlation coefficient was calculated between the expression of the *cis*-eQTL gene and all other genes. Given the large zero-inflation of scRNA-seq data, we only tested those 7,975 genes that showed variance in expression for each of the 45 samples. As a consequence we could study 102 eQTL genes out of the 145 unique genes that showed a significant *cis*-eQTL effect in CD4+ T-cells. For each of these combinations, a weighted linear model was used (*co-expression ~ genotype*, where weight is $\sqrt{cellCount}$), in which the explained variable is a Spearman correlation coefficient that describes the co-expression between the two genes and the genotype is the predictor and the weights are the square root of the number of CD4+ T cells within the given sample (Suppl. Fig. 5).

In order to determine for how many *cis*-eQTL genes we had identified a significant co-expression QTL we performed 100 permutations (see Data availability). For the real analysis we denoted for each of the tested 102 eQTL genes what was the most significant co-expression QTL P-Value (Suppl. Table 4). For each permutation we shuffled the genotype identifiers and reran the above analysis and also determined for each of the 102 eQTL genes what was the most significant co-expression QTL P-Value (see Data availability). This subsequently enabled us to calculate an eQTL-gene level FDR[2] (using exactly the same multiple testing correction procedures as we employ for the detection of *cis-eQTLs*, see paragraph "eQTL analysis"). An eQTL gene-level FDR of 0.05 was considered significant, i.e. the p-value threshold of the most significant co-expression QTL p-values at which 5% of the co-expression QTLs are significant in the permuted compared to the real data.

All significant co-expression QTLs were discovered using non-imputed gene expression data. We then assessed whether these co-expression QTLs were also significant when using the MAGIC-imputed gene expression data. Subsequently, we

tested whether these co-expression QTLs replicated using a large whole blood bulk RNA-seq dataset[11] (Suppl. Table 4). We finally attempted to falsify the observed co-expression QTL for rs7297175 on the co-expression between *RPS26* and *RPL21*, by checking the following potential confounders:

- Potential sequence homology: no evidence was found for sequence homology between *RPS26* and *RPL21.*
- Genotype-dependent mapping problems of RNA sequence reads: no evidence was found that the *RPS26 cis*-eQTL SNP rs7297175 has any SNP proxies (r^2 > 0.8) that are coding and that map within *RPS26*. As such this suggests that potential genotype-dependent mapping biases of sequence-reads are unlikely.
- Multi-mapping of RNA sequence reads: no differences were found between individuals with regards to the amount of sequence reads that were discarded due to multi-mapping of sequence reads to *RPS26*.
- Unexpected *trans*-eQTL on *RPL21*: no evidence was found that the *RPS26 cis*-eQTL SNP rs7297175 is affecting the expression of *RPL21* in *trans*.
- Genotype-dependent subcell-type composition effects: the *RPS26-RPL21* co-expression QTL is unlikely the result of a subcell-type within the $CD4^+$ T cell population, as this co-expression QTL effect is also significant within $CD8^+$ T cells, within monocytes and within NK cells (Suppl. Figure 4).

Data availability

Raw gene expression counts, MAGIC imputed $CD4^+$ T cell gene expression, and eQTL and co-expression QTL summary statistics can be found under “Supplementary Data” at the website accompanying this paper (https://molgenis58.target.rug.nl/scrna-seq/).

Processed (deanonimyzed) single-cell RNA-seq data, including a text file that links each cell barcode to its respective donor, has been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001002560. Gene expression and genotype data can be obtained and requested by filling in a single and short web form at https://molgenis58.target.rug.nl/scrna-seq/. This form is subsequently reviewed by a single Data Access Committee, who will be able to approve access to both the raw gene expression and genotype data within 5 working days (during the holiday season there might be a slight delay). Once the proposed research is approved, access to the relevant gene expression or genotyped data will be free of charge. Access to the genotype and gene expression data is facilitated via the Lifelines workspace and the EGA, respectively. Sample metadata (age, gender, processing batch) is presented in Suppl. Table 8.

Code availability

The original R code for Seurat[16] (https://github.com/satijalab/seurat), Demuxlet[33] (https://github.com/statgen/demuxlet), MAGIC[22] (https://github.com/pkathail/magic) and our in-house eQTL pipeline[2] (https://github.com/molgenis/ systemsgenetics/ tree/master/eqtl-mapping-pipeline) can be found at Github. All custom-made code is made available via GitHub (https://github.com/molgenis/scRNA-seq).

Supplementary material

Supplementary material is available at https://doi.org/10.1038/s41588-018-0089-9

References

30. Loh, P. R. *et al*. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443-1448 (2016).

31. McCarthy, S. *et al*. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279-1283 (2016).

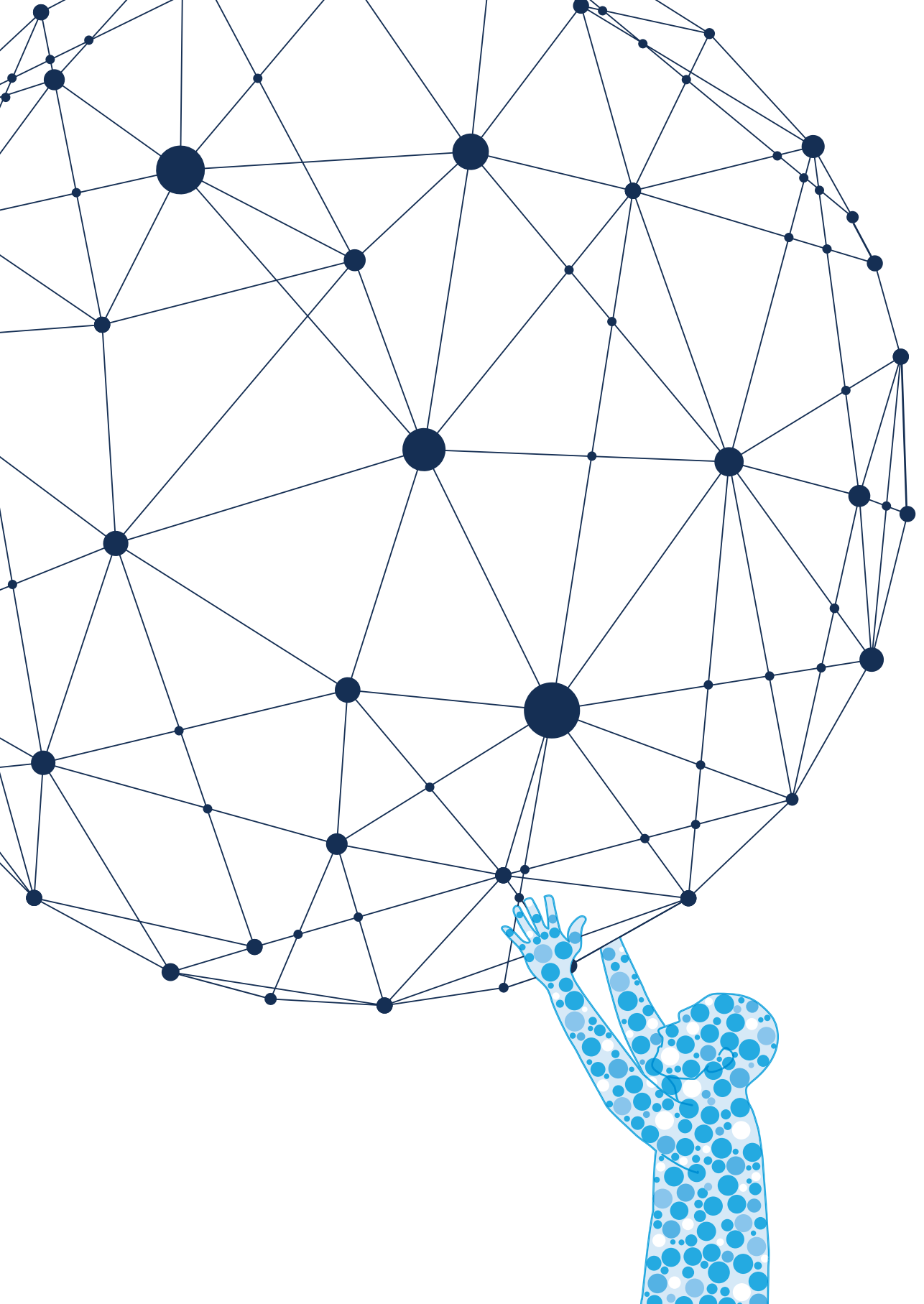
32. Das, S. *et al*. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284-1287 (2016).

33. Kang, H. M. *et al*. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* (2017).

34. Rosser, Z. H., Balaresque, P. & Jobling, M. A. Gene conversion between the X chromosome and the male-specific region of the Y chromosome at a translocation hotspot. *Am. J. Hum. Genet.* **85**, 130-134 (2009).

35. Ilicic, T. *et al*. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17**, 29-016-0888-1 (2016).

36. van de Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605 (2008).



CHAPTER 4

Integrating GWAS with bulk and single-cell RNA-sequencing reveals a role for *LY86* in the anti-*Candida* host response

Dylan H. de Vries[1], Vasiliki Matzaraki[1,2], Olivier B. Bakker[1], Harm Brugge[1], Harm-Jan Westra[1], Mihai G. Netea[2,3], Lude Franke[1,#,*], Vinod Kumar[1,2,#], Monique G.P. van der Wijst[1,#,*]

[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 (RCI), Radboud University Medical Center, Nijmegen, the Netherlands.
[3] Human Genomics Laboratory, Craiova University of Medicine and Pharmacy, Craiova, Romania.
[#] Shared last authorship
[*] Correspondence to MW: m.g.p.van.der.wijst@umcg.nl or LF: lude@ludesign.nl

Published in PLOS Pathogens. https://doi.org/10.1371/journal.ppat.1008408

Abstract

Candida bloodstream infection, i.e. candidemia, is the most frequently encountered life-threatening fungal infection worldwide, with mortality rates up to almost 50%. In the majority of candidemia cases, *Candida albicans* is responsible. Worryingly, a global increase in the number of patients who are susceptible to infection (e.g. immunocompromised patients), has led to a rise in the incidence of candidemia in the last few decades. Therefore, a better understanding of the anti-*Candida* host response is essential to overcome this poor prognosis and to lower disease incidence. Here, we integrated genome-wide association studies with bulk and single-cell transcriptomic analyses of immune cells stimulated with *Candida albicans* to further our understanding of the anti-*Candida* host response. We show that differential expression analysis upon *Candida* stimulation in single-cell expression data can reveal the important cell types involved in the host response against *Candida*. This confirmed the known major role of monocytes, but more interestingly, also uncovered an important role for NK cells. Moreover, combining the power of bulk RNA-seq with the high resolution of single-cell RNA-seq data led to the identification of 27 *Candida*-response QTLs and revealed the cell types potentially involved herein. Integration of these response QTLs with a GWAS on candidemia susceptibility uncovered a potential new role for *LY86* in candidemia susceptibility. Finally, experimental follow-up confirmed that *LY86* knockdown results in reduced monocyte migration towards the chemokine MCP-1, thereby implying that this reduced migration may underlie the increased susceptibility to candidemia. Altogether, our integrative systems genetics approach identifies previously unknown mechanisms underlying the immune response to *Candida* infection.

Author summary

Candida albicans is a fungus that can cause a life-threatening infection in individuals with an impaired immune system. To improve the prognosis and treatment of patients with such an infection, a better understanding of an individual's immune response against *Candida* is required. However, small patient group sizes have limited our ability to gain such understanding. Here we show that integrating many different data layers can improve the sensitivity to detect the effects of genetics on the response to *Candida* infection and the roles different immune cell types have herein. Using this approach, we were able to prioritize genes that are associated with an increased risk of developing systemic *Candida* infections. We expand on the gene with the strongest risk association, *LY86*, and describe a potential mechanism through which this gene affects the immune response against *Candida* infection. Through experimental follow-up, we provided additional insights into how this gene is associated with an increased risk to develop a *Candida* infection. We expect that our approach can be generalized to other infectious diseases for which small patient group sizes have restricted our ability to unravel the disease mechanism in more detail. This will provide new opportunities to identify treatment targets in the future.

Introduction

Candida albicans (*C. albicans*) is an opportunistic fungus colonizing the skin and/or mucosae of approximately 70% of the population (1). Disruption of the mucosal barrier or a compromised immune system of the host can increase susceptibility to *Candida* infections. This makes it the most common cause of hospital-acquired invasive fungal infections globally (2), with high mortality rates between 33% and 46% (3,4). The most common form of invasive candidiasis occurs in the blood, known as candidemia (2). Despite the severity of candidemia and its accompanying research interest, the ability to improve the outcomes for affected individuals has stagnated in recent years. Adjuvant immunotherapy has been suggested as an important strategy to improve patient outcomes, but to implement this a better understanding of the immune response to *Candida* is required (5,6). As genetics have a great impact on an individual's immune response (7), knowledge on its impact to the anti-*Candida* response will be important as well for the implementation of such therapies.

Genome-wide association studies (GWAS), linking genetic variants to disease risk, have been a commonly used approach to increase disease understanding. However, in the context of candidemia and other infectious diseases, conducting a GWAS is challenging due to the limited size of patient cohorts (8). Moreover, GWAS studies provide limited insight into the underlying biology of how these genetic variants are linked to *Candida* infection susceptibility. Thus additional approaches are required.

Integrative strategies that combine different molecular datasets in the context of *Candida* infection have been suggested as alternative approaches to prioritize cell types, genes and pathways. These can then be used for follow-up functional studies to better understand candidemia susceptibility. For instance, Smeekens et al. integrated gene expression array data of *Candida*-stimulated PBMCs with genetic information and cytokine measurements from both healthy volunteers and patients with increased susceptibility to *Candida* infections (9). Using this integrative approach, they identified the interferon pathway as being a crucial host response pathway against *Candida* infection. In a follow-up study, the additive value of integrating multiple molecular datasets became even more apparent as suggestive genetic associations together with transcriptomic data could prioritize novel pathways implicated in candidemia susceptibility, including the complement and hemostasis pathways (10).

However, further integration is required to understand the mechanism through which genetic variants lead to increased candidemia-susceptibility. These disease-associated variants can be linked to effects on gene expression levels through so-called expression quantitative trait loci (eQTL) analysis. Since disease-associated genetic variants are often regulated in a context-specific manner (11), such eQTL analyses should be performed in such a way that the context-specific nature, i.e. pathogen- and cell type-specificity, can be revealed. With the advent of single-cell RNA-sequencing (scRNA-seq) it now becomes possible to profile the expression of tens of thousands of individual cells at the same time in an unbiased manner (12).

This now allows capturing the context-specific nature of disease-associated genetic variants with increased resolution, while retaining the intercellular dynamics.

Here, we used an integrative approach combining GWAS with bulk and scRNA-seq transcriptomic analyses on *Candida*-stimulated and RPMI control peripheral blood mononuclear cells (PBMCs). By leveraging the sensitivity of bulk RNA-seq data with the context-specific information acquired from scRNA-seq, this integrative approach further improves our understanding of the host response against *Candida.*

Results and discussion

Cell type-specific transcriptional response to *Candida albicans*

To reveal the cell type-specific immune response against *Candida*, scRNA-seq analysis was performed on PBMCs from 6 individuals that were stimulated with *Candida* or RPMI control for 24h. After QC, a total of 15,085 cells remained, of which 7,925 cells were RPMI control and 7,160 cells were *Candida*-stimulated. These cells were classified as one of the following six immune cell types: B cells, $CD4^+$ T cells, $CD8^+$ T cells, monocytes, natural killer (NK) cells or plasmacytoid dendritic cells (pDC).

As pathogen-stimulation can potentially affect the cellular state or induce active recruitment of specific cell types, we first determined whether *Candida*-stimulation affected the relative abundance of immune cell types. At baseline, the largest differences in relative abundance of individual cell types varied between 1.6-fold for the $CD4^+$ T cells up to 8.3-fold for the $CD8^+$ T cells (Suppl. Fig. 1). However, upon stimulation these abundances remained constant within an individual. Overall, $CD4^+$ T cells were the most abundant cell type (61.2%), whereas pDCs were observed the least (1.3%) (Table 1). Even though changes in relative abundances were not detected, we cannot exclude that this is not happening *in vivo*, as our *in vitro* stimulation of PBMCs does not allow detection of active recruitment. Active recruitment of monocytes towards the lymph nodes is part of the host immune response towards *Candida,* as recently shown in mice (13).

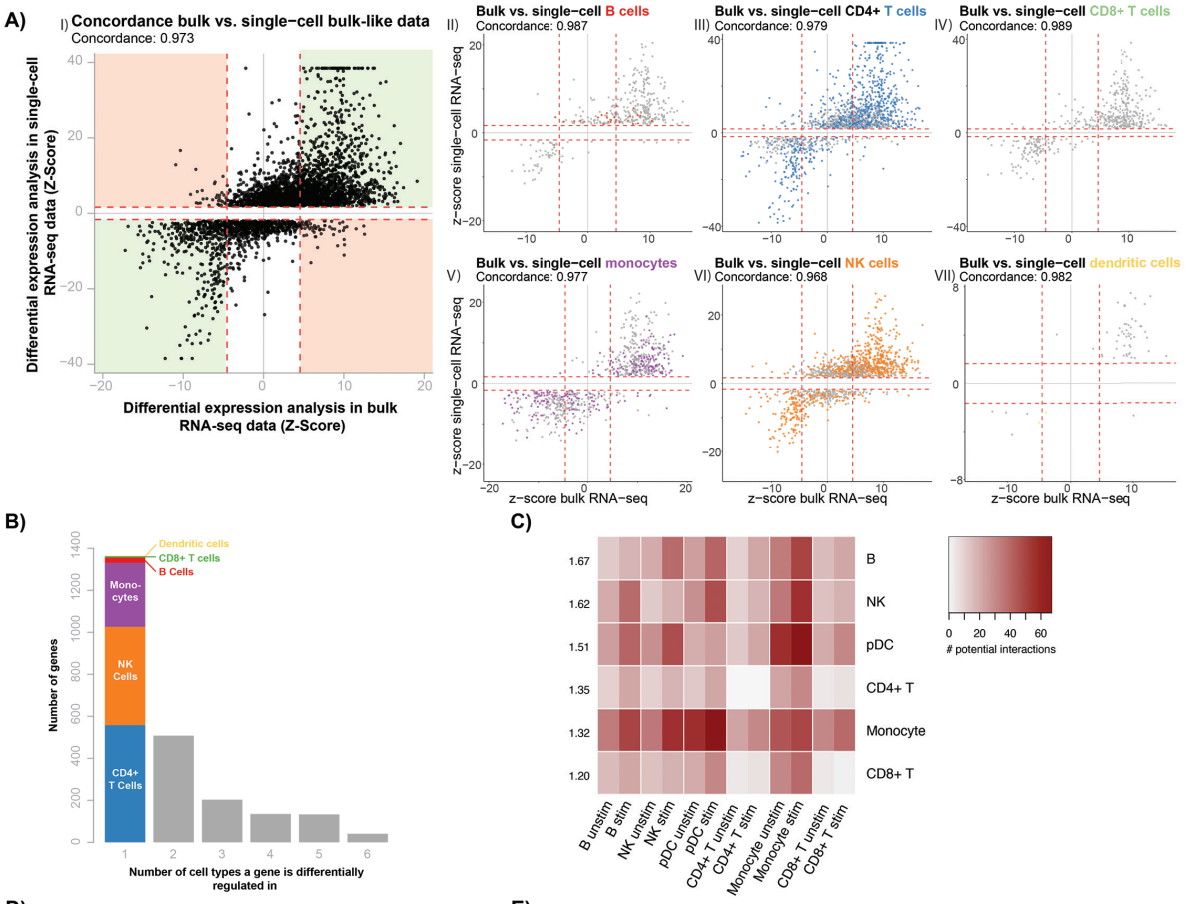
Table 1. Differentially expressed genes per cell type within PBMC single-cell RNA-seq data

Cell type	# cells	# DE genes	# up-regulated	# down-regulated
CD4+ T	9,236	1,459	1,095	364
CD8+ T	2,300	453	334	119
NK	1,807	1,313	927	386
B	789	392	329	63
Monocyte	757	767	418	349
pDC	196	56	49	7

DE, differentially expressed; NK, natural killer cell; pDC, plasmacytoid dendritic cell; PBMC, peripheral blood mononuclear cell.

Secondly, we identified differentially expressed (DE) genes upon stimulation per cell type separately as well as in all cells together (bulk-like) performing DE analysis with MAST (14). This analysis identified a total of 2,384 DE genes in the individual cell types and 3,568 DE genes in the bulk-like sample (Table 1, Suppl. Table 1). However, the noisiness and sparseness of single-cell data could potentially introduce artifacts in the DE analysis, resulting in false-positives (15). To determine the extent to which this occurs, we compared the DE genes identified in the scRNA-seq data with their differential response in a previously described bulk RNA-seq dataset generated from *Candida*-stimulated PBMCs isolated from 70 individuals (7). This comparison showed that 97.3% of the DE genes from the bulk-like scRNA-seq sample (Fig. 1A-I) and at least 96.8% of the DE genes from the individual cell types (Fig 1A-II-VII) could be replicated in the bulk-RNA seq data (Suppl. Table 1). Thus, the DE genes identified in scRNA-seq data reflect true biology rather than artifacts and can be used to uncover the cell type-specific immune response against *Candida*. However, please note that during this prolonged incubation of 24h, it is not possible to distinguish between direct and indirect responses upon *Candida* stimulation.

***Candida* induces large gene expression differences in CD4+ T cells, NK cells and monocytes** Continuing with the 2,384 DE genes identified in the individual cell types (Fig. 1B, Suppl. Table 2), we found that 71% of these genes are being upregulated upon stimulation. The majority of these DE genes (1,364) are only found in one cell type, of which the largest part in $CD4^+$ T cells, NK cells and monocytes (558, 468 and 304 DE genes, respectively). The remaining three cell types have very few uniquely identified DE genes, with 27, 5 and 2 DE genes for B cells, $CD8^+$ T cells and pDCs, respectively. As the power to detect DE genes for a cell type is strongly correlated with the number of cells for that particular cell type (Pearson correlation = 0.71) (Suppl. Fig. 2A), part of these differences can be attributed to differences in cell numbers (Table 1). However, even when taking this into account, a disproportionately large number of DE genes are specifically identified in the monocytes and NK cells (Suppl. Fig. 2B).


A)
I) Concordance bulk vs. single−cell bulk-like data
Concordance: 0.973
Differential expression analysis in single-cell RNA-seq data (Z-Score)
Differential expression analysis in bulk RNA-seq data (Z-Score)
II) Bulk vs. single−cell B cells
Concordance: 0.987
III) Bulk vs. single−cell CD4+ T cells
Concordance: 0.979
IV) Bulk vs. single−cell CD8+ T cells
Concordance: 0.989
V) Bulk vs. single−cell monocytes
Concordance: 0.977
VI) Bulk vs. single−cell NK cells
Concordance: 0.968
VII) Bulk vs. single−cell dendritic cells
Concordance: 0.982
z−score single−cell RNA−seq
z−score bulk RNA−seq
B)
Dendritic cells
CD8+ T cells
B Cells
Mono-cytes
NK Cells
CD4+ T Cells
Number of genes
Number of cell types a gene is differentially regulated in
C)
1.67
1.62
1.51
1.35
1.32
1.20
B
NK
pDC
CD4+ T
Monocyte
CD8+ T
B unstim
B stim
NK unstim
NK stim
pDC unstim
pDC stim
CD4+ T unstim
CD4+ T stim
Monocyte unstim
Monocyte stim
CD8+ T unstim
CD8+ T stim
potential interactions

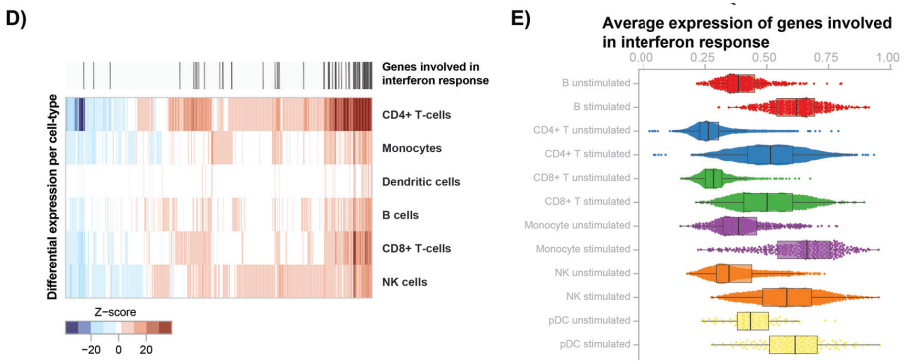


Figure 1. Single-cell RNA-seq differential expression analysis reveals the cell type-specific response to *Candida* stimulation. (A) Comparison of differentially expressed (DE) genes upon *Candida* stimulation identified in 6 individuals for whom single-cell RNA-seq (scRNA-seq) data is generated (y-axis) as compared to the effect in 70 bulk RNA-seq samples (x-axis). Each dot represents a DE gene and the dotted red lines indicate the significance thresholds. In panel I (DE genes in bulk-like scRNA-seq sample, which contains all cells from an individual), concordant DE genes are shown in the green area and discordant genes in the red area. In panels II-VII (DE genes in specific cell type), color indicates whether a DE gene is cell type-specific. (B) Bar plot showing the sharedness of DE genes across cell types. The first bar, with cell type-specific DE genes, is colored based on the cell type in which the DE gene is found. (C) Heatmap of the total number of ligand-receptor interactions between cells of the same or different cell types. Each cell type is compared to cell types of the same condition (RPMI control left, 24h *Candida*-stimulation right). Each row has a number showing the average fold enrichment in ligand-receptor pair interactions between that cell type and all cell types. (D) Heatmap of DE gene Z-scores per cell type (y-axis) for genes that are identified as DE in more than one cell type (x-axis). Red colors indicate upregulation and blue colors show downregulation upon Candida stimulation. Above the heatmap, genes found within the interferon pathway are highlighted. (E) Box plots showing the mean expression of interferon pathway-associated genes (x-axis) for each cell type and stimulation condition (y-axis). Box plots show the median, first and third quartiles, and 1.5×the interquartile range and each dot represents the expression of a single cell.

To follow-up on these findings, we determined whether the connectivity between each of the major cell types changed upon stimulation with *Candida.* For this, we calculated for each cell type their potential to interact with cells from the same or another cell type by analyzing the expression of cell type-specific receptor and ligand pairs per condition (*Candida*-stimulated and RPMI control), using the computational framework CellPhoneDB (16). This analysis revealed that especially the B cells (on average 1.67-fold increase) and NK cells (on average 1.62-fold increase) gain additional potential cell-cell interactions upon stimulation with *Candida* (Fig. 1C).

Previous studies have reached a consensus that monocytes play an important role in candidemia (17,18), but the contribution of NK cells is less clear (19,20). Interestingly, specifically in immunocompromised mice the depletion of NK cells increased the susceptibility to candidemia (21). As in humans, candidemia mainly affects immunocompromised patients, we hypothesize that NK cells are likely to play an important role in the human candidemia response as well. Through the DE and ligand-receptor expression analysis we show that, in addition to monocytes, also NK cells are strongly activated and are increasingly connected to other cells. This provides extra evidence for their importance in the immune response against *Candida*.

In addition to these unique responses, 1,020 DE genes were identified across multiple cell types, of which a core of 41 DE genes was shared between all six cell types. Of these shared DE genes, 89.8% of effects have the same direction across all responding cell types (Fig. 1D). Moreover, these shared DE genes showed the strongest differential effect upon stimulation (Fig. 1D). Pathway analysis on the core set of 41 DE genes revealed strong enrichment of the interferon pathway ($p = 10^{-22}$) (Suppl. Table 2). This is in line with previous findings in PBMC bulk expression data that showed strong differential expression of the interferon pathway upon 24h *Candida* stimulation (9). Notably, when taking the average expression of all interferon pathway-associated genes per cell, the strength of upregulation of the interferon I pathway after *Candida* stimulation is consistent across all cell types (Fig. 1E).

Identification of Candida-response QTLs using bulk RNA sequencing

In addition to identifying cell type-specific responses to *Candida* infection, we also studied the effect of genetic variants on gene expression levels before and after *Candida* stimulation using previously published bulk RNA-seq data from PBMCs (7). The rather small sample size of this study limits its predictive power, in part by the large multiple testing burden of genome-wide eQTL analysis (22). To reduce the multiple testing burden, we limited our single nucleotide polymorphism (SNP)-gene combinations to only the 16,990 top *cis* SNP-gene pairs identified in the largest eQTL meta-analysis to date (23), containing whole blood samples of 31,684 individuals. However, by confining our analysis only to previously reported *cis*-eQTLs in unstimulated blood samples, we might miss out on eQTLs that only show up

after stimulation. Nevertheless, if there is a weak effect without stimulation that is strengthened by *Candida* stimulation, restricting ourselves to previously identified top SNP-gene pairs will increase our chance to detect the eQTL effect. Using this approach, a total of 1,563 and 1,637 eQTLs were found in 72 *Candida*-stimulated and 75 RPMI control samples, respectively (Fig. 2A, Suppl. Table 3). Whilst many (44%) of these eQTLs were found both before and after stimulation, the majority of eQTLs were condition-specific (Fig 2A). By subtracting per individual and per gene the *Candida*-stimulated expression from the RPMI control expression, we also tested whether certain SNPs affected the expression of a particular gene with different effect sizes before and after stimulation. This so-called response QTL analysis was performed in the 67 individuals for which both *Candida*-stimulated and RPMI control conditions were assessed and revealed 27 response QTLs (Suppl. Table 3). Subsequently, scRNA-seq data was used to pinpoint the potential cell type in which the response QTL effects manifest themselves (Suppl. Fig. 3). Annotation of the cell type- and context-specificity of eQTLs may help to understand their involvement in human disease.

Prioritization of LY86 as a potential key driver gene for candidemia

Previously, it was shown that integrating multiple molecular datasets can help prioritize disease-relevant genes, cell types and pathways (9,10). Therefore, as a next step, we took the GWAS summary statistics of a previously published candidemia cohort of 161 cases and 152 disease-matched controls (8) and overlaid this with our 27 response QTLs (Fig. 2B). This revealed an enrichment of candidemia-susceptibility SNPs within the *Candida*-response QTL SNPs ($\lambda_{inflation}$ = 1.49) (Fig. 2C). The top enriched response QTL SNP was rs9405943 ($p=1.2 \times 10^{-3}$, OR= 0.594), and was in near perfect linkage disequilibrium with rs2103635, the SNP showing the strongest association with candidemia in this region ($P = 7 \times 10^{-4}$, OR = 0.58) ($r^2 = 0.94$) (Suppl. Fig. 4). SNP rs9405943 showed a strong effect on expression levels of *LY86* after *Candida* stimulation (β=0.58, $P= 1.5 \times 10^{-7}$), but not in the RPMI control condition (β=0.05, P = 0.68) (Fig. 3A).

The expression of *LY86* is strongly downregulated upon *Candida* stimulation in the bulk RNA-seq dataset ($P = 7.2 \times 10^{-28}$). Additionally, we see that the candidemia-risk allele *A* at rs9405943 is associated with stronger downregulation of *LY86* after stimulation. This suggests that high expression of *LY86* has a protective function against candidemia. Single-cell gene expression data shows that both B-cells and monocytes express *LY86*. However, only expression in monocytes is affected by the stimulation ($P = 1.9 \times 10^{-14}$) (Fig. 3B, 3C), suggesting that this gene contributes to candidemia susceptibility through monocytes.

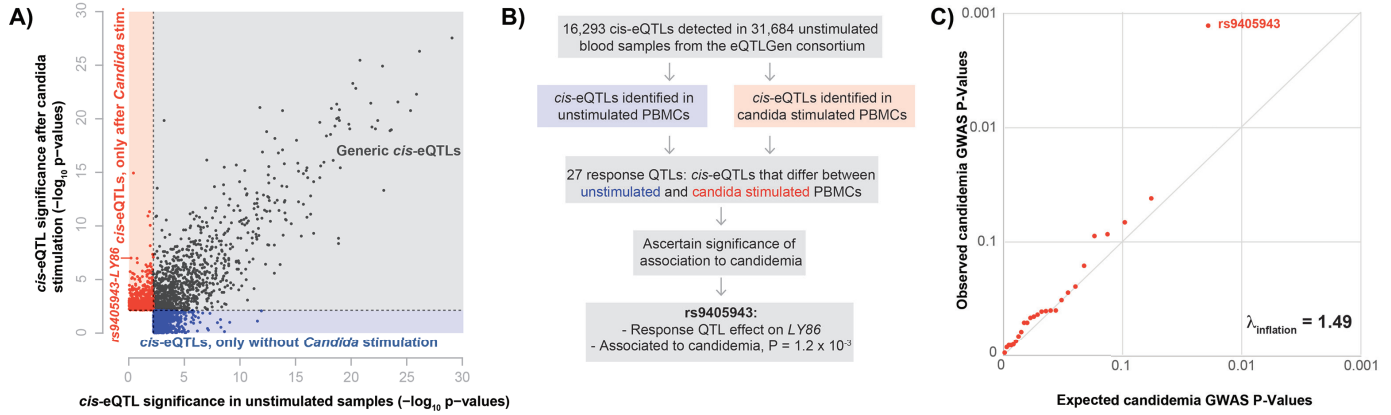


Figure 2. Integration of GWAS with eQTL analysis allows for prioritization of potential key driver genes. (A) Comparison of -log10 p-values of identified eQTLs in individuals without *Candida* stimulation (n=75) and eQTLs identified in individuals after *Candida* stimulation (n=72). Red points show eQTLs that are significant only with *Candida* stimulation, blue points show eQTLs that are significant only without *Candida* stimulation and black points depict eQTLs that are significant in both conditions. The eQTL analysis was restricted to top SNP-gene combinations identified in the eQTLGen consortium (23). (B) The performed work flow to identify potential key driver genes in Candida response. (C) QQ-plot of 27 *Candida*-response QTL SNPs in a GWAS on candidemia susceptibility, comparing expected GWAS P-values (x-axis) with observed GWAS P-values (y-axis). The dots show deviation from the expected line ($I_{inflation}$=1.49) with the strongest GWAS association found for rs9405943.

It is known that LY86 forms a complex with Toll-like receptor protein RP105 and is involved in several immune disorders (24-26). Depending on the cell type, this complex has opposite regulatory effects on TLR4 signaling (27,28); while TLR4 signaling is activated and stimulates proliferation and antibody production in B-cells, it is negatively regulated in myeloid cells. These opposite effects likely reflect the engagement of different cell type-specific co-receptors (28). While previous studies have shown the importance of the RP105/LY86 complex in mediating the TLR4-mediated innate immune response against bacterial lipopolysaccharides (LPS) (29,30), its role in the anti-*Candida* response is unknown.

In monocytes, both increased signaling activity of TLR4 and absence of RP105 are associated with downregulation of the chemokine receptor CCR2, leading to their reduced migratory capacity (25,31). Through complex formation with LY86, RP105 inhibits TLR4 signaling in monocytes (28). Therefore, we hypothesize that the rs9405943 candidemia-risk allele *A*, which lowers *LY86* expression in monocytes upon *Candida* stimulation, will decrease the migratory capacity of monocytes, which ultimately increases susceptibility to candidemia (Fig. 3D). Of note, the TLR4 signaling pathway has been shown to be involved in the innate immune responses of several microbial and fungal infections (32-35). In addition, a previous study, in which PBMCs from 8 individuals were stimulated for 24h with microbial and fungal pathogens, showed reduced expression of *LY86* after stimulation with *Mycobacterium Tuberculosis*(-1.20 fold, $p = 1.53 \times 10^{-7}$), *Borrelia* (-1.34 fold, $p = 7.31 \times 10^{-13}$), *Pseudomonas Aeruginosa* (-1.31 fold, $p = 9.45 \times 10^{-9}$) and *Streptococcus Pseudomoniae* (-1.54 fold, $p = 2.01 \times 10^{-19}$), but not *Aspergillus Fumigatus* (-0.012 fold, p = 0.98) (7). Altogether, this indicates that the differential regulation of *LY86* in monocytes, as seen in response to *Candida,* could also affect the susceptibility towards other blood-based bacterial infections.

Functional validation of the role of LY86 in monocytes

To test our hypothesized mechanism of action (Fig. 3D), we conducted experimental follow-up studies in THP-1 monocytes. As the candidemia risk allele is associated with reduced expression of *LY86*, we used siRNA knockdown of *LY86* to mimic this effect. 72h after siRNA treatment, we confirmed efficient knockdown of *LY86* (14.3-fold lower expression, p = 0.001) by qPCR. In line with our hypothesis, the expression of *CCR2* was also reduced (33.3 fold, p = 0.0006) upon knockdown of *LY86* (Fig. 3E). These 72h *LY86* or control siRNA treated cells were then used in a migration assay to assess their migratory capacity towards the chemokine MCP-1 or serum-free medium as a control. After 3h incubation, we only observed migration towards MCP-1 and not the serum-free medium. Notably, the migratory capacity towards MCP-1 of the *LY86* siRNA treated cells was reduced (2.0 fold, p = 0.01) as compared to control siRNA treated cells (Fig. 1F). Summarized, these results indicate that reduced expression of *LY86* can reduce the migratory capacity of monocytes, potentially through reduced expression of *CCR2*, and thereby, may increase the susceptibility to candidemia.

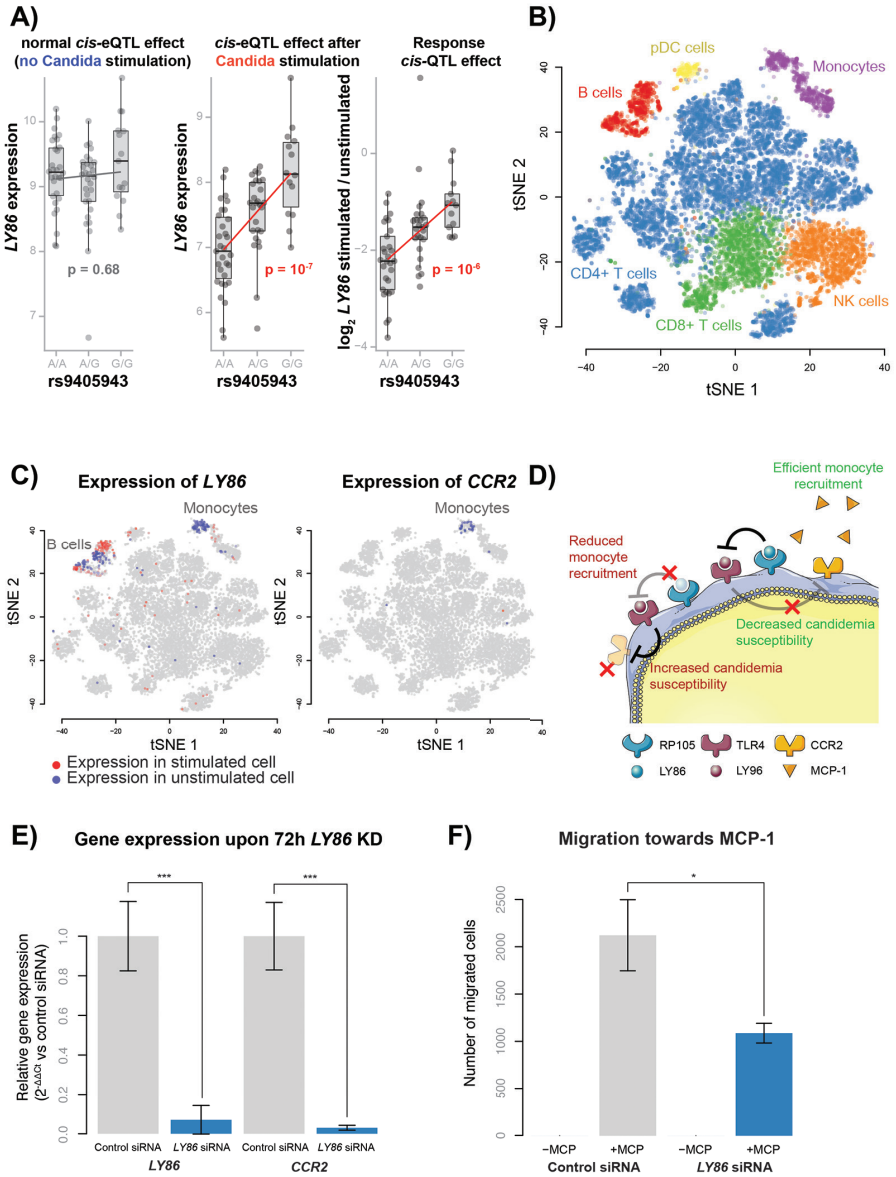


Figure 3. Proposed mechanism of *LY86* in candidemia susceptibility. (A) Box plots showing the effect of rs9405943 genotype on *LY86* expression without *Candida* stimulation (left), after *Candida* stimulation (center) and the response difference to *Candida* stimulation (right), as calculated in bulk RNA-seq data. Box plots show the median, first and third quartiles, and 1.5 × the interquartile range and each dot represents the expression of an individual. The x-axis shows the rs9405943 genotype and the y-axis shows the expression or expression response difference for *LY86*. Red p-values indicate significant effects. (B) A tSNE plot generated with single-cell expression data with and without *Candida* stimulation, colored by cell type. (C) Two tSNE plots

colored by the expression of *LY86* (left) and *CCR2* (right). Red cells indicate expression in a stimulated cell, blue cells indicate expression in an unstimulated cell and gray cells have no expression. (D) The proposed working mechanism for LY86 on candidemia susceptibility in monocytes. Due to competitive binding between LY86 and LY96 onto the TLR4 receptor, lower expression of *LY86* leads to increased activity of TLR4. As a consequence, TLR4-mediated chemokine receptor 2 (CCR2) repression increases, which reduces monocyte recruitment and increases candidemia susceptibility. (E) Normalized *LY86* and *CCR2* gene expression levels upon 72h *LY86* siRNA or control siRNA treatment in THP-1 monocytes. Each bar represents the mean ± SD of three independent experiments, *** $p < 0.001$ (F) Migration rate of 72h *LY86* versus control siRNA treated THP-1 cells towards MCP-1 or RPMI medium without serum. Each bar represents the mean ± SD of three independent experiments, *** $p < 0.001$

Final discussion and conclusion

In summary, we present an integrative approach of GWAS, bulk RNA-seq and scRNA-seq data to extract important knowledge about candidemia susceptibility. Such an integrative approach is valuable in the context of infectious diseases, such as candidemia, for which the limited size of patient cohorts limits the power of the GWAS. Otherwise, a GWAS alone would require much larger sample sizes in order to extract useful information from such studies. Moreover, a GWAS alone cannot explain how genetic variation affects disease or which cell type will be affected, and therefore, a systematic integration of different molecular datasets may be the only avenue to reveal this information. By combining these data layers, we corroborate the previously identified importance of the IFN pathway and of monocytes in *Candida* infections (9). In addition, we provide new evidence for a strong response in NK cells against *Candida* and a potential novel role for the *LY86* gene in candidemia susceptibility.

Our integrative approach is not limited to *Candida* infection, but can also be applied to gain a better insight into other infectious diseases for which the progress of disease understanding is hindered by small patient cohorts. We expect that in the near future, the cell type-specific and context-specific resolution of this integrated approach can be further improved as large-scale scRNA-seq datasets become readily available in many different individuals, stimulation conditions and diseases through large-scale consortia such as the single-cell eQTLGen (https://eqtlgen.org/single-cell.html) and LifeTime consortium (https://lifetime-fetflagship.eu). Such increased resolution would allow reconstruction of personalized, disease-specific gene regulatory networks that could provide us with new insights that could guide new treatment opportunities (36).

Materials and Methods

PBMC collection and *Candida*-stimulation

Whole blood from 6 individuals of the northern Netherlands population cohort Lifelines Deep (37) was drawn into EDTA-vacutainers (BD). PBMCs were isolated and maintained as previously described (38). In short, PBMCs were isolated using Cell Preparation Tubes with sodium heparin (BD) and were cryopreserved until use in RPMI 1640 containing 40% FCS and 10% DMSO. After thawing and a 1h resting period, $50x10^4$ cells were seeded in 200 µl RPMI1640 supplemented with 50 µg/mL gentamicin, 2 mM L-glutamine, and 1 mM pyruvate in a nucleon sphere 96-round bottom well plate. Cells were either stimulated or kept unstimulated for 24h with $1x10^6$ heat-killed *C. albicans* blastoconidia (strain ATCC MYA-3573, UC 820) CFU/ml at 37°C in a 5% CO_2 incubator. After 24h, cells were washed twice in medium supplemented with 0.04% bovine serum albumin. Cells were counted using a haemocytometer and cell viability was assessed by Trypan Blue.

Single-cell library preparation and sequencing

Three, sex-balanced sample pools were prepared each aimed to contain 1750 cells/donor from 6 donors (10,500 cells). One pool contained only unstimulated cells, one pool only stimulated cells and one pool contained a 50/50 mixture of both. Each sample pool was loaded into a different lane of a 10x chip (Single Cell A Chip Kit, 120236). The 10x Chromium controller (10x Genomics) in combination with v2 reagents was used to capture the single cells and generate sequencing libraries according to the manufacturer's instructions (document CG00026) and as previously described (38). Sequencing was performed using the Illumina HiSeq 4000 with a 75-bp paired-end kit, performed by GenomeScan (Leiden, the Netherlands).

Single-cell RNA-seq alignment, preprocessing and QC

Alignment, demultiplexing and cell type classification of the scRNA-seq data was performed as previously described (38), but now using the 2.3.0 version of Seurat (39). After QC, 15,085 cells remained of which 7,160 were stimulated and 7,925 were unstimulated. The stimulated and unstimulated cells were combined into a single dataset using Canonical Correlation Analysis (CCA) (39), by taking the first 20 dimensions. Clusters were identified using the FindClusters function from Seurat, using the first 20 dimensions in the CCA space. Expression of known marker genes was assessed to assign cell types to each cluster, resulting in the identification of six major cell types.

Single-cell RNA-seq differential expression analysis

Differential expression (DE) between *Candida*-stimulated and RPMI control cells was calculated for each cell type separately and in a bulk-like analysis using the MAST

implementation of the Seurat package (14). All genes without expression in at least 1 cell were removed, leaving 20,236 genes. Bonferroni multiple testing correction was applied, yielding a significance threshold of 2.47e-06. Genes that were differentially expressed in all cell types (i.e. core genes) and each cell type individually were used as input for the ToppFun functional enrichment analysis using the REACTOME pathway (40). P-values were calculated using the probability density function and were Bonferroni corrected.

Cell-to-cell interaction potential analysis

The potential of cell-to-cell communication through ligand/receptor pair interactions was studied using version 2 of CellPhoneDB (16). This software uses the normalized expression data and the cell type classifications to see which cell types have expression of known ligands and receptors to estimate whether there is an interaction potential between cells of the same or different cell types. The analysis was performed on each cell type and condition (24h *Candida*-stimulated versus RPMI control) separately. CellPhoneDB was run using the default database of ligand-receptor interactions provided with the software and was run using default settings for p-value thresholds (0.05), expression threshold (expression in >= 10% cells) and permutations (1,000).

Bulk RNA-seq data on *Candida*-stimulated PBMCs

All bulk RNA-seq data from PBMCs was previously generated (7) in 70 individuals from the GONL cohort (41). This data was generated from PBMCs that were stimulated for 24h with *Candida* or remained unstimulated (RPMI control condition). Details of the stimulation are similar to the scRNA-seq data on *Candida*-stimulated PBMCs as mentioned above, and have been previously described (9). The differentially expressed genes upon stimulation were previously identified (7) through DESeq2 (42). The differential expressed genes identified in the scRNA-seq data were compared with the differential response in this bulk RNA-seq data.

Bulk RNA-seq eQTL analysis

Of the same bulk RNA-seq cohort, eQTLs were identified in the data from 72 individuals and 75 individuals for *Candida*-stimulated and RPMI control conditions, respectively. The response QTLs were identified in the 67 individuals for which both conditions were assessed and genotype information is available. To calculate this, we subtracted per individual and per gene the *Candida*-stimulated expression from the RPMI control expression and tested whether certain SNPs affected the expression of a particular gene with different effect sizes before and after stimulation. All expression data were log2 transformed before being used in the 1.2.4F version of the QTL pipeline as described before (43). To reduce the multiple testing burden, analysis was restricted

to the list of 16,989 top SNP-gene combinations identified in the largest whole blood eQTL meta-analysis to date containing 31,684 whole blood samples (23). This list of top SNP-gene combinations contains SNPs with minor allele frequencies (MAF) >0.01, Hardy-Weinberg P-values >0.0001, call rate >0.95, and MACH r2 > 0.5 within a 1Mb window of the gene. An FDR threshold of 0.05 was used as significance cut-off, using the permutation strategy described in Westra et al. with 100 permutations (44).

GWAS on candidemia susceptibility

The GWAS on candidemia susceptibility was previously described (8). In short, this GWAS was performed in a cohort of 161 candidemia cases and 152 disease-matched controls of European ancestry whose demographic and clinical characteristics have been previously described (45). DNA was genotyped using Illumina HumanCoreExome-12 v1.0 and HumanCoreExome-24 v1.0 BeadChip SNP chips. Genotypes were imputed using the human reference consortium reference panel (46) using the Michigan imputation server (47). In total, 5,426,313 SNPs were tested for disease association using Fisher's exact test with PLINK v1.9 (48). The lambda inflation was calculated by taking the GWAS p-values for each of the 27 response-QTL SNPs, regardless of whether the GWAS p-value was significant.

siRNA treatment

Before starting the experiment, THP-1 monocytes were maintained in RPMI medium supplemented with 10% FBS and 1% Pen-Strep at 37°C in a humidified 5% CO_2 incubator. 50,000 THP-1 monocytes were seeded in round-bottom 96-wells plates. During seeding, 1 µM Accell human *LY86* siRNA SMARTpool (Dharmacon) or 1 µM Accell Green non-targeting siRNA control (Dharmacon) was delivered to these cells in 100 ul Accell delivery medium. After 24h, this procedure was repeated by adding an additional 100 ul to each well. After 72h, *LY86* and *CCR2* mRNA levels were quantified using qRT-PCR and migration rate was quantified using a migration assay.

Quantitative real-time PCR (qRT-PCR)

RNA was isolated using QIAzol lysis reagent according to manufacturer's instructions. RNA was quantified using a Nanodrop 1000 spectrophotometer (Thermo Scientific). 400 ng RNA was reverse transcribed into cDNA using random hexamer primers with the RevertAid H Minus First Strand cDNA Synthesis Kit (Thermo Scientific) following manufacturer's protocol. Each qRT-PCR reaction contained 500 nM of each primer pair (Table 1), 10 ng of cDNA and 1x iTaq universal SYBR green supermix (Bio-Rad). qRT-PCR reactions were conducted on the Quantstudio 7 Flex real time PCR (Thermo Fischer) for 10 min at 95 °C, followed by 40 cycles of 15 sec at 95 °C and 30 sec at 60 °C. GAPDH was used as housekeeping gene. Data and melting curves were analyzed using Quantstudio Real-time PCR software v1.3 and

relative expression compared to controls was calculated using the ΔΔCt method (49). Significance was calculated using an unpaired t-test.

Table 1. qRT-PCR primer sequences

Target gene	Forward primer (5'-3')	Reverse primer (5'-3')
LY68	TGTGGAAGAAGGAAAGGAGAGCA	GTACAGTTCCAGCAAAACCTGG
CCR2	AGTTGCTGAGAAGCCTGACA	TCTCTGTTCAGCTTGTGGCT
GAPDH	CCACATCGCTCAGACACCAT	GCGCCCAATACGACCAAAT

Migration assay

The Boyden Chamber transwell migration assay was used to determine the migration rate towards MCP-1 upon *LY86* KD (50). A polycarbonate membrane insert with a 5 μM pore size (Cell Biolabs) was placed in a well of a 24-wells plate filled with 500 μl Accell delivery medium supplemented with 0.5% BSA with or without 100 ng/ml human MCP-1 (Prospec). The insert was filled with 100 μl Accell delivery medium supplemented with 0.5% BSA and 100,000 THP-1 monocytes treated for 72h with *LY68* siRNA or Green non-targeting siRNA. Cells were placed in a humidified incubator with 5% CO2 at 37 °C. After 3h, the number of migratory cells was quantified in the bottom well using a hemocytometer. Significance was calculated using an unpaired t-test.

Data availability

A Seurat object (39), containing the processed single-cell RNA-seq data after QC and cell type assignment, is made available through the website accompanying our manuscript: https://eqtlgen.org/candida.html.

Code availability

The original R code for Seurat (39) (https://github.com/satijalab/seurat) and our in-house eQTL pipeline (43) (https://github.com/molgenis/ systemsgenetics/tree/master/eqtl-mapping-pipeline) can be found at Github. All custom-made code is made available via GitHub (https://github.com/molgenis/scRNA-seq).

Ethics statement

The LifeLines DEEP study was approved by the ethics committee of the University Medical Center Groningen, document number METC UMCG LLDEEP: M12.113965. All participants signed an informed consent form before study enrollment. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Author contributions

Conceptualization: Dylan H. de Vries, Vasiliki Matzaraki, Mihai G. Netea, Lude Franke, Vinod Kumar, Monique G.P. van der Wijst
Data curation: Dylan H. de Vries, Olivier Bakker, Harm Brugge, Harm-Jan Westra
Formal analysis: Dylan H. de Vries, Vasiliki Matzaraki
Funding acquisition: Lude Franke
Investigation: Dylan H. de Vries, Monique G.P. van der Wijst
Project administration: Lude Franke, Monique G.P. van der Wijst
Software: Dylan H. de Vries
Supervision: Lude Franke, Vinod Kumar, Monique G.P. van der Wijst
Visualization: Dylan H. de Vries, Harm-Jan Westra, Lude Franke
Writing - Original Draft Preparation: Dylan H. de Vries, Vasiliki Matzaraki, Vinod Kumar, Monique G.P. van der Wijst
Writing - Review & Editing: Olivier Bakker, Mihai G. Netea, Lude Franke

Supplementary material

Supplementary material is available at https://doi.org/10.1371/journal.ppat.1008408

References

(1) Mavor AL, Thewes S, Hube B. Systemic fungal infections caused by Candida species: epidemiology, infection process and virulence attributes. Curr Drug Targets 2005 Dec;6(8):863-874.

(2) Quindos G. Epidemiology of candidaemia and invasive candidiasis. A changing face. Rev Iberoam Micol 2014 Jan-Mar;31(1):42-48.

(3) Leroy O, Gangneux JP, Montravers P, Mira JP, Gouin F, Sollet JP, et al. Epidemiology, management, and risk factors for death of invasive Candida infections in critical care: a multicenter, prospective, observational study in France (2005-2006). Crit Care Med 2009 May;37(5):1612-1618.

(4) Moran C, Grussemeyer CA, Spalding JR, Benjamin DK,Jr, Reed SD. Comparison of costs, length of stay, and mortality associated with Candida glabrata and Candida albicans bloodstream infections. Am J Infect Control 2010 Feb;38(1):78-80.

(5) Johnson MD, Plantinga TS, van de Vosse E, Velez Edwards DR, Smith PB, Alexander BD, et al. Cytokine gene polymorphisms and the outcome of invasive candidiasis: a prospective cohort study. Clin Infect Dis 2012 Feb 15;54(4):502-510.

(6) Delsing CE, Gresnigt MS, Leentjens J, Preijers F, Frager FA, Kox M, et al. Interferon-gamma as adjunctive immunotherapy for invasive fungal infections: a case series. BMC Infect Dis 2014 Mar 26;14:166-2334-14-166.

(7) Li Y, Oosting M, Deelen P, Ricano-Ponce I, Smeekens S, Jaeger M, et al. Inter-individual variability and genetic influences on cytokine responses to bacteria and fungi. Nat Med 2016 Aug;22(8):952-960.

(8) Jaeger M, Matzaraki V, Aguirre-Gamboa R, Gresnigt MS, Chu X, Johnson MD, et al. A genome-wide functional genomics approach identifies susceptibility pathways to fungal bloodstream infection in humans. jid 2019.

(9) Smeekens SP, Ng A, Kumar V, Johnson MD, Plantinga TS, van Diemen C, et al. Functional genomics identifies type I interferon pathway as central for host defense against Candida albicans. Nat Commun 2013;4:1342.

(10) Matzaraki V, Gresnigt MS, Jaeger M, Ricano-Ponce I, Johnson MD, Oosting M, et al. An integrative genomics approach identifies novel pathways that influence candidaemia susceptibility. PLoS One 2017 Jul 20;12(7):e0180824.

(11) Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015 Feb 19;518(7539):317-330.

(12) Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc 2018 Apr;13(4):599-604.

(13) Blecher-Gonen R, Bost P, Hilligan KL, David E, Salame TM, Roussel E, et al. Single-Cell Analysis of Diverse Pathogen Responses Defines a Molecular Roadmap for Generating Antigen-Specific Immunity. Cell Syst 2019 Feb 27;8(2):109-121.e6.

(14) Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol 2015 Dec 10;16:278-015-0844-5.

(15) Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics 2019 Jan 18;20(1):40-019-2599-6.

(16) Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, et al. Single-cell reconstruction of the early maternal-fetal interface in humans. Nature 2018 Nov;563 (7731):347-353.

(17) Smeekens SP, van de Veerdonk FL, Joosten LA, Jacobs L, Jansen T, Williams DL, et al. The classical CD14(+)(+) CD16(-) monocytes, but not the patrolling CD14(+) CD16(+) monocytes, promote Th17 responses to Candida albicans. Eur J Immunol 2011 Oct;41(10):2915-2924.

(18) Ngo LY, Kasahara S, Kumasaka DK, Knoblaugh SE, Jhingran A, Hohl TM. Inflammatory monocytes mediate early and organ-specific innate defense during systemic candidiasis. J Infect Dis 2014 Jan 1;209(1):109-119.

(19) Romani L, Mencacci A, Cenci E, Spaccapelo R, Schiaffella E, Tonnetti L, et al. Natural killer cells do not play a dominant role in CD4+ subset differentiation in Candida albicans-infected mice. Infect Immun 1993 Sep;61(9):3769-3774.

(20) Whitney PG, Bar E, Osorio F, Rogers NC, Schraml BU, Deddouche S, et al. Syk signaling in dendritic cells orchestrates innate resistance to systemic fungal infection. PLoS Pathog 2014 Jul 17;10(7):e1004276.

(21) Quintin J, Voigt J, van der Voort R, Jacobsen ID, Verschueren I, Hube B, et al. Differential role of NK cells against Candida albicans infection in immunocompetent or immunocompromised mice. Eur J Immunol 2014 Aug;44(8):2405-2414.

(22) GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)-Analysis Working Group, Statistical Methods groups-Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, et al. Genetic effects on gene expression across human tissues. Nature 2017 Oct 11;550(7675):204-213.

(23) Võsa U, Claringbould A, Westra H, Bonder MJ, Deelen P, Zeng B, et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv 2018 Cold Spring Harbor Laboratory:447367.

(24) Tada Y, Koarada S, Morito F, Mitamura M, Inoue H, Suematsu R, et al. Toll-like receptor homolog RP105 modulates the antigen-presenting cell function and regulates the development of collagen-induced arthritis. Arthritis Res Ther 2008;10(5):R121.

(25) Wezel A, van der Velden D, Maassen JM, Lagraauw HM, de Vries MR, Karper JC, et al. RP105 deficiency attenuates early atherosclerosis via decreased monocyte influx in a CCR2 dependent manner. Atherosclerosis 2015 Jan;238(1):132-139.

(26) Chen X, Pan H, Li J, Zhang G, Cheng S, Zuo N, et al. Inhibition of myeloid differentiation 1 specifically in colon with antisense oligonucleotide exacerbates dextran sodium sulfate-induced colitis. J Cell Biochem 2019 May 19.

(27) Divanovic S, Trompette A, Atabani SF, Madan R, Golenbock DT, Visintin A, et al. Negative regulation of Toll-like receptor 4 signaling by the Toll-like receptor homolog RP105. Nat Immunol 2005 Jun;6(6):571-578.

(28) Schultz TE, Blumenthal A. The RP105/MD-1 complex: molecular signaling mechanisms and pathophysiological implications. J Leukoc Biol 2017 Jan;101(1):183-192.

(29) Ogata H, Su I, Miyake K, Nagai Y, Akashi S, Mecklenbrauker I, et al. The toll-like receptor protein RP105 regulates lipopolysaccharide signaling in B cells. J Exp Med 2000 Jul 3;192(1):23-29.

(30) Kimoto M, Nagasawa K, Miyake K. Role of TLR4/MD-2 and RP105/MD-1 in innate recognition of lipopolysaccharide. Scand J Infect Dis 2003;35(9):568-572.

(31) Parker LC, Whyte MK, Vogel SN, Dower SK, Sabroe I. Toll-like receptor (TLR)2 and TLR4 agonists regulate CCR expression in human monocytic cells. J Immunol 2004 Apr 15;172(8):4977-4986.

(32) Loures FV, Pina A, Felonato M, Araujo EF, Leite KR, Calich VL. Toll-like receptor 4 signaling leads to severe fungal infection associated with enhanced proinflammatory immunity and impaired expansion of regulatory T cells. Infect Immun 2010 Mar;78(3):1078-1088.

(33) Meier A, Kirschning CJ, Nikolaus T, Wagner H, Heesemann J, Ebel F. Toll-like receptor (TLR) 2 and TLR4 are essential for Aspergillus-induced activation of murine macrophages. Cell Microbiol 2003 Aug;5(8):561-570.

(34) Netea MG, Gow NA, Joosten LA, Verschueren I, van der Meer JW, Kullberg BJ. Variable recognition of Candida albicans strains by TLR4 and lectin recognition receptors. Med Mycol 2010 Nov;48(7):897-903.

(35) Mukherjee S, Karmakar S, Babu SP. TLR2 and TLR4 mediated host immune responses in major infectious diseases: a review. Braz J Infect Dis 2016 Mar-Apr;20(2):193-204.

(36) van der Wijst MGP, de Vries DH, Brugge H, Westra HJ, Franke L. An integrative approach for building personalized gene regulatory networks for precision medicine. Genome Med 2018 Dec 19;10(1):96-018-0608-4.

(37) Tigchelaar EF, Zhernakova A, Dekens JA, Hermes G, Baranska A, Mujagic Z, et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open 2015 Aug 28;5(8):e006772-2014-006772.

(38) van der Wijst MGP, Brugge H, de Vries DH, Deelen P, Swertz MA, Franke L. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat Genet 2018 04/02;50(4):493-497.

(39) Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018 Apr 2;36(5):411-420.

(40) Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res 2009 Jul;37(Web Server issue):W305-11.

(41) Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 2014 Aug;46(8):818-825.

(42) Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550-014-0550-8.

(43) Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet 2017 print;49(1):139-145.

(44) Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet 2013 Oct;45(10):1238-1243.

(45) Kumar V, Cheng SC, Johnson MD, Smeekens SP, Wojtowicz A, Giamarellos-Bourboulis E, et al. Immunochip SNP array identifies novel genetic variants conferring susceptibility to candidaemia. Nat Commun 2014 Sep 8;5:4675.

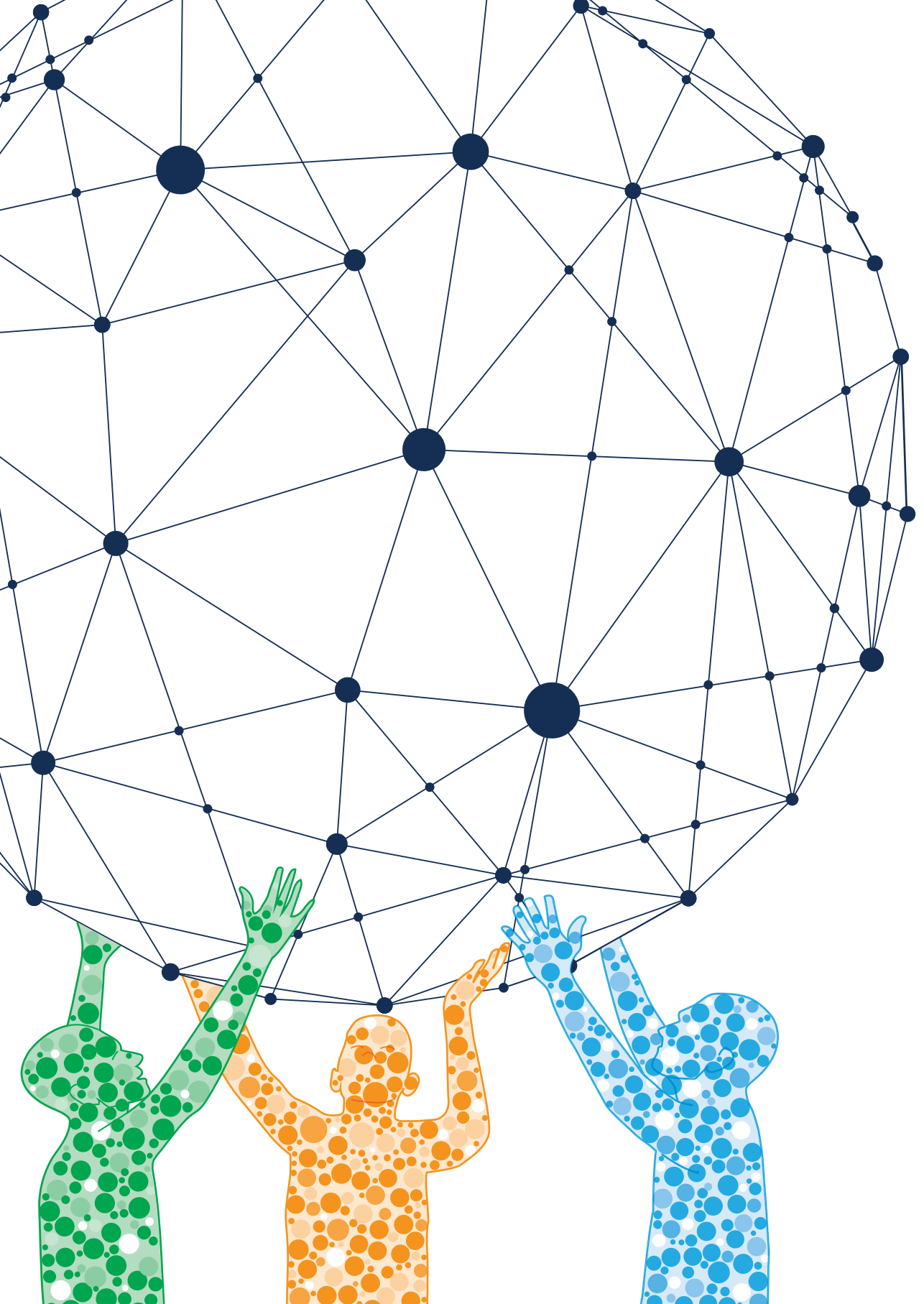
(46) McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 2016 Oct;48(10):1279-1283.

(47) Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet 2016 Oct;48(10):1284-1287.

(48) Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007 Sep;81(3):559-575.

(49) Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 2001 Dec;25(4):402-408.

(50) Boyden S. The chemotactic effect of mixtures of antibody and antigen on polymorphonuclear leucocytes. J Exp Med 1962 Mar 1;115:453-466.



CHAPTER 5

Single-cell RNA-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure

Roy Oelen[1,2#], Dylan H. de Vries[1,2#], Harm Brugge[1,2#], M. Grace Gordon[3-6], Martijn Vochteloo[1], single-cell eQTLGen consortium[$], BIOS Consortium[$], Chun J. Ye[4,6-10], Harm-Jan Westra[1], Lude Franke[1,2^`], Monique G.P. van der Wijst[1,2^`]

[1] Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.
[2] Oncode Institute, Utrecht, The Netherlands
[3] Biological and Medical Informatics Graduate Program, University of California San Francisco, San Francisco, CA, USA.
[4] Institute for Human Genetics, University of California San Francisco, CA, USA.
[5] Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA.
[6] UCSF Division of Rheumatology, Department of Medicine, University of California San Francisco, San Francisco, CA, USA.
[7] Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, CA, USA.
[8] Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, USA.
[9] Parker Institute for Cancer Immunotherapy, San Francisco, CA, USA.
[10] Chan Zuckerberg Biohub, San Francisco, CA, USA
[#] These authors contributed equally
* These authors jointly supervised this work
` Corresponding authors (LF: l.h.franke@umcg.nl, MW: m.g.p.van.der.wijst@umcg.nl)
[$] A full list of members and their affiliations appears in the Supplementary Information.

Published in Nature Communications. https://doi.org/10.1038/s41467-022-30893-5

Abstract

The host's gene expression and gene regulatory response to pathogen exposure can be influenced by a combination of the host's genetic background, the type of and exposure time to pathogens. Here we provide a detailed dissection of this using single-cell RNA-sequencing of 1.3M peripheral blood mononuclear cells from 120 individuals, longitudinally exposed to three different pathogens. These analyses indicate that cell-type-specificity is a more prominent factor than pathogen-specificity regarding contexts that affect how genetics influences gene expression (i.e. eQTL) and co-expression (i.e. co-expression QTL). In monocytes, the strongest responder to pathogen stimulations, 71.4% of the genetic variants whose effect on gene expression is influenced by pathogen exposure (i.e. response QTL) also affect the co-expression between genes. This indicates widespread, context-specific changes in gene expression level and its regulation that are driven by genetics. Pathway analysis on the *CLEC12A* gene that exemplifies cell-type-, exposure-time- and genetic-background-dependent co-expression interactions, shows enrichment of the interferon (IFN) pathway specifically at 3-hour post-exposure in monocytes. Similar genetic background-dependent association between IFN activity and *CLEC12A* co-expression patterns is confirmed in systemic lupus erythematosus by *in silico* analysis, which implies that *CLEC12A* might be an IFN-regulated gene. Altogether, this study highlights the importance of context for gaining a better understanding of the mechanisms of gene regulation in health and disease.

Introduction

Over a decade of genome-wide association studies (GWAS) has revealed thousands of genetic variants associated with disease risk[1], most of them single nucleotide polymorphisms (SNPs). Despite this, the cascade of events through which these variants change disease risk remains largely unclear. One way to dissect this cascade is by linking disease-associated SNPs to downstream gene expression through so-called expression quantitative trait locus (eQTL) analysis.[2] However, recent work by Yao et al. indicated that, on average, only 11% ± 2% of disease heritability is mediated by *cis*-eQTLs, i.e. SNPs affecting the expression of nearby genes.[3] One explanation for this relatively low contribution could be that many of these eQTL effects are cell type-specific and context-dependent[4,5], which means that their disease contribution cannot be accurately estimated using steady-state expression in bulk-averaged tissues. In other words, the relevant context for a particular disease-associated SNP may not have been studied yet, meaning that many of the true downstream effects of these SNPs remain hidden.[6] In a first effort to identify tissue-specific eQTLs, the GTEx consortium performed eQTL analysis in 44 different human tissues across 449 individuals (70-361 individuals/tissue).[7] However, this study was limited by the relatively small number of donors for many of the tissues and the lack of cell type-specific resolution. More recently, with the advent of high-throughput, cost-efficient single-cell RNA sequencing (scRNA-seq) technologies[8,9], it has become possible to assess both the cell type-specific and context-dependent effects of risk SNPs on downstream gene expression.[10–12]

While the tissue or cell type is one context that can affect the association between a SNP genotype and gene expression, many other contexts can also be of influence. For the immune system, for example, exposure to specific pathogens commonly occurs and the immune response following exposure can create the environmental context required to change specific interactions between genetics and downstream gene expression.[4,13–17] In turn, these context-specific interactions may explain why exposure to specific pathogens has been associated with the development of autoimmune diseases in individuals with a genetic predisposition.[18] For example, a reovirus can disrupt intestinal immune homeostasis and initiate loss of tolerance to gluten in individuals expressing HLA-DQ2 or HLA-DQ8, leading to celiac disease.[19] Another example is the strong indications that enteroviral infections in the pancreas, such as with coxsackievirus, in genetically predisposed individuals may accelerate the development of type I diabetes (T1D).[20–22] Several T1D-associated risk genes affect the antiviral response through regulation of type I interferon (IFN) signaling.[23] When the insulin-producing pancreatic β cells of genetically predisposed individuals are then exposed to such viruses, incomplete viral clearance and chronic infection of these β cells may be the consequence. This could then induce β cell apoptosis that contributes to the development of T1D.[24,25] Overall, it is estimated that 11-30% of autoimmune risk loci involve *cis*-eQTLs in blood, and it is hypothesized that trait-

associated eQTLs have increased context-specificity.[26–28] Given this hypothesized context-specificity, it is important to study eQTLs in a variety of different contexts to determine the possible effect of environment on the interplay between genetic variation and gene expression in disease.

This study aims to disentangle the gene expression and gene regulatory processes that are driven by differences in genetics and/or pathogen exposures, and that could explain how interindividual differences can contribute to disease risk. Moreover, we show how the properties of scRNA-seq data (i.e. cell- and context-specific resolution, high number of cellular observations per individual) can be employed to disentangle the molecular mechanisms that underlie the context-specificity of the genetic regulation. By disentangling these mechanisms, we provide novel insights how genetics can contribute to disease risk aiding us to reduce such risk in the future.

Results

Single-cell profiling of immune cells upon pathogen stimulation

Here we present the 1M-scBloodNL study in which we performed 10x Genomics scRNA-seq on 120 individuals from the Northern Netherlands population cohort Lifelines. For each individual, we sequenced peripheral blood mononuclear cells (PBMC) in an unstimulated condition and after 3h and 24h *in vitro*-stimulation with *C. albicans* (CA), *M. tuberculosis* (MTB) or *P. aeruginosa* (PA), totaling approximately 1.3 million cells (**Fig. 1**, **Table S1**). A combination of 10X Genomics v2 and v3 chemistry reagents were used to capture an average of 1,226 cells per individual per condition (v2: 907 genes/cell, v3: 1,861 genes/cell) (**Table S2**). Souporcell[29] was used to identify the doublets coming from different individuals, followed by sample demultiplexing using Demuxlet.[11] This revealed, on average, 12.0% of cells as doublets. Due to differences in gene amplification between v2 and v3 chemistry, determination of quality control (QC) thresholds and analyses were performed separately per chemistry (**Supplementary Fig. 1a-d**). Results from both chemistries were then meta-analyzed for interpretation. Low quality cells were excluded, leaving 928,275 cells in the final dataset used for analysis (see Methods, **Supplementary Table 3**). UMAP dimensionality reduction and KNN-clustering was then applied on the normalized, integrated count data, allowing the identification of six main cell types: B, CD4+ T, CD8+ T, monocytes, natural killer (NK) and dendritic cells (DCs) (**Supplementary Fig. 1e-g, Supplementary Table 4**), for which the latter five were further subdivided in two subcell types each: naïve and memory CD4+ T and CD8+ T cells, classical (cMono) and non-classical monocytes (ncMono), NKdim and NKbright, myeloid (mDC) and plasmacytoid DCs (pDC).

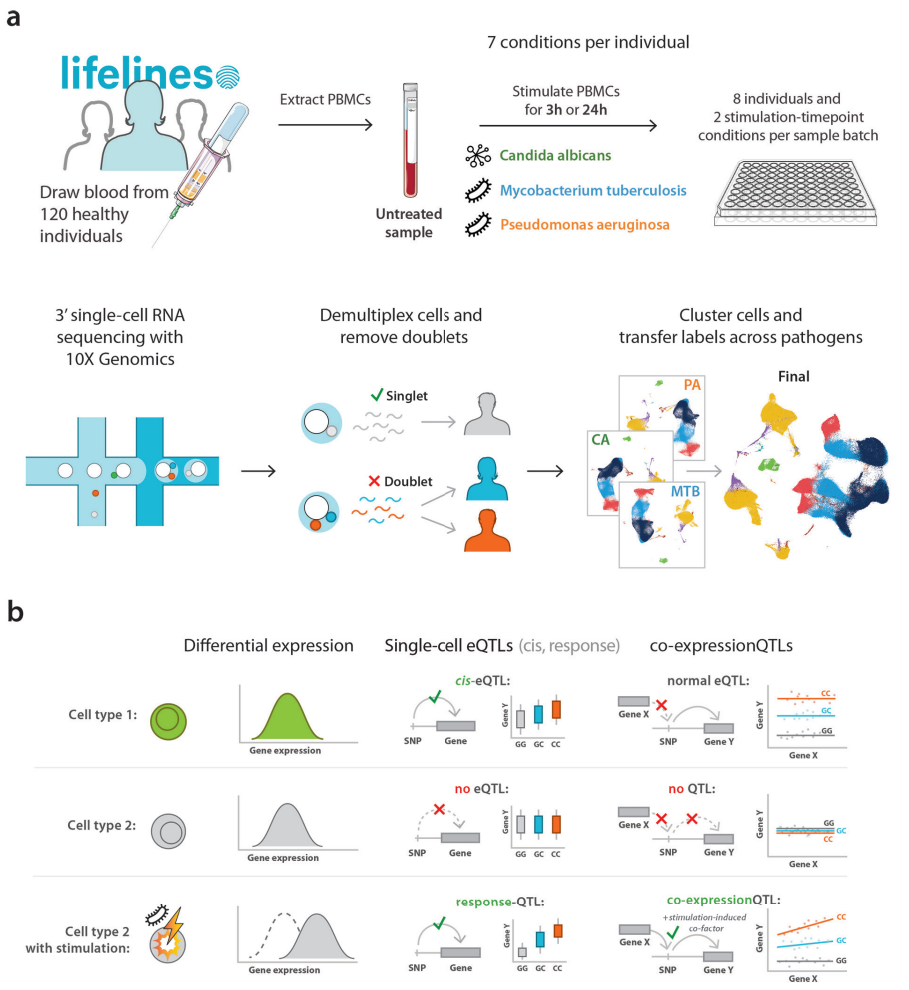


Figure 1. Study overview. a. *Blood was drawn from 120 individuals from the Northern Netherlands participating in the population-based cohort Lifelines. PBMCs were isolated within 2h of blood collection and cryopreserved in liquid nitrogen until further use. For each scRNA-seq experiment, PBMCs from 16 individuals were thawed. Per individual, these PBMCs were left untreated (UT) or were stimulated with C. Albicans (CA), M. Tuberculosis (MTB) or P. Aeruginosa (PA) for 3h or 24h in a 96-well plate. In total, this resulted in 7 stimulation-timepoint combinations per 120 individuals = 840 different conditions processed. Multiplexed 3'-end scRNA-seq was performed using the v2 and v3 chemistries of the 10X Genomics platform. Per experiment, two sample batches of a 10x chip were loaded, each containing a mixture of eight individuals and a combination of two different stimulation-timepoint combinations. After sequencing, samples were demultiplexed and doublets were identified. Cell type classification was performed on the QCed dataset by clustering the cells per pathogen, mixed with cells of the unstimulated condition. The cell type labels were subsequently transferred back to the dataset containing all cells.* ***b.*** *This study was conducted to identify cell type-specific or pathogen-stimulation dependent: 1. differentially expressed genes, 2. eQTLs and response-QTLs and 3. co-expression QTLs.*

Gene expression response upon pathogen stimulation reveals stronger cell type-specificity than pathogen-specificity

To assess the transcriptional changes upon pathogen stimulation with CA, MTB and PA, we performed differential expression (DE) analysis using MAST in each of the major cell types and their subcell types (**Supplementary Table 5**).[30] For the major cell types, pairwise comparisons between the untreated and pathogen-stimulated conditions revealed between 688 to 2,022 DE genes after 3h stimulation, further increasing to 1,052 to 2,616 DE genes after 24h stimulation (**Fig. 2a**). The number of DE genes was comparable between the different pathogen stimulations at the same timepoint but differed strongly between some cell types. Myeloid cells (monocytes and DCs) showed the highest number of DE genes, whereas both CD4+ and CD8+ T cells showed the fewest DE genes. This is consistent with the innate immune cells being the first responders during pathogen stimulation.[31]

A total of 5,516 unique DE genes were identified over all conditions and major cell types, and an additional 1,621 DE genes were identified in the subcell types (**Table S5**). This indicates that most DE genes can already be identified at the major cell type level. However, since the statistical power to detect such DE effects is correlated with the number of cells within a subcell type[32], likely some of the subcell type specificity remains undetectable. Of the 5,516 DE genes within the major cell types, 31.1% were cell type-specific and 15.1% were shared across all major cell types (**Fig. 2b**). The fraction of DE genes that were cell type-specific was comparable for each of the cell types, but, in absolute numbers, monocytes and DCs had the most unique DE genes. Sharing between different pathogen stimulations at the same timepoint was more prominent than sharing between different timepoints within the same pathogen stimulation (**Fig. 2c, Fig. S1f**): 39.8% of the total unique DE genes were shared across the same timepoint (7.4% at 3h and 32.4% at 24h), whereas only 10.3% of DE genes were unique to a specific pathogen stimulation and 41.3% were shared across all stimulation-timepoint combinations. This indicates that the immune response to our pathogen stimulations of both bacterial and fungal origin was more specific to timepoint after stimulation than to type of pathogen. Consequently, the genetic control of these responsive genes is expected to be more time-dependent than pathogen-dependent.

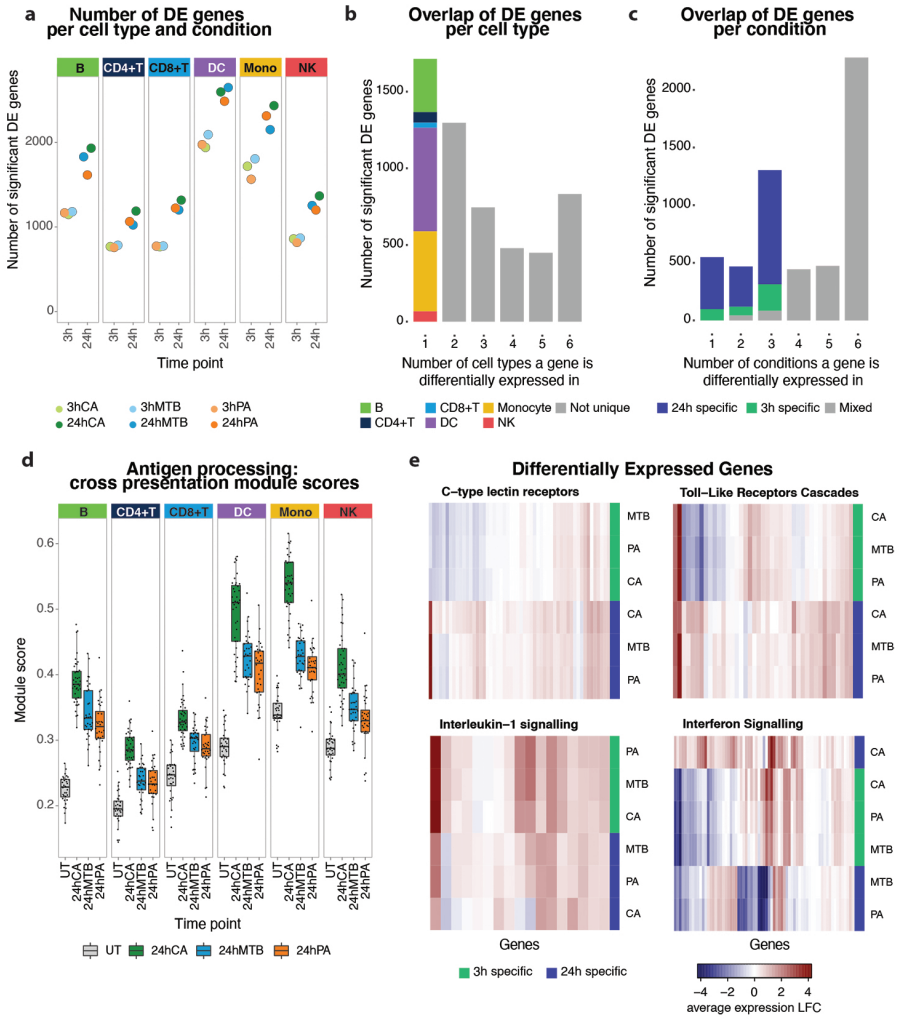


Figure 2. Differentially expressed genes and pathways upon pathogen stimulation. a. *Number of DE genes per cell type upon 3h (light colors) or 24h (dark colors) stimulation with C. Albicans (CA, green), M. Tuberculosis (MTB, blue) or P. Aeruginosa (PA, orange).* ***b.*** *Bar plot showing the overlap of DE genes across cell types. The first bar, depicting the cell type-specific DE genes, is colored based on the cell type in which the DE gene is found.* ***c.*** *Bar plot showing the overlap of DE genes across pathogen-timepoint combinations (3h vs 24h stimulation with CA, PA or MTB). Bars are colored based on the length of stimulation.* ***d.*** *Boxplots (showing median, 25th and 75th percentile, and 1.5 ×the interquartile range) representing the module score of the antigen-processing cross-presentation pathway across all individuals per cell type and per 24 hours pathogen-stimulated condition. Each dot shows the average module score per individual (V3 chemistry is shown, the* ***Source Data*** *file includes the individual data points).* ***e.*** *Heatmaps showing the immune-related DE genes in monocytes (V3 chemistry is shown), split by their involvement in either one of the four selected immune pathways associated with pathogen recognition and its downstream signaling. C-type lectin and*

*toll-like receptors show more general activation upon pathogen stimulation, whereas interleukin-1 and interferon signaling show a more specific expression pattern with timepoint (3h stimulation) or stimulation (CA), respectively. DE summary statistics can be found in **table S5**. The number of individuals and cells included in each analysis can be found in the **Source Data** file.*

To evaluate the DE results and confirm proper activation of the cells upon stimulation, we performed two different analyses. In the first analysis, we measured the activity of a general stimulation-responsive pathway - the antigen processing-cross presentation pathway (REACTOME R-HSA-1236975) - that should become activated in each of the cell types and upon each of the pathogen-stimulations. This analysis revealed increased activity of the antigen-processing pathway-associated genes across all cell types after 24h stimulation and for each of the pathogens (**Fig. 2d**). In the second analysis, we focused on DE genes identified upon 24h stimulation with CA. We had previously performed similar analyses in a smaller scRNA-seq study[15], so we could use this study for comparison purposes. This analysis revealed a high concordance between DE genes in our current study and those from our previous study, varying from 73% for the monocytes up to 93% for the B cells (**Fig. S2**). In general, these analyses showed that CA stimulation resulted in the highest activation of genes associated with the antigen-processing pathway and that monocytes were the cell type with the strongest response. These two analyses confirmed proper activation of the cells and stimulation responses that were in line with previous literature.[15,33]

Next, we determined which pathways were enriched within the upregulated DE genes for each cell type and each pathogen-timepoint combination (**Supplementary Table 6**). In line with the DE results, most of the enriched pathways were shared across the different pathogen stimulation conditions within the same timepoint (**Fig. 2e**). To highlight relevant pathways involved in pathogen recognition and downstream immune response, we filtered the enriched pathways for those related to the 'Immune system' REACTOME pathway parent term. For this illustrative example, we selected monocytes because this was the cell type in which we observed the most DE genes (**Fig. 2e**). Here we observed a general activation of pathogen-recognition receptors and downstream signaling, including the C-type lectin and toll-like receptors. Some pathways, such as interleukin-1 (IL-1) signaling, were clearly enriched at a specific timepoint (3h stimulation), whereas others, such as the IFN pathway, showed a notable difference between different pathogen stimulations (more prominently activated in CA compared to the other two pathogens). These findings corroborate literature describing IFN as an important signaling pathway in response to all three pathogens[33–35] and that IL-1 family molecules are part of the early stages (<14h) of the inflammatory response in monocytes with their expression decreasing again at later stages.[36]

For the subcell types, we were mainly interested in those pathways that were differentially activated upon pathogen stimulation between the two subcell types of each major cell type. For this, we visualized the top 10 most enriched pathways with the largest difference in significance between both subcell types (**Supplementary Fig. 3**).

This revealed that most pathways were enriched in both subtypes, but that the relative activation could differ. For example, several pathways associated with interferon signaling were more significantly enriched in the ncMono as opposed to the cMono (**Supplementary Fig. 3, Supplementary Table 6**).

The number of eQTLs decrease in cells with stronger stimulation response

Our experimental set-up, in which we analyzed pathogen-stimulated PBMCs using scRNA-seq, allowed us to investigate the extent to which SNPs affect gene expression in different contexts. To maximize the power to detect eQTLs, we took advantage of a previously conducted genome-wide *cis*-eQTL meta-analysis in 31,684 whole blood bulk samples (eQTLGen[37]) by only testing their top SNP-gene combinations, i.e. lead-eSNPs. Due to the power of eQTLGen, they could identify even *cis*-eQTL effects with a small effect size. We therefore expected that many of the context-specific effects, to which only a subset of individuals or cell types might have been exposed, should have resulted in an eQTL effect identified in eQTLGen. However, compared to the eQTLGen bulk whole-blood dataset, our pathogen-stimulated scRNA-seq data has the additional benefit that it can identify the cell types and contexts in which these eQTL effects manifest themselves.

We performed the eQTLGen lead-eSNP *cis*-eQTL discovery analysis per cell type and for each stimulation-timepoint combination separately (**Supplementary Table 7**). When determining the concordance between eQTLGen's bulk whole-blood eQTLs and those identified in our study, we observed that the concordance was high in general despite the compositional differences between whole blood and the PBMCs or cells in this study that were pathogen-stimulated. As expected, we obtained the highest concordance (95.5%) with eQTLGen when comparing to our bulk-like unstimulated PBMC scRNA-seq data, i.e. taking the average gene expression across all cells from one individual in the untreated condition (**Fig. 3a**). We then saw only a minor drop to 94.7% concordance when comparing eQTLs from eQTLGen with our pathogen-stimulated (24h CA) bulk-like scRNA-seq dataset (**Fig. 3b**) and a further decrease to 92.6% when comparing to our pathogen-stimulated and cell-type-specific scRNA-seq dataset (24h CA in monocytes) (**Fig. 3c**). Finally, to verify that our initial selection of eQTLGen lead-eSNPs did not confound our conclusions, we also compared the output of a genome-wide *cis*-eQTL discovery (**Supplementary Table 8**) in pathogen-stimulated and cell type-specific scRNA-seq data (24h CA in monocytes) with eQTLGen. In this analysis, the concordance decreased a little bit further to 87.4% (**Fig. 3d**). Although up to 19.6% of the eQTLs were only detected in the genome-wide discovery (and not in the eQTLGen lead-eSNP-confined *cis*-eQTL discovery), these unique *cis*-eQTL gene sets were not enriched for specific biological pathways. Altogether, this indicated that the eQTLGen lead-eSNP confinement only had a minimal impact on our observations and confirmed our initial assumption that

the majority of context-specific eQTLs identified by our current study can already be detected in very large bulk RNA-seq datasets. However, we still require single-cell data to pinpoint their relevant context. As the eQTLGen lead-eSNP *cis*-eQTL analysis identified 1.5x more eQTLs, while showing no clear bias towards common eQTLs rather than cell type-specific or context-dependent eQTLs (**Supplementary Fig. 4a**), we continued our analysis with these results.

The CD4+ T cells revealed the most eQTL effects, followed by the monocytes and CD8+ T cells (**Supplementary Fig. 4a**). The cell types with the lowest frequencies, the DCs and B cells, also showed the lowest number of eQTLs (**Supplementary Fig. 4b**). This large difference in the number of identified eQTLs per cell type is, at least in part, explained by the differenced in power, given the number of cells of each cell type (**Supplementary Fig. 1g**). When overlapping the identified eQTL genes in the major cell types with each of their two subcell types (over all stimulation-timepoint combinations combined), we observed that the majority of eQTL genes identified in the subcell types were already detected in the corresponding major cell type (**Supplementary Fig. 4c**). Nevertheless, 4.6% (for the NKdim) up to 24.5% (for the pDCs) additional eQTL genes were uniquely identified in such a subcell type.

In addition to differences between cell types, we also observed differences between stimulation-timepoint combinations (**Supplementary Fig. 4d**). However, direct comparisons of the number of eQTLs between conditions within the same cell type were complicated because the number of included individuals varied among the stimulation-timepoint combinations as a result of QC dropouts (UT: 104 individuals, 3h CA: 120 individuals, 3h MTB: 104 individuals, 3h PA: 112 individuals, 24h CA: 119 individuals, 24h MTB: 112 individuals, 24h PA: 111 individuals). Most interestingly, when comparing the effect of pathogen stimulation on the number of identified eQTLs between cell types, we observed an inverse correlation with the responsiveness of that cell type to pathogen stimulation (**Supplementary Fig. 4e**). For example, the myeloid cells showed the largest DE response upon pathogen stimulation (**Fig. 2a**) but a consistent reduction in the number of eQTLs identified after stimulation (**Supplementary Fig. 4a**). In contrast, the lymphoid cells showed a much smaller DE response upon pathogen stimulation (**Fig. 2a**) but an increase in the number of eQTLs identified after stimulation, in about half of the conditions (**Supplementary Fig. 4a**). This could indicate that, for at least a subset of the genes, the influence of genetics on gene expression may become more restricted when cells have to orchestrate a response to an environmental stimulus.[38]

To identify eQTLs for which the strength of the eQTL effect was affected by pathogen stimulation, we performed a response-QTL (re-QTL) analysis.[39] We systematically looked for re-QTLs in all major cell types and stimulation conditions compared to the untreated condition (**Supplementary Table 9**). Most re-QTLs were specific to a particular timepoint or cell type, but less so to a particular pathogen (**Fig. 3e**). We observed that most re-QTLs were in the monocytes for each of the stimulation-timepoint combinations (**Supplementary Fig. 4a**), likely the direct

consequence of the combination of a high number of DE genes upon stimulation (**Fig. 2a**) and the relatively high number of monocytes per individual (**Supplementary Fig. 1g**). We also observed that most re-QTLs describe eQTL effects that became weaker after stimulation (**Fig. 3f**). Of those eQTL effects that became stronger after stimulation, 26.3% on average showed a significant effect that was already present in the unstimulated samples, whereas those effects were not yet present for 63.7%. Moreover, we observed clear enrichment of DE genes within the set of eQTL and re-QTL genes, but this enrichment was not consistently greater for re-QTL in comparison to eQTL genes (**Supplementary Fig. 4f**).

Finally, when linking the eSNP loci identified in each of the major cell types to GWAS output of immune-mediated diseases (see Methods), we observed a strong genomic inflation across all conditions (**Supplementary Table 10**). This genomic inflation increased further for the re-QTLs (in monocytes over all immune-mediated GWASes: p=0.024) (**Supplementary Table 10**). These findings confirmed previous studies showing that stimulation-responsive eQTL effects provide additional explanation of immune-mediated disease risk over baseline eQTLs[4,40]. Additionally, it has been shown that the effect size of GWAS-associated SNPs becomes larger in the disease-relevant context (e.g. immune-mediated disease patients as opposed to the healthy controls).[41] Therefore, also the power to detect these disease-associated effects will be larger in the disease relevant context.

In summary, we observed that 20.9% of our eQTL genes that were identified in the major cell types were influenced by a combination of genetics and environment (**Supplementary Table 7, Supplementary Table 9**). We expect this percentage is an underestimate, as the power to detect re-QTLs is inherently lower than that of eQTLs and exposure to additional environmental stimuli may reveal additional context-dependency. Altogether, our findings indicate that, in addition to cell type-specificity, context-dependency is also a major driver of genetic regulation of gene expression and provides additional explanation of disease risk.

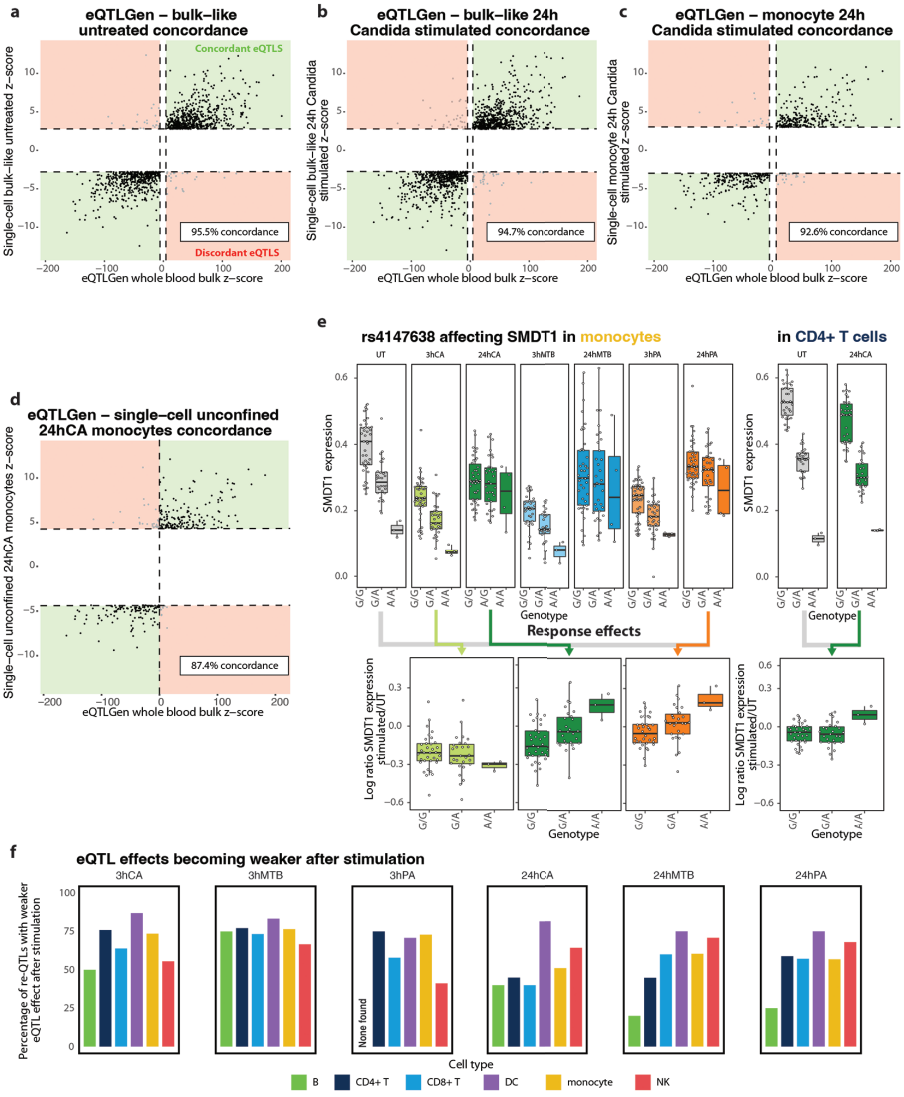


Figure 3. eQTLs and re-QTLs upon pathogen stimulation. *Concordance between the eQTLs identified in 31,684 bulk whole blood samples of the eQTLGen consortium and:* ***a.*** *those identified in our eQTLGen lead-eSNP discovery of bulk-like unstimulated PBMC scRNA-seq data,* ***b.*** *those identified in our eQTLGen lead-eSNP discovery of bulk-like 24h C. Albicans (CA)-stimulated PBMC scRNA-seq data,* ***c.*** *those identified in our eQTLGen lead-eSNP discovery of monocyte 24h CA-stimulated PBMC scRNA-seq data, and* ***d.*** *those identified in our genome-wide eQTL discovery of monocyte 24h CA-stimulated PBMC scRNA-seq data. e. Box plots showing the effect of the rs4147638 genotype on SMDT1 expression in the untreated (UT) condition and each of the six stimulation-timepoint combinations in the monocytes (left) or for the UT and 24h CA condition in the CD4+ T cells (right). Box plots show median, first and third quartiles, and 1.5 × the interquartile range, and each dot represents the average expression of all cells per cell type and individual. Stars indicate a significant*

effect (FDR < 0.001). The log ratio of SMDT1 expression in the UT cells vs a specific stimulation-timepoint combination is shown in the bottom. Colored arrows indicate which specific stimulation-timepoint combination was selected for the corresponding re-QTL boxplot. ***f.*** *The proportion of re-QTLs of which the eQTL effect became weaker after stimulation, split per cell type and stimulation-timepoint combination. eQTL summary statistics for eQTLGen-confined analysis, genome-wide analysis and response-QTL analysis can be found in* ***Supplementary Table 7****,* ***Supplementary Table 8*** *and* ***Supplementary Table 9****, respectively. The number of individuals and cells included in each analysis can be found in the* ***Source Data*** *file.*

Pathogen stimulation induces widespread context-specific gene regulation

We have previously shown that genetics can influence the co-expression relationship between genes and that scRNA-seq data is uniquely suitable to do so by taking the individual cells per cell type per donor as observations over which the individual-specific co-expression is calculated.[12] In contrast, bulk RNA-seq data usually contains a single measurement per donor, and therefore, co-expression in bulk data cannot be calculated at the individual level. As a consequence, the co-expression between two genes as calculated from bulk RNA-seq data may be different from the true individual-specific co-expression relationship as extracted from scRNA-seq data (due to Simpson's paradox[42]).

In addition, studies that compared co-expression in healthy versus disease states have indicated that environmental conditions may also impact these gene-gene interactions.[43] Here, we took the next step by determining whether and how the combination of genetics and environment may affect how genes are interacting with one another by performing co-expression QTL analysis, i.e. a SNP genotype affecting the co-expression relationship of a gene pair. For this purpose, we selected a subset of 49 SNP-gene combinations that we then tested against up to 5,772 genes. To enrich for SNP-gene combinations in which we expect an interaction with the environment, we selected these based on: 1. the gene being DE and 2. the SNP-gene combination being a re-QTL in at least one of the stimulation-timepoint combinations; 3. the gene being expressed in at least 50% of the individuals (in each 10X chemistry). For this analysis, we focused solely on the monocytes because this was the cell type that showed a strong response to pathogen stimulations and for which we had sufficient cells per individual (i.e. hundreds) to perform a robust co-expression QTL analysis. By making this pre-selection of 49 SNP-gene combinations, we could reduce the multiple testing burden from over 10^{14} in a genome-wide analysis to fewer than 283,000 tests.

Across the unstimulated condition and each of the six stimulation-timepoint combinations, we found at least one co-expression QTL for 35 SNP-gene combinations and more than 100 co-expression QTLs in at least one condition for 9 SNP-gene combinations. For each of these 9 SNP-gene combinations with a high number of co-expression QTLs, we observed an interaction between genotype and stimulation condition (**Fig. 4a**, **Supplementary Table 11**). One of these co-expression QTLs described an interaction between RPS26 and rs1131017, which was an effect in high

LD with one we had identified as a co-expression QTL in CD4+ T cells in our previous study (rs7297175, $R^2 = 0.92$). [12] rs1131017 was previously associated with rheumatoid arthritis[44] ($p = 1.3x10^{-8}$) and is in high LD with a type I diabetes GWAS SNP[45] (rs11171739, $R^2 = 0.94$). For this *RPS26*-rs1131017 SNP combination, we found 1,701 co-expression QTLs in the unstimulated condition. Of the 106 *RPS26* co-expression QTLs that we had previously identified in CD4+ T cells[12], 72 (67.9%) were also found in the unstimulated monocytes in our current study (91.7% with the same allelic direction) (**Supplementary Fig. 5a**). Any discrepancy between these two cell types might be the consequence of distinct regulatory mechanisms that are active in those cell types. Next, looking at the effect of stimulation, the number and strength of the detected *RPS26* co-expression QTLs reduced greatly after stimulation and was related to the duration of stimulation: on average we observed 459 co-expression QTLs after 3h stimulation and 112 after 24h stimulation (**Fig. 4a**, **Supplementary Fig. 5b**).

We also observed this general decrease in the strength and number of co-expression QTLs with increasing duration of pathogen stimulation for the *HLA-DQA2* co-expression QTLs, but not for any of the other 7 SNP-gene combinations (**Fig. 4a**). These other 7 co-expression QTL effects increased in strength and numbers upon stimulation (**Fig. 4a**, **Supplementary Table 11**). Interestingly, for some of these co-expression QTL genes, we observed the most prominent increase at 3h stimulation (i.e. *CLEC12A*, *CTSC* and *NDUFA12*), whereas others were more prominent at 24h stimulation (i.e. *TMEM176A/B*, *DNAJC15* and *HLADQA1*). The observation of different numbers of co-expression QTLs for a specific gene over the 7 stimulation-timepoint combinations was not fully explained by the expression level of that gene. Beyond this variation over the timepoints, we also observed clear differences between the various pathogen stimulations. At gene-level, there was little overlap between the co-expressed gene sets between the different pathogen stimulations (**Supplementary Table 11**), whereas this overlap was much larger at pathway-level (**Supplementary Fig. 5**). The low gene-level overlap is likely a consequence of power and is something that will be largely overcome in the near future with the increase in the number of cells per dataset.[46,47] Together, these results indicate that specific environmental conditions can fulfill the requirements needed for a specific co-expression QTL interaction to occur.

Previously, re-QTL analyses in cells exposed to highly specific stimuli were used to disentangle the environmental conditions that underlie specific genetic regulation of gene expression.[4,16] However, this has the disadvantage that either many highly specific stimuli have to be applied, or, in the case of applying broad stimuli, the exact environmental context relevant for the interaction remains vague. Here, we propose using co-expression QTL analysis upon stimulation with a few broad stimuli to gain this detailed insight in a more unbiased way, without the need to apply many highly specific stimuli. As a first example of how co-expression QTL analysis can help us understand the underlying mechanisms of gene regulation, we focused on the *CLEC12A* co-expression QTLs affected by SNP rs12230244, which were most prominent at 3h of pathogen stimulation (**Fig. 4a**, **4b**). *CLEC12A*, also known as *MICL*, encodes for an inhibitory

C-type lectin-like receptor and is mostly expressed in myeloid cells such as monocytes and DCs. CLEC12A signaling can be activated by the binding of uric acid crystals, which are the byproduct of nucleic acids that can be released from damaged or dying cells.[48,49] Activation of CLEC12A signaling can result in inhibition of the activating C-type lectin receptors and can prevent hyperinflammation during necrosis.[50]

To identify the potential causal factor underlying the *CLEC12A* co-expression QTL, we performed a pathway analysis on the associated co-expressed gene set of each of the stimulation-timepoint combinations. We hypothesized that co-expressed genes linked to the same co-expression QTL mostly describe the same (or only a few) biological processes that are driven by a single (or a few) causal factors being directly involved, and that most of these co-expressed genes are themselves just a consequence of being highly co-expressed with the causal factor. An important category of causal factors are transcription factors. However, average expression levels of transcription factors are generally low and, particularly in dynamic situations such as a pathogen response, mRNA levels might not correlate well with the nuclear protein expression levels (i.e. the functional proportion).[51,52] Consequently, it can be difficult to define the direct causal factor solely using co-expression QTL analysis. Nevertheless, we expected that by taking a pathway-level view, the downstream genes of transcription factors would have a high correlation with the functional protein level of the transcription factor and would be more easily picked up than a single gene.

The pathway analysis of the *CLEC12A* genotype-dependent co-expressed gene set after 3h stimulation (Benjamini-Hochberg (BH) corrected $p = 2.9x10^{-5}$, $8.7x10^{-7}$ and $4.3x10^{-4}$ for 3h CA, 3h PA and 3h MTB, respectively), but not in the untreated or 24h stimulation conditions, revealed enrichment of the IFN pathway (**Fig 4c**). This result hinted that a component within or regulating the IFN pathway could be the causal factor that is regulating the different *CLEC12A* co-expression responses per genotype after 3h stimulation. To provide additional support for this hypothesis, we performed a functional enrichment analysis for putative transcription factor binding sites (TFBSs) (TRANSFAC database[53]) on the *CLEC12A* genotype-dependent co-expressed genes upon 3h stimulation. We divided this gene set into a subset in which individuals with the TT as opposed to the AA genotype showed a more positive rather than a more negative co-expression relationship between *CLEC12A* and its co-expressed genes, as potentially different mechanisms could be underlying these gene sets. This analysis revealed no enriched TFBSs in the negative-strength gene set, but a clear enrichment of various IFN regulatory factors (IRFs) in the positive-strength gene set, including *IRF1*, *2*, *4*, *5*, *7* and *8* (**Fig. 4d**). Additionally, when overlapping the *CLEC12A* co-expression QTL SNP rs12230244 and its accompanying (near-)perfect LD SNPs with putative TFBSs**54**, we observed several transcription factors that may bind to the genomic location of these SNPs. Most noticeably, the predicted binding site of *IRF1* was shown to be enriched in the genomic location of two SNPs that are in near-perfect LD with the *CLEC12A* co-expression QTL SNP: rs999185 ($R^2 = 0.9943$) and rs57106602 ($R^2 = 0.9$).

Finally, we used two external datasets and slightly different approaches to further strengthen our hypothesis that IFN activity is regulating the *CLEC12A* co-expression QTL effects. First, we used the BIOS consortium bulk RNA-seq dataset containing whole blood data from 3,553 individuals.[55] For each of those individuals, we calculated a polygenic risk score (PRS) for the autoimmune disease systemic lupus erythematosus (SLE), a disease characterized by increased type I IFN activity.56–59 We reasoned that the genetic risk captured by the SLE PRS could be used as a proxy for IFN activity. Consequently, the difference in the co-expression relationship between the SLE PRS and the *CLEC12A* per rs12230244 genotype indicated the involvement of IFN signaling in this interaction (**Fig. 4e**). Second, we used an independent scRNA-seq dataset generated in 68 healthy controls and 117 SLE patients from European (EUR) and East Asian (EAS) origin. We reasoned that since IFN activity is characteristic for SLE[56–59], SLE patients would mimic the 3h pathogen stimulation state in which high IFN activity seems to drive the observed *CLEC12A* co-expression QTL effects. We also reasoned that the healthy controls would mimic the untreated cells in our study and therefore show fewer *CLEC12A* co-expression QTL effects driven by IFN activity. To define whether the SLE patients mimicked the results as observed after 3h pathogen stimulation, we performed a co-expression QTL analysis for *CLEC12A* and SNP rs12230244 in the monocytes of SLE patients and healthy controls (**Supplementary Table 12**). Pathway analysis on the *CLEC12A* co-expression QTL genes revealed stronger enrichment for the IFN pathway in the SLE patients (FDR = 1.965×10^{-7}) compared to the healthy controls (FDR = 1.203×10^{-3}), again supporting that this pathway is involved in the regulation of *CLEC12A* through the locus with the rs12230244 SNP.

As a second example, we applied a similar strategy to learn the underlying regulatory mechanism by which the co-expression QTLs identified for SNP rs6945636 affect the heat-shock protein response gene *ZFAND2A*. The heat-shock protein response is a pathway that, amongst others, can be activated by bacterial and viral infections.[60] We selected this specific SNP-gene combination for further analysis because the co-expression QTLs identified were both pathogen- and timepoint-specific (only at the 24h timepoint, 96% of the genes being detected in CA only). Pathway analysis of the co-expressed genes revealed 'Intracellular pH reduction' (GO:0051452) as the top associated biological process (adjusted p = $3.8\text{x}10^{-4}$). Interestingly, *HSF1*, a known regulator of *ZFAND2A*[61] was shown to be pH sensitive in yeast.[62] Moreover, the *ZFAND2A*-associated co-expression QTL SNP rs6945636 was in almost perfect LD with previously identified *HSF1* binding sites in K562 cells (R^2 = 0.99, rs715188378; R^2 = 0.99, rs79849558; R^2 = 0.99, rs11767061, retrieved from dbSNP release 153).[63] Together, this indicates that CA-induced pH regulation activated *HSF1,* which in turn bound with stronger (TT genotype) or weaker (AA genotype) strength to rs6945636 SNP locus, and thereby strongly or weakly activated *ZFAND2A*, respectively.

These two examples provide clear use cases for how co-expression QTL analysis can be applied to gain detailed insights into the underlying context of gene expression regulation. For example, in the case of *CLEC12A*, without co-expression QTL analysis we could only reveal that *CLEC12A* is a re-QTL regulated by a factor active 3h downstream of pathogen stimulation (**Supplementary Table 9**). In contrast, using co-expression QTL analysis, we were directed to the causal regulatory factor for this re-QTL. This enabled follow-up analyses that gathered solid evidence for the following mechanism of action through which the rs12230244 SNP locus affects *CLEC12A* expression specifically upon 3h pathogen stimulation: 1. pathogen-associated molecular patterns bind to a pattern recognition receptor (PRR) and initiate a signaling cascade that eventually results in phosphorylation of interferon regulatory factors (IRFs), 2. phosphorylated IRF then translocates to the nucleus where it binds to specific DNA motifs such as IFN-stimulated response elements, and 3. this can then activate transcription of IFNs and IFN-stimulated genes (ISGs). Additionally, IRF is expected to bind to a region containing SNP rs12230244 (or any another SNP in high LD), thereby regulating *CLEC12A* expression. In this case, depending on the SNP genotype, the IRF binding and induction of *CLEC12A* expression is expected to be stronger (TT genotype) or weaker (AA genotype) (**Fig. 4f**).

Interestingly, we identified a number of (near-)genome-wide significant PheWAS traits related to immune cell composition and size to be associated with these two co-expression QTL loci (extracted from 452,264 White British individuals of the UK Biobank[64]): platelet counts ($p = 2.1x10^{-8}$), monocyte percentage ($p = 1.8x10^{-5}$) and eosinophil counts ($p = 6.4x10^{-5}$) for *CLEC12A* and mean corpuscular volume ($p = 1.5x10^{-14}$) and mean sphered cell volume ($p = 2.6x10^{-9}$) for *ZFAND2A*. However, no direct association was found for any of the immune-related GWAS tested (SLE[65], inflammatory bowel disease[66], celiac disease[67], rheumatoid arthritis[44], multiple sclerosis[68], type I diabetes mellitus[69] and candidemia[70], which had 10-2,541-fold smaller sample sizes than the PheWAS. This overlap with immune-related PheWAS traits indicate the relevance of these SNPs for immune function.

Moreover, looking at the function of the affected genes, we also expect immunological consequences of the identified co-expression QTLs. For example, previously *CLEC12A* was shown to act as an early adaptor molecule for antibacterial autophagy, and in mice, complete knockout of Clec12a resulted in higher susceptibility to Salmonella infection.[71] Additionally, *CLEC12A* is known to contribute to the pathogenesis of rheumatoid arthritis. For example, upon collagen-induced arthritis, *CLEC12A* knockdown mice show increased joint inflammation[72] and in monocytes of early rheumatoid arthritis patients reduced expression of *CLEC12A* correlated with more severe disease 6 months later.[73] Together, this suggests that individuals with the AA allele on rs12230244 may be at increased risk of bacterial infection and of developing joint inflammation, acting through reduced induction of *CLEC12A* expression when exposed to pathogens or other factors inducing IFN signaling.

Summarized, using co-expression QTL analysis, we can now dissect the underlying mechanism by which such an effect is regulated. This information will help explain the downstream consequences on immune function, and potentially enable new routes for medical intervention.

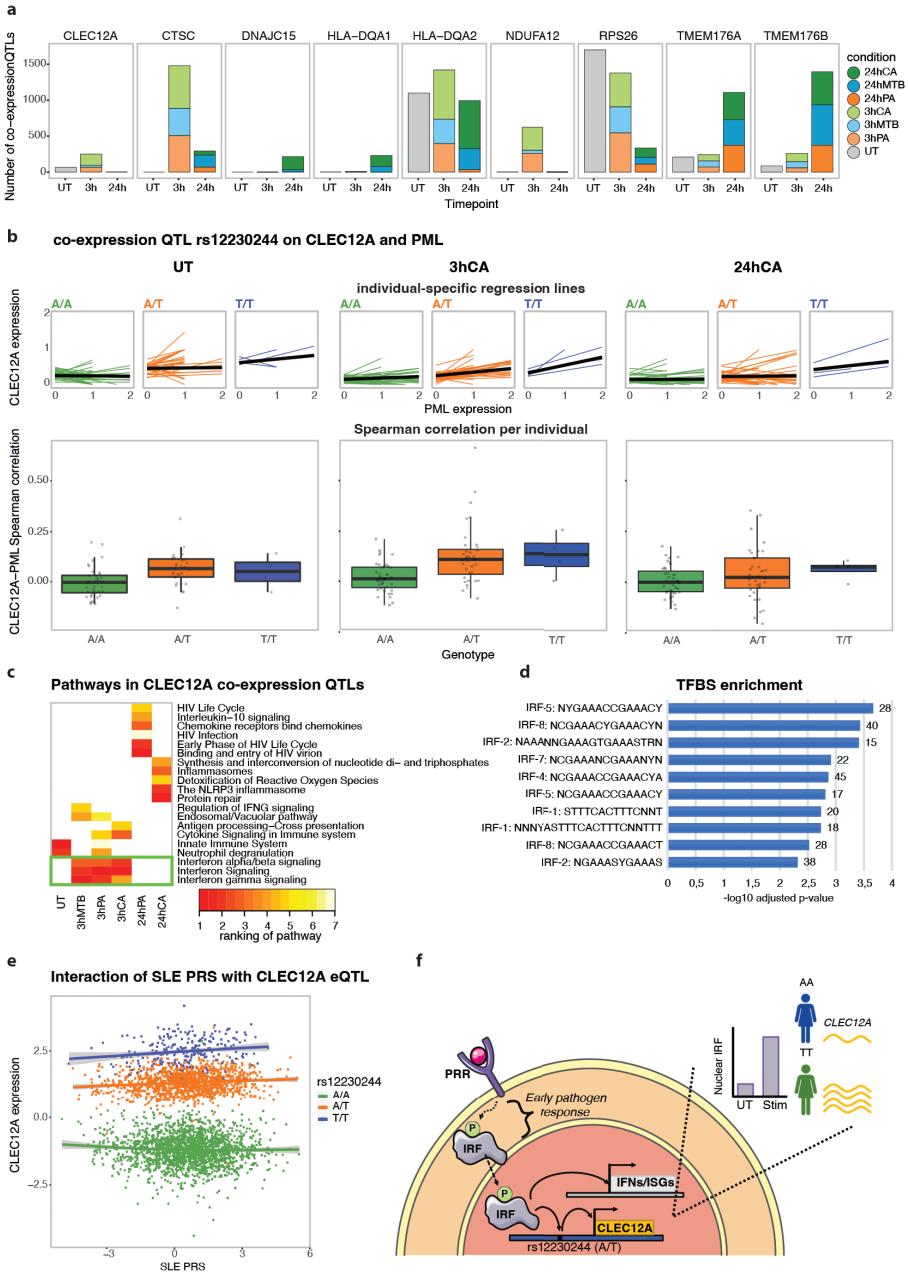


Figure 4. Interferon regulatory factor affects CLEC12A co-expression QTLs upon 3h pathogen stimulation in monocytes. *a. Number of co-expression QTLs identified in each of the stimulation-timepoint combinations for those co-expression QTLs with over 100 co-expression QTLs in at least one condition. The 3h and 24h timepoint are colored by pathogen stimulation (green: C. Albicans (CA), blue: M. Tuberculosis (MTB), orange: P. Aeruginosa (PA). Co-expression QTL summary statistics can be found in* ***Supplementary Table 11****. b. The lines in the top plots show co-expression between CLEC12A and PML (most significant co-expression QTL across the 3h stimulation conditions) for individual cells in the untreated (left), 3h CA (middle) and 24h CA (right) condition. In these plots, individual-specific regression lines are shown, split by genotype. The average genotype-specific regression lines are shown in black. The bottom boxplots depict Spearman's rank correlation between CLEC12A and PML expression, stratified by SNP rs12230244 genotype in the monocytes per individual, in the untreated (left), 3h CA (middle) and 24h CA (right) stimulated cells (the V2 chemistry data is plotted). Each data point shows a single individual. Box plots show median, first and third quartiles, and 1.5× the interquartile range.* ***c.*** *Heatmap of the top-5 enriched pathways within the co-expressed CLEC12A co-eQTL genes per stimulation-timepoint combination. Per combination, pathways are ranked based on significance. White indicates that the pathway was not found to be enriched in that specific stimulation-timepoint combination. The green box highlights pathways that are associated with all 3h stimulation conditions.* ***d.*** *Top 10 enriched putative transcription factor binding sites within the CLEC12A co-expression QTL genes that: 1. showed a more positive strength of the co-expression relationship in individuals with the TT as opposed to the AA genotype and 2. were identified in the 3h stimulated (outer join) monocytes using the TRANSFAC database. Enrichment of putative transcription factor binding sites was defined using a g:SCS multiple testing correction method, applying a significance threshold of 0.05.* ***e.*** *Co-expression QTL analysis for CLEC12A-SNP rs12230244 against the SLE PRS (calculated using those SLE GWAS SNPs with a P-value threshold of $<5x10^{-8}$) using whole blood bulk expression data from 3,553 individuals (BIOS consortium). A one-tailed F-test (coefficient=0.04, standard error=0.01, f-value=19.60, p-value=$9.84x10^{-6}$, R^2=0.84) was used to determine whether the distribution of the squared residuals with the SLE PRS as interaction term was significantly smaller than without.* ***f.*** *Proposed mechanism of action of CLEC12A co-expression QTLs. When pathogen-associated molecular patterns bind to a pattern recognition receptor (PRR), a signaling cascade is initiated that eventually results in phosphorylation of interferon regulatory factors (IRFs). Phosphorylated IRF then translocates to the nucleus, where it binds to specific DNA motifs such as IFN-stimulated response elements. This can then activate transcription of IFNs and IFN-stimulated genes (ISGs). Additionally, IRF is expected to bind to a region containing SNP rs12230244 (or any another SNP in high LD), thereby regulating CLEC12A expression. In this case, depending on the SNP genotype, the IRF binding and activation of CLEC12A expression is expected to be stronger (TT genotype) or weaker (AA genotype). Many of the identified CLEC12A co-expression QTL genes are involved in the IFN pathway (see* ***4b****). This has to be the result of a common upstream factor (i.e. IRF) of CLEC12A transcription that can also activate IFNs and ISGs, but cannot be the result of a downstream regulator because this would have led to trans-eQTL effects for the same SNP rs12230244 (which we do not observe). The number of individuals and cells included in each analysis can be found in the* ***Source Data*** *file.*

Discussion

GWAS studies have provided important insights into the genetic architecture of phenotypic traits and diseases.[1] However, the exact mechanisms by which genetic variation leads to these traits or diseases largely remain a black box. To uncover these mechanisms, various approaches have been successfully applied, for example coupling the trait-associated risk factor to the nearest positional gene[74], downstream gene expression[55], or gene regulation.[12] Nevertheless, a large knowledge gap remains that may, in part, be filled by taking into consideration the context in which the genetic variant can lead to disease.[7,16,17]

To uncover the interplay between genetics and cellular and environmental context, we single-cell RNA-sequenced PBMCs from 120 individuals from Lifelines, a large population-based cohort from the Northern Netherlands, that had been exposed to various pathogens or left untreated. Subsequent DE, eQTL and co-expression QTL analysis revealed that there are widespread interactions between an individual's genetics and the cellular and environmental context, both at the level of gene expression and in its regulation. We identified hundreds of eQTLs in the individual cell types and upon pathogen-stimulation and observed strong context-specificity for 25.7% of the identified co-expression QTLs. In general, we observe more interactions between genetics and cell type-specific context, as opposed to context induced by pathogen stimulation. However, some of these differences may have been the result of differences in detection power. Contrary to expectations, in the cell types with the strongest response to pathogen stimulation (i.e. the myeloid cells), the total number of eQTLs was reduced after stimulation. Moreover, in all cell types, we observed that eQTLs more often became weaker rather than stronger after stimulation and that neither category of eQTL genes was associated with a specific pathway. In contrast, for the co-expression QTL genes, the number of co-expressed genes more often increased upon pathogen stimulation. However, this might in part have been the result of our selection, i.e. choosing re-QTLs in monocytes as the starting point for the co-expression QTL mapping. Moreover, we observed genomic inflation of eQTLs that further increased when focusing solely on the re-QTLs. Altogether, these observations indicate that context, here the pathogen-stimulation condition, is an important contributor that affects the association between SNPs and gene expression or co-expression, and that taking this context in consideration further improves our understanding of disease risk.

A major advantage of co-expression QTL analysis as opposed to re-QTL analysis is that we do not require many highly specific stimuli to disentangle the mechanisms that underlie the context-specificity of the genetic regulation. Instead, in this study, we have shown that, after applying a broad stimulation (i.e. whole-pathogen stimulation), a wide range of contexts are activated, and that, through subsequent co-expression QTL analysis, the specific context and mechanism of action could be uncovered. For example, we revealed that an interferon-regulated transcription factor was affecting the SNP rs12230244-dependent downstream activation of *CLEC12A*. Additionally, we showed how pH-dependent regulation of the heat shock protein response transcription factor *HSF1* affected the SNP rs6945636-dependent downstream activation of ZFAND2A. Even though the causal SNP cannot be conclusively determined using co-expression QTL analysis, understanding the underlying mechanism can help to further fine-map the genetic signal. These examples clearly show the potential of the technology and provide an outlook into where the field will be moving as more population-scale scRNA-seq datasets become available. We foresee that newly developed methodology, such as inCITE-seq[52] and NEAT-seq[75], combining measurements of multiple omics layers from the same cell,

including RNA and nuclear protein levels (which allows measuring active transcription factors levels), will further enhance the interpretability of the identified co-expression QTLs in the future.

Importantly, this study was conducted in European individuals with a white background. Although we do not expect general conclusions to be different in other populations, it may be that the upstream regulators or downstream consequences of some of the specific genetic variants act differently across populations. Moreover, as the infection history with the three pathogens under study is unknown for the individuals included in our study, there is a small chance that this may have introduced additional noise or confounding in our analyses.

In the last few years, scRNA-seq has become a mature, high-throughput technology.[8,9] This has led to several initiatives aiming to study population genetics at single-cell resolution, such as the sc-eQTLGen consortium[46] and others.[76] Such efforts bring together many single-cell eQTL studies, conducted in individuals from different ethnicities and exposed to different environments or diseases. This will not only increase the power to detect eQTLs and co-expression QTLs, it will also further extend our findings to additional contexts and enable genome-wide cell-type and context-specific *trans*-eQTL mapping. Moreover, instead of linking individual genetic variants, linking of polygenic risk scores to cell-type-specific gene expression (i.e. eQTS analysis[37]) may provide a more disease-focused insight into how the combination of disease-associated variants together contribute to changes in gene expression levels. By integrating GWAS signals, PRS scores and context-specific QTL information, we expect that these efforts can drive major leaps forward in disease understanding and precision medicine.[77]

Methods

Ethics approval and informed consent

The LifeLines DEEP study was approved by the ethics committee of the University Medical Centre Groningen, document number METC UMCG LLDEEP: M12.113965. All participants signed an informed consent from prior to study enrollment. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

PBMC collection and stimulations

Whole blood from 120 European white background individuals of the northern Netherlands population cohort Lifelines Deep[78] was drawn into EDTA-vacutainers (BD). PBMCs were isolated and maintained, as previously described.[12] In short, PBMCs were isolated using Cell Preparation Tubes with sodium heparin (BD) and were cryopreserved until use in RPMI1640 containing 40% FCS and 10% DMSO. After

thawing and a 1h resting period, unstimulated cells were washed twice in medium supplemented with 0.04% BSA and directly processed for scRNA-seq. In contrast, for stimulation experiments, $5x10^5$ cells were seeded in a nucleon sphere 96-well round bottom plate in 200 μl RPMI1640 supplemented with 50 μg/mL gentamicin, 2 mM L-glutamine and 1 mM pyruvate. Then, *in vitro* stimulations were applied for either 3h or 24h using $1x10^6$ CFU/ml heat-killed *C. albicans blastoconidia* (strain ATCC MYA-3573, UC 820), 50 μg/ml heat-killed *M. Tuberculosis* (strain H37Ra, Invivogen) or $1x10^7$ heat-killed *P. Aeruginosa* (Invivogen) while incubating the cells at 37˚C in a 5% CO_2 incubator. After stimulations, cells were washed twice in medium supplemented with 0.04% BSA. Cells were then counted using a haemocytometer, and cell viability was assessed by Trypan Blue.

Single-cell library preparation and sequencing

105 sample pools were prepared, each aimed to yield 1,400 cells/individual from 8 individuals (11,200 cells). In general, pools contained a mixture of both sexes and two different stimulation conditions. Each sample pool was loaded into a different lane of a 10x chip (Single Cell A Chip Kit for v2 or Single Cell B Chip Kit for v3 reagents). The 10x Chromium controller (10x Genomics), in combination with v2 (72 libraries) or v3 (33 libraries) reagents, was used to capture the single cells and generate sequencing libraries, according to the manufacturer's instructions (document CG00052 and CG000183 for v2 and v3, respectively) and as previously described.[12] Sequencing was performed with a 150 bp paired-end kit using a custom program (V2: 27-9-0-150, V3: 28-8-0-150) on the Illumina NovaSeq 6000 at BGI (Hong Kong).

scRNA-seq alignment, preprocessing and QC

CellRanger v3.0.2 was used with default parameters to demultiplex, generate FASTQ files, align reads to the hg19 reference genome, filter both cell- and unique molecular identifier (UMI) barcodes and count gene expression per cell. To assign cells to one of the eight individuals in a lane, Demuxlet was used.[11] The genotype information used by Demuxlet was previously generated as described in Tigchelaar et al.[78] and was phased with Eagle v2.322 using the HRC reference panel and the Michigan Imputation Server. Only exonic variants with a MAF of at least 0.02 were used for demultiplexing. Subsequently, Souporcell v1.0[29] was used to remove doublets coming from different individuals, by looking for the different genotypes within a single cell assignment. We limited the SNP calling to positions that were also used for demultiplexing.

Version 3.1 of the Seurat[79] package was used for further quality control and processing. Due to mRNA capture differences between the v2 and v3 chemistries, a version-chemistry-specific maximum mitochondrial gene content percentage of 8% and 15% was used, respectively. Cells with less than 200 detected genes were

discarded, as well as cells with more than 9 UMIs mapping to the hemoglobin subunit beta (HBB) gene (representing red blood cells), and other low-quality cells (i.e. clusters of cells with a low number of expressed genes and a relatively high mitochondrial content, or missed, likely same-individual, doublets) (**Supplementary Table 3**).

For annotating the cell types, we first log-normalized the count matrices for each of the seven timepoint-stimulation conditions and two chemistries separately using Seurat's LogNormalize function (scale.factor = 10,000).[79] The log-normalized count matrices of the unstimulated data were then integrated separately for each of the three pathogen stimulations. For this, we used the first 30 dimensions from a Canonical Correlation Analysis to identify integration anchors in Seurat's FindIntegrationAnchors function. These anchors were then used for integration using Seurat's IntegrateData function.[79] We performed principal component analysis (PCA) and selected the first 30 PCs to identify the cell clusters using k-nearest neighbor clustering and visualized this in UMAP space (using the default settings). Cell types were assigned to each cluster based on marker gene expression, resulting in a set of six major cell types and ten subcell types (**Supplementary Fig. 1B, Supplementary Table 4**). A small fraction of the cells could not be classified at higher resolution, and therefore, where omitted from the subcell type analyses (**Source Data** file). For each version chemistry, gene expression counts were then SCT normalized using Seurat's SCTransform function, and cell type labels obtained from the integrated data were transferred to non-integrated data (**Fig. 1A**), to preserve the stimulation response at the gene expression level.

Differential expression: mapping, pathway enrichment and module scoring

For each pathogen-timepoint combination, major an subcell type and 10X chemistry, differential expression (DE) analysis was performed between the pathogen-stimulated and the untreated condition using the MAST implementation of Seurat.[30] Testing was limited to genes with a log-fold change (LFC) >0.1 and with expression in at least 10% of the cells. We used MetaVolcanoR[80] to perform a meta-analysis for each cell type, taking the results of the v2 and v3 chemistries as inputs for Fishers Combined Probability Test.[81] Significance was determined by taking a Bonferroni-corrected p-value of <0.05 within the meta-analysis. When an analysis could only be performed in one version chemistry, only that output is reported.

Per cell type, the resulting DE gene set was split in up- and downregulated genes after stimulation, which was then used as input for a pathway enrichment analysis with ToppFun, selecting the REACTOME database.[82] To calculate statistical significance, the probability density function was used, selecting those pathways that had a BH-corrected p-value <0.05.

For the major cell types, the enriched pathways were visualized by calculating the LFC in average gene expression in all pathogen-timepoint conditions compared to

the untreated condition and clustered these results using hierarchical clustering with the complete linkage method. For the subcell types we made the comparisons only within the subtypes that fall within the same major cell type (**Supplementary Table 4**). We visualized up to 10 enriched pathways that showed the largest difference between the two subtypes (within the same major cell type), and ordered these pathways by the difference in log10 transformed significance between the cell types. The fraction of genes that were found to be differentially expressed versus the total annotated genes in the gene sets, was determined by dividing the differentially expressed genes found for each gene set, by the total number of genes of a gene set.

Calculation of pathway activity was done using the module score function of Seurat[83], by calculating, per cell, the combined activity of a specific gene set annotated to be part of a pathway in the REACTOME database. This score was then averaged per donor for each condition and cell type.

eQTL and re-QTLs: mapping and GWAS enrichment

The mapping of eQTLs was performed in a bulk-like and cell-type-specific manner. We limited the analysis to the top independent effects identified in the eQTLGen meta-analysis on 31,684 individuals, resulting in the testing of 16,987 possible SNP-gene pair combinations.[55] These SNP-gene combinations identified by eQTLGen were the result of genome-wide *cis*-eQTL mapping of SNPs within a 100 kb distance to the gene midpoint, MAF >0.1, call rate >0.95 and Hardy-Weinberg equilibrium p-value >0.001. These 16,987 SNP-gene pairs were then further filtered to only include SNPs with a minor allele frequency (MAF) >0.1 or genes that were expressed in least three cells in our single-cell data. Filtering of SNP-gene combinations and mapping of eQTLs were done separately for each cell type and reagent version chemistry using the averaged, normalized gene expression values per individual, cell type and stimulation-timepoint combination. This was followed by a sample-weighted meta-analysis[84] over the v2 and v3 chemistry outputs per cell type and stimulation-timepoint combination. When an analysis could only be performed in one version chemistry, only that output is reported. eQTLs with a gene-level FDR < 0.05 were considered statistically significant, and a permutation-based strategy (n = 10) we had described before was used to control this FDR.[2] Using the same parameters described above, but without eQTLGen SNP-gene pair filtering, we also performed a genome-wide *cis*-eQTL discovery analysis.

Next, we performed re-QTL mapping, confining ourselves to the total gene set of FDR < 0.05 significant eQTLs across all cell types and conditions. For this, we calculated the log-ratio of the averaged expression of the unstimulated condition and the stimulated condition per sample, cell type and chemistry, and then applied the same mapping strategy we used to identify regular eQTLs.

To determine whether eQTLs and re-QTLs were genetically inflated, eQTLgen lead eSNPs were matched to the top GWAS SNP per locus for each of the following immune-

mediated disease GWAS studies: celiac disease[67], type 1 diabetes[45], multiple sclerosis[68], inflammatory bowel disease[66], candidemia susceptibility[70] and rheumatoid arthritis[44]. For this, the LD between eSNPs and GWAS SNPs was calculated from genotypes of the 503 European individuals in the 1000g phase3 reference panel at R2 >0.8 using Plink 1.9-beta6.[85] Lambda inflation was calculated using all GWAS p-values matched to the eQTL or re-QTL SNPs. To determine whether there is a difference in genomic inflation for those SNPs whose eQTL effect changes upon stimulation (re-QTLs), we compared the genomic inflation of the re-QTL SNPs with the non-re-QTL overlapping eQTL SNPs that were tested in both the unstimulated and relevant stimulation condition and significant in either. Using the different conditions and GWASes, specifically for the monocytes, the distributions of lambda values for the re-QTL and non re-QTL sets were compared using a two-sided Wilcoxon Rank Sum Test. This statistical testing was solely performed in monocytes, as this was the cell type with a strong pathogen response and the largest set of identified re-QTL SNPs, expecting largest effects on genomic inflation and allowing for the most robust genomic inflation analysis.

Co-expression QTLs: mapping, pathway enrichment, TFBS and GWAS overlap

Co-expression QTL mapping was performed in the monocytes on a subset of SNPs and genes, selected based on their being: 1. DE and 2. a re-QTL in at least one of the stimulation-timepoint combinations; 3. expressed in at least 50% of the individuals (for each 10X chemistry tested). This selection resulted in 49 SNP-gene combinations for which we calculated the Spearman correlation with every other gene per individual and per stimulation-timepoint condition. A weighted linear model was used in which the genotype predicts the strength of the correlation between the two genes, using the square root of the number of cells as a weight. Analysis was performed separately for the different 10X chemistries, after which betas and standard errors were meta-analyzed. When an analysis could only be performed in one version chemistry, only that output is reported. The statistical significance threshold was then determined using a permutation-based (n = 100) FDR approach. The most significant co-expression QTL p-value per stimulation-timepoint condition was then compared with the one coming from re-running the same permutations after randomly shuffling the genotype identifiers. This allowed us to calculate an eQTL gene-level FDR.[2] An FDR < 0.05 was considered statistically significant. Separate thresholds were determined for each re-QTL SNP-gene combination and each stimulation-timepoint condition.

Pathway analysis was performed on the co-expression QTL genes associated with the selected eQTL gene per stimulation-timepoint combination using Toppfun with similar settings to those described in the ‘DE and pathway analysis’ section. Significant pathways (BH-corrected p-value < 0.05) were then ranked by p-value. The rankings of the pathways for each condition were then clustered using hierarchical clustering using the complete linkage method.

Transcription factor motif enrichment analysis was performed on the 3h stimulation outer join *CLEC12A* co-expressed gene set split by having either a more positive or more negative correlation with the minor versus major allele. For this, we took information from the TRANSFAC database release 2020.2 v2[53] and used g:Profiler (version e102_eg49_p15_7a9b4d6)[86] with the g:SCS multiple testing correction method, applying a significance threshold of 0.05. Additionally, the *CLEC12A* co-expression QTL SNP rs12230244 and its accompanying (near-)perfect LD SNPs were overlapped with putative TFBSs, as defined by SNP2TFBS.**54**

Overlap of co-expression QTL SNPs (or SNPs within a 1 Mb window with LD >0.8) with disease-associated GWAS SNPs was determined by searching the GWAS catalog (https://www.ebi.ac.uk/gwas/) and an additional set of immune-related GWAS studies (celiac disease[67], type 1 diabetes[45], multiple sclerosis[68], inflammatory bowel disease[66], candidemia susceptibility[70] and rheumatoid arthritis[44]).

CLEC12A co-expression QTL validation and replication: SLE PRS interaction analysis, SLE scRNA-seq co-expression QTL analysis

Using the summary statistics of the SLE GWAS by Bentham et al.[65], we calculated the PRS for SLE in 3,553 samples from the BIOS consortium using a custom Java program, GeneticRiskScoreCalculator-v0.1.0c, as described previously.[55] Briefly, to account for LD between variants, our approach included a double clumping strategy where we first clumped variants within a 250 kb window and then within a 10 Mb window using an LD threshold $R^2 = 0.1$. We then calculated the PRS for each individual by summing the products of the number of risk alleles and the GWAS effect-size (i.e. beta) for each SLE-associated variant. We constructed the PRS using a p-value threshold for the SLE GWAS of $p < 5\times10^{-8}$. The resulting PRS was scaled between 0 and 2 for compatibility with the eQTL mapping software. We then determined whether the co-expression between *CLEC12A* and an individual's SLE PRS was modulated by SNP rs12230244. For this, we fitted a generalized linear model with and without the SLE PRS as an interaction term and determined how far the predicted model deviated from the true observation by taking the residuals of the observation. A one-tailed F-test was then used to determine whether the distribution of the squared residuals with the SLE PRS as interaction term was significantly smaller than without, meaning that the SLE PRS interacts with the *CLEC12A* co-eQTL.

We used an independent cohort of SLE patients and healthy controls (GEO accession number: GSE174188) to replicate our findings of a clear enrichment for IFN-related genes within the co-expressed gene set of the *CLEC12A*-SNP rs12230244 co-expression-QTL. This cohort contained individuals of EUR and EAS descent, including healthy individuals (EAS: 18, EUR: 58) and individuals diagnosed with SLE (EAS: 58, EUR:59) who were not in an active disease state when samples were collected. For all individuals, PBMCs were collected and cryopreserved until further use. The SLE samples were collected through the California Lupus Epidemiological Study

(CLUES) cohort. Healthy controls were collected at the UCSF Rheumatology Clinic and through the Immune Variation Consortium (ImmVar) in Boston. All UCSF samples were genotyped using the Affymetrix World LAT Array, while samples collected in Boston were genotyped using the Illumina OmniExpressExome Array. The Michigan Imputation Server was used for imputation with the Haplotype Reference Consortium version 1.1 reference. The samples collected at UCSF and Boston were processed using established protocols[11,27]. ScRNA-seq was performed using 10X Chromium Single Cell 3' V2 chemistry, as described previously.[11] Libraries were sequenced on the HiSeq4000 or NovaSeq6000 at a depth of 6,306-29,862 reads/cell. Freemuxlet was used to assign cells to individuals and, together with Scrublet[87], for the identification of doublets. Marker gene expression was used to assign the major cell types. Only the monocytes were taken for this independent discovery analysis. Monocytes with less than 1500 UMIs were removed, as were donors with fewer than 200 cells remaining after applying this cutoff. Co-expression QTL analysis was performed as described in the co-expression QTL mapping paragraph above, but only testing the *CLEC12A*-SNP rs12230244 co-expression-QTL and doing this analysis separately in each cohort, ancestry and disease state (SLE versus healthy). A meta-analysis over the cohorts and ancestries was then performed, and pathway analysis using the REACTOME database was conducted to determine whether the IFN pathway was differently enriched in SLE compared to the healthy controls.

Data availability

The number of individuals and cells included in each analysis can be found in the **Source Data** file. Raw gene expression counts, eQTL and co-expression QTL summary statistics can be found under "Supplementary Data" at the website accompanying this paper (https://eqtlgen.org/sc/datasets/1m-scbloodnl.html). Processed (de-anonymized) scRNA-seq data, including a text file that links each cell barcode to its respective individual, has been deposited at the European Genome-Phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001005376 (https://ega-archive.org/studies/EGAS00001005376). Gene expression and genotype data can be obtained and requested by filling in a short web form at https://eqtlgen.org/sc/datasets/1m-scbloodnl.html. This form is subsequently reviewed by a single Data Access Committee, who will be able to approve access to both the raw gene expression and genotype data within 5 working days (during the holiday season there might be a slight delay). Once the proposed research is approved, access to the relevant gene expression or genotyped data will be free of charge. Access to the genotype and gene expression data is facilitated via the HPC cluster of the UMCG and the EGA, respectively. Access to this data is restricted to comply with the European Union General Data Protection Regulation for protection of privacy-sensitive data. Sample metadata (age, gender) is presented in **Supplementary Table 1**. The REACTOME and TRANSFAC release 2020.2 v2[53]

database can be accessed through https://reactome.org/ and https://biit.cs.ut.ee/gprofiler/gost, respectively.

Code availability

The original code for Seurat v3.1[79] (https://github.com/satijalab/seurat), Eagle v2.322 (https://github.com/poruloh/Eagle), Demuxlet[11] 85dca0a4d648d18e-6b240a2298672394fe10c6e6 (Mar 25 2019) (https://github.com/statgen/demuxlet), Souporcell v1.0[29] (https://github.com/wheaton5/souporcell), Freemuxlet v1 as part of the Popscle suite of statistical genetics tools (https://github.com/statgen/popscle), Scrublet[87] v0.2 (https://github.com/swolock/scrublet), the GeneticRiskScoreCalculator v0.1.0c[55] (https://github.com/molgenis/systemsgenetics/tree/master/GeneticRiskScoreCalculator) and our in-house eQTL pipeline[2] v1.4.0 (https://github.com/molgenis/systemsgenetics/tree/master/eqtl-mapping-pipeline) can be found at GitHub. All custom-written code is made available via GitHub (https://github.com/molgenis/1M-cells).

Supplementary material

Supplementary material is available at https://doi.org/10.1038/s41467-022-30893-5

References

1. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics* **101**, 5–22 (2017).
2. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
3. 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).
4. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science (80-.).* **343**, 1246949 (2014).
5. Romanoski, C. E. *et al.* Systems Genetics Analysis of Gene-by-Environment Interactions in Human Cells. *Am. J. Hum. Genet.* **86**, 399–410 (2010).
6. Tang, L. Disease heritability explained by eQTLs. *Nat. Methods* **17**, 655 (2020).
7. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).
8. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 1–12 (2017).
9. Datlinger, P. *et al. Ultra-high throughput single-cell RNA sequencing by combinatorial fluidic indexing 1*.
10. Cuomo, A. S. E. *et al.* Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* **11**, 1–14 (2020).
11. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
12. Van Der Wijst, M. G. P., Brugge, H., De Vries, D. H., Deelen, P. & Swertz, M. A. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).
13. Davenport, E. E. *et al.* Discovering in vivo cytokine-eQTL interactions from a lupus clinical trial. doi:10.1186/s13059-018-1560-8
14. Le, K. T. T. *et al.* Functional annotation of genetic loci associated with sepsis prioritizes immune and endothelial cell pathways. *Front. Immunol.* **10**, (2019).
15. de Vries, D. H. *et al.* Integrating GWAS with bulk and single-cell RNA-sequencing reveals a role for LY86 in the anti-Candida host response. *PLOS Pathog.* **16**, e1008408 (2020).
16. Ye, C. J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science (80-.).* **345**, (2014).
17. Lee, M. N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science (80-.).* **343**, 1246980 (2014).
18. Wucherpfennig, K. W. Mechanisms for the induction of autoimmunity by infectious agents. *J. Clin. Invest.* **108**, 1097–1104 (2001).
19. Bouziat, R. *et al.* Reovirus infection triggers inflammatory responses to dietary antigens and development of celiac disease. *Science (80-.).* **356**, 44–50 (2017).
20. Yeung, W. C. G., Rawlinson, W. D. & Craig, M. E. Enterovirus infection and type 1 diabetes mellitus: Systematic review and meta-analysis of observational molecular studies. *BMJ* **342**, 421 (2011).
21. Nyalwidhe, J. O. *et al.* Coxsackievirus-induced proteomic alterations in primaryhuman islets provide insights for the etiology of diabetes. *J. Endocr. Soc.* **1**, 1272–1286 (2017).
22. Hyöty, H. Viruses in type 1 diabetes. *Pediatric Diabetes* **17**, 56–64 (2016).

23. Pociot, F. *et al.* Genetics of type 1 diabetes: What's next? *Diabetes* **59**, 1561–1571 (2010).
24. De Beeck, A. O. & Eizirik, D. L. Viral infections in type 1 diabetes mellitus-why the β cells? *Nature Reviews Endocrinology* **12**, 263–273 (2016).
25. Qaisar, N., Jurczyk, A. & Wang, J. P. Potential role of type I interferon in the pathogenic process leading to type 1 diabetes. *Current opinion in endocrinology, diabetes, and obesity* **25**, 94–100 (2018).
26. Gutierrez-Arcelus, M., Rich, S. S. & Raychaudhuri, S. Autoimmune diseases — connecting risk alleles with molecular traits of the immune system. *Nat. Publ. Gr.* (2016). doi:10.1038/nrg.2015.33
27. Raj, T. *et al.* Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science (80-.).* **344**, 519–523 (2014).
28. Hu, X. *et al.* Regulation of Gene Expression in Autoimmune Disease Loci and the Genetic Basis of Proliferation in CD4+ Effector Memory T Cells. *PLoS Genet.* **10**, e1004404 (2014).
29. Heaton, H. *et al.* Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17**, 615–620 (2020).
30. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. doi:10.1186/s13059-015-0844-5
31. Li, Y. *et al.* A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans. *Cell* **167**, 1099-1110.e14 (2016).
32. Schmid, K. T. *et al.* scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies. *Nat. Commun. 2021 121* **12**, 1–18 (2021).
33. Smeekens, S. P. *et al.* Functional genomics identifies type i interferon pathway as central for host defense against Candida albicans. *Nat. Commun.* **4**, 1–10 (2013).
34. Parker, D. & Prince, A. Type I interferon response to extracellular bacteria in the airway epithelium. *Trends in Immunology* **32**, 582–588 (2011).
35. Blankley, S. *et al.* A 380-gene meta-signature of active tuberculosis compared with healthy controls. *European Respiratory Journal* **47**, 1873–1876 (2016).
36. Italiani, P. *et al.* Profiling the Course of Resolving vs. Persistent Inflammation in Human Monocytes: The Role of IL-1 Family Molecules. *Front. Immunol.* **11**, 1426 (2020).
37. 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. 2021 539* **53**, 1300–1310 (2021).
38. Hagai, T. *et al.* Gene expression variability across cells and species shapes innate immunity. *Nature* **563**, 197–202 (2018).
39. Gat-Viks, I. *et al.* Deciphering molecular circuits from genetic variation underlying transcriptional responsiveness to stimuli. *Nat. Biotechnol.* **31**, 342–349 (2013).
40. Barreiro, L. B. *et al.* Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1204–1209 (2012).
41. Ota, M. *et al.* Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell* **184**, 3006-3021.e17 (2021).
42. Simpson, E. H. The Interpretation of Interaction in Contingency Tables. *Source J. R. Stat. Soc. Ser. B* **13**, 238–241 (1951).
43. Gao, L. *et al.* Identifying noncoding risk variants using disease-relevant gene regulatory networks. *Nat. Commun.* **9**, 1–12 (2018).
44. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

45. Burton, P. R. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
46. van der Wijst, M. G. P. *et al.* The single-cell eQTLGen consortium. *Elife* **9**, (2020).
47. Regev, A. *et al.* The human cell atlas. *Elife* **6**, (2017).
48. Jeha, S. Tumor lysis syndrome. in *Seminars in Hematology* **38**, 4–8 (W.B. Saunders, 2001).
49. Sancho, D. & Reis e Sousa, C. Signaling by myeloid C-Type lectin receptors in immunity and homeostasis. *Annual Review of Immunology* **30**, 491–529 (2012).
50. Li, K. *et al.* The uric acid crystal receptor Clec12A potentiates type I interferon responses. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 18544–18549 (2019).
51. Jovanovic, M. *et al.* Dynamic profiling of the protein life cycle in response to pathogens. *Science (80-.).* **347**, (2015).
52. Chung, H. *et al. Simultaneous single cell measurements of intranuclear proteins and gene expression*.
53. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, (2006).
54. Kumar, S., Ambrosini, G. & Bucher, P. SNP2TFBS-a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.* **45**, D139–D144 (2017).
55. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv* 447367 (2018). doi:10.1101/447367
56. Crow, M. K. Type I Interferon in the Pathogenesis of Lupus. *J. Immunol.* **192**, 5459–5468 (2014).
57. Santana-de Anda, K., Gómez-Martín, D., Díaz-Zamudio, M. & Alcocer-Varela, J. Interferon regulatory factors: Beyond the antiviral response and their link to the development of autoimmune pathology. *Autoimmunity Reviews* **11**, 98–103 (2011).
58. Sharma, S. *et al.* Widely divergent transcriptional patterns between SLE patients of different ancestral backgrounds in sorted immune cell populations. *J. Autoimmun.* **60**, 51–58 (2015).
59. Banchereau, R. *et al.* Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell* **165**, 551–565 (2016).
60. Bolhassani, A. & Agi, E. Heat shock proteins in infection. *Clinica Chimica Acta* **498**, 90–100 (2019).
61. Kovács, D. *et al.* HSF1Base: A comprehensive database of HSF1 (heat shock factor 1) target genes. *Int. J. Mol. Sci.* **20**, (2019).
62. Triandafillou, C. G., Katanski, C. D., Dinner, A. R. & Allan Drummond, D. Transient intracellular acidification regulates the core transcriptional heat shock response. *Elife* **9**, 1–30 (2020).
63. Vihervaara, A. *et al.* Transcriptional response to stress in the dynamic chromatin environment of cycling and mitotic cells. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E3388–E3397 (2013).
64. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593–1599 (2018).
65. 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).
66. De Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
67. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* **43**, 1193–1201 (2011).
68. Patsopoulos, N. A. *et al.* Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science (80-.).* **365**, (2019).

69. 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).

70. Jaeger, M. *et al.* A genome-wide functional genomics approach identifies susceptibility pathways to fungal bloodstream infection in humans. *J. Infect. Dis.* (2019). doi:10.1093/infdis/jiz206

71. Begun, J. *et al.* Integrated Genomics of Crohn's Disease Risk Variant Identifies a Role for CLEC12A in Antibacterial Autophagy. *Cell Rep.* **11**, 1905–1918 (2015).

72. Redelinghuys, P. *et al.* MICL controls inflammation in rheumatoid arthritis. *Ann. Rheum. Dis.* **75**, 1386–1391 (2016).

73. Vaillancourt, M. *et al.* Expression of the myeloid inhibitory receptor CLEC12A correlates with disease activity and cytokines in early rheumatoid arthritis. *Sci. Reports* | **11**, 11248 (123AD).

74. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun. 2017 81* **8**, 1–11 (2017).

75. Chen, A. F. *et al.* NEAT-seq: Simultaneous profiling of intra-nuclear proteins, chromatin accessibility, and gene expression in single cells. doi:10.1101/2021.07.29.454078

76. Deciphering Intra- and Inter-Individual Variation at Single Cell Resolution - Chan Zuckerberg Initiative. Available at: https://chanzuckerberg.com/science/programs-resources/single-cell-biology/seednetworks/deciphering-intra-and-inter-individual-variation-at-single-cell-resolution/. (Accessed: 23rd April 2021)

77. 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).

78. Tigchelaar, E. F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).

79. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

80. Cesar Prada, Diogenes Lima, H. N. MetaVolcanoR: Gene Expression Meta-analysis. *R Packag. version 1.6.0.* (2021).

81. Fisher, R. A. Statistical methods for research workers. *Oliver Boyd Edinburgh* (1925).

82. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305–W311 (2009).

83. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species A n A ly s I s. *Nat. Biotechnol. Vol.* **36**, (2018).

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

85. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

86. Raudvere, U. *et al.* G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).

87. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* **8**, 281-291.e9 (2019).

Acknowledgements

We are very grateful to all the volunteers who participated in this study and would like to thank K. McIntyre for proofreading the manuscript. This work was carried out on the computer cluster of the Genomics Coordination Center, hosted at the University of Groningen Center for Information Technology. M.G.G. was supported by the National Institutes of Health under Ruth L. Kirschstein National Research Service Award F31HG011007. C.J.Y. is further supported by the NIH grants R01AR071522, R01AI136972, U01HG012192, and the Chan Zuckerberg Initiative, and is an investigator at the Chan Zuckerberg Biohub and is a member of the Parker Institute for Cancer Immunotherapy (PICI). L.F. and M.W. are supported by grants from the Dutch Research Council (ZonMW-VIDI 917.14.374 and ZonMW-VICI to L.F., NWO-VENI 192.029 to M.W.), and by an ERC Starting Grant, grant agreement 637640 (ImmRisk) and through a Senior Investigator Grant from the Oncode Institute. 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). The images in Fig. 4f are created using Servier Medical Art, which we are thankful to for providing free online images.

Author contributions

M.W. collected blood samples and generated the scRNA-seq data. RO, DV and HB performed bioinformatics and statistical analyses. R.O., D.V., H.B. and M.W. generated figures. HB built the website accompanying the manuscript. M.G.G. and C.J.Y. performed and provided SLE scRNA-seq data for co-expression QTL replication analysis. M.V. and H.W. provided critical input for the statistical analyses. The BIOS consortium provided samples to conduct the SLE PRS co-expression QTL analysis. D.V., M.W. and L.F. designed the study. R.O., D.V. and M.W. wrote the manuscript and all other authors provided critical feedback. All authors discussed the results and commented on the manuscript.

Competing Interests

C.J.Y. is a Scientific Advisory Board member for and hold equity in Related Sciences and ImmunAI, a consultant for and hold equity in Maze Therapeutics, and a consultant for TReX Bio. C.J.Y. has received research support from Chan Zuckerberg Initiative, Chan Zuckerberg Biohub, and Genentech.

The remaining authors declare no competing interestests f

CHAPTER 6

Identification of genetic variants that impact gene co-expression relationships using large-scale single-cell data

Shuang Li [1,2] *, Katharina T. Schmid [3,4] *, Dylan de Vries [1] *, Maryna Korshevniuk [1], Roy Oelen [1], Irene van Blokland [1,5], BIOS Consortium, sc-eQTLgen Consortium, Hilde E. Groot [5], Morris Swertz [1,2], Pim van der Harst [6], Harm-Jan Westra [1], Monique van der Wijst [1], Matthias Heinig [3,4] **, Lude Franke [1] **

[1] Genetics Department, University Medical Center Groningen, the Netherlands
[2] University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, the Netherlands
[3] Institute of Computational Biology, Helmholtz Center Munich, Munich, Germany
[4] Department of Informatics, Technical University Munich, Munich, Germany
[5] Department of Cardiology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[6] Department of Cardiology, University Medical Center Utrecht, Utrecht, the Netherlands

* These authors contributed equally
** These authors contributed equally

Keywords: Co-expression QTLs, scRNA-seq, eQTL

Under review at Genome Biology. Preprint available on bioRxiv.
https://doi.org/10.1101/2022.04.20.488925

Abstract

Background

Expression quantitative trait loci (eQTL) studies have shown how genetic variants affect downstream gene expression. To identify the upstream regulatory processes, single-cell data can be used. Single-cell data also offers the unique opportunity to reconstruct personalized co-expression networks—by exploiting the large number of cells per individual, we can identify SNPs that alter co-expression patterns (co-expression QTLs, co-eQTLs) using a limited number of individuals.

Results

To tackle the large multiple testing burden associated with a genome-wide analysis (i.e. the need to assess all combinations of SNPs and gene pairs), we conducted a co-eQTL meta-analysis across four scRNA-seq peripheral blood mononuclear cell datasets from three studies (reflecting 173 unique participants and 1 million cells) using a novel filtering strategy followed by a permutation-based approach. Before analysis, we evaluated the co-expression patterns to be used for co-eQTL identification using different external resources. The subsequent analysis identified a robust set of cell-type-specific co-eQTLs for 72 independent SNPs that affect 946 gene pairs, which we then replicated in a large bulk cohort. These co-eQTLs provide novel insights into how disease-associated variants alter regulatory networks. For instance, one co-eQTL SNP, rs1131017, that is associated with several autoimmune diseases affects the co-expression of *RPS26* with other ribosomal genes. Interestingly, specifically in T cells, the SNP additionally affects co-expression of *RPS26* and a group of genes associated with T cell-activation and autoimmune disease. Among these genes, we identified enrichment for targets of five T-cell-activation-related transcriptional factors whose binding sites harbor rs1131017. This reveals a previously overlooked process and pinpoints potential regulators that could explain the association of rs1131017 with autoimmune diseases.

Conclusion

Our co-eQTL results highlight the importance of studying gene regulation at the context-specific level to understand the biological implications of genetic variation. With the expected growth of sc-eQTL datasets, our strategy—combined with our technical guidelines—will soon identify many more co-eQTLs, further helping to elucidate unknown disease mechanisms.

Background

In recent years, genome-wide association studies (GWAS) have revealed a large number of associations between genetic variation and disease (1). Many of these variants also change downstream gene expression, as identified using expression quantitative trait locus (eQTL) analysis (2). However, even with many such connections now identified, the upstream biological processes that regulate these eQTLs often remain hidden. Such knowledge is important for better understanding the underlying processes that lead to specific disease, which would aide in drug development (3).

One way to study the biological processes in which eQTL genes are involved is to construct gene co-expression networks. In these networks, genes (nodes) involved in shared biological processes are expected to be connected through co-expression (edges) (4). Traditionally, these networks have been reconstructed with bulk RNA sequencing (RNA-seq) data, using a variety of computational tools (5–7). However, whether certain biological processes are active can depend on various factors, such as cell type, environmental exposures and even single nucleotide polymorphisms (SNPs) (2,8,9). With single-cell technologies, many of these highly specific contexts can now be captured at high resolution. Single-cell RNA-seq (scRNA-seq) not only allows for cell-type-specific analyses, it does so without the technical biases introduced by the cell-sorting required to perform similar analyses with bulk RNA-seq.

In addition to capturing the cell-type-specific contexts, scRNA-seq can also be used to construct personalized co-expression networks using the repeated measurements (i.e. multiple single-cell gene expression profiles) for each individual. This enables quantification of the covariance between genes, and thus their co-expression strengths, within an individual (10). These personalized co-expression networks can then be used to study the effects of genetic variation on network properties. Some of these network changes can be linked to individual SNP genotypes, called co-expression quantitative trait loci (co-eQTLs).

While we have previously shown that co-eQTLs can be both cell-type-specific and stimulation-specific, several challenges to systematic identification remain (10,11). Firstly, it is unclear how to best construct gene regulatory networks (GRNs) with scRNA-seq data. Co-expression patterns identified from bulk RNA sequencing data have been shown to be informative for physical and functional gene–gene interactions (5–7), but it is unclear whether the co-expression patterns identified with scRNA-seq data also reflect gene–gene functional interactions given technical challenges of scRNA-seq data such as sparseness and low signal-to-noise ratios (12,13). These issues are caused by a combination of low mRNA counts in cells, imperfect capture efficiencies and the inherent stochasticity of mRNA expression (14). Many methods have been proposed to account for this issue. A recent benchmark paper (15) suggested ‘rho proportionality’ (16) as an association measure because of its consistent performance. Also complementary strategies could be beneficial, such as combining association measures with *MetaCell,* a recently proposed method

that groups homogeneous cells to reduce sparsity, but to our knowledge it has not yet been evaluated in benchmark studies (17). Moreover, a recent benchmark paper concluded that different GRN construction methods show moderate performance that is often dataset-specific (18), indicating that many challenges remain in GRN reconstruction. Therefore, validation of the robustness and functional relevance of the network is warranted.

Secondly, there is no consensus method for co-eQTL mapping and personalized GRN construction. In bulk data with only one measurement per individual, it is not possible to identify co-eQTLs directly. To carry out a similar type of analysis in bulk data, we previously used a linear regression model with an interaction term to identify interaction QTLs in bulk data from whole blood (8). This approach can reveal co-eQTLs using the expression levels of individual genes as interaction terms. However, as bulk data nearly always comprises a mixture of cell types, it is not straightforward to unequivocally conclude that eQTLs showing an interaction effect reflect co-eQTLs (genetic variants that affect the co-expression between pairs of genes). A further compounding problem is that very large numbers of samples are required to identify co-eQTLs, and effects that manifest in specific (rare) cell types can easily be missed because they are masked by more common cell types. In theory, single-cell data allows direct estimation of cell-type-specific and individual-specific co-expression strength and should reduce the sample size requirement compared to bulk datasets. However, in practical terms, there are currently no datasets large enough to provide the statistical power to do genome-wide co-eQTL mapping, as this involves a large multiple testing burden due to billions of tests for every SNP and every possible gene pair combination. As such, there is a clear need for a robust co-eQTL strategy that can overcome the severe multiple testing issues and deal with the aforementioned issues with regards to the construction of reliable personalized co-expression networks.

In this work, we studied the genetic regulation of gene co-expression by conducting the largest-to-date co-eQTL meta-analysis in 173 peripheral blood mononuclear cell (PBMC) scRNA-seq samples. Before conducting this co-eQTL analysis, we determined the best strategy to identify cell-type-specific co-expression relationships in scRNA-seq data by benchmarking various methods and studying them in several independent datasets, including bulk RNA-seq and a CRISPR-coupled scRNA-seq screen knockout dataset. We then studied the effects of cell-type and inter-individual differences in gene co-expression networks by reconstructing personalized and cell-type-specific networks. We subsequently developed a robust co-eQTL mapping strategy with a novel filtering approach and an adapted permutation-based multiple testing procedure to deal with the correlation structure in the co-expression networks. By applying this strategy, we could perform a co-eQTL meta-analysis using data from three different scRNA-seq studies. We provided a comprehensive analysis of the different factors that affect the quality and quantity of co-eQTLs, including the number of cells, gene expression levels and filtering strategy. We then studied which biological processes and genes are regulated by the identified

co-eQTLs by performing different enrichment analyses and exploring common biological functions, transcription factor (TF) binding and disease associations to try and pinpoint potential direct regulators of the co-eQTL genes. In sum, our results suggest that the combination of a robust method and a large sample size is crucial for identification of genetic variants that affect co-expression networks.

Results

Overview of the study

To uncover the contexts and biological processes that affect gene expression regulation, this study took advantage of both the resolution of single-cell data and the directionality captured by co-eQTLs. First, we constructed cell-type-specific co-expression networks from three recently generated PBMC scRNA-seq studies totaling 187 individuals and approximately one million cells. In two of the studies, donors were measured using two different versions of 10X Genomics chemistry (version 2 or version 3). To avoid batch effects due to these technical differences, we split both studies into two datasets, depending on the chemistry, leading to five datasets in total: 1) two datasets from the Oelen study (11) that collected unstimulated PBMCs from 104 healthy individuals from the Northern Netherlands, a dataset measured using version 2 chemistry (hereafter called the Oelen v2 dataset) and one measured using version 3 chemistry (called the Oelen v3 dataset), 2) a dataset from the van der Wijst study (10) that collected unstimulated PBMCs from 45 healthy individuals from the Northern Netherlands measured using version 2 chemistry (called the 'van der Wijst' dataset) and 3) two datasets from the van Blokland study (19) that collected unstimulated PBMCs from 38 individuals 6–8 weeks after having a heart attack, one dataset measured using version 2 chemistry (called the van Blokland v2 dataset) and one measured using version 3 chemistry (called the van Blokland v3 dataset) (Figure 1a, Supplementary Table 1).

We focused on the six major cell types (B cells, CD4+ T cells, CD8+ T cells, dendritic cells (DCs), monocytes and natural killer (NK) cells), of which CD4+ T cells, CD8+ T cells and monocytes were the most frequent cell types (**Supplementary Figure 1**). We compared commonly used measures of correlation and those previously reported to be particularly suitable for capturing co-expression in scRNA-seq data, including rho proportionality (16), Spearman correlation and GRNBoost2 (20), and tested complementary strategies such as MetaCell (17). We validated that the co-expression patterns from our single-cell dataset are enriched for actual gene regulatory relationships by benchmarking the concordance of the co-expression patterns across the three single-cell studies (10,11,19) and three cell-type-specific or whole-blood bulk RNA-seq datasets (2,21,22) (**Figure 1b**). Furthermore, we validated identified connections with a CRISPR dataset (23).

Next, we evaluated the concordance of the co-expression networks between the major blood cell types and between different individuals within each cell type (**Figure 1b**). To identify the genetic contribution to such common and cell-type-specific effects, we performed a constrained co-eQTL meta-analysis. For this, we filtered SNPs that exhibit an eQTL effect (with the corresponding gene referred to as an eGene below) and tested all genes with sufficient co-expression strength with the eGene (called co-eGenes below) among different individuals (**Figure 1c**).

For the co-eQTL interpretation, we considered different scenarios that can lead to detection of co-eQTL. One is that the genetic variant changes the binding affinity of a TF and thus the regulation of its target gene, which would cause a co-eQTL between the variant, the TF and the target gene (**Figure 1c**). However, a co-eQTL will also occur for all genes in strong correlation with this TF (**Figure 1c**), so we tried to identify directly interacting TFs via additional annotations and enrichment analyses. Other scenarios include genetic variants that change the structure of the TF and thereby its binding affinity and genetic variants that affect sub-cell-type composition and thus the correlation pattern of sub-cell-type-specific genes.

We then replicated the identified co-eQTLs in a large bulk study (2), explored technical factors influencing the identification of co-eQTLs (sample size, number of cells, different filtering approaches) and biologically interpreted several examples of co-eQTLs (**Figure 1d**).

Correlation validation

Co-expression correlations can be assessed using various dependency measures. A recent benchmark study (15) reported that the proportionality measure from the propr package (16) outperforms several other methods in the identification of functional, coherent biological clusters. We observed high correlations between rho proportionality and Spearman correlations ($r = 0.68$) for genes expressed in $> 5\%$ of the cells (**Supplementary Figure 2a**), but for genes expressed in fewer cells, rho proportionality gave arbitrarily high values while the Spearman correlation remains near zero (**Supplementary Figure 2b**). The reason for the stark differences for very lowly expressed genes is probably that rho proportionality changes zero values to the next lowest value of the gene pair, which may result in false positive associations (i.e. very high rho values) for lowly expressed gene pairs. Another drawback of rho proportionality is the high computational demand (24), which makes it challenging to evaluate all gene pairs. As the differences between Spearman correlation and rho proportionality are very small for highly expressed genes and Spearman correlation calculation is far more efficient and handles zero values better, we chose to use Spearman correlation over rho proportionality.

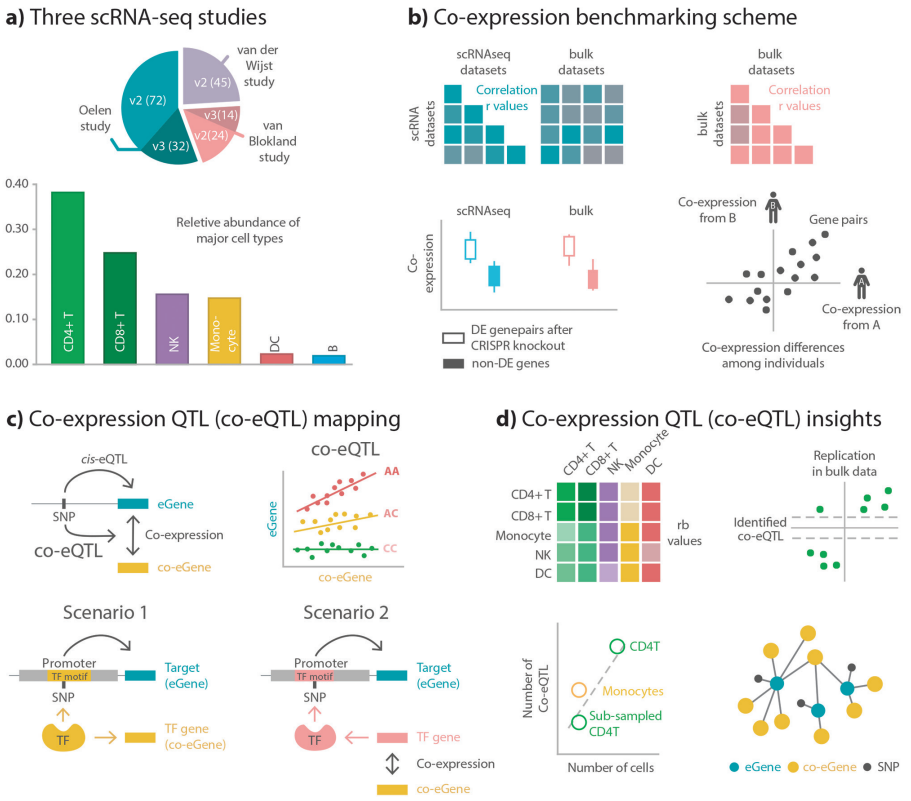


Figure 1. Study overview. a) *Overview of the three PBMC scRNA-seq studies used in our study. The studies, the version of the used chemistry for data generation (version 2, referred as v2, and version 3, referred as v3), number of individuals involved (indicated as the number in the parenthesis), and relative composition of the major blood cell types used in this study.* ***b)*** *Co-expression benchmarking scheme. We first benchmarked co-expression patterns among the three scRNA-seq studies and compared them to co-expression patterns in different bulk datasets. As an additional external validation, we benchmarked both the scRNA-seq and bulk co-expression patterns with a CRISPR knockout dataset. After benchmarking, we evaluated differences in co-expression patterns among cell types and among individuals within a cell type.* ***c)*** *Co-expression QTL (co-eQTL) mapping. Building on the benchmarked co-expression pattern, we developed a novel strategy to identify co-eQTLs (genetic variants changing co-expression). Part of the strategy is a strict filtering of tested SNP–eGene–co-eGene triplets, where the SNP is required to be an eQTL for one of the genes and the genes show significant correlation in at least a certain number of individuals.* ***d)*** *Co-expression QTL (co-eQTL) insights. co-eQTL mapping was conducted for each major blood cell type, then replicated in a bulk dataset. To evaluate technical influences, we assessed the impact of cell number, number of tests and the number of individuals on the number of significant co-eQTLs. Lastly, we interpreted the biological relevance of the co-eQTLs and reconstructed the gene regulatory network using identified co-eQTLs.*

We also tested other approaches, including GRNBoost2 (20), grouping cells into *MetaCells* (17) before calculation of Spearman correlation, and testing pseudotime ordering (25) and RNA velocity (26), but these did not yield more reliable results than Spearman correlation (**Supplementary Figures 3,4,5; Supplementary Text**). We therefore selected Spearman correlation to measure the co-expression patterns in scRNA-seq data for its robustness and simple interpretability. However, although we determined that Spearman correlation was optimal for the single-cell PBMC datasets that we studied, we cannot exclude that the other methods might be optimal for other single-cell datasets.

We then evaluated whether the co-expression patterns obtained from scRNA-seq data are robust and reproducible across different single-cell datasets and whether they reflect functional interactions among genes. Benchmarking the co-expression patterns obtained from scRNA-seq data is difficult because, to our knowledge, there is no clear gold-standard dataset of known functional gene–gene interactions for different cell types. As an alternative approach to assess the reliability of the identified co-expression relationships, we compared to what extent we could replicate the co-expression patterns found in one dataset in another dataset.

We first compared the cell-type-specific co-expression patterns among the five scRNA-seq datasets in our study (10,11,19). For this, we inferred the co-expression strength using Spearman correlation for each gene pair in each dataset and cell type, where gene pairs were only considered when both genes were expressed in at least 50% of the cells. We summarized the concordance between datasets by calculating the Pearson correlation on the gene pair correlation values. Overall, there was high concordance across all cell types (median r = 0.80 across all cell types). CD4+ T cells, the most abundant cell type in our dataset, had a high correlation across the different 10X chemistries and datasets, with values ranging from 0.67 to 0.86 and a median of 0.81 (**Figure 2a**). For CD8+ T cells and NK cells, we observed a comparably high correlation (CD8+ T cells median r = 0.86, NK cells median r = 0.80), while the correlation was slightly lower for the other cell types (monocytes median r = 0.69, B cells median r = 0.70, DCs median r = 0.71) (**Supplementary Figure 6**). The number of genes expressed in 50% of the cells varied between dataset and chemistry, so it was not always possible to test the same set of genes. In general, this filtering strategy is quite stringent, yielding a limited number of tested genes (at most 766 genes for the Oelen v3 dataset in CD4+ T cells, **Figure 2a**), which ensured a high-quality gene set to test due to the sparse single-cell data. A detailed evaluation of the expression cutoff follows in the next sections.

Next, we compared the co-expression patterns from the single-cell datasets to three different bulk datasets from BLUEPRINT (21), ImmuNexUT (22) and the BIOS Consortium (2). The BLUEPRINT dataset contains fluorescence-activated cell sorting (FACS)-sorted expression data from naive CD4+ T cells and classical monocytes for up to 197 individuals. The ImmuNexUT study collected gene expression data from 337 patients for 28 FACS-sorted immune cell subsets. The BIOS dataset contains

whole-blood expression data from 3,198 individuals. Notably, the co-expression correlation between the single-cell and bulk-based datasets (**Figure 2b**) was much lower than those between the single-cell datasets alone (**Figure 2a**).

Comparing our single-cell data with ImmuNexUT, the only dataset with cell-type-specific expression for all evaluated cell types, CD8+ T cells showed the highest correlation (median r = 0.570) and monocytes (median r = 0.395) and DCs (median r = 0.259) showed the lowest correlations (**Figure 2b**, **Supplementary Figure 7**). The correlations from BLUEPRINT were slightly lower but in the same range (CD4+ T cells median r = 0.356, monocytes median r = 0.339) (**Figure 2b**, **Supplementary Figure 7**). Finally, we observed that the whole blood bulk data from the BIOS dataset correlated reasonably with the different single-cell cell types (median r between 0.265 and 0.458 across cell types; **Figure 2b**, **Supplementary Figure 7**).

We studied this seemingly low correlation between bulk and single-cell data, and identified multiple factors that play a role. One is the sparseness of the single-cell data, which could introduce noise and therefore lead to less stable co-expression values. To test this, we correlated the co-expression from the Oelen v3 dataset with that from ImmuNexUT using varying expression cutoffs based on the number of cells expressing a gene (**Figure 2c**). Indeed, the sparseness of the single-cell data affects the correlation. We observed increased concordance with increasing gene expression levels: the correlation increased from r = 0.21 for an expression cutoff of 10% to r = 0.71 at a cutoff of 90%. However, the number of genes that can be tested dropped from 4,482 at an expression cutoff of 10% to 172 at a cutoff of 90%. The same trends were observable when comparing the Oelen v3 dataset with the BLUEPRINT dataset for different cutoffs (**Supplementary Figure 8**). For this reason, we chose a cutoff of 50% as a trade-off between both extremes in our benchmarking study (**Figure 2a,b,e,f**).

Other aspects that may affect correlations between genes are the difference in resolution and potential biases introduced by acquiring cell-type-specific data, such as the gene expression changes induced by FACS and the technical complications of deconvoluting cell types. Furthermore, the validity of bulk-based correlations is affected by the possibility of Simpson's paradox occurring. Simpson's paradox describes the incorrect introduction or removal of correlations by averaging expression levels. This can potentially occur in bulk datasets, whereas single-cell data can accurately identify the co-expression value since we can calculate co-expression values per cell type and per individual (**Supplementary Figure 9a**). To estimate the effects of this phenomenon, we recalculated co-expression from the single-cell data using a bulk-like approach, compared it to the normal single-cell co-expression values and observed several examples of highly expressed genes in which Simpson's paradox occurs (**Supplementary Figure 9b,c**). However, taking the average gene expression over many cells also results in more robust expression estimates, which can generate less noisy co-expression estimates, especially for lowly expressed genes. For this reason, we cannot differentiate for all genes which co-expression differences between single-cell and bulk are caused by Simpson's paradox and which are caused by noisy single-cell data.

To contextualize the correlation values between single-cell and bulk data, we also compared the bulk datasets with each other and assessed whether bulk datasets actually capture gene co-expression consistently. Surprisingly, the co-expression correlation similarity between bulk datasets was quite low (r between 0.47 and 0.52 for CD4+ T cells and between 0.35 and 0.42 for monocytes) (**Figure 2d, Supplementary Figure 10**). Given that these correlations are expected to be an upper bound when comparing bulk datasets with single-cell datasets, our observed correlations in those comparisons are very reasonable.

Given the imperfect correlation between the different bulk datasets, we used gene expression data from CRISPR-knockouts as an additional evaluation criterion. For this purpose, we benchmarked the co-expression patterns from our single-cell datasets against a CRISPR knockout scRNA-seq dataset in CD4+ T cells (23). While a unique single-guide RNA barcode reveals which gene was targeted in which cell, some cells may escape from successful CRISPR perturbation. To account for this, we used Mixscape to assign a perturbation status to each cell (27). For each knockout, we then determined other genes that were differentially expressed (DE) in successfully perturbed cells compared to wild-type cells. We then selected genes for which perturbation resulted in at least 10 DE genes and compared the correlation of these DE genes with non-DE genes using the Wilcoxon rank-sum test (see **Methods**). For four out of five gene knockouts, we observed significantly higher correlation of the knockout gene with the DE genes than with non-DE genes ($p < 0.05$) in the single-cell dataset (**Figure 2e**). In contrast, the bulk naive CD4+ T cell data from ImmunNexUT showed a weaker connection between correlation and DE genes, with only two out of five knockout genes having significantly higher correlation with the DE genes ($p < 0.05$) (**Figure 2f**).

As another line of evidence, we tested whether pairs of genes known to interact on the protein level showed higher co-expression correlation compared to other pairs of genes. Here we found that gene pairs with protein interactions listed in the STRING database (28) had a higher co-expression correlation than gene pairs not in STRING, both when using the single-cell dataset and the bulk dataset (for both Wilcoxon rank-sum test, $p < 0.05$, **Supplementary Figure 11**).

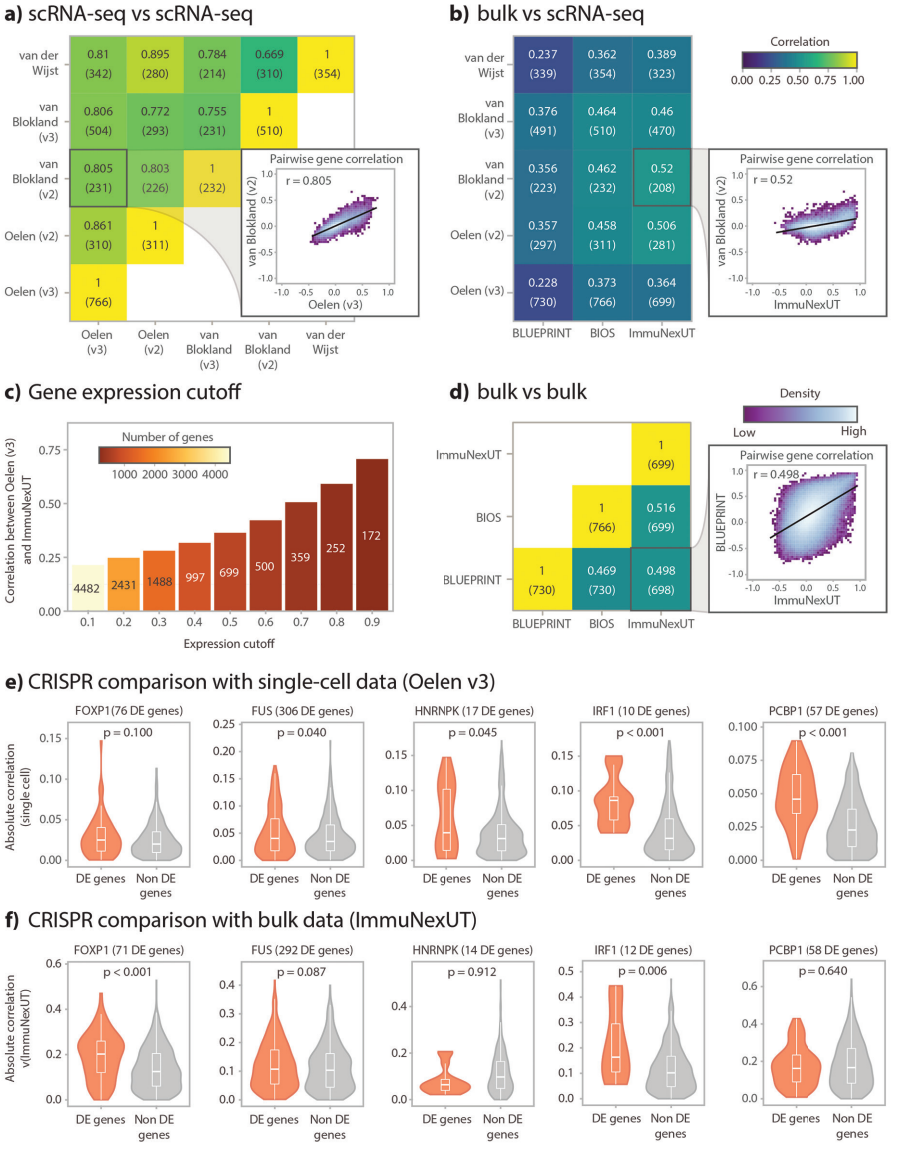


Figure 2. Evaluation of correlation metrics. a) *Comparison of the co-expression profiles among the different single-cell datasets in this study. Spearman correlation of the Oelen v2 and v3 datasets, the van Blokland v2 and v3 datasets and the van der Wijst dataset were compared with each other, always taking the CD4+ T cells and genes expressed in at least 50% of the cells in the corresponding datasets. The number of genes tested is shown in parentheses below the exact Spearman correlation value.* ***b)*** *Comparison of the co-expression profiles between the single-cell datasets and with the bulk RNA-seq datasets from BLUEPRINT, ImmuNexUT (both measuring FACS-sorted naive CD4+ T cells) and BIOS (whole blood). Again, we only assessed genes expressed in at least 50% of the cells for the single-cell dataset (number of tested genes shown in parentheses*

below the Spearman correlation value). ***c)*** *Relationship between the co-expression similarity between the ImmuNexUT naive CD4+ T cells and Oelen v3 dataset CD4+ T cells and increasing gene expression cutoffs (the ratio of cells with non-zero expression for a given gene). The number of genes tested are indicated by color scale and the numbers in the bar plot.* ***d)*** *Comparison of the co-expression profiles between the bulk RNA-seq datasets, taking the same gene subset as in* ***a)*** *and* ***b)****. The number of tested genes is shown in parentheses below the exact Spearman correlation value.* ***e)*** *Enrichment of correlated genes in scRNA-seq (Oelen v3 dataset) among associated genes identified by CRISPR knockout. For the enrichment, genes differentially expressed after knockout of FOXP1, FUS, HNRNPK, IRF1 and PCBP1 were identified and tested for enrichment. P-values in the plot show the significance level of the Wilcoxon rank-sum test.* ***f)*** *Enrichment of correlated genes in bulk RNA-seq (ImmuNexUT) among associated genes identified by CRISPR knockout. See* ***e)*** *and* ***Methods*** *for further details.*

Overall, we have shown that single-cell data can identify true gene co-expression relationships as co-expression patterns from scRNA-seq data are highly replicable among different datasets and are supported by functional interactions among genes identified by CRISPR perturbations and the STRING database.

Cell type and donor differences in co-expression pattern

Next, we examined cell-type-specific and individualized co-expression patterns. As expected, lymphoid cell types (B, T and NK cells, $r > 0.73$) were more alike with each other but they are less alike with myeloid cell types (monocytes and DCs, $r > 0.45$) (**Figure 3a**, **Supplementary Figure 12a**). However, myeloid cell types were not as alike to each other as lymphoid cell types. This is possibly due to the fact that DCs are one of the least abundant cell types (**Supplementary Figure 1**), which would have resulted in less accurate co-expression estimations. Overall, the correlation between different cell types within one scRNA-seq dataset (for Oelen v3 dataset median $r = 0.64$, **Figure 3a**) was generally lower than the correlation between different scRNA-seq datasets when studying a single cell type (median $r = 0.80$ across all cell types, **Figure 2a, Supplementary Figure 6**). These differences highlight cell-type-specific differences in the correlation pattern, further confirming the biological aspects captured by scRNA-seq co-expression values. We also explored the distribution of co-expression among cell types (**Figure 3b, Supplementary Figure 12b**). Typically, the correlations between gene pairs were rather low, with only a small proportion of gene pairs (median 12.4%) showing correlations above 0.1. However, we did observe cell-type-specific differences, with DCs possessing a higher proportion of co-expressed gene pairs compared to the other cell types (32.3% of gene pairs with $r > 0.1$).

In addition to detecting cell-type-specific associations, scRNA-seq enables direct calculation of co-expression correlations per individual as it provides many measurement points per donor. When we calculated the correlation separately for each donor and cell type, we observed overall strong correspondence of co-expression networks between different donors for the more frequent cell types (CD4+ T cells median $r = 0.56$, CD8+ T cells median $r = 0.48$, monocytes median $r = 0.47$) (**Figure 3c, Supplementary Figure 12c**). As a result of noisier estimates, the correlation

between individuals was drastically lower for the less frequent cell types (DCs median r = 0.24, B cells median r = 0.06). Moreover, these correlations were much smaller than comparing one cell type across entire datasets (i.e. including all individuals at once), which showed correlations of at least 0.81 for CD4+ T cells, 0.64 for CD8+ T cells, 0.49 for monocytes, 0.66 for NK cells, 0.62 for B cells and 0.38 for DCs (**Figure 2a**). This decline is potentially caused by the number of cells used to calculate the correlation, which is drastically lower when comparing donors within one dataset. The number of cells could also explain the differences between the cell types. To test this, we subsampled the number of cells for each cell type and indeed observed that the correlation increased when the number of cells increased (**Figure 3d**). Apart from the number of cells, we also observed potential cell type differences. The similarities between individuals were significantly smaller in NK cells compared to monocytes and T cells, when the same number of cells was used (**Figure 3d**). We also confirmed these observations in another scRNA-seq dataset (**Supplementary Figure 12d**).

We further explored the relationship between the number of cells per individual and the correlation between individuals by fitting a logarithmic curve for the four most frequent cell types: CD4+ T cells, CD8+ T cells, monocytes and NK cells (**Supplementary Figure 13**). Each of the observed trends could be fit well with the logarithmic curve (adjusted R^2 values between 0.86 and 0.98). We then extrapolated the trend to 1,000 cells, showing that a correlation > 0.80 would be expected for T cells and monocytes with this number of cells and a correlation of 0.65 for NK cells (**Supplementary Figure 13**). We acknowledge, however, that the exact upper bound for the correlation between donors cannot be estimated accurately with our current dataset. For example, the correlation close to 100% for CD4+ T cells and 1,500 cells is likely too high considering that donor-specific differences such as genetics and environment will remain independent of the number of cells. Nevertheless, our fits highlight the value of having measurements from many cells for accurate correlation estimates as well as cell-type-specific differences in the correlation pattern.

During this comparison, we observed a few gene pairs that showed a high variance in correlation across donors within one cell type (median fraction of gene pairs with correlation Z-score variance > 2 across cell types: 4.9% for Oelen v2 dataset and 3.3% for Oelen v3 dataset, **Supplementary Figure 14**). This high variance could, in theory, be caused by different sources, e.g. technical factors or environmental influences, but could also reflect genetic differences between individuals. Since we observed low co-expression variance between different individuals for the same cell type and similar numbers of cells (**Figure 3d**), we concluded that these differences are not likely to originate from technical factors, and thus we next looked into genetic variation as one of the other potential major influences.

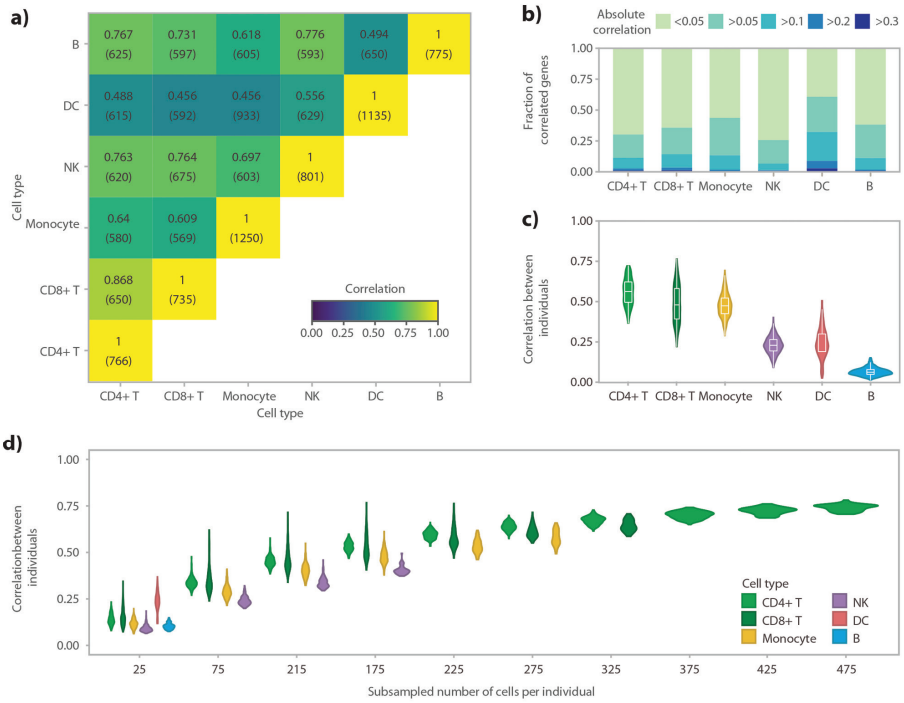


Figure 3. Comparison of correlation across cell types and donors. *Each analysis was performed in the Oelen v3 dataset for all genes expressed in at least 50% of the cells of the respective cell type.* ***a)*** *Comparing co-expression patterns across cell types within the Oelen v3 dataset for genes expressed in 50% of the cells for both cell types in each pair-wise comparison. The number of tested genes is shown in parenthesis below the Spearman correlation value.* ***b)*** *Correlation distribution within each cell type.* ***c)*** *Correlation between different individuals within each cell type showing the distribution of all pair-wise comparisons between individuals.* ***d)*** *Relationship between the number of cells per individual and cell type and correlation between individuals separately for each cell type. In each subsampling step, we assessed all individuals who have at least this number of cells and subsampled to exactly this number (this leads to removal of some individuals for higher number of cells and thus, a direct comparison with the correlation values in* ***c)*** *is not possible).*

Establishing a method to identify co-expression QTLs

To assess how strongly genetic variation influences the correlation between pairs of genes, we performed a co-eQTL analysis. In contrast to classical eQTL analysis, co-eQTL analysis not only reveals the downstream target gene whose expression is affected by a genetic variant, it can also help identify the upstream regulatory factors that affect these eQTLs, as discussed in the Overview.

Compared to an eQTL analysis, a full co-eQTL analysis with all SNP–gene pair combinations would massively increase the multiple testing burden. Previously, we showed the necessity of filtering the SNP–gene pair combinations to reduce the multiple testing burden associated with a genome-wide co-eQTL analysis on all possible triplets while not missing true co-eQTLs (11). For example, in our current

study, testing all pairs of genes expressed in monocytes would lead to 1.96×10^8 tests when considering only one SNP per pair and to a very limited power to detect small effect sizes (power of 1.4% to detect a significant effect for a phenotype (here the co-expression relationship) with a heritability of 10% that is explained by a single locus, **Supplementary Figure 15**).

In this study, we aimed to define a generally applicable filtering strategy that yields a large number of highly confident co-eQTLs. First, we decided to focus on *cis*-eQTL SNPs and genes because we expect a SNP influencing the co-expression of two genes to also influence the expression of one of the genes directly (a strategy we applied successfully before in (10,11)). To identify these *cis*-eQTLs, we first performed a *cis*-eQTL meta-analysis across four of the five scRNA-seq datasets. We excluded the van Blokland v3 dataset from this eQTL analysis and all subsequent analyses because the small sample size (N = 14) provided very few variants above the minor allele frequency (MAF) cutoff (>10%), which made it unsuitable for this meta-analysis. To reduce the multiple testing burden and maximize the number of *cis*-eQTLs detected given the relatively low number of individuals (N = 173) used for the eQTL mapping, we confined ourselves to 16,987 lead *cis*-eQTLs previously identified in a large (N = 31,684) bulk blood eQTL study (2). Depending on cell type, we identified between 917 (for CD4+ T cells) and 51 (for B cells) eQTLs (FDR<0.05; **Supplementary Table 2, 3**).

As filtering for the eQTL effects still resulted in a large number of tests (e.g. for CD4+ T cells, n = 12,137,281, **Supplementary Table 4**) and consequently a large multiple testing burden, we imposed additional filtering on the co-eGenes to study. Here, we used a filtering strategy based on the co-expression significance, selecting co-eGenes for which we observed a significant (nominal $p \leq 0.05$) correlation with the eGene in at least 10% of the individuals (**Methods**). We assumed this captures genuine co-expression effects that are present in at least one of the genotype groups (i.e. homozygous reference/heterozygous/homozygous alternative allele). Note that the filtering strategy we used here is less stringent than the cutoff used in the co-expression benchmarking analyses (**Figure 2,3; Methods**). This is because the two analyses have very different goals, while the benchmarking was more technical in nature, we aimed to uncover new biology in the co-eQTL analyses. Thus we used a less stringent selection in the co-eQTL analysis to ensure that we did not miss out on detecting relevant biological processes underlying gene regulation.

An additional challenge is the large number of missing co-expression values for gene pairs within individuals. This is introduced by the sparsity of the scRNA-seq data: correlation is missing when the expression of one gene is zero in all cells of an individual. We argue that these missing co-expression values may not reflect true null correlations between gene pairs because zero values in single-cell data can also be caused by lowly expressed genes not being quantified accurately. As we observed that replacing missing values with 0 can lead to spurious co-eQTL results (**Supplementary Figure 16**), we removed such missing correlations when calculating co-eQTL correlations rather than imputing them with 0 correlation values.

Finally, we applied a customized permutation strategy for each gene pair. Since common upstream regulators might lead to co-expression of many co-eGenes, we expect correlated test statistics among the family of tests carried out for each SNP–eGene pair. Therefore, we applied a customized permutation strategy per SNP–eGene pair and an adapted multiple testing correction strategy based on fastQTL (29,30) (see **Methods** for details).

With our co-eQTL mapping strategy, we conducted a meta-analysis with four of the five single-cell datasets (Oelen v2 and v3, van Blokland v2 and the van der Wijst dataset). This identified cell-type-specific co-eQTLs for 72 independent SNPs, affecting 946 unique gene pairs in total (**Supplementary Table 5, Supplementary Table 6**). We identified the maximum number of 500 co-eQTLs in CD4+ T cells, comprising 30 SNPs, 500 gene pairs and 420 unique genes. We identified the minimum number of 35 co-eQTLs in B cells, comprising 1 SNP, 35 gene pairs and 36 unique genes.

We first examined the cell-type-specificity of these co-eQTLs. This analysis is limited by the fact that, due to our filtering strategy, we used a different set of cell-type-specific eQTLs and tested a different set of co-eGenes. Consequently, this resulted in very different sets of tested triplets for biologically different cell types, which could explain the small overlap of significant co-eQTLs between cell types (**Supplementary Figure 17, 18a; Supplementary Table 6**). Therefore, to give a complete picture of the cell-type-specificity of co-eQTLs, we replicated co-eQTLs from each cell type in all other cell types and quantified this with two different measures: 1) the ratio of co-eQTLs that could be tested in the replication cell type (**Supplementary Figure 18a**) and 2) the r_b concordance measure (31), which reflects the correlation of the effect sizes for the co-eQTLs that were tested in the replication cell type (**Figure 4a, Supplementary Figure 18b,** details in **Methods**). Consistent with the co-eQTL overlap results, the ratio of tested co-eQTLs are generally small, ranging between 5% to 97% (**Supplementary Figure 18a**). However, for the SNP–eGene–co-eGene triplets that were tested in the replication cell type, their effect sizes and directions were generally highly concordant, with a median r_b value of 0.85 (**Figure 4a, Supplementary Figure 18b, Supplementary Table 7**). The highest r_b were observed between CD4+ T cells and CD8+ T cells (0.97 for co-eQTLs identified in CD4+ T cells replicated in CD8+ T cells, 0.99 for co-eQTLs identified in CD8+ T cells replicated in CD4+ T cells).

To validate our co-eQTL results, we first examined the effect sizes and directions among the datasets used in the meta-analysis and observed high correlations (**Supplementary Figure 19**). Next, we replicated them in the BIOS bulk whole blood dataset (N = 2,491 excluding common individuals, see **Methods**) (2), using the ratio of tested co-eQTLs and r_b value (see **Methods**). For this replication, we used a linear regression model with an interaction term to model the associations between the expression level of eGenes and the product of genotype and the expression level of co-eGenes (see **Methods** for detailed explanation), as we have done before (8). We tested all identified co-eQTLs in the BIOS data. and their effect sizes and directions showed r_b values between 0.30 to 0.61 (**Figure 4b, Supplementary Table 8, 9**), with the highest

concordance achieved for CD4+ T cells, with an r_b value of 0.61 (SE = 0.06).

After we established a baseline for the number of co-eQTLs identified and their replication rates, we used this to evaluate various technical factors such as filtering strategy, sub-cell-type composition, sample size and cell number. We first compared the analysis to a set of co-eQTLs identified when omitting the filtering step for significantly correlated gene pairs, which increased the number of tests (**Supplementary Table 4**). While this led to detection of an increased number of co-eQTLs for the more abundant cell types (CD4+ T, CD8+ T, monocytes and NK cells) and a decreased number of co-eQTLs for less abundant cell types (B cells and DCs) (**Supplementary Tables 4, 10**), we also observed a general decrease in concordance among cell types compared to the co-eQTLs obtained with the filtering strategy (**Supplementary Figure 18, 20; Supplementary Table 11**). We then repeated the BIOS replication procedures for co-eQTLs found without the filtering strategy and observed a decrease in effect concordance compared to the set of co-eQTLs identified with the filtering strategy (**Supplementary Figures 21-23; Supplementary Tables 12, 13**), indicating that the filtering increases the robustness of the co-eQTLs.

We additionally explored the correlation mean and variance, as well as the non-zero ratio for co-eQTLs compared to non-significant triplets, in the scenarios with and without additional filtering (**Supplementary Figure 24**). Here we observed that significant co-eQTLs show both a higher co-expression correlation mean and variance and a higher non-zero ratio for their expression (**Supplementary Figure 24**) compared to non-significant triplets. This is to be expected as gene pairs with a high average co-expression correlation more likely reflect true biological associations and gene pairs with a high correlation variance likely reflect true co-expression network polymorphisms. This trend is also much clearer for the filtered set compared to the non-filtered set (**Supplementary Figure 24**), suggesting that alternative preselection strategies could be envisioned that are based on specific expression values or co-expression correlation variance thresholds.

Sub-cell-type composition is a potential confounder that might introduce false positive co-eQTLs, similar to cell type-composition in bulk studies (32). If a genetic variant is associated with sub-cell-type composition, co-eQTLs with sub-cell-type-specific genes might be identified even when there is no direct association between the SNP and the co-expression. To assess this, we analyzed co-eQTLs found among classical monocytes, non-classical monocytes and the whole set of all monocytes. Here we found that co-eQTL effect sizes are highly concordant ($r_b \geq 0.9$) (**Supplementary Figure 25**) for co-eQTLs tested in one of the subtypes and in the major cell type (>82% of co-eQTL identified in monocytes were tested in both classical monocytes and non-classical monocytes). This suggests that the co-eQTLs are not generally driven by sub-cell-type composition, although individual co-eQTLs could still be caused by sub-cell-type differences.

To highlight how future co-eQTL analyses can benefit from the expected expansion of population-based scRNA-seq datasets with available genotype data,

we determined how the number of identified co-eQTLs is related to the number of individuals and cells per individual. To test the influence of the number of cells, we randomly subsampled the CD4+ T cells and monocytes per individual and repeated the co-eQTL mapping pipeline (**Figure 4c**). For the influence of the number of individuals, we randomly subsampled the individuals for CD4+ T cells (**Figure 4d**). We observed that the number of co-eQTLs is linearly and positively correlated with both the number of cells and the number of individuals, although the number of individuals had a stronger effect than the number of cells (**Figure 4c,d; Supplementary Table 5**).

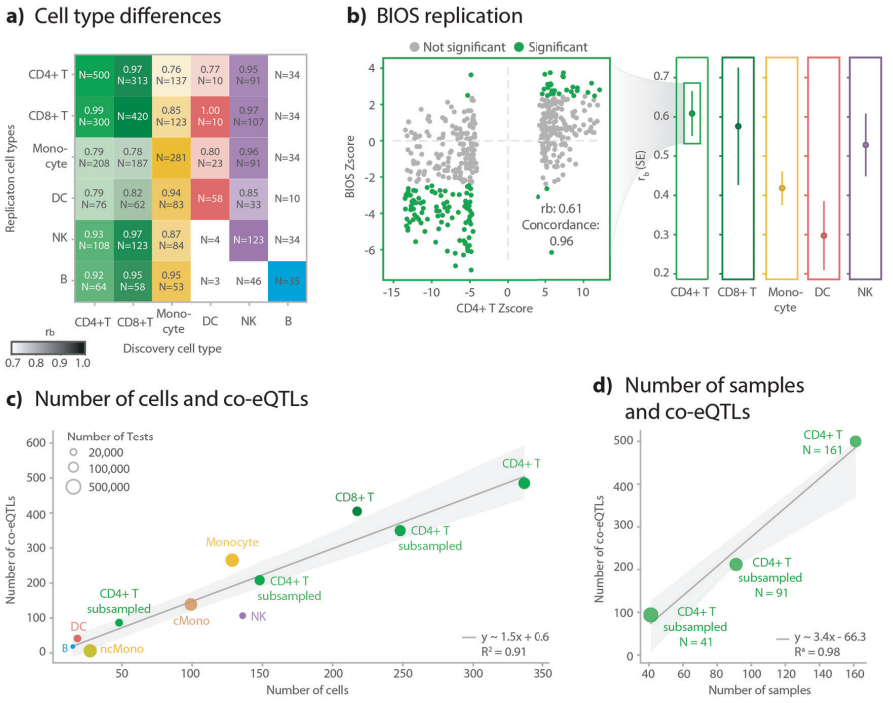


***Figure 4. General characteristics of identified co-eQTLs. a**) Replication of discovered co-eQTLs across the major cell types. Correlation of the effect sizes in replications among different cell types, measured by r_b value. Text inside each block indicates the r_b value, standard error and number of replicated co-eQTLs. Color intensity indicates r_b value. **b)** Replication in BIOS dataset for different cell types, indicated by the r_b values. Scatter plot shows the detailed Z-score comparison between the co-eQTL meta-analysis and the Z-score from the BIOS replication for CD4+ T cells. **c)** Number of significant co-eQTLs for varying cell numbers. Dot color indicates the cell type, as indicated in the text next to each dot. "cMono" means classical monocytes. "ncMono" means non-classical monocytes. "CD4+ T Subsampled cells" means that this analysis was done for CD4+ T cells, but for every individual we randomly downsampled cells to the desired cell number as indicated in the x-axis. **d**) Number of significant co-eQTLs for varying sample numbers. "CD4+ T Subsampled Individuals" indicates that this analysis was done for CD4+ T cells, but we randomly subsampled for the individuals.*

Annotating identified co-expression QTLs

After we successfully validated the identified co-eQTLs by exploring different technical aspects and replicating them in the BIOS dataset (2), we examined to what extent the co-eQTLs could provide interesting biological insights into genetic regulation, which could be relevant for the interpretation of disease variants. As discussed in the Overview, we hypothesize that among the co-eGenes identified for each SNP–eGene pair there are direct regulator genes or genes co-expressed with the direct regulators for the eGene. Even if the direct upstream regulatory factor was not evaluated in the co-eQTL analysis, due to the limited capturing efficiency of the single-cell data, the biological function of the co-eQTLs could still be inferred by the other co-eGenes in strong co-expression with the unknown upstream regulator as they presumably share the same biological function and potentially also a common role in disease. To assess these hypotheses, we combined different lines of evidence: functional enrichment based on gene ontology (GO) terms, enrichment of TF binding sites and enrichment of GWAS annotations.

Each enrichment analysis was run separately per cell type and for all co-eGenes associated with the same SNP–eGene pair (see **Methods** for details). To increase the power of enrichment analyses, we restricted ourselves to SNP–eGenes pairs with at least five co-eGenes, which covered 25% of SNP–eGenes pairs in at least one cell type (19 out of 76 unique SNP–eGene pairs). GO enrichment analysis revealed shared functional pathways for the co-eGenes. For 18 of the 19 SNP–eGene pairs, we found enrichment among the associated co-eGenes for at least one GO term (**Supplementary Table 14**). Moreover, we assessed potential common TFs regulating the shared function of these co-eGenes using ChIP-seq data processed by ReMap 2022 (33) and found enrichment of TF binding sites among the co-eGenes for 7 of the 19 SNP–eGene pairs (**Supplementary Table 15**). For four of the SNP–eGene pairs, the co-eQTL SNP itself or a SNP in high linkage disequilibrium (LD) ($R^2 \geq 0.9$) lay in the binding region of the enriched TFs (**Supplementary Table 15**), making these likely candidates for the direct regulator.

We also explored whether co-eQTLs and the respective sets of co-eGenes could enhance our understanding of disease-associated variants. For this, we annotated co-eQTL SNPs with GWAS loci, identifying approximately half the SNPs to be in high LD ($R^2 \geq 0.8$) with a GWAS locus (41 out of 72 SNPs, **Supplementary Table 16**). To assess if sets of co-eGenes for a specific SNP-eGene share a common role in disease, we explored if the co-eGenes show higher gene level trait association for GWAS traits that are also associated with the respective co-eQTL SNP. We identified overlapping GWAS traits for two co-eQTL SNPs and their co-eGenes for at least one GWAS trait and cell type, with many of the traits covering blood cell counts and immune-mediated diseases (GWAS SNP p-value $< 5\times10^{-8}$, FDR <0.05, **Supplementary Table 17**), further strengthening the biological connection of the co-eGenes with the eQTL.

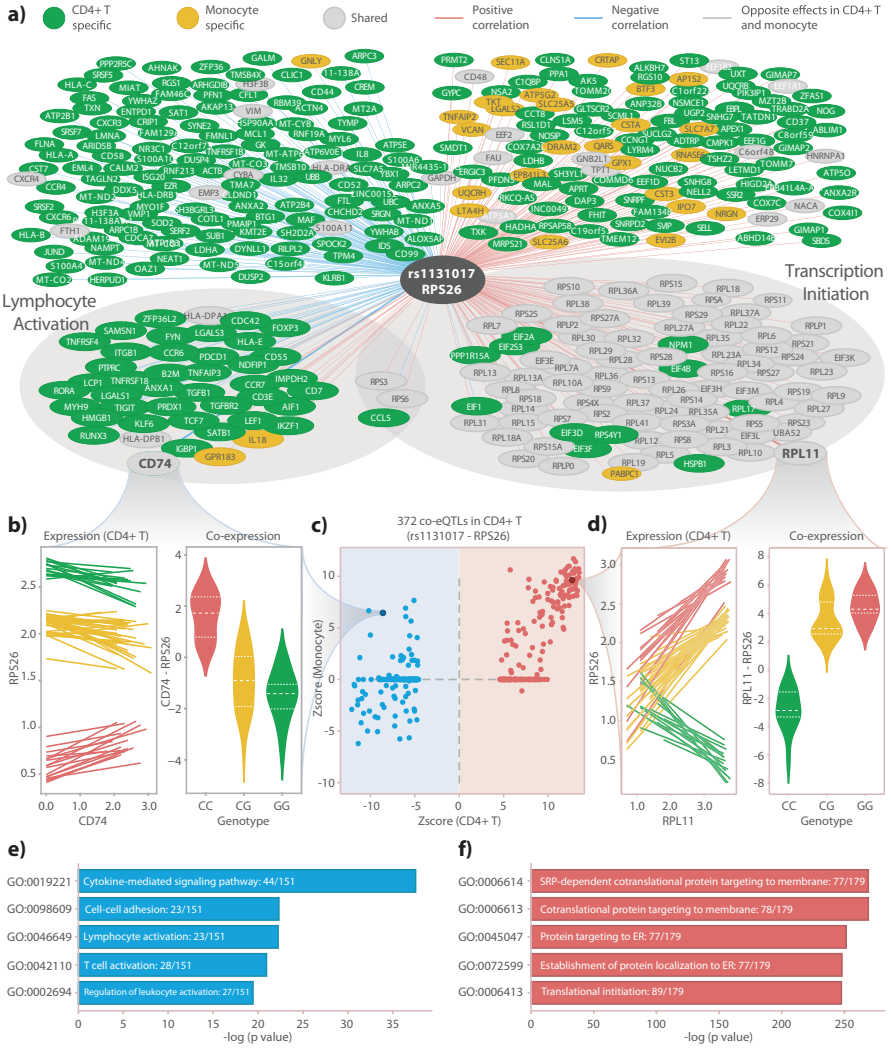


Figure 5. Annotation of co-eQTLs. ***a****) Union network constructed with co-eQTLs found in CD4+ T cells or monocytes that are associated with the SNP–eGene: rs1131017–RPS26. The two circled clusters contain co-eGenes that are in those corresponding GO terms.* ***b****) Example of one co-eQTL: rs1131017–RPS26-CD74. Left plot indicates the co-expression patterns from all individuals in the Oelen v3 dataset. Each regression line was fitted with expression data from one individual. Right plot indicates the co-expression values from the three genotype groups.* ***c****) Comparison between z-scores from monocytes and z-scores from CD4+ T cells. Red dots indicate positive co-eQTLs from CD4+ T cells. Blue dots indicate negative co-eQTLs from CD4+ T cells.* ***d****) Example of one co-eQTL: rs1131017–RPS26-RPL11 with the same layout as (b).* ***e****) GO term enrichment results for the co-eGenes in negative co-eQTLs from CD4+ T cells.* ***f****) GO term enrichment results for the co-eGenes in positive co-eQTLs from CD4+ T cells.*

Furthermore, we observed that the direction of effect of the co-eQTLs can be helpful in grouping genes sharing the same functions. For this, we compared the direction of effect of the co-eQTL with the direction of the associated eQTL, choosing the same reference allele in both cases. If the direction matched, we classified it as concordant. In these co-eQTLs, increasing expression of the eGene led to increasing co-expression. If the directions did not match, we said the direction of the co-eQTL is discordant. Between 37% and 97% of the co-eQTLs showed a concordant direction of effect across cell types (**Supplementary Figure 26**), but the majority of co-eGenes were associated with rs1131017–*RPS26* and thus the observed distributions are probably not generalizable for future larger studies that identify more co-eQTL.

In the following section, we highlight some examples of how these co-eQTL can help to better understand the molecular functional consequences of genetic variants associated to disease.

When grouping co-eQTLs based on their associated eQTL, eQTL rs1131017–*RPS26* had the most significantly associated co-eGenes in all cell types except for DCs (between 372 co-eGenes for CD4+ T cells and 35 for B cells) (**Figure 5a,b,c,d**). *RPS26*, encoding a ribosomal protein, showed strong correlation with other ribosomal proteins, and we had previously reported a few *RPS26* co-eQTLs in CD4+ T cells (10) and monocytes (11). Our new methodology and the larger sample size in the current study allowed us to now compare the genes part of the rs1131017–*RPS26* co-eQTLs across cell types.

In monocytes, NK cells and B cells, nearly all the associated genes showed a positive direction of effect, concordant with the eQTL direction (95% of all co-eGenes for monocytes, 90% for NK cells and 97% for B cells), while in CD4+ T cells and CD8+ T cells, several genes showed a negative direction of effect, discordant with the eQTL direction (46% of all co-eGenes for CD4+ T cells and 43% for CD8+ T cells).

The positively associated genes replicated well across all cell types (**Figure 5c, Supplementary Figure 27**) and were enriched for functions associated with translation (**Figure 5e**), which is consistent with the fact that many co-eGenes were ribosomal proteins from both the large and the small subunit (for CD4+ T cells: 46 of 47 tested RPL genes and all 31 tested RPS genes were associated). In contrast, the negatively associated genes only replicated well between CD4+ T cells and CD8+ T cells (**Supplementary Figure 27; Figure 5c,d**), despite the fact that these genes were well expressed in the other cell types. This negatively associated set of genes showed enrichment in functions associated with immune response and T cell activation (**Figure 5f**).

TF enrichment analysis identified six TFs—*RBM39, TCF7, LEF1, KLF6, CD74* and *MAF*—whose binding sites were enriched in the promoter region of the rs1131017–*RPS26* co-eGenes, that had a binding site overlapping with rs1131017 and that were among the rs1131017–*RPS26* co-eGenes themselves (**Supplementary Table 15**). This led us to the assumption that one or more of these TFs represent the direct regulators of the eQTL, as described in the Overview (**Figure 1c, Scenario 1**). Five of the TFs (*TCF7,*

LEF1, KLF6, CD74 and *MAF*) are also connected with lymphocyte activity (the first four based on GO annotations, *MAF* based on a recent study (34)), further strengthening the link with T cell activation. Of these, *MAF* and *CD74* were specifically enriched not only among all co-eGenes but additionally among co-eGenes with a negative effect direction (**Supplementary Table 15**).

GWAS enrichment analysis showed enrichment for several different blood cell counts, but only in CD4+ T cells and CD8+ T cells we also observed enrichment for immune-mediated disease (rheumatoid arthritis, Crohn's disease (CD), multiple sclerosis (MS) and hay fever), further connecting the co-eGenes especially with T cells (**Supplementary Table 17**). Interestingly, several studies have highlighted a connection of RPS26 with T cell activation and survival (35,36), and the associated co-eQTL SNP rs1131017 is associated with the enriched immune-mediated diseases (rheumatoid arthritis, CD, MS, hay fever), and additionally associated with type 1 diabetes (T1D) and with other autoimmune traits (37).

We examined whether the large number of co-eQTLs for rs1131017 were confounded by sub-cell-types in CD4+ T cells. We cannot exclude the possibility that this variant showed this effect in CD4+ and CD8+ T cells by specifically affecting the amount of circulating CD4+ or CD8+ sub-cell types whose marker genes would subsequently show up as co-eQTLs in our analysis, where we have not distinguished between sub-cell types. To test whether this is a possibility, we associated SNP rs1131017 and the ratio between CD4+/CD8+ TEM cells and CD4+/CD8+ naive T cells, but we did not see a significant correlation (**Supplementary Figure 28**). Together, these results suggest that RPS26 plays a dual-function role, both in general transcription and specifically in lymphocytes in T cell activation. This points to a potential working mechanism in the role of rs1131017 in the manifestation of autoimmune diseases.

Another set of promising co-eQTLs are those associated with rs7806458–*TMEM176A* in monocytes (11 co-eGenes) and rs7806458–*TMEM176B* in monocytes (6 co-eGenes) and DCs (1 co-eGene) as they connect the co-eQTL SNP rs7806458 that has been associated with MS (38) with blood coagulation. This is interesting as this disease has previously been connected to disturbances in blood coagulation (39). The relevance of the co-eGenes to MS is supported by two lines of evidence. Firstly, GO enrichment suggested that the six co-eGenes associated with rs7806458–*TMEM176B* in monocytes are enriched for complement component C3b binding (**Supplementary Table 14**), which is closely related to the blood coagulation system (40). When looking closely at the exact gene functions, we found three genes (*ITGB1, FCN1* and *CFP*) that contribute to the local production of complement (41). Secondly, GWAS enrichment analysis showed MS enrichment for co-eGenes associated with rs7806458–*TMEM176A* in monocytes (**Supplementary Table 17**). Intriguingly, the eGene *TMEM176B* was previously found to be associated with the maturation of DCs (42), and it has been shown that white blood cells, including DCs, can act as a local source of certain complement proteins (43,44). Though we could not identify (in)direct regulator genes for these co-eQTL in our TF enrichment analysis with the ReMap

database, we argue that these co-eGenes, supported by several lines of evidence, provide valuable mechanistic insights for the MS SNP rs7806458.

For several of the other co-eQTLs, we could not provide as strong and coherent evidence for the interpretation but nevertheless found promising connections to biological functions and disease that can be explored in further studies. One is the SNP–eGene pair rs9271520–*HLA-DQA2*. We found co-eQTL effects for it in CD4+ and CD8+ T cells, monocytes and DCs, with the number of co-eGenes ranging from 7 to 17. Interestingly, rs9271520 is in LD with several immune disease SNPs where we also found enrichment for the co-eGenes in the same GWAS traits. The most significant (sorted by GWAS SNP p-values) enriched traits include rheumatoid arthritis, MS and asthma (see **Supplementary Table 17** for full GWAS enrichment results). However, we found several other genes in the HLA region being co-eGenes associated with rs9271520–*HLA-DQA2*, and, when we removed those HLA genes, the GWAS enrichment signals disappeared. This indicateed that the enriched signal could be due to the LD structure in the HLA region and a confident mapping of the causal regulatory connections is not possible with our dataset. Other interesting co-eQTL examples and their interpretations are discussed in **Supplementary Text**.

In general, our study is still underpowered in finding a lot of associated co-eGenes (**Figure 4c,d**, **Supplementary Figure 15**). This limits the set of SNP-eGenes, for which we can perform a well powered enrichment analysis and so the biological interpretation of these co-eQTLs. One of the potentially interesting SNP-eGenes, with too few co-eGenes for the enrichment analysis, is rs393727 - *RNASET2*, which is associated with four co-eGenes (*B2M, ITGB1, ALOX5AP, CRIP1*). The SNP rs393727 is in very high LD with two previously described SNPs associated with Crohn's disease (CD) and inflammatory bowel disease (IBD) (**Supplementary Table 16**), eGene *RNASET2* has also been previously associated with IBD (45), and among the four co-eGenes, *ITGB1* was associated with CD (46) and *CRIP1* is associated with gut immunity (47).

Intriguingly, we found a number of overlapping co-eGenes associated with different SNP–eGene pairs, indicating potential common upstream regulatory pathways. For example, all the co-eGenes positively associated with rs4147638–*SMDT1* are also found to be positively associated with rs11311017–*RPS26*, while the four co-eGenes negatively associated with rs393727–*RNASET2* are also negatively associated with rs1131017–*RPS26* (**Supplementary Figure 29**).

Discussion

In this study, we validated the use of scRNA-seq data to identify cell-type-specific co-expression patterns and developed a novel approach to extend the discovery of co-eQTLs. Applying this to a large meta-analysis with 173 samples, we identified 72 independent SNPs leading to co-eQTLs for 946 unique gene pairs across different cell types. These co-eQTLs shed light on the biological processes upstream of individual *cis*-eQTLs, such as that seen for rs1131017, which affects

RPS26 expression levels and is associated to autoimmune diseases. We observed that this variant affects T cell activation genes, providing a potential explanation for the association of this variant to autoimmune diseases.

In this study, we used Spearman correlation to quantify the co-expression patterns from scRNA-seq data because of its straightforward interpretability, scalability, robustness against outliers and high reproducibility among different scRNA-seq and bulk RNA-seq datasets. However, we acknowledge that such correlations do not take into account the sparseness of scRNA-seq data, and it is difficult to infer direct regulator genes. This of course also depends on the quality of the single-cell data. Direct interactions can only be distinguished from indirect interactions when the direct upstream target was measured, which is currently not always the case. As other association methods (16,20) that were top-performing in independent benchmarking studies (15,18) did not perform better in our validation and a reliable temporal ordering of the cells (25,26) was not possible in our dataset, we applied Spearman correlation as a solid basis for the co-eQTL analysis. Future work may find that other association measures are equally suitable or more suitable, and this may potentially depend on the specific single-cell dataset under investigation.

We also found that scRNA-seq and bulk RNA-seq data do not always correlate well for all gene pairs and explored different factors that could explain this. Part of the variable correlation could be explained by the sparsity of the single-cell data, as higher expressed gene pairs correlated better, but at least a few example cases showed the potential occurrence of Simpson's paradox. With regards to cell type–composition, however, the FACS-sorted datasets did not correlate better with single-cell datasets than the whole-blood bulk dataset, which could either be caused by the smaller sample size of the single-cell data, technical changes introduced by FACS or specific differences in the (sub-)cell types, as we had naive CD4+T cells and classical monocytes (subsets of CD4+T cells and monocytes, respectively) for BLUEPRINT and ImmuNexUT that we tested for the single-cell data. Another interpretation is that scRNA-seq and bulk RNA-seq data capture different functional gene clusters, as a previous study showed in tumor samples (48). One possible explanation for this is that bulk and single-cell capture different sources of variability. Whereas single-cell data captures between-cell variability, bulk data captures between-person variability, which is affected by additional factors like genetics and environment. Therefore, a statistical framework combining both data types could be beneficial in the future.

Our study sheds light on several important considerations for future scRNA-seq study design regarding personalized network construction and co-eQTL mapping. Firstly, we showed that several factors, including cell number and gene selection, greatly influence the stability of co-expression patterns. We observed a clear trend indicating that a certain minimum number of cells from one individual is needed to achieve a stable co-expression pattern (**Figure 3d**). Secondly, we also explored factors influencing the number and quality of co-eQTLs. We showed that the number of significantly detected co-eQTLs can be greatly increased by either increasing the

number of individuals or by increasing the number of cells per individual (**Figure 4c, d**). We believe that future larger single-cell datasets such as two very recent studies (49,50) and the sc-eQTLGen consortium (51) will improve statistical power to identify more robust co-eQTLs.

Furthermore, we showed that a sophisticated filtering strategy of tested SNP–gene–gene triplets is essential to maximize the number of reliable co-eQTLs. However, we also suggest that the filtering strategy should be designed for the specific goals of the respective analysis. In this study, we systematically searched for robust co-eQTLs and adapted our strategy to balance the trade-off between achieving a stable co-expression pattern and enlarging the search space. For this reason, we first selected SNP–gene pairs and then used co-expression strength as an additional criterion rather than the very stringent expression cutoff criterion we used in our benchmarking analysis. In contrast, in our previous study (11), we focused specifically on co-eQTLs among the eQTLs that changed after pathogen stimulation and performed a strict pre-filtering for a highly targeted analysis. In the current study, we were, in particular, able to replicate the most significant co-eQTLs from the targeted analysis (**Supplementary Figure 30**). While the targeted analysis identified additional lower significance co-eQTLs that are below our much stricter multiple testing-corrected significance threshold, we were able to quantify the number of co-eQTLs more broadly for several additional SNPs and to include, for the first time, a comparison across cell types. In other cases, a selection of known TF–target pairs or pathway information could be desirable, e.g. for prioritizing TFs connected with diseases for experimental validation purposes.

We showed that the annotated co-eQTLs could identify potential direct regulators of the associated eQTLs as well as the affected biological processes, with several examples based on a combination of different enrichment analyses. We identified several TFs either directly as co-eGenes or via enriched binding sites among the co-eGenes of a SNP–eGene pair, providing potential regulatory mechanisms for explaining the co-eQTL. For the eQTL rs1131017–*RPS26*, six enriched TFs were themselves co-eGenes in CD4+ T cells, providing compelling evidence to support the hypotheses that direct regulators can be identified among co-eQTLs. Among these six TFs, five are associated with lymphocyte activation, further strengthening the connection of the eQTL with lymphocyte activation and through this to autoimmune diseases.

Another interesting aspect of the rs1131017–*RPS26* example is that we revealed a potential mechanism for a previously described GWAS signal by showing cell-type-specific genetic regulation of a multi-functional gene. By comparing T cells and monocytes, we identified that *RPS26* may be involved in two distinct biological functions. Interestingly, these two distinct functional co-eQTL clusters are characterized by opposite effect directions. Moreover, while *RPS26* showed enough variation to be picked up as an eQTL effect, it did not show high correlation with either gene cluster (**Supplementary Figure 31**), which may be why understanding its role in multiple functions has been challenging up to now (36,52). We envision that more multi-functional eGenes could operate in such a cell-type-specific manner, with

variation in expression that could be explained as the downstream consequences of many other conserved or highly co-expressed gene clusters, and this understanding could assist in interpreting GWAS signals. We also observed that different eGenes could have shared upstream genes/pathways as we identified four common immune-related co-eGenes associated with rs393727–*RNASET2* and rs1131017–*RPS26*, and both SNPs were in LD with immune diseases (T1D and CD), suggesting a shared upstream process for these two eQTL effects. By providing cell-type-specific gene regulation backgrounds through co-eQTLs, we expect more eQTLs and GWAS signals to be explained in the relevant cell type via future large-scale co-eQTL studies.

For the other co-eQTL examples, no enriched TF was in the co-eGene list, potentially because the TFs were not measured in the scRNA-seq datasets due to low expression. Here, the enrichment allowed us to still identify relevant TFs for further exploration. A third group of co-eQTL examples were supported by GWAS or GO enrichment analysis but not TF enrichment analysis. Here, the co-eGenes revealed part of the disease-relevant network, but we could not pinpoint the direct regulatory TFs. One explanation for this may be that our study is still underpowered to discover co-eGenes, while the enrichment strategy works best when there are a substantial number of co-eGenes as for rs1131017–*RPS26*. Based on our evaluation, we estimated that future studies with larger sample size and more cells will identify many more co-eQTLs (**Figure 4c,d**). This can help identify the direct regulators for some of our other examples, where the current enrichment analyses provided no clear interpretation, as well as co-eQTLs associated with other SNP–eGenes.

There are also several challenges to interpret the identified co-eQTLs. Firstly, as discussed earlier, it is difficult to determine the direct and indirect regulators that work through co-expression among correlated co-eGenes. This creates problems in using correlation-based metrics to quantify replication performance. For example, all the co-eQTLs we identified in B cells were associated with the rs1131017–RPS26 pair, making the correlation-based r_b measure invalid for this case. Also, to reduce the multiple testing burden, we only tested the top-SNP, a choice that could pose additional challenges for follow-up analysis such as colocalization to identify the causal SNP. Moreover, comparison of co-eQTLs between cell types remains challenging. We showed that the number of co-eQTLs is strongly driven by the number of cells (**Figure 4c**), so that it is not meaningful to only compare the absolute number of co-eQTLs between cell types in the current study. Furthermore, the sparsity of the single-cell data lead to the removal of many lowly expressed genes which, combined with the strict filtering our analysis required, meant only a small number of genes were tested in all cell types. In addition, sub-cell-type composition can introduce false positive co-eQTLs within a cell type if a genetic variant influences the sub-cell-type composition and one of the tested genes shows sub-cell-type-specific differences in expression. However, in our evaluation of classical and non-classical monocytes, we observed no strong confounding of monocyte co-eQTLs by the sub-cell types (**Supplementary Figure 24**). We also found several SNPs in LD with GWAS SNPs for traits such as monocyte counts, but there was

no additional evidence that these SNPs have an effect on (sub-)cell type composition. Still, these analyses were limited to a smaller number of samples and a small number of cells in the sub-cell types, so we cannot exclude that some co-eQTLs were caused by sub-cell-type composition effects of the co-eQTL SNPs.

Several of the limitations of our current analysis will be overcome by on-going technological developments. First of all, we expect that follow-up analyses with larger sample sizes and more cells per person will identify many additional co-eQTLs. This can be further enhanced by improvements in single-cell technologies that lead to better capture efficiency of expressed genes. CITE-seq (53) and similar technologies (54,55) allow improved cell type and sub-cell type classification that can show the effect of sub-cell type differences more accurately. The combination of multiple-omics, such as scRNA-seq, scATAC-seq and/or single-cell proteomics (56–58), will enable us to capture regulation happening outside the mRNA level, and lead to improved association analysis of gene pairs above standard Spearman correlation.

Conclusion

Through our co-eQTL mapping strategy we identified a robust set of co-eQTLs that provides insight into cell-type-specific gene regulation and leads for future functional testing. Among these results, we uncovered a potential mechanism for a previously identified GWAS signal and a multi-functional gene. Our evaluation of different technical factors provides valuable suggestions for future experimental study design. We believe that more co-eQTLs will be uncovered by applying our general co-eQTL mapping pipeline to future large-scale scRNA-seq data. We envision that these co-eQTLs will in the future help to position eQTL and GWAS signals into cell-type-specific GRNs by annotating which regulatory edges are affected by which genetic variants. This knowledge is important for interpreting the effects of genetic variants in general, but also specifically for improve personalized medicine through better genetic risk prediction for diseases and personalized drug treatment based on genotype (51).

Methods

Single-cell datasets

Three different scRNA-seq datasets were included in this study, both for benchmarking the associations and for combined meta-analysis of co-expression QTLs. All five datasets from the three studies were generated from human PBMCs and are referred to by their first author: the Oelen dataset (n = 104 donors) (11), the van der Wijst dataset (n = 45 donors) (10) and the van Blokland dataset (n = 38 cardiac patients) (19). Further specifications can be found in **Supplementary Table 1** and the respective manuscripts. The processed versions of the datasets from the original publications were used, including quality control and cell type identification (for details,

see the respective publications). The Oelen dataset also contains cells stimulated with different pathogens, but we only included the unstimulated cells in this analysis to improve comparison with the other datasets. For the van Blokland dataset, we included the data from the time point 6–8 weeks after the individual was admitted to the hospital for myocardial infarction, again to improve comparison across datasets.

For cell type classification, we took the annotation for the Oelen data from their original publication (11) and annotated the van Blokland and van der Wijst datasets using the Azimuth classification method (59). For Azimuth classification, we used the following settings: 1) the FindTransferAnchors function to find anchors using the reference from publication (59), normalization method "SCT", reference reduction "spca" and first 50 dimensions and 2) the MapQuery function to annotate cell types using the same reference and parameters such as reference.reduction = "spca" and reduction.model = "wnn.umap". We then compared the annotation from the Oelen publication and the Azimuth classification and found high correspondence (**Supplementary Figure 32**). For analyses using the sub-cell-type classification, we always refer to the Azimuth classification results.

Single-cell co-expression

We calculated the Spearman correlation of gene pairs in the three different single-cell studies (Oelen dataset (11), van Blokland dataset (19) and van der Wijst dataset (10)) and then compared between datasets and 10XGenomics chemistry. In the benchmarking section, correlation was calculated separately per cell type but together over all individuals and only for gene pairs for which both genes were expressed in at least 50% of the cells from the respective cell type. For the comparison between two datasets, the gene pair–wise Spearman correlation values from each dataset were compared using Pearson correlation.

Rho calculation

Rho proportionality was calculated using the "propr" function in R, from the "propr" package, with the symmetrized value set to true. We used the v3 unstimulated monocytes to compare the rho proportionality values to Spearman-rank correlations of the same data. We filtered out genes expressed in fewer than 5% of cells, leaving 8,634 genes to be assessed. Concordance between rho values and Spearman correlations was assessed with Pearson correlation.

We also explored rho proportionality values for very lowly expressed genes because the log-normalization of the method potentially introduces false associations for these genes (60). However, the computational demand to run the method was so high that we could not evaluate all expressed genes at once. Instead, we subsampled a set of 50 very lowly expressed genes (expressed in 0–5% of the cells) and 50 very highly expressed genes (expressed in at least 90% of the cells) and calculated the rho proportionality and Spearman correlation values for each combination of these 100

genes. We then compared gene pairs for which both genes were lowly expressed, pairs for which both genes were highly expressed, and mixed pairs, for which one gene was lowly and one highly expressed.

Alternative association metrics besides rho proportionality and Spearman correlation are discussed in **Supplementary Text**.

Validation in bulk datasets

Spearman correlations from single-cell data were compared to Spearman correlations made with three different bulk datasets: the BLUEPRINT Epigenome consortium data (21), the ImmuNexUT dataset (22) and the BIOS dataset(2). For BLUEPRINT, we further removed the first principal component from the monocyte dataset to remove any uncorrected covariates. For the ImmuNexUT dataset, preprocessing was performed as described in the publication: we filtered out genes with less than 10 counts in 90% of the samples, performed TMM normalization with edgeR and scaling to CPM, batch corrected with combat and removed samples with a mean correlation coefficient smaller than 0.9. For the BIOS dataset, we corrected for 20 RNA Alignment metrics and then calculated the co-expression values using all individuals.

We then calculated the Pearson correlation across all gene pair-wise correlation values. As BLUEPRINT and ImmuNexUT are cell type-sorted datasets, we matched the cell types between bulk and single-cell data in these cases in the comparison. Again, we used only genes expressed in at least 50% of the cells from the cell type. This threshold was chosen after our initial evaluation of different thresholds from 10% to 90% in the comparison of BLUEPRINT and Oelen v3 dataset, with 50% chosen to balance the number of genes that can be used against the correlation strength between the datasets.

Validation using CRISPR knockout data

To further validate the correlation values, we used CRISPR knockout data from (23). Mixscape was used to identify perturbed vs unperturbed cells for each CRISPR perturbation (27). We selected five knockout genes for which a sufficient number of successful CRISPR-perturbed cells were identified and that were expressed in our single-cell dataset (Oelen v3 dataset, CD4+ T cells) in > 50% of cells. The publication identified DE genes in wild-type vs perturbed cells and wild-type vs non-perturbed cells, as labeled by Mixscape. We selected a credible set of DE genes that were expressed in the single-cell dataset and significant in the wild-type vs perturbed cells but not in the wild-type vs non-perturbed cells. For this, we applied FDR-correction based on all genes expressed in the single-cell dataset. The correlation of these genes was compared to the correlation of non-DE genes, i.e. all other genes expressed in the single-cell dataset, using the Wilcoxon rank-sum test (one-sided test with “greater” in DE genes). The same test was done using the naive CD4+ T cells from the ImmuNexUT dataset.

Validation using STRING annotations

Following the same approach used for the CRISPR knockout data, we explored if gene pairs whose proteins are interacting show higher correlation. We used the STRING database (version 11) (28), processed by the (18) benchmark study**,** to identify interacting gene pairs. We compared the correlation of gene pairs in STRING versus gene pairs not listed in STRING via Wilcoxon rank-sum test (one-sided test with "greater" for Gene pairs in STRING): once using the correlation estimates from the Oelen v3 dataset and once using the estimates from the ImmuNexUT dataset, both times for the CD4+ T cells and filtered for genes expressed in > 50% of single cells.

Exploring Simpson's paradox

To identify whether our strategy to identify single-cell co-expression is affected by Simpson's paradox and whether bulk-based approaches would suffer from it, we studied the co-expression outcomes for two different strategies. In both strategies, we only included genes with non-zero expression in at least half of all monocytes in the Oelen v3 dataset. In the first strategy, we calculated Spearman correlations for gene pairs per individual separately for each gene pair. In the second strategy, we calculated the average expression of genes per individual and then calculated the Spearman correlation between genes. To identify potential Simpson's paradox events, we looked into the gene pairs that had the largest deviation in co-expression estimate between the two strategies.

Comparison between cell types

After successful validation of the Spearman correlation values, we compared differences between cell types within one dataset for the Oelen v2 and v3 dataset. Here we applied the same strategy as in the dataset comparison. We selected genes expressed in 50% of the cells from both cell types for each corresponding comparison, calculated Spearman correlation per gene pair within each cell type and followed up with Pearson correlation to compare both cell types. We also explored the absolute distribution of correlation coefficients between the cell types.

Comparison between individuals

Again, we applied the same strategy as for the cell type and dataset comparison. We calculated gene pair–wise Spearman correlation values for each cell type and donor separately, taking all genes expressed in 50% of cells from the cell type in general (not per donor). We then compared each donor with each other donor by calculating Pearson correlation over the gene pair-wise correlation values to get a distribution of how well donors match per cell type.

To explore the effect of the number of cells per donor on this distribution, we subsampled each cell type to different numbers of cells (depending on the frequency of the cell type). For this, we take all individuals with at least this number of cells in this cell type and subsample the cell number to exactly this value for each individual. We stop subsampling at a threshold for the cell type when more than 75% of all measured individuals have fewer cells than the threshold. For the four most abundant cell types (CD4+ T cells, CD8+ T cells, monocytes and NK cells), we additionally fitted a logarithmic curve separately for each cell type to better quantify the connection:

correlation_individuals ~ log(number_cells) (with log being the natural logarithm).

We then used the fitted formulas to extrapolate up to 1,500 cells for each cell type.

Power calculation

For power calculation, we use an F-test, as implemented in (61), with a sample size of 173 (the total size of the combined cohorts), a heritability between 10% and 30% and a Bonferroni-corrected significance threshold of 0.05. The range for the heritability was chosen based on previously detected co-eQTLs (11). The number of tests influences the Bonferroni-corrected thresholds and depends on the selected gene–gene–SNP triplets. Here we assumed only one SNP per gene pair and all genes are tested against each other. Then, we increased the non-zero ratio threshold for gene selection from 0 to 0.95 (monocytes, Oelen v3 dataset), got the number of tests and calculated the power.

Testing multiple SNPs per pair would further increase the total number of tests and reduce the overall power.

eQTL mapping

We performed a meta-analysis to identify significant eQTL in four out of the five single-cell datasets (Oelen v2 and v3 dataset (11), van Blokland v2 dataset (19) and van der Wijst dataset (10)). We excluded the van Blokland v3 dataset because the sample size was so small that few variants lay above the MAF threshold (see below). Due to the limited sample size, we chose to perform a constrained eQTL mapping rather than a genome-wide mapping. To select the SNP–gene pair to test for eQTL mapping, we took the eQTL results from the largest meta eQTL analysis study in whole blood (2) and selected the most significant SNP for each gene. This resulted in 16,987 SNP–gene pairs to test. For these selected SNP–gene pairs, we performed eQTL mapping using eQTLPipeline v1.4.9 (62) within a *cis*-window of 100 kb, using 10 permutation rounds for determining FDR as described in (2) and a MAF of 0.1.

Co-expression QTL (co-eQTL) mapping and the filtering strategy

First, we generated all possible combinations of the cell-type-specific eQTL findings (denoted as SNP–eGene) from the constrained eQTL mapping procedure in the respective cell type (as explained in the eQTL mapping method section above) and all other genes (denoted as co-eGene) that are expressed in the corresponding cell types. This resulted in the full list of SNP–eGene–co-eGene triplets for co-eQTL mapping analysis. We then calculated co-expression using Spearman correlation for the unique eGene–co-eGene pairs for each individual using untreated cells of the six major cell types (CD4+ T and CD8+ T cells, monocytes, B cells, NK cells and DCs) and the sub-cell-types in monocytes (classical monocytes and non-classical monocytes). For each gene pair, we counted the ratio of individuals who exhibit a significant correlation (nominal p-value from Spearman correlation < 0.05). If at least 10% of individuals showed a significant co-expression correlation for the specific eGene–co-eGene, we took this gene pair further into follow-up analysis. The total number of tests for each cell type can be found in **Supplementary Table 4**.

To assess the impact of cell numbers and sample numbers on the quality and quantity of co-eQTLs, we artificially created a few scenarios with fewer cells per individual and fewer individuals using a random subsampling strategy. To examine the impact of cell numbers, we randomly subsampled the CD4+ T cells per individual to three different levels (50, 150 and 250 cells). In each level, we kept the individuals with fewer cells, randomly subsampled those with a cell number higher than the corresponding level and performed the co-eQTL analysis using the strategies mentioned. Similarly, to examine the impact of sample numbers, we randomly subsampled 50 and 100 individuals, and excluded nine individuals with fewer than 10 CD4+ T cells for both scenarios.

Multiple testing correction strategy for co-eQTL

To account for the correlation structure for gene pairs with one common gene and genome-wide, we modified and applied the permutation-based multiple testing correction strategy from fastQTL (29), implementing the method as follows. For each SNP–eGene–co-eGene triplet, we performed 100 permutations. Then, for each SNP–eGene pair, we determined the lowest p-values per permutation over all the genes (co-eGene) tested for the SNP–eGene pair. This resulted in the 100 lowest permuted p-values per SNP–eGene pair. For each SNP–eGene pair, we fitted a beta-distribution over the 100 permuted lowest p-values, which enabled us to subsequently establish the empirical p-value for the lowest non-permuted p-value. Through this procedure we ensured that under the null test statistic each SNP–eQTLGen pair has a uniform p-value distribution. Finally, for all SNP–eGene pairs, we calculated Benjamini-Hochberg FDR over the empirical p-values.

For each SNP–eGene pair, we also derived a p-value cutoff that indicates which of the co-eGenes are significant for that SNP–eGene pair via the following steps. After

determining the FDR for all SNP–eGene pairs, we determined the empirical p-values that are closest to FDR = 0.05. Using the beta distributions for each SNP–eGene pair, we then determined its nominal p-value threshold. All co-eGenes with a nominal p-value lower than the corresponding p-value threshold for that SNP–eGene pair were considered significant.

Replication in BIOS dataset

We replicated the co-eQTL findings in bulk whole-blood RNA-seq data from the BIOS Consortium, using the same method described in a previous study (8). Briefly, we implemented the following ordinary least squares model with the Python package statsmodels (63): *eGene ~ SNP + co-eGene + SNP:co-eGene*. We then examined the effect sizes of the interaction term SNP:co-eGene and used Benjamini-Hochberg procedures for multiple testing correction.

Calculation of rb values and allelic concordance

We used the same evaluation metrics to quantify the cell-type-specificity and replication performance in the BIOS data set of the co-eQTLs. First we used the rb method with modification. We followed the same procedures as the original study (31) but chose a suggested alternate strategy to estimate errors across gene pairs between two tissues. Whereas the original paper used null SNPs per each eQTL for this purpose, we tested only the significant eQTL SNP for SNP–eGene–co-eGene triplets and therefore we did not have information for the null SNPs. Thus, we used the alternative approach indicated in the original paper with **Equation 1)**, where r_e is the estimation errors across gene pairs between two tissues, r_p is the correlation of co-expression levels between two cell types in the overlapping sample, n_s is the number of overlapping samples, n_i and n_j are the number of samples in cell typed *i* and *j*, respectively. For the BIOS replication, we excluded overlapping individuals from the BIOS RNA-seq dataset for the replication analysis. Additionally, in cases where fewer than 10 co-eQTLs were tested in the replication analysis, we could not get a robust estimation of the r_b value and hence represent them as NAs in the results section.

Equation 1 $$r_e = r_p \times \frac{n_s}{\sqrt{n_i \times n_j}}$$

Due to our filtering strategy, we did not always test the same set of SNP–eGene–co-eGene triplets in all cell types. Therefore we also need to compare the tested ratio when quantifying the cell-type-specificity. The tested ratio means the SNP–eGene–co-eGene triplets that were also tested in another cell type or the BIOS replication analysis.

A third evaluation metric that we used is allelic concordance between the discovered co-eQTLs and the results in the replication study. This is defined as the ratio of co-eQTLs with concordant effect direction by the number of significant co-eQTLs identified in the replication study.

Biological interpretation based on enrichment of GO terms, TF binding sites and GWAS variants

We explored the biological function of the co-eQTLs based on different enrichment analyses that all tested if co-eGenes associated with the same SNP–eGene pair in the same cell type show similar functional properties. For this, we selected all SNP–eGene pairs that had at least five significant co-eGenes in the same cell type.

First, we performed GO enrichment analysis separately for each co-eGene set, grouped by SNP–eGene and cell type, applying the R package clusterProfiler (ver 4.0.5) (64) and performing FDR multiple testing correction separately for each SNP–eGene pair across the different GO terms (defining enrichment below FDR < 0.05 as significant). As the background set for the enrichment, we used all genes tested in the co-eQTL analysis in the respective cell type.

Next, we explored if these co-eGene sets were enriched for certain TF binding sites. TF annotations were taken from ChIP-seq peaks processed in the ReMap 2022 database (33), which we filtered for cell lines associated with blood cell lines. We tested the overlap of these peaks with the promoter regions of the co-eGenes tested, defining the promoter region as the region 2kB upstream and downstream of the first transcription start site of the gene. Enrichment was tested based on Fisher's exact tests for each TF, using all genes tested in the co-eQTL analysis in the respective cell type as the background set. We performed FDR multiple testing correction separately for each SNP–eGene pair over all TFs (defining enrichment below FDR < 0.05 as significant). Furthermore, we explored if the enriched TF itself was a co-eGene associated with the respective SNP–eGene pair and if the co-eQTL SNP or a SNP in high LD ($R^2 \geq 0.9$) lies in a binding site of the enriched TF. The SNPs in high LD were obtained from SNiPA (65) using the variant set from the 1000 Genomes Project, Phase 3 v5, European population, Genome assembly GRCH37 and genome annotations from Ensembl 87.

For the GWAS annotations, we considered two different strategies. In the first approach, we annotated SNPs or SNPs in high LD ($R^2 \geq 0.8$) with GWAS loci from the GWAS Catalog (1), with the last updated timestamp being 3/1/2022, 07:13 AM (GMT+0100). LD information for this was taken from LDtrait (66) with the following parameters: window size = 500KB, reference population = 1000 Genomes CEU, GRCh37). In the second approach, we used the magma method (67) to assess enrichment of GWAS associations among co-eGenes. We obtained uniformly processed GWAS summary statistics for 114 traits that were used for the GWAS analysis of the GTEx consortium (68,69). We then followed the strategy previously described by (67). We defined gene sets for each co-eQTL SNP in each tissue as the set of significant co-eGenes associated with the SNP, as done for the GO and TF enrichment analysis. Protein names/gene symbols were converted to Entrez gene ids and mapped to the corresponding annotations on the human genome assembly 38. We performed individual magma analyses for each trait based on summary statistics and LD structure from the

1000 genomes European reference panel for all gene sets compared to the background set of genes tested for co-eQTL, always conditioning on default gene-level covariates (for example, gene length). Subsequently, we applied the Benjamini-Hochberg method and selected gene set–trait associations with FDR < 5%.

After we observed different distributions of co-eQTLs for rs11311017–*RPS26* with regards to the direction of effect in the different cell types, we repeated all enrichment analysis (GO, TF and GWAS) separately for the positively associated co-eGenes and negatively associated co-eGenes in CD4+ T cells.

Direction of effect

We compared the direction of effect in eQTLs and co-eQTLs by comparing the direction of the zscores. After ensuring that the reference allele aligns in the eQTL and co-eQTL analysis, co-eQTLs for which the sign of the zscore matches the sign of the eQTL zscore are called concordant. If otherwise, they are called discordant.

Declarations

Ethics approval and consent to participate

Ethics approval was requested and approved for each of the data sets. The Lifelines DEEP study was approved by the ethics committee of the University Medical Centre Groningen, document number METC UMCG LLDEEP: M12.113965. All participants signed an informed consent prior to study enrollment. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Consent for publication

All authors have read and approved the submission of this manuscript.

Availability of data and materials

All code is available on github at https://github.com/sc-eQTLgen-consortium/co-expressionQTLs
BIOS: https://www.bbmri.nl/acquisition-use-analyze/bios
Oelen datasets: https://eqtlgen.org/sc/datasets/1m-scbloodnl.html
Van Blokland data set: manuscript in preparation.
Van der wijst data set: https://eqtlgen.org/sc/datasets/vanderwijst2018.html

Competing interests

The authors declare that they have no competing interests.

Funding

Horizon2020 №860895 (MK)
NWO-VENI 192.029 (MW)
NWO-VICI, 917.14.374 Oncode Investigator grant (LF)
NWO-VIDI grant number 917.164.455 (MAS)
Chan Zuckerberg Initiative grant number 2019- 202666 (MH)

Authors' contributions

SL, KTS, DV, and MK implemented the pipeline, analyzed the data, and drafted the manuscript. HW implemented the QTL mapping and permutation tool. LF, MH, MW, HW, and MAS contributed to the project design, results interpretation, and supervision of the work. RO, IB, HG, PH, and MW provided the data. IB, HG, and PH generated the van Blokland data set. All authors reviewed the manuscript and approved the final manuscript.

Acknowledgements

We are very grateful to all the volunteers who participated in this study. We thank Kate Mc Intyre for editing the manuscript systematically and extensively.
We thank Martijn Vochteloo for the BIOS replication data preparation.
We thank the UMCG Genomics Coordination Center, the UMCG Research IT programme, the UG Center for Information Technology and their sponsors BBMRI-NL & TarGet for storage and compute infrastructure. We thank the Biobank-Based Integrative Omics Studies (BIOS) Consortium, funded by the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL), a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO) under award number 184.021.007.

Supplementary material

Supplementary material is available at https://doi.org/10.1101/2022.04.20.488925

References

1. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D1005–12.
2. Võsa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021 Sep;53(9):1300–10.
3. van der Wijst MGP, de Vries DH, Brugge H, Westra HJ, Franke L. An integrative approach for building personalized gene regulatory networks for precision medicine. Genome Med. 2018 Dec 19;10(1):96.
4. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene–disease predictions. Brief Bioinform. 2018 Jul 20;19(4):575–92.
5. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008 Dec;9(1):559.
6. Deelen P, van Dam S, Herkert JC, Karjalainen JM, Brugge H, Abbott KM, et al. Improving the diagnostic yield of exome- sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. Nat Commun. 2019 Dec;10(1):2837.
7. Mordelet F, Vert JP. SIRENE: supervised inference of regulatory networks. Bioinformatics. 2008 Aug 15;24(16):i76–82.
8. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2017 Jan;49(1):139–45.
9. Kim-Hellmuth S, Aguet F, Oliva M, Muñoz-Aguirre M, Kasela S, Wucher V, et al. Cell type-specific genetic regulation of gene expression across human tissues. Science. 2020 Sep 11;369(6509):eaaz8528.
10. van der Wijst MGP, Brugge H, de Vries DH, Deelen P, Swertz MA, LifeLines Cohort Study, et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat Genet. 2018 Apr;50(4):493–7.
11. Oelen R, de Vries DH, Brugge H, Gordon G, Vochteloo M, BIOS Consortium, et al. Single-cell RNA-sequencing reveals widespread personalized, context-specific gene expression regulation in immune cells [Internet]. Genetics; 2021 Jun [cited 2021 Aug 23]. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.06.04.447088
12. Crow M, Paul A, Ballouz S, Huang ZJ, Gillis J. Exploiting single-cell expression to characterize co-expression replicability. Genome Biol. 2016 Dec;17(1):101.
13. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat Methods. 2017 Jun;14(6):565–71.
14. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014 Jul;11(7):740–2.
15. Skinnider MA, Squair JW, Foster LJ. Evaluating measures of association for single-cell transcriptomics. Nat Methods. 2019 May;16(5):381–6.
16. Quinn TP, Richardson MF, Lovell D, Crowley TM. propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. Sci Rep. 2017 Dec;7 (1):16252.
17. Baran Y, Bercovich A, Sebe-Pedros A, Lubling Y, Giladi A, Chomsky E, et al. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol. 2019 Dec;20(1):206.
18. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020 Feb;17(2):147–54.

19. van Blokland I, Oelen R, de Groot H, van der Harst P, Franke L, van der Wijst MGP. Single-cell dissection of the immune response after a myocardial infarction. manuscript in preparation.
20. Moerman T, Aibar Santos S, Bravo González-Blas C, Simm J, Moreau Y, Aerts J, et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics. 2019 Jun 1;35(12):2159–61.
21. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martín D, et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell. 2016 Nov;167(5):1398-1414.e24.
22. Ota M, Nagafuchi Y, Hatano H, Ishigaki K, Terao C, Takeshima Y, et al. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. Cell. 2021 May;184(11):3006-3021.e17.
23. Gate RE, Kim MC, Lu A, Lee D, Shifrut E, Subramaniam M, et al. Mapping gene regulatory networks of primary CD4 + T cells using single-cell genomics and genome engineering [Internet]. Genomics; 2019 Jun [cited 2021 Dec 22]. Available from: http://biorxiv.org/lookup/doi/10.1101/678060
24. propr: vignettes/b_visualization.Rmd [Internet]. [cited 2022 Apr 13]. Available from: https://rdrr.io/cran/propr/f/vignettes/b_visualization.Rmd
25. Cannoodt R, Saelens W, Sichien D, Tavernier S, Janssens S, Guilliams M, et al. SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development [Internet]. Bioinformatics; 2016 Oct [cited 2021 Dec 16]. Available from: http://biorxiv.org/lookup/doi/10.1101/079509
26. Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. Nat Biotechnol. 2020 Dec;38(12):1408–14.
27. Papalexi E, Mimitou EP, Butler AW, Foster S, Bracken B, Mauck WM, et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. Nat Genet. 2021 Mar;53(3):322–31.
28. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019 Jan 8;47(D1):D607–13.
29. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics. 2016 May 15;32(10):1479–85.
30. Klein N de, Tsai EA, Vochteloo M, Baird D, Huang Y, Chen CY, et al. Brain expression quantitative trait locus and network analysis reveals downstream effects and putative drivers for brain-related diseases [Internet]. bioRxiv; 2021 [cited 2022 Feb 25]. p. 2021.03.01.433439. Available from: https://www.biorxiv.org/content/10.1101/2021.03.01.433439v2
31. Qi T, Wu Y, Zeng J, Zhang F, Xue A, Jiang L, et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. Nat Commun. 2018 Jun 11;9(1):2282.
32. Aguirre-Gamboa R, de Klein N, di Tommaso J, Claringbould A, van der Wijst MG, de Vries D, et al. Deconvolution of bulk blood eQTL effects into immune cell subpopulations. BMC Bioinformatics. 2020 Dec;21(1):243.
33. Hammal F, de Langen P, Bergon A, Lopez F, Ballester B. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. Nucleic Acids Res. 2022 Jan 7;50(D1):D316–25.
34. Imbratta C, Hussein H, Andris F, Verdeil G. c-MAF, a Swiss Army Knife for Tolerance in Lymphocytes. Front Immunol. 2020 Feb 14;11:206.
35. Šulić S, Panić L, Barkić M, Merćep M, Uzelac M, Volarević S. Inactivation of *S6* ribosomal protein gene in T lymphocytes activates a p53-dependent checkpoint response. Genes Dev. 2005 Dec 15;19(24):3070–82.
36. Chen C, Peng J, Ma S, Ding Y, Huang T, Zhao S, et al. Ribosomal protein S26 serves as a checkpoint of T-cell survival and homeostasis in a p53-dependent manner. Cell Mol Immunol.

2021 Jul;18(7):1844–6.

37. Kasela S, Kisand K, Tserel L, Kaleviste E, Remm A, Fischer K, et al. Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+ versus CD8+ T cells. Lappalainen T, editor. PLOS Genet. 2017 Mar 1;13(3):e1006643.
38. Nickles D, Chen HP, Li MM, Khankhanian P, Madireddy L, Caillier SJ, et al. Blood RNA profiling in a large cohort of multiple sclerosis patients and healthy controls. Hum Mol Genet. 2013 Oct 15;22(20):4194–205.
39. La Starza S, Ferraldeschi M, Buscarinu MC, Romano S, Fornasiero A, Mechelli R, et al. Genome-Wide Multiple Sclerosis Association Data and Coagulation. Front Neurol. 2019 Feb 14;10:95.
40. Amara U, Rittirsch D, Flierl M, Bruckner U, Klos A, Gebhard F, et al. Interaction between the coagulation and complement system. Adv Exp Med Biol. 2008;632:71–9.
41. Lubbers R, van Essen MF, van Kooten C, Trouw LA. Production of complement components by cells of the immune system. Clin Exp Immunol. 2017 Apr 6;188(2):183–94.
42. Condamine T, Le Texier L, Howie D, Lavault A, Hill M, Halary F, et al. Tmem176B and Tmem176A are associated with the immature state of dendritic cells. J Leukoc Biol. 2010 Sep;88(3):507–15.
43. Li K, Sacks SH, Zhou W. The relative importance of local and systemic complement production in ischaemia, transplantation and other pathologies. Mol Immunol. 2007 Sep 1;44(16):3866–74.
44. Dixon KO, O'Flynn J, Klar-Mohamad N, Daha MR, Kooten C van. Properdin and factor H production by human dendritic cells modulates their T-cell stimulatory capacity and is regulated by IFN-γ. Eur J Immunol. 2017 Mar;47(3):470.
45. Gonsky R, Fleshner P, Deem RL, Biener-Ramanujan E, Li D, Potdar AA, et al. Association of Ribonuclease T2 Gene Polymorphisms With Decreased Expression and Clinical Characteristics of Severity in Crohn's Disease. Gastroenterology. 2017 Jul;153(1):219–32.
46. Dotan I, Allez M, Danese S, Keir M, Tole S, McBride J. The role of integrins in the pathogenesis of inflammatory bowel disease: Approved and investigational anti-integrin therapies. Med Res Rev. 2020 Jan;40(1):245–62.
47. Cai H, Chen J, Liu J, Zeng M, Ming F, Lu Z, et al. CRIP1, a novel immune-related protein, activated by Enterococcus faecalis in porcine gastrointestinal epithelial cells. Gene. 2017 Jan 20;598:84–96.
48. Wang J, Xia S, Arand B, Zhu H, Machiraju R, Huang K, et al. Single-Cell Co-expression Analysis Reveals Distinct Functional Modules, Co-regulation Mechanisms and Clinical Outcomes. Zhou XJ, editor. PLOS Comput Biol. 2016 Apr 21;12(4):e1004892.
49. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease [Internet]. [cited 2022 Apr 20]. Available from: https://www.science.org/doi/10.1126/science.abf3041
50. Perez RK, Gordon MG, Subramaniam M, Kim MC, Hartoularos GC, Targ S, et al. Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. Science. 376(6589):eabf1970.
51. van der Wijst M, de Vries D, Groot H, Trynka G, Hon C, Bonder M, et al. The single-cell eQTLGen consortium. eLife. 2020 Mar 9;9:e52155.
52. Polychronakos C, Li Q. Understanding type 1 diabetes through genetics: advances and prospects. Nat Rev Genet. 2011 Nov;12(11):781–92.
53. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017 Sep;14(9):865–8.
54. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, et al. Multiplexed quantification of proteins and transcripts in single cells. Nat Biotechnol. 2017 Oct;35(10):936–9.
55. Frei AP, Bava FA, Zunder ER, Hsieh EWY, Chen SY, Nolan GP, et al. Highly multiplexed simultaneous detection of RNAs and proteins in single cells. Nat Methods. 2016 Mar;13(3):269–75.

56. NEAT-seq: Simultaneous profiling of intra-nuclear proteins, chromatin accessibility, and gene expression in single cells | bioRxiv [Internet]. [cited 2022 Apr 20]. Available from: https://www.biorxiv.org/content/10.1101/2021.07.29.454078v1.full

57. Joint single-cell measurements of nuclear proteins and RNA in vivo - PubMed [Internet]. [cited 2022 Apr 20]. Available from: https://pubmed.ncbi.nlm.nih.gov/34608310/

58. Ma A, McDermaid A, Xu J, Chang Y, Ma Q. Integrative Methods and Practical Challenges for Single-Cell Multi-omics. Trends Biotechnol. 2020 Sep 1;38(9):1007–22.

59. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun;184(13):3573-3587.e29.

60. Quinn TP, Richardson MF, Lovell D, Crowley TM. propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. Sci Rep. 2017 Nov 24;7(1):16252.

61. Schmid KT, Höllbacher B, Cruceanu C, Böttcher A, Lickert H, Binder EB, et al. scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies. Nat Commun. 2021 Dec;12(1):6625.

62. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013 Oct;45(10):1238–43.

63. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. In 2010.

64. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021 Aug;2(3):100141.

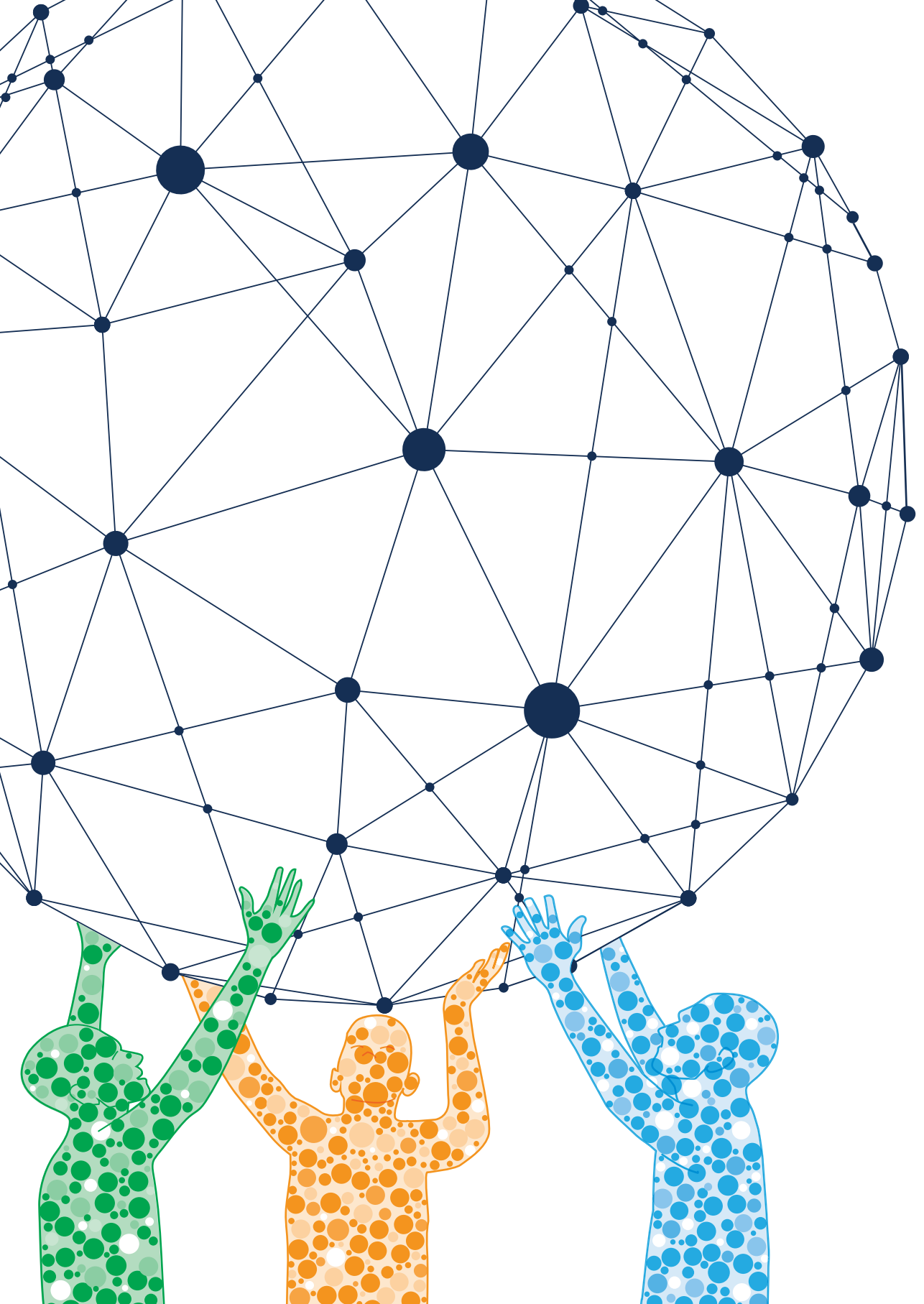
65. Arnold M, Raffler J, Pfeufer A, Suhre K, Kastenmüller G. SNiPA: an interactive, genetic variant-centered annotation browser. Bioinformatics. 2015 Apr 15;31(8):1334–6.

66. Lin SH, Brown DW, Machiela MJ. LDtrait: An Online Tool for Identifying Published Phenotype Associations in Linkage Disequilibrium. Cancer Res. 2020 Aug 14;80(16):3443–6.

67. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. Tang H, editor. PLOS Comput Biol. 2015 Apr 17;11(4):e1004219.

68. Barbeira AN, Bonazzola R, Gamazon ER, Liang Y, Park Y, Kim-Hellmuth S, et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci [Internet]. Genetics; 2019 Oct [cited 2022 Apr 13]. Available from: http://biorxiv.org/lookup/doi/10.1101/814350

69. Barbeira AN, Bonazzola R, Gamazon ER, Liang Y, Park Y, Kim-Hellmuth S, et al. GWAS and GTEx QTL integration [Internet]. Zenodo; 2019 [cited 2022 Apr 13]. Available from: https://zenodo.org/record/3518299



CHAPTER 7

The single-cell eQTLGen Consortium

Van der Wijst MG[1*], de Vries DH[1*], Groot HE[2], Trynka G[3], Hon CC[4], Bonder MJ[5], Stegle O[5], Nawijn MC[6], Idaghdour Y[7], van der Harst P[2], Ye CJ[8], Powell J[9], Theis FJ[10], Mahfouz A[11], Heinig M[12], Franke L[1]

[1] Department of Genetics, Oncode Institute, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
[2] Department of Cardiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
[3] Wellcome Sanger Institute, Open Targets, Wellcome Genome Campus, Hinxton, Cambridge, UK
[4] RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
[5] Genome Biology Unit, European Molecular Biology Laboratory and Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany.
[6] Department of Pathology and Medical Biology, GRIAC Research institute, University of Groningen, University Medical Center Groningen, the Netherlands
[7] Program in Biology, Public Health Research Center, New York University Abu Dhabi, UAE
[8] Institute for Human Genetics, Bakar Computational Health Sciences Institute, Bakar ImmunoX Initiative, Division of Rheumatology, Department of Medicine, Department of Bioengineering and Therapeutic Sciences, Department of Epidemiology and Biostatistics, Chan Zuckerberg Biohub, University of California San Francisco, San Francisco, USA.
[9] Garvan-Weizmann Centre for Cellular Genomics, Garvan Institute, UNSW Cellular Genomics Futures Institute, University of New South Wales, Sydney, Australia
[10] Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany
Department of Mathematics, Technical University of Munich, Garching bei München, Germany
[11] Leiden Computational Biology Center, Leiden University Medical Center, Leiden, the Netherlands
Delft Bioinformatics Lab, Delft University of Technology, Delft, the Netherlands
[12] Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany
Department of Informatics, Technical University of Munich, Garching bei München, Germany
* These authors contributed equally to the manuscript

Published in eLife. https://doi.org/10.7554/eLife.52155

Abstract

In recent years, functional genomics approaches combining genetic information with bulk RNA-sequencing data have identified the downstream expression effects of disease-associated genetic risk factors through so-called expression quantitative trait locus (eQTL) analysis. Single-cell RNA-sequencing creates enormous opportunities for mapping eQTLs across different cell types and in dynamic processes, many of which are obscured when using bulk methods. Rapid increase in throughput and reduction in cost per cell now allow this technology to be applied to large-scale population genetics studies. To fully leverage these emerging data resources, we have founded the single-cell eQTLGen consortium (sc-eQTLGen), aimed at pinpointing the cellular contexts in which disease-causing genetic variants affect gene expression. Ultimately, this information can enable development of personalized medicine. Here, we outline the goals, approach and potential utility of the sc-eQTLGen consortium. We also provide a set of study design considerations for future single-cell eQTL studies.

Interindividual variation needs to be studied at the single-cell level

Genetic variants, most commonly single nucleotide polymorphisms (SNPs), can contribute to disease in a plethora of ways. In monogenic diseases, one single variant is sufficient to result in a disease phenotype. In complex diseases, tens to hundreds of variants each independently contribute to disease risk and an accumulation of risk alleles – often in combination with specific environmental exposures – is required to develop the disease phenotype. The overwhelming evidence showing enrichment of disease-associated variants in regulatory regions suggests that regulation of gene expression is likely a dominant mediator for disease risk. Expression quantitative trait loci (eQTL) analysis is commonly used for linking disease risk-SNPs to downstream expression effects on local (*cis*) or distal (*trans*) genes. Large-scale eQTL efforts such as GTEx[1], PsychENCODE[2], ImmVar[3], BLUEPRINT[4], CAGE[5], and eQTLGen[6] have proven highly valuable to identify downstream transcriptional consequences. All these efforts together lead to ever growing sample sizes that now allow us to start identifying both *cis*- and *trans*-eQTLs.

An important next step is to precisely define the cellular contexts in which disease risk-SNPs affect gene expression levels. This will help to better understand the molecular and cellular mechanisms by which disease risk is conferred and to inform therapeutic strategies. This is particularly important, as recent analyses have shown that many eQTL effects are tissue-[1, 7] and cell type-specific[8, 9]. Additionally, many eQTLs are conditional, and only revealed after specific stimuli that, for example, change the activation or differentiation of specific cell types[3, 10]. Beyond the ability to annotate individual disease associations, cell type-specific eQTLs have been shown to be strongly enriched for heritability across complex traits[11]. Sorting[9, 12] and computational deconvolution[13, 14] of cell types from bulk samples have been used to

uncover context-specificity of eQTLs. However, these methods are biased towards known cell types defined by a limited set of marker genes[15], are of limited use for less abundant cell types, and do not capture any heterogeneity within a sorted population. In contrast, single-cell RNA-sequencing (scRNA-seq) enables the simultaneous and unbiased estimation of cellular composition and cell type-specific gene expression[16], and is particularly well positioned to investigate rare cell types[17]. As opposed to using bulk data, single-cell data allows us to also link genetics to phenomena such as cell-to-cell expression variability[10], cell type heterogeneity[18], and gene regulatory network differences[16]. As such, single-cell analyses in a population-based setting will likely become mainstream in the next few years. However, we envision that most scientific value will be obtained by unifying these efforts. Additionally, to utilize the aforementioned developments in the single-cell field most efficiently and effectively, a coordinated effort from multiple research groups is urgently needed.

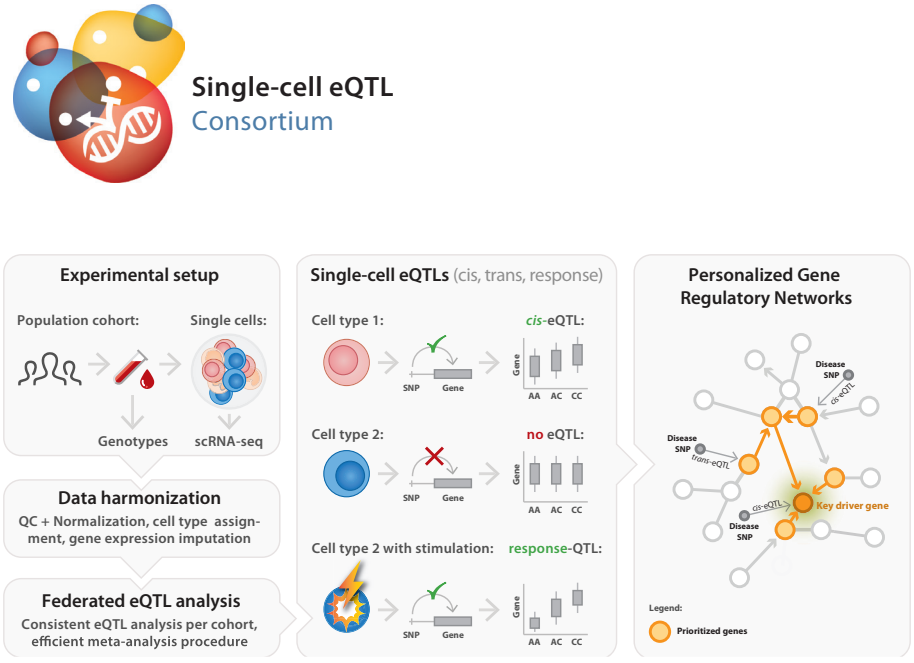


Figure 1. Set-up of the single-cell eQTLGen (sc-eQTLGen) consortium. The sc-eQTLGen consortium combines an individual's genetic information with single-cell RNA expression (scRNA-seq) data of peripheral blood mononuclear cells (PBMCs) in order to identify effects of genetic variation on downstream gene expression levels (eQTLs) and to enable reconstruction of personalized gene regulatory networks. Right panel is adapted from[44].

Here we introduce the single-cell eQTLGen consortium (sc-eQTLGen), a large-scale, international collaborative effort that has been set up to identify the upstream interactors and downstream consequences of disease-related genetic variants in individual immune cell types (https://eqtlgen.org/single-cell.html, Figure 1). In this consortium we will attain a sufficiently large sample size to have the statistical power to unbiasedly identify cell type-specific effects on both local (*cis*) and distal (*trans*) genes. Moreover, we aim to reconstruct context-specific gene regulatory networks (GRNs) by combining single-cell and bulk RNA-seq datasets for increased resolution. We expect a broad impact of the results of sc-eQTLGen that ranges from prioritizing disease-risk genes to predicting drug efficacy through the reconstruction of personalized GRNs.

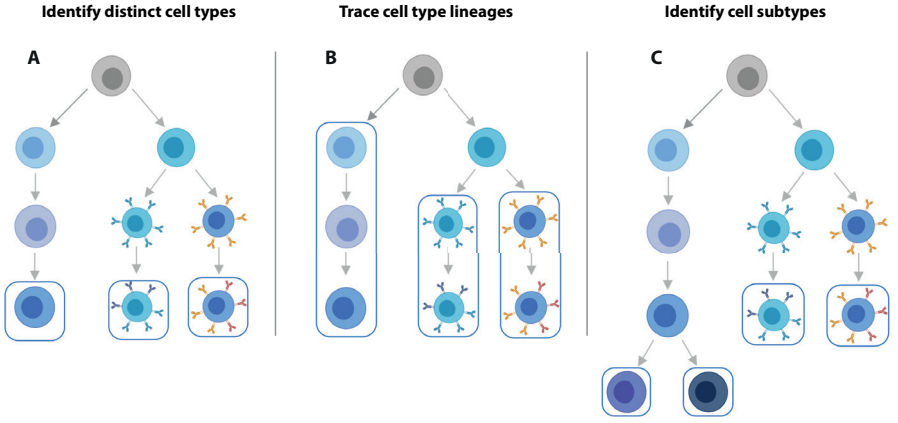


Figure 2. scRNA-seq data offers increased flexibility in the eQTL analysis strategy over bulk RNA-seq data. Using scRNA-seq data for eQTL mapping offers numerous advantages over bulk RNA-seq based approaches, of which the flexibility in analysis strategy is a major one. **(A)** From single cell data, individual cell types can be identified and we can map eQTLs for each of these. **(B)** Alternatively, lineages based on either knowledge of cell developmental lineages or through pseudo-time based approaches can be constructed. By positioning cells across a trajectory dynamic changes in the allelic effects on gene expression levels as a function of trajectory position can be integrated. **(C)** Finally, as the discoveries of new cell subtypes are made or cell type definitions are being refined, the analysis can be revisiting by re-classifying cells and determining how the genetic effects on gene expression vary on these new annotations.

Integration of sc-eQTLGen within the scientific landscape

Large numbers of single cell expression profiles from many individuals are required to reach our goals. The accessibility and clinical relevance of peripheral blood mononuclear cells (PBMCs) have made them the most studied cell types in current population-based scRNA-seq datasets. Therefore, to have such datasets from the same tissue type readily available, we have chosen to focus on PBMCs. It also allows for continuation of the knowledge acquired through the eQTLGen consortium, which performed the largest eQTL meta-analysis to date using whole

blood bulk gene expression data of over 30,000 individuals to reveal the influence of genetics on gene expression[6]. The sc-eQTLGen consortium now allows us to take the next step by systematically assessing the cell types and contexts in which the eQTL effects manifest. Beyond resolving the influence of genetics on individual genes, the consortium will also take advantage of the unique features of scRNA-seq data to learn the directionality of GRNs and uncover how genetics is affecting co-expression relationships[16]. We expect that the infrastructure and best practices developed within sc-eQTLGen can serve as a basis for studying population genetics at the single-cell level in solid tissues in the future.

Other large-scale efforts such as the Human Cell Atlas (HCA)[19] or Lifetime FET flagship consortium (https://lifetime-fetflagship.eu) mainly focus on mapping all cells of the human body or a disease context in a limited number of individuals. The sc-eQTLGen consortium is an important addition to those efforts by putting a unique focus on deciphering the impact of genetic variation on gene expression and its regulation. Different to experimental designs that aim to generate an extensive map on a low number of individuals, we require larger numbers of individuals, whereas the number of cells per individual can be lower. This enables accurately capturing both the genetic variation and cell type heterogeneity. By building on the data and harmonized cell type annotations generated within the HCA, our results will be easily transferable to other datasets as well. We will share best practices of the HCA consortium with regard to data acquisition, analysis and reporting. We also share standards for open science and the infrastructure and legal frameworks for data sharing while accounting for the privacy issues specific to genetic, health record and demographic information.

Single-cell eQTL analysis: the new era of population genetics

The practice of identifying eQTLs is shifting from bulk to single-cell analyses. Considering only its ability to identify eQTLs, scRNA-seq data has a lower statistical power compared to bulk RNA-seq data on the same number of donors[10, 20], likely due to increased sparsity of the single-cell data. Nevertheless, there are several clear benefits of single-cell over bulk expression data for QTL analysis. First, scRNA-seq data enables the simultaneous estimation of the composition and expression profiles of discrete cell populations including cell types and their activation states[16] (Figure 2). Second, scRNA-seq data provides a flexible, unbiased approach that has increased resolution to define cell states along continuous dynamic processes in which the eQTL effects manifest themselves[10]. Third and fourth, single-cell data allows estimating the variability in gene expression across individual cells[21, 22], which could be used to improve mean estimations for eQTL analysis. At the same time, the single-cell nature now also enables us to look at the effect of genetic variation on transcriptomic traits other than average gene expression level, such as dispersion QTLs that alter the variance independently of the mean expression[20] or cell type proportion QTLs[23], providing a new angle on how genetic variation may impact disease pathogenesis. Fifth, the large number of observations

per individual (i.e. cells) enable the generation of personalized co-expression networks, which vastly reduces the number of individuals required to identify SNPs altering co-expression relationships (i.e. co-expression QTLs[16]). Finally, and paradoxically, is the potential benefit of lower experimental costs compared to bulk experiments on sorted cells: such experiments require a library to be generated for each sorted population, whereas a single scRNA-seq library of just one sample contains all this information and can easily be multiplexed across multiple individuals[23].

So far, only a limited number of papers have performed eQTL analysis using scRNA-seq data[10, 16, 20, 23]. In the earliest single-cell eQTL studies, bulk-based eQTL analysis approaches, such as Spearman rank correlation[24, 25] and linear regression[26, 27], were applied to the average expression level of all cells from a particular cell type per individual. However, the underlying assumptions of these bulk-based approaches may not be applicable to scRNA-seq data. Therefore, these bulk-based methods will lose statistical power when applied to scRNA-seq data, because of the inflation of zero values (i.e. sparsity). More recently, single-cell-specific eQTL methods have been developed that, for example, take into account zero-inflated gene expression[20, 28] or take advantage of pseudotime (i.e. statistically inferred time from snapshot data) to increase the resolution by which response-/differentiation-associated eQTLs (dynamic eQTLs, i.e. eQTLs that dynamically change along pseudotime) can be identified[10]. Instead of averaging gene expression levels across all cells from a particular cell type, some of these approaches look at the fraction of zero expression and the non-zero expression separately for each gene[28]. Other approaches take dynamic pseudotime-defined instead of statically-defined cell types into consideration for the eQTL analysis[10]. This latter approach was shown to uncover hundreds of new eQTL variants during iPSC differentiation that had not been detected when static differentiation time points would have been used[10]. In line with this, we expect that some of these methodological advances, as opposed to bulk-based approaches, will further improve the power and resolution of single-cell eQTL analysis. However, there are two initial challenges that need to be carefully addressed for single-cell eQTL mapping: firstly, the normalization of data to remove technical variation in sequencing depth per cell, while avoiding the removal of biological variation; and secondly, the identification or classification of a cell into a cell type or state.

During library preparation and sequencing, technical and stochastic factors will lead to variation in cell-to-cell sequencing depth. However, simply normalizing to equal sequencing depth per cell will remove important biological variation – for example a $CD4^+$ T cell is expected to have lower RNA contents than a plasma B cell. Therefore, we need to employ normalization strategies that can account for traditional batch effects, such as sample run or sequencing lane, while retaining biological differences [29, 30].

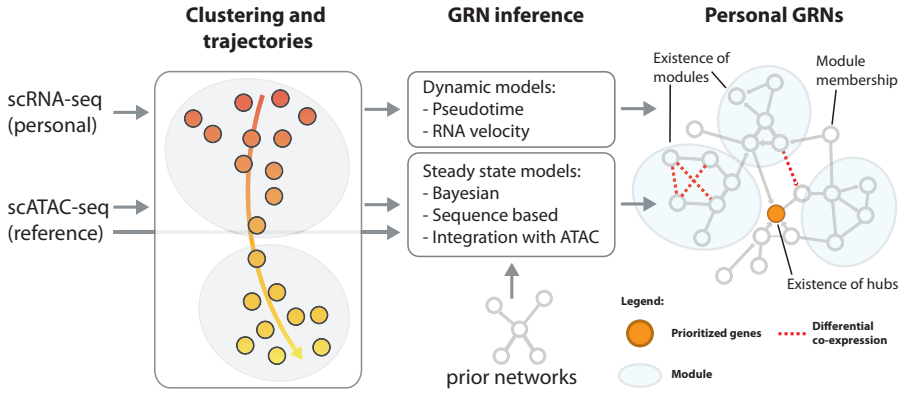


Figure 3. Reconstruction of personalized gene regulatory networks. Individual and cell-type specific scRNA-seq data will be used to construct personalized gene regulatory networks. Some single cell datasets allow for the inference of trajectories, for instance in response to a stimulus. These can be used as input to dynamic models to infer causal (directed) interactions. Steady state datasets, characterized by cell type clusters can be analyzed with models that exploit co-expression, prior networks or cell type-specific reference scATAC-seq datasets in combination with sequence motifs to infer directed transcription factor-target relations. Topological comparison between personalized networks of groups of individuals can reveal coordinated differences, for instance the change of connectivity in densely connected modules, change of connectivity of hub genes or changes of module membership of individual genes.

Once normalized, each cell needs to be accurately annotated into a cell type and/or cell state to maximize the statistical power to detect cell type-specific eQTLs. We encourage the use of individual cell classification approaches, rather than cluster-labeling methods. Clustering approaches are powerful ways of identifying a subpopulation of cells that share similar expression levels. However, while most cells placed in a specific cluster will likely be the same cell type, clusters can also contain alternative cell types. Labeling all cells in a cluster based on a high percentage of the expression of a canonical marker(s) will therefore lead to the incorrect classification of some cells[31]. To acquire a reliable classification model, large scRNA-seq datasets from various contexts are required. Such datasets have been collected within large-scale efforts such as our consortium and the HCA. We expect these will help to develop a gold standard classification model that can classify each cell independently. This will ensure a higher accuracy in cell labeling and thus will maximize power to detect cell type-specific effects.

After solving these challenges, eQTLs can be mapped by either averaging the normalized expression levels on a per gene, per cell type, per individual basis. Alternatively, each cell from an individual can be taken as a repeated measure which can then be used to fit a statistical model to all cells, while including a random effect of the individual.

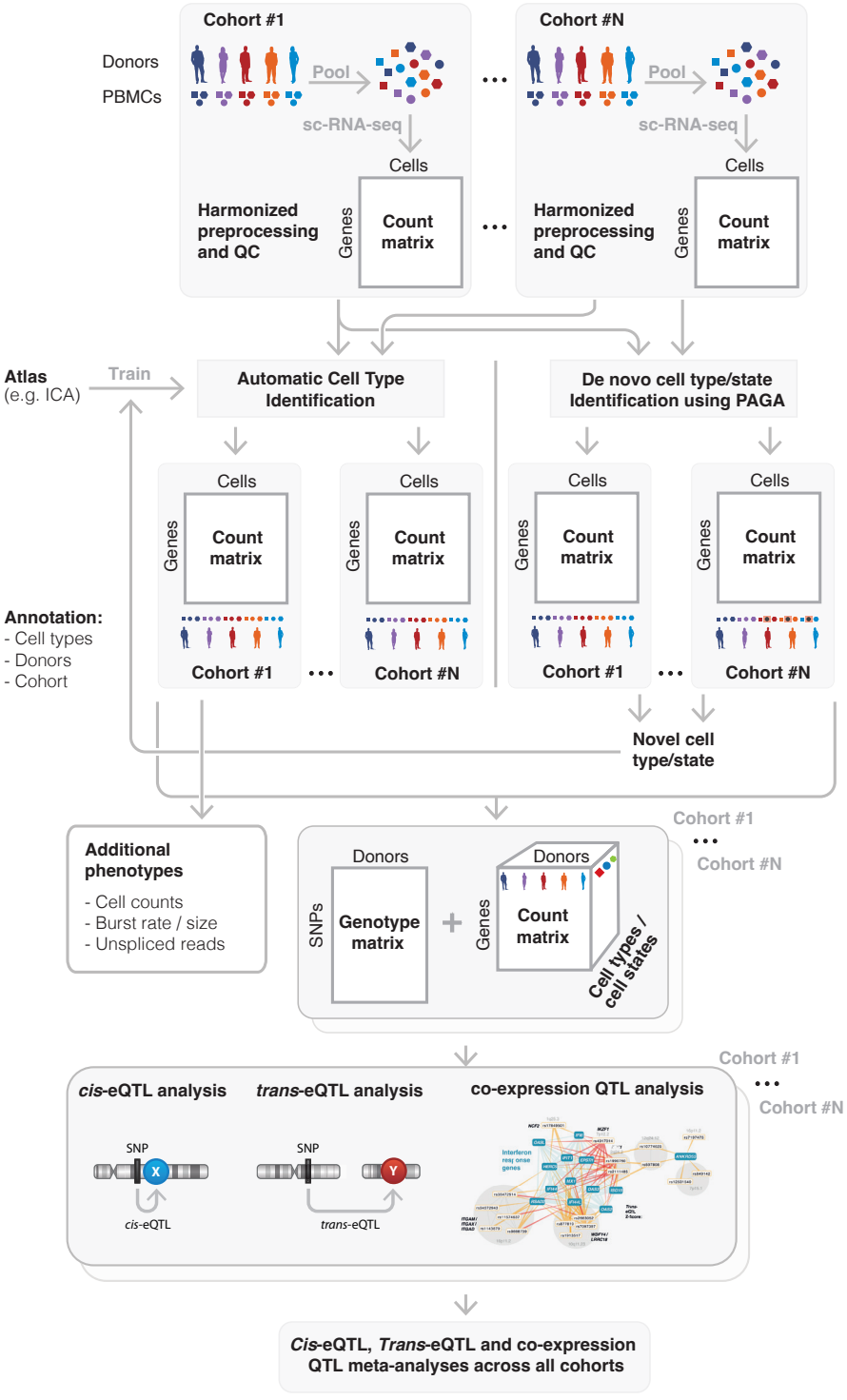

Cohort #1
Donors
PBMCs
Pool
sc-RNA-seq
Cells
Genes
Count matrix
Harmonized preprocessing and QC
Cohort #N
Atlas
(e.g. ICA)
Train
Automatic Cell Type Identification
De novo cell type/state Identification using PAGA
Annotation:
- Cell types
- Donors
- Cohort
Novel cell type/state
Additional phenotypes
- Cell counts
- Burst rate / size
- Unspliced reads
SNPs
Genotype matrix
Donors
Cell types / cell states
cis-eQTL analysis
trans-eQTL analysis
co-expression QTL analysis
SNP
cis-eQTL
trans-eQTL
Cis-eQTL, Trans-eQTL and co-expression QTL meta-analyses across all cohorts

Figure 4. Overview of the sc-eQTLGen proposed federated approach. sc-eQTLGen aims to identify the downstream consequences and upstream interactors of gene expression regulation. To increase the resolution and power of this analysis, datasets of multiple cohorts need to be combined while taking privacy issues into account. This will be done using a federated approach in which we will first harmonize all preprocessing and quality control (QC) steps across cohorts. Subsequently, shared gene expression matrices will be normalized and cell types will be classified based on a trained reference dataset (e.g. Immune Cell Atlas (ICA)). Any cells that cannot be classified using this trained classifier, representing new cell types or previously unknown cell states, can then be manually annotated based on marker genes, and then be used to further train the classifier. Each cohort will then separately perform a *cis*- and *trans*-eQTL and co-expression QTL analysis using their genotype and expression matrix, while using appropriate statistical models to account for effects such as gender, population structure and family-relatedness that can alter the genotype-expression relationship in a cohort-specific manner. The summary statistics will be shared and analyzed in one centralized place. Finally, these results will be used for reconstruction of personalized and context-specific gene regulatory networks. Bottom panel is reproduced from[6].

Instead of using observational studies, eQTLs could also be identified through experimental approaches that use single cells as individual units of experimentation[32]. Sample multiplexing (Box 1) can be combined with experimental perturbation to more efficiently characterize the genetic architecture of gene expression. For example, synthetic genetic perturbations with CRISPR/Cas9 may allow precise control of the expression levels of target gene regulators enabling the validation of detected *trans*-eQTLs and the establishment of upper and lower bounds of *trans* effects. Encoding environmental and genetic perturbations across large population cohorts also enables new designs for studying genetic interactions, both gene-by-environment and gene-by-gene (epistasis). Historically, characterizing these effects in human cells has been plagued by the lack of power and the susceptibility to technical confounding of bulk experiments. Recent work that knocked out ~150 regulators in primary human T cells of nine donors illustrates a proof of concept of how single-cell sequencing across individuals can be combined with experimental perturbations to detect these genetic interactions[33].

Another promising avenue that has become available in recent years to gain increased insights in the link between genetics and disease, is through the use of spatial transcriptomics technologies, including MERFISH, seqFISH+, Slide-seq and 10x Visium[34, 35]. While for PBMCs this approach may not be applicable, in solid tissues and organs this extra layer of spatial information is extremely valuable. For example, it can help to disentangle *trans*-eQTL interactions that are modulated through cell-cell communication (e.g. a SNP affects ligand expression in one cell type, and thereby affects downstream receptor signaling in a second cell type). Despite not having this spatial information available in PBMCs, other approaches that consider receptor-ligand expression pairs do provide insights in potential cell-cell interactions. These approaches have been successfully applied before to uncover how the ligand expression in one cell type can affect the frequency[36] or the downstream signaling[37] of another cell type expressing the corresponding receptor.

Single-cell GRN reconstruction: taking eQTLs one step further

In the case of complex diseases, it is not the disruption of a single gene that causes the disease phenotype. In fact, hundreds of variants can contribute to the disease and converge into just a few key disrupted regulatory pathways[38, 39]. Therefore, for a better disease understanding and to take eQTLs one step further, one has to look beyond the disruption of individual genes and determine how the interaction of genes changes based on cell type[13, 40, 41], environment[42, 43] and an individual's genetic makeup[15, 16]. The sc-eQTLGen consortium will do so by reconstructing personalized, cell type-specific GRNs[44] (Figure 3). The unique features of scRNA-seq data, among which the inference of pseudotime and RNA velocity (i.e. the ratio between spliced and unspliced mRNA that allows prediction of the future state of a cell)[45], enable learning the directionality of network connections[46]. We expect that such personalized GRNs will help explain for example differences in interindividual drug responses, and thereby, will aid in precision medicine in the future.

Reconstruction of GRNs from single cell data (reviewed in [47]) is complicated by the sparsity of the data as a consequence of the stochasticity underlying gene expression[48] and dropouts, i.e. genes that are not detected in some cells as a consequence of technical limitations[49]. This sparsity leads to lower correlation estimates that obscure the identification of true edges in the GRNs. Several solutions have been developed to overcome this problem, including the implementation of prior information[50, 51], gene expression imputation[50, 52] and usage of alternative measurements of correlation[53, 54].

Firstly, prior information encoded in the DNA sequence can be used to overcome these complications[55, 56]. Such priors on regulatory interactions can be derived from, for example, ChIP-seq data[57], ATAC-seq data[58], spatial information[34, 35] or from perturbation experiments[33, 50]. Implementation of such priors was shown to improve bulk GRN reconstruction[58-60], and similarly, it is expected to also improve GRNs reconstructed from single-cell data[50, 51]. However, caution is warranted when using this information, as their effect on GRN reconstruction depends on the quality of these data priors[61, 62] and priors derived from bulk data may not hold true at the single-cell level[63]. Recent technological advances enable studying chromatin accessibility[64, 65] and expression of enhancers RNAs[66, 67] at the single cell level, which will make it possible to implement single-cell derived priors in GRN reconstruction in the future, though these quantifications come with their own limitations and challenges.

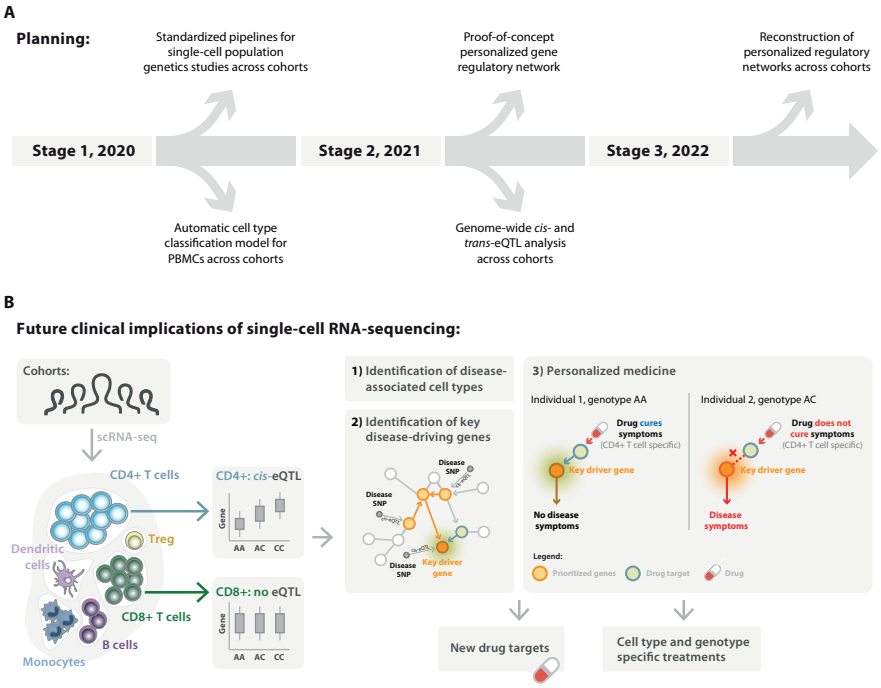


Figure 5. Deliverables of the single-cell eQTLGen consortium in relation to their future clinical implications. A) In the coming 3 years the sc-eQTLGen consortium aims to deliver 1. Standardized pipelines and guidelines for single-cell population genetics studies (2020); 2. Cell type classification models for PBMCs (2020); 3. Summary statistics of *cis*- and *trans*-eQTLs, co-expression QTLs, cell count QTLs and variance QTLs (2021); 4. Reconstruction of personalized, cell type-specific gene regulatory networks (2022). **B)** These efforts of the consortium will lead to the (**1**) identification of disease-associated cell types and (**2**) key disease-driving genes, which together will aid (**3**) the implementation of personalized medicine and the development of new therapeutics that take all this information into account (cell type- and genotype-specific treatments). Panel B2 is adapted from[44].

Secondly, gene expression imputation may be used to restore the underlying correlation structure. However, current gene expression imputation methods become more unreliable as the dropout rates increase[52, 53]. After gene expression imputation, more network edges are identified, but with a higher chance of detecting false positives[50, 52]. Nevertheless, by combining prior information with imputation, GRN reconstruction can be improved both in the bulk[58] and single cell setting[50]. For example, one can replace transcription factor expression with inferred transcription factor activities based on the collective expression patterns of their target genes or take advantage of cross-omics relationships[68].

Finally, alternative correlation measures are being explored to overcome the complications associated with data sparsity, including measures of proportionality[54] and by calculating the correlations on measures other than the normalized expression

counts[53]. For example, Z-scores of the gene expression distributions of highly similar cells have been used to calculate the co-expression relationships. This approach could reveal the true correlation structure that was otherwise hidden by technical artifacts[53]. In addition to these computational tools, technological advances, such as single-cell multi-omics approaches[69, 70] and improved experimental protocols, are expected to alleviate these complications. Moreover, being able to assess multiple layers of information within the same cell, e.g. chromatin accessibility, DNA methylation, gene and protein expression, opens unique opportunities for developing new methodology for GRN reconstruction and validation. Altogether, this will further improve the accuracy of GRNs reconstructed using single-cell data in the future.

The incorporation of dynamic information extracted from time series or pseudotime[71, 72] is another promising avenue to further improve single-cell GRN reconstruction. However, not all datasets are equally well suited to identify temporal trajectories. For example, PBMCs are usually in steady state, and only after stimulation such trajectories would appear.

Summarized, the ideal GRN reconstruction tool can efficiently manage large amounts of single-cell data, incorporate prior information, model non-linear relationships and take dynamic information into account. Early benchmark studies, performed for a limited number of methods on rather small datasets[49] or on simulated data[73] show that current tools usually only work well in specific situations. As such, there is a clear need for the development of all-round tools that work well in every situation.

sc-eQTLGen: a federated single-cell eQTL meta-analysis consortium

Combining data of numerous groups increases the resolution and power by which downstream analyses, such as eQTL identification and personalized GRN reconstruction, can be performed. Ideally, all scRNA-seq datasets should be jointly analyzed at one centralized location. This is particularly helpful to align each group's approaches for preprocessing, quality control (QC) and cell type classification. However, it also eases for instance benchmarking different statistical and computational methods. While this concept of 'bringing the data to the algorithm' is preferred from an analytical perspective, it is usually very difficult to do so when handling privacy-sensitive scRNA-seq and genotype data from human individuals[5, 74].

To overcome this, a federated approach could be used instead, which has the aim of 'bringing the algorithm to the data': each participating cohort will run the analyses themselves (adhering to predefined criteria for preprocessing and QC), and will only share summary statistics that are not privacy-sensitive. Finally, one site takes responsibility for performing the overall meta-analysis using these provided summary statistics. For genome-wide association studies this is a common strategy[75, 76], and for eQTL studies this procedure has been shown to be effective as well[6, 38]. In the following sections we will expand on all steps that have to be taken and what

considerations should be made when conducting such a federated approach for single-cell population genetics studies (Figure 4).

Preprocessing, quality control

The first challenge of federated analyses is the need to have a standardized protocol on how each group should perform their analyses. While such a protocol helps to ensure reproducibility of the data analysis, it requires that all methods and tools used have been rigorously tested before. For scRNA-seq data such protocols are still under development, while in other fields such as that of genome-wide association studies, standardized protocols have been available for years.

Several initiatives are now being undertaken to define best practices in the scRNA-seq field[77]. For example, Tian et al. have compared 3,913 combinations of different scRNA-seq data analysis pipelines to define best practices in the field[78]. Such initiatives could provide the basis for defining the optimal preprocessing, QC and cell type classification steps for our consortium. Additionally, in population-based scRNA-seq studies special attention is required to account for ethnic variation and population stratification (Box 1)[79, 80]. In the event of presence of relatedness in a given cohort, a genetic relatedness matrix will be included in a mixed model to account for the effect, such as in [80, 81]. Adjustments of cohort-level genetic differences will be made in the framework of meta-analysis using summary statistics of the individual cohorts. Once all protocols are established, we can harmonize the preprocessing steps across all groups in the consortium, such as the genome build to use, alignment tool and sample demultiplexing strategy. Due to the cohort-specific characteristics of each dataset, the QC steps cannot be harmonized to the same extent as the preprocessing. Nevertheless, the parameters used for QC can be coordinated across all groups, such as the cutoffs for number of detected genes per cell and mitochondrial fraction. Both the preprocessing and the QC do not require exchanges of data and can be performed independently.

Cell type classification

To facilitate the eQTL meta-analysis, we need to ensure that the cell type annotations are consistent across the different cohorts. To ensure reproducibility of annotations across the different cohorts, we will employ a classification scheme to identify canonical cell types in each cohort separately. Performing cell type labeling using classification models does not only increase the reproducibility, but also constitutes a privacy-safe way of annotating cell types that does not require the sharing of raw or processed gene expression data.

Reference datasets with labeled cells, such as those available from the Immune Cell Atlas (http://immunecellatlas.net/) will be used to train a classifier for automatic cell type classification in each cohort. Our recent comparisons of single cell classification methods showed that simple linear models can yield good results[82, 83]. Despite the

wide availability of reference datasets, we expect that some cohorts will contain novel unknown cell types or states that cannot be identified using the trained classifier. For this, we will use a classification scheme with a rejection option that can flag unknown cells whenever the confidence in cell type assignment is low[82]. The rejected cells can then be manually annotated based on marker gene expression.

To capitalize on the large number of cells and individuals to be profiled in each cohort, an unsupervised clustering approach will be used to analyze the count matrix of each cohort, in parallel to the supervised approach described earlier. This unsupervised approach will serve two purposes: (1) it will help annotate unassigned cells by the classifier, and (2) it will allow refining the resolution at which cells are annotated. Varying levels of granularity of the clustering may reveal cell types, as well as particular cell states or subtypes. This level of granularity required to separate particular cell states is not known a priori. Therefore, novel unbiased approaches such as partition based graph abstraction[84] or metacells, i.e. disjoint, homogenous and highly compact groups of cells that each exhibit only sampling variance[85], provide a framework to reconcile discrete states at different levels of granularity with continuous cell states. These novel annotations can feed back into an iterative online learning approach of supervised classification models, where we could refine cell type prediction models on the available datasets. Once new datasets become available within the consortium these can be annotated based on current models and updated labels can be used in the next round of training. An important consideration here is to preserve the hierarchy of cell annotations, so that if new annotations are added to the classifier, they are subclasses of existing classes. In this way, any downstream analysis based on older annotations remains valid at the older level of granularity. This would yield a coherent approach of labelling over time as the dataset grows. For inference of continuous cell states, we require data integration across multiple centers, as this would ensure the usage of a similar pseudotime scale between individuals. Currently, ordering cells along pseudotime is challenging and best practices are being evaluated[78, 86].

Ultimately, integrating all expression data in a privacy-preserving manner, i.e. as gene expression matrices, will produce a dataset with unprecedented numbers of cells. Such a dataset allows discovery of novel rare cell types or states using clustering approaches as described above. This valuable dataset will then be shared with the community through platforms like the HCA data portal.

eQTL and co-expression QTL analysis

After cell type assignment, annotated gene expression matrices can be returned to each of the cohorts. Each cohort will then map genome-wide cell type-specific *cis*- and *trans*-eQTLs by combining the cell type-specific gene expression matrices with the privacy-sensitive bulk-assessed genotype information using appropriate statistical models. The resulting summary-statistics can then be safely shared without privacy-issues.

One challenge with federated eQTL analyses is that the amount of summary statistics that need to be shared is substantial. For instance, when assuming there are 10 cohorts and for each of these cohorts cells have been assigned to 10 major cell types, a genome-wide *trans*-eQTL analysis (testing the effect of 10,000,000 common SNPs on 20,000 protein coding genes for each of the 10 cell types), where only the correlation for a SNP-gene combination is stored as a 64 bit double value, would require each cohort to exchange 10,000,0000 x 20,000 x 10 x 8 bytes = 146 terabytes of data. To overcome this problem, several frameworks have recently been proposed that take advantage of the fact that these summary statistics matrices reflect the product of a normalized genotype matrix and a normalized gene expression matrix. For instance, the HASE framework[87] recodes genotype and phenotype (i.e. gene expression) data, along with a covariate matrix, in such a way that privacy is ensured and only those matrices, making up only a few gigabytes of data, need to be exchanged.

While protocols exist that explain how eQTL data needs to be processed, harmonized and QCed to perform a federated eQTL analysis (e.g. eQTLGen used the eQTLMappingPipeline[6]), not all steps can be completed immediately: for instance, to identify effects of polygenic risk scores on gene expression levels (ePRS), gene expression data first needs to be corrected for *cis*-eQTL effects[6]. Therefore, the full *cis*-eQTL meta-analysis has to precede calculations of ePRSs. Such iterations take considerable time and are also inconvenient, since it requires a lot of coordination with each of the participating cohorts. For sc-eQTLGen we will first conduct a federated, cell-type specific *cis*- and *trans*-eQTL analysis. After this is completed, we will proceed with a co-expression QTL (co-eQTL) analysis. This analysis will be limited to a predefined set of genes or SNPs, such as the SNP-gene combinations extracted from the identified *cis*- and *trans*-eQTLs or the SNPs located within open chromatin regions that show high interindividual variability, as otherwise trillions of statistical tests have to be conducted (e.g. in [16]: 7,975 variable genes * 7,975 variable genes * 4,027,501 SNPs (MAF ≥ 0.1) = 256,151,580,788,125 tests). Finally, all these results will be combined to reconstruct personalized, cell type-specific GRNs. This multi-step approach will require us to go back and forth between the different cohorts at least twice. Therefore, easy-to-use analysis scripts that can be run efficiently on different high-performance cluster infrastructures are essential to limit the amount of hands-on time.

Gene regulatory network reconstruction

Finally, the scRNA-seq data will be used to reconstruct GRNs. Two strategies will be explored in the context of sc-eQTLGen. The first approach makes use of the large number of bulk RNA-seq datasets for specific cell types that are available in public RNA-seq repositories[88, 89]. Using this publicly available bulk RNA-seq data, reference co-expression networks will be constructed using cell type-specific data. Subsequently, scRNA-seq data will be used to implement directionality and specify

the connections in the network that are affected by specific contexts[44]. The second approach will directly use scRNA-seq data to build cell type-specific GRNs, thereby enabling to immediately take the context-specificity into account. However, the number of genes that can confidently be taken into account by this second approach may be lower due to the sparsity of scRNA-seq data. For both strategies, we will make use of prior information (e.g. ATAC-seq data[58], TF binding information), dynamic information (e.g. information extracted from time series data[71], pseudotime[72] in combination with RNA velocity[45, 46]) and experimental validation (e.g. perturbation experiments[33, 50]) to go from a co-expression to a gene regulatory network. Before implementation, the additional benefit of using such information, extracted from either bulk or single-cell data[50, 51], and using gene expression imputation[50, 52] will be assessed. We expect that the optimal strategy will depend on the amount of available bulk data and prior information that is available for a particular cell type. We will extract this prior information from existing large-scale efforts, such as ENCODE[2] and BLUEPRINT[4]. Additionally, we will make use of single-cell information beyond gene expression levels that is or will be collected within subsets of cohorts within the consortium, including information on chromatin accessibility[64, 65] and expression of enhancers RNAs[66, 67].

Additionally, recent advances have made it possible to measure multi-omics data from the very same cell[70, 92]. However, current approaches are very time- and cost-consuming, and therefore limited to only a few hundred cells. As such, currently, this type of single-cell multi-omics data is of limited use for reconstructing personalized GRNs. Nevertheless, as single-cell multi-omics approaches mature, this combined information of gene expression and additional data layers has the potential to improve GRN inference beyond correlating separate omics layers and allows for direct measurements instead.

Once reconstructed, these GRNs can be used to determine how for example, genetic differences or disease status change the architecture of the network. These networks consist of nodes, representing genes, that are connected through edges, representing the relationship between genes. The context-specific changes in the network can be identified on different levels, such as on the level of individual edges or nodes, topological properties of individual nodes, such as their connectivity (degree) or module membership[90], subnetwork properties, such as the existence and size of modules, or global topological properties, such as degree distribution (Figure 3). Comparing topological features such as node degree to genotypes may identify polymorphisms altering the function of master regulators (highly connected ‘hub’ genes). Interestingly, implementation of network information was shown to be complementary to the identification of eQTLs; using this network information, novel SNPs were identified that could not be identified through single- or multi-tissue eQTL analyses of GTEx[91]. This clearly shows the complementarity of both eQTL and network-based analyses for understanding the impact of genetic variation.

Ultimately, CRISPR perturbations will be coupled to scRNA-seq to validate or improve reconstructed GRNs. To optimize the number of perturbations required for extracting the most useful information from such experiments, an iterative approach will be taken that feeds back the experimental data to the GRN. This approach will make use of active machine learning to select those perturbations that are required to further improve the model[93, 94]. These well validated, personalized and context-specific GRNs will provide us with a better understanding of disease and can be the starting point of applying this knowledge for precision medicine in the future.

Future clinical implications

The goal of the sc-eQTLGen consortium is to identify how genetic and environmental factors interact to affect gene expression in the context of both health and disease. With ever increasing sample sizes, eQTLs have now been detected for almost every gene[1, 6]. It is likely this will become even more pronounced through our initiative in which we will study many different cell types and contexts, and pose the question to what extent extensive eQTL maps will help to better understand disease. For *cis*-eQTLs this will not be straightforward: although it is known that disease-associated SNPs are enriched for showing *cis*-eQTL effects, this enrichment is quite modest[15, 95]. It is therefore not sufficient to simply look up and catalogue which disease-associated SNPs show which cell type- and context-specific eQTL effects, since this can lead to incorrect inferences on the likely causal gene(s) per locus[96]. To partly overcome this, several colocalization and mendelian randomization approaches have been published that help to better infer likely causal genes[97-100]. Once these methods are able to account for multiple cell type- and context-dependent, causal regulatory variants per locus, we expect increased statistical power to prioritize causal genes. Additionally, we envision that such methods in conjunction with our cell type- and context-specific eQTL maps will help to determine which genetic variants have pleiotropic effects, affecting the expression levels of several genes in multiple cell types and conditions.

Nevertheless, we expect that most statistical power to pinpoint causal genes will be gained through the other goals of the sc-eQTLGen consortium: the reconstruction of cell type-specific gene regulatory networks (expected by the end of 2022), the mapping of cell type-specific *trans*-eQTLs and co-expression QTLs (expected by the end of 2021). These efforts will enable us to ascertain how the prioritized *cis*-eQTL genes (expected by the end of 2021) work together. Moreover, it permits us to study the effect of all disease-associated SNPs of a particular disease in a gene network structure, which helps prioritizing the key disrupted genes and pathways in that disease. In line with recent findings of the eQTLGen consortium that applied eQTL analysis in 31,684 bulk samples, we expect that the (majority of) causal genes within disease-associated loci will converge onto only a few key pathways per disease. One strategy to identify those key driver genes is to consider all associated variants

for a specific disease jointly, and ascertain whether most of these variants show (small) downstream effects on an overlapping set of downstream genes. We recently showed proof-of-concept in eQTLGen that this holds for independent systemic lupus erythematosus (SLE)-associated SNPs: many of these variants show downstream *trans*-eQTL effects on genes involved in the type I interferon pathway[6], indicating an important role for this pathway in SLE development. Recently, success has been reported of a type I interferon-targeted therapy in SLE patients[101], highlighting the value of using *trans*-eQTLs for identifying key genes and pathways that are amenable for pharmaceutical intervention. We expect that our single-cell eQTL initiative will aid such analyses substantially: by performing large-scale eQTL mapping in specific cell types that are in a specific cellular state or are exposed to a particular stimulus, we will be able to more accurately determine where and when these downstream effects manifest. Moreover, single-cell studies will also help to overcome the problem associated with cell type composition differences across individuals in bulk-based eQTL studies: many variants exist that affect the proportion of specific cell types that for instance circulate in blood[102]. If this is not fully accounted for, trans-eQTLs will be identified in genes that are specifically expressed in such cell types in bulk analyses. Single-cell studies allow us to distinguish between effects of genetic variants on cell type composition and effects on intracellular gene expression levels. Therefore, we expect scRNA-seq data will be vital to gain insight into the downstream consequences of disease-associated genetic variants, and to identify the key pathways and genes that drive disease.

Altogether, we expect these approaches will provide us with the information required to reveal new targets for disease prevention and treatment (Figure 5). For example, a novel subset of tissue-resident memory T cells has recently been identified in the setting of asthma using scRNA-seq[103]. This study also showed that mostly T helper 2 cells are dominating the cell-cell interactions in the asthmatic airway wall, whereas in healthy controls mostly epithelial and mesenchymal cell types are communicating with each other. Integration of the gene expression of this asthma-associated cell type with asthma-associated genetic risk variants would further increase our understanding of the disease and such knowledge would greatly accelerate the development of personalized/precision treatments in the future. It is this information about how genes interact differently between individuals as a function of their genetic predisposition that will be obtained through the results of our consortium (Figure 5). One of the major benefits of such personalized treatments is in prescribing the correct drug based on the individual (mechanism that underlies) susceptibility to disease. Currently only between 4% and 25% of the people respond to commonly prescribed drugs[104], showing the need to better predict drug responsiveness and thereby avoid unnecessary exposure to side-effects.

This high interindividual variability in drug response is a consequence of genetic and environmental exposure differences between individuals, which can result in differences in drug metabolism, absorption and excretion (pharmacodynamics)[105].

For example, a variant in the *CYP2C19* gene changes the response to the anti-blood clotting drug clopidogrel. The *CYP2C19* gene encodes for an enzyme in the bioactivation of the drug. *CYP2C19* poor metabolizers were shown to exhibit higher cardiovascular event rates after acute coronary syndrome, or percutaneous coronary intervention, as compared to patients with normal *CYP2C19* function[106].

While previous efforts have mainly focused on pharmacodynamic variation, recent single-cell analyses have revealed that gene-gene interactions can also be changed by genetic[16] and environmental variation[10]. For example, two closely related SNPs (linkage disequilibrium $R^2 = 0.92$) affected both gene-gene interactions (*RPS26* and *RPL21*)[16] and gene-environment interactions (*RPS26* and the respiratory status of the cell)[10]. This shows that gene regulatory network changes may underlie part of the interindividual variation in drug responsiveness. However, such effects have never been studied in detail before and the extent to which such interactions affect drug responsiveness are unknown. The sc-eQTLGen consortium is able to study both how gene-gene interactions and gene-environment interactions are affected by genetic variation, giving insight into where and when they occur. Importantly, the applied methodologies will be easily transferable to single-cell data that is collected in other cell types and disease context through other large-scale efforts[19] (https://lifetime-fetflagship.eu). Moreover, several partners within our consortium have generated scRNA-seq data in cohorts with extensive information on individuals' health records and drug usage (e.g. the Lifelines Deep cohort[107] and the OneK1K cohort). With such information, we will be able to validate the link between changes in the gene regulatory network and the drug responsiveness of an individual. This allows us to determine the predictive value of gene networks for determining responsiveness of specific drugs and the applicability of such networks in precision medicine.

As such, the sc-eQTLGen consortium will not only increase our basic knowledge about the contribution of genetics in gene expression and its regulation, but will also be a valuable resource for drug target identification and validation. To increase the impact of this work, all code, guidelines and summary statistics (including all non-significant results) will become freely available to the community through the sc-eQTLGen website (https://eqtlgen.org/single-cell.html). For any additional information, please visit the contact page (https://eqtlgen.org/contact.html).

Funding

Currently, each group within the sc-eQTLGen consortium is separately funded through their own funding. Based on data generated in the pilot study that will be conducted in 2020, the consortium aims to apply for consortium-wide funding through initiatives like the Chan Zuckerberg Initiative, national (e.g. NIH) or transnational (e.g. H2020) funding. M.W. and L.F. are supported by grants from the Dutch Research Council (NWO-Veni 192.029 to M.W., ZonMW-VIDI 917.14.374 to L.F.), L.F. is supported by a European Research Council Starting Grant (Immrisk

637640 to L.F.), L.F. is supported by the Oncode institute, and J.P. is supported by the National health and Medical Research Council Investigator grant (1175781). The funding bodies did not have any role in the content or writing of this manuscript.

References

1. GTEx Consortium *et al*. Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).
2. Wang, D. *et al*. Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, 10.1126/science.aat8464 (2018).
3. Ye, C. J. *et al*. Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* **345**, 1254665 (2014).
4. Chen, L. *et al*. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e24 (2016).
5. Lloyd-Jones, L. R. *et al*. The Genetic Architecture of Gene Expression in Peripheral Blood. *Am. J. Hum. Genet.* **100**, 228-237 (2017).
6. Võsa, U. *et al*. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv*, 447367 (2018).
7. Fu, J. *et al*. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.* **8**, e1002431 (2012).
8. Brown, C. D., Mangravite, L. M. & Engelhardt, B. E. Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genet.* **9**, e1003649 (2013).
9. Fairfax, B. P. *et al*. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502-510 (2012).
10. Cuomo, A. S. *et al*. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *bioRxiv*, 630996 (2019).
11. Hormozdiari, F. *et al*. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041-1047 (2018).
12. Ishigaki, K. *et al*. Polygenic burdens on cell-specific pathways underlie the risk of rheumatoid arthritis. *Nat. Genet.* (2017).
13. Westra, H. *et al*. Cell Specific eQTL Analysis without Sorting Cells. *PLOS Genetics* **11**, e1005223 (2015).
14. Venet, D., Pecasse, F., Maenhaut, C. & Bersini, H. Separation of samples into their constituents using gene expression data. *Bioinformatics* **17 Suppl 1**, S279-87 (2001).
15. Zhernakova, D. V. *et al*. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139-145 (2017).
16. van der Wijst, M. G. P. *et al*. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50(4)**, 493-497 (2018).
17. Villani, A. C. *et al*. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, 10.1126/science.aah4573 (2017).
18. Donovan, M. K. R., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. A. Cellular deconvolution of GTEx tissues powers eQTL studies to discover thousands of novel disease and cell-type associated regulatory variants. *bioRxiv*, 671040 (2019).
19. Regev, A. *et al*. The Human Cell Atlas. *Elife* **6**, 10.7554/eLife.27041 (2017).
20. Sarkar, A. K. *et al*. Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLoS Genet.* **15**, e1008045 (2019).
21. Brennecke, P. *et al*. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093-1095 (2013).
22. Eling, N., Richard, A. C., Richardson, S., Marioni, J. C. & Vallejos, C. A. Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data. *Cell. Syst.* **7**, 284-294.e12 (2018).

23. Kang, H. M. *et al*. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36(1)**, 89-94 (2017).

24. Heap, G. A. *et al*. Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med. Genomics* **2**, 1-8794-2-1 (2009).

25. Stranger, B. E. *et al*. Population genomics of human gene expression. *Nat. Genet.* **39**, 1217-1224 (2007).

26. Michaelson, J. J., Loguercio, S. & Beyer, A. Detection and interpretation of expression quantitative trait loci (eQTL). *Methods* **48**, 265-276 (2009).

27. Stranger, B. E. *et al*. Genome-wide associations of gene expression variation in humans. *PLoS Genet.* **1**, e78 (2005).

28. Hu, Y. & Zhang, X. SCeQTL: an R package for identifying eQTL from single-cell parallel sequencing data. *bioRxiv*, 499863 (2018).

29. Bacher, R. *et al*. SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods* **14**, 584-586 (2017).

30. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *bioRxiv*, 576827 (2019).

31. Alquicira-Hernández, J., Sathe, A., Ji, H. P., Nguyen, Q. & Powell, J. E. scPred: Cell type prediction at single-cell resolution. *bioRxiv*, 369538 (2018).

32. Gasperini, M. *et al*. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**, 377-390.e19 (2019).

33. Gate, R. E. *et al*. Mapping gene regulatory networks of primary $CD4^+$ T cells using single-cell genomics and genome engineering. *bioRxiv*, 678060 (2019).

34. Burgess, D. J. Spatial transcriptomics coming of age. *Nat. Rev. Genet.* **20**, 317-019-0129-z (2019).

35. Maynard, K. R., Jaffe, A. E. & Martinowich, K. Spatial transcriptomics: putting genome-wide expression on the map. *Neuropsychopharmacology* **45**, 232-233 (2020).

36. Smillie, C. S. *et al*. Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178**, 714-730.e22 (2019).

37. Arneson, D. *et al*. Single cell molecular alterations reveal target cells and pathways of concussive brain injury. *Nat. Commun.* **9**, 3894-018-06222-0 (2018).

38. Westra, H. J. *et al*. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238-1243 (2013).

39. Fagny, M. *et al*. Exploring regulation in tissues with eQTL networks. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E7841-E7850 (2017).

40. GTEx Consortium *et al*. Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).

41. Battle, A. *et al*. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14-24 (2014).

42. Knowles, D. A. *et al*. Allele-specific expression reveals interactions between genetic variation and environment. *Nat. Methods* **14**, 699-702 (2017).

43. Fave, M. J. *et al*. Gene-by-environment interactions in urban populations modulate risk phenotypes. *Nat. Commun.* **9**, 827-018-03202-2 (2018).

44. 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 Med.* **10**, 96-018-0608-4 (2018).

45. La Manno, G. *et al*. RNA velocity of single cells. *Nature* (2018).

46. Qiu, X. *et al*. Towards inferring causal gene regulatory networks from single cell expression Measurements. *bioRxiv*, 426981 (2018).

47. Fiers, M. W. E. J. *et al*. Mapping gene regulatory networks from single-cell omics data. *Brief Funct. Genomics* (2018).

48. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA Synthesis in Mammalian Cells. *PLOS Biology* **4**, e309 (2006).

49. Chen, S. & Mar, J. C. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics* **19**, 232-018-2217-z (2018).

50. Jackson, C. A., Castro, D. M., Saldi, G., Bonneau, R. & Gresham, D. Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *bioRxiv*, 581678 (2019).

51. Aibar, S. *et al*. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083-1086 (2017).

52. Andrews, T. S. & Hemberg, M. False signals induced by single-cell imputation. *F1000Res* **7**, 1740 (2018).

53. Iacono, G., Massoni-Badosa, R. & Heyn, H. Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome Biol.* **20**, 110-019-1713-4 (2019).

54. Skinnider, M. A., Squair, J. W. & Foster, L. J. Evaluating measures of association for single-cell transcriptomics. *Nat. Methods* **16**, 381-386 (2019).

55. Budden, D. M. *et al*. Predicting expression: the complementary power of histone modification and transcription factor binding data. *Epigenetics Chromatin* **7**, 36-8935-7-36. eCollection 2014 (2014).

56. Angermueller, C., Lee, H. J., Reik, W. & Stegle, O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **18**, 67-017-1189-z (2017).

57. Angelini, C. & Costa, V. Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems. *Front. Cell. Dev. Biol.* **2**, 51 (2014).

58. Miraldi, E. R. *et al*. Leveraging chromatin accessibility for transcriptional regulatory network inference in T Helper 17 Cells. *Genome Res.* **29**, 449-463 (2019).

59. Qin, J., Hu, Y., Xu, F., Yalamanchili, H. K. & Wang, J. Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods* **67**, 294-303 (2014).

60. Ghanbari, M., Lasserre, J. & Vingron, M. Reconstruction of gene networks using prior knowledge. *BMC Syst. Biol.* **9**, 84-015-0233-4 (2015).

61. Azizi, E. *et al*. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* **174**, 1293-1308.e36 (2018).

62. Siahpirani, A. F. & Roy, S. A prior-based integrative framework for functional transcriptional regulatory network inference. *Nucleic Acids Res.* **45**, e21 (2017).

63. Simpson, E. H. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society.Series B (Methodological)* **13**, 238-241 (1951).

64. Buenrostro, J. D. *et al*. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490 (2015).

65. Lareau, C. A. *et al*. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916-924 (2019).

66. Hayashi, T. *et al*. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat. Commun.* **9**, 619-018-02866-0 (2018).

67. Kouno, T. *et al*. C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. *Nat. Commun.* **10**, 360-018-08126-5 (2019).

68. Lin, D. *et al*. An integrative imputation method based on multi-omics datasets. *BMC Bioinformatics* **17**, 247-016-1122-6 (2016).

69. Stoeckius, M. *et al*. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865-868 (2017).

70. Cao, J. *et al*. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380-1385 (2018).

71. Moignard, V. *et al*. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.* **33**, 269-276 (2015).

72. Ocone, A., Haghverdi, L., Mueller, N. S. & Theis, F. J. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics* **31**, i89-96 (2015).

73. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *bioRxiv*, 642926 (2019).

74. Lukowski, S. W. *et al*. Genetic correlations reveal the shared genetic architecture of transcription in human peripheral blood. *Nat. Commun.* **8**, 483-017-00473-z (2017).

75. Yengo, L. *et al*. Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641-3649 (2018).

76. Xue, A. *et al*. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* **9**, 2941-018-04951-w (2018).

77. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

78. Tian, L. *et al*. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods* **16**, 479-487 (2019).

79. Price, A. L. *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904-909 (2006).

80. Zeng, B. & Gibson, G. PolyQTL: Bayesian multiple eQTL detection with control for population structure and sample relatedness. *Bioinformatics* **35**, 1061-1063 (2019).

81. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821-824 (2012).

82. Abdelaal, T. *et al*. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**, 194-019-1795-z (2019).

83. Köhler, N. D., Büttner, M. & Theis, F. J. Deep learning does not outperform classical machine learning for cell-type annotation. *bioRxiv*, 653907 (2019).

84. Wolf, F. A. *et al*. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59-019-1663-x (2019).

85. Baran, Y. *et al*. MetaCell: analysis of single cell RNA-seq data using k-NN graph partitions. *bioRxiv*, 437665 (2018).

86. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547-554 (2019).

87. Roshchupkin, G. V. *et al*. HASE: Framework for efficient high-dimensional association analyses. *Sci. Rep.* **6**, 36076 (2016).

88. Silvester, N. *et al*. The European Nucleotide Archive in 2017. *Nucleic Acids Res.* **46**, D36-D40 (2018).

89. Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.* **39**, D19-21 (2011).

90. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559-2105-9-559 (2008).

91. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660 (2015).

92. Clark, S. J. *et al*. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781-018-03149-4 (2018).

93. Sverchkov, Y. & Craven, M. A review of active learning approaches to experimental design for uncovering biological networks. *PLoS Comput. Biol.* **13**, e1005466 (2017).

94. Ud-Dean, S. M. & Gunawan, R. Optimal design of gene knockout experiments for gene regulatory network inference. *Bioinformatics* **32**, 875-883 (2016).

95. Aguet, F. *et al*. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv*, 787903 (2019).

96. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768-769 (2019).

97. Giambartolomei, C. *et al*. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

98. Gusev, A. *et al*. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245-252 (2016).

99. Porcu, E. *et al*. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.* **10**, 3300-019-10936-0 (2019).

100. Mancuso, N. *et al*. Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.* **51**, 675-682 (2019).

101. Kemp, A. Anifrolumab Phase III trial meets primary endpoint in systemic lupus erythematosus., https://www.astrazeneca.com/media-centre/press-releases/2019/anifrolumab-phase-iii-trial-meets-primary-endpoint-in-systemic-lupus-erythematosus-29082019.html (29 August 2019).

102. Astle, W. J. *et al*. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).

103. Vieira Braga, F. A. *et al*. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* **25**, 1153-1163 (2019).

104. Schork, N. J. Personalized medicine: Time for one-person trials. *Nature* **520**, 609-611 (2015).

105. Mukherjee, D. & Topol, E. J. Pharmacogenomics in cardiovascular diseases. *Prog. Cardiovasc. Dis.* **44**, 479-498 (2002).

106. Johnson, M. D. *et al*. Cytokine gene polymorphisms and the outcome of invasive candidiasis: a prospective cohort study. *Clin. Infect. Dis.* **54**, 502-510 (2012).

107. Tigchelaar, E. F. *et al*. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772-2014-006772 (2015).

108. Heaton, H. *et al*. souporcell: Robust clustering of single cell RNAseq by genotype and ambient RNA inference without reference genotypes. *bioRxiv*, 699637 (2019).

109. Bycroft, C. *et al*. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).

110. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499-511 (2010).

111. Xu, J. *et al*. Genotype-free demultiplexing of pooled single-cell RNA-seq. *bioRxiv*, 570614 (2019).

112. McCarthy, S. *et al*. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279-1283 (2016).

113. Gravel, S. *et al*. Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 11983-11988 (2011).

114. Carlson, C. S. *et al*. Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. *PLoS Biol.* **11**, e1001661 (2013).

115. Martin, A. R. *et al*. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635-649 (2017).

116. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26-31 (2019).

117. Wojcik, G. L. *et al*. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514-518 (2019).

118. Hsiao, C. L., Lian, I., Hsieh, A. R. & Fann, C. S. Modeling expression quantitative trait loci in data combining ethnic populations. *BMC Bioinformatics* **11**, 111-2105-11-111 (2010).

119. Mandric, I. *et al*. Optimal design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. *bioRxiv*, 766972 (2019).

Box 1: Guidelines for creating a population-based single-cell cohort

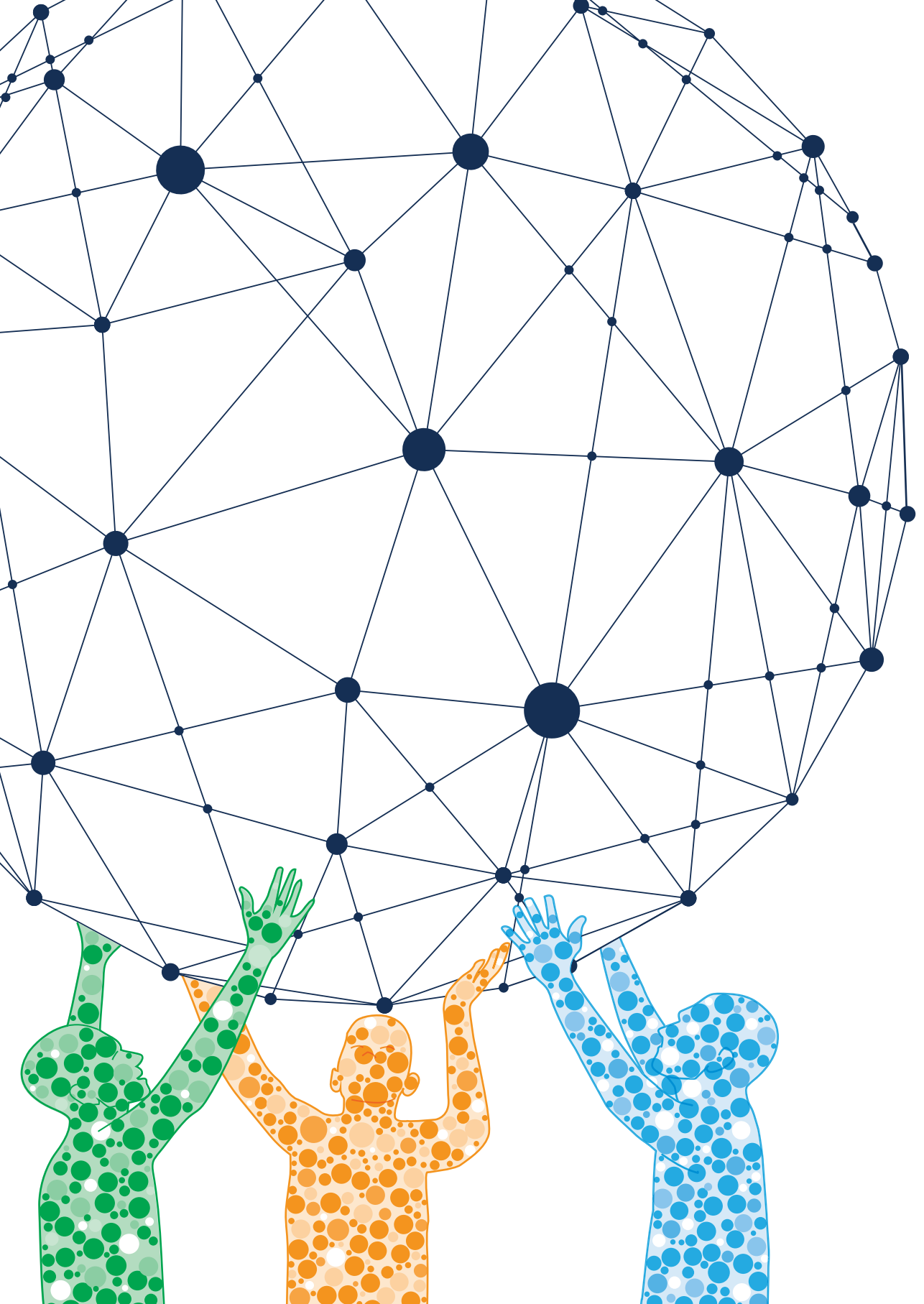
Even though a single-cell eQTL dataset has less discovery power than an equal-sized bulk RNA-seq eQTL dataset (6.9 fold difference based on the lowest correlation that led to the identification of a significant eQTL from single-cell[16] vs bulk RNA-seq data[15]), it does provide insights that cannot easily be extracted from bulk data. For example, single-cell data allows for the unbiased detection of cell type- and context-dependent eQTLs and has more power to detect co-expression QTLs[16]. This makes population-based single-cell datasets a valuable addition to bulk-based datasets for studying the effects of genetic variation on gene expression and its regulation[16, 23]. In comparison to 'standard' single-cell datasets, generating such population-based single-cell datasets require some additional aspects to be taken into account.

First of all, the genetic information that is available for each of the individuals in such cohorts can be used to demultiplex pools of multiple individuals within the same sample. This approach allows to properly randomize experiments, while also significantly reducing cost and confounding effects[23, 108]. This genetic information can either be efficiently generated using genotype arrays[109] in combination with imputation-based approaches[110], or extracted from the scRNA-seq data itself[108, 111]. Within the consortium all reads will be aligned to the **GRCh38/hg38** reference genome and genotypes will be imputed using the Haplotype Reference Consortium reference panel[112]. The basic principle behind genetic multiplexing is that enough transcripts harboring SNPs are expressed and detected in each single cell such that cells can be accurately assigned to the donor of origin. Furthermore, as the number of multiplexed individuals increases, the probability that a droplet harbors multiple cells from different individuals increases, thus allowing the detection of multiplets using genetic information. This enables the overloading of cells into standard droplet-based workflows and overall reduction of cost per cell up to about 10-fold (https://satijalab.org/ costpercell). As the cost of sequencing and the background multiplet rate reduce, the benefits of multiplexing increase. We anticipate that future workflows will allow for even higher throughput.

Secondly, accounting for ethnicity variation and population stratification will be required when single-cell data of diverse populations are being analyzed. It is known that a different genetic architecture exists between different populations. Nevertheless, practical considerations have limited the majority of eQTL studies to cohorts of European origin. As an undesirable consequence of this bias in population representation, certain variants may not have been detected before[113] or the effect sizes and associated polygenic risk scores based on the European population may not be translatable to other populations[114, 115]. Therefore, inclusion of datasets from different ethnic populations will help reduce long-standing disparities in genetic studies and has many analytical advantages[116, 117]. For example, the increased genotype frequency diversity will enhance the range over which gene expression varies, and thereby, will further increase statistical power. To implement multi-population sc-eQTL

analysis, several challenges have to be addressed. Handling data from populations with different levels of population genetic properties such as LD structure, relatedness and multiple genetic origins that result in the presence of genetic covariance remains important and requires appropriate adjustments to avoid spurious signal and to manage the bias in estimating genetic *cis*- and *trans*-effects[80, 118]. This is particularly important when differences in cohort-specific genetic characteristics are enhanced such as when family-based and unrelated cohorts or cohorts of different ancestries are analyzed. Failing to account for these effects affects the accuracy of mapping and results in false positives.

Finally, studying genetic variation at the single-cell level adds some extra requirements for the number of cells per individual and the number of individuals to be included in the study. The number of cells per individual will mainly define for which cell types in a heterogeneous sample such as PBMCs eQTL and co-eQTL analyses can be performed. In contrast, the number of individuals will mainly define the number of genetic variants for which effects on gene expression can be confidently assessed. A recent analysis showed that, with a fixed budget, the optimal power for detecting cell type-specific eQTLs is obtained when the number of reads is spread across many individuals[119]. Even though a lower sequencing depth per cell results in a lower accuracy of estimating cell type-specific gene expression levels, many more individuals and cells per individual can be included for the same budget. As a result, the optimal experimental design with a fixed budget provides up to three times more power than a design based on the recommended sequencing depth of 50,000 reads per cell (for 10X Genomics scRNA-seq). In contrast, for co-eQTL analysis there is a different trade-off between sequencing depth, number of individuals and number of reads per cell; while for eQTL analysis gene expression levels among cells of the same cell type can be averaged, for co-eQTL analysis you cannot as this would prohibit you from calculating a gene-gene correlation per individual. Therefore, for co-eQTLs the sequencing depth will be a major limiting factor that determines the number of genes for which you can confidently calculate gene-gene correlations. Altogether, depending on the goal of your study, the optimal balance between sequencing depth and number of individuals and cells per individual will be different. By the end of 2020, the sc-eQTLGen consortium will provide standardized pipelines and guidelines for single-cell population genetics studies.



CHAPTER 8

Discussion

In this thesis we have described several strategies to study how genetic variation impacts gene expression levels in a cell-type- and context-specific manner. In this **Discussion**, I place the research that I have conducted in the context of past, current and future developments in the field of single-cell technologies and provide an outlook on where this field of research is likely to move in the years to come.

Availability of single-cell technologies

Technological advancements in the single cell field continue to happen at a rapid pace, with each advance allowing new research questions to be answered that help us better understand health and disease. When I started my PhD, single-cell technology was still in its infancy, but the technology and field advanced greatly during my PhD. In this thesis, I described how single-cell RNA sequencing (scRNA-seq) can be applied to look into the context-specificity of gene expression and how it is influenced by genetic variation. In this chapter, I discuss how the new technological developments can alleviate some of the limitations we encountered and how I envision these techniques can be used to continue the research we have initiated in the future.

Understanding single-cell data for context-specific analyses

While single-cell data holds very strong potential for answering exciting research questions that could not be easily studied before, it still has several limitations. As a consequence, analyses require that certain assumptions be made with regards to the underlying data, thereby limiting most currently available computational methods to specific use-cases. Nevertheless, just as tremendous technological advances have been made in generating single-cell data, the tools to analyze them have also continuously improved. Here, I describe the current state of the field and discuss several of the unsolved limitations and (implicit) assumptions.

Assigning cell types

Even though the genetic information in all cells of a single individual is identical (excepting germ cells and the presence of somatic variation), gene expression varies strongly between individual cells, even when considering cells of the same cell type[1]. Current tools mostly require cells to have a distinct cell type classification for analysis, both to compare between cell types and to identify patterns within a cell type[2]. For example, in differential expression analysis, a cell type A is compared to cell type B to identify genes that are significantly differentially expressed between them. Cell types are currently classified either by clustering analysis with manual annotation, or more recently by automatic cell type annotation tools.

While most comparative analyses require distinct cell types, there are ongoing discussions about what defines a cell type[3]. The practical definition of a cell type uses its function, morphology, tissue of origin and gene and protein expression patterns. However, actual cell types cannot be defined as easily because their functions and expression patterns depend on their cell state. As some cells can transition between states with various functions, it is important to take cell state, function and lineage into account to define cell types. Interestingly, Svensson et al. showed that there is a correlation between the number of reported cell types and the number of cells sequenced in a study[4]. This is not due to biology but rather due to the necessity of having sufficient cells of each cell type for analyses. In **Chapters 5** and **6**, we deliberately reported a less detailed cell type classification than we were able to make so that we could retain the intended robustness for the analyses. I expect this problem to be alleviated in the future in two ways. The first is through the development of tools that take more than just categorical cell type classifications as input, using instead a combination of cell type and the position on a cell state trajectory[5], or cell type probabilities[6], or creating overlapping cell neighborhoods[7]. Tools that use trajectories are already being created. For example, Van den Berge et al.[8] developed a tool that uses trajectory information for differential expression analysis and Cuomo et al.[9] used such information for eQTL analysis. The second way that number of cells of a defined type is likely to become less important in the future is by the continued improvements on the number of genes for which gene expression will be measured per captured cell, with costs per cell remaining constant or even decreasing.

Dropouts in single-cell data

Dropouts are instances where there is incomplete capturing of the transcriptome, often resulting in no expression being measured for that gene even though the cell did express it[10], an effect that can be shown when comparing RNA FISH to scRNA-seq data[11]. Differences are observed between different single-cell technologies that affect the number of dropouts and the robustness of the results of analyses. For example, Wang et al. performed a direct comparison between 10X Genomics Chromium and Smart-seq2 in which they found that there are approximately 1.4 times more dropouts in 10X data than in Smart-seq2 data[12]. For the data used throughout this thesis, this means that at least some of the zero values can be explained by technical deficiencies. In another example, Yamawaki et al. did a systematic comparison of different high-throughput scRNA-seq techniques and showed how capture efficiencies, library pool efficiencies and the number of genes detected improved between the 10x v2 and v3 chemistry[13]. I expect similar improvements to continue in the foreseeable future throughout all single-cell technologies, leading to a reduction in the number of dropouts. Improvements are also being made into mapping strategies. Pool et al. showed that by including intronic regions and correcting for overlapping gene regions, they were able to identify expression for 20% more genes[14].

Another frequently used technique to reduce dropouts is gene expression imputation, which will greatly benefit from improvements to capturing efficiencies. Currently, the performance of imputation strategies suffers due to high sparsity, and it has been shown that imputation algorithms work better when the percentage of zeros decreases[15]. While testing different gene expression imputation strategies in Chapter 2, two issues occurred frequently among tested strategies (data not shown). The first was that no gene expression could be imputed because that gene only showed expression in a very small number of cells, making it impossible to identify correlated but more abundantly expressed genes that could be used to make inferences for the low abundance gene. The second phenomenon we observed was over-imputation, where expression found in one cell would be imputed into all other cells. This sometimes resulted in the abundant expression of multiple genes that were markers for different, mutually exclusive, cell types. I expect more advances to come from single-cell technology that allows measurements of various omics data from the same cell. Combining the measurements of these different omics will make it possible to infer a missing value via another correlated measurement, greatly reducing the number of zero values in all of omics layers[16].

Studying the interplay between genetic variation and pathogen infection in blood

Using scRNA-seq to study the genetic effects on gene expression

Since the first genome-wide association studies (GWAS) were completed around 2006, there has been an increased need to systematically study how genetic variants affect biological processes in order to better explain disease pathology in humans[17]. Expression quantitative trait loci (eQTLs) have been used to directly link changes in genotype to differences in gene expression, but they have not been able to identify likely causes for disease for the vast majority of variants[18]. We hypothesized that this was partly due to context-specific effects that were missed in the majority of studies, which often used bulk RNA-seq[19,20]. To address this, we set out to identify context-specific eQTLs using scRNA-seq.

This started in **Chapter 3**, where we performed a cell-type-specific eQTL analysis and introduced the concept of co-expression QTLs. We found that data from just 45 individuals allowed us to identify cell-type-specific eQTLs, but not in a genome-wide fashion due to statistical power limitations. In **Chapter 5**, we attempted this with a much larger dataset, using data from 120 individuals taken at seven timepoints and, while we identified significantly more context-specific eQTLs, we still lacked the statistical power needed to accurately detect these eQTLs genome-wide. We restricted the single-nucleotide polymorphisms we tested to those for which the eQTLGen consortium had already identified a *cis*-eQTL effect and required a minor allele frequency (MAF) above 10% to ensure we had sufficient coverage of those

variants. Both the MAF cut-off and our sample size prevented us from studying the effects of rare variants, even though pathogenic variants are more often rare[21]. With that in mind, I expect much progress to be made in the future as larger sample sizes allow for genome-wide eQTL analyses and less stringent MAF filtering enables the study of rare variants.

Improving context-specific eQTL analysis

One way to resolve these issues is to increase sample sizes, and we therefore initiated the single-cell eQTLGen Consortium, as described in **Chapter 7**. Within this consortium, we have already increased the number of donors to almost 20 times the number in **Chapter 5**. While this may be large enough to do a first genome-wide cis-eQTL analysis, I expect we will need significantly more samples to be able to identify the effects of rare genetic variants. In the eQTLGen Consortium, a genome-wide *cis*-eQTL analysis was run using 31,684 individuals[20]. Within this study, we identified *cis*-eQTLs for most genes, and it seems that an increase in power will only yield additional findings in very lowly expressed genes. However, we have observed that overall statistical power is lower in scRNA-seq data than in bulk RNA-seq for datasets with the same number of individuals, indicating that many more samples will be required to run a comprehensive genome-wide *cis*-eQTL analysis, and even more would be needed for a genome-wide *trans*-eQTL analysis[9,22], though others have reported similar power between single-cell and bulk eQTL analyses[23].

Improvements can also be made on a technical level. Thus far, most single-cell eQTL studies have used tools from bulk studies and used the average expression across all cells of the context of interest for an individual[24,25]. Now, eQTL mapping tools are being developed that are specially tailored to scRNA-seq data, and these use the information of all cells instead of taking the average[5]. However, it remains to be seen whether such tools can be used in large-scale meta-analyses. Improvements are also being made in other contexts, e.g. the single-cell eQTLGen consortium, that allow for the integration of covariates such as the number of unique molecular identifiers, read depth, number of sequenced cells per individual and biological covariates such as sex, ethnicity and age[23,26].

Finding the right context

In theory, there are an infinite number of potential contexts in which an eQTL might exclusively manifest, but not all of them are likely to be relevant for disease. Finding the right context to study is therefore imperative for understanding disease. While many tissue-specific eQTL effects have been extensively mapped within the Genotype-Tissue Expression (GTEx) project, the project mostly looked at healthy tissues from deceased donors and was unable to assess individual cell types[27].

In **Chapters 4** and **5**, we used ex-vivo pathogen stimulation to study the molecular responses to potentially infectious pathogens in peripheral blood mononuclear cells (PBMCs). By looking at ex-vivo responses to fungal, gram-positive bacterial and gram-negative bacterial, we could investigate a wide range of genes associated to immune responses, although our measurement timepoints meant we were mainly able to investigate the immediate innate response. The extent to which the responses to these three different types of potentially infectious pathogens were similar was quite surprising. We saw a larger difference between timepoints for the same pathogen (3 versus 24 hours) than between pathogens at the same timepoint, although this could be different for other types of infectious agents such as by viruses, intracellular bacteria and parasites[28].

While several pathogen-dependent eQTLs could be identified, their individual importance to the pathogen response remains unclear. While we normally link eQTLs to significantly associated GWAS variants, this was not possible here. Sample sizes of GWASes for infectious diseases are often limited[29]. Generally, a p-value threshold of $5*10^{-8}$ is used to define statistically significant loci[30], but no loci reached this threshold for the infectious diseases associated to the stimuli that we studied, despite earlier evidence of there being a genetic component to the infection susceptibility[31,32]. To overcome this, we looked for several lines of evidence (from differential expression, co-eQTL analysis and experimental follow-up) pointing in the same direction to identify potentially disease-relevant variants and their effects.

Another complication of studying the response of white blood cells on potentially infectious disease agents in a population-based cohort like Lifelines is that an actual infection is difficult to model *ex vivo*. In our stimulation studies, dead pathogens were added to PBMCs, and in much higher concentrations than would occur in nature. This may overstimulate some parts of the immune response, while not triggering other parts at all. In most stimulation experiments, the *ex vivo* experiment deviates even further from an actual infection because specific stimulatory factors (e.g. lipopolysaccharide, various interleukins or TNF[33,34]) were used instead of whole pathogens.

The issue of overstimulation is not present when comparing a diseased population to healthy controls, but the potential problem of sample size remains. One group that looks at a specific disease population is dr. Jimmy Ye's group at UCSF, who performed scRNA-seq on 162 systemic lupus erythematosus (SLE) patients and 99 controls to identify disease-specific eQTLs[35]. While such studies are incredibly valuable for the target disease, it is impossible for every disease to be studied in that detail due to the lack of appropriate disease cohorts and the high costs of generating the data. Furthermore, these studies run into the issue that it is hard to prove causality for the identified effects due to the wide variety of factors that can differ between a disease population and a healthy control population. To reduce costs, it is possible to use large-scale population-based cohorts such as Lifelines[36] and UK Biobank[37] (UKBB). These cohorts not only have genetic and transcriptional data available, they also include a wide variety of other data modalities, as well as extensive medical records. With

all these data layers, these cohorts are also a great resource for replicating findings for a wide variety of studies. Additionally, in exceptional circumstances, such as the Covid-19 pandemic, in which a high percentage of the population become ill, new data can be generated for those individuals who become infected to directly compare the same individuals before and after infection. Beyond such exceptions, the information and opportunities these cohorts provide make it important investments and should be continuously updated with data generated using the latest technologies.

Future perspectives

The recurring theme of this thesis is the need to perform studies in the correct disease-relevant contexts. *Ex vivo* systems with human cells that recapitulate the dynamics, microenvironments and architectures of the native tissues would allow us to study these disease-relevant contexts. To this end, there have been exciting developments over the past decade in the field of stem cell—derived tissues that can be combined with organs-on-a-chip (OoC) technology. OoCs are microfluidic devices in which tissue functions can be closely mimicked. These technologies have already been shown to enable modeling of complex human tissues and organs in a controllable yet highly physiologically relevant system[38]. Using human induced pluripotent stem cells (hiPSCs), potentially any human tissue consisting of multiple cell types can be recapitulated on these chips. Many human tissues have already been successfully reconstituted, including gut[39], lung[40], bone marrow[41], heart[42] and liver[43]. However, a current limitation of OoC technology is the lack of large amounts of well-genotyped hiPSC cells covering a wide range of phenotypes, polygenic risk scores and ethnicities and that it is currently not standardized, with many academic groups making their own low-throughput systems[44]. As the technology matures and standards arise, it is expected that OoCs will have high-throughput systems to model most tissues in a way that changes the way we model disease, develop and test drugs and use precision medicine[45]. I expect that by coupling multiple organ-specific OoCs, we may be able to better mimic the complexity of an organism while still being able to experimentally control for factors of interest. I expect this will eventually allow us to perform context-specific eQTL analyses in a highly controlled fashion, increasing our power to detect such effects and maybe better capturing the *in vivo* context.

To be able to study eQTLs, we need to capture the various genotypes for each variant we want to study in sufficient amounts to make inferences on the effect of individual genetic variants on gene expression. Using CRISPR-Cas genome editing tools, it is becoming increasingly feasible to change single bases, allowing us to make cell lines or OoCs with specifically tailored genotypes recapitulating monogenic disease[46]. Previously, changes were limited to cytosine base editors and adenine base editors, allowing only the four transition mutations (C«T, A«G)[47]. With prime editing (PE), such limitations no longer exist, and even small insertions and deletions can be introduced[48]. Using PE, Anzalone et al.[48] estimated that up to 89% of pathogenic

human variance recorded in the clinical variance database (ClinVar) can be corrected. Both off-target effects and the byproducts of the gene editing were reduced using PE, while retaining or even increasing the efficiency compared to base editors, and this will likely improve further in the near future. This will enable the use of CRISPR-Cas genome editing to specifically assess those variants that are likely disease-causing and identify their effects on transcription within the OoC systems, vastly improving the process of identifying and interpreting the downstream molecular effects of individual genetic variants. This holds most potential for monogenic diseases, but even in complex disease, it can help understand how the individual variants may independently contribute to the disease.

I also expect gene regulatory network (GRN) reconstruction to benefit greatly from coming advancements in single-cell technologies. In **Chapter 6**, we used scRNA-seq data to reconstruct GRNs and look for personalized components, but we were unable to make large directed networks. We showed that gene expression levels strongly affected our ability to identify co-expression and co-expression QTLs (co-eQTLs), limiting us to only reconstructing portions of the networks. Despite the challenges due to sparsity, scRNA-seq does remove the issue of Simpson's paradox, which occurs when using bulk RNA-seq data, showing scRNA-seq's potential and importance in reconstructing GRNs. In **Chapter 2**, we hypothesized how personalized, directed, cell-type-specific GRNs can be used to identify key driver genes of disease. By making small genotypic alterations with CRISPR-Cas genome editing in an OoC, we should see transcriptional changes in a controlled environment that would allow us to more accurately reconstruct such networks and not only identify eQTLs with high precision[49]. We would be able to modify the strongest complex disease-associated variants in these systems, enabling the identification of the most disease-relevant pathways and potential key disease-driving genes that could be used as drug targets. However, we have seen that network structure is not only dependent on cell type and genetic make-up, it can also change upon exposure to environmental stimuli.

When a stimulus is added to an OoC system, the cells' exposures are dependent on the position within the tissue, unlike when a stimulus is added to cells in suspension outside the body. To account for this difference, it is important to know the spatial position of the cell when assessing its transcriptional profile. Using spatial transcriptomics, we are now able to take a slice of a tissue, retaining its spatial structure, and study gene expression[50]. While this technique is not intended for non-solid tissues like circulating immune cells, one can use, for example, bone marrow[41] or lymph node tissue[51] as alternatives to study immune cells. When looking at response to environmental triggers, the spatial information may help identify a gradient of response that helps distinguish between strongly responding, weakly responding and even unexposed cells, revealing which genes are transcribed first upon activation and which follow. Such information can help in adding directionality to the networks as another layer of evidence in addition to eQTLs and RNA velocity, as described in **Chapter 2**. To overcome some of the challenges with RNA velocity

that we identified in **Chapter 6**, it would be possible to use full-length transcriptome sequencing, for instance with Smart-seq3[52], which would also enable us to look into various other factors such as allele-specific expression, splice-QTLs and alternative polyadenylation QTLs.

Together, OoCs, PE, directed GRNs and spatial transcriptomics promise a great future for understanding (context-specific) gene regulation. Many efforts are ongoing to improve these techniques and to provide large resources for the scientific community to use, such as the human Organ and Disease Model Technologies (hDMT) consortium for integrating OoC as disease models[53], the Human Cell Atlas (HCA) to create a complete reference map of human cells[54], the GTEx project to map variation in gene expression levels across individuals in diverse tissues[27] and the eQTLGen and single-cell eQTLGen consortia to create a comprehensive map of eQTLs in human blood[20,55]. As the amount of data continues to increase in both volume and complexity, I expect that research groups will more frequently collaborate in these kinds of large international consortia. With that growth comes an opportunity and a responsibility to include more ethnicities in our research and to allow for more diverse teams to conduct the research. I fully believe that the output of such consortia will eventually lead to the implementation of more personalized healthcare.

Conclusions

The work in this thesis has laid some of the groundwork for doing eQTL analysis and GRN reconstruction in scRNA-seq data and expanded our understanding of context-specific transcriptional variation in blood. We have also provided resources for cell-type-specific eQTLs and stimulation-dependent eQTLs. In addition, we introduced the concept of co-expression QTL mapping and extensively tested strategies to systematically identify them. For GRN reconstruction, we have shared an outline of how we envision GRNs can be built with scRNA-seq data and have done extensive testing on the best methods to identify co-expression and gene clusters. Finally, through the sc-eQTLGen consortium, we have developed the infrastructure for an international, collaborative effort to map context-specific eQTLs. I expect that these and similar efforts will contribute to a better understanding of functional genomics and through that help realize personalized healthcare systems in the future.

References

1. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
2. Wang, T., Li, B., Nelson, C. E. & Nabavi, S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* **20**, 40 (2019).
3. Clevers, H. Cell Systems Voices What Is Your Conceptual Definition of '"Cell Type"' in the Context of a Mature Organism? What Is an Adult Cell Type, Really? (2017). doi:10.1016/j.cels.2017.03.006
4. Svensson, V., da Veiga Beltrame, E. & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database* **2020**, (2020).
5. Nathan, A. *et al.* Modeling memory T cell states at single-cell resolution identifies in vivo state-dependence of eQTLs influencing disease. *bioRxiv* 2021.07.29.454316 (2021). doi:10.1101/2021.07.29.454316
6. Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q. & Powell, J. E. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol. 2019 201* **20**, 1–17 (2019).
7. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol. 2021 402* **40**, 245–253 (2021).
8. Van den Berge, K. *et al.* Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun. 2020 111* **11**, 1–13 (2020).
9. Cuomo, A. S. E. *et al.* Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* **11**, 1–14 (2020).
10. Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nat. Commun. 2020 111* **11**, 1–9 (2020).
11. Huang, M. *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods 2018 157* **15**, 539–542 (2018).
12. Wang, X., He, Y., Zhang, Q., Ren, X. & Zhang, Z. Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2. *Genomics. Proteomics Bioinformatics* (2021). doi:10.1016/J.GPB.2020.02.005
13. Yamawaki, T. M. *et al.* Systematic comparison of high-throughput single-cell RNA-seq methods for immune cell profiling. *BMC Genomics 2021 221* **22**, 1–18 (2021).
14. Pool, A.-H., Poldsam, H., Chen, S., Thomson, M. & Oka, Y. Enhanced recovery of single-cell RNA-sequencing reads for missing gene expression data. doi:10.1101/2022.04.26.489449
15. Steinheuer, L. M., Canzler, S. & Hackermüller, J. Benchmarking scRNA-seq imputation tools with respect to network inference highlights deficits in performance at high levels of sparsity. *bioRxiv* 2021.04.02.438193 (2021). doi:10.1101/2021.04.02.438193
16. Hu, Y. *et al.* Single cell multi-omics technology: Methodology and application. *Front. Cell Dev. Biol.* **6**, 28 (2018).
17. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications* **11**, 1–3 (2020).
18. 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).
19. Strunz, T. *et al.* A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. *Sci. Reports 2018 81* **8**, 1–11 (2018).
20. 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. 2021 539* **53**, 1300–1310 (2021).

21. Claringbould, A., de Klein, N. & Franke, L. The genetic architecture of molecular traits. *Curr. Opin. Syst. Biol.* **1**, 25–31 (2017).
22. Sarkar, A. K. *et al.* Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLOS Genet.* **15**, e1008045 (2019).
23. Nathan, A. *et al.* Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. doi:10.1038/s41586-022-04713-1
24. Van Der Wijst, M. G. P., Brugge, H., De Vries, D. H., Deelen, P. & Swertz, M. A. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).
25. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
26. Cuomo, A. S. E., Alvari, G., Azodi, C. B., McCarthy, D. J. & Bonder, M. J. Optimizing expression quantitative trait locus mapping workflows for single-cell studies. *Genome Biol.* **22**, 1–30 (2021).
27. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nat. 2017 5507675* **550**, 204–213 (2017).
28. Verhoef, J., van Kessel, K. & Snippe, H. Immune Response in Human Pathology: Infections Caused by Bacteria, Viruses, Fungi, and Parasites. (2019). doi:10.1007/978-3-030-10811-3_10
29. Kwok, A. J., Mentzer, A. & Knight, J. C. Host genetics and infectious disease: new tools, insights and translational opportunities. *Nat. Rev. Genet. 2020 223* **22**, 137–153 (2020).
30. Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur. J. Hum. Genet. 2016 248* **24**, 1202–1205 (2016).
31. Smeekens, S. P., Veerdonk, F. L. van de, Kullberg, B. J. & Netea, M. G. Genetic susceptibility to Candida infections. *EMBO Mol. Med.* **5**, 805 (2013).
32. Farhat, M. R. *et al.* GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nat. Commun. 2019 101* **10**, 1–11 (2019).
33. M, S., SL, P., FM, B. & M, F. Cytokine stimulation of T lymphocytes regulates their capacity to induce monocyte production of tumor necrosis factor-alpha, but not interleukin-10: possible relevance to pathophysiology of rheumatoid arthritis. *Eur. J. Immunol.* **27**, 624–632 (1997).
34. Bakker, O. B. *et al.* Integration of multi-omics data and deep phenotyping enables prediction of cytokine responses. *Nat. Immunol. 2018 197* **19**, 776–786 (2018).
35. Perez, R. K. *et al.* Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).
36. Tigchelaar, E. F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).
37. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.* **12**, e1001779 (2015).
38. G, V.-N., K, R.-B. & M, R. Organs-on-a-chip models for biological research. *Cell* **184**, 4597–4611 (2021).
39. Moerkens, R., Mooiweer, J., Withoff, S. & Wijmenga, C. Celiac disease-on-chip: Modeling a multifactorial disease in vitro. *UEG J.* **7**, 467–476 (2019).
40. Huh, D. *et al.* Reconstituting Organ-Level Lung Functions on a Chip. *Science (80-.).* **328**, 1662–1668 (2010).
41. Chou, D. B. *et al.* On-chip recapitulation of clinical bone marrow toxicities and patient-specific pathophysiology. *Nat. Biomed. Eng. 2020 44* **4**, 394–406 (2020).
42. JM, S., CL, M. & M, B. Engineered models of the human heart: Directions and challenges. *Stem cell reports* **16**, 2049–2057 (2021).

43. Kostrzewski, T. *et al.* A Microphysiological System for Studying Nonalcoholic Steatohepatitis. *Hepatol. Commun.* **4**, 77–91 (2020).
44. AR, V. *et al.* Facilitating implementation of organs-on-chips by open platform technology. *Biomicrofluidics* **15**, 051301 (2021).
45. Low, L. A., Mummery, C., Berridge, B. R., Austin, C. P. & Tagle, D. A. Organs-on-chips: into the next decade. doi:10.1038/s41573-020-0079-3
46. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol. 2020 387* **38**, 824–844 (2020).
47. Kantor, A., Mcclements, M. E. & Maclaren, R. E. Molecular Sciences CRISPR-Cas9 DNA Base-Editing and Prime-Editing. doi:10.3390/ijms21176240
48. Anzalone, A. V *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149 (2019).
49. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens HHS Public Access. *Cell* **176**, 377–390 (2019).
50. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, (2021).
51. Hallfors, N. *et al.* Multi-Compartment Lymph-Node-on-a-Chip Enables Measurement of Immune Cell Motility in Response to Drugs. (2021). doi:10.3390/bioengineering8020019
52. Hagemann-Jensen, M. *et al.* Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol. 2020 386* **38**, 708–714 (2020).
53. Mastrangeli, M. & van den Eijnden-van Raaij, J. Organs-on-chip: The way forward. *Stem Cell Reports* **16**, 2037–2043 (2021).
54. Regev, A. *et al.* The human cell atlas. *Elife* **6**, (2017).
55. van der Wijst, M. G. P. *et al.* The single-cell eQTLGen consortium. *Elife* **9**, (2020).

APPENDIX

Summary

The human immune system is a complex system that we still do not fully understand despite its importance and the amount of effort humanity has put into studying it. No two bodies react the exact same way when a bacteria, virus or fungus attacks the body. We know of many different factors that lead to these unique responses, such as the DNA, the type of pathogen that attacks, previous exposure to the pathogen and many other factors. The impact of each of these factors on the overall immune response is extremely complex already, but each of them also impact each of the components that make up our immune system differently. Some factors will affect T cells, but not B cells, others affect both or neither. The impact of all these important factors on the many different cell types involved in our immune systems is yet to be understood. As technology evolves and our knowledge expands, we are able to study new scientific questions and increase our understanding of the human immune system.

One of the largest technological evolutions was regarding DNA sequencing. Scientists started to gather DNA from large groups of people and then studied which positions in the DNA affected the risks to diseases, bodily characteristics and even unexpected things such as your income. This type of study is called a genome-wide association study and with it many association were identified. However, it quickly became clear that these associations needed further investigation before they could help us understand the phenotypes they associated with. One of those types of investigation attempts to learn the relation between the DNA variants and the gene expression levels (eQTL) of the over 20,000 genes scattered across the human genome. However, the relation between these DNA variants and gene expression levels are not necessarily constant throughout our bodies. They can have very different effects in one cell type compared to another. Due to the restraints of technology at the time, it was very difficult to overcome this problem and it wasn't until single-cell RNA sequencing (scRNA-seq) became available and affordable that people could account for this in their study designs.

This thesis uses scRNA-seq to study the relations between genetic variants and gene expression levels in a cell-type-specific manner in an attempt to better understand the human immune system. In chapters 1 and 2, we explain how the scientific field evolved to get to the point where we could do these cell-type-specific analyses and what equipment is needed for it. We share our expectations of how the unique features of scRNA-seq data not only allows increased resolution for eQTL analyses, but also how it can be leveraged to understand the relations between genes to reconstruct cell-type-specific gene expression networks. These gene networks often represent important biological processes that happen within our cells that can be used to help understand disease and develop new drugs.

Chapter 3, is our first and one of the very first single-cell eQTL analyses ever performed and was important to the field to show that scRNA-seq technology could be used to study the effects of genetic variants at never-before-seen resolutions.

Although it was expected, we could finally show that the effects of these genetic variants was different in different cell types and thereby provided extra evidence to the hypothesis that many disease-associated genetic variants likely worked through changing gene expression level and that those genes had not been identified yet because the right context of those effects had not yet been identified. Interestingly, we showed that genetic variants not only change gene expression levels, but can also change how genes co-express (co-eQTL). This means that your genetics can change the gene networks that we use to understand disease and develop drugs with. However, this analysis required more data to be performed systematically in a genome-wide manner, but did inspire the research in the other chapters of this thesis.

In chapter 4, we added an extra level of context to our analyses. We exposed the blood cells we had collected to a *Candida albicans* fungus and then studied its effects on gene expression across cell types. We saw that the exposure to the fungus also changed the expression levels for many genes in a cell-type-specific manner. More importantly, we showed that there was an interaction between the genetic variants and the fungal exposure, where for a gene we found that only people with a specific variant had the desired immune response and others did not. We performed an extra experiment to validate our findings and could show that the genetic variant changed whether monocytes, required to attack the fungus once it is in the bloodstream, would move towards the location of the fungus or not.

Chapter 5 combined the analyses of chapters 3 and 4, but made increased the number of cells and people (and thus the genetic variation) substantially. We also studied the effects of two extra bacterial pathogens in addition to the effects of Candida. For all three of these conditions we also added an extra timepoint, to see how the immune response would change over time. We found that the length of exposure also affected the expression levels and had interaction effects with genetics. Additionally, we saw that the three different pathogens triggered had some unique aspects in the associated immune response, but found that this was not as strong as the effect of the length of exposure. This new dataset also allowed us to study whether co-eQTLs were affected by pathogen exposure and indeed saw that this was the case.

In chapter 6 we leverage this huge dataset to reconstruct gene networks for different cell types and more systematically study the effects of genetics on these networks. Unfortunately, despite the size of this data, we still could not do a full co-eQTL analysis and realise the dream we described in chapter 2. Therefore, we focused on giving the field a guideline on how best to conduct co-expression analyses with scRNA-seq data and ran simulations to show what kind of data would be required to do a full co-eQTL analysis for when the right datasets are generated and such analyses can be performed.

One way to overcome the limitations of analyses that result of having insufficient data, is by combining datasets into a meta-analysis. For chapter 7, we established the single-cell eQTLGen Consortium to bring together all the research groups that are

working with similar datasets as us, so that we can improve and expand the analyses that we can run. This chapter explains the process of how we will combine all this data and how we can ensure the privacy of all research participants, while still allowing for genetic analyses to be performed.

Finally, in chapter 8, I describe how I expect the single-cell field to evolve and how other technologies will be integrated into single-cell analyses to provide further contexts to study how some genetic variants lead to disease and how this information will be used for precision medicine.

Samenvatting

Het menselijk immuunsysteem is een complex systeem dat we nog steeds niet volledig begrijpen ondanks dat het zo belangrijk is en dat het enorm veel bestudeerd is. Elk lichaam reageert anders wanneer deze in aanraking komt met een bacterie, virus of schimmel. We kennen veel verschillende factoren die bijdragen aan het uniek maken van de immuunrespons, zoals het DNA, het type pathogeen, eerdere blootstelling aan pathogenen en nog veel meer. De impact van al deze factoren op de gehele immuunrespons is al enorm complex, maar elk van deze factoren heeft ook nog eens een uniek effect op de losse componenten die samen ons immuunsysteem maken. Zo zullen sommige factoren een effect laten zien in T cellen, maar niet in B cellen, en andere factoren juist in beide en weer andere in geen van beide celtypen. The volledige impact van deze factoren op de verschillende celtypen en de gevolgen daarvoor op het immuunsysteem zijn nog niet bekend. Terwijl technologie verder ontwikkelt en onze kennis steeds dieper wordt zijn wij steeds verder in staat om nieuwe wetenschappelijke vragen te beantwoorden die ons het menselijke immuunsysteem steeds beter leren te begrijpen.

Een van de grootste technologische evoluties was die van het DNA sequencen. Wetenschappers begonnen het DNA van grote groepen mensen te verzamelen en konden zo bestuderen welke posities in het DNA effect hadden of ziekterisico's, lichamelijke eigenschappen en soms zelfs onverwachte dingen zoals je inkomen. Dit type studie heet een genome-wide assocation study en hiermee werden enorm veel associaties gevonden. Het werd echter al snel duidelijk dat dit soort associaties meer onderzoek nodig hadden voordat we deze konden gebruiken om alle fenotypes mee te begrijpen waarmee de varianten geassocieerd werden. Een van de types vervolgonderzoeken bestudeerd de relatie tussen de DNA variant en de genexpressie levels (eQTL) van meer dan 20.000 genen die in het menselijk genoom zitten. Maar ook dit soort analyses hadden limitaties, bijvoorbeeld doordat de effecten van de DNA varianten op de genexpressie niet constant zijn door ons hele lichaam heen. Zo kunnen ze heel anders zijn in het ene celtype dan in het andere. Met de technologie van toen was het enorm lastig om hiervoor te kunnen corrigeren en het werd pas grootschalig mogelijk toen single-cell RNA sequencing (scRNA-seq) beschikbaar en betaalbaar werd.

In mijn proefschrift heb ik scRNA-seq data gebruikt om de relaties tussen genetische varianten en genexpressie levels op een celtype specifieke manier te bestuderen in een poging om het menselijke immuunsysteem beter te leren begrijpen. In hoofdstukken 1 en 2 leg ik uit hoe het wetenschappelijke veld ontwikkeld is om tot het punt te komen waarin wij dit soort celtype specifieke analyses kunnen doen en welke apparatuur ervoor nodig is om dit te kunnen doen. We delen ook onze verwachtingen over hoe de unieke aspecten van scRNA-seq data gebruikt kan worden om niet alleen eQTL analyses met hogere resolutie te kunnen uitvoeren, maar ook om de relaties tussen genen te kunnen bestuderen in celtype specifieke gennetwerken. Dit soort netwerken

representeren meestal biologische processen in de cel en kunnen ons helpen begrijpen hoe een ziekte werkt en daardoor ook helpen nieuwe medicatie te ontwikkelen.

Hoofdstuk 3 is onze eerste en zelfs een van de eerste single-cell eQTL studies die ooit uitgevoerd is en was belangrijk om aan het veld te laten zien dat scRNA-seq data gebruikt kan worden om de effecten van genetische varianten te bestuderen met een veel hogere resolutie dan hiervoor. Zoals ook wel verwacht konden we laten zien dat de effecten van genetische varianten werkelijk verschillend was in verschillende celtypes en gaven hiermee extra bewijs voor de hypothese dat veel ziekmakende genetische varianten werken door genexpressies te veranderen, maar dat deze mogelijk nog niet bestudeerd zijn in de juiste context om dit waar te nemen. Ook lieten we zien dat genetische varianten niet allen genexpressie niveaus veranderen, maar ook dat ze de co-expressie van genen met elkaar kunnen veranderen (co-eQTL). Dit betekend dat de genetica de gennetwerken die wij gebruiken om ziekte mee te begrijpen en medicijnen mee te ontwikkelen kan veranderen. Ondanks deze bevinding te kunnen maken was de dataset te klein om dit op een systematische manier te doen door ons hele genoom en de volledige impact te kunnen begrijpen, maar het stuurde ons in de rest van de hoofdstukken in welke onderzoeksvragen wij wilden stellen.

In hoofdstuk 4 voegden wij een extra level van context toe aan onze analyse. Wij stelden de bloedcellen die wij verzameld hadden bloot aan de schimmel, *Candida albicans*, en bestudeerden de effecten die dit had op genexpressie in de verschillende celtypes. We lieten zien dat de blootstelling aan deze schimmel effect had op de genexpressie, maar belangrijker dan dat, we lieten zien dat er een interactie effect was tussen het hebben van bepaalde genetische varianten en de blootstelling aan de schimmel. Hierbij zagen we dat de mensen met de ene variant een gewenste immuunrespons hadden terwijl die met een andere variant dit niet hadden. We hebben dit opgevolgd met een experiment om te bevestigen dat deze genetische variant invloed uitoefende op het aantrekken van monocyten, die nodig zijn om de schimmel mee weg te halen, naar de locatie van de schimmel in het bloed.

Hoofdstuk 5 combineerde de analyses van hoofdstukken 3 en 4, maar deed dit op een veel grotere hoeveelheid cellen en mensen, waardoor wij ook naar meer genetische varianten konden kijken. Daarbij hebben wij ook gekeken naar twee extra bacteriële pathogenen als aanvulling op *Candida*. Voor elk van de drie condities hebben wij ook een extra tijdspunt toegevoegd om te zien op welke manier de duur van de blootstelling effect had op de immuunrespons. We zagen dat de lengte van de blootstelling een effect had op genexpressie levels en dat het ook een interactie had met de genetica. Verder zagen we ook dat de drie verschillende stimulaties unieke en gedeelde effecten hadden op genexpressie niveau, maar dat dit een kleiner effect bleek te hebben dan de duur van de blootstelling. Deze nieuwe dataset stelde ons ook in staat om te kijken of co-eQTLs veranderd werden door blootstelling aan pathogenen en zagen dat dit inderdaad het geval was.

In hoofdstuk 6 hebben we deze grote dataset gebruikt om celtype specifieke gennetwerken te bouwen en om op een meer systematische manier de effecten

van genetica op deze netwerken te bestuderen. Helaas, ondanks de grootte van deze data, waren we niet in staat om een volle co-eQTL analyse uit te voeren en de verwachtingen van hoofdstuk 2 te realiseren. Daardoor besloten wij te focussen op het maken van een handleiding voor het veld over hoe op de beste manier gennetwerken uit scRNA-seq data te reconstrueren en hebben wij simulaties gedraaid waarmee wij konden laten zien wat nodig is om een volledige co-eQTL analyse uit te voeren voor wanneer de juiste datasets gegenereerd zijn.

Ten slotte leg ik in hoofdstuk 8 uit hoe ik verwacht dat het single-cell veld zich verder gaat ontwikkelen en hoe andere nieuwe technologieën geïntegreerd gaan worden in single-cell analyses om nog diepere context specifieke informatie te kunnen bestuderen, met als doel om te leren hoe genetische varianten tot ziekte leiden en hoe wij deze informatie kunnen gebruiken voor gepersonaliseerde medicatie ontwikkeling.

Dankwoord

Context matters and though my name is on the cover of this thesis, none of it would have been possible without the amazing people around me. In science nothing is ever truly achieved alone. I've had the pleasure to work with people all over the world, with incredibly diverse backgrounds and personalities. The push for collaborative science is such a good development and I can only hope that it will be encouraged even more in the future. Beyond all collaborators, I also have to thank a lot of people who helped me throughout everything else in life. I want to thank everyone who was directly or indirectly involved in making this thesis happen.

Department of Genetics

Beste Lude, ik wil je enorm bedanken voor de kans die je mij gaf na mijn stage om mijn PhD met je te mogen doen. Ik heb enorm veel van je geleerd, niet alleen over hoe ik onderzoek moet doen, maar ook over hoe ik dit kan presenteren zodat de boodschap van het werk duidelijk over komt. Ik heb de vrijheid die ik van je gekregen heb erg gewaardeerd, al was het soms nodig om mij weer even terug op het juiste pad te krijgen. Ik ga alle lessen die ik onder jou geleerd heb voor de rest van mijn leven met mij mee dragen.

Monique, ik denk dat onze werkverdeling heel goed voor ons pasten en ik heb hierdoor vanaf het begin kunnen zien hoe ik het best mijn onderzoek kon opstellen terwijl ik jou een beetje kon helpen met hoe je kunt programmeren. Het verbaasd mij ook niks dat je nu zo snel al zo veel vooruitgang boekt in je carrière en weet zeker dat dit niet gaat stoppen voordat je professor bent. Je gaat het heel ver schoppen en ik kijk uit naar alle coole dingen die je gaat ontdekken!

Ondanks dat je halverwege mijn PhD een nieuwe rol kreeg als rector heb ik enorm veel van je kunnen leren Cisca. Het was enorm inspirerend om te zien hoe jij altijd meteen een onderzoek van top tot teen begreep en kon laten zien waar de zwaktepunten lagen door de perfecte kritische vragen te stellen. Ook buiten het onderzoek om heb ik veel van je mogen leren tijdens onze koffiepauzes en wanneer wij even konden bijkletsen op een terras. Ik hoop ooit half zo goed te worden als jij bent.

To my single-cell buddies, you were always there for me whenever I needed any help, advice or just to have a chat. You were all integral in getting all research done, so that I can have this thesis today. Harm, you were there from the beginning and you were my first partner in crime. I still have many fond memories to the days where we locked ourselves in room 12 and programmed all day every day to meet the deadline. Roy, I don't think I've ever seen someone work as hard as you do. Your dedication to your work is really admirable and has produced many great things for our department. Please don't overdo it, because you are far too good and important to everyone in this

group for us to lose you. Shuang, Katharina and Maryna, though our collaboration was supposed to be a short final project for my thesis, it ended up being a colossal task to do all the work and make sense of any of it. The results often didn't show us what we were expecting or hoping for, but in the end with your perseverance I think we can be proud of what we accomplished. I also want to thank the rest of the single cell team; Marjolein, Irene, Aaron, Werna and all the others scattered across departments in the UMCG that helped elevate everyone's understanding of the ins and outs of doing single cell analyses. I also want to thank everyone in the FrankeSwertz group for their support, advice and all the fun we had together.

I don't even know how to start thanking all the people of the Genetics Department that made my time at the UMCG into what it was. So many of you became close friends that I still hang out with, even though I've gone to work somewhere else. Over the years we have done so many different things; watched sport matches, movies and series, played sports, videogames, board games and dungeons and dragons, sung karaoke, filled in stupid Dutch tax forms, cooked amazing meals and even more cakes, gone camping, went on holidays, done many pubquizes, went to fancy parties and of course had many, many coffee breaks together. These last few years all of you were an incredibly important part of what made me love my job, but also so much more. It was because of this that when Covid came and much of this fell away, that I struggled to find the same enjoyment that I had before. I'm happy to see that despite Covid, many of you are still working hard to organise fun events and group activities. I truly believe that they are part of what made our department great, not only because of the fun environment it creates, but also to bridge the gaps between the many different research groups we have to support collaborative science. I loved to be able to organise the PhD lunches with the amazing PIs and management to help bridge the gap even further, and want to thank everyone for participating and so openly sharing, while we asked all kinds of silly questions. I don't have the right words to thank everyone who was involved in all these things we did together without coming across as being repetitive or insincere, so I'll limit myself to a couple of you who directly helped with the work part of my PhD.

Kate, as the mum of all PhDs you have not only helped everyone to get their theses to not read like it was written by distracted kindergartners, but you were also always there to talk through any of the problems we encountered in our daily PhD lives. You are an amazing person with a great heart and I wish for every PhD student to have someone like you to help them out.

Janneke, niemand begint bij ons op de afdeling zonder eerst een heel warm welkom te krijgen van jou. Jij helpt ons door de chaos heen en geeft ons de sturing die wij nodig hebben om daadwerkelijk te mogen werken aan onze PhDs. Ik hoop dat nu Covid wat minder ons leven lijkt te dicteren, iedereen weer van alle gezelligheid die jij brengt mag gaan genieten.

Nine, jij hebt mij vanaf het moment dat je bij ons op de afdeling kwam gesteund wanneer ik weer eens bezig was met gekke dingen, waarmee ik mijzelf verder kon ontwikkelen op gebieden die niet direct gerelateerd waren aan het doen van een PhD. Ik heb met jou de PhD lunches, de activiteitencommissie, PhD mentoring en nog allemaal losse activiteiten kunnen organiseren. Dit heb ik altijd enorm leuk gevonden en gaf mij een gevoel van betrokkenheid bij de afdeling die heel belangrijk voor mij was. Jouw inzet maakt de afdeling een mooiere plek om te zijn en ik hoop dat je hier nog lang mee door blijft gaan.

Ten slotte moet ik uiteraard Olivier ook nog even bedanken. Niet alleen voor al je hulp bij het doen van het onderzoek, maar vooral ook voor de compleet gestoorde missie die wij aangegaan zijn om de data op ons cluster te organiseren. Wat een werk heeft daar in gezeten om alles te ontrafelen en ik zal altijd trots blijven op hoe wij de herkomst van run01 in de head directory terug hebben weten te vinden. Ik weet niet in hoeverre mensen zich nu houden aan ons datamanagement plan, maar ik vond het prachtig om met jou op te stellen. Jij bent een van de scherpste onderzoekers die ik ken en ik hoop dat jij nog lang de wetenschap versterkt met je talent.

Collaborations

Katharina and Matthias, thanks for helping us make sense of all the networks that we built together. Our work together was a constant battle to understand the data, the algorithms and the biology, but despite that I have really enjoyed working together with you.

Vicky, Vinod and Mihai, your many insights into the finer details of the immune system and the workings of Candida really made us be able to use our data to the fullest. Without your help we never would have found LY86 and made our hypothesis. Thanks for all your help and your patience with me.

It became clear very quickly what potential the single-cell data had and we all knew through the eQTLGen consortium that we could use it so much better if we worked together. I'm grateful to see so many of the greatest mind of my field come together to make this consortium. Thank you all for working with us and I look forward to all the exciting findings in the future.

I also want to thank everyone at Lifelines and especially all the participants of Lifelines. Without your willingness to participate in our research, we would not have been able to perform any of the studies we did.

Friends and family

My dear family, getting to finish my PhD is due to so much more than just the work of these past couple of years. You made it possible for me to even start this incredible journey and I will be forever grateful for that. Your support, advice and love has made me the person I am today and I wouldn't know what to do without it. Even now I still have to ask all kinds of stupid questions and requests to you and you always help me out, no matter what, and are always patient with me. Thank you for all you've done and continue to do for me.

Elke week weer kijk ik uit naar alle gekkigheid die we beleven bij onze DnD sessies. Ik heb met zo veel van jullie mogen genieten van de raarste verhalen en ik ben er nog steeds niet op uitgekeken. Soms als speler, soms als DM, maar altijd met enorm veel plezier. Bedankt voor het samen *nerden* met een berg snacks.

Mijn week is niet compleet zonder een avondje naar de bios te gaan. Eerst moeten we natuurlijk snel slecht eten naar binnen werken en even bijpraten over alles van de afgelopen week en dan door naar de meest random films die er zijn. Ik hoop dat onze traditie nog lang door blijft gaan en dat we nog vele verborgen pareltjes kunnen ontdekken.

De tijd waarin ik elk weekend tot diep in de nacht ging *Magiccen* en Dominion ging spelen zijn voorbij, maar ik blijf het heerlijk vinden om lekker met jullie te discussiëren over de wereld en alle gekkigheden die daarin plaatsvinden. Ik kan mij geen andere app groepsnaam verzinnen die slechter overeen komt met de werkelijkheid dan die van ons en weet zeker dat wij met zijn allen meer bij elkaar *gespammed* hebben dan al mijn andere chats bij elkaar. Jullie zijn allemaal enorm belangrijk geweest voor mij om mij door de moeilijkste periode van mijn leven te krijgen, toen mijn gezondheid ook steeds tegenwerkte en ik ben jullie enorm dankbaar daarvoor. Laten we nog jaren lang door spelen en door *spammen*.

Hoe veel mensen kunnen zeggen dat ze nog om gaan met hun vrienden van de basis en middelbare school? Het is niet gek dat we inmiddels zo verspreid door het land en zelfs de wereld zijn, maar toch doen we ons Kamperen als Beren uitje gewoon elk jaar weer. Tenminste, totdat Douwe en Dorien hunzelf werkelijk een keer kapot gaan maken en ik ze niet weer in elkaar kan plakken. Soms is er iemand weer een tijd weg en komt er weer een nieuw iemand bij, maar de sfeer in de groep blijft altijd hetzelfde en ik voel mij er altijd thuis. Bedankt voor alle mooie feestjes, gala's, kampeertripjes, vakanties, Guitar Hero avonden en alle andere dingen die ik met jullie deel.

Mijn liefde voor bordspelletjes is niet zo geconcentreerd in groepjes als alles hiervoor, maar toch speel ik met veel van jullie regelmatig bordspelletjes. Het maakt niet uit hoe vaak wij samen Spirit Island, Gloomhaven, Great Western Trail, Pandemic

of wat dan ook spelen, het blijft altijd mooi. Bedankt voor alle mooie competitie en samenwerking. Ooit moet ik eens een bordspel gaan maken en dan gebruik ik jullie allemaal als proefkonijn!

Ik heb het geluk gehad veel mensen vrienden te mogen noemen tijdens mijn PhD, maar Joram, Marije, Kevin, Renée en Johanne, jullie blijven toch speciaal. Ik heb zo veel steun van jullie gehad tijdens mijn PhD, dat ik zeker weet dat ik het niet had gekund zonder jullie. Ik kon met jullie altijd praten, even mijn frustraties kwijt en zelfs support sessies krijgen in het ML2 lab voor al mijn problemen. We zijn samen op vakanties geweest, op weekend tripjes, leuke dagjes uit en hebben natuurlijk ook gewoon allemaal gezellige avondjes thuis gehad. Misschien dat wij eigenlijk te oud zijn voor pyjamafeestjes, fristi muffins of dierentuinen, maar ik kan er alleen maar van genieten.

I hope I have captured most of you within the groups above, but some friendships cannot be captured in those groups. I don't know how to thank everyone individually, but know that you've all meant the world to me.

"Dit is Dylan, hij heeft issues". Sheree, ondanks deze prachtige introductie hebben we het toch voor elkaar gekregen goede vrienden te worden. Onze gedeelde liefde voor Mean Girls , chick flicks en slechte kerstfilms heeft toch een sterke band gevormd. Bereid je maar voor op nog heel veel filmmarathons en gezellige girls nights samen.

Thamar, eerst samen onze stages doen bij Eriba en daarna tegelijk onze PhDs doen. Jij was mijn buddy met wie ik het kon hebben over alle stress en problemen wanneer ik even een onafhankelijke blik nodig had, maar ook met wie ik het over de eigenlijk belangrijke dingen in het leven kon hebben. Ik wou dat ik ervoor kan zorgen dat jij evenveel vertrouwen hebt in jezelf als ik in jou heb, want dan zou je eindelijk begrijpen hoe fantastisch je bent.

Tijdens een PhD leer je kritische vragen te stellen over je onderzoek, maar mijn gesprekken met jou, Nilouq, dwongen mij altijd om ook kritisch na te denken over mijzelf als persoon. Ik vind het altijd heerlijk om met jou te filosoferen over het leven en goed na te denken over waarom wij dingen doen op de manier dat wij ze doen. Je bent natuurlijk nog zo veel meer dan dat en ik waardeer onze gezellige avondjes altijd enorm.

Anne-Grete, even though you already dumped on my nerdy way of life and the hours I spent behind the computer on the first night we met, you've become an amazing friend who I cherish very much. We've had so many dinners, movie nights in which you fell asleep and deep conversations together, that I'm sure that much of my personality is based on our time spent together. You're an amazing person, and even though you're now far away, I hope we can continue our friendship for a long, long time.

Joram en Marije, naast ons groepje, ga ik natuurlijk nog heel vaak met jullie los om. Jullie zijn echt fantastisch. Onze vakantie samen naar de VS was gewoon echt perfect en zal ik altijd met veel geluk op terug kijken. Als paranimfen helpen jullie mij nu bij het afronden van mijn PhD met een knaller, maar eigenlijk zijn jullie al jaren lang paranimfen voor mij geweest waarin jullie alles een feestje voor mij maken. Ik moet jullie veel meer bedanken dan ik hier in woorden kan doen, maar ik ga mijn best doen dat te doen door jullie dezelfde steun te geven die jullie mij altijd geven.

Johanne, I don't know what to write for you. You have meant so much for me these last few years. You've supported me through all the highs, but more importantly, you were also there for me when I went through some deep lows. I love how funny, caring, passionate and incredibly creative you are. You always inspire me to be a better person and actively help me in becoming one. Whenever I'm starting to feel down, you are the one to pull me back up and help me get on with things. With you by my side, I feel ready to explore the world and I can't wait to experience it all together.

Curriculum vitae

Dylan Hanraoi de Vries was born on May 27th 1994 in Gouda, the Netherlands. After completing his HAVO at the Zernike in Haren in 2011, he went on to study bioinformatics at the Hanze University of Applied Sciences in Groningen. During his studies, he was involved in doing extracurricular activities in which he learned to set up and perform his own research and expand his knowledge into other fields by following courses from other study programs. During his 12-month internship at the Wellcome Trust Sanger Institute in Cambridge he developed tools to identify copy number variation (CNV) data from whole exome sequencing data to identify disease-causing CNVs in sick children. While in Cambridge, he became aware of the existence of single cell technologies and was immediately enthralled by them.

In 2015, he started his Bioinformatics and Systems Biology MSc programme at the Vrije Universiteit and Universiteit van Amsterdam, in Amsterdam for which he would graduate *cum laude*. To learn more about single cell technologies, he did both his minor and major internships on the topic. In his minor internship he studied the sequencing biases of a newly developed single cell DNA sequencing method in the European Research Institute for the Biology of Aging in Groningen. His major internship was in the research group of Lude Franke in the Department of Genetics at the University Medical Centre Groningen, in which he studied the effects of *Candida albicans* on gene expression in human peripheral blood mononuclear cells.

In 2017, Dylan started his PhD as a continuation of his major internship project under the supervision of prof. Lude Franke, prof. Cisca Wijmenga and dr. Monique van der Wijst at the Department of Genetics of the University Medical Centre Groningen. He studied the effects of genetic variation and exposure to pathogens on the human immune system, resulting in this thesis. In 2018, he won the young investigator award for outstanding science at the European Conference for Human Genetics for his presentation titled “Personalized co-expression networks reveal genetic risk factors that change the regulatory wiring of cells”.

Dylan now works at Ancora Health as a bioinformatics scientist, where he helps develop models to do disease risk prediction and to explore avenues to improve health through digital lifestyle interventions.

List of Publications

First Author Publications

- Oelen, R.*, de Vries, D.H.*, Brugge, H.*, et al. Single-cell RNA-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure. Nature Communications 13, 3267 (2022). https://doi.org/10.1038/s41467-022-30893-5
- Li, S.*, Schmid, K.T.*, de Vries, D.H.*, et al. Identification of genetic variants that impact gene co-expression relationships using large-scale single-cell data. bioRxiv (2022). doi: https://doi.org/10.1101/2022.04.20.488925
- de Vries, D.H., et al. Integrating GWAS with bulk and single-cell RNA-sequencing reveals a role for LY86 in the anti-Candida host response. PLOS Pathogens (2020). https://doi.org/10.1371/journal.ppat.1008408
- van der Wijst, M.G.P.*, de Vries, D.H.*, et al. Science Forum: The single-cell eQTLGen consortium. eLife 9 (2020). https://doi.org/10.7554/eLife.52155
- van der Wijst, M.G.P.*, de Vries, D.H.*, et al. An integrative approach for building personalized gene regulatory networks for precision medicine. Genome Med 10, 96 (2018). https://doi.org/10.1186/s13073-018-0608-4

Second Author Publications

- van der Wijst, M.G.P., Brugge, H.*, de Vries, D.H.*, et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nature Genetics 50, 493–497 (2018). https://doi.org/10.1038/s41588-018-0089-9

Co-author Publications

- Võsa, U., Claringbould, A., Westra, HJ. et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nature Genetics 53, 1300–1310 (2021). https://doi.org/10.1038/s41588-021-00913-z
- Spreckels, J.E., et al. Lactobacillus reuteri Colonisation of Extremely Preterm Infants in a Randomised Placebo-Controlled Trial. Microorganisms 9, 915 (2021). https://doi.org/10.3390/microorganisms9050915
- Aguirre-Gamboa, R.*, de Klein, N.*, di Tommaso, J.*, et al. Deconvolution of bulk blood eQTL effects into immune cell subpopulations. BMC Bioinformatics 21, 243 (2020). https://doi.org/10.1186/s12859-020-03576-5
- O'Donnell-Luria, A.H., et al. Heterozygous Variants in KMT2E Cause a Spectrum of Neurodevelopmental Disorders and Epilepsy. American Journal of Human Genetics (2019). https://doi.org/10.1016/j.ajhg.2019.03.021
- Gorman, K.M., et al. Bi-allelic Loss-of-Function CACNA1B Mutations in Progressive Epilepsy-Dyskinesia. American Journal of Human Genetics (2019). https://doi.org/10.1016/j.ajhg.2019.03.005

- Blok, L.S., et al. Author Correction: CHD3 helicase domain mutations cause a neurodevelopmental syndrome with macrocephaly and impaired speech and language. Nature Communications 10, 883 (2019). https://doi.org/10.1038/s41467-019-08800-2
- Faundes, V., et al. Histone Lysine Methylases and Demethylases in the Landscape of Human Developmental Disorderes. American Journal of Human Genetics (2018). https://doi.org/10.1016/j.ajhg.2017.11.013
- Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. Nature 542, 433–438 (2017). https://doi.org/10.1038/nature21062

A