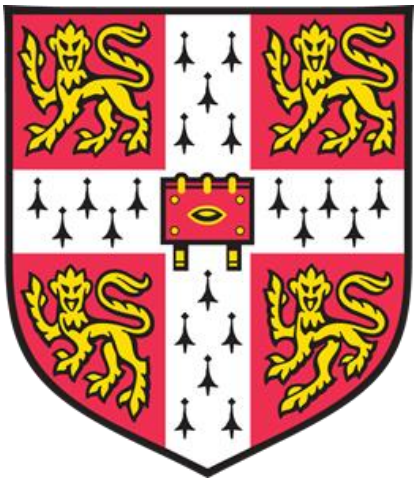
# Understanding the role of CD4+ T cells in common immune-mediated diseases




**Eddie Cano Gámez**

University of Cambridge
Selwyn College

May, 2021

This dissertation is submitted for the degree of
*Doctor of Philosophy*

## Declaration of originality

I hereby declare that, except for the instances specifically indicated in the collaboration note, this dissertation is the result of my own original work and contains nothing which is the outcome of work done by others or in collaboration with others. Moreover, I declare that this work has not been considered or is being considered for any other academic degree or qualification at this or any other university. Finally, I declare that this dissertation contains **46,404 words** (exclusive of tables, references, and appendices), which is below the 60,000 word limit prescribed in the Special Regulations of the PhD examination for which I am a candidate.

Eddie Cano Gámez

May, 2021

**Title:** Understanding the role of CD4+ T cells in common immune-mediated diseases

**Candidate:** Eddie Cano Gámez

**Degree:** PhD in Biological Science (Sanger Institute)

**Summary:** Immune-mediated diseases such as autoimmunity are complex traits which collectively affect around 10% of the European population. Genome-wide association studies (GWAS) have demonstrated that genetic susceptibility to these diseases is explained by thousands of loci spread throughout the genome, most of which lie within non-coding DNA. Moreover, these loci are enriched in CD4+ T cell regulatory elements, which suggests they might disrupt the expression of nearby genes in T cells. Nonetheless, the target genes of most immune disease loci have yet to be discovered. In this dissertation, I describe three studies designed to further our understanding of the relationship between genetic variation, CD4+ T cell function, and disease risk.

I first introduce a large epigenetic study which profiled active promoters and enhancers in 55 different CD4+ T cell and macrophage states. By integrating these data with GWAS loci with a novel statistical approach, I conclude that immune disease loci are enriched in enhancers and promoters specifically active during early memory T cell activation. In a second study, I proceed to characterize memory CD4+ T cells at single-cell resolution by profiling cells in the resting state and after stimulation with 11 different cytokine combinations. My observations reveal that CD4+ T cells are formed of a continuum of cell states which reflect a naïve-to-memory progression, and that as cells advance in this progression they express increasingly higher levels of cytokines, chemokines, and other effector molecules. Finally, I describe the results from a single-cell expression quantitative trait locus (sc-eQTL) mapping study performed on CD4+ T cells undergoing activation. I identify over 6,000 genes regulated by an eQTL, of which approximately 2,000 show evidence of a gene-by-environment interaction, where the eQTL effect size changes as a function of T cell activation time. Integration with GWAS associations demonstrates that immune disease loci alter the expression of genes *in cis* at specific stages of T cell activation. This results in the prioritization of 139 candidate disease genes which could be relevant for drug target identification. These results expand our understanding of CD4+ T cells and suggest that dysregulation of gene expression dynamics during T cell activation could be a hallmark of disease.

*“The stars, we are given. The constellations, we make. That is to say, stars exist in the cosmos, but constellations are the imaginary lines we draw between them, the readings we give the sky, the stories we tell.”*

Rebecca Solnit

*“Our genes are the pen and paper we are born with, but the story we will write is up to us.”*

Leena Peltonen-Palotie

*“If you can’t give me poetry, can’t you give me poetical science?”*

Ada Lovelace

*For years I worked hard to attain at least a glimmer of objectivity. I tried to distill it with flasks and cell culture plates, with sketches and computer programs. I strived to set right the biases and misconceptions of previous generations, of those who came before me. But in writing I realise that there is no objectivity; that even the word* ***T cell*** *comes from the initial for* ***thymus****, an organ named after its resemblance to a bud of* ***thyme****, itself named after the greek word* ***thuein****, meaning smoke, because thyme was burnt as a sacrifice. I don't offer objectivity. This dissertation is but a large work of fiction.*

# Collaboration note

The work described in this dissertation is highly interdisciplinary and was the product of multiple collaborations between individual researchers, groups, and institutions. These collaborations are listed below.

This dissertation forms part of a study designed by Dr Gosia Trynka and Dr Blagoje Soskic, funded by the public-private initiative Open Targets (grant code OTAR040). Open Targets brings together pharmaceutical companies and academic institutions working on drug target identification and prioritisation. Under this scheme, the firms GlaxoSmithKline (GSK) and Biogen were involved in the study design and provided constant and valuable feedback throughout the interpretation of results. In particular, Dr David Wille, a current employee at GSK, contributed substantially to the development of the statistical method *CHEERS*, described in chapter 2. Furthermore, Paola Bronson, a current employee of Biogen, provided valuable feedback on the colocalization results presented in chapter 4. The work described in chapters 2, 3, and 4 was led by Blagoje Soskic and myself, with the invaluable help of several other members of the immune genomics group, especially Deborah Plowman, who performed a large proportion of the experimental work. Moreover, the staff at the sequencing and flow cytometry facilities of the Wellcome Sanger Institute made the single-cell and high-throughput sequencing work described in this dissertation possible. Finally, the proteomics work described in chapter 3 was carried out in collaboration with the Functional Proteomics group at the Institute of Cancer Research (ICR) in London, UK, where Dr Theodoros Roumeliotis carried out all Mass Spectrometry experiments and upstream proteomics data analysis under the supervision of Dr Jyoti Choudhary.

## Publication note

The work described in chapters 1, 2, and 3 of this dissertation has been published as research articles in several academic journals. Where this is the case, a note has been added at the beginning of the chapter to indicate that the results are public. A full list of the publications derived from this dissertation is detailed below in chronological order:

Soskic B, **Cano-Gamez E**, Smyth DJ, et al. (2019). Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. *Nature Genetics*. 51, 1486-1493. doi: https://doi.org/10.1038/s41588-019-0493-9

**Cano-Gamez E**, Soskic B, Roumeliotis TI, et al. (2020). Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4+ T cells to cytokines. *Nature Communications*. 11, 1801. doi: 10.1038/s41467-020-15543-y

**Cano-Gamez E** and Trynka G. (2020). From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases. *Frontiers in Genetics*. doi: 10.3389/fgene.2020.00424

# Acknowledgements

Few moments in my PhD can compare to these final hours of thesis proof-reading, when I can look at the entirety of my work, with its achievements and its drawbacks, and think of the people who made it possible. Mostly, I feel a deep sense of gratitude that extends to countless individuals who contributed their ideas, intellectual and emotional support to this long endeavour.

I first and foremost want to thank my supervisor, Gosia Trynka, for her unconditional guidance throughout what surely were the four most challenging years of my life. Her scientific mentorship was unparalleled and allowed me to grow so much in so many aspects, that I can hardly recognise myself. This dissertation would have looked very bleak without all the discussions we had during impromptu meetings in her office or, once the pandemic hit, over zoom calls. Nonetheless, important as this is, it is not her scientific input that I value the most, but rather her utterly selfless way of caring for people. I want to thank her because, when the pandemic hit and I was figuring out how to survive, she asked how I was doing and when I would take time to rest. Because she always gave much more credit to my work than I did myself, proudly presented it to others at any possible chance, and pushed me to apply to many opportunities I did not believe I even qualified for. As a scientist in training, few things are as valuable and motivating as someone who believes in you. I also want to thank Blagoje Soskic, with whom I share credit for everything here presented. During my MPhil degree, Blagoje was like a second supervisor to me, teaching me everything from how T cells function to how to operate a cytometer. As I made my way through my training years, Blagoje and I became colleagues who designed, carried out, and interpreted experiments together, and I am happy to say that, at the end of this stage of career, I leave with a lifelong friend who shares my passion for science, languages, and many other little things. This is already the best result I could ask for. I would also like to acknowledge the support I received from everyone in the Immune Genomics group at the Sanger Institute. Though I don't have the space to list all your names, you know who you are and how much I appreciated both your intellectual support and your ability to build a lovely working environment. I want to make special mention here of Deborah Plowman, a truly gifted scientist whose stellar work enabled us to turn the experiments in our minds into physical realities.

In addition, I would like to thank all the funding bodies who believed in this work and the scientists who supervised any of my PhD work. Duncan Odom and Carl Anderson, my PhD thesis committee, as well as Leopold Parts, who evaluated my first year work, provided invaluable feedback. The Gates Cambridge Scholarship, Open Targets, and the Sanger

Institute all believed in me enough to fund my work, and I hope I made them justice. Last but not least, I would like to thank everyone who came before me: the endless network of scientists and thinkers which begins in the 330 citations listed in my reference list and extends uninterrupted until the very dawn of human thinking. It is wonderful and humbling in equal parts to join such a noble human enterprise and to realise that, even though my contribution might be negligible, it links with that of everybody else to form a beautiful cobweb. I hope future generations will extend these threads deeper into the darkest realms of the unknown.

I would finally like to take some time to recognise the importance of all the support I received from the ones I love the most. Firstly, I want to thank my parents for believing in me and sharing my excitement and frustrations, even though they weren't involved in any of the science. Despite being far away, they managed to make me feel at home, anywhere I was. Thanks especially to my mom for sending me chocolates and puzzles during the pandemic to ensure I knew I was not alone. I also want to thank my sister, Tabata, whom I admire and miss very much, and whose devotion to helping others and to the field of medicine always gave me the motivation I needed to continue. It is thanks to her support that I am in Cambridge in the first place, and it is also thanks to her that I can reach the end of this PhD. I also want to thank my friends in Cambridge and elsewhere, who believed in me and cared for me. Finally and most importantly, I would like to thank Francesca Brown, the most special person in my life. I very much doubt I would have made it out of this PhD sane were it not for her. Her patience and loving kindness, even in the most difficult moments, reminded me that life is beautiful and worth living, and helped me keep my mental wellbeing. In addition, she always questioned my narrow scientific thinking with her understanding of social phenomena, and her beautiful mind helped me become a better geneticist and a better person. I dedicate all of this work to her.

## Brief note on mental health

During the course of my PhD, I was diagnosed with dysthymia, the lay term for persistent depressive disorder. I spent many afternoons in bed, wrapped in my duvet, convinced the world was the darkest of places. I struggled to find enough motivation to do anything at all. I still feel this some days. The fact that I completed such a productive PhD seems to me a true miracle, and I owe it to those who supported me, be it consciously or unconsciously. Thank you in particular to my family and my partner, who convinced me to seek therapy, to my supervisor, who was incredibly kind and understanding of my struggle, and to my colleagues, who made work easier.

Though many hesitate to admit it, mental health issues are widespread in academia. If you too are struggling, know you are not alone. Please seek help, take time for yourself, and cultivate a balanced life. And if you ever think you are not doing enough, remember this dissertation and the fact that, for the most of it, I completed my PhD working 8 hours a day, five days a week. You are doing more than enough.

# List of acronyms and abbreviations

| | |
|---|---|
| AD | Alzheimer's disease |
| AO/PI | Acridine orange/propidium iodide stain |
| APC | Antigen-presenting cell |
| ATAC | Assay for transposase-accessible chromatin |
| bp | Base pairs |
| CAD | Coronary artery disease |
| CCA | Canonical correlation analysis |
| CD | Crohn's disease |
| CD[number] | Cluster of differentiation number |
| CDR3 | Complementarity-determining region 3 |
| CeD | Celiac disease |
| CHEERS | Chromatin element enrichment ranking by specificity |
| ChIP | Chromatin immunoprecipitation |
| DAPI | 4',6-diamidino-2-phenylindole |
| DNA | Deoxyribonucleic acid |
| DHS | DNAse hypersensitivity |
| FACS | Fluorescence-activated cell sorting |
| FDR | False discovery rate |
| FRiP | Fraction of reads in peaks |
| GEM | Gel bead-in-emulsion |
| GLIPH | Grouping of lymphocyte interactions by paratope hotspots |
| GTEx | Genotype-tissue expression atlas |
| GWAS | Genome-wide association study |
| HDL | High density lipoprotein |
| HLA | Human leukocyte antigen |
| H3K4me1 | Monomethylation of histone 3 at lysine 4 |
| H3K4me3 | Trimethylation of histone 3 at lysine 4 |
| H3K27ac | Acetylation of histone 3 at lysine 27 |
| H3K27me3 | Trimethylation of histone 3 at lysine 27 |
| IBD | Inflammatory bowel disease |
| IFN | Interferon |
| IL | Interleukin |
| kb | Kilobases |
| LC-MS/MS | Liquid chromatography-tandem mass spectrometry |
| LD | Linkage disequilibrium |
| LDL | Low density lipoprotein |
| LFC | Log-fold change |
| MAF | Minor allele frequency |
| MHC | Major histocompatibility complex |
| MPRA | Massively parallel reporter assay |
| MR | Mendelian randomization |
| MS | Multiple sclerosis |
| PBMC | Peripheral blood mononuclear cell |
| PBS | Phosphate buffered saline solution |
| PC | Principal component |

| | |
|---|---|
| PCA | Principal component analysis |
| pMHC | Peptide complexed with major histocompatibility complex molecules |
| PRS | Polygenic risk score |
| QTL | Quantitative trait locus |
| QTS | Quantitative trait score |
| RA | Rheumatoid arthritis |
| RNA | Ribonucleic acid |
| RIN | RNA integrity number |
| sc | Single-cell |
| seq | Sequencing |
| SLE | Systemic lupus erythematosus |
| SNP | Single-nucleotide polymorphism |
| $T_{CM}$ | Central memory T cell |
| TCR | T cell receptor |
| $T_{EM}$ | Effector memory T cell |
| $T_{EMRA}$ | Effector memory T cell re-expressing CD45RA |
| TF | Transcription factor |
| TFBS | Transcription factor binding site |
| TGF | Transforming growth factor |
| Th | T helper |
| $T_N$ | Naive CD4+ T cell |
| TNF | Tumor necrosis factor |
| $T_M$ | Memory CD4+ T cell |
| Treg | Regulatory T cell |
| TSS | Transcriptional start site |
| T1D | Type 1 diabetes |
| T2D | Type 2 diabetes |
| UC | Ulcerative colitis |
| UMAP | Uniform manifold approximation and projection |

# Table of contents

# Genetic susceptibility to complex immune-mediated diseases and its underlying molecular mechanisms

*A modified version of this chapter was previously published in Frontiers in Genetics (Cano-Gamez and Trynka, 2020), as detailed in the Publication Note.*

## Understanding the molecular causes of complex immune diseases

Immune-mediated diseases such as autoimmunity and autoinflammation are common non-communicable diseases which affect approximately 8% of individuals of European ancestry (Cooper et al., 2009), representing one of the most pressing challenges in present day healthcare. These conditions are influenced by the interaction between a genetic predisposition and environmental or lifestyle factors (Smith et al., 2005). Importantly, as opposed to rare disorders such as primary immunodeficiencies, which are often caused by the dysfunction of a single gene (Stray-Pedersen et al., 2017; Thaventhiran et al., 2020), common immune-mediated diseases are complex traits. This means that genetic predisposition to them arises from the contribution of thousands of common genetic variants, each having a small individual effect on disease (Hindorff et al., 2011). This makes the study of complex immune diseases challenging, as their genetic architecture follows a polygenic rather than a Mendelian model (Visscher & Goddard, 2019).

Genome-wide association studies (GWAS) are a statistical framework designed to map the polygenic architecture of complex traits by identifying common genetic variants present at a significantly higher frequency in individuals with disease than in the healthy population (i.e. risk variants) (Wellcome Trust Case Control Consortium, 2007). Most of the risk variants identified by GWAS are single-nucleotide polymorphisms (SNPs), one-letter changes in the DNA sequence found at genomic positions with substantial inter-individual variation. Over the last 12 years, GWAS have grown substantially both in sample size and in the number of investigated traits (Visscher et al., 2017), with over 9,000 risk variants and over 600 publications reported in the GWAS catalog under the category of immune system disease (MacArthur et al., 2017).

Despite the success of GWAS in identifying immune disease risk variants, the clinical insights derived from these studies are limited. This is due to the difficulty in interpreting the statistical associations derived from GWAS. Firstly, neighbouring genetic variants are often correlated with one another because they tend to be segregated and inherited together during meiosis,

a phenomenon known as linkage disequilibrium (LD) (Slatkin, 2008). Due to LD, a single disease causal variant can result in many neighbouring variants in the locus appearing to be more common in individuals with the disease purely due to co-segregation. This makes it difficult to pinpoint the causal variants underlying each association. Secondly, it is often unclear which cell types are involved in disease, since the pathophysiology of these traits implicates interactions between multiple cell types and it is unclear which of these cells are causally involved and which are altered as a consequence of disease progression. Finally, over 90% of GWAS variants fall in non-coding regions of the genome and thus do not directly affect the coding sequence of any gene. The accumulation of these variants in DNA regulatory elements (Maurano et al., 2012) and the observation that they can disrupt binding sites for transcription factors (Musunuru et al., 2010) suggests that these variants act by regulating the expression levels of nearby genes. However, disease-associated loci often contain many genes, making it challenging to distinguish the affected ones. Consequently, follow-up studies are necessary to interpret GWAS results and to infer the exact disease-causal variants, their target genes, and the cell types in which they act (**Figure 1.1**).

Several statistical methods have been designed to tackle these challenges. Most of these work by integrating GWAS signals with functional genomics data sets such as gene expression or chromatin activity profiles assayed across a range of cell types and tissues. Three families of approaches are of particular importance: 1) fine-mapping aims to identify the causal variants at each GWAS locus, 2) SNP enrichment analysis aims to prioritize disease relevant cell types, and 3) colocalization aims to nominate the most likely target genes (**Figure 1.1**). In this chapter, I review a selection of these methods with a particular focus on studies which have used them to identify the causal cell types and genes in complex immune-mediated diseases. Finally, I reflect on some of the challenges and opportunities of post-GWAS research, such as the availability of high-throughput single-cell sequencing technologies, the integration of GWAS signals into polygenic risk scores, and the systematic use of genetic engineering for investigating GWAS loci.

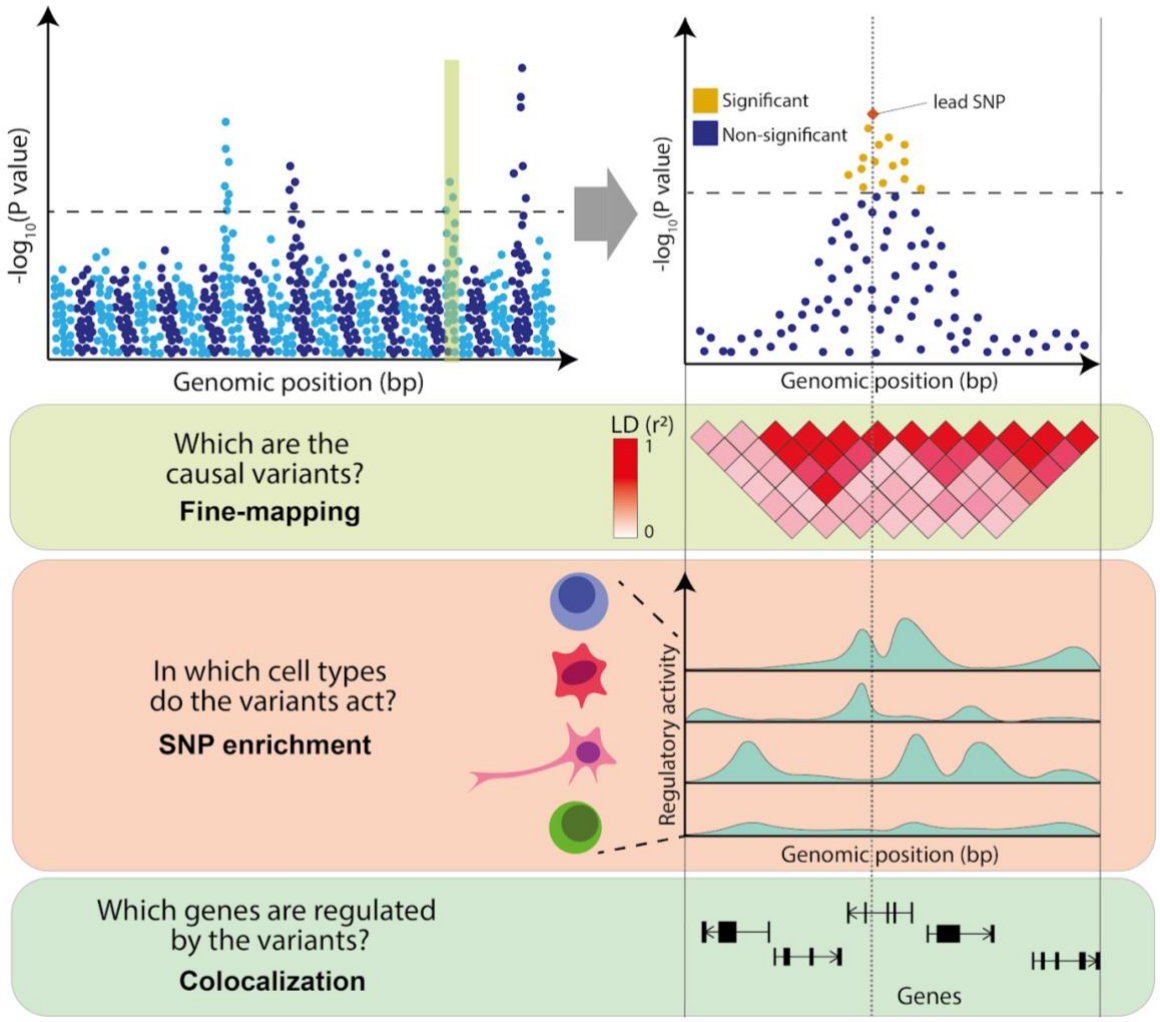



**Figure 1.1 Challenges in interpreting GWAS associations** The top panel shows a Manhattan plot illustrating the association between genetic variants and a trait (e.g. a disease) at a genome-wide level (left panel) and within an example locus (right panel). Variants above the dotted line represent genome-wide significant associations. The panels below illustrate the main challenges in interpreting GWAS associations: high LD between variants (encoded in shades of red), variable levels of regulatory activity of the genomic regions across cell types (peaks of different heights represent different levels of activity of chromatin marks) and multiple genes within the associated locus.

## Approaches to identify disease causal cell types

The loci identified by GWAS provide a strong anchoring point to the regions of the genome involved in disease. However, going from genetics to function requires robust model systems in which disease-causal cells and tissues can be probed and manipulated to understand their pathogenic role. For example, tumor-derived human cell lines have been extensively used for the systematic identification of novel drug targets in cancer (Behan et al., 2019). Such model systems provide valuable clues for drug development, as they enable the elucidation of molecular mechanisms of disease, the identification of relevant genes, and the screening of compounds with therapeutic potential at high-throughput. However, for many complex immune-mediated diseases it is unclear which cells are causal. For instance, independent

studies have proposed that rheumatoid arthritis (RA) is caused by cells as diverse as T cells (Cope et al., 2007), B cells (Bugatti et al., 2014), macrophages (Udalova et al., 2016), and synoviocytes (Beatrix Bartok, 2010). This problem is not specific to immune disease and encompasses most complex traits. For example, psychiatric traits, which involve dysregulation of the central nervous system, pose a similar challenge due to the complex histological structure of the brain, so that over 20 different cellular models have been used to study bipolar disorder (Viswanath et al., 2015). This lack of ground truth causal cell types makes the functional validation of GWAS variants challenging, as dozens of tissues could potentially be involved in the development of a trait. Statistical methods that integrate GWAS variants either with transcriptomic or chromatin annotations assayed across a range of different tissues can help us find the most disease-relevant cell types.

### SNP enrichment analysis based on genome-wide significant risk variants

If a cell type is causally involved in a disease, then we would expect risk variants to accumulate in genomic regions specifically active in that cell type. We refer to this overrepresentation of disease-associated SNPs in annotations derived from causal cell types as GWAS SNP enrichment. Methods to detect the presence of SNP enrichment integrate GWAS signals with different genomic annotations and prioritize the cell types in which risk variants tend to overlap the annotations more frequently than expected by chance.

There are multiple ways to define which genomic regions are active in a given cell type. For example, cell type specific activity of a locus can be defined based on the expression levels of nearby genes. As an example, an approach proposed by Hu et al. (*SNPsea*) defined cell type specific genes as those with higher expression in one cell type compared to the rest (Hu et al., 2011). If GWAS loci for a trait tended to be in close proximity to genes specifically expressed in a given cell type, then that cell type was said to be enriched and could be prioritized in future studies. The statistical significance of this enrichment was derived from a permutation-based test in which disease-associated loci were compared to random loci of similar properties (Slowikowski et al., 2014). The authors used this approach to study Crohn's disease, systemic lupus erythematosus (SLE) and RA, testing for enrichment in gene expression across 79 human and 223 mouse tissues. While SLE-associated variants were enriched in genes specifically expressed in B cells, RA variants were enriched in genes specific to CD4+ T cells (Hu et al., 2011). This demonstrated that SNP enrichment is a valid approach for cell type prioritization and suggested that variants associated with immune-mediated diseases affect the function of the adaptive immune system.

However, most GWAS variants are not located within the body of a gene. Thus, gene expression-based methods are constrained to use an arbitrary definition of which genes contribute to the SNP enrichment calculation at each locus, either selecting a single gene with the highest cell type specific expression or including all genes in the region (Hu et al., 2011). A caveat of these two approaches is that the first one can select the wrong gene and does not account for the effects of multiple causal genes, while the second one can dilute the signal by including many genes which are likely not relevant to the tested trait. Alternatively, GWAS variants can be integrated with chromatin annotations such as open chromatin regions (assayed by DNase-hypersensitivity or using the assay for transposase-accessible chromatin using sequencing, ATAC-seq) (A. P. Boyle et al., 2008; Buenrostro et al., 2013), histone modifications (e.g. H3K4me1, H3K4me3, H3K27ac, and H3K27me3, assayed by chromatin immunoprecipitation followed by sequencing, ChIP-seq) (Bannister & Kouzarides, 2011) or DNA methylation (Frommer et al., 1992). These annotations tag genomic elements with high levels of regulatory activity, often referred to as peaks. For example, DNA accessibility peaks indicate regions available for transcription factor (TF) binding, H3K4me3 peaks highlight gene promoters (Barski et al., 2007), and H3K27ac peaks mark active enhancer and promoter regions (Creyghton et al., 2010). As opposed to gene expression, chromatin marks can be physically overlapped with non-coding GWAS variants and therefore enrichment analysis can be estimated directly from the SNPs located within the annotations (**Figure 1.2**). Initiatives like the Encyclopedia of DNA Elements (ENCODE) (ENCODE Project Consortium, 2012), Roadmap Epigenomics (Roadmap Epigenomics Consortium et al., 2015) and the BLUEPRINT project (Chen et al., 2016) have profiled tens of epigenetic marks across dozens of human tissues, providing rich resources for SNP enrichment analysis.

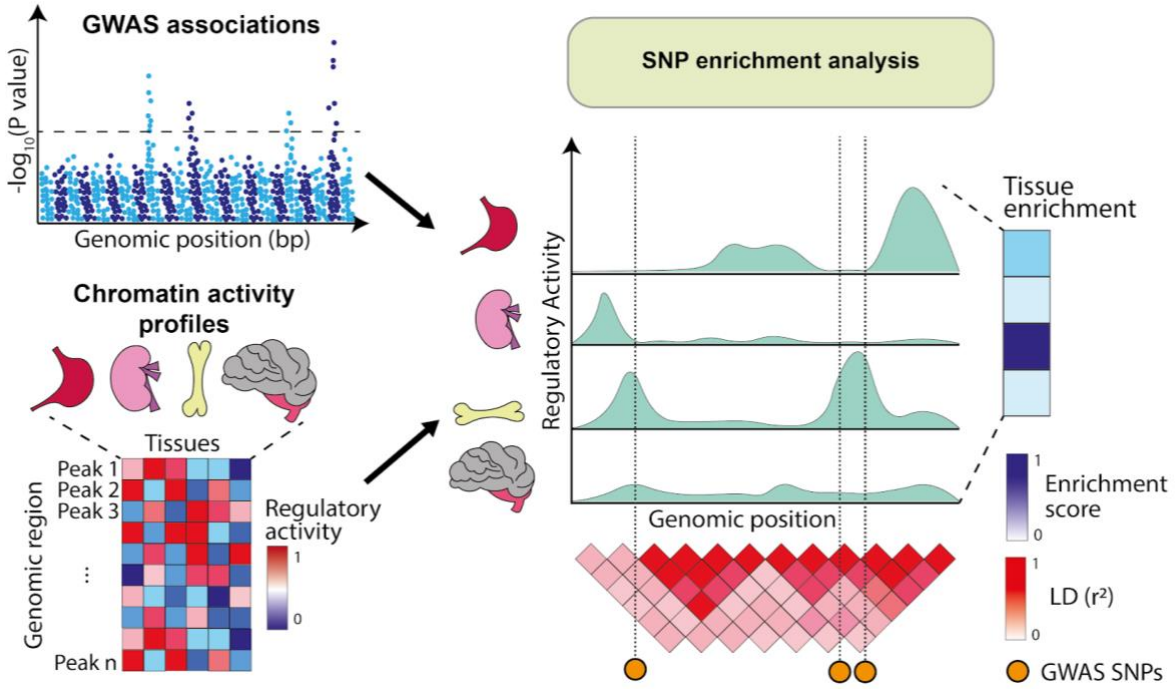



**Figure 1.2. Overview of SNP enrichment analysis using chromatin annotations** SNP enrichment analysis integrates association signals from GWAS (Manhattan plot at the top left) with functional genomics data such as chromatin annotations (heatmap at the bottom left). GWAS SNPs are overlapped with regulatory elements (right panel) and the overlap occurs if in a given tissue more frequently than expected by chance, the tissue is assigned a high enrichment score.

An early example of SNP enrichment analysis based on chromatin annotations used GWAS variants for 447 traits and overlapped them with DNase-hypersensitive (DHS) regions from 348 tissues (Maurano et al., 2012). Using a simple binomial test, this study found that GWAS SNPs were enriched in DHS regions compared to a background set of common SNPs. Importantly, this enrichment was tissue-specific. For example, risk variants for coronary heart disease and body mass index were enriched in DHS regions active in fetal cells. Conversely, genetic variants associated with age-related traits (e.g. cancer and immune-mediated diseases) were significantly depleted from fetal DHS regions. These findings showed that GWAS variants could modify the regulatory activity of non-coding elements in a cell-type specific manner.

However, many GWAS loci reside in regions of high gene density, which also show a high density of chromatin regulatory elements. This can confound enrichment estimates. To address this, a further study tested for enrichment of disease variants in DHS regions by comparing GWAS SNPs to random sets of SNPs which were matched for gene density, among other properties (i.e. LD, gene density, and distance to the nearest transcriptional start site, TSS) in a permutation-based framework (*GREGOR*) (Schmidt et al., 2015). By matching

SNPs, this approach is robust to both gene and annotation density. Results from this study confirmed that GWAS SNPs are generally enriched in active regulatory regions compared to random SNPs.

In addition to the binary overlap between SNPs and annotations, SNP enrichment analysis can also take into account other peak properties, such as the position of a variant within a peak and the height of the peak (reflecting the absolute levels of regulatory activity of that genomic region). Moreover, SNP enrichment analysis can be extended to chromatin marks other than DHS. For example, *epiGWAS* is a method which tests for the accumulation of GWAS variants in chromatin regions defined using ChIP-seq for histone modifications (Trynka et al., 2013). In this approach, variants within each GWAS locus are scored for their distance to the summit of the nearest peak and for the height of such peak (i.e. the height to distance ratio, h/d). The contribution of each locus to the final enrichment score is determined by a single variant with the highest h/d score, and statistical significance is inferred by comparison to a matched set of random SNPs. This approach is especially suitable for narrow histone marks, where peak summits can be reliably defined. The authors of this method confirmed that variants associated with low-density lipoprotein (LDL) cholesterol levels were enriched in gene promoters active in the liver, while type 2 diabetes (T2D) variants were enriched in gene promoters active in liver cells and pancreatic islets. In both cases, the tissues are well understood to play a role in disease. The authors also used this approach across immune-mediated traits, where pathogenic cell types are less well characterised. This revealed an enrichment of RA and type 1 diabetes (T1D) variants in CD4+ T cell subsets, particularly in regulatory T cells. These observations align well with results from previous enrichment methods (Hu et al., 2011), which also propose CD4+ T cells as key players in disease development.

One limitation of the above methods, which all rely on random sampling of SNPs to derive a null distribution, is that they make assumptions about the SNP parameters that need to be controlled for during random sampling (e.g. proximity to TSS, minor allele frequency, gene density, etc). However, the presence of hidden confounders could bias the enrichment statistics if uncontrolled for. One approach to address this, the *GoShifter* method (Trynka et al., 2015), derives statistical significance by shifting the location of functional annotations within the tested regions while preserving the distance between them. The result is a null distribution of SNP-annotation overlaps due to chance. This approach maintains the local genomic architecture, including the number of tested SNPs in LD, the number of annotations and the distance between the features, therefore controlling for hidden confounders. The authors of *GoShifter* reported a significant enrichment of breast cancer risk variants in human

mammary epithelial cells (Trynka et al., 2015), which are known to be involved in disease. Moreover, risk variants for RA were enriched in promoter regions specific to CD4+ memory T cells, significantly strengthening previous evidence that CD4+ T cells are causally involved in rheumatoid arthritis (Hu et al., 2011; Trynka et al., 2013).

Additional studies have also argued for a role of CD4+ T cells in immune mediated diseases. Onengut-Gumuscu et al., asked if credible sets of T1D SNPs, derived using fine-mapping, were enriched in functional annotations from the ENCODE (ENCODE Project Consortium, 2012) and Roadmap (Roadmap Epigenomics Consortium et al., 2015) projects (Onengut-Gumuscu et al., 2015). They did so by comparing the proportion of disease-associated SNPs and non-disease SNPs which overlapped functional elements, stratifying variants by their minor allele frequency. Interestingly, T1D credible sets were strongly enriched in immune cell enhancers, particularly enhancers active in CD4+ and CD8+ T cells. Conversely, there was no detectable enrichment in enhancers active in pancreatic islets, in agreement with T1D being an immune-mediated pathology. In contrast, a separate study profiled open chromatin, transcription factor binding and gene expression in human pancreatic islets and integrated these profiles with GWAS loci for T2D and fasting glycemia (Pasquali et al., 2014). The authors used a permutation-based test to estimate enrichments and concluded that glycemia and T2D SNPs were strongly enriched in pancreatic islet enhancers, where they disrupted DNA binding by key islet transcription factors. This illustrates how GWAS SNP enrichment can distinguish different disease etiologies based solely on genetic associations, despite the traits sharing similar physiological manifestations.

Once the disease-relevant cell types are identified, subsequent experiments can be carried out to further refine the observed enrichments and identify the most relevant cell states. For example, a recent study performed SNP enrichment analysis for variants associated with nine different immune mediated diseases using chromatin accessibility profiles from 25 immune cell types, both in the resting state and after *in vitro* stimulation (Calderon et al., 2019). The authors observed that disease variants were strongly enriched in open chromatin regions specific to stimulated T cells. This accumulation of GWAS variants in regions that are only observed upon stimulation suggests that the cell states and functions involved in disease could be missed by looking at primary cells from blood, which are often in the resting state. This is in line with the observation that SNPs have regulatory effects which are only detectable when cells are exposed to a given stimulus, as has been shown in monocytes exposed to interferon gamma (IFNγ) and lipopolysaccharide (LPS) (Fairfax et al., 2014). Thus, translating immune disease GWAS variants to function will require the generation of more detailed functional maps

which include immune cells exposed to disease relevant stimuli such as pathogens, antigen or cytokines.

### SNP enrichment analysis based on common variants

The approaches described so far leverage the signal from genome wide significant risk variants. However, complex traits result from the added contribution of thousands of risk loci, the majority of which remain undiscovered at current GWAS sample sizes (Visscher et al., 2017). Thus, restricting SNP enrichment analysis to genome-wide significant variants could limit the statistical power to detect biologically important enrichments. This has motivated the development of a number of methods which leverage all common variants in the genome to estimate enrichments.

In a method called *fGWAS,* Pickrell reasoned that if GWAS variants are enriched in a given functional category, then SNPs that belong to that category should be more likely to have an effect on the trait (Pickrell, 2014). Using whole genome variants derived from imputation (Pasaniuc et al., 2014), he modelled the probability of each locus being associated with a disease as a function of its annotations, using a hierarchical Bayesian model. When applied to chromatin regulatory maps from 402 tissues and 18 complex traits, *fGWAS* identified enrichment of variants associated with high-density lipoprotein (HDL) levels in enhancers specifically active in the liver. Moreover, variants were generally depleted from repressed chromatin regions across all traits. By integrating functional annotations with GWAS statistics, *fGWAS* can also “re-weigh” variants and discover association signals for loci which did not originally reach genome-wide significance (Pickrell, 2014). An example is the SNP rs6659176, upweighted by *fGWAS* and confirmed to be associated with HDL in an independent study (Global Lipids Genetics Consortium et al., 2013).

**Table 1.1 Methods for SNP enrichment analysis** List of the methods for identification of GWAS SNP enrichment in functional annotations reviewed in this chapter.

| Method | Publications | Hypothesis tested | Input data |
|---|---|---|---|
| SNPsea | Slowikowski et al., 2014<br><br>Hu et al., 2011 | Accumulation of GWAS variants near genes with high tissue specificity | Gene expression, GWAS index variants |
| EpiGWAS | Trynka et al., 2013 | Accumulation of GWAS variants near highly active regulatory elements | Chromatin marks, GWAS index variants |
| GREGOR | Schmidt et al., 2015 | Accumulation of GWAS variants in regulatory elements | Chromatin marks, GWAS index variants |
| GoShifter | Trynka et al., 2015 | Intersection of GWAS variants with regulatory annotations (based on local-shifting of annotations) | Functional annotations, GWAS index variants |
| fGWAS | Pickrell, 2014 | Higher GWAS effect sizes observed if a loci and a SNP overlap a functional annotation | Functional annotations, GWAS summary statistics |
| GARFIELD | Iotchkova et al., 2019 | Higher GWAS effect sizes observed in variants that overlap regulatory annotations | Chromatin annotations, full GWAS summary statistics |
| RolyPoly | Calderon et al., 2017 | Higher GWAS effect sizes observed near highly expressed genes | Gene expression, full GWAS summary statistics |
| LDSC | Finucane et al., 2015 | Accumulation of heritability in variants overlapping a functional annotation | Chromatin annotations, full GWAS summary statistics |
| LDSC-SEG | Finucane et al., 2018 | Accumulation of heritability near tissue specific genes | Gene expression, full GWAS summary statistics |

In another study, Iotchkova et al. used a logistic regression framework to assess SNP enrichment (*GARFIELD*) (Iotchkova et al., 2019). The authors modeled the trait association status of each SNP as a probability which depends on the variant's features (i.e. whether the SNP overlaps a functional annotation, its distance to the nearest TSS, and its number of LD proxies). The presence or absence of an association for each SNP with the trait was tested at several significance thresholds, thus allowing more SNPs to be included in the calculation. The authors applied *GARFIELD* to DHS regions and functional annotations from ENCODE (Ernst & Kellis, 2012) and found that variants associated with height were enriched in DHS elements across all tissues, while ulcerative colitis (UC) variants showed tissue-specific enrichment mostly in blood cell types. Interestingly, the authors observed some of the enrichments only at lower significance thresholds. For example, variants associated with beta cell activity index were enriched in pancreatic islets enhancers only at lower significance thresholds (P value $< 1 \times 10^{-5}$). This suggests that including more trait-associated variants can improve enrichment estimates.

### Heritability enrichment analysis

Heritability is defined as the proportion of a trait's variance that is explained by genetic variation. In particular, SNP heritability is the amount of phenotypic variance explained by a given set of SNPs (J. Yang et al., 2017). A number of methods have been developed to estimate the SNP heritability of a trait using either individual-level genotypes (Bulik-Sullivan et al., 2015; J. Yang et al., 2010) or GWAS summary statistics (Bulik-Sullivan et al., 2015; J. Yang et al., 2010). This gave rise to partitioning heritability approaches, which test for a significant accumulation of trait heritability in different functional categories of the genome.

The authors of stratified LD-score regression (*LDSC*) (Finucane et al., 2015) argue that if GWAS variants are enriched in a functional category, then variants falling within that category will explain more trait heritability than other variants. To test this, Finucane et al. partitioned all common SNPs into categories based on the functional elements that they overlapped. These categories included 24 unspecific annotations (coding regions, promoters, enhancers, introns, conserved elements, and DHSs, among others) as well as histone modification profiles acquired from a variety of cell types. The authors calculated the SNP heritability of variants in each category using GWAS data for 17 traits and defined an enrichment score as the proportion of SNP heritability in a category divided by the proportion of SNPs in the category (Finucane et al., 2015). The authors found that, in general, conserved regions of the genome explained more heritability. Moreover, variants within enhancers specific to disease-relevant cell types also explained a substantial proportion of heritability. For example, liver-specific

enhancers were enriched for HDL heritability and enhancers active in the central nervous system captured more SNP heritability of psychiatric traits (e.g. schizophrenia and bipolar disorder) than variants residing in enhancers present in other cell types.

However, one limitation of the *LDSC* method is its dependency on chromatin activity profiles, which are not always available. In contrast, gene expression profiles are available for a far greater number of cell types, including the less abundant ones. *LD-score regression applied to specifically expressed genes* (*LDSC-SEG*) extends the *LDSC* framework to partition heritability using gene expression profiles (Finucane et al., 2018). It first identifies the top 10% most specific genes expressed in each tissue and extends the regions on each side of the genes by 100 kb. The resulting regions are used as tissue-specific annotations in which variants are partitioned. Because gene expression is available for a wider set of tissues than epigenetic data, this enabled the analysis of less common cell types. The authors used *LDSC-SEG* to integrate expression profiles form the genotype-tissue expression (GTEx) project with GWAS data for psychiatric traits and showed evidence of differential heritability enrichment across brain regions. For example, while only cells from the cortex were enriched for schizophrenia SNP heritability, both the cortex and the cerebellum were enriched for bipolar disorder SNP heritability. Subsequent application of *LDSC-SEG* to brain expression data from the PsychENCODE project (The PsychENCODE Consortium et al., 2015) revealed that schizophrenia SNP heritability enrichment was driven by glutamatergic neurons, while bipolar disorder SNP heritability enrichment was driven by GABAergic neurons. Importantly, these psychiatric traits had not been analyzed for SNP enrichment before because of the insufficient number of genome-wide significant risk variants. This highlights the increased statistical power enabled by including all common variants in the analysis.

Finally, the *RolyPoly* method models the polygenic architecture of complex traits to estimate SNP enrichment (Calderon et al., 2017). In brief, the authors reasoned that variants with higher GWAS effect sizes would tend to be nearby genes with higher expression in the causal tissues. Using a regression model, *RolyPoly* estimates the influence of cell type specific gene expression on the variance of GWAS effect sizes in each tissue. The authors applied *RolyPoly* to tissue-specific expression data from GTEx and confirmed a significant enrichment of variants affecting cholesterol levels in genes expressed in the liver and the small intestine. Moreover, they integrated GWAS data with single-cell gene expression profiles from brain tissue (Darmanis et al., 2015) and found a significant enrichment of risk variants for Alzheimer's disease (AD) in genes specific to microglia (Calderon et al., 2017). This agrees with increasing evidence suggesting the immune system is involved in AD pathology (Gosselin et al., 2017).

In summary, SNP enrichment analysis leverages GWAS signals and functional annotations to pinpoint disease-relevant cell types. Multiple approaches have been proposed to estimate enrichment, such as integrating genome-wide significant variants with chromatin or gene expression profiles, as well as partitioning the SNP heritability of a trait based on the functional annotations of the genome. A brief list of these approaches is available in **Table 1.1**. When applied to common complex immune diseases, all of these methods revealed the involvement of cells from the adaptive immune system, particularly CD4+ T cells. Moreover, recent evidence suggests this enrichment is higher in stimulated than in resting T cells. The increasing availability of expression and chromatin data for more cell types and states is expected to improve the granularity of these enrichment signals. This will allow us to confidently nominate the cell types and cell states causally involved in immune-mediated disease.

## Approaches to identify disease causal genes

Once the most relevant cell types are identified, it becomes possible to study the genes involved in disease. Identification of candidate causal genes is most straightforward for coding variants, which directly disrupt the structure of a protein. One notable example is the locus containing the *TYK2* gene, a kinase that mediates signal transduction downstream of various cytokine receptors. Variants at this locus have been associated with a number of immune mediated diseases such as RA, ankylosing spondylitis, multiple sclerosis (MS), and inflammatory bowel disease (IBD) (Franke et al., 2010; International Genetics of Ankylosing Spondylitis Consortium (IGAS) et al., 2013; Jostins et al., 2012; Okada et al., 2014). Importantly, a number of these SNPs are *TYK2* missense variants. There are three independent disease association signals at this locus, of which at least one is entirely explained by a coding SNP which confers disease protection (Diogo et al., 2015). This SNP induces a proline to alanine substitution in the catalytic domain of the TYK2 protein (Dendrou et al., 2016). This substitution significantly impairs cytokine signaling, thus altering the communication between immune cells. Surprisingly, even though this variant protects against more than ten different autoimmune diseases, complete knock-out of *TYK2* causes severe susceptibility to infections (Kreins et al., 2015). This led to the theory that TYK2 function constitutes a spectrum, with complete abrogation causing immunodeficiency and augmented function causing autoimmunity (Dendrou et al., 2016). Thus, a compound able to modulate the kinase activity of TYK2 could be a successful drug candidate for autoimmune disorders.

Despite the success in translating associations to function at the *TYK2* locus, approximately 90% of the variants identified by GWAS are non-coding (Farh et al., 2015) and cannot be easily linked to a candidate causal gene. These variants are thought to regulate gene expression via mechanisms such as modification of promoter and enhancer activity or disruption of binding sites for transcription factors. An example is the 1q13 locus, which contains a variant significantly associated with LDL cholesterol levels and myocardial infarction (Myocardial Infarction Genetics Consortium et al., 2009; Teslovich et al., 2010). This variant was shown to create a new transcription factor binding site, which in turn causes the recruitment of an enhancer-binding protein, sharply increasing the expression of the nearby gene *SORT1*, a regulator of lipoprotein levels in plasma (Musunuru et al., 2010). SORT1 in turn downregulates the levels of LDL. This makes SORT1 an interesting drug target in myocardial infarction.

Most of the variants associated with complex immune mediated diseases are thought to act by mechanisms analogous to those operating at the *SORT1* locus. However, GWAS loci often contain multiple genes and identifying the causal ones is challenging. Profiling molecular traits (e.g. gene expression, DNA methylation, transcription factor binding) and integrating them with GWAS signals enables the connection of non-coding variants to their target genes, unveiling the underlying regulatory events.

### Colocalization analysis

The quantification of molecular traits such as gene expression across thousands of individuals with different genotypes makes it possible to test for the association between these traits and common genetic variation. For example, the expression level of a gene or the amount of histone acetylation deposited at an enhancer can be correlated with the presence of specific alleles at nearby loci. If the level of the trait is higher in individuals carrying an allele, and if this increase is proportional to the genetic dosage (i.e. higher in homozygous compared to heterozygous individuals for the allele in question), then the locus is said to be a quantitative trait locus (QTL). Because the genetic variants carried by an individual are themselves not influenced by the measured trait, reverse causality can be largely excluded here. This means that QTLs represent regulatory elements likely to causally control the levels of the measured traits. The systematic identification of QTLs across the genome is referred to as QTL mapping (**Figure 1.3A**).

The decreasing costs of high-throughput sequencing have resulted in dozens of QTL-mapping studies, profiling traits as diverse as gene expression (eQTLs) (Nica & Dermitzakis, 2013), protein expression (pQTLs) (Melzer et al., 2008; B. B. Sun et al., 2018; Yao et al., 2018), exon

splicing (sQTLs) (Y. I. Li et al., 2018; Monlong et al., 2014; Ongen & Dermitzakis, 2015), DNA methylation (mQTLs) (Banovich et al., 2014; Hannon et al., 2016), chromatin acetylation (acQTLs) (Pelikan et al., 2018; W. Sun et al., 2016), and chromatin accessibility (caQTLs) (Degner et al., 2012; Kumasaka et al., 2016). Of these, eQTLs are the most common due to the widespread use of RNA-sequencing technologies. One of the most comprehensive eQTL resources is the GTEx project, which profiled 53 tissues across nearly 1,000 individuals (GTEx Consortium, 2013; Melé et al., 2015). Another initiative, the BLUEPRINT project, measured the transcriptome, together with DNA methylation and histone modifications, in the most abundant cell types in peripheral blood from 197 individuals (Chen et al., 2016).

The integration of QTLs with GWAS loci is a straightforward strategy to pinpoint the regulatory mechanisms potentially altered in disease. The first studies using this approach simply asked whether GWAS loci also tended to be eQTLs. A study by Nicolae et al. combined GWAS results for several diseases with eQTLs from human lymphoblastoid cell lines and concluded that GWAS SNPs are almost twice as likely to be eQTLs than random sets of SNPs (Nicolae et al., 2010). Similarly, a study by Dubois et al. concluded that 20 out of 38 (52%) risk loci for celiac disease (CeD) are eQTLs in primary immune cells (Dubois et al., 2010). However, eQTLs are abundant (Lappalainen et al., 2013), with 48% of common genetic variants estimated to act as eQTLs for at least one gene in at least one tissue (B. Liu et al., 2019), which makes the overlap between GWAS and eQTL signals likely to emerge by chance.

Additionally, these early approaches did not sufficiently control for the genetic architecture underlying complex and molecular traits, resulting in high numbers of false positives findings. In particular, the LD between nearby SNPs makes it challenging to identify which variants within a GWAS locus or a QTL are driving the associations. Consequently, the overlap between a QTL and a GWAS signal could be explained by three different scenarios: 1) two independent causal SNPs (one for the complex trait and one for the molecular trait) in LD with each other (linkage), 2) a single-causal SNP which affects the complex trait via changes in the molecular trait (causality), or 3) a single-causal SNP with independent effects on both traits (pleiotropy) (**Figure 1.3B** and **Figure 1.3C**). Distinguishing between these scenarios is crucial in order to appropriately interpret GWAS associations. This motivated the development of formal statistical tests that estimate the probability that an overlap between two association signals is due to a shared causal variant, rather than to chance or linkage. These approaches are referred to as colocalization tests.

A study by Plagnol et al. (Plagnol et al., 2009) focused on a potentially causal relationship between the 12q13 locus, associated with T1D, and the nearby gene *RPS26*. The authors

reasoned that if the locus in question increased disease susceptibility via regulation of *RPS26* expression, then the effect sizes inferred from GWAS and eQTL mapping (i.e. odds ratios and regression coefficients, respectively) should be proportional. In other words, the SNPs with the highest effects on T1D would tend to also have the highest effects on *RPS26* expression, and the direction of effects would be consistent. The authors developed a statistical test for this proportionality (*QTLmatch*) and concluded that there was no evidence for colocalization at the 12q13 locus. Subsequently, Wallace et al. revisited this approach and implemented a generalized version of the method (Wallace et al., 2012).

An alternative approach described by Nica et al. first identified loci with potential colocalizations and then regressed out the most significant GWAS SNP in each locus from the eQTL effect (Nica et al., 2010). The eQTL association was then re-tested using the residuals from this regression. To account for LD, the procedure was repeated for all the SNPs within the region and the impact of the top GWAS SNP was compared to that of all other variants. In the presence of a true colocalization, regressing the top GWAS SNP is expected to have a significantly larger impact than regressing any other variant in the region. This process was implemented into a method called *Regulatory-Trait Concordance* (*RTC*).

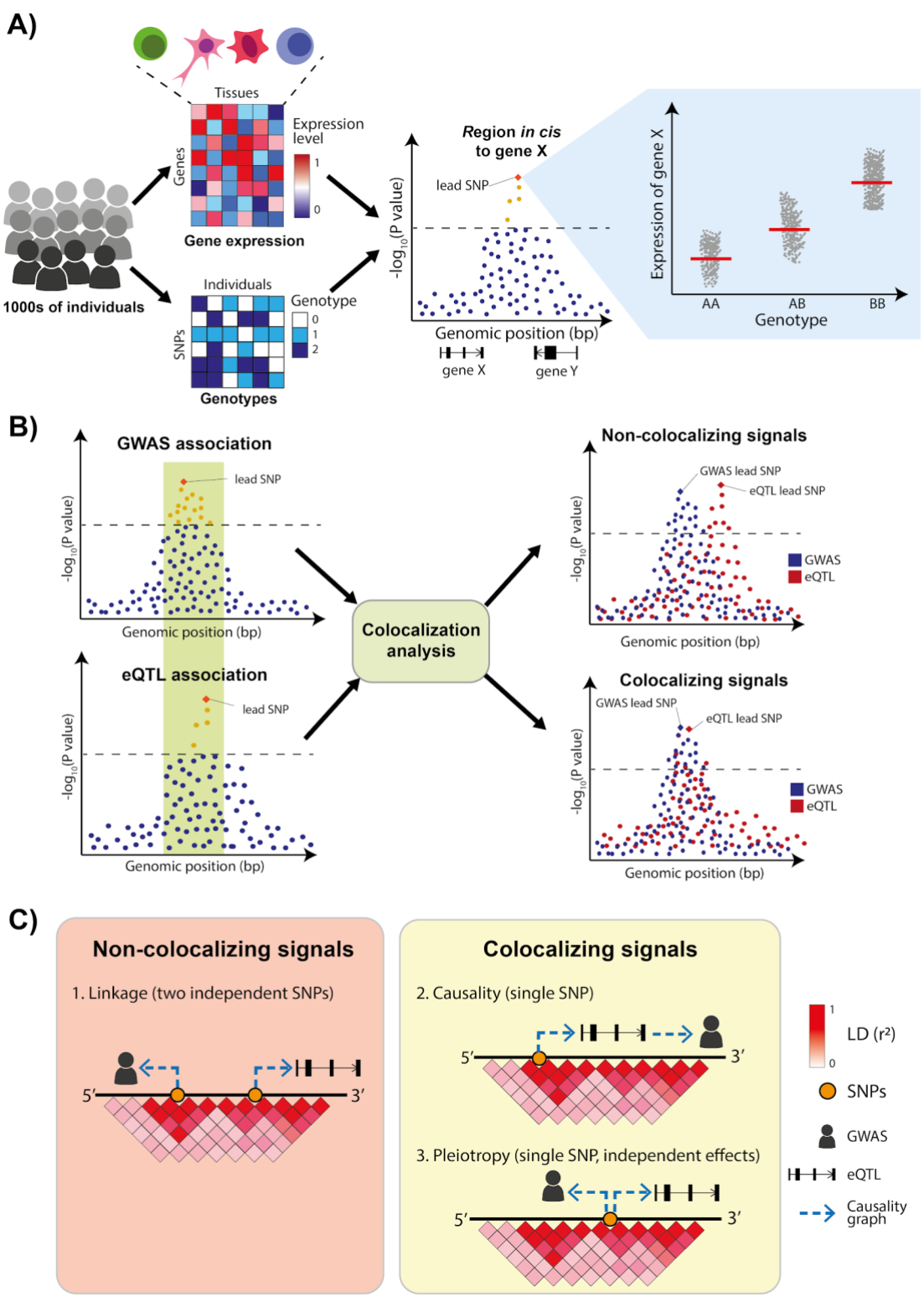



**Figure 1.3. Overview of eQTL-mapping and colocalization A)** In eQTL mapping, gene expression is profiled across thousands of individuals and the expression level of each gene is tested for association with genotypes at nearby (*cis*) SNPs. **B)** Colocalization compares the association patterns of a GWAS locus and an eQTL to find if both signals are driven by the same causal variants. **C)** GWAS and eQTL signals can overlap for three reasons: two independent causal variants in LD (linkage), a single causal variant affecting the GWAS trait via gene expression modulation (causality) or a single causal variant affecting both traits independently (pleiotropy). A positive colocalization supports causality or pleiotropy over linkage.

Despite the usefulness of these approaches, neither of the two formally compares the odds of colocalization versus a null hypothesis. Instead, they are based on the proportionality of effects or the conditional association between two traits, which can be biased by LD and by the fact that only pre-selected SNPs are being used for testing (i.e. winner's curse) (Wallace, 2013). This can result in a large proportion of false positives. Additionally, both approaches require individual-level genotype data, which is seldom available. This motivated the development of methods which could be applied to more commonly available data such as variant effect sizes and P values (i.e. GWAS summary statistics). Giambartolomei et al. proposed a test (*COLOC*) which computes the odds of colocalization compared to the null hypothesis using GWAS summary statistics (Giambartolomei et al., 2014). The authors identified five mutually exclusive scenarios at any given locus: either 1) the locus is not associated with any of the traits (the null hypothesis, $H_0$), 2) the locus is only significant in the GWAS ($H_1$), 3) the locus is only a significant QTL ($H_2$), 4) the locus is associated with both traits due to two independent signals (linkage, $H_3$) or 5) the locus is associated with both traits due to a single causal SNP (colocalization, $H_4$). The probability of each of these scenarios is estimated using a Bayesian framework, and any locus where the probability of causality ($H_4$) is significantly higher than that of linkage ($H_3$) is said to colocalize.

Since its release, *COLOC* has become a reference method for colocalization testing. However, a limitation is that it only tests for two traits at a time. In contrast, the full chain of events that connects sequence variation to organismal phenotypes is thought to involve more than one molecular trait. For example, a variant which increases DNA methylation could in turn reduce the expression of nearby genes, impairing cell function and increasing disease risk. Disentangling this chain of effects would require a joint colocalization test which leverages signals from DNA methylation, gene expression, and cell function. *MOLOC* expanded the original formulation of *COLOC* to include multiple traits (Giambartolomei et al., 2018). These traits can either be independent complex phenotypes (i.e. GWAS signals), molecular traits (i.e. QTLs) or a combination of both. To show the utility of their framework, the authors considered an example case with three traits: GWAS variants for schizophrenia, gene expression (eQTLs), and DNA methylation (mQTLs) in the human brain. They showed that adding a third trait significantly increased the power to link variants to genes, as evidenced by 39 new candidate target genes which could only be identified when integrating mQTLs with eQTLs. However, these improvements come at the expense of interpretability, increasing the number of possible hypotheses at a locus to 15. Further increases in the number of traits would make the interpretation of colocalization results even more challenging.

A trait association signal can also arise from multiple causal variants at the same locus (i.e. allelic heterogeneity, AH) and recent studies estimate that as much as 20% of the loci identified by GWAS or eQTL-mapping could show AH (Hormozdiari et al., 2017). Methods which assume a single causal variant could thus misclassify AH as colocalization (Giambartolomei et al., 2014). One method that accounts for multiple causal SNPs per locus is *eCAVIAR* (Hormozdiari et al., 2016), an approach based on statistical tools originally designed for fine-mapping. Hormozdiari et al. proposed that fine-mapping could be applied independently to GWAS and QTL associations, and then integrated (Hormozdiari et al., 2016). Specifically, they defined the probability of a colocalization as the product of the probabilities that the variant was causal in the GWAS and in the eQTL (i.e. the product of the posterior probabilities derived from fine-mapping). Because this approach estimates a posterior probability for each SNP, it does not assume a single causal variant per locus. Instead, *eCAVIAR* can be extended to find colocalizations under the assumption of any number of causal SNPs while accounting for LD.

Colocalization can also be combined with SNP enrichment analysis, as demonstrated by the statistical method *ENLOC* (Wen et al., 2017). In brief, the authors reasoned that if the majority of GWAS SNPs for a trait are also eQTLs in a given cell type (i.e. if GWAS SNPs are enriched in eQTLs), then most overlaps between the two traits will be driven by true colocalizations. In contrast, if GWAS SNPs are not enriched in eQTLs in that cell type, more of the overlaps are expected to be due to chance. Thus, the authors first estimate a SNP enrichment score and then weigh the priors of their Bayesian model by the inferred scores. The authors argue that this approach significantly improves the performance of both fine-mapping and colocalization.

**Table 1.2 Methods for colocalization analysis** Selected approaches and methods used to test for colocalization between GWAS and QTL signals included in this review.

| Method | Publication | Approach | Input data |
|---|---|---|---|
| Regulatory trait cconcordance (RTC) | Nica et al., 2010 | Conditional regression | Individual genotypes |
| Proportionality test | Wallace et al., 2012 | Test for concordance of effects | Individual genotypes |
| Sherlock | He et al., 2013 | Genome-wide comparison of association “signatures” | Summary statistics |
| COLOC | Giambartolomei et al., 2014 | Bayesian test | Summary statistics |
| gwas-pw | Pickerell et al., 2016 | Bayesian test | Summary statistics |
| eCAVIAR | Hormozdiari et al., 2016 | Bayesian fine-mapping and colocalization | Summary statistics |
| ENLOC | Wen et al., 2017 | Bayesian test for enrichment, fine-mapping and colocalization | Summary statistics |
| MOLOC | Giambartolomei et al., 2018 | Bayesian test for multiple traits | Summary statistics |

Finally, GWAS variants could regulate not only genes in close proximity, but also distal genes (*trans* eQTLs). For example, a GWAS variant could affect the expression of a transcription factor, which would result in a cascade of effects on downstream genes. *Trans* eQTLs are located far away from their target genes and tend to have small effect sizes, which makes them extremely challenging to map at moderate sample sizes due to the burden imposed by multiple testing. In addition, *trans* eQTLs are estimated to be substantially more numerous than *cis* eQTLs (X. Liu et al., 2019), potentially leading to many false positive colocalizations. However, He et al. reasoned that, while a colocalization between one *trans* eQTL and one GWAS SNP is very likely to be a false positive, the repeated presence of colocalizations between multiple *trans* eQTLs for the same gene and multiple SNPs from the same GWAS is unlikely to be due to chance (He et al., 2013). Thus, they proposed that the association signals for two traits (e.g. a complex trait and a molecular trait) could be compared not locally but genome-wide, analogously to comparing two “fingerprints” or “signatures”. If two traits tend to have the same signature, they are said to colocalize. The authors applied their method (*Sherlock*) to integrate summary statistics from a GWAS for T2D with 3,210 *cis* and 242 *trans* eQTLs specific to the liver (Schadt et al., 2008). This analysis identified four candidate genes regulated by T2D variants, two of which acted *in trans* and would thus have been missed by

traditional colocalization approaches. Importantly, three of these four genes (*TSPAN8*, *GNB5* and *JAZF1*) were supported by previous functional studies. The increasing sample sizes of gene expression studies are allowing us to systematically map *trans* eQTLs (Võsa et al., 2018; Westra et al., 2013) and will provide more statistical power to detect meaningful colocalizations between GWAS and *trans* eQTLs. A brief list of the colocalization approaches reviewed in this chapter is available in **Table 1.2**.

Colocalization analysis has been particularly informative in the study of complex immune-mediated diseases. A study by Fortune et al. used colocalization to investigate the shared etiology of four complex immune diseases (Fortune et al., 2015). The authors investigated 126 GWAS loci associated with T1D, RA, CeD, and MS. Of these loci, 33 were shared across the four traits. Colocalization analysis showed evidence of a single causal variant affecting more than one trait in at least 19 of these regions. For example, the associations at the *CTLA4* locus colocalized between three of the tested diseases. Interestingly, the authors also found three significant colocalizations between T1D and T2D loci, suggesting that these diseases could share certain aspects of their etiology, despite T1D having an immune origin.

A further study by Huang et al. fine-mapped variants associated with IBD and integrated them with eQTLs mapped in immune cells. The authors found that a large number of IBD variants colocalized with eQTLs in CD4+ T cells (H. Huang et al., 2017), corroborating the observations from SNP enrichment analyses presented in the previous section of this chapter. In contrast, a separate study tested for colocalization of risk loci for immune-mediated diseases (including IBD) with eQTLs across three immune cell types (lymphoblastoid cells, CD4+ T cells, and monocytes) and found that, surprisingly, the majority of loci did not colocalize with eQTLs (Chun et al., 2017). The authors proposed that GWAS variants could act via more complicated mechanisms and regulate molecular traits other than gene expression. This is supported by a study by Bossini-Castillo et al. which mapped QTLs for gene expression, histone modifications, and chromatin accessibility in regulatory CD4+ T cells, a rare cell type which plays a central role in suppressing the immune response (Bossini-Castillo et al., 2019). The authors integrated chromatin and gene expression QTLs with GWAS loci for 14 immune-mediated diseases, identifying 253 colocalizations. Surprisingly, the majority of these colocalization events did not implicate an eQTL, but a histone acetylation (H3K27ac) QTL (acQTL). Over 70% of these acQTLs had no eQTL effects at all, i.e. the loci were associated with local chromatin remodelling but not with the expression of nearby genes. The authors proposed that a proportion of these colocalizations could represent context-specific eQTLs, which would only be detected upon exposure of the cells to the correct environmental cues. This is known to be the case for other immune cells such as human monocytes or

macrophages, where exposure to cytokines or pathogens has been shown to induce context-specific chromatin accessibility and expression QTLs (Alasoo et al., 2018; Fairfax et al., 2014).

## Novel approaches for translating GWAS to function

SNP enrichment and colocalization analyses have prioritized cell types and genes likely to be causally involved in complex immune diseases. However, these approaches are largely limited by the availability of comprehensive reference functional data sets. In particular, most of the published studies using these methodologies have so far relied on gene expression data obtained from bulk tissues. As a consequence, their results could be dominated by the most abundant cell types and thus fail to capture information about cell composition and cell type frequencies (Trapnell, 2015). In addition, the majority of studies have focused on the effect of disease variants on RNA expression levels, while neglecting other equally important aspects of gene expression regulation such as splicing, promoter usage, or mRNA polyadenylation. Finally, colocalization methods are purely observational and cannot establish causality. For example, a SNP could affect both a gene and a trait via independent mechanisms (i.e. pleiotropy), and colocalization is unable to conclusively distinguish this scenario from causality. Thus, candidate genes require further experimental validation in order to unambiguously establish their functional role in disease. In this section, I discuss a variety of novel approaches to tackle these issues, including the possibility of quantifying molecular phenotypes at single-cell resolution, the integration of GWAS signals with traits other than RNA expression, and the validation of candidate disease genes and loci via genome editing.

### Integration of GWAS signals with single-cell molecular traits

Over the last five years, advances in the development of single-cell sequencing assays have enabled the quantification of molecular traits at the single-cell level with high accuracy. For example, it is now possible to profile gene expression (Kimmerling et al., 2016; Macosko et al., 2015; Picelli et al., 2013; G. X. Y. Zheng et al., 2017), chromatin accessibility (Buenrostro et al., 2015), and transcription factor occupancy (Grosselin et al., 2019; Rotem et al., 2015) with single-cell resolution. These assays can unbiasedly resolve the cellular composition of complex organs and tissues, and have been used to assemble cells into reference tissue atlases (Regev et al., 2017). Moreover, they make it possible to order differentiating cells into time-course trajectories that span different stages of development, an approach referred to as trajectory inference or pseudotime ordering (Saelens et al., 2019).

The high resolution of such single-cell omics data sets makes them a promising resource for SNP enrichment analysis. This is illustrated by a recent GWAS of hematological traits, including hematocrit, hemoglobin, and blood cell counts (Ulirsch et al., 2019). In this study, the authors integrated fine-mapped GWAS variants for these traits with bulk and single-cell chromatin accessibility profiles spanning a large number of hematopoietic and progenitor cell lineages. The authors developed a SNP enrichment test (*g-chromVAR*) which integrates the quantitative levels of chromatin accessibility in each single cell with the posterior probabilities of causality of each single variant, as inferred from fine-mapping. Enrichment estimates varied throughout the differentiation trajectory, concentrating at specific stages of hematopoiesis. For example, variants associated with platelet counts showed increasingly higher enrichment in open chromatin as cells differentiated into megakaryocytes, the precursors of platelets. Conversely, this enrichment decreased along differentiation towards the lymphoid lineage. With the rapid increase in number, depth and size of single-cell datasets, more studies like this will soon be possible and applicable to complex immune traits. However, single-cell sequencing approaches introduce new challenges to the current statistical methods, such as data size, sparsity and high dropout rates (Lähnemann et al., 2020). Thus, it will be essential to develop new statistical methods designed to deal with the intricacies of single-cell data.

Single-cell technologies can also expand the current scope of colocalization. Because the throughput of these assays is growing at an unprecedented scale, it is now feasible to profile single-cell transcriptomes across large populations of individuals, which makes it possible to map single-cell eQTLs (sc-eQTLs). One such study profiled gene expression in 45,000 single-cells isolated from peripheral blood of 45 healthy individuals (van der Wijst et al., 2018), revealing eQTLs which have opposite effects in different blood cell types. For example, rs4804315 increased the expression of *ZNF414* in NK cells but decreased it in T cells. Moreover, the authors also recapitulated two previously reported monocyte eQTLs for the *HLA-DQA1* and *CTSC* genes and showed that they were specific to the classical monocyte subpopulation (van der Wijst et al., 2018). These results would be difficult to obtain from bulk gene expression measurements. This study serves as a proof of concept and shows how single-cell eQTL associations could rapidly become available for integration with GWAS loci.

An additional advantage of single-cell sequencing is the possibility of ordering cells into time-course trajectories, thus adding a temporal component to the association models used for eQTL-mapping. This permits the identification of eQTLs with different effect sizes at different stages of differentiation (dynamic eQTLs). Two recent studies have mapped dynamic eQTLs during the differentiation of human induced pluripotent stem cells (iPSCs). The first study investigated iPSC differentiation into endoderm (Cuomo et al., 2020). The authors profiled

single-cell gene expression at four time points across 125 iPS cell lines and ordered cells into a time-course trajectory spanning distinct cell states. This uncovered 785 dynamic eQTLs. Interestingly, this study also identified eQTLs with cell cycle-dependent effect sizes. The second study focused on cardiomyocyte differentiation and mapped eQTLs at 16 time points across 19 iPS cell lines (Strober et al., 2019). Here, the authors ordered samples in a time-course trajectory based on bulk RNA expression profiles and identified modules of genes which increase or decrease along differentiation. Next, they performed eQTL-mapping using a Gaussian model which accounted for the interaction between genetic regulation and differentiation time. This resulted in the identification of 550 genes with linear and 693 genes with non-linear dynamic eQTL effects. Interestingly, two dynamic eQTLs which regulated the expression of *SCN5A* (a gene altered in dilated cardiomyopathy) were also GWAS variants for QRS and QT interval duration, thus suggesting that dysregulation of gene expression dynamics could have important phenotypic consequences. Until now, these approaches have not been used to study immune cells in the context of complex immune-mediated diseases. However, as the variety of cell types and sample sizes of sc-eQTL and dynamic eQTL catalogues grow, these data sets will become an increasingly important resource for identifying subtle changes in gene expression dynamics which lead to immune-mediated disease.

### Integration of GWAS signals with molecular traits other than gene expression

So far, most colocalization studies have assessed the effect of disease variants on the mRNA levels of genes *in cis* to them. However, gene expression regulation is achieved by a series of processes which include chromatin remodelling, RNA transcription, splicing, polyadenylation, RNA degradation, and protein translation, many of which can be independent from the absolute levels of mRNA. All of these could play a role in human disease. In a landmark study, Li and colleagues quantified a series of traits spanning most steps in the gene expression regulation cascade, successfully mapping QTLs for chromatin accessibility, histone acetylation, RNA transcription rate, steady state RNA levels, splicing, translation, and protein levels (Y. I. Li et al., 2016). This study concluded that, while most genetic variants have effects which cascade from one layer of regulation to the next one, many QTL effects are not detectable when only measuring mRNA levels. In particular, the authors identified a large number of splicing QTLs (sQTLs) which do not manifest as eQTLs. As opposed to eQTLs, which are enriched near TSSs, sQTLs localise to gene bodies and are often more than 10 kb away from the lead eQTL signal for the gene in question, suggesting that they act via an independent mechanism. Surprisingly, around 17% of chromatin QTLs had sQTL activity but did not affect RNA levels, thus suggesting a mechanism by which chromatin remodelling can regulate splicing (Y. I. Li et al., 2016).

This has motivated the study of alternative splicing and in particular the elucidation of its role in human diseases. The study by Li et al. already demonstrated that sQTLs are enriched in disease loci from GWAS, with a p value distribution inflated even above that of eQTLs (Y. I. Li et al., 2016). Similar observations have since been confirmed by others, reinforcing the belief in a previously underappreciated role for splicing in complex traits. For example, a colocalization and TWAS study based on data from the GTEx project revealed that approximately 30% of eQTLs and 15% of sQTLs are predicted to share a causal variant with complex human diseases (Barbeira et al., 2021). A separate study which mapped sQTLs using a model of *in vitro* stimulated macrophages also estimates a large contribution of sQTLs to disease, with splicing accounting for up to 50% of colocalizations with immune disease associations from GWAS (Alasoo et al., 2019). Importantly, there is growing evidence that sQTLs can be cell type specific (Kim-Hellmuth et al., 2020), a phenomenon of particular importance to tissues with highly complex splicing patterns, such as the nervous system (Raj & Blencowe, 2015). This is supported by a recent study which mapped sQTLs in the human prefrontal cortex (Takata et al., 2017). The authors found an enrichment of brain sQTLs in GWAS loci for schizophrenia and proposed a molecular mechanism for three loci associated with the disease.

The studies discussed so far strongly suggest alternative splicing as a key mechanism linking genetic and phenotypic variation. However, alternative splicing can have different functional impacts on the affected transcript. In particular, while splicing often results in differential inclusion of exons within the gene's body, it can also manifest as a change in promoter usage (Landry et al., 2003) or as inclusion or excision of polyadenylation sites, which in turn modify the length of the transcript (B. Tian et al., 2007). A detailed understanding of the different consequences of splicing at population scale has been limited by the fact that most studies are based on RNA-sequencing with short read technologies, which are often unable to reconstruct full transcripts. However, recent developments in long-read technology have resulted in the characterization and annotation of thousands of new transcripts (Amarasinghe et al., 2020; Glinos et al., 2021), and computational methods such as pseudo aligners (Bray et al., 2016; Patro et al., 2017) and tools for the quantification of reads spanning exon-exon junctions (Y. I. Li et al., 2018) have made it possible to study splicing in more depth. A recent study used a combination of these tools to assess the relevance of different splicing events in human disease (Alasoo et al., 2019). The authors found that promoter usage and 3' end splicing are often regulated by different genetic variants, and thus follow independent mechanisms. The two categories also showed different contributions to disease, with promoter

usage QTLs (puQTLs) accounting for approximately 25% of the observed colocalizations with GWAS signals for immune mediated diseases.

In addition to splicing, quantifying other mechanisms of gene expression regulation could help interpret a larger proportion of disease associations. It is estimated that as many as 30% of eQTLs are not mediated by a chromatin QTL, the majority of which concentrate within gene bodies and could thus act by disrupting post transcriptional regulation (Y. I. Li et al., 2016). An interesting example are RNA decay QTLs, which tend to overlap binding sites for microRNAs (miRNAs), in turn regulating RNA degradation (Pai et al., 2012). In addition, two recent studies have assessed the role of genetic variants in mRNA polyadenylation. One study used 3' RNA-seq to identify alternative polyadenylation QTLs (apaQTLs) (Mittleman et al., 2020). The authors found over 600 apaQTLs and showed that they often manifest as protein QTLs (pQTLs) in an eQTL-independent manner. Moreover, 19% of these were linked to complex diseases. Similarly, a second study used RNA-seq data from the GTEx project to map apaQTLs across human tissues (L. Li et al., 2021). The authors estimate that apaQTLs could explain close to 16% of trait associated variants from GWAS, suggesting a disease mechanism which is independent of absolute mRNA transcription.

In summary, while RNA expression (eQTLs) is a key mechanism linking non-coding variants to complex phenotypes, there is a growing body of evidence suggesting that other gene expression regulation mechanisms could be equally important. In particular, splicing could account for as much as half of colocalizations with disease, often in a cell type specific manner. Additionally, alternative polyadenylation and RNA decay could also be disease relevant. Further developments in long-read technology and larger studies designed to assess the role of genetic variation in each of these processes could help bridge the gap between genetic associations and functional understanding of human phenotypes.

### Integration of polygenic risk scores with molecular and intermediate traits

In addition to revealing disease-associated cell types and genes, GWAS variants can also be used to identify individuals at high risk of disease. This can be achieved by combining the hundreds of disease associated-variants carried by an individual into a single score that reflects their overall genetic risk, a polygenic risk score (PRS) (Chatterjee et al., 2016). The integration of PRSs with epidemiological risk factors such as age, sex, smoking status, diet, or family history of disease could improve the stratification of individuals, potentially resulting in more effective clinical interventions (Torkamani et al., 2018). To build a PRS, a subset of variants is selected based on their GWAS association. Next, each variant is assigned a weight, which corresponds to its standardized effect size (i.e. the odds ratio from the GWAS multiplied

by the effect direction). Finally, the genetic dosage of each individual variant (i.e. 0, 1, and 2 according to the number of risk alleles carried) is multiplied by its weight, and all loci across the genome are added into a single score. PRSs are often normally distributed and individuals can be grouped by PRS decile, with those in the top deciles being at highest risk (Chatterjee et al., 2016).

PRS performance has improved as GWAS increased in sample sizes and larger validation cohorts became available, as shown in coronary artery disease (CAD) (Abraham et al., 2016; Khera et al., 2016; Mega et al., 2015; Ripatti et al., 2010) and cancer (Garcia-Closas et al., 2013; Maas et al., 2016; Mavaddat et al., 2015). The availability of large-scale biobanks (Bycroft et al., 2018; Gaziano et al., 2016; Nagai et al., 2017) has enabled unparalleled improvements in this area by linking genetic information with electronic health records for hundreds of thousands of individuals. For example, two of the largest PRS studies available to date leveraged UK BioBank data to estimate CAD risk using up to 6.6 million SNPs (Abraham et al., 2016; Khera et al., 2018). In one of these studies, Khera et al. demonstrated that individuals at the highest PRS percentiles were at a risk equivalent to that of carrying a monogenic mutation for familial hypercholesterolemia (Khera et al., 2018). Another related study used 2.1 million SNPs to build an obesity PRS (Khera et al., 2019), demonstrating that PRSs can stratify individuals before phenotypic differences appear. In particular, while the authors observed no differences in the birthweight of individuals at different PRS deciles, these became apparent when individuals reached puberty.

In addition to their clinical utility in risk prediction and patient stratification, PRSs are powerful tools for studying the functional mechanisms underlying complex traits. In particular, the theory of omnigenic inheritance proposes that polygenic risk arises from the contribution of thousands of genetic variants evenly distributed across the genome (E. A. Boyle et al., 2017). Of these variants, only a selected few with strong effects are thought to directly regulate disease-causal genes (Sinnott-Armstrong et al., 2020). In contrast, the vast majority of variants that drive polygenic risk are thought to act as weak *trans* eQTLs, having only a secondary effect on disease-causal genes (X. Liu et al., 2019). However, despite their effects being too small to study them individually, these *trans* eQTLs could accumulate in disease-relevant pathways and networks (X. Liu et al., 2019), making it possible to use polygenic scores as a means to reveal the biological networks altered in disease.

The availability of large scale biobanks will be instrumental in furthering our functional understanding of polygenic risk. By collating electronic health records, clinical measurements, and genetic data into a single database, biobanks make it possible to systematically test for

the association of a PRS with a large number of organismal phenotypes. For example, Richardson et al. used GWAS variants and UK BioBank data to build 162 PRSs, spanning traits as varied as anthropometric measurements, cardiovascular traits, and ICD10 diagnostic codes (Richardson et al., 2019). They identified traits correlated with each other based on their polygenic scores and used Mendelian randomization (MR) to infer causality. Polygenic scores for triglyceride levels, urate levels, LDL, and gout were significantly correlated with each other, and MR analysis revealed that elevated triglycerides are causally related to higher urate production, which in turn increases the risk of gout. A similar study derived PRSs for quantitative blood traits such as hematocrit and cell counts (Xu et al., 2020) and correlated them with PRSs for complex diseases. This pinpointed disease-relevant traits such as eosinophil counts, which were associated with polygenic risk for allergies.

Gene expression measurements can also be integrated with PRSs. Vosa et al. mapped *cis* and *trans* eQTLs in a meta-analysis of 31,684 samples from 37 cohorts (Võsa et al., 2018). They subsequently identified genes affected by dozens of *trans* eQTLs and proposed that such genes could act as hubs where biological processes converge, potentially accumulating a disproportionate amount of genetic risk for disease. To identify these hubs, the authors mapped quantitative trait scores (QTS), which they defined as a significant association between the expression of a gene and the PRS for a disease. They identified 2,658 eQTS genes, including a group of IFN-regulated genes associated with polygenic risk for lupus. In the future, increases in the sample size of eQTL studies may enable the systematic identification of cell type specific eQTSs in the context of complex immune-mediated diseases.

### Dissection of GWAS loci using genome editing

The last decade has seen a rapid expansion in the number and efficacy of genome editing tools. In particular, CRISPR/Cas9 allows the disruption of specific sections of the genome with high accuracy (Wang et al., 2016). CRISPR-based approaches have been used to systematically knock out genes genome-wide, an approach referred to as CRISPR screening (Koike-Yusa et al., 2014). The applications of CRISPR screening are numerous. For example, it can be used to investigate which genes are essential for cancer growth, which in turn provides a platform for drug target identification (Behan et al., 2019).

Coupling CRISPR-editing platforms with disease-relevant functional readouts can be a powerful approach to functionally translate GWAS associations. For example, a recent study asked which genes were essential for T cell activation by systematically knocking out all genes in primary human CD8+ T cells and measuring cell proliferation upon stimulation (Shifrut et al., 2018). A second study used a similar approach to investigate T helper cell differentiation

in mice (Henriksson et al., 2019). These studies are relevant in the context of complex immune-mediated diseases, for which GWAS variants are thought to act during T cell activation and differentiation (Calderon et al., 2019).

Nonetheless, using CRISPR to follow-up candidate genes requires previous knowledge regarding which functional assays are the most disease-relevant. For example, neuronal cell types are thought to be implicated in psychiatric traits (Finucane et al., 2018), but it is not known which specific neuronal functions are compromised in disease, and thus it is uncertain what the best readout for a CRISPR-screen would be. Selecting informative assays may require mapping the genetic architecture of cellular and intermediate traits. As an example, a recent study showed that variants which modulate secretion of monocyte cytokines (cytokine-QTLs) tend to be associated with susceptibility to infection (Y. Li et al., 2016). Thus, a CRISPR-screen to validate infection susceptibility genes could assess cytokine secretion.

Alternatively, single-cell gene expression can also be used as a readout for CRISPR-screens. Due to its high resolution, single-cell sequencing is able to match the transcriptome of a cell with the guide RNA sequence used to edit it. This is the basis of methods like CROP-seq and Perturb-seq (Datlinger et al., 2017; Dixit et al., 2016), which have been used to investigate which genes are essential in processes such as dendritic cell response with single-cell resolution. In the future, high-throughput phenotyping of human cells will be crucial for identifying the best assays to validate candidate genes.

Genome-editing approaches can also be used to study the non-coding genome. Because the target genes of many regulatory elements are unknown, approaches to dissect the non-coding space have centred around the idea of cloning regulatory elements into vectors specifically designed to assess their influence on the transcription of a reporter gene (Inoue & Ahituv, 2015). This in turn makes it possible to edit such elements and evaluate the impact of each nucleotide position on transcription. Massively parallel reporter assays (MPRAs) are a prominent example (Melnikov et al., 2012). In MPRAs, putative regulatory elements are cloned into plasmids containing an inert open reading frame (ORF) and a unique nucleotide tag. Sequence libraries built in this way can then be transfected into relevant cell types, where RNA can be sequenced and nucleotide tags counted. Tag counts are then compared to their expected values based on the library composition, which enables the identification of elements which enhance or repress transcription (Melnikov et al., 2012). Several variations of this method exist (Kinney & McCandlish, 2019). For example, s*elf-transcribing active regulatory region sequencing* (STARR-seq) is a technique which leverages the ability of enhancers to regulate transcription regardless of their position or orientation (Arnold et al., 2013). In this

method, enhancers are cloned into a vector containing a minimal ORF and a polyadenylation site, where they are positioned downstream of the TSS. This causes them to be transcribed as a part of the ORF, which in turn makes it possible to clone the libraries into relevant cell types and quantify the transcriptional activity of each construct based on RNA-seq, by quantifying the number of reads which contain the enhancer sequence of interest (Arnold et al., 2013; Muerdter et al., 2015). These methods can be used to characterise non-coding elements and assess the effect of different sequence features on regulatory activity.

Several recent studies have used the MPRA methodology to dissect GWAS variants located within regulatory elements. One such study investigated loci associated with hematological traits such as red blood cell (RBC) counts using fine-mapping followed by validation with an MPRA (Ulirsch et al., 2016). Of the 75 loci analysed in this study, the authors managed to identify 32 functional variants which alter the activity of a regulatory element. This represents 30% of the loci investigated. Importantly, the majority of variants prioritised were found to fall within ChIP-seq peaks for the erythroid transcription factor GATA1, thus suggesting a disease mechanism where transcriptional programs downstream of GATA1 could be dysregulated in disease. This MPRA was performed in an erythroid cell line, since erythroid progenitors are the cell type most enriched in variants associated with RBC traits. This illustrates how SNP enrichment analysis can be used to select the most suitable models for experimental validation.

One of the limitations of MPRAs is that most assays have so far been performed in cell lines, which can differ substantially from the tissues involved in disease. Nonetheless, experimental advances such as the use of a more diverse set of promoters in vector design are enabling the use of this methodology in primary tissues. For example, in the context of immune-mediated diseases, a recent study adapted the MPRA methodology to the study of primary CD4+ T cells (Bourges et al., 2020). This was achieved by substituting the minimal promoter commonly used in MPRA vectors with a Rous sarcoma virus (RSV) promoter, which achieves better transcription in T cells. The authors applied this MPRA methodology to dissect 14 loci associated with immune-mediated diseases located at gene deserts (Bourges et al., 2020) This resulted in successful identification of candidate functional variants for loci which cannot be fine-mapped due to the presence of many variants in perfect LD. Most strikingly, a functional variant was identified at a locus associated with six different immune diseases. This variant was found to reduce the enhancer activity of the locus via disruption of an NFkB binding site, which in turn prevents the formation of a super enhancer. The authors then used CRISPR/Cas9-mediated editing and identified *TNFAIP3*, a negative regulator of NFkB, as the target gene of this enhancer. In individuals carrying a risk allele, levels of *TNFAIP3* are lower,

resulting in increased NFkB signaling and uncontrolled T cell activation (Bourges et al., 2020). These observations illustrate how gene expression dysregulation during T cell activation can alter cellular phenotypes, ultimately increasing disease risk.

These studies demonstrate the potential of MPRAs for dissecting the role of non-coding GWAS variants. However, MPRAs suffer from a number of limitations. Firstly, they are only informative if carried out in relevant cell types, and they assess the activity of non-coding elements outside their original genomic contexts. Secondly, they rely on DNA synthesis and are thus limited in the number of perturbations that can be studied at a time. The use of CRISPR-based technologies as an additional layer of information could bypass the first of these limitations by perturbing regulatory elements in their endogenous environment. For example, CRISPR-interference (CRISPRi) uses guide RNAs and a modified version of the Cas9 enzyme (either catalytically dead or fused to a repressor of transcription) to disrupt enhancer-promoter contacts (Qi et al., 2013). In contrast, CRISPR-activation (CRISPRa) uses a transcriptional activator fused to a catalytically dead Cas9 protein to enhance transcription (Bikard et al., 2013). These tools can be used to map the function of disease-associated regulatory elements and identify their target genes. In addition, deep mutagenesis, a method which employs error-prone PCR to randomly mutate all the nucleotides in a regulatory sequence one at a time (McCullum et al., 2010), could circumvent the second limitation. For example, a recent study used deep-mutagenesis followed by sequencing to study the function of each nucleotide in 20 regulatory elements associated with rare and common diseases (Kircher et al., 2019), including the well-known LDL-associated locus near the *SORT1* gene (Musunuru et al., 2010). This enabled the identification of clusters of nucleotides for which mutation significantly alters gene expression. Importantly, these sites often contained known GWAS SNPs and corresponded to transcription factor binding sites, thus suggesting a molecular mechanism for the implicated variants.

In summary, genome editing can be used to dissect the impact of GWAS variants by perturbing either coding or non-coding sequences. These approaches have already revealed key insights into disease biology. However, some challenges remain. In particular, genome editing should be performed in disease-relevant cell types (for example, in cells prioritized by SNP enrichment analysis), but current approaches are mostly limited to cell lines. The reasons for this are varied. The application of mutagenesis to primary cells is hindered by the large numbers of cells required and the need to keep cells in culture for prolonged periods of time, while CRISPR-editing is limited by the p53-dependent cellular toxicity which accompanies Cas9-induced double-strand breaks (Ihry et al., 2018). Methodological advances such as better systems for Cas9 delivery (DeWitt et al., 2017; Shifrut et al., 2018) and platforms such

as base editors, which do not depend on double-strand breaks to edit (Anzalone et al., 2019; Gaudelli et al., 2017; Komor et al., 2016), will likely overcome some of these limitations. However, further technological development is needed to routinely apply genome editing as a follow up strategy for GWAS.

## Conclusions

The integration of GWAS associations with cell type-specific functional data sets has significantly furthered our understanding of how genetic variation leads to increased susceptibility to immune-mediated diseases. Firstly, SNP enrichment approaches have enabled the prioritization of cell types and tissues based on their disease-relevance. These methods work by testing for the accumulation of variants in regulatory elements specific to a given cell type, and have revealed that immune disease risk variants tend to accumulate in regions of the genome specifically active in cells of the adaptive immune system, particularly in CD4+ T cells. Importantly, this enrichment is higher in activated than in resting T cells. Secondly, colocalization analysis integrates eQTLs and GWAS loci to identify disease causal genes, leveraging LD information and association patterns. These approaches have revealed genes involved in complex immune-mediated diseases such as RA and T1D. Importantly, a number of these candidate genes seem to only be expressed under specific cellular contexts such as immune cell activation or the cellular response to pathogens. However, colocalization is limited by the resolution of currently available functional datasets and cannot establish causality. This could be solved by the integration of GWAS signals with molecular traits for a wider variety of cell types and cell states, as well as with single-cell molecular traits. Additionally, genome editing, high-throughput phenotyping, and methods that explicitly model polygenic risk will help translate GWAS signals for immune-mediated diseases into clinically actionable gene sets.

## Thesis layout

Based on the observations described in the previous sections of this chapter, which show that immune disease risk variants accumulate in regulatory DNA active in CD4+ T cells, here I focus on the identification of the T cell states, genes and functions altered in disease. Chapter 2 presents a novel statistical method to disentangle the enrichment of GWAS variants in regulatory elements when these elements are shared between closely related cell states. When applied to epigenetic profiles from a variety of cytokine-induced CD4+ T cell states, this method reveals that polygenic risk for immune-mediated diseases accumulates during the early stages of memory CD4+ T cell activation. Chapter 3 describes how CD4+ T cells respond to stimulation in the presence of cytokines, based on observations from single-cell and bulk

gene expression. These results reveal that CD4+ T cells form a transcriptional continuum which spans the naive and memory phenotypes, and that cells at different stages of this continuum show different responses to stimulation. Chapter 4 describes the generation of a CD4+ T cell eQTL map with single-cell resolution, specifically designed to capture the temporal changes that accompany T cell activation. Colocalization of these eQTLs with GWAS loci for immune-mediated diseases demonstrates that immune disease risk variants can alter gene expression dynamics during specific stages of CD4+ T cell activation. Finally, chapter 5 summarises the overall findings from this dissertation and discusses their implications in the broader context of complex trait genetics.

# Immune disease risk loci accumulate in promoters and enhancers involved in CD4+ T cell activation

*A modified version of this chapter has been previously published in Nature Genetics (Soskic, Cano-Gamez et al., 2019), as detailed in the Publication Note.*

## Overview

Studying the role of risk variants for complex immune-mediated diseases is challenging, as the majority of loci mapped through GWAS reside in non-coding regions of the genome. Previous studies have addressed this by integrating GWAS loci with regulatory elements such as regions of open chromatin or regions tagged by histone modifications (Chen et al., 2016; ENCODE Project Consortium, 2012; Finucane et al., 2018; Maurano et al., 2012; Trynka et al., 2013), which suggests that these loci are involved in gene expression regulation. Moreover, the integration of GWAS signals with cell type specific chromatin marks has revealed disease-relevant cell types (Maurano et al., 2012; Trynka et al., 2013, 2015). These approaches have proposed that CD4+ T cells (Farh et al., 2015; Trynka et al., 2013) and monocytes (Alasoo et al., 2018; Fairfax et al., 2014) are involved in the pathobiology of complex immune-mediated diseases. Nonetheless, the functional impact of non-coding GWAS loci can be specific not just to a cell type, but also a cell state, such as different stages of cell activation (Fairfax et al., 2014). Thus, understanding the molecular processes that underlie susceptibility to immune-mediated diseases requires profiling closely-related cell states within the T cell and monocyte populations.

CD4+ T cells are key regulators of the immune response and are crucial in the protection against pathogens. One of the hallmarks of CD4+ T cells is their plasticity, which stems from their ability to differentiate into a range of cell states in response to environmental signals (Federica Sallusto, 2016). CD4+ T cells undergo activation when they recognise an antigen displayed by antigen-presenting cells (APCs) in the context of class II major histocompatibility complex (MHC) molecules and co-stimulatory signals (Guermonprez et al., 2002). Subsequently, activated T cells undergo proliferation and can differentiate into distinct T helper (Th) phenotypes in response to cytokines (Zhu et al., 2010). The best defined Th phenotypes are Th1, Th2, Th17, and induced regulatory T cells (iTregs), each of which exerts different functions in the immune response (Fantini et al., 2004; Harrington et al., 2005; Mosmann et al., 1986). Effector Th phenotypes are defined by the cytokines they secrete, which in turn instruct other immune cells to acquire different functions. For example, the Th1 cytokine IFN-γ polarizes macrophages to a proinflammatory (M1) phenotype with increased pathogen killing

ability, while the Th2 cytokine IL-4 induces a tissue remodelling macrophage phenotype (M2) (Martinez & Gordon, 2014). As such, the differentiation of T cells and macrophages following cytokine signals are interrelated and represent a crucial step in eliciting an appropriate immune response.

Although it is established that immune disease variants are enriched in chromatin regions specific to CD4+ T cells and monocytes, it is not yet known if this enrichment is specific to a particular cytokine-induced Th cell state or monocyte-derived macrophage population. In this chapter, I describe the results of a study carried out by myself and others (as specified in the collaboration note) to identify whether immune disease risk variants regulate cellular responses to cytokine polarization. We profiled chromatin accessibility as well as active enhancers and promoters in naive CD4+ T cells, memory CD4+ T cells, and monocyte-derived macrophages across 55 cell states which comprised early and late responses to activation and cytokine polarization. Furthermore, we proposed a novel statistical method to assess the enrichment of GWAS SNPs in cell state-specific chromatin marks. Results from this study show an accumulation of immune disease risk variants in promoters and enhancers involved in the early activation of memory CD4+ T cells.

## Results

### Experimental design

GWAS studies have suggested that CD4+ T cells are at the centre of immune disease pathobiology. Some of the key ways in which CD4+ T cells regulate the immune response include their initial activation and differentiation and their subsequent interaction with downstream effector cells such as macrophages, whose activity is regulated by T cell-derived factors. In this study we focused on dissecting the role of immune disease risk variants in regulating this circuitry.

Firstly, we differentiated monocytes from peripheral blood into the macrophage lineage (i.e. monocyte-derived macrophages). Following differentiation, we stimulated monocyte-derived macrophages with T cell derived cytokines associated with inflammation and autoimmunity, including IFNγ, TNFα, IL-4, IL-23, and IL-26 (**Figure 2.1A** and **Supplementary Table 2.1**). Given that macrophages are part of the fast-responding innate immune system, we profiled their chromatin landscape at six hours and 24 hours following stimulation, corresponding to their early and late responses to polarization.

Secondly, we mimicked T cell activation *in vitro* by isolating CD4+ T cells and delivering T cell receptor (TCR) and costimulatory (CD28) signals using beads coated with anti-CD3 and anti-

CD28 antibodies. In addition, we exposed CD4+ T cells to cytokine cocktails that promote differentiation towards the Th1, Th2, Th17 or iTreg cell fates, as well as to individual cytokines relevant to autoimmunity (IL-10, IL-21, IL-27, TNFα, and IFNβ) (Feldmann, 2002; Hall & Rosen, 2010; Khor et al., 2011; Mascanfroni et al., 2013; Yoshizaki et al., 2012) (**Supplementary Table 2.1** and **Methods**). These stimuli were applied independently to naive and memory CD4+ T cells, the two major T cell subsets, since they have been shown to differ in their response to stimulation (Glinos et al., 2018). Given that the response to stimulation also develops over time (Ye et al., 2014), we profiled T cells during both early and late activation, which we defined as 16 hours (prior to the first cell division) and five days (when T cells have acquired a fully effector phenotype).

For both CD4+ T cells and macrophages, we profiled active enhancers and promoters marked by H3K27 acetylation using ChIPmentation-seq (Schmidl et al., 2015), and chromatin accessibility using ATAC-seq (Buenrostro et al., 2013) (**Figure 2.1A**). We then integrated these chromatin profiles with disease-associated variants to identify the most disease-relevant cell states.

### Stimulation alters the chromatin landscape of immune cells

We used H3K27ac and ATAC-seq reads to define active chromatin elements (peaks) in the profiled cell types. We detected an average of 25,613 H3K27ac peaks and 26,290 ATAC-seq peaks in T cells, and 34,708 H3K27ac peaks and 73,474 ATAC-seq peaks in macrophages (**Supplementary Figure 2.1B**). Importantly, the number of peaks detected in T cells increased after stimulation (**Supplementary Figure 2.1C**). Moreover, the identified peaks showed the expected characteristics of a high quality epigenomic map. Firstly, they were highly reproducible across biological replicates (**Supplementary Figure 2.1D** and **Supplementary Figure 2.1E**). Secondly, H3K27ac peaks were substantially larger than ATAC-seq peaks, with median peak sizes being 3,205 and 410 bp, respectively (**Supplementary Figure 2.2A**). Thirdly, a larger proportion of ATAC-seq peaks concentrated at transcriptional start sites (TSS) compared to H3K27ac, reflecting that ATAC-seq peaks are more enriched in promoter regions (**Supplementary Figure 2.2B** and **Supplementary Figure 2.2C**). Fourthly, ATAC-seq peaks tagged mostly introns and promoters, while H3K27ac peaks were present in introns and intergenic regions (**Supplementary Figure 2.2C**). Finally, most H3K27ac peaks overlapped one or more ATAC-seq peaks (**Supplementary Figure 2.2D**). Taken together, these observations showed that our maps of chromatin activity were high quality and appropriately captured regulatory elements across the profiled cytokine-induced cell states.

We next investigated the effects of stimulation on the chromatin landscape of T cells and macrophages. We observed that the main factors explaining epigenetic variation were cell type and stimulation (**Supplementary Figure 2.3**), with stimulation inducing thousands of peaks across all cell types (**Figure 2.1B**). However, only a small proportion of peaks were specific to any individual cell state. In contrast, the majority of peaks were shared by many cell states (**Figure 2.1C**). For example, only 2% of H3K27ac and 0.8% of ATAC-seq peaks in T cells were condition specific.

Having shown that the majority of peaks detected in a cell type are present in many of the assayed conditions, we asked if their quantitative levels of activity tended to vary between cell states. We assessed the variability in levels of activity of each regulatory element across cell states by quantifying the coefficient of variation of reads in peaks. We found that quantitative levels of chromatin activity were highly variable across different cytokine-induced cell states. In particular, both ATAC and H3K27ac peaks showed a continuous spectrum of variability (**Figure 2.1C**). The most variable peaks were in proximity to genes involved in cell differentiation, while lowly variable peaks were in proximity to genes involved in metabolic processes (**Supplementary Figure 2.2E**). Interestingly, the levels of chromatin accessibility were less variable than those of H3K27ac. For instance, loci harbouring hallmark genes for cytokine-induced cell states showed marked differences in their H3K27ac (**Figure 2.1D**) but not in their ATAC-seq profiles (**Supplementary Figure 2.4**). This is in line with other reports showing that chromatin activity marked by H3K27ac is more informative in discriminating between closely related cell states than chromatin accessibility (Minoux et al., 2017; B. Mueller et al., 2017)

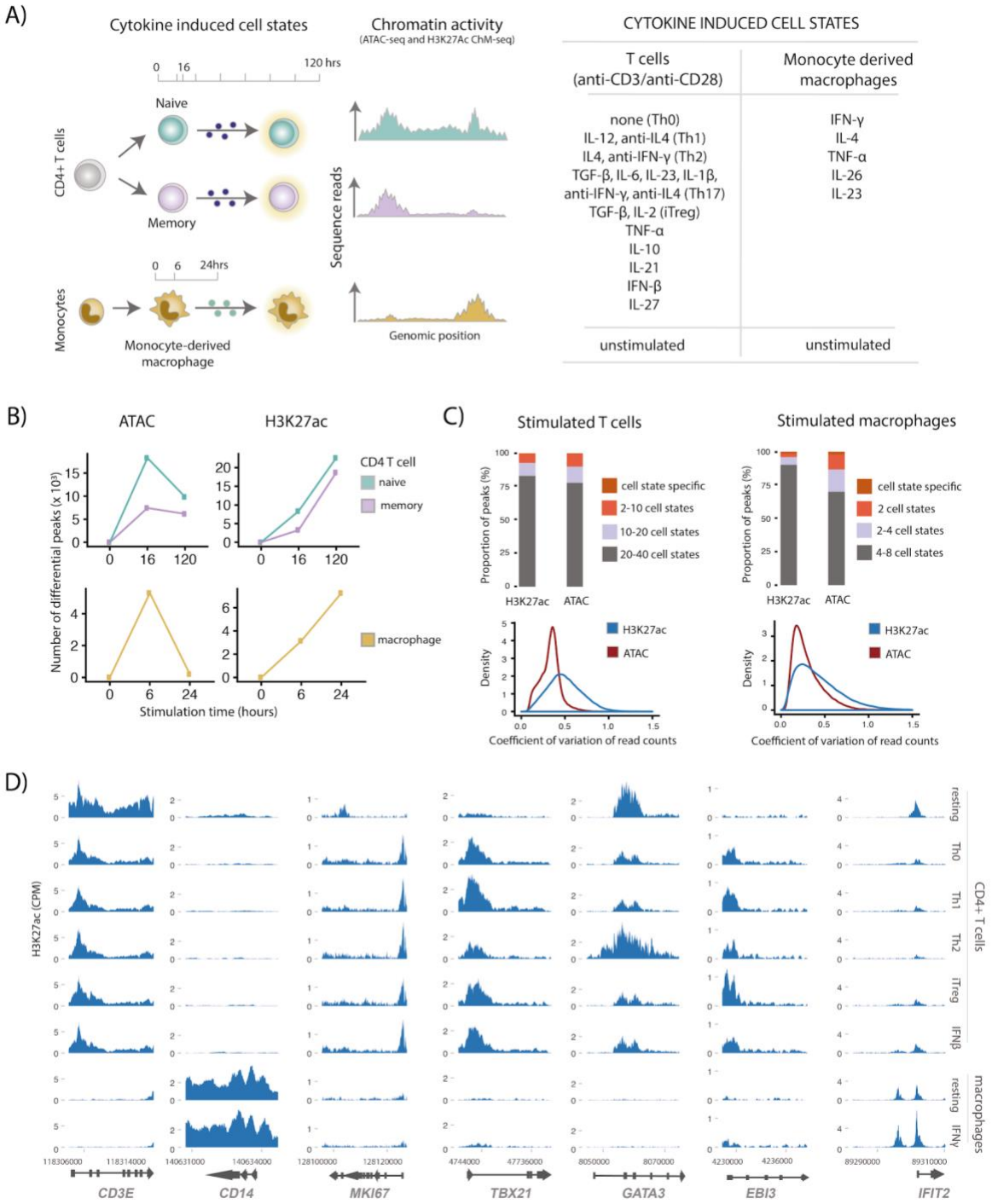



**Figure 2.1. Quantitative changes in chromatin activity distinguish cytokine induced cell states.** **A)** Overview of the study design**. B)** Number of differential H3K27ac and ATAC peaks upon T cell activation and macrophage stimulation with TNFa. **C)** Proportion of activation-induced peaks shared between all macrophage or T cell states. Different shades of gray represent the extent of sharing. Density plots show the coefficient of variation of reads within ATAC and H3K27ac peaks. **D)** H3K27ac read pile-ups in proximity to condition-specific hallmark genes: CD3 (T cells), CD14 (macrophages), MKI67 (also known as KI-67; T cell activation), TBX21 (Th1), GATA3 (Th2), IL35B (also known as EBI3; activated by iTregs), IFIT2 (IFN-induced). Genomic coordinates (GRCh38) for each gene are labeled. H3K27ac tracks were generated after merging three biologically independent samples per cell state. CPM, counts per million.

## Accounting for peak properties refines GWAS SNP enrichments

We next assessed if immune disease risk variants were enriched in chromatin elements specific to any of the cell states in our data set. Typically, GWAS SNP enrichment analyses rely on the presence or absence of an overlap between associated variants and regulatory regions (Maurano et al., 2012; Pickrell, 2014; Trynka et al., 2015). However, in our data set the majority of the peaks were present in many cell states and therefore such a binary SNP-peak overlap would be unsuitable to discriminate between the different cellular conditions (**Supplementary Figure 2.5**). A similar observation was previously made using partitioning heritability analysis of neuropsychiatric and metabolic disorders in highly correlated chromatin annotations of different brain regions (Finucane et al., 2018). Thus, we developed a new SNP enrichment method, *CHEERS* (Chromatin Element Enrichment Ranking by Specificity). In addition to SNP-peak overlaps, our method takes into account the quantitative levels of read counts within peaks in one cell state compared to others, which reflects the variability in the level of H3K27ac or chromatin accessibility (**Figure 2.2** and **Methods**).

Briefly, we constructed a matrix of quantile normalized read counts across peaks (**Figure 2.2A**). For each peak, we then defined specificity scores across all cell states. A high specificity score means that the peak has a higher read count in that cell state compared to the others. Within each cell state, peaks were then ranked based on their specificity scores (specificity rank; **Figure 2.2B**). To assess disease SNP enrichment across the different cell states, we used the index variant at each GWAS locus and identified variants in strong LD ($R^2$>0.8). Next, we identified the peaks that overlap these variants (**Figure 2.2C**) and retrieved their specificity ranks. Importantly, our method is peak-centric, which means that a peak that overlaps multiple disease associated variants in a locus contributes to the calculation only once. Alternatively, if the associated variants within a locus intersect multiple peaks, each independent peak contributes to the final enrichment score. Finally, we calculated the mean specificity rank of all peaks that intersect disease associated variants and tested if it was higher than expected by chance by comparing it to a theoretical distribution (**Methods**).

We used simulations to assess the sensitivity of *CHEERS* (**Methods**). These simulations showed that, when more than 23% of the GWAS loci overlap peaks in the top 10th percentile of specificity, *CHEERS* has over 80% power to detect a significant enrichment (p-value < 0.01). However, if GWAS loci overlap peaks with lower specificity ranks (e.g. among the top 60th to 70th percentile by specificity), then over 80% of the loci need to overlap the peaks for *CHEERS* to detect a significant enrichment (**Supplementary Figure 2.6**). This indicates that *CHEERS* can detect a significant cell type enrichment only when a sufficient proportion of trait associated variants overlap peaks with high cell type specificity.

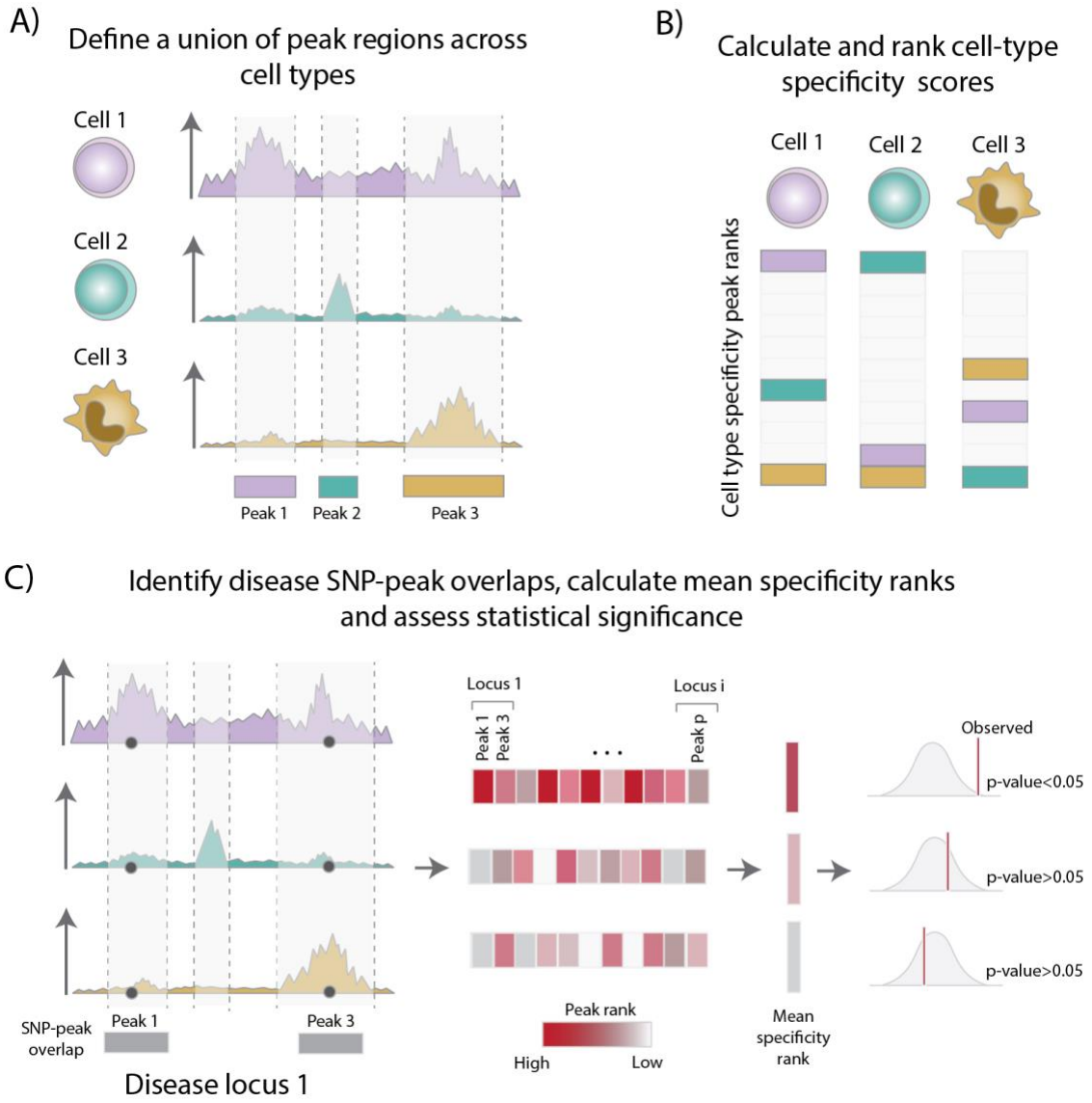



**Figure 2.2. Overview of the *CHEERS* method. A)** We first define a set of regulatory elements present in at least one cell state and construct a matrix of normalized read counts. **B)** We then calculate and rank specificity scores for each element in each cell state. **C**) To test for disease SNP enrichment, we take index variants from GWAS and identify variants in strong LD ($r^2 > 0.8$) with them. We then identify the peaks overlapping these variants, calculate their mean specificity ranks, and assess if this mean rank is higher than expected by chance.

## *CHEERS* identifies cell type relevant for immune-mediated diseases

To further validate *CHEERS,* we used a catalogue of 19 primary immune cell types assayed with H3K27ac ChIP-seq in the BLUEPRINT (Chen et al., 2016) and tested the enrichment of SNPs associated with 12 complex immune diseases: RA, CeD, IBD, CD, UC, MS, T1D, SLE, psoriasis, allergies, asthma, and systemic sclerosis. Concordant with previous results, we observed enrichment of RA variants in chromatin regions active in memory CD4+ T and CD8+

T cells (Hu et al., 2011; Trynka et al., 2013) (**Figure 2.3A**). Risk variants for allergies and asthma were also enriched in chromatin regions specifically active in memory CD4+ and CD8+ T cells. Furthermore, we saw enrichment of risk variants for SLE and systemic sclerosis in chromatin elements specific to B cells (**Figure 2.3A**), as previously reported (Farh et al., 2015). Importantly, *CHEERS* not only recapitulated the enrichment of SLE variants in B cells, but further showed that this signal is driven by class-switched and not unswitched B cells (**Figure 2.3B**). Based on these observations, we divided these immune-mediated diseases into three groups (**Figure 2.3B**): diseases with a strong T cell component (asthma, allergies, IBD, RA, CED, and T1D), diseases with a strong B cell component (psoriasis, SLE and systemic sclerosis), and diseases with both B and T cell components (CD, MS and UC). Based on these observations, we concluded that *CHEERS* can reliably test for the enrichment of risk variants in active chromatin elements across related cell types.

In order to ensure that *CHEERS* is applicable to non-immune traits, we tested its performance using epigenomic data from the Roadmap Epigenomics dataset, which contains H3K27ac data for a collection of 50 cell types, including a variety of non-immune tissues. We included three traits in this analysis: MS, CAD and Parkinson's disease (**Figure 2.3D**). We identified strong enrichment of GWAS loci for Parkinson's disease in cell types from different brain regions and the adrenal gland. Conversely, GWAS loci for MS showed significant enrichment in immune cell types. This showed that *CHEERS* was able to correctly disentangle the biological differences between MS and Parkinson's disease even though both affect the central nervous system. Moreover, we also identified a significant enrichment of risk variants for CAD in chromatin regions specific to the aorta (top enriched tissue), lungs and muscle (both skeletal and smooth). These results demonstrated that *CHEERS* was able to identify relevant cell type specific enrichments of GWAS loci across both immune and non-immune traits.

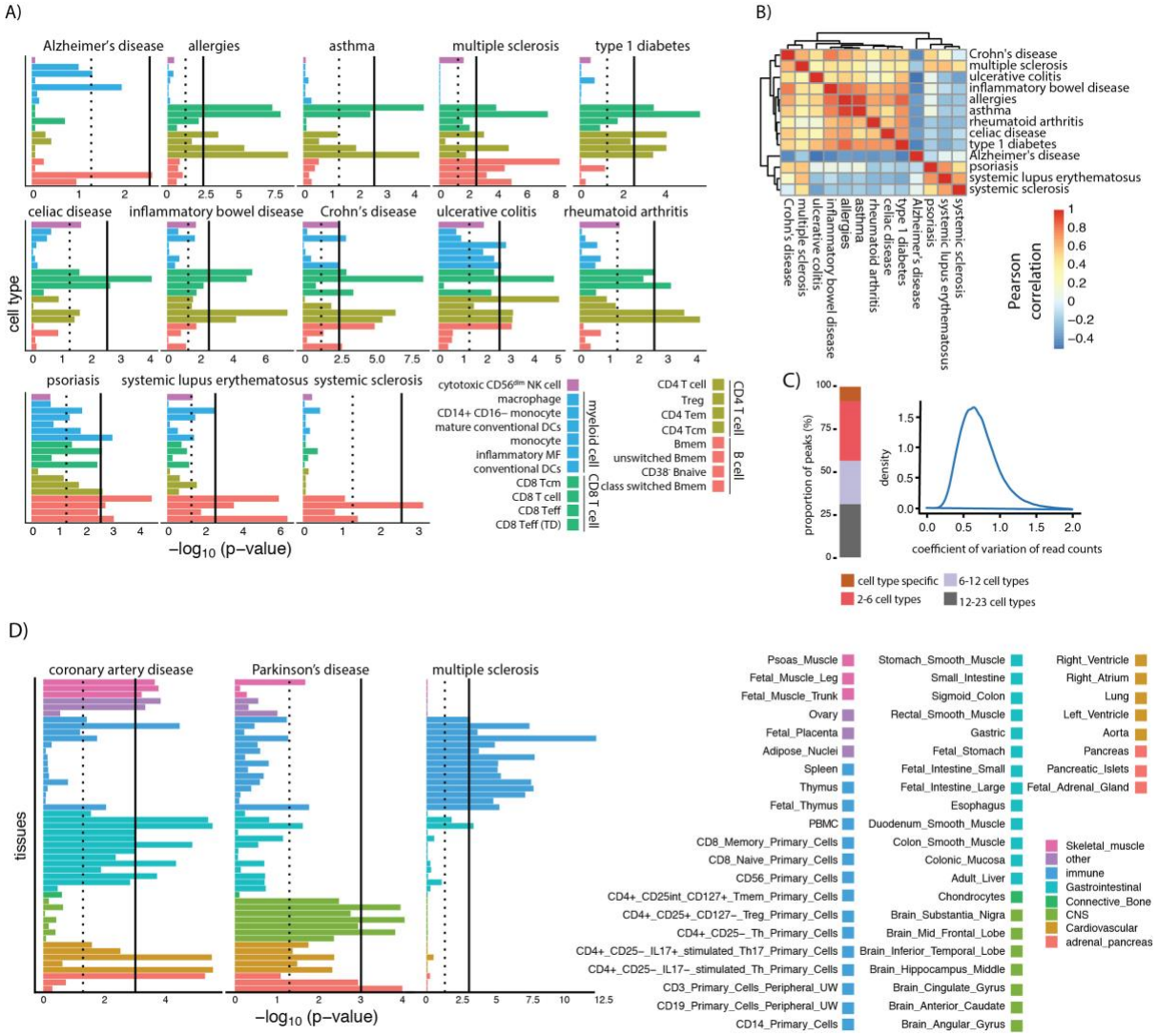



**Figure 2.3. Disease SNP enrichment in H3K27ac regions in blood cell types from the BLUEPRINT and ROADMAP projects. A)** One-sided p-values are reported from a discrete uniform distribution. CHEERS was performed after merging one to six biological replicates per cell type, resulting in 236,222 peaks. GWAS data and the statistical methods are described in the "GWAS data processing" and “CHEERS” sections of the Methods. The dotted gray line marks the nominal p-value threshold of 0.05, while the solid gray line represents the Bonferroni-corrected significance threshold (p-value ≤ 0.0026). **B)** Correlation of -log10(p-values) across diseases. Pearson’s correlations between 13 diseases were calculated using 20 cell types. **C)** Proportion of peaks shared by multiple immune cell types (barplot) and coefficient of variation of read counts per peak between cell types in the BLUEPRINT data. Number of peaks is 236,222. **D)** One-sided p-values are reported from a discrete uniform distribution. CHEERS was performed on one sample resulting in 174,296 peaks. GWAS data and the statistical method are described in the "GWAS data processing" and “CHEERS” sections of the Methods.

## Refining cell states relevant for immune-mediated diseases

We next tested if GWAS variants for immune-mediated diseases were enriched in regulatory elements specific to any of the cell states in our dataset. Risk variants for AD were also included in this analysis, since microglia, a subtype of macrophages, are thought to play a role in Alzheimer’s disease (Calderon et al., 2017; Tansey et al., 2018). The tested traits had

between 17 and 132 reported GWAS associations, resulting in between 25 (asthma) and 317 (IBD) SNP-peak overlaps that contributed to the final enrichment scores.

We did not observe any T cell enrichment in control traits such as CAD and LDL cholesterol (**Figure 2.4**). In contrast, variants associated with MS, RA, CeD, IBD, CD, UC, T1D, psoriasis, allergy and asthma were predominantly enriched in T cells, in line with our observations from the BLUEPRINT data (**Figure 2.4**). This enrichment was particularly strong in chromatin regions specifically active during early activation of memory T cells (16 hours; **Figure 2.4**). Importantly, this was not explained by any differences in the quality of H3K27ac peaks (i.e. q-value) between early and late activation (**Supplementary Figure 2.1**), and the enrichments were predominantly driven by peaks proximal to a TSS (**Supplementary Figure 2.7**). Additionally, resting memory T cells showed no significant enrichment in any of these diseases, suggesting that the enrichments are driven by the chromatin remodelling events that accompany T cell activation.

When inspecting individual loci, we found that the enrichment of immune variants in early memory T cell activation was driven by a group of H3K27ac peaks highly specific to this cell state (**Figure 2.5A** and **Supplementary Figure 2.8**). For example, a block of risk variants for allergies overlapped a peak close to *CCL20* gene, which showed higher levels of acetylation during early activation of memory T cells (**Figure 2.5B**). Allergy variants also overlapped a small group of peaks with higher activity during early activation of naive T cells, explaining the secondary enrichment in activated naive T cells (**Figure 2.5A**). Furthermore, a proportion of immune disease variants localised to peaks specific to early activation of both naive and memory CD4 T cells. For instance, RA associated variants overlapped a peak near acyl-CoA Oxidase-like protein (*ACOXL*) gene, which is involved in T cell metabolism (**Figure 2.5B**). Collectively, genes near peaks driving the enrichment in early activation of memory T cells across all diseases were enriched in pathways related to T cell activation, T cell differentiation, and leukocyte activation (**Figure 2.5D**). These results suggest that immune disease variants overlap enhancers and promoters that likely regulate gene expression programmes during early activation of memory T cells.

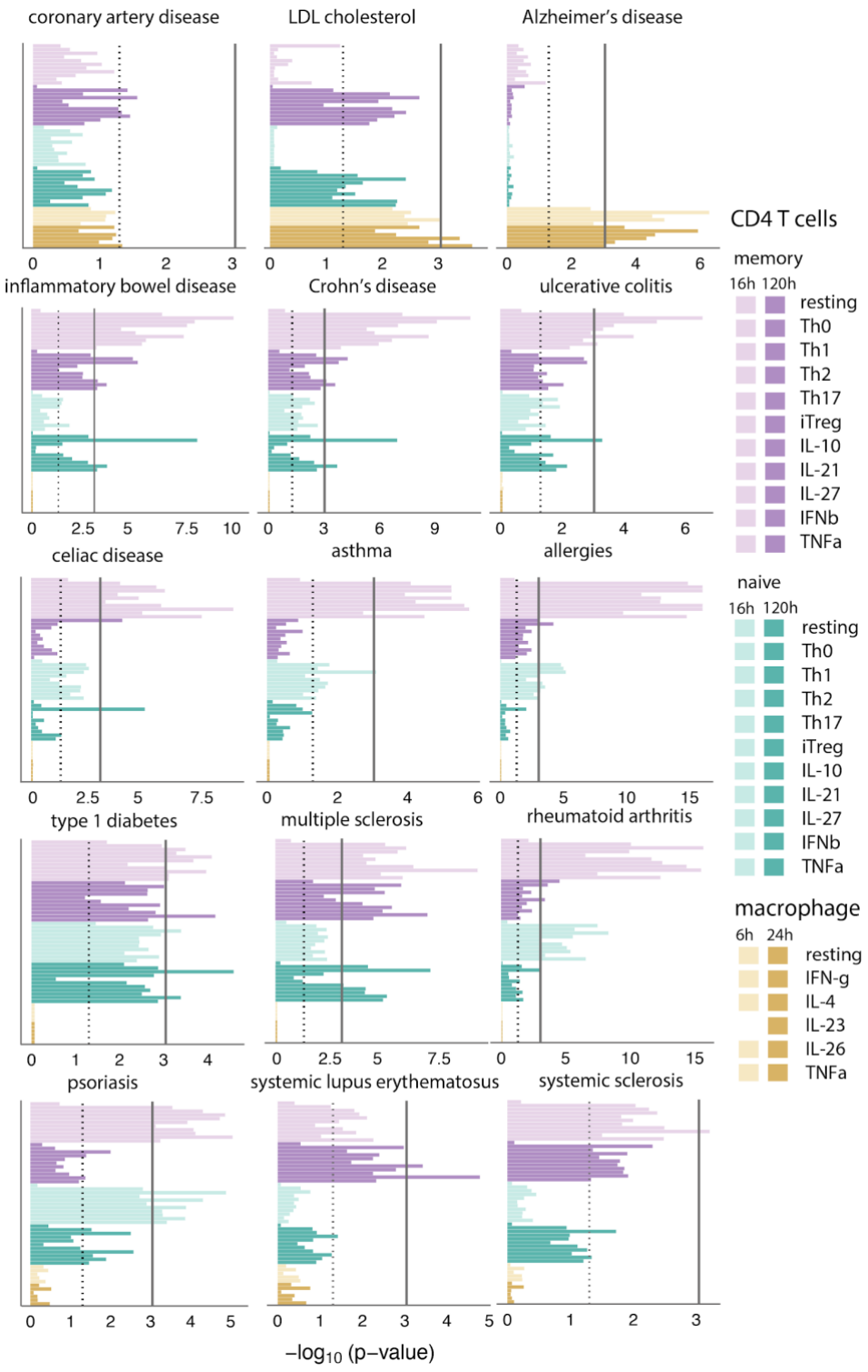



**Figure 2.4**. **Enrichment of disease associated SNPs in H3K27ac regions in cytokine-induced cell states**. One-sided P values were estimated based on a discrete uniform distribution. CHEERS was performed after merging three biologically independent samples resulting in 127,723 peaks. GWAS data and the statistical methods are described in the Methods sections of this chapter. The dotted line marks the nominal P-value threshold of 0.05, while the solid gray line represents the Bonferroni-corrected significance threshold (P ≤ 9 × 10−4). There is no data for IL-23-stimulated macrophages for 6 h, owing to the lack of biological material.

Overall, individual cytokine conditions and late stages of T cell activation were only enriched among a few selected T cell states and diseases. For example, MS variants were significantly enriched post IL-27 stimulation in early memory T cell activation ($p=5.4x10^{-10}$). This is concordant with previous studies that reported elevated levels of IL-27 in the cerebrospinal fluid of MS patients (Lalive et al., 2017; Sénécal et al., 2016). Moreover, naive T cells polarized with Th1 cytokines particularly stood out as strongly significant for IBD, CD, UC and CeD ($p = 9.92x10^{-9}$, $p = 1.11x10^{-7}$, $p = 4.89x10^{-4}$ and $p = 1.01x10^{-5}$, respectively). The H3K27ac peaks intersecting risk variants for these diseases and with high specificity in Th1 cells were in close proximity to genes involved in Th1 biology, such as IFN-γ and IL-12R (**Figure 2.5C**). Collectively, these genes were enriched in hallmark pathways of Th1 cells such as IFN-γ, IL-12 and JAK-STAT signaling (**Figure 2.5D**).

Of all the traits tested in our study, only AD variants were significantly enriched in macrophages (**Figure 2.4**). Surprisingly, we did not observe any individual cytokine condition being more significant than others, which suggests that macrophage function as a whole is relevant in Alzheimer's. Genes in proximity to peaks overlapping AD associated variants driving the macrophage signal were enriched in pathways such as regulation of amyloid beta clearance and amyloid beta formation, processes extensively studied in the context of AD pathology (**Figure 2.5D**).

Finally, we observed that the enrichment estimates derived using ATAC-seq peaks were significantly weaker than their H3K27ac counterparts (**Supplementary Figure 2.9A**), even if the general pattern of enriched cell states was concordant between the two (**Supplementary Figure 2.9B**). This agreed with our observations that ATAC-seq peaks were less variable. Similarly to H3K27ac data, ATAC-seq enrichments were mostly driven by peaks in close proximity to a TSS (**Supplementary Figure 2.10**).

In summary, our results suggest that variants associated with immune-mediated diseases likely play a role in genetic regulation during the early activation of CD4+ T cells, predominantly in memory T cells. We found that individual cytokine-induced cell states were enriched for variants associated with a handful of immune diseases. For example, risk variants for IBD, its two subphenotypes (UC and CD), and CeD were enriched in chromatin elements specific to late activation of naive T cells with Th1 polarizing cytokines. Our approach provides a valuable framework for identifying disease relevant cell types and cell states and to inform the design of subsequent large scale studies for functional interpretation of GWAS variants for complex immune-mediated diseases.

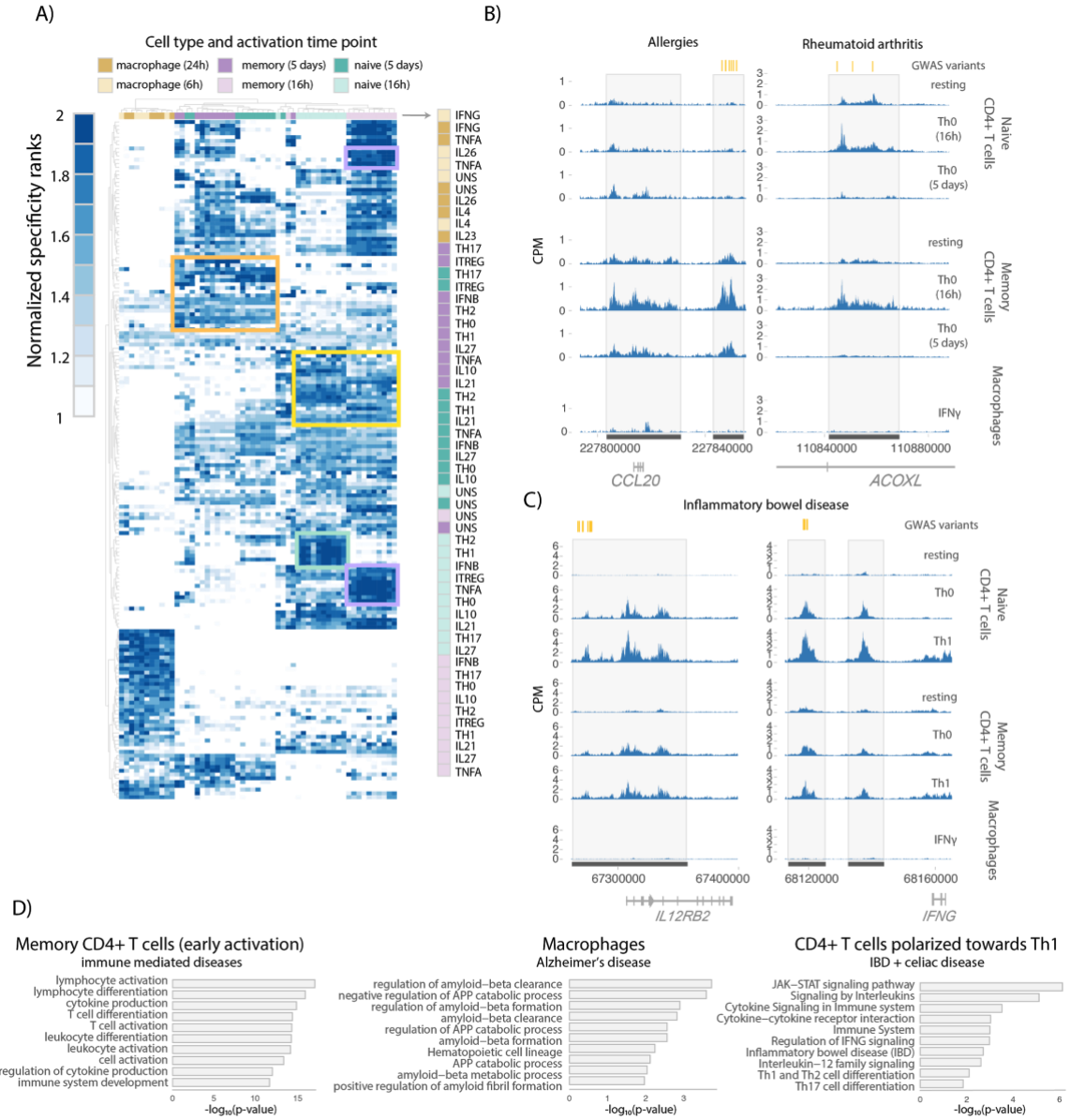



**Figure 2.5. Example loci driving immune disease SNP enrichment in cytokine induced cell states. A)** Loci driving enrichment of allergy variants in early activation of memory T cells. Each row corresponds to a H3K27ac peak that intersects GWAS variants for allergy, while each column corresponds to a different cytokine induced cell state. Shades of blue represent how specific each peak is to each cell state (specificity rank of the peak normalized to the mean specificity rank of all peaks). Different boxes highlight groups of peaks specific to different cell states. **B)** Read pileups at selection of H3K27ac peaks driving the enrichment of allergy and RA variants in early T cell activation. Genomic coordinates (GRCh38) for each gene are labeled. H3K27ac tracks were generated after merging three biologically independent samples per cell state. **C)** Read counts at a selection of H3K27ac peaks driving the enrichment of IBD variants in Th1-stimulated T cells. Genomic coordinates (GRCh38) for each gene are labeled. H3K27ac tracks were generated after merging three biologically independent samples per cell state. **D)** Pathway enrichment analysis using all the genes in proximity to H3K27ac peaks driving the enrichment of GWAS variants in early T cell activation (all immune mediated diseases), Th1 cells (IBD and CED) and macrophages (AD). To determine these enrichments, we selected peaks with top 10% specific peaks at the enriched cell states and assigned them to the closest gene. P-values were calculated using g:ProfileR with default parameters. The number of genes used for enrichment analysis was 215 for memory CD4+ T cells (early activation), 27 for macrophages and 53 for CD4+ T cells polarized towards Th1.

## Discussion

Identifying the cellular contexts in which disease associated variants function is critical for designing meaningful functional follow-up studies. Here, we used an *in vitro* system to identify cytokine-induced cell states relevant to the pathobiology of immune-mediated diseases. We observed that most chromatin remodelling events in CD4+ T cells were caused by TCR/CD28 stimulation. In contrast, cytokines had only a subtle effect on the quantitative levels of chromatin accessibility and histone acetylation.

This similarity between the chromatin landscape of different immune cell subsets could suggest that, at the epigenetic level, these cells are not as distinct from each other as previously suggested. A recent epigenomic study of resting and activated human immune cells reported similar observations, with most T helper subsets clustering together in PCA space (Calderon et al., 2019). Several biological explanations could account for this. For example, it has been shown that while Th1 and Th2 cells show similar active regulatory elements, GATA3 binds a different set of enhancers depending on whether T-bet is expressed (Kanhere et al., 2012). Thus, similar chromatin landscapes can still result in distinct transcriptomes. In addition, cell heterogeneity upon polarization and cell states which do not conform with the classic paradigm of T helper subsets are increasingly recognised. In particular, a recent study used scATAC-seq and scRNA-seq to profile CD4+ T cells from mouse colonic lamina propria (Kiner et al., 2021) and showed that, even when infected with pathogens known to bias the T helper response towards specific phenotypes, T cells from mice exposed to different pathogens clustered together rather than forming distinct groups of cells. While the authors observed a slight enrichment of cells expressing hallmark cytokines of the expected lineage, T cells from other lineages and even dual IFNγ/IL17-expressing cells were detected in response to all pathogens (Kiner et al., 2021). These observations from human and mouse support the idea that, upon exposure to polarizing cytokines, the T cell response is heterogeneous and cells retain active chromatin marks across loci associated with other subsets, possibly accounting for their plasticity (Kanno et al., 2012).

In addition to these mechanisms, it is worth noting that some technical factors could also account for the observed similarity of chromatin landscapes. In particular, human T cells are known to polarize less efficiently than mouse T cells (Mestas & Hughes, 2004), which could result in incomplete polarization *in vitro*. For example, human T cells often fail to acquire a suppressive phenotype in response to stimulation with TGFβ and IL-2, and the levels of IL-17 secretion upon Th17 polarization can vary depending on the amount of TGFβ, IL-23 and IL-6 used for polarization (Schubert et al., 2014). We believe that these technical factors are unlikely to confound our results, since we previously performed optimization experiments to

ensure that cells treated with these cytokine cocktails expressed their hallmark transcription factors (T-bet, GATA3, RORγt, and FOXP3, respectively) and produced their lineage-defining cytokines (as measured by RNA-seq and mass spectrometry) in a highly specific manner (Cano-Gamez, 2017). Nonetheless, without an in-depth assessment of the function of these cells, the influence of incomplete polarization cannot be entirely ruled out.

Regardless of the underlying mechanism, the observed similarity between chromatin landscapes from different cell states meant that currently available SNP enrichment methods could not distinguish between them. To address this, we developed *CHEERS*, a statistical method that takes into account quantitative changes in chromatin activity. Using *CHEERS*, we refined the enrichment of immune disease variants observed in previous studies (Hu et al., 2011; Okada et al., 2014; Trynka et al., 2015) to specific cellular contexts. For ten of the twelve immune diseases that we tested, we observed that the associated variants were enriched in early activation of memory CD4+ T cells. A recent study which profiled open chromatin regions across subsets of immune cells in resting and stimulated states also identified enrichment of immune disease variants in T cell activation (Calderon et al., 2018). Our study now provides further evidence for the importance of memory T cells in the biology of complex immune-mediated diseases and proposes dysregulation of specific cellular processes during T cell activation as an important disease mechanism.

Immunophenotyping studies have identified elevated cytokine levels in patients with immune diseases, subsequently leading to development of disease-modifying therapies targeting different cytokines (Babaloo et al., 2013; Jamshidian et al., 2013; I. B. McInnes et al., 2016; Schett et al., 2013). This implies that effector T cells (roughly corresponding to five day stimulations in our dataset) and the cytokines that they secrete drive autoimmune inflammation. In contrast, our results suggest that GWAS variants for immune-mediated diseases affect regulation mostly during the initial phase of memory T cell activation, but less so during the effector response seen at a later time point. Therefore, while cytokine-induced cell states could be key in the later stages of pathologic inflammation, they might not be triggering immune disease. This emphasises the importance of tight regulation of T cell activation, suggesting that many subtle effects of immune variants lead to dysregulated early cell responses. This agrees with observations from severe immune disorders where, for instance, deficiency in the expression of one of the main regulators of T cell activation, CTLA-4, has been associated with the development of autoimmune diseases due to uncontrolled T cell activation (Kuehn et al., 2014; Lo et al., 2015; Schubert et al., 2014; Tivol et al., 1995). Furthermore, it is worth noting that immune-mediated diseases are more often diagnosed in adults, which correlates with a shift in the frequency of T cells from naive to predominantly memory. Therefore, our results suggest that focusing on regulation of early activation of

memory CD4+ T cells may provide an important axis for development of new treatment options.

In addition to early activation of memory CD4+ T cells, we observed selected diseases where the associated variants were enriched in distinctive cytokine polarized cell states. For example, IBD, CD, UC and CeD risk variants were highly enriched in T cells polarized with Th1 cytokines. This agrees with previous findings from immunophenotyping studies showing that lamina propria T cells express high levels of STAT4, T-bet and IL-12R, which are all induced by Th1-polarizing cytokines (Neurath et al., 2002). This could indicate that a proportion of IBD variants regulate the function of Th1 cells, suggesting the involvement of this cell type in IBD development. Another example was asthma, where we detected GWAS SNP enrichment in naive T cells polarized towards the Th2 phenotype.

Across the tested immune diseases, we did not observe any significant enrichment in the macrophage stimulatory cell states, suggesting that autoimmune inflammation is mostly driven by T cells. In contrast, variants associated with AD point to a crucial role of macrophages in Alzheimer's. Recent studies have highlighted the role of microglia in AD (Calderon et al., 2017; Gosselin et al., 2017; Tansey et al., 2018), a type of resident macrophage in the central nervous system (CNS). Our results suggest that a proportion of AD associated variants could be studied in an *in vitro* model of monocyte-derived macrophages. This can have significant implications, as cells from the CNS are challenging to collect. Finally, the lack of stratification of AD variants between the different cytokine-induced macrophage cell states could indicate that some AD variants might regulate more general functions of the macrophage lineage. These functions could be closely linked to amyloid beta clearance, as suggested by our results.

An important limitation of our study is that the cell types prioritized by *CHEERS* are based on the ensemble of all loci associated with a trait of interest. As a consequence, it is not possible to use *CHEERS* to determine the tissue of action of each individual GWAS locus, and it remains plausible that secondary tissues other than the top findings from *CHEERS* play an important role in disease. This issue can be tackled using some recently developed methods. In particular, chromatin conformation data (e.g. HiC or PC-HiC) can be used to identify variants which are not in LD with a GWAS locus, but which physically interact with it in three dimensional space (Corradin et al., 2016). These so called 'outside variants' tend to be different across cell types, and thus cell type specific outside variants can be retested for association with disease in the context of the original GWAS, in turn making it possible to elucidate the most likely tissue of action of each disease-associated locus. This methodology was recently applied to GWAS associations for MS, revealing that, while T cells account for

the majority of GWAS signals, a small subset of loci acts via oligodendrocytes and has a direct impact on remyelination (Factor et al., 2020). A second approach, tissue of action scores (TOAs), partitions the credible sets of disease associated SNPs obtained for each locus according to whether each SNP overlaps functional annotations profiled across a variety of cell types (Torres et al., 2020). This enables the calculation of scores representing the probability that each tissue is responsible for the association at a given locus. Despite the benefits of these methods, they are designed to answer a different question and thus none of them takes into account quantitative levels of chromatin activity. However, by combining *CHEERS* with these approaches, it could be possible to identify the main tissues involved in disease while still prioritizing potential secondary cell types through which a smaller set of disease associations could be mediated.

In conclusion, we developed a novel statistical approach to estimate SNP enrichment across similar cell types and used it to identify T cell activation as a key process dysregulated in immune disease. Our study is the first to systematically profile changes in chromatin regulatory landscape induced by cytokines during activation of human immune cells. This provides a valuable resource to identify appropriate cell models for studying how genetic variants lead to diseases.

## Methods

### Sample processing and cell activation

Blood samples were obtained from three healthy individuals. The human biological samples were sourced ethically and their research use was in accord with the terms of the informed consents under an IRB/EC approved protocol. Peripheral blood mononuclear cells (PBMCs) were isolated using Ficoll-Paque PLUS (GE healthcare, Buckingham, UK) density gradient centrifugation. Naive and memory CD4+ T cells were isolated from PBMCs using EasySep® naive CD4+ T cell isolation kit and memory CD4+ T cell enrichment kit (StemCell Technologies, Meylan, France) according to the manufacturer's instructions. T cells were stimulated with anti-CD3/anti-CD28 human T-Activator Dynabeads® (Invitrogen) at 1:2 bead to T cell ratio. Cytokines were added at the same time as the stimulus and cells were harvested after 16h and 5 days (for a list of cytokines and their concentration refer to **Supplementary Table 2.1**).

Monocytes were isolated using EasySep monocytes isolation kit according to the manufacturer's instructions. In order to generate macrophages, monocytes were plated in a 100mm x 20mm dish and cells were treated with 800 U/ml M-CSF (PeproTech) for seven

days. Following macrophage differentiation, cells were stimulated with cytokines for 6 and 24 hours (for a list of cytokines and their concentration refer to **Supplementary Table 2.1**).

### ChIPmentation-sequencing

In order to profile active enhancers and promoters, cells were washed using RPMI and Dynabeads were removed using a DynaMag magnet (Thermo Fisher). Next, cells were resuspended at $1x10^6$ cells/ml in fluorescence-activated cell sorting (FACS) buffer (phosphate buffered solution, i.e. PBS buffer, supplemented with 10% FCS and 1 mM EDTA) and chromatin was cross-linked by adding 1% formaldehyde, incubating at 37°C for five minutes. To quench the reaction, glycine was added and cells were washed using cold PBS buffer. Cross-linked cell pellets were frozen in liquid nitrogen and stored at -80°C until further processing.

To perform ChIPmentation, cross-linked pellets were thawed and processed using the iDeal® ChIP-seq kit for histones (Diagenode) according to the manufacturer's instructions. Briefly, cells were lysed and sonicated, and the resulting material was used for immunoprecipitation (IP). Sonication was performed using a Bioruptor Pico (Diagenode), with 6 sound pulse cycles of 30 seconds each. For immunoprecipitation (IP), we used ChIP grade antibodies specific to human H3K27ac histones (Diagenode). A negative control undergoing no IP (input) was also generated for each cell type. After IP, the recovered chromatin fragments were tagmented as previously described (Schmidl et al., 2015). Briefly, 2 µl of IP material were resuspended in 30 ul ChIP-seq buffer (Diagenode) supplemented with 1 µl Tn5 (Illumina). Samples were then reverse cross-linked using the iDeal ChIP-seq kit for histones according to the manufacturer's instructions. The resulting DNA was purified using SPRI magnetic beads (AMPure XP A63881 Beckman Coulter). Enrichment of active chromatin regions was verified by qPCR.

Sequencing libraries were constructed using the Nextera DNA library preparation kit according to the manufacturer's instructions. Briefly, DNA was amplified by PCR and fragments of inappropriate sizes were removed using Agencourt AMPure XP beads (BD). Finally, samples were pooled and loaded into an Illumina HiSeq 2500 instrument for 75 bp paired-end sequencing. In order to minimise batch effects, samples were randomized before sequencing. We obtained on average 63 million paired-end reads per sample.

### ATAC-sequencing

In order to profile open chromatin regions, stimulated cells were washed with RPMI and Dynabeads were removed using a DynaMag magnet (Thermo Fisher). Next, tagmentation was performed using the fast ATAC protocol (Corces et al., 2016). Briefly, 50,000 cells were

washed using cold PBS buffer and resuspended in 50 µl of Nextera tagmentation buffer supplemented with 0.01% digitonin and 2.5 µl Tn5 transposase (Nextera). Samples were then incubated at 37°C and 800 rpm for 30 minutes. After tagmentation, DNA was purified using MinElute PCR columns (Qiagen) according to the manufacturer's instructions and stored at -80°C until library preparation.

Sequencing libraries were generated from tagmented DNA using the Nextera DNA library preparation kit according to the manufacturer's instructions. Briefly, DNA was amplified by PCR and fragments of inappropriate sizes were removed using Agencourt AMPure XP beads (BD). Finally, samples were pooled and loaded into an Illumina HiSeq 2500 for paired-end sequencing. In order to minimise batch effects, samples were randomized before library preparation and before sequencing. We obtained on average 58 million paired-end reads per sample.

### ATAC-sequencing and ChIPmentation-sequencing analysis

We assessed the quality of reads using fastx and trimmed the adapters using skewer (v0.2.2) (Jiang et al., 2014). We then mapped reads to the human genome reference GRCh38 using bwa mem (v0.7.9a) (H. Li & Durbin, 2010), keeping only uniquely mapped reads. We also removed PCR duplicates and mitochondrial reads in ATAC-seq using samtools (v0.1.9) (H. Li et al., 2009; H. Li & Durbin, 2010). This resulted in final BAM files containing uniquely mapped, non-mitochondrial reads that were used for peak calling. Finally, we calculated insert size distributions using PICARD tools (v2.6.0) to remove samples with over- or under- sonicated chromatin and removed samples with a skewed distribution of insert sizes. Peaks were called using MACS2 (v2.1.1) (Y. Zhang et al., 2008) -q 0.05 --nomodel --extsize 200 --shift -100 for ATAC; and --broad --broad-cutoff 0.1 --nomodel --extsize 146 for H3K27ac ChIPmentation. Next, we calculated the fraction of reads in peaks (FRiP) to investigate the quality of our data. The average FRiP for H3K27ac ChIPmentation-seq was 46.64% and for ATAC was 17.03%. Samples with FRIP <5%, skewed distribution of insert sizes, or <20 million QCed reads were removed from the downstream analysis.

To call peaks per cell state, QCed BAM files corresponding to biological replicates of the same condition were merged using samtools and peaks were called as described above but with two additional parameters. First, when pooling reads across individuals and within the cell state we expected to see a proportion of identical sequence reads in independent donors by chance, therefore we used the --keep-dup flag in MACS2, as PCR duplicated reads had already been removed. Second, we increased the q-value threshold to 0.1. In ATAC-seq,

where the quality of the data was lower compared to ChIPmentation, we wanted to ensure that noisy peaks were excluded from our enrichment analysis. Therefore we used stringent parameters and kept all peaks with fold enrichment > 4 and q-value < $10^{-4}$. In ChIPmentation, where peaks are broad and FRiP values higher, we kept all peaks with fold enrichment > 2 and q-value < $10^{-2}$. This resulted in the final BED files which were used for GWAS enrichment analysis.

To define differentially accessible regions and differentially modified histone regions, we used DESeq2. We compared all conditions to the resting state or to Th0 and used a Benjamini-Hochberg controlled false discovery rate (FDR) of 5% and an absolute fold-change ≥ 1.

### Disease enrichment with GoShifter

We ran GoShifter on ATAC and H3K27ac peaks as described previously (Trynka et al., 2015). We ran 10,000 permutations. The command used was: *python goshifter.py -s SNP_list -a annotation/file/path -p 10000 -l LD/file/path -o output_name*.

### Chromatin Element Enrichment Ranking Specificity (CHEERS)

We first merged the peaks across all cell types and cell states using the merge option in bedtools (v2.22.0) in order to get a unified set of peaks. Then, for each cell state we quantified the number of reads within each of the merged peak regions using featureCounts (v1.5.1) (Liao et al., 2014). We normalized for library size by scaling the read counts per peak to the sample with the largest count of informative sequence reads:

$$Z_{i,j} = C_{i,j} \frac{max_i \{\sum_j \quad C_{i,j}\}}{\sum_j \quad C_{i,j}}$$

where $Ci,j$ is the number of reads falling within peak $j$ in cell state $i$.

To ensure that our analysis was not confounded by the low confidence peaks, we removed the bottom 10th percentile of peaks with the lowest read counts and obtained a final count of 132,236 peaks for H3K27ac ChIPmentation-seq and 740,221 for ATAC-seq.

In order to compare the peaks across cell types and conditions, we quantile normalized the library size-corrected peak counts. We then transformed the read count of each peak into a score that reflects cell type specificity (specificity score). For that, we divided the normalized read counts of each peak in each condition by the Euclidean norm for that peak across all conditions, as described by the following formula:

$$S_{i,j} = \frac{Z_{i,j}}{\sqrt{\sum_i \quad Z_{i,j}^2}}$$

where $Si,j$is the specificity score, and $Zi,j$ is the normalized number of reads within a peak $j$ in a condition $i$. This score is a number from 0 to 1, where 1 means a peak has high read counts in only one cell state and 0 means the peak shows no reads in that cell state.

Then, within each cell state, peaks were ranked by specificity score from the highest to the lowest score, where the peak least specific to the cell state was ranked 1. If multiple peaks have equal specificity scores, the same lower rank is assigned to all of them.

To test for disease SNP enrichment, we take all the index variants, identify variants in LD (r2>0.8) and overlap them with peaks. Importantly, our method is peak-centric, i.e. a peak that overlaps multiple disease associated variants in a locus contributes to the final cell-type specificity score only once. On the other hand, if within a locus the associated variants intersect multiple peaks, each independent peak contributes to the final enrichment score. We then calculate the mean specificity rank of all peaks which intersect disease associated variants. We inferred the significance of the observed enrichment by fitting a discrete uniform distribution. Within each cell state, all ranks (1, 2, 3...*N*) can be observed with an equal probability and thus they follow a discrete uniform distribution (Kolmogorov-Smirnov test p=1) with mean ($\mu$):

$$\mu = \frac{1 + N}{2}$$

and variance:

$$\sigma^2 = \frac{(max \ - \ min \ + 1)^2 \ - 1}{12}$$

where *max* and *min* are the maximal and minimal ranks in the dataset. When we substitute *max* with number of peaks (*N*) and *min* with the minimal rank (1) we obtain the following formula:

$$\sigma^2 = \frac{N^2 - 1}{12}$$

which, under the central limit theorem converges to:

$$\underline{\sigma}^2 = \frac{N^2 - 1}{12n}$$

where $\underline{\sigma}^2$ is the variance of the mean of the $n$ peak ranks overlapping the GWAS SNPs assuming that these overlap at random.

Finally, we calculate p-values as:

$$p = 1 - \Phi\left(\frac{x - \mu}{\sigma}\right)$$

where the $\Phi\left(\frac{x-\mu}{\sigma}\right)$ is a cumulative normal distribution for $x \sim N(\mu, \sigma^2)$; x is the observed rank, and μ and σ are the expected values under the null hypothesis that all ranks occurred as random.

To ensure that this method accounts for possible properties of the data, such as correlations, we also assessed the significance of the enrichment using an empirical, permutation based strategy. For that, within each of the tested cell states, we randomly sampled sets of peaks, matching for the number of peaks overlapping GWAS variants, and calculated the mean of their ranks. We repeated this process N times, assessed the frequency at which the mean of permuted ranks was greater or equal to the mean of observed ranks, and derived an empirical p-value:

$$p_{value} = \frac{num(permuted \ > \ observed)}{N}$$

where N is a number of permutations. Both approaches yielded similar results ($R^2$ = 0.96, p-value < 2.2e-16), but the p-values derived based on the discrete uniform distribution were not limited by the number of permutations.

### Power calculations for *CHEERS*

To estimate the power of *CHEERS*, we simulated 100 SNP-peak overlaps and tested for enrichment. Each simulation was repeated 100 times and power was estimated as the percentage of simulations that yielded a significant enrichment (p-value < 0.01). We designed the simulations such that a given percentage of SNPs (ranging from 0% to 100%), but not the remaining SNPs, always overlapped with peaks in the top 10th percentile of specificity. Finally, in order to test how specific the peaks needed to be for our method to detect enrichment, we repeated the simulations at lower specificity percentiles.

### Pathway enrichment analysis

Pathway enrichment analysis and statistics were performed using g:ProfileR (v 0.6.1).

### Code and data availability

All of the raw-sequencing data presented in this chapter are deposited in the European Genome-Phenome Archive (EGA) with accession numbers EGAS00001003501 (ATAC-seq) and EGAS00001002749 (H3K27ac ChIPmentation-seq). The *CHEERS* software is publicly available in GitHub via https://github.com/trynkaLab/CHEERS.

# Naive and memory CD4+ T cells form a transcriptional continuum with a spectrum of effector responses

*The contents of this chapter have been previously published in Nature Communications (Cano-Gamez, Soskic et al., 2020), as detailed in the Publication Note.*

## Overview

The communication between immune cells is mediated by cytokines, which promote the differentiation of cells into effector cell types (Federica Sallusto, 2016; Turner et al., 2014). In particular, upon activation, naïve CD4+ T cells are polarized by cytokines into phenotypes like Th1, Th2 and Th17 (Harrington et al., 2005; Le Gros et al., 2008; Mosmann et al., 1986; Park et al., 2005). T helper cells in turn coordinate the downstream response of other immune cells (e.g. CD8+ T cells, macrophages and B cells) (Zhu et al., 2010). Previous studies have increased our understanding of cytokine-induced polarization (Aijö et al., 2012; Fantini et al., 2004; Kanhere et al., 2012; Purvis et al., 2010; Samstein et al., 2012; S. J. Szabo et al., 2000; Ubaid Ullah et al., 2018; Vahedi et al., 2012; X. O. Yang et al., 2008; W. Zheng & Flavell, 1997). Nonetheless, most studies focus exclusively on naïve CD4+ T cells. This is due to the premise that, once polarized, the phenotype acquired by CD4+ T cells remains stable. Recent studies have challenged this idea, showing that cytokines can reprogram previously polarized cells (Nistala et al., 2010; Federica Sallusto, 2016; Veldhoen et al., 2008). For example, IL-6 can convert Treg cells to a pathogenic Th17-like phenotype under arthritic conditions (Komatsu et al., 2014). Furthermore, Th17 cells upregulate *TBX21* and IFN-γ in response to Th1-polarizing cytokines (Cohen et al., 2011), and infection-induced Th17 cells can secrete Th1 cytokines (Omenetti et al., 2019). These observations highlight the plasticity of CD4+ T cells and suggest that memory cells respond to cytokines. Furthermore, several genetic studies, including the one described in chapter two, have implicated memory T cells in many complex immune diseases (Hu et al., 2011; Trynka et al., 2013). This makes it crucial to understand their response to cytokines. However, studying the effects of cytokines on memory T cells is challenging because memory cells comprise multiple subpopulations, each of which could respond differently to the same environmental cues (F. Sallusto et al., 1999; Federica Sallusto et al., 2004; Y. Tian et al., 2017).

In this chapter, I describe the results of a study carried out by myself and others (as specified in the collaboration note) to characterize the response of naïve and memory CD4+ T cells to five different cytokine combinations at two different time points following stimulation, profiling bulk and single-cell gene expression. At the single-cell level, we show that CD4+ T cells form

a transcriptional continuum which spans the naive, central memory, and effector memory phenotypes. Progression in this continuum is accompanied by increased expression of effector molecules and can influence the response to activation and cytokine-polarization. Our results provide a new framework for studying naive and memory T cell activation.

## Results

### Study design

To investigate the effects of cytokines on human naive ($T_N$) and memory ($T_M$) CD4+ T cells (**Supplementary Figure 3.1A**), we stimulated cells with anti-CD3/anti-CD28 coated beads in the presence of different cytokine cocktails (**Figure 3.1A**, **Figure 3.1B** and **Supplementary Table 3.1**). We polarized $T_N$ and $T_M$ towards four T helper phenotypes (Th1, Th2, Th17 and iTreg), as well as with IFN-β due to its role in multiple sclerosis (Ebers, 1998; Paty & Li, 1993). To distinguish T cell responses to TCR/CD28-activation from responses induced by cytokines, we also stimulated cells with anti-CD3/anti-CD28 beads in the absence of cytokines (Th0). Further, we cultured cells in the absence of stimulation or cytokines (resting cells). We profiled gene expression (RNA-seq) 16 hours and 5 days after stimulation. To comprehensively characterise cellular states at the latter time point, we also profiled the whole proteome (liquid chromatography-tandem mass spectrometry, LC-MS/MS), and single-cell transcriptomes (single-cell RNA-sequencing, scRNA-seq) (**Methods**).

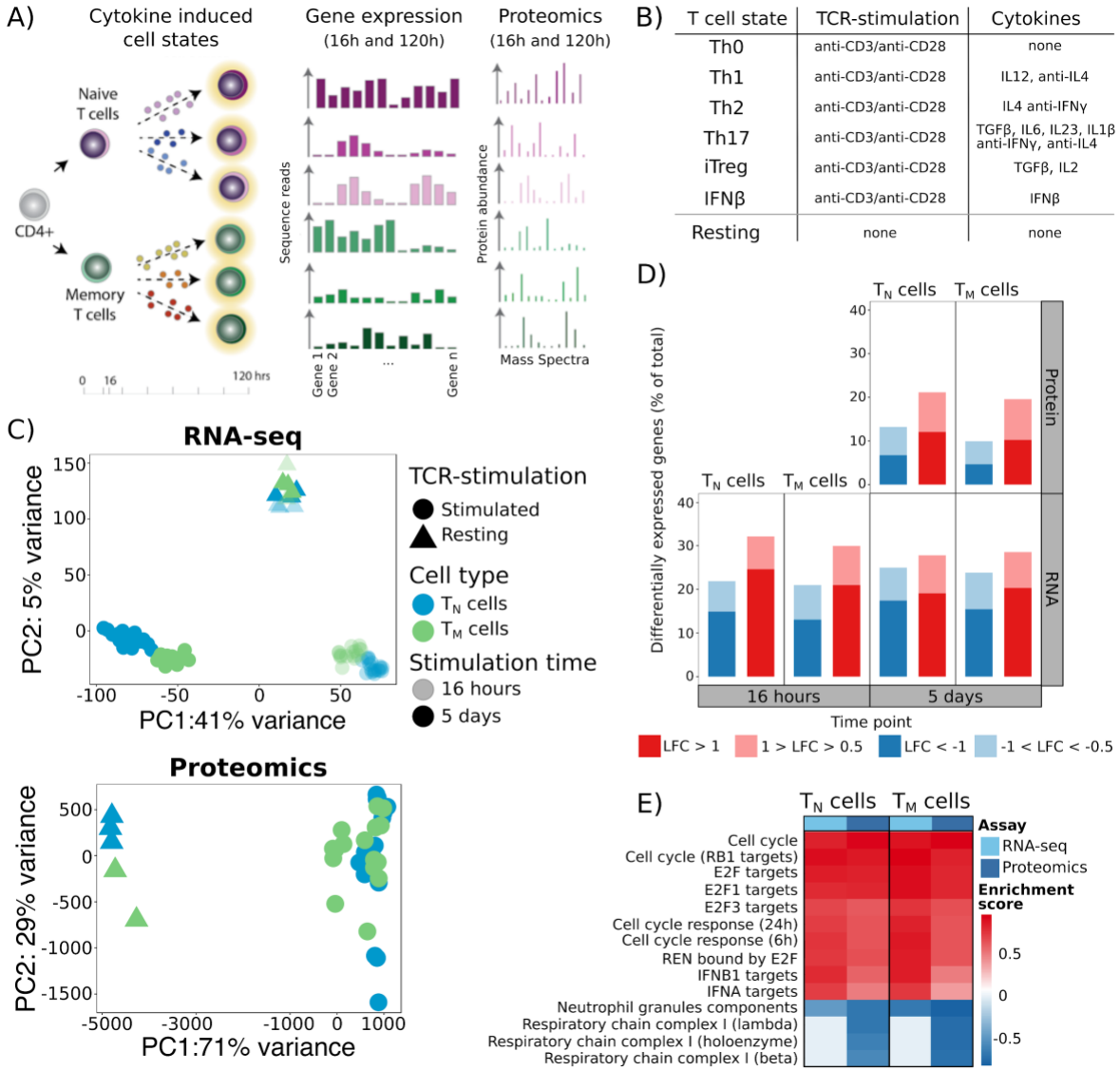



**Figure 3.1. TCR/CD28-activation induces cell type specific gene expression programs in CD4⁺ T cells. A)** Overview of the experimental design. **B)** List of cytokine conditions. **C)** PCA plots derived from whole transcriptome (upper panel) and whole proteome (lower panel) profiles of $T_N$ and $T_M$ cells. Different colors correspond to cell types and different shades to stimulation time points. PCA plots were derived using 47 naive and 47 memory T cell samples for RNAseq and 21 naive and 19 memory T cell samples for proteomics. **D)** Gene expression changes at the RNA and protein levels obtained by comparing TCR/CD28-activated (Th0) cells to resting cells. Up-regulated genes are shown in red and down-regulated genes in blue. Different shades indicate different fold-change thresholds. **E)** A selection of significantly enriched pathways (with enrichment scores > 0.7) estimated based on genes and proteins differentially expressed after five days of activation using the 1D enrichment method.

## Activation induces cell type specific responses in naive and memory T cells

To understand $T_N$ and $T_M$ responses to T cell activation (TCR/CD28-activation), we compared the transcriptomes of activated cells with those of resting cells. The main source of variation across the transcriptome and proteome was T cell activation, with resting cells separating from activated cells in PCA space (**Figure 3.1C**). Activated cells clustered by duration of stimulation (16 hours and five days) and cell type ($T_N$ and $T_M$), suggesting that the response to T cell activation is dynamic and cell type specific (**Figure 3.1C**). We then assessed differential gene expression between resting and activated (Th0-stimulated) $T_N$ and $T_M$ (**Figure 3.1D,** **Supplementary Tables 3.2** and **Supplementary Table 3.3**). At the RNA level, 8,333 and 7,181 genes (40% of genes) were differentially expressed after 16 hours in $T_N$ and $T_M$, respectively (log-fold change, LFC > 1 at 0.05 false discovery rate, FDR). This number was comparable after five days (7,705 and 7,544 in $T_N$ and $T_M$). At the protein level, 4,009 and 3,443 proteins were differentially expressed (35% of proteins) after five days in $T_N$ and $T_M$, respectively (LFC > 0.5 at 0.1 FDR). These genes were enriched in cell cycle progression and type I IFN response (**Figure 3.1E** and **Supplementary Table 3.4**). Conversely, $T_N$ and $T_M$ downregulated components of the respiratory chain complex (**Figure 3.1E**), in line with previous observations that T cell activation induces proliferation and a metabolic switch to support effector responses (Menk et al., 2018).

## Cytokines induce cell type specific responses in naive and memory T cells

We next investigated how cytokines modulate gene expression in $T_N$ and $T_M$. This was achieved by performing PCA on the proteome and transcriptome, treating time points and cell types independently. While there were few cytokine effects at 16 hours (**Supplementary Figure 3.1B**), we observed clear clustering by cytokine condition at five days (**Figure 3.2A**) in both the transcriptome and the proteome. We next compared stimulated cells exposed to cytokines to Th0-stimulated cells (**Supplementary Table 3.2** and **Supplementary Table 3.3**). With the exception of IFN-β, all cytokine induced changes were apparent only after five days of stimulation (**Figure 3.2B**). For example, Th17-stimulation induced 42 differentially expressed genes after 16h in $T_N$, compared to 1,818 differential genes after 5 days. In contrast, IFN-β induced a large number of early transcriptional changes (357 genes at 16h and 329 after 5 days in $T_N$), reflecting its role in the fast antiviral response. These results suggest that early changes in gene expression are dominated by T cell activation, while the expression programs of differentiated Th cells are only apparent at later stages of stimulation. This implies that cytokine polarization occurs after the initiation of T cell activation.

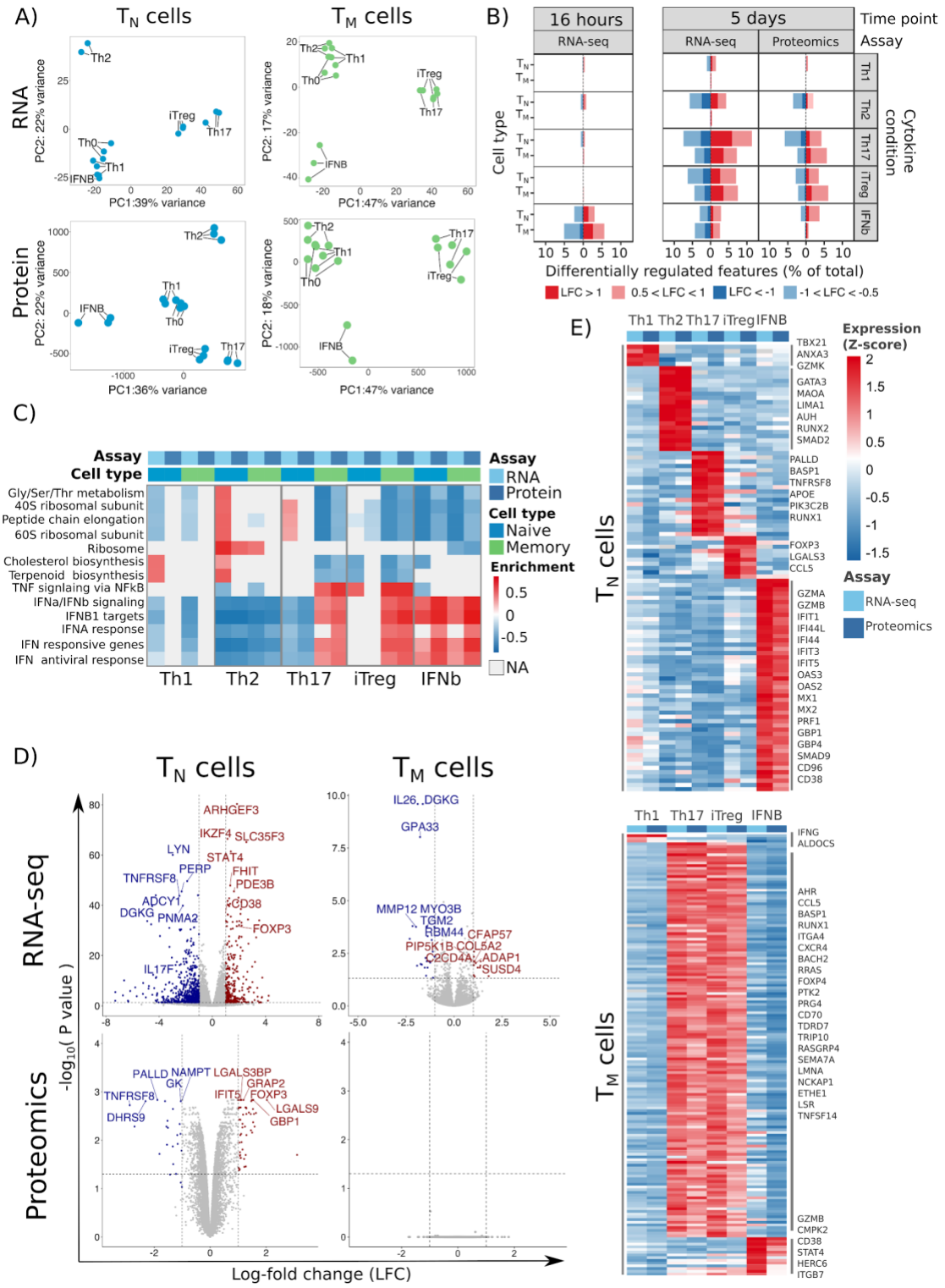



**Figure 3.2. Cytokines induce cell type specific gene expression programs in CD4⁺ T cells. A)** PCA plot based on the full transcriptome and full proteome of $T_N$ and $T_M$ cells following five days of cytokine stimulation. Only stimulated cells were included in this analysis. PCA plots were derived using 20 $T_N$ and 21 $T_M$ samples for RNAseq, and 18 $T_N$ and 17 $T_M$ samples for proteomics. **B)** Gene expression changes at the RNA and protein levels were inferred from pairwise comparisons between cytokine-stimulated cells and Th0-stimulated cells. Up-regulated genes are shown in red and down-regulated genes in blue. Different shades indicate different fold-change thresholds. **C)** Selection of significantly enriched pathways (with enrichment scores > 0.7 in at least one cytokine condition)

estimated using differentially expressed genes and proteins with the 1D enrichment method. Colors correspond to the enrichment score estimated for each pathway. For the pathways and conditions indicated in light grey, not enough genes were detected to reliably estimate enrichment scores. **D)** Volcano plots highlighting significant differences in RNA and protein expression between Th17 and iTreg-stimulated $T_N$ and $T_M$ cells. Red indicates upregulation in iTreg with respect to Th17-stimulation, blue indicates upregulation in Th17 with respect to iTreg-stimulation. Labels were added to *IL17*, *FOXP3* and the top 20 most differentially expressed genes. **E)** Cell state specific gene signatures defined using joint RNA and protein expression. Colours encode normalized (Z-scored) gene and protein expression levels. Example genes for each signature are labeled.

Since cytokine-induced effects manifested five days after stimulation, we focused on this time point to further elucidate changes in RNA and protein expression driven by different cytokines. The number of cytokine-induced changes in RNA and protein was comparable between $T_N$ and $T_M$ (**Figure 3.2B**). However, Th2-stimulation triggered different responses between the two cell types, resulting in differential expression of 944 genes in $T_N$ compared to 49 in $T_M$. We observed the same trend at the protein level, where 290 proteins were differentially expressed in $T_N$ but no differences were detected in $T_M$ (**Figure 3.2B**), although $T_M$ expressed comparable levels of the IL-4 receptor (**Supplementary Figure 3.2A**). This suggested that $T_M$ cannot polarize towards the Th2 phenotype.

We next sought to translate these observations to cellular functions, noting that the transcripts and proteins differentially expressed upon cytokine stimulation were enriched in relevant pathways (**Figure 3.2C** and **Supplementary Table 3.4**). Stimulation with IFN-β induced upregulation of the type I IFN response in both $T_N$ and $T_M$, while Th2-polarization of $T_N$ suppressed this pathway, likely reflecting the fact that Th2-polarization involves IFN-γ blockade. These effects were concordant between RNA and protein expression (**Supplementary Figure 3.3**). Furthermore, Th1-stimulation of $T_N$ induced metabolic changes such as increased cholesterol and terpenoid synthesis, while Th2-stimulation increased the expression of genes involved in amino acid metabolism (**Figure 3.2C**). Interestingly, some pathways showed opposite effects between $T_N$ and $T_M$ . For example, while Th17-stimulation of $T_N$ induced downregulation of the type I IFN response, Th17-stimulation of $T_M$ increased the activity of this same pathway. We observed a similar pattern upon iTreg-stimulation, with type I IFN response upregulated in $T_M$ but not in $T_N$ (**Figure 3.2C**). These observations suggest that Th17 and iTreg-stimulation induce different cell states in $T_N$ than in $T_M$.

Th17 and iTreg cells have been linked to autoimmune inflammation and immune suppression, respectively. Polarization to both of these cell states requires the presence of transforming growth factor β (TGF-β) and there is evidence of interconversion between the two phenotypes (Komatsu et al., 2014). Consistent with these observations, we found that Th17 and iTreg-stimulated $T_M$ cells were more similar to each other than to any other cell state and overlapped

in PCA space (**Figure 3.2A**). This similarity was captured by both proteome and transcriptome. In contrast, in $T_N$ the two cytokine-induced cell states formed separate groups. Importantly, both cell types expressed comparable levels of the TGF-β and IL2 receptors (**Supplementary Figure 3.2B-C**). To test whether Th17 and iTreg-stimulation induced the same phenotype in $T_M$, we compared gene expression between the two cell states (**Figure 3.2D,** **Supplementary Table 3.2** and **Supplementary Table 3.3**). Only 42 genes and no proteins were differentially expressed between the two cytokine conditions in $T_M$ (LFC > 1 at 0.05 FDR for RNA-seq and LFC > 0.5 at 0.1 FDR for proteomics). In contrast, in $T_N$ 733 genes and 455 proteins were differentially expressed between iTreg and Th17-stimulated cells (**Figure 3.2D**). In particular, iTreg-stimulated $T_N$ expressed higher levels of *FOXP3*, *IKZF4* and *LGALS3*, while Th17-stimulated $T_N$ expressed higher levels of *IL17F*, *TNFRSF8* and *PALLD*. Therefore, while $T_N$ acquire different phenotypes upon Th17 and iTreg-polarization, both cytokine conditions polarize $T_M$ towards the same cell state.

## Cell state specific gene signatures

Our results suggested that cytokines act in a cell type specific manner to induce five cell states in $T_N$ (Th1, Th2, Th17, iTreg and IFN-β) and three cell states in $T_M$ (Th1, Th17/iTreg and IFN-β, with no detectable Th2 response). As we observed high correlation between RNA and protein expression (**Supplementary Figure 3.3A-C**), we applied a multi-omics approach which leveraged both layers of information to derive cell state gene signatures (**Methods**). In brief, we identified differentially expressed RNA-protein pairs and asked if any of these pairs were present at a higher level in one cell state compared to the rest, obtaining cytokine specific *proteo-transcriptomic signatures* (**Methods**). The identified signatures were sensitive to relative changes in both RNA and protein, thus increasing our confidence that these are true cytokine-induced effects.

In $T_N$ we identified 105 signature genes across the five cell states (5 genes for Th1, 20 for Th2, 20 for Th17, 10 for iTreg, and 50 for IFN-β) (**Figure 3.2E** and **Supplementary Table 3.5**). The $T_N$ IFN-β signature contained known antiviral genes involved in RNAse L induction (*OAS2*, *OAS3*), GTPase activity (*MX1*, *MX2*) and cell lysis (*GZMA*, *GZMB*) (Randall & Goodbourn, 2008). Signatures of other $T_N$ states also included known hallmark genes, such as *GATA3* (Th2), *TBX21* (Th1) and *FOXP3* (iTreg) (**Figure 3.2E**), validating our approach. Additionally, we observed signature genes for Th1 (*ANXA3*), Th2 (*MAOA, LIMA1, MRPS26*), Th17 (*TNFRSF8, RUNX1, PALLD*) and iTreg (*LMCD1, LGALS3, CCL5*) which have not been previously described in the context of cytokine polarization.

We performed the same analysis for $T_M$, where we identified 162 signature genes across the three cell states (3 for Th1, 145 for Th17/iTreg and 14 for IFN-β genes) (**Figure 3.2E** and **Supplementary Table 3.5**). Since Th17 and iTreg-stimulated $T_M$ overlapped on both the RNA and protein levels, we treated them as a single group (Th17/iTreg). Several Th17/iTreg $T_M$ signature genes were present in the iTreg and Th17 signatures derived from $T_N$ (*CCL5, LGALS3, TNFRSF8*) or had been previously associated with one of the two phenotypes in the literature (*BACH2, BATF3, AHR*), suggesting that Th17/iTreg-stimulated $T_M$ might have overlapping functions with the Th17 and iTreg states in $T_N$. These signatures provide a valuable resource for future follow-up studies in specific biological contexts or disease settings.

## Single-cell RNA-seq reveals a T cell effectorness gradient

Our results showed that the gene expression programs induced in response to cytokines differ substantially between $T_N$ and $T_M$. Whilst $T_N$ constitute a uniform cell population, $T_M$ are known to be composed of numerous subpopulations including central ($T_{CM}$) and effector ($T_{EM}$) memory cells, as well as effector memory cells re-expressing CD45RA ($T_{EMRA}$). Given this heterogeneity, we speculated that the observed differences in cytokine responses could be explained by two scenarios: i) $T_M$ as a whole are unresponsive to certain cytokines, or ii) a subset of $T_M$ responds to cytokines, but bulk gene expression profiles are dominated by a larger proportion of unresponsive cells. To disentangle between these possibilities, we profiled single-cell gene expression in 43,112 $T_N$ and $T_M$, which included resting, Th0, Th2, Th17, and iTreg-stimulated cells.

First, we isolated $T_N$ and $T_M$ from four healthy individuals and quantified gene expression in the resting state using droplet-based scRNA-seq (G. X. Y. Zheng et al., 2017), with all replicates of the same cell type pooled together in a single droplet capture reaction (**Methods** and **Supplementary Figure 3.4**). In total, we profiled 5,269 resting CD4+ T cells (2,159 $T_N$ and 3,110 $T_M$ respectively), with an average of 1,146 genes detected per cell. We identified 64 highly variable genes, which we used for dimensionality reduction, embedding with the uniform manifold approximation (UMAP) (L. McInnes et al., 2018), and unsupervised clustering (**Methods**). We identified five distinct groups of cells (**Figure 3.3A**) which we annotated as $T_N$, $T_{CM}$, $T_{EM}$, $T_{EMRA}$ and natural T regulatory (nTreg) cells based on their expression of well established cell type markers (**Figure 3.3B** and **Supplementary Table 3.6**). $T_{EMRA}$ cells showed a distinct cytotoxic transcriptional profile (eg. *PRF1*, *CCL4*, *GZMA*, *GZMH*), consistent with previous observations (Y. Tian et al., 2017). The proportions of cells classified to each subpopulation were comparable across individuals (**Figure 3.3A**). Importantly,

although we confirmed that $T_N$ were a homogeneous group of cells, a small percentage of $T_{EMRA}$ cells were originally isolated with $T_N$ (as they re-express CD45RA) and could only be correctly identified at the single-cell level.

Strikingly, our results showed that CD4+ T cells do not comprise discrete subpopulations. Instead, CD4+ T cells formed one major population with multiple interrelated cell states which suggested the existence of a transcriptional continuum. To investigate the relationships between these cell states, we applied pseudotime analysis (Qiu et al., 2017). We observed that cells formed a continuous progression starting in $T_N$ and gradually progressing towards $T_M$ (**Figure 3.3C**). The cells at the beginning of this trajectory expressed high levels of naïve markers (e.g. *SELL, CCR7* and *LRRN3*), while the end of the trajectory was enriched in cells expressing cytotoxic molecules (e.g. *GZMA, GZMB* and *PRF1*), as well as cytokines and chemokines (*IL32*, *CCL4* and *CCL5)* (**Figure 3.3D** and **Supplementary Table 3.7**). A comparison of this trajectory with the previously assigned cluster labels confirmed that CD4+ T cells form a natural progression that starts with naive T cells ($T_N$), and advances towards central ($T_{CM}$), then effector ($T_{EM}$) and finally highly effector ($T_{EMRA}$) memory T cells (**Figure 3.3C**), with nTreg cells branching out as a separate cell state. Based on these observations, we reasoned that the observed CD4+ T cell continuum reflected the potential of cells to initiate a rapid and robust response upon stimulation, where cells that express more chemokine and cytokine transcripts in the resting state are able to rapidly secrete them upon stimulation. We refer to this property as *effectorness*.

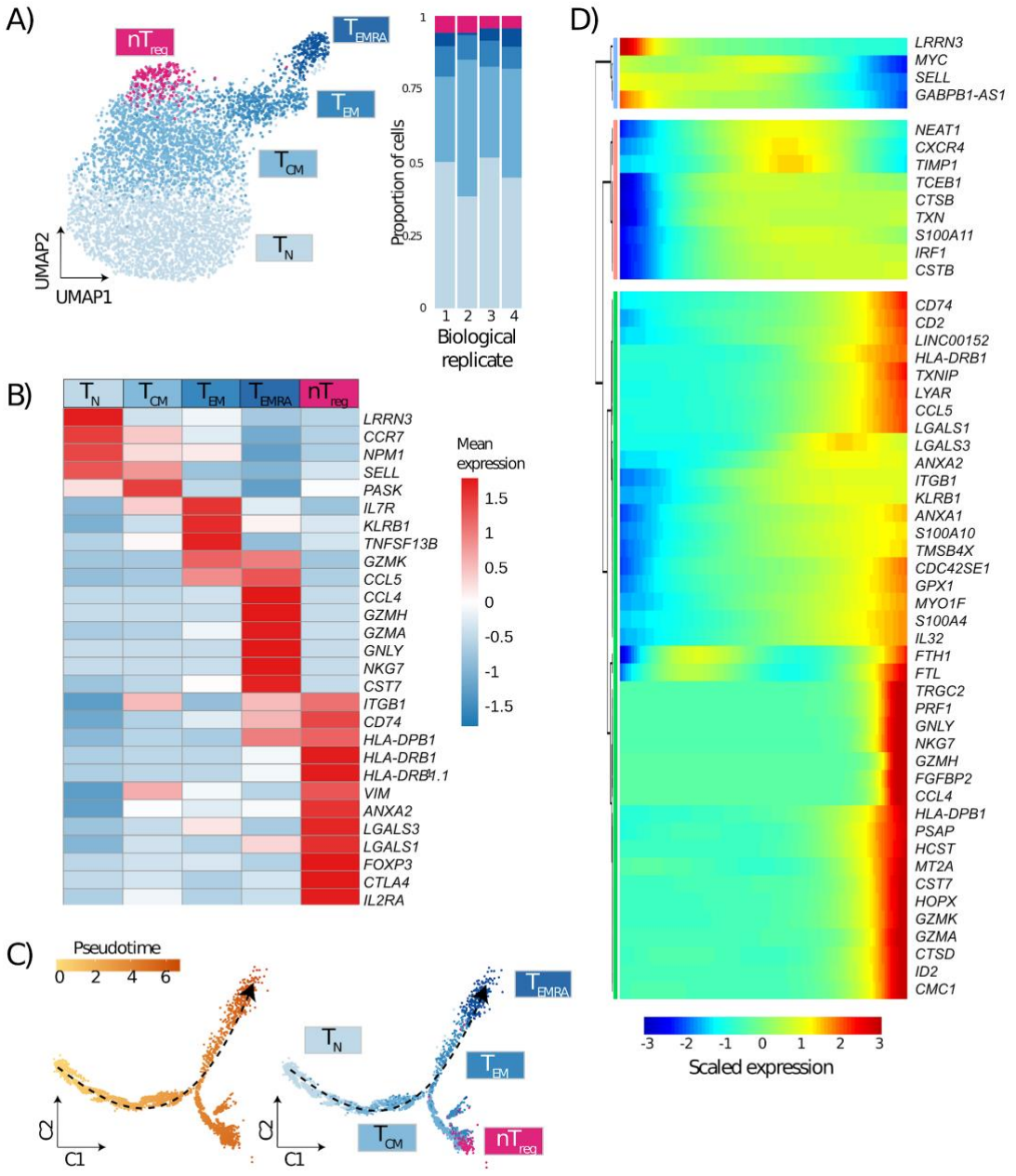



**Figure 3.3. Effectorness gradient in resting CD4+ T cells. A)** UMAP embedding of single-cell RNA-seq data from resting T cells. Colors represent cells assigned to five clusters, defined using top variable genes and unsupervised clustering. Bar plots represent the proportion of cells assigned to different clusters in each biological replicate. **B)** Expression of marker genes for each cell cluster (obtained using the Wilcoxon rank sum test) and well known cell type markers from the literature. Colors encode the mean expression of each gene in each cluster. **C)** Branched pseudotime trajectory. Each cell is colored by its pseudotime value (left panel) or its cluster label (right panel), as determined in panel A. **D)** Heatmap of genes that vary along the pseudotime trajectory (obtained using Monocle). The X axis represents cells ordered by pseudotime (from left to right) and different colors correspond to the scaled (Z-scored) expression of each gene in each cell.

Based on these observations, we reasoned that T cell effectorness could be linked to the activation history of cells and thus be accompanied by increased TCR clonality. We tested this

using a public data set containing matched single-cell transcriptomes and paired TCR-sequences of CD4+ T cells isolated from peripheral blood of colorectal cancer patients (L. Zhang et al., 2018) (**Methods**). We retrieved all cells in this study which corresponded to one of the subpopulations identified in our data-set ($T_N$, $T_{CM}$, $T_{EM}$ $T_{EMRA}$ and nTreg; **Supplementary Figure 3.5A**) and performed pseudotime ordering. The analysis revealed a similar progression to the one observed in our data (**Supplementary Figure 3.5B**), which was driven by a shared set of genes (**Supplementary Figure 3.5C**). Based on this progression, we assigned an effectorness value to each single T cell (**Methods**) and leveraged the paired TCR-sequences to test if effectorness was associated with clonal expansion. Cells with high effectorness values had reduced TCR clonotype diversity and higher numbers of expanded clones (**Methods** and **Supplementary Figure 3.5D**). To assess if this correlated with antigen specificity (i.e. effectorness being driven by a small number of clones specific for the same antigen), we used grouping of lymphocyte interactions by paratope hotspots (GLIPH) (Glanville et al., 2017), which groups together TCR-sequences predicted to recognize the same peptide-MHC complex (**Methods**). We observed that even cells in the highest effectorness range (effectorness > 0.7, which mostly marks $T_{EMRA}$ cells) comprised multiple independent TCR clonotypes, each of which was predicted to interact with a different peptide (**Supplementary Figure 3.5E**). Thus, we concluded that effectorness is a consequence of T cell activation history and is accompanied by clonal expansion, irrespective of antigen-specificity.

## Effectorness shapes the response of CD4+ T cells to activation

We next assessed if T cell effectorness influenced cell responses to TCR/CD28-activation, quantifying single-cell RNA expression in Th0-stimulated $T_N$ (2,543 cells) and $T_M$ (4,766 cells), with an average of 3,677 genes detected per cell. Pseudotime ordering of Th0-stimulated cells revealed an equivalent trajectory to that observed in resting cells, with $T_N$ gradually progressing towards $T_M$. This progression was accompanied by an increased expression of a similar gene set to that observed in the resting state (i.e. *CCL3, CCL4, GZMB, GMZA*; **Figure 3.4A and Figure 3.4B**). However, some genes were associated with this progression only upon TCR/CD28-activation (i.e. *IFNG* and *IL2)*.

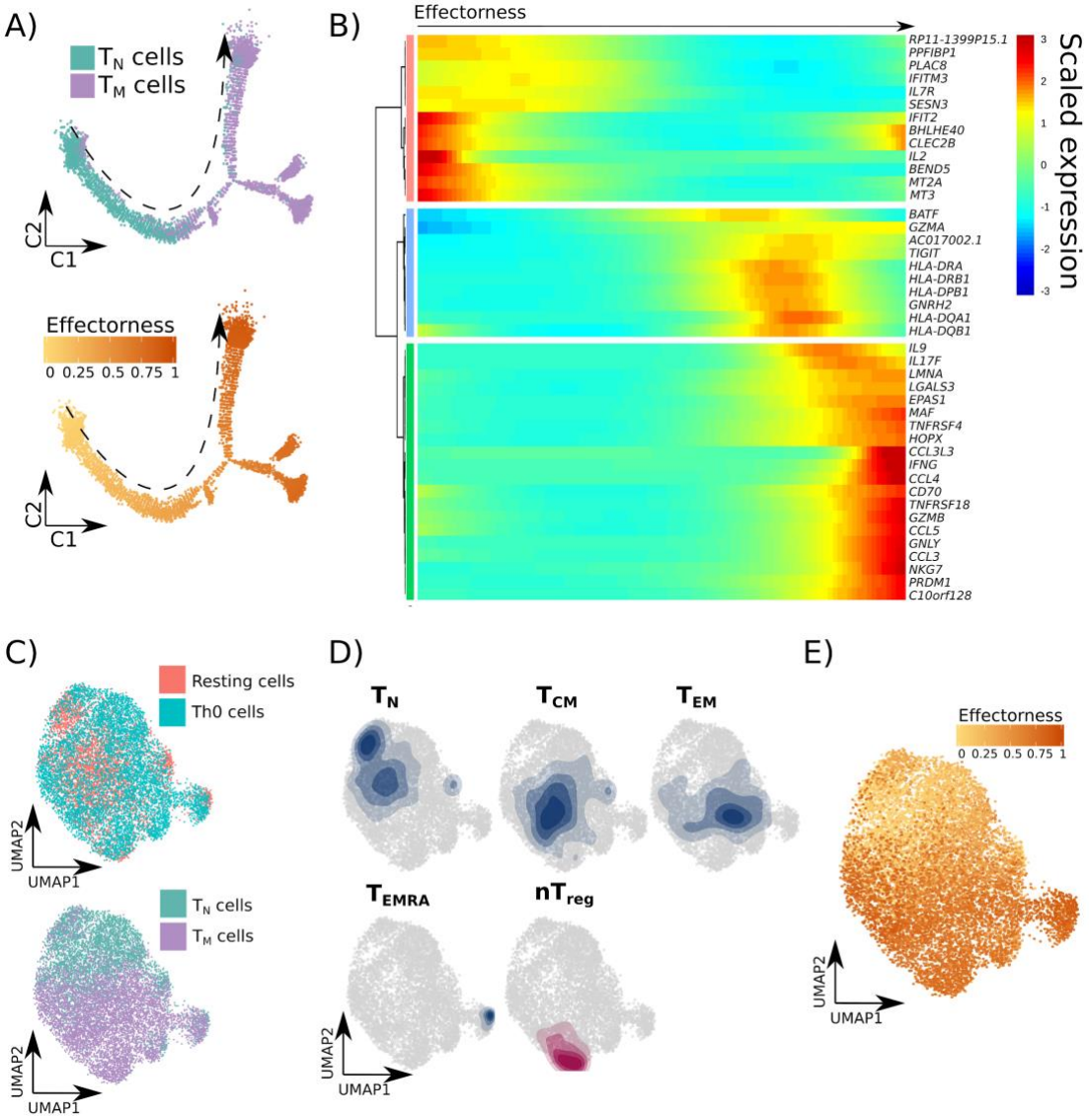



**Figure 3.4. The effectorness gradient is preserved after T cell activation. A)** Th0-stimulated $T_N$ and $T_M$ cells ordered into a branched pseudotime trajectory. Cells are colored by cell type (top panel) or effectorness (bottom panel). **B)** Heatmap of the most variable genes along the Th0 pseudotime trajectory (**Methods**). The X axis represents cells ordered by effectorness (from left to right) and colors correspond to the scaled (Z-scored) expression of each gene in each cell. **C)** Mapping of Th0-stimulated cells to their corresponding resting T cell populations. Integration of resting and Th0-stimulated cells using canonical correlation analysis (**Methods**) followed by UMAP embedding. Cells are colored by activation status (resting vs Th0, top panel) or cell type (bottom panel). **D)** Density plots highlighting the location of cell clusters as defined in resting state. **E)** UMAP embedding color-coded by the effectorness values of resting and stimulated cells.

The expression of classical $T_{CM}$ and $T_{EM}$ markers, such as *SELL* and *CCR7,* significantly changes following activation (**Supplementary Figure 3.6**), making the annotation of $T_M$ subpopulations challenging. Therefore, we used the transcriptome of resting T cells as a reference and mapped Th0-stimulated cells to their most similar unstimulated equivalent in order to annotate T cell subpopulations upon stimulation (**Methods**). Following integration of

resting and stimulated cells (**Figure 3.4C**), we observed that the continuum observed in the resting state persisted after activation, with a gradual progression of cells from $T_N$ to $T_{CM}$, $T_{EM}$ and $T_{EMRA}$ (**Figure 3.4D** and **Supplementary Table 3.7**). Finally, we integrated these results with effectorness scores, which were calculated independently for resting and Th0-stimulated cells. Independently of stimulation, cells with the lowest effectorness localized to the $T_N$ area, while cells with increasingly higher effectorness localized to the $T_{CM}$, $T_{EM}$ and $T_{EMRA}$ areas (**Figure 3.4E**). Thus, we concluded that T cell effectorness is detectable before and after T cell activation. Moreover, we found that T cells of different effectorness respond differently to TCR/CD28-stimulation, differentially regulating cytokines such as *IFNG* and *IL2*.

## Effectorness shapes the response of CD4+ T cells to cytokines

We next assessed if T cell effectorness influenced cell responses to cytokine polarization. To achieve this, we analyzed single-cell RNA expression in Th2, Th17 and iTreg-stimulated $T_N$ and $T_M$. We separately ordered Th2, Th17 and iTreg cells into branched pseudotime trajectories (**Supplementary Figure 3.7**) and showed that cells exposed to cytokine polarization can also be ordered by their effectorness. (**Supplementary Figure 3.7** and **Supplementary Table 3.7**). We then merged the data obtained from Th0 and cytokine conditions into a single dataset. To compare the effectorness values after merging, we derived a unified effectorness measurement by scaling the values inferred from independent cytokine-specific trajectories (**Methods**). The final data set contained transcriptomes of 37,843 cells, of which 18,786 were $T_N$ and 19,057 were $T_M$. $T_N$ and $T_M$ formed a single group of cells in UMAP space, but separated by cell type into two different areas (**Figure 3.5A**). This separation correlated with the effectorness gradient (**Figure 3.5A**), suggesting that T cell effectorness is amongst the strongest drivers of gene expression variation.

In addition to their separation by effectorness, $T_N$ and $T_M$ exposed to different cytokines also localized to different areas of the UMAP space (**Figure 3.5B**), suggesting that cytokine polarization generates distinct T cell states and is a second major driver of transcriptional variation (**Figure 3.5C**). For example, iTreg-stimulated $T_N$ cells localized to an area with high expression of *CTLA4*, while the area associated with Th17-stimulated $T_N$ showed high *RORA* expression. Similarly, the area enriched in Th17-stimulated $T_M$ showed higher levels of *IL17F* (**Figure 3.5C**). Cells in these areas also showed higher expression of the corresponding signature genes derived from bulk RNA and protein expression (**Supplementary Figure 3.8**).

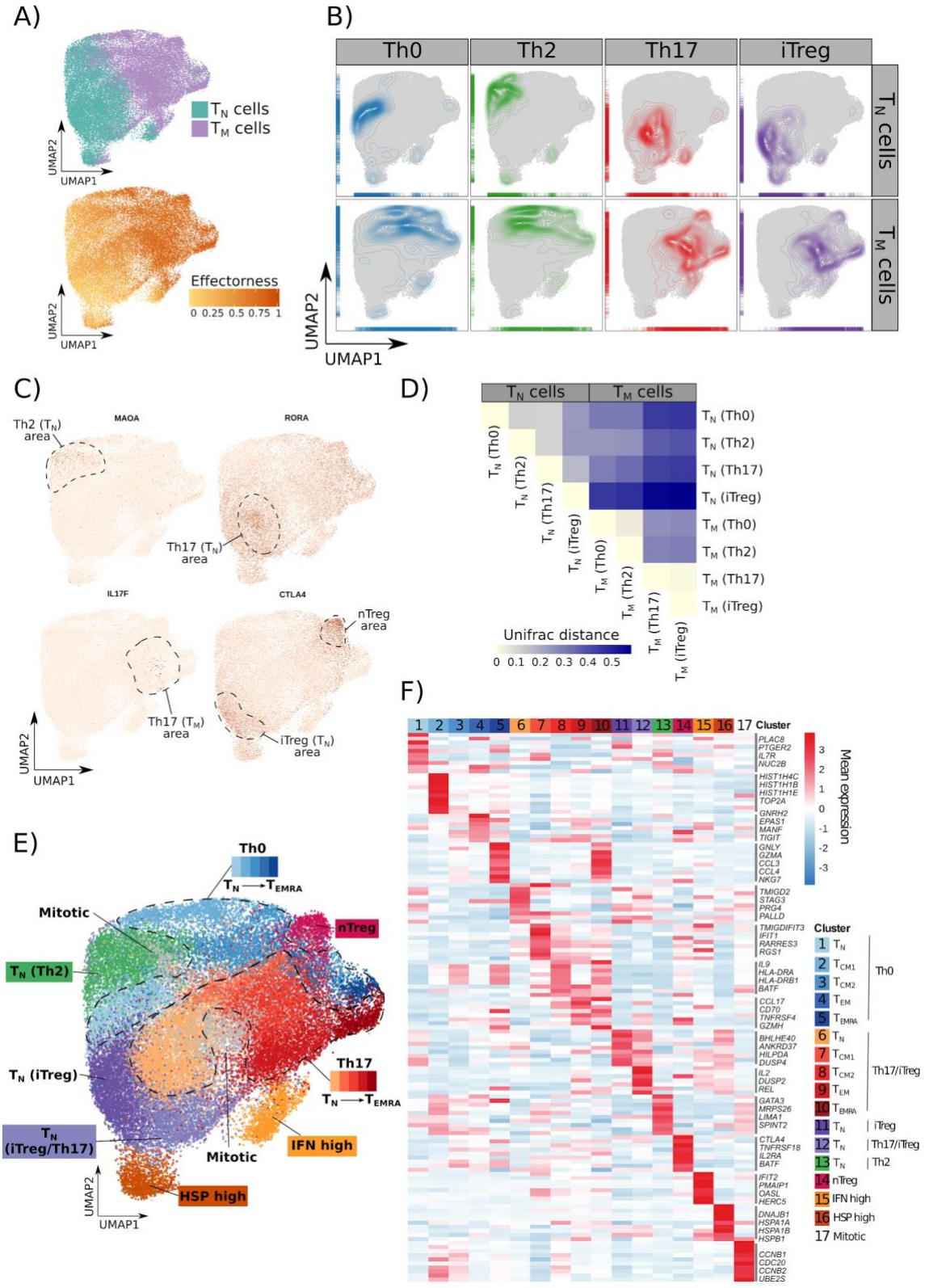



**Figure 3.5. The effectorness gradient persists after cytokine-induced T cell polarization. A)** UMAP embedding of stimulated T cells. Cells are colored by cell type (top panel) or effectorness value (bottom panel) **B)** Density plots highlighting cells based on the cytokine conditions they come from. **C)** Expression of selected cytokine-specific markers obtained from our bulk RNA-seq analysis. Cells are colored by their expression level of each marker gene. **D)** UniFrac distances between $T_N$ and $T_M$ cells exposed to different cytokines, summarized in a correlation plot. **E)** Annotation of 17 cell clusters identified from unsupervised clustering using the top variable genes. Each cluster is annotated based

on either the genes with highest expression or the effectorness and cytokine condition of the cells contained within it. **B)** Heatmap of the top 10 markers of each cluster (Wilcoxon rank sum test). Colors encode the mean expression of each gene in each cluster. Labels were added to a number of example genes for each cluster.

We next asked whether the absence of response to Th2-stimulation in $T_M$ was characteristic of the entire population of cells or if a small group of cells responded to Th2-stimulation but was masked by a majority of unresponsive cells. Interestingly, Th0 and Th2-stimulated $T_M$ localized to the same UMAP area (**Figure 3.5B**). We used UniFrac distances (Lozupone & Knight, 2005), to formally test if Th0 and Th2-stimulated $T_M$ overlapped or formed different groups. A UniFrac distance of 0 indicates that cells from two conditions localise to the same cluster, while a distance of 1 indicates that the two conditions form entirely separate clusters. We confirmed that Th0 and Th2-stimulated $T_M$ overlapped substantially (UniFrac distance = 0.047) (**Figure 3.5D**), indicating that none of the subpopulations of $T_M$ were capable of responding to Th2-stimulation. Thus, the observed lack of response was a general characteristic of all $T_M$.

Our observations from the bulk data also showed that $T_M$ polarize to the same cell state in response to Th17 and to iTreg-stimulation. We confirmed this at the single cell level, where cells from these two conditions localized to the same UMAP areas. The UniFrac distance between these cell states was 0.015 in $T_M$, compared to 0.164 in $T_N$ (**Figure 3.5D**). Thus, we concluded that in response to Th17 and iTreg-stimulation, $T_M$ converge on the same cell state. This is not driven by any subpopulation of $T_M$, but is a general characteristic of memory T cell biology. Interestingly this population expressed high levels of *IL17F,* suggesting that iTreg-stimulation in $T_M$ induces a Th17-like phenotype.

These results show that the transcriptome of T cells is shaped by the combination of two factors: T cell effectorness and cytokine-stimulation. Despite being separate biological variables, we hypothesized that these two axes of variation could interact to determine the transcriptional profile of each single cell. Thus, we performed unsupervised clustering of the entire data set and annotated the resulting clusters based on their average effectorness values as well as the cytokines to which they had been exposed. This resulted in 17 clusters. (**Figure 3.5E** and **Supplementary Table 3.6**). For instance, we identified clusters of Th0, Th2, Th17 and iTreg-stimulated $T_N$, as well as a cluster formed of a mix of Th17 and iTreg-stimulated $T_N$, characterised by high expression of TNF-signaling molecules (eg. *IL2, DUSP2, REL, TNF*) (**Figure 3.5E-F**). Moreover, we identified four clusters of Th0-stimulated $T_M$, which we annotated as stimulated $T_{CM1}$, $T_{CM2}$, $T_{EM}$, and $T_{EMRA}$ cells. The same was true for Th17/iTreg-stimulated $T_M$, which localized into four groups with different effectorness ($T_{CM1}$, $T_{CM2}$, $T_{EM}$, and

$T_{EMRA}$). We also identified a group of nTreg cells, which expressed canonical markers such as *FOXP3, CTLA4* and *TNFRSF8*. This cluster contained a comparable number of cells from all cytokine conditions, suggesting that these cytokines do not affect the nTreg transcriptome. Finally, we observed a small cluster formed of $T_N$ and $T_M$ expressing high levels of IFN-induced genes, as well as a cluster of cells expressing heat shock proteins (HSPs) and other markers of cellular stress (**Figure 3.5E-F**). In conclusion, T cell effectorness and cytokine-induced polarization act jointly to modify the transcriptome of single T cells. This suggests that T cell effectorness shapes the response of T cells to cytokines.

To understand how effectroness shapes the T cell response to cytokines, we modelled gene expression as a function of effectorness, cytokine condition (i.e. Th0, Th2, Th17 or iTreg), and the interaction between them (**Methods**) using linear models. Our model accounted for four possible mechanisms (**Figure 3.6A**): i) gene expression is modulation by cytokines irrespective of effectorness, ii) gene expression is a function of effectorness irrespective of cytokines, and gene expression is modulated by both effectorness and cytokines acting iii) independently or iv) jointly (interaction effect). We identified 210 genes significantly associated with effectorness (estimated FDR-corrected p value for $\beta < 0.05$, see **Methods** and **Supplementary Table 3.8**). Of these, the majority (203 genes) were further regulated by cytokines. In particular, 12 genes showed independent effects of cytokine-stimulation and effectorness (estimated FDR-corrected p value for $\gamma < 0.05$, see **Methods** and **Supplementary Table 3.8**), while 191 showed an interaction effect (estimated FDR-corrected p value for $\delta < 0.05$ , see **Methods** and **Supplementary Table 3.8**). Within the genes with interaction effects, 12 showed an effectorness dependency only in the presence of a given cytokine, while 179 showed effectorness dependency ubiquitously (across all cytokine conditions, **Supplementary Table 3.8**), with the strength of this effect varying across cytokine conditions.

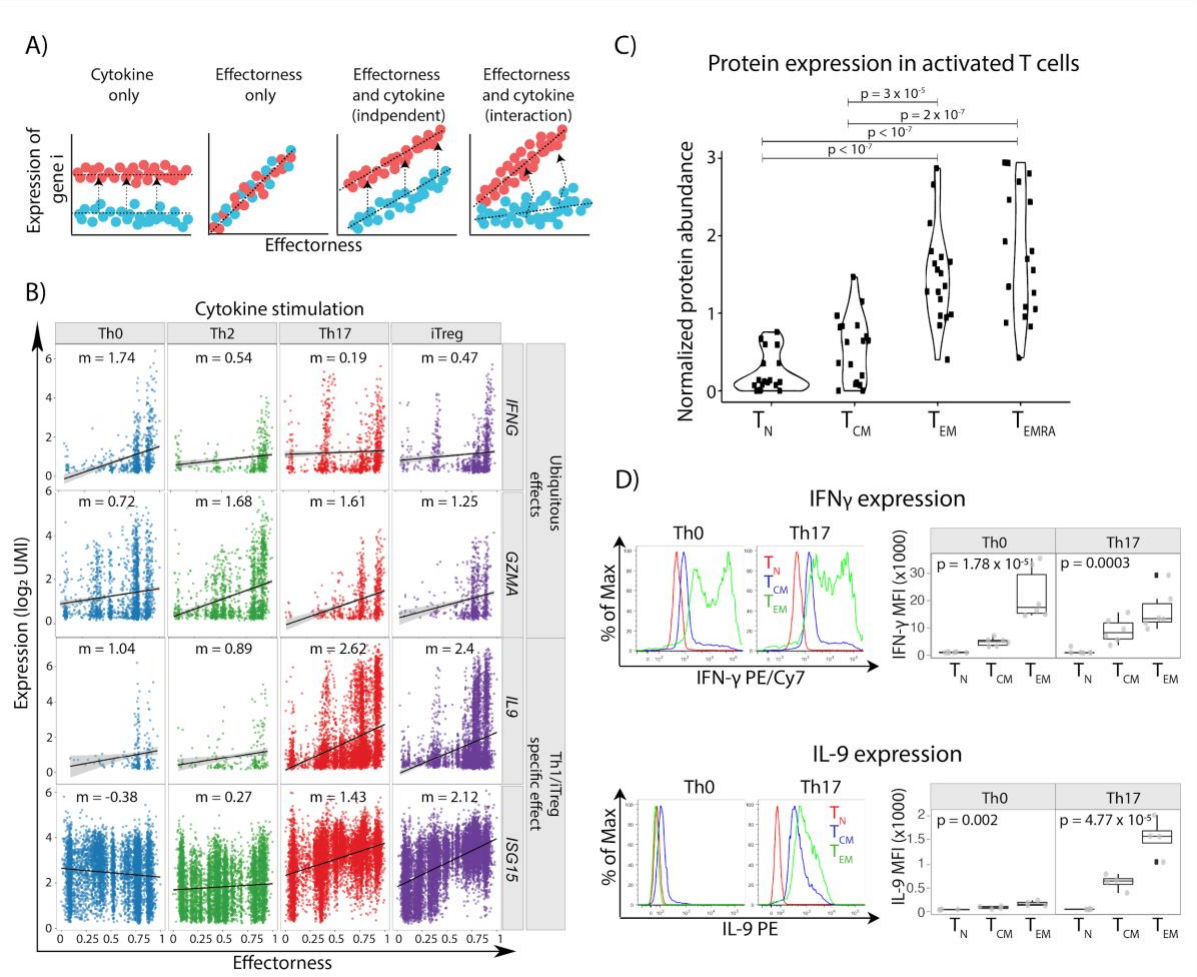



**Figure 3.6. Interactions between effectorness and cytokine condition regulate gene expression.** **A)** Schematic representation of a gene expression interaction model. Effectorness and cytokine conditions were incorporated into a linear model with an interaction term (**Methods**). Genes were assigned to four groups: genes induced by cytokine-stimulation regardless of effectorness (left panel), genes which correlate with effectorness regardless of cytokine-stimulation (central left panel), and genes which correlate with both effectorness and cytokine-stimulation either independently (central right panel) or through an interaction (right panel). **B)** Plots of gene expression (Y axis) as a function of effectorness (X axis), with cells stratified by cytokine condition. Two example genes significantly associated with effectorness regardless of cytokine conditions (top panels) and two example genes with a strong interaction between effectorness and Th17 or iTreg-stimulation (bottom panels). Each dot represents a single cell. **C)** Protein expression levels for 18 genes associated with effectorness in activated $T_N$, $T_{CM}$ and $T_{EM}$ and $T_{EMRA}$ cells regardless of cytokines condition (genes used in the analysis: *ACTB, CCL3, CCL5, CTSW, GNLY, GZMA, GZMB, HLA-DPB1, HLA-DQA1, HLA-DRA, HLA-DRB1, HOPX, IFNG, LGMN, LMNA, NFKBIA, TMEM173, TNFRSF18*). Expression values were obtained from a publicly available quantitative proteomics data set (**Methods**). Significance was calculated using one-way ANOVA ($p=3.42x10^{-12}$) and group means were compared using Tukey's Honest Significant Difference test. Only significant p-values are shown ($p<0.05$). Each dot represents an individual gene (n=18). **D)** Levels of IFNγ and IL-9 protein in Th0 and Th17-stimulated $T_N$, $T_{CM}$ and $T_{EM}$ cells as assessed by flow cytometry. Representative cytometry histograms of IFNγ and IL-9 expression (left panels) and mean fluorescence intensity (MFI) of cytokine-expressing cells (right panels). P-values were calculated using one-way ANOVA. Each dot represents a biologically independent sample.

Although statistically significant, some of these effects were too small to be biologically meaningful. Thus, we filtered genes by their effect sizes and identified 24 genes with a strong effectorness dependency across all cytokine conditions (**|β|** > 0.5, see **Methods**). These

included the *TNFRSF4* (encoding for OX40), which is known to be critical in the maintenance of memory T cell responses (Mousavi et al., 2008), as well as effector molecules such as granulysin (*GNLY*), *GMZA*, *CCL3* and *IFNG* (**Figure 3.6B and Supplementary Table 3.8**). The expression of these genes increased proportionally to T cell effectorness. In addition, we identified 37 and 16 genes strongly associated with effectorness and with particularly large effects upon Th17 and iTreg-stimulation, respectively ($|\boldsymbol{\delta}| > 0.5$ for the respective cytokine, see **Methods**). These genes included cytokines like *IL2* (which decreased with effectorness upon Th17 and iTreg-stimulation) and *IL9* (which increased with effectorness in these conditions) (**Figure 3.6B** and **Supplementary Table 3.8**). Moreover, genes induced by type I IFNs (eg. *ISG15, IFIT1, IFIT2, IFIT3*) also increased with effectorness upon iTreg and Th17-stimulation. This is in line with our observations from bulk RNA and protein expression, where we found that the type I IFN response was differentially regulated in $T_N$ and $T_M$ in response to Th17 and iTreg-stimulation (**Figure 3.2C**).

To validate these results, we identified all the genes which contribute to the effectorness gradient and which have a large effect across all cytokine conditions (**Supplementary Table 3.8**). We then assessed their expression levels in a previous proteomics study which profiled resting and activated immune cell populations (Rieckmann et al., 2017). Of the 24 effectorness-associated genes with large and ubiquitous effects, 18 were present in this data set, all of which followed the same trend observed in our scRNA-seq data (**Figure 3.6C**). Moreover, we experimentally validated two cytokines (IFNγ and IL-9), which showed strong association with effectorness either across all conditions or upon Th17/iTreg-stimulation. We isolated CD4+ $T_N$, $T_{CM}$ and $T_{EM}$ directly from blood (**Supplementary Figure 3.9**), performed Th0 and Th17-stimulation, and quantified the production of IFNγ and IL-9 protein upon restimulation (**Methods**). As observed in scRNAseq, the levels of secreted IFNγ increased proportionally to effectorness in both Th0 and Th17-stimulated cells (**Figure 3.6D**). In contrast, the levels of secreted IL9 only marginally correlated with effectorness in Th0 cells, but substantially increased with effectorness in Th17-stimulated cells (**Figure 3.6D**). This confirmed our observations from the transcriptome and suggested that key T cell functions such as cytokine secretion are under the control of both effectorness and environmental cues, and that these two factors can interact.

In summary, we showed that CD4+ $T_N$, $T_{CM}$, $T_{EM}$, and $T_{EMRA}$ cells form a transcriptional continuum characterized by an effectorness gradient. This gradient is determined by the expression of cytokine, chemokine and granzyme genes and affects the response of cells to cytokine polarization.

## Discussion

Cytokines have been extensively studied in the context of naïve T cells, but the response of memory T cells to cytokines remains understudied. Here we analyzed the effects of cytokines on memory T cell gene expression and compared these to the responses of naive T cells. We demonstrate that early gene expression changes in both $T_N$ and $T_M$ cells are dominated by the response to TCR and CD28, while cytokine-induced changes are apparent only at later stages of stimulation. This suggests that polarization to T helper subsets occurs after the initiation of T cell activation and fine-tunes the response of T cells.

Our multi-omic data further show that the response to cytokines differs between $T_N$ and $T_M$. While $T_N$ responded to all tested cytokine conditions, acquiring a distinctive phenotype, $T_M$ did not respond to Th2 polarization. Furthermore, while $T_N$ induce hallmark markers such as *FOXP3* and *CTLA4* upon iTreg-stimulation, $T_M$ converge on the same cell state in response to both Th17 and iTreg-stimulation. This state is characterized by high levels of *IL17F,* suggesting that $T_M$ do not acquire a regulatory phenotype upon iTreg polarization. This is particularly relevant in the context of immune-mediated diseases, given that $T_M$ increases with age (Saule et al., 2006), potentially leading to a pro-inflammatory response to TGF-β.

In contrast, the response to IFN-β was conserved between $T_N$ and $T_M$ and was apparent within 16h of stimulation. This is in line with type I interferon responses being rapidly triggered to prevent viral replication. Upon IFN-β stimulation, both $T_N$ and $T_M$ upregulated genes involved in RNAse L induction (*OAS2*, *OAS3*) (Sadler & Williams, 2008), GTPase activity (*MX1*, *MX2*) (Verhelst et al., 2012), and cell lysis (*GZMA*, *GZMB*) (Randall & Goodbourn, 2008).

Using single cell transcriptomics we further show that CD4+ T cells form a continuum, characterized by a progression with the structure $T_N \rightarrow T_{CM} \rightarrow T_{EM} \rightarrow T_{EMRA}$, with nTregs branching out separately. This progression is accompanied by upregulation of chemokine and cytokine genes, suggesting that T cells at the end of this trajectory are poised to initiate a rapid effector response upon activation. We thus refer to this property as effectorness. A similar gradient is present in innate T cells, as shown by a recent study where higher expression of cytokines, chemokines and granzymes negatively correlated with ribosome synthesis and proliferative capacity (Gutierrez-Arcelus et al., 2019). Additionally, we show that high T cell effectorness is accompanied by clonal expansion and therefore reflects the activation history of each single-cell. These results agree with observations of CD8+ T cells upon viral infection, where exposure to CMV results in T cells that express high levels of CCL4 and perforin (Appay et al., 2000). The similarity of this profile with the most effector cells in our study suggests that

an equivalent gradient exists in CD8+ T cells, where chronic exposure to CMV drives repeated clonal expansion, generating highly effector memory cells. Importantly, previous studies suggest this process could be impaired in HIV (Appay et al., 2000; Champagne et al., 2001). The effectorness gradient described here also recapitulates observations from tissue-resident immune cells. In particular, a previous study described the generation of memory T cells in the fetal intestine, with an equivalent trajectory from $T_N$ to $T_M$ (N. Li et al., 2019) accompanied by downregulation of *CCR7* and upregulation of cytokines like *IL32*. Thus, our results could explain how $T_M$ adapt to inflammation within tissues.

It is particularly interesting to speculate whether our data could give further insights into memory formation. At first glance, the transcriptional continuum observed in our study would seem to resemble the previously proposed linear model of memory T cell differentiation. Briefly, $T_{CM}$ and $T_{EM}$ cells were first described in 1999 as two subsets of memory cells in peripheral blood which differ in their tissue homing properties (F. Sallusto et al., 1999). While $T_{CM}$ cells express CCR7 and CD62L (SELL), which enables them to home to lymphoid tissues, $T_{EM}$ cells lack these markers and instead migrate to non-lymphoid tissues, where they are able to rapidly respond to secondary challenges but at the expense of proliferative capacity. This landmark study further showed that $T_{CM}$ cells can acquire a $T_{EM}$ phenotype upon restimulation (F. Sallusto et al., 1999). This resulted in the proposal of a linear differentiation model, where naive T cells first encounter antigen and differentiate into effector cells, a small fraction of which differentiate into $T_{CM}$, which ultimately give rise to $T_{EM}$ cells in secondary responses (Kaech et al., 2002). While this indeed agrees with our observations of transcriptional similarity between T cell subsets, caution is warranted, since the data from our study could also be compatible with alternative models of memory differentiation. In particular, two competing models have since been proposed. Firstly, the divergent differentiation model suggests that, upon stimulation, naive cells can form both effector and memory cells at the same time (Gasper et al., 2014; Kaech et al., 2002). According to this model, $T_{CM}$ and $T_{EM}$ cells could also arise simultaneously from the same progenitor. A recent single-cell study which followed memory T cell development in mice upon malaria infection seems to support this theory (Soon et al., 2020). Divergent differentiation is thought to be mediated by asymmetric cell division, where a single naive cell would generate two distinct daughter cells that differentially inherit proteins and cytoplasmic components, analogously to the mechanisms operating in stem cells (Arsenio et al., 2015). Interestingly, there is evidence that the capacity to perform asymmetric division inversely correlates with the capacity to produce effector molecules in CD8+ T cells (Borsa et al., 2019). This model could be compatible with our observations of increased levels of effector molecules in $T_{EM}$ and $T_{EMRA}$ cells, both of which are similar to the CD8 counterparts reported to perform the least amount of asymmetric division (Borsa et al., 2019). This model

could also explain early observations suggesting that a small fraction of $T_{EM}$ cells can convert into $T_{CM}$ in the absence of antigen (Wherry et al., 2003) Secondly, the decreasing potential model proposes that naive cells differentiate into either effector or memory cells depending on the strength of stimulation they experience (Gasper et al., 2014; Kaech et al., 2002). In particular, cells experiencing prolonged interactions with APCs or high levels of activation signals would become effector cells, while cells receiving intermediate levels of signal would become memory cells and retain their proliferative capacity. While the impact of stimulation strength is difficult to assess in our system, it should not be disregarded. For example, the increase in clonality we observe in $T_{EM}$ and $T_{EMRA}$ cells could either support the linear differentiation model, with $T_{EM}$ cells progressively expanding and becoming $T_{EMRA}$ upon serial restimulation, or the progressive differentiation model, where $T_{EMRA}$ cells could have arisen in individuals previously exposed to chronic viral infection due to high antigen load. In sum, the data collected from our study could be explained by any of the three main models of memory T cell differentiation and we do not advocate for any model in particular. Future work applying single-cell sequencing and lineage tracking to systems where memory development can be followed *in vivo* would help settle this question.

Having characterized the heterogeneity of the CD4+ T cell compartment and demonstrated the presence of an effectorness gradient, we then showed that different levels of effectorness result in differences in the way CD4+ T cells respond to TCR/CD28 signals. In particular, we identified genes that increase proportionally to effectorness upon T cell activation, such as *IFNG*. These differences, as well as differences between naive and memory T cells more generally, could be mediated by several previously proposed mechanisms. Firstly, epigenetic differences between naive and memory T cells are well documented (Kanno et al., 2012), and it could be possible for subpopulations such as $T_{EM}$ and $T_{EMRA}$ to have more regions of open chromatin, as well as poised regulatory elements ready to rapidly produce effector molecules. Secondly, memory cells are also known to store pre-formed costimulatory molecules, which can be readily used upon stimulation (Gasper et al., 2014; Koguchi et al., 2012). The amount of these molecules could be higher in memory subsets with higher effectorness. Finally, it has been reported that, while TCR receptors are distributed homogeneously across the surface of naive cells, in memory cells TCRs cluster together in patches across the membrane, where they form oligomeric receptor complexes (Kumar et al., 2011). This pre-clustering of receptors increases sensitivity and reduces the time needed to induce activation (Castro et al., 2014). It is possible that higher effectorness could correlate with higher levels of pre-formed TCR complexes, ultimately helping explain the enhanced effector response observed when stimulating subsets of cells such as $T_{EM}$ and $T_{EMRA}$.

Finally, we showed that the response to cytokine-induced polarization also differs between T cell subsets depending on their effectorness level. For example, *IL9* showed a strong effectorness dependency in response to Th17/iTreg-stimulation. $T_M$ are known to upregulate IL-9 in response to TGF-β (Putheti et al., 2010), and this cytokine is also known to reprogram Th2 cells to an IL-9 secreting phenotype (Veldhoen et al., 2008). Here, we refine this observation, showing that IL-9 upregulation is driven by $T_M$ cells of higher effectorness (i.e. $T_{EM}$ and $T_{EMRA}$). This is important given the role of TGF-β in Th17 cell biology and the substantial Th17 diversity *in vivo* (Omenetti et al., 2019). Our study suggests that cells with high effectorness which infiltrate tissues might respond strongly to the local cytokine environment. Future scRNA-seq studies of inflamed tissues will provide an opportunity to investigate these effects in greater detail. Understanding this will be key in the development of drug targets for autoimmune disease, as Th17-cytokines are known to promote inflammation in immune-mediated diseases such as MS (Babaloo et al., 2013; Jamshidian et al., 2013; Lin & Edelson, 2017).

## Methods

### Cell isolation and *in vitro* stimulation

Blood samples were obtained from six individuals for the bulk assays (with naïve and memory T cells being isolated from three independent individuals, respectively) and from four additional individuals for single-cell RNA-seq. All individuals were healthy males living in the UK of 56.4 years of age on average (sd = 12.41 years). Human biological samples were ethically sourced and their research use was in accordance with the terms of the informed consents under an IRB/EC approved protocol (15/NW/0282). PBMCs were isolated using Ficoll-Paque PLUS (GEhealthcare, Buckingham, UK) density gradient centrifugation. Naïve (CD25- CD45RA+ CD45RO-) and memory (CD25- CD45RA- CD45RO+) CD4+ T cells were isolated from PBMCs using EasySep® naïve CD4+ T cell isolation kit and memory CD4+ T cell enrichment kit (StemCell Technologies, Meylan, France) according to the manufacturer's instructions. T cells were then stimulated with anti-CD3/anti-CD28 human T-Activator Dynabeads® (Invitrogen) at a 1:2 ratio of beads to T cells. Cytokines were added at the same time as the stimulus (see **Supplementary Table 3.1** for a full list of the cytokines used with product details and exact concentrations). Cells were harvested after 16 hours and 5 days of stimulation.

### Bulk RNA-sequencing

A total of 3 x $10^5$ cells were resuspended in 500 µl of TRIzol™ and stored at -80°C until further processing. For RNA isolation, samples were thawed at 37°C, 100 µl chloroform were added and samples were centrifuged for 15 min at 4°C and 10,000g. The aqueous phase was

collected and mixed at a 1:1 ratio with 70% ethanol (Qiagen). RNA was isolated from this mixture using the RNeasy MinElute Kit (Qiagen), with RNA quality assessed using a Bioanalyzer RNA 6000 Nano Chip (Agilent Technologies). All samples had an RNA integrity number (RIN) higher than 7, with a mean RIN of 9.35. Finally, sequencing libraries were prepared using the Illumina TruSeq protocol and sequenced on an Illumina HiSeq 2500 platform using v4 chemistry and 75 bp paired-end reads.

## Proteomics

Pellets formed of up to 3 x $10^6$ cells were isolated and washed twice with PBS buffer, dried and stored at -20℃ until protein extraction. Cell pellets were then lysed in 150 µl 0.1 M triethylammonium bicarbonate (TEAB) buffer (Sigma Aldrich) supplemented with 0.1% SDS and Halt protease and phosphatase inhibitor cocktail (100X, Thermo #78442). Pulse probe sonication (40% power, 4℃ and 20 seconds) was performed twice using EpiShear™, after which samples were incubated for 10 minutes at 96 °C. Protein from cell lysates was quantified using the quick start Bradford protein assay (Bio-Rad) as specified by the manufacturer's instructions. Protein samples were finally divided into aliquots of up to 100 µg. Protein aliquots were reduced with 5 mM tris-2-carboxymethyl phosphine (TCEP) buffer (Sigma Aldrich) and incubated for 1 hour at 60 °C to reduce disulfide bonds. Iodoacetamide (IAA) was added to a final concentration of 10 mM and samples were incubated for 30 minutes at room temperature in the dark. Pierce Trypsin (Thermo Scientific) was then added at a mass ratio of 1:30, and samples were incubated overnight for peptide digestion. Digested protein samples were diluted to a total volume of 100 µl in 0.1 M TEAB buffer. TMT reagents (Thermo Scientific) supplemented with 41 µl anhydrous acetonitrile were added to the corresponding protein samples. After 1 hour, the reaction was quenched using 8 µl 5% hydroxylamine. Samples were then combined into a single tube and dried using a speedvac concentrator. Dry samples were stored at -20˚C until fractionation. High pH Reverse Phase (RP) peptide fractionation was performed with the Waters XBridge C18 column (2.1 x 150 mm, 3.5 µm) on a Dionex™ UltiMate 3000 HPLC system. A 0.1% solution of ammonium hydroxide was used as mobile phase A, while mobile phase B was composed of acetonitrile with 0.1% ammonium hydroxide. The TMT-labelled samples were reconstituted in 100 µl mobile phase A, centrifuged and injected into the column, which operated at 0.2 ml/min. The fractions collected from the column were dried with the SpeedVac concentrator and stored at -20℃ until the MS analysis.

Liquid Chromatography-Mass Spectrometry (LC-MS) was performed using a Dionex™ UltiMate 3000 HPLC system (Thermo Scientific) coupled with the Orbitrap Fusion Tribrid Mass Spectrometer (Thermo Scientific). Dried samples were reconstituted in 40 µl 0.1% formic acid, of which 7 µl were loaded to the Acclaim PepMap 100 trapping column (100 µm x 2 cm, C18, 5 µm, 100Ä) at a flow rate of 10 µl/min. Multi-step gradient elution was performed at 45 °C using the Dionex™ Acclaim PepMap RSLC capillary column (75 µm x 50 cm, 2 µm, 100Ä). A 0.1% solution of formic acid was used as mobile phase A, and a 80% acetonitrile, 0.1% formic acid solution as mobile phase B. Precursors were selected with mass resolution of 120k, AGC $4 \times 10^5$ and IT 50 ms were isolated for CID fragmentation with quadrupole isolation width of 0.7 Th. Collision energy was set at 35%. Furthermore, MS3 quantification spectra were acquired with 50k resolution via further fragmentation for the top 7 most abundant CID fragments in the Synchronous Precursor Selection (SPS) mode. Targeted precursor ions were dynamically excluded for 45 seconds.

Raw data were processed in Proteome Discoverer (v2.2) with SequestHT search engine (Thermo Scientific) using reviewed UniProt human protein entries for protein identification and quantification (The UniProt Consortium, 2017). The precursor mass tolerance was set at 20 ppm and the fragment ion mass tolerance was 0.02 Da. Spectra were searched for fully tryptic peptides with maximum 2 miss-cleavages. TMT 6-plex at N-terminus/K and Carbamidomethyl at C were used as static modifications. Dynamic modifications included oxidation of M and deamidation of N/Q. Peptide confidence was estimated with the Percolator node. Peptide FDR was set at 0.01 and validation was based on q-value and decoy database search. The reporter ion quantifier node included a TMT10plex quantification method, with an integration window tolerance of 15 ppm and integration method based on the most confident centroid peak at the MS3 level. Only unique peptides for the protein groups were used for quantification. Peptides with average reporter signal-to-noise less than 3 were excluded from protein quantification.

### Single-cell RNA-sequencing

Cells were resuspended in RPMI media to obtain a single-cell suspension with high cell viability. Next, cells were stained with the live/dead dye 4',6-diamidino-2-phenylindole (DAPI) and dead cells were removed using fluorescence-activated cell sorting (FACS). Live cells were resuspended in PBS buffer and recounted using acridine orange/propidium iodide (AO/PI) staining and the Nexcelom Cellometer Auto 2000 Cell Viability Counter. Finally, cells from four independent biological replicates were pooled in equal cell numbers into a single cell suspension for each condition. Cell suspensions were processed for single-cell RNA-sequencing using the 10X-Genomics 3' v2 kit (G. X. Y. Zheng et al., 2017), as specified by

the manufacturer's instructions. In brief, 1 x $10^4$ cells from each condition were loaded in separate inlets of a 10X-Genomics Chromium controller in order to create gel bead-in-emulsions (GEMs). The targeted recovery was 3,000 cells per condition. Emulsions were used to perform reverse transcription, cDNA amplification and RNA-sequencing library preparation. Libraries were sequenced on the Illumina HiSeq 4000 platform, using 75 bp paired-end reads and loading one sample per sequencing lane.

### Flow cytometry

Cells were washed using FACS buffer (PBS buffer supplemented with 1% FCS and 1 mM EDTA) by centrifugation and stained with the respective antibodies. Reactions were incubated for 30 minutes at 4°C. Following two washes, samples were resuspended in 200 µl of FACS buffer and data was acquired using a Fortessa analyser (BD Bioscience). All data were processed with FlowJo (v9.9). The antibodies used in this study were: anti-human CD4, APC (BioLegend; Clone: OKT4, Catalog No: 317416, Dilution: 1:100), anti-human CD45RA, Brilliant Violet 785 (BioLegend; Clone: HI100, Catalog No: 304140, Dilution: 1:100), anti-human CD45RO, PE-Cyanine7 (BioLegend; Clone: UCHL1, Catalog No: 304229, Dilution: 1:100), anti-human CD197 (CCR7), (BD Bioscience; Clone: 150503, Catalog No: 561271, Dilution: 1:100), anti-human IFN gamma, PE-Cyanine7 (eBioscience; Clone: 4S.B3, Catalog No: 25-7319-82, Dilution: 1:50), anti-human IL-9, PE (BD bioscience; Clone: MH9A3, Catalog No: 560814, Dilution: 1:50).

### Intracellular cytokine staining

CD4+ T cells were obtained from PBMCs using the EasySep human CD4+ T cell enrichment kit (StemCell Technologies, Meylan, France). Next, CD4+CCR7+CD45RA+ ($T_N$), CD4+CCR7+CD45RA- ($T_{CM}$) and CD4+CCR7-CD45RA- ($T_{EM}$) cells were isolated from CD4+ T cells via FACS using a MoFlo XDP cell sorter (Beckman Coulter) (**Supplementary Figure 3.9**) and polarized to the Th0 and Th17 phenotypes as described above. After five days of stimulation, activated naive and memory T cells were restimulated with 50 ng/ml phorbol 12-myristate 13-acetate (PMA) (Sigma) and 1 µM Ionomycin (Sigma) for five hours in the presence of 10 µg/ml of Brefeldin A (Sigma) at 37°C. After five hours, cells were fixed and permeabilized using the eBioscience™ Foxp3/Transcription Factor Staining Buffer Set (Thermo Fisher Scientific), according to the manufacturer's instructions. Cells were resuspended in 50 µl of permeabilization solution and stained for cytokines, then flow cytometry was performed.

### RNA-seq data analysis

Sequencing reads were aligned to the reference human genome using STAR (Dobin et al., 2013) (v2.5.3) and annotated using the hg38 build of the genome (GRCh38) and Ensembl (v87). Next, the number of reads mapping to each gene was quantified using featureCounts (Liao et al., 2014) (v1.22.2). After quantification, reads mapping to the Y chromosome and the major histocompatibility complex (HLA) region (chr6:25,000,000-47,825,000) were removed from the analysis. The final result from this process was a counts table of RNA expression in each sequenced sample.

RNA counts were imported into R (v3.5.1) where normalization for library size and regularized-logarithmic transformation of counts was performed using DESeq2 (Love et al., 2014) (v1.19.52). We identified and removed batch effects using limma (Ritchie et al., 2015) (v3.35.15). Exploratory data analysis was performed using ggplot2 (v3.0.0) and the base R functions for principal component analysis. Differential expression analysis was performed with DESeq2. More specifically, pairwise combinations were performed between any two conditions of interest, usually setting either resting or Th0-stimulated cells as controls. Differentially expressed genes were defined as any genes with absolute LFC larger than 1 at a FDR of 0.05.

### Proteomics data analysis

After quantification, protein abundances were normalised in order to allow comparisons between samples and plexes (mass spectrometry batches). Namely, protein abundance values were normalized to the total abundance of the respective sample (sample-wise normalization) and then scaled to the maximum abundance of the respective protein (protein-wise scaling). Data were then imported into R, where principal component analysis was performed using all the proteins with no missing values (proteins detected in all batches and samples) with base R functions. Finally, differential protein expression was analyzed by performing pairwise comparisons between any two conditions of interest. This was done using the moderated T test implemented in limma's eBayes function (Ritchie et al., 2015). When testing for differential protein expression, only proteins detected in at least two biological replicates per condition were kept. Multiple testing correction was performed using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). Finally, differentially expressed proteins were defined as any proteins with an absolute log-fold change larger than 0.5 at an FDR of 0.1.

### Pathway enrichment analysis

Pathway enrichment analysis was performed using proteomics and RNA-seq data. To do so, genes detected at both the RNA and protein level were identified by matching gene names. Next, genes were ranked by differential gene or protein expression, respectively, compared to either resting or Th0-stimulated $T_N$ and $T_M$ cells. Finally, pathway enrichment analysis was performed independently in the RNA and protein data using the Perseus software (Tyanova et al., 2016) (v1.6) and the 1D-annotation enrichment method (Cox & Mann, 2012). The enrichment scores indicated whether the RNAs and proteins in a given pathway tended to be systematically up-regulated or down-regulated based on a Wilcoxon-Mann-Whitney test. A term was defined as differentially enriched if it had a Benjamini-Hochberg FDR < 0.05. Firstly, all significantly enriched pathways with absolute enrichment scores higher than 0.25 in both RNA and protein data for the same cell types and cytokine conditions were retrieved. The enrichment scores estimated for these pathways were used to estimate the correlation between RNA and protein using Pearson correlation coefficients (**Supplementary Figure 3.3C-D**). Next, a subset of strongly enriched pathways was selected for visualization in R using the pheatmap package (v1.0.10). This selection included all pathways with an absolute enrichment score higher than 0.7, at an FDR < 0.05 which were included in either Reactome, KEGG or CORUM (Fabregat et al., 2016; Kanehisa & Goto, 2000; Ruepp et al., 2008) and were of relevance to CD4+ T cell biology.

### Cell state specific gene signatures

The correlation between RNA and protein expression was evaluated by estimating LFCs, with respect to the control (Th0) in each cytokine condition and computing the Pearson correlation between RNA and protein LFCs. This was done both sample-wise and gene-wise. Resting T cells were excluded from this analysis. Furthermore, correlation estimates were also computed at the pathway level based on the pathway enrichment analysis results.

Proteomics and RNA-seq data were used jointly to identify gene signatures associated with each cytokine-induced cell state. First, both data sets were matched by gene name to identify a common set of genes detected at both the RNA and protein level. Next, the f-divergence cut-off index (fCI) method (Tang et al., 2016) was used to identify genes (RNA-protein pairs) with significant evidence of differential expression given their RNA counts and protein abundances. For any genes detected as significant by fCI, their normalized regularized-log (rlog) RNA counts (Love et al., 2014) and scaled protein abundances were used to calculate specificity scores in RNA and protein datasets, respectively. To achieve this, replicates from each condition were first averaged. Next, the specificity of each gene in each cytokine-induced

cell state was defined by normalizing the expression of each gene or protein to the Euclidean mean across its different cell states, as described elsewhere (Hu et al., 2011).

RNA specificity was defined as:

$$s_{i,j,RNA} = \frac{X_{i,j}}{\sqrt[2]{\sum_{j=1}^{n} X_{i,j}^2}}$$

Protein specificity was defined as:

$$s_{i,j,prot} = \frac{Y_{i,j}}{\sqrt[2]{\sum_{j=1}^{n} Y_{i,j}^2}}$$

Where $\mathbf{X_{i,j}}$ and $\mathbf{Y_{i,j}}$ are the average RNA expression and protein abundance of gene **i** in cytokine condition **j**, respectively, and **n** is the number of cytokine conditions assessed. As the RNA expression and protein abundance are both non-negative values, $S_{i,j,RNA}$ and $S_{i,j,prot}$ are both ≥0.

Proteo-transcriptomic specificity scores were defined as the weighted sum of RNA and protein specificities for each gene:

$$\boldsymbol{S}_{i,j} = W_{RNA} S_{i,j,RNA} + W_{prot} S_{i,j,prot}$$

Where $\mathbf{S_{i,j}}$ is the specificity score of gene i in condition j. In order to give the same weight to proteomic and transcriptomic evidence, the RNA and protein weights ($\mathbf{W_{RNA}}$ and $\mathbf{W_{prot}}$) were set to 0.5.

To test which genes were more specific to one cell state than expected by chance, sample labels were randomly permuted and the specificity score was recalculated. Empirical P values were computed as the proportion of times the observed specificity score of a gene in a given cell state was larger than the corresponding permuted value. P values were corrected for the number of genes tested using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). A total of 10,000 permutations were performed. Finally, proteo-transcriptomic signatures for each cytokine condition were defined as any genes with a specificity score larger than 0.7 at an FDR of 0.1. This analysis was performed separately for naïve and memory T cells.

### Single-cell RNA-sequencing data analysis

Single-cell RNA-sequencing data were processed using the Cell Ranger Single-Cell Software Suite (G. X. Y. Zheng et al., 2017) (v2.2.0, 10X-Genomics). Namely, reads were assigned to cells and then aligned to the human genome using STAR (Dobin et al., 2013), based on the hg38 build of the human genome (GRCh38). Reads were annotated using Ensembl (v87). Gene expression was then quantified based on any reads assigned to cells and confidently mapped to the genome.

As each of the samples consisted of a pool of four individuals, natural genetic variation was used to identify which cells corresponded to which person. A list of common genetic variants was collected, defined as any SNP included in gnomAD (Lek et al., 2016) with a minor allele frequency higher than 1% in the Non-Finish European (NFE) population. Next, cellSNP (v0.99) (Y. Huang et al., 2019) was used to generate pileups at these SNPs, resulting in one VCF file per sample. This information was then used by Cardelino (v0.99, now Vireo) (Y. Huang et al., 2019) to infer which cells belonged to the same individual. Any cells which remained unassigned (with less than 0.9 posterior probability of belonging to any individual) or were flagged as doublets were discarded. In general, over 85% of cells were unambiguously assigned to an individual (**Supplementary Figure 3.4**). This analysis was performed separately for each sample. To identify which individual from a given sample corresponded to an individual in a different sample, results from Cardelino were hierarchically clustered by genotypic distances between individuals. Clustering separated genotypes into four distinct groups, each group corresponding to one of the profiled individuals.

Results from RNA quantification and genotype deconvolution were imported into R and analysed using Seurat (v2.3.4) (Butler et al., 2018). Cells with less than 500 genes detected or with more than 7.5% mitochondrial genes were removed from the data set. Cells expressing high levels of hemoglobin genes (i.e. *HBA* and *HBB*) were also removed from the data set, as they likely represent contamination during cell culture or sample processing. Counts were next normalized for library size and log-transformed using Seurat's default normalization parameters. Next, a publicly available list of cell cycle genes (Kowalczyk et al., 2015) was used to perform cell cycle scoring and assign cells to their respective phase of the cell cycle. Cell cycle scores, as well as additional known sources of unwanted variation (mitochondrial content, cell size as reflected by UMI content, biological replicate and library preparation batch) were regressed using Seurat's built-in regression model. Highly variable genes were identified using Seurat and used to perform PCA. The first 30 PCs were used as an input for clustering and for embedding using UMAP (L. McInnes et al., 2018). Marker genes for each

cluster were identified computationally using the Wilcoxon rank sum test implemented in Seurat. Multiple testing correction was performed using FDR with cell cycle genes excluded. Moreover, marker genes were required to be expressed by at least 10% of the cells in the cluster at a minimum fold change of 0.25. A total of 5 clusters were identified in resting cells and 17 clusters were found in stimulated cells. Clusters were manually annotated according to their gene expression pattern, the cytokines which cells in the cluster were exposed to and the presence or absence of hallmark genes compiled from the literature.

UniFrac distance analysis (Lozupone & Knight, 2005) was used to test if cells exposed to two different cytokine conditions tended to colocalize in the same clusters. Pairwise UniFrac distances were computed for all combinations of cytokine conditions using all the cells captured for the respective conditions. The R package scUnifrac (v0.9.6) (Q. Liu et al., 2018) was used, as it was specifically adapted to deal with scRNA-seq data. All parameters were set to the default values (1,000 permutations, nDim = 4, ncluster = 10).

### Pseudotime ordering

Cells were ordered into a branched pseudotime trajectory using Monocle (v2.12.0) and restricting the analysis to the highly variable genes identified by Seurat. This was done separately for each cytokine condition (resting, Th0, Th2, Th17 and iTreg), including both $T_N$ and $T_M$. This resulted in five condition-specific pseudotime trajectories. Monocle was used to test for a significant correlation between gene expression and pseudotime in each trajectory. A gene was defined as significantly associated with pseudotime if its estimated q value was lower than 0.01.

### Mathematical definition of T cell effectorness

The four pseudotime trajectories derived from TCR/CD28 or cytokine-stimulated T cells (Th0, Th2, Th17 and iTreg) were combined into a single numeric variable. To do this, the pseudotime values of cells within each condition were scaled to the range [0,1] and all cells were combined into a single data set

### Alignment of resting and Th0-stimulated cells

The single-cell transcriptomes of resting and Th0-stimulated T cells were analyzed separately according to the methods described above. This allowed the identification and annotation of

clusters in resting T cells, as well as the estimation of an effectorness value for every cell in both data sets. Next, canonical correlation analysis (CCA) (Butler et al., 2018), as implemented in Seurat v2.3.4, was used to identify correlated features between the two conditions and to align resting and Th0-stimulated cells into a common space of lower dimensionality. The first 30 CCA dimensions were used to perform UMAP embedding and visualization of cells. The cluster labels defined for resting T cells, as well as the effectorness values independently estimated for resting and Th0-stimulated cells were examined in these visualizations and used to annotate Th0-stimulated cells based on the corresponding resting T cell annotations.

### Modelling the interaction between effectorness and cytokines

The association between gene expression, effectorness and cytokine-stimulation was tested with the *lm()* function from base R. The expression of each gene was modelled as a linear function of T cell effectorness (a numeric variable in the [0,1] range) and cytokine-stimulation (a categorical variable with levels Th0, Th2, Th17 and iTreg). An additional term was incorporated which accounted for potential interactions between these two variables, as specified in the following equation:

$$X_{i,j} = \alpha E_j + \beta C_j + \gamma E_j * C_j + \varepsilon$$

Where **X** is the expression of gene **i** in cell **j** (log2 of normalized UMIs), **E** the effectorness of cell **j**, **C** the cytokine cocktail which cell **j** was exposed to and **ε** a random error term, which was assumed to follow a normal distribution with a mean of zero. The regression coefficients for the intercept, effectorness, cytokine stimulation and the effectorness-cytokine interaction were represented, respectively, by **α**, **β**, **γ** and **δ**. An estimate and a P value were derived for each of these coefficients in each tested gene. P values were corrected for the number of genes tested using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). This analysis was restricted to the top variable genes identified by Seurat. All cells with zero-expression for a given gene were omitted. A coefficient was defined as significant if its corresponding FDR-adjusted P value was lower than 0.05.

### Effectorness and TCR repertoire analysis

Pre-processed scRNA-seq data were obtained from a study profiling CD4+ and CD8+ T cells from colorectal cancer patients (L. Zhang et al., 2018). This data set contained matched full-length transcriptomes and paired TCR-sequences for 10,805 single-cells isolated from 12 patients. First, cells annotated by the authors as CD4+ $T_N$, $T_{CM}$, $T_{EM}$/$T_{EMRA}$ or Treg cells (i.e.

CD4_C01-CCR7, CD4_C02-ANXA1, CD4_C03-GNLY and CD4_C10-FOXP3, according to the cluster labels reported in the study) were retrieved. This resulted in 1,513 single-cells which corresponded to the populations detected in our study. Of these cells, 94% were from peripheral blood of patients, the remaining 6% being either T cells from tumors or from adjacent normal tissues. Cells were next processed using the analysis pipelines described above. Namely, highly variable genes were estimated and used for dimensionality reduction, UMAP embedding and unsupervised ordering of cells in branched pseudotime trajectories. Finally, the effectorness value of each single T cell in the data set was defined as its estimated pseudotime value scaled to the [0,1] range.

TCR clonotype IDs for each single-cell in the study were retrieved. These clonotypes were next used to estimate three diversity metrics: 1) the fraction of unique clones (i.e. the number of unique TCR clonotypes divided by the total number of clonotypes), 2) the Shannon entropy (Rosati et al., 2017) and 3) the expansion index (defined as 1 - Shannon entropy) (L. Zhang et al., 2018) at different ranges of T cell effectorness. To achieve this, cells were ordered by increasing effectorness (i.e. from the least to the most effector) and the diversity metrics were estimated using a sliding window of size 100 (i.e. first analyzing the cells ranked 1 to 100, next the cells ranked 2 to 101, and so on until reaching the end of the data set). The three diversity metrics, along with the mean effectorness value, were repeatedly calculated within each window. Importantly, the fixed window size allowed us to avoid the need for any normalization by cell numbers. Sliding windows were created using the *rollapply()* function included in R's Zoo package (Zeileis & Grothendieck, 2005).

Finally, the amino acid sequence of the TCRβ1 complementarity-determining region 3 (CDR3) region of each cell (Stubbington et al., 2016) was retrieved from the study supplementary data. The GLIPH algorithm (Glanville et al., 2017) was used to cluster TCR sequences by pMHC specificity based on their CDR3 sequences. The TCR network and clonotype groups inferred by GLIPH were imported into R and visualized using the igraph package (Csárdi & Nepusz, 2006), focusing on the clonotypes of cells with an estimated effectorness value > 0.7.

### Analysis of public proteomics data

We downloaded proteomics data from a public data set (Rieckmann et al., 2017) and calculated the mean expression for each protein across donors. Of the 24 effectorness-associated genes with large and ubiquitous effects we observed in our study, 18 were present in Rieckmann et al data set (ACTB, CCL3, CCL5, CTSW, GNLY, GZMA, GZMB, HLA-DPB1, HLA-DQA1, HLA-DRA, HLA-DRB1, HOPX, IFNG, LGMN, LMNA, NFKBIA, TMEM173,

TNFRSF18). The protein abundance of these genes in each cell type was normalized to the mean protein abundance across cell types. Significance of the association between protein abundance and cell type was calculated using one-way ANOVA and group means were compared using Tukey's Honest Significant Difference test.

## Code and data availability

The raw sequencing data used in this chapter is publicly available via the European Genome-Phenome Archive, with the accession numbers listed below. **RNA-seq:** study accession number EGAS00001003823 and sample accession number EGAD00001005291 [https://www.ebi.ac.uk/ega/studies/EGAS00001003823]. **scRNA-seq:** study accession number EGAS00001003215 and sample accession number EGAD00001005290 [https://www.ebi.ac.uk/ega/studies/EGAS00001003215]. Furthermore, the raw mass spectrometry data used in this chapter is publicly available via the Proteomics Identifications Database (PRIDE) under the sample accession number PXD015315 [https://www.ebi.ac.uk/pride/archive/projects/PXD015315]. Finally, processed RNA-seq counts, protein abundances, scRNA-seq UMI counts and cell annotations, are available via the Open Targets website [https://www.opentargets.org/projects/effectorness]. This website also contains interactive applications to visualize bulk RNA/protein expression, and single-cell RNA expression profiles of resting and cytokine-polarized T cells.

All the codes used for processing and analyzing the data presented in this chapter were compiled into a single and publicly available GitHub repository [https://github.com/eddiecg/T-cell-effectorness]. In addition, the functions used for deriving cell state-specific proteo-transcriptomic signatures are available as an R package in GitHub: [https://github.com/eddiecg/proteogenomic].

# Immune disease risk variants modify gene expression dynamics during CD4+ T cell activation

## Overview

Gene expression dysregulation is thought to underlie many human diseases, including immune-mediated traits like autoimmunity and autoinflammation. Genetic susceptibility to immune diseases is caused by thousands of genetic loci with individually small effects, most of which lie in the non-coding genome (Visscher et al., 2017). Variants at these loci are thought to alter the expression of nearby genes via mechanisms like modification of local chromatin architecture (Bossini-Castillo et al., 2019; ENCODE Project Consortium, 2012; Kumasaka et al., 2016) or disruption of transcription factor binding (Musunuru et al., 2010; Pasquali et al., 2014; Sinnott-Armstrong et al., 2021). However, understanding the mechanisms that operate at each of these loci is not easy, and we currently lack a thorough understanding of which genes mediate genetic risk.

One way to tackle this issue is by integrating disease associations from genome-wide association studies (GWAS) with genetic associations for molecular traits such as gene expression (expression quantitative trait loci, eQTLs). Colocalization methods can be used to test if both traits are explained by a shared causal variant, thereby identifying candidate disease genes (Fortune et al., 2015; H. Huang et al., 2017). This approach is particularly useful if gene expression is measured in disease-relevant cell types and cell states, which are often unknown. In the case of immune diseases, it has been shown that GWAS loci are enriched in regulatory elements tagged by chromatin marks in CD4+ T cells (Hu et al., 2011; Trynka et al., 2013, 2015). This suggests that colocalization studies for immune traits would benefit from studying CD4+ T cell gene expression and how it is affected by genetic variation. However, most eQTL studies to date have only profiled CD4+ T cells in bulk, which limits their results to an average measurement across cell subpopulations (Chen et al., 2016; Schmiedel et al., 2018; Ye et al., 2014). This contrasts starkly with the complexity and plasticity of CD4+ T cells, known to be a heterogeneous cell type formed of distinct subpopulations and cell states (S. N. Mueller et al., 2013; Federica Sallusto, 2016; Trzupek et al., 2020). As such, profiling T cells using single-cell technologies would enable us to better characterize the effects of genetic variation in T cell biology.

In addition, CD4+ T cells perform their function *in vivo* by responding to antigens presented by antigen-presenting cells (APCs) in secondary lymphoid organs (Bousso, 2008; Luckheeram et al., 2012; Smith-Garvin et al., 2009). However, most studies to date have only

profiled CD4+ T cells isolated from peripheral blood, where any gene expression changes arising during T cell activation would be missing (Chen et al., 2016; Schmiedel et al., 2018). This is particularly important in light of the observations described in chapter 2, which showed that the enrichment of immune disease GWAS loci is particularly strong in regulatory regions active during early activation of memory T cells. This suggests that currently available datasets could be missing many of the effects relevant for disease, and that we are in need of more detailed eQTL maps encompassing CD4+ T cell activation at single-cell resolution.

In this chapter, I describe the results of a study carried out by myself and others (as specified in the collaboration note) to map eQTLs throughout CD4+ T cell activation at unprecedented resolution using scRNA-seq. To our knowledge, we present the largest single-cell data set of activated T cells to date, with over 655,000 high quality transcriptomes spanning four time points and comprising 38 cell populations. We leverage this data to reconstruct detailed activation trajectories for naive and memory CD4+ T cells, which enables us to map eQTLs at different points in time and across subpopulations of cells. We identify 6,407 genes with evidence of genetic regulation, of which 2,265 are regulated by eQTLs that change as a function of time during T cell activation. By testing for colocalization between these eQTLs and genetic associations for immune-mediated diseases, we prioritise 139 candidate disease causal genes. Importantly, these genes are enriched for time-dependent eQTLs, suggesting that disruption of gene expression dynamics during T cell activation could be a key mechanism underlying immune disease. Our observations provide new insights into the molecular mechanisms underlying immune-mediated disease, and could be especially useful for identification of novel drug targets.

## Results

### Study design

To investigate the gene expression changes induced by CD4+ T cell activation, we isolated naive and memory CD4+ T cells from 119 healthy individuals of British ancestry (**Supplementary Figure 4.1** and **Supplementary Table 4.1**) and stimulated them using anti-CD3/anti-CD28 coated beads (**Figure 4.1A** and **Methods**). We then profiled single-cell gene expression using droplet-based scRNA-seq (G. X. Y. Zheng et al., 2017) at 16 hours, 40 hours, and 5 days of stimulation, as well as in cells kept in culture without any stimulus (resting cells, 0 hours of activation). In order to increase throughput and minimize technical variation between experimental batches, we mixed together between four and six biological replicates per droplet capture reaction (**Supplementary Figure 4.2, Supplementary Table 4.1** and **Methods**). This resulted in 655,349 CD4+ T cell transcriptomes spanning four activation time

points and two cell types, which to our knowledge constitutes the largest single-cell map of T cells to date. In addition, we performed genotyping on all the individuals in the study, which enabled us to link the presence of common genetic variants at disease-associated loci with changes in the expression dynamics of neighbouring genes.

## Characterising the response to T cell activation at single-cell resolution

CD4+ T cell activation is known to induce large transcriptional changes that unfold gradually over time (Ye et al., 2014). Thus, we performed a high level analysis to identify major gene expression and cellular composition changes induced by activation. In brief, we first removed low quality genes and cells, as well as cellular contaminations (cells expressing markers of any cell type other than CD4+ T cells; **Supplementary Figure 4.3** and **Methods**). Over 90% of cells in the data set passed these quality filters, demonstrating that our data are high quality. We next performed dimensionality reduction and embedding using the uniform manifold approximation (UMAP) (L. McInnes et al., 2018) (**Methods**). We observed that cells separated by time point of stimulation, forming a gradual progression from the resting state towards the early and late time points of activation (**Figure 4.1B**). This progression was accompanied by changes in known T cell activation markers. For example, *CD69*, a marker of early activation, was upregulated at 16 hours but downregulated at later time points. In contrast, upregulation of *IL2RA*, a marker of late activation, was not observed until 40 hours, remaining highly expressed at 5 days (**Figure 4.1C**).

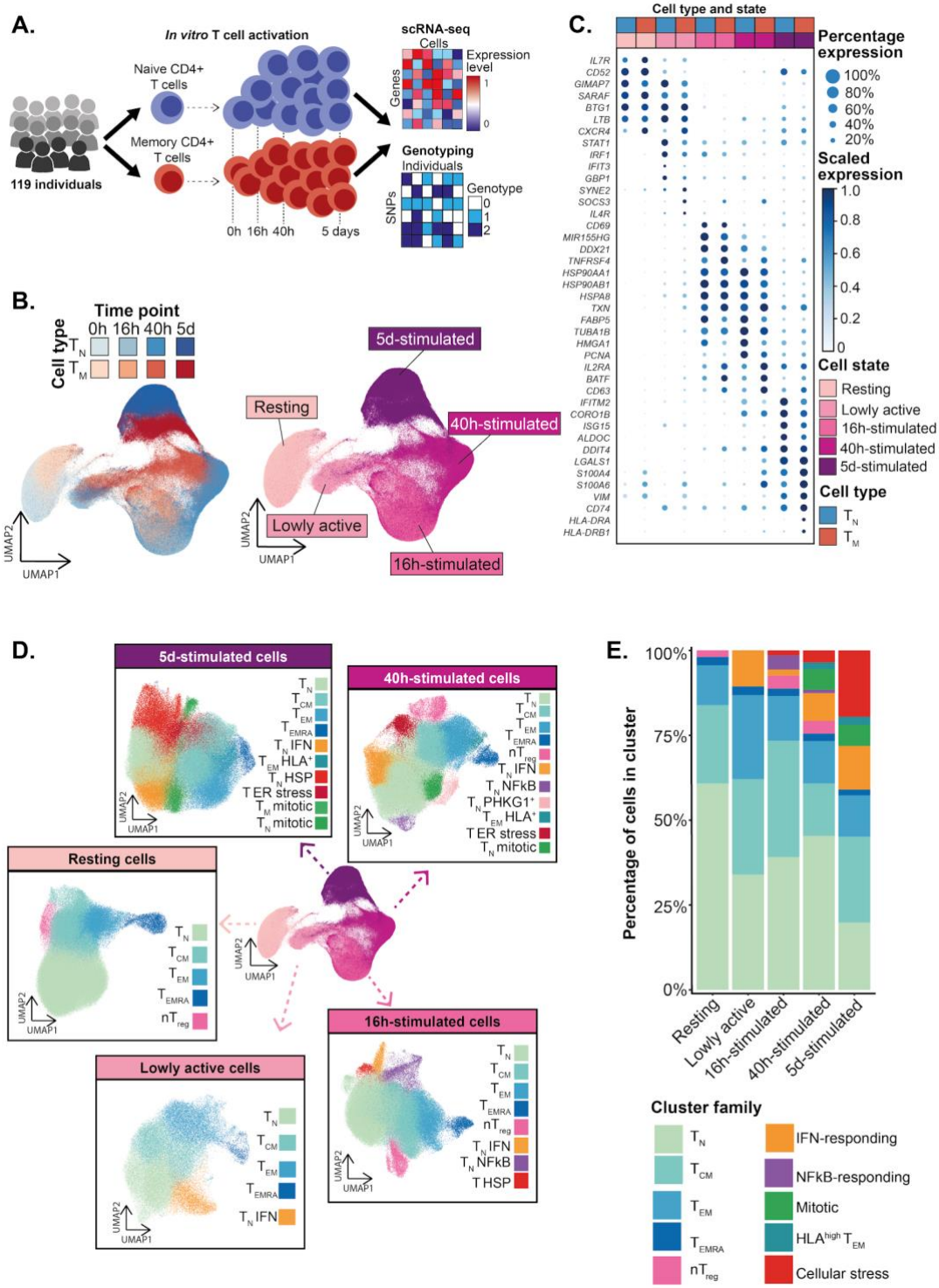



**Figure 4.1. A single-cell transcriptional map of CD4+ T cell activation. A)** Schematic of study design. **B)** UMAP embedding of scRNA-seq data for CD4+ T cells at four time points after activation. Colours represent either cell types ($T_N$ and $T_M$), with shades of colour indicating time points (left panel), or five broad cell states identified upon exploratory data analysis (right panel) **C)** Dotplot of genes which vary throughout T cell activation. Shades of blue represent average expression in each cell population. Dot sizes represent the proportion of cells expressing each gene. **D)** Separate UMAP embeddings for the four broad cell states. Colours represent cell populations derived from unsupervised clustering. **E)** Proportion of different cluster groups present at each time point. Cell populations defined from clustering were classified into one of 10 families, represented in different colours.

Moreover, scRNA-seq also enabled us to capture cell states which were not directly sampled. We identified a large population of cells that localized between the resting and 16 hour-stimulated cell groups in UMAP space (**Figure 4.1B**). The majority of cells in this group had been collected at 16 hours of activation, with a smaller proportion coming from the 40 hour time point. We hypothesised that this group represented an early cell state through which cells transitioned when they activated. To test this, we analysed cells from the 16 and 40 hour time points independently. We confirmed that, at each of these time points, cells separated into two clear groups, one of which corresponded to this early transitional state (**Supplementary Figure 4.4A-C**). Cells in this group expressed 4-fold less genes than the remaining cells at their respective time point and showed lower expression of the CD4+ T cell activation markers defined in chapter 3 (**Supplementary Table 3.2** and **Supplementary Figure 4.4A-B**). Furthermore, they had a unique profile which differed from that of resting or fully activated cells, characterised by high expression of *STAT1*, *IFIT3*, and *GBP1* (**Figure 4.1C**). Based on these observations, we concluded that these cells represent a distinct, early activation state and annotated them as lowly active cells. We treat lowly active cells as a separate group for most analyses in this study. Importantly, since a small proportion of the cells collected at 16 and 40 hours were lowly active, this cell state would have been missed using bulk RNA-seq.

These observations gave us a high-level understanding of T cell activation dynamics. However, as shown in the previous chapter, CD4+ T cells are formed of multiple subpopulations with non-overlapping effector functions, each of which could potentially be mounting a different response. Thus, we used unsupervised clustering to map the subpopulations present at each of the broad cell states described above (resting, lowly active, 16 hours, 40 hours, and 5 days). This revealed a total of 51 cell clusters, which were merged based on their correlated patterns of expression into 38 well-defined cell populations (**Supplementary Figure 4.5** and **Methods**). We annotated these populations using their marker genes (**Supplementary Table 4.2**) and their overlap with descriptions in the literature (**Supplementary Figure 4.6**).

Of the 38 populations detected, 25 (66%) represented stable T cell subpopulations. These belonged to one of five phenotypes that were consistently detected at all time points in approximately equal proportions: naive ($T_N$), central memory ($T_{CM}$), effector memory ($T_{EM}$), effector memory re-expressing CD45RA ($T_{EMRA}$), and regulatory ($nT_{reg}$) CD4+ T cells (**Figure 4.1D-E**). Importantly, these subpopulations were interconnected, forming a naive-to-memory progression. This agrees with the description of a T cell effectorness gradient presented in chapter 3. In contrast, the remaining 13 (34%) populations were only observed upon activation, often at specific time points (**Figure 4.1D-E**). This latter group represents transient

cell states. For example, we observed a population of cells expressing high levels of type 1 interferon (IFN)-induced genes (eg. *IFI6*, *IFIT3*, *ISG15, MX1*) which appeared during early activation (**Supplementary Figure 4.6**). In contrast, other populations peaked at mid-stages of activation but disappeared altogether at five days, such as a group of cells expressing targets of the transcription factor NFκB (eg. *NFKBID*, *REL*, *BCL2A1*; **Supplementary Figure 4.6**). Finally, two cell populations were only observed during late activation: cells undergoing mitosis, and cells expressing high levels of cellular stress markers (e.g. *HSPA1A*, *HSPA1B*, *DNAJB1*; **Supplementary Figure 4.6**). In addition, we noticed a subset of $T_{EM}$ cells which dramatically upregulated MHC class II molecules (eg. *HLA-DRA*, *HLA-DPA1*, *HLA-DRB1*) during late activation (**Supplementary Figure 4.6**), which is in agreement with previously described changes in expression dynamics at the HLA locus (Gutierrez-Arcelus et al., 2020). Taken together, we characterised the transcriptional response of CD4+ T cells to activation in an unbiased manner, identifying changes in both stable subpopulations and transient cell states.

## Identification of subpopulation and time point-specific gene modules

We next analysed the different patterns of gene expression present during CD4+ T cell activation. To do so, we leveraged the sample size of our study to compute gene-wise correlations across all individuals in the cohort and across the 38 identified cell populations. We focused on a subset of 11,130 genes which were highly expressed and variable (**Methods**). Finally, we used this information to build a gene co-expression network that captures T cell activation changes at subpopulation resolution (Langfelder & Horvath, 2008). Our analysis revealed 12 different gene modules, representing the main patterns of gene expression present in the dataset (**Figure 4.2A** and **Supplementary Table 4.3**). For example, a group of genes involved in the regulation of cell cycle checkpoints and DNA repair mechanisms (**Methods** and **Figure 4.2B-C**) was highly expressed at 40 hours and five days after activation. This agrees with the previously published reports that the first cell division occurs two days after activation (Hawkins et al., 2007), which roughly corresponds to our 40h activation time point. Further, we found a module of genes whose expression peaked in the lowly active and 16h-stimulated cell groups. These genes were involved in IFN-induced antiviral mechanisms such as OAS and ISG15-signalling (**Figure 4.2B-C**). These pathways are known to be induced rapidly upon viral infection and have been reported to follow a similar behaviour in $CD4^+$ T cells profiled with bulk RNA-seq (Ye et al., 2014). Finally, we identified a module of genes involved in NOTCH4, RUNX2 and RUNX3-signaling which had the highest measured expression at 16 hours, after which they were downregulated, ultimately returning

to baseline levels (**Figure 4.2B-C**). This pattern of transient upregulation was also observed for a module of genes involved in RNA metabolism, which were highly expressed after 16 hours of activation.

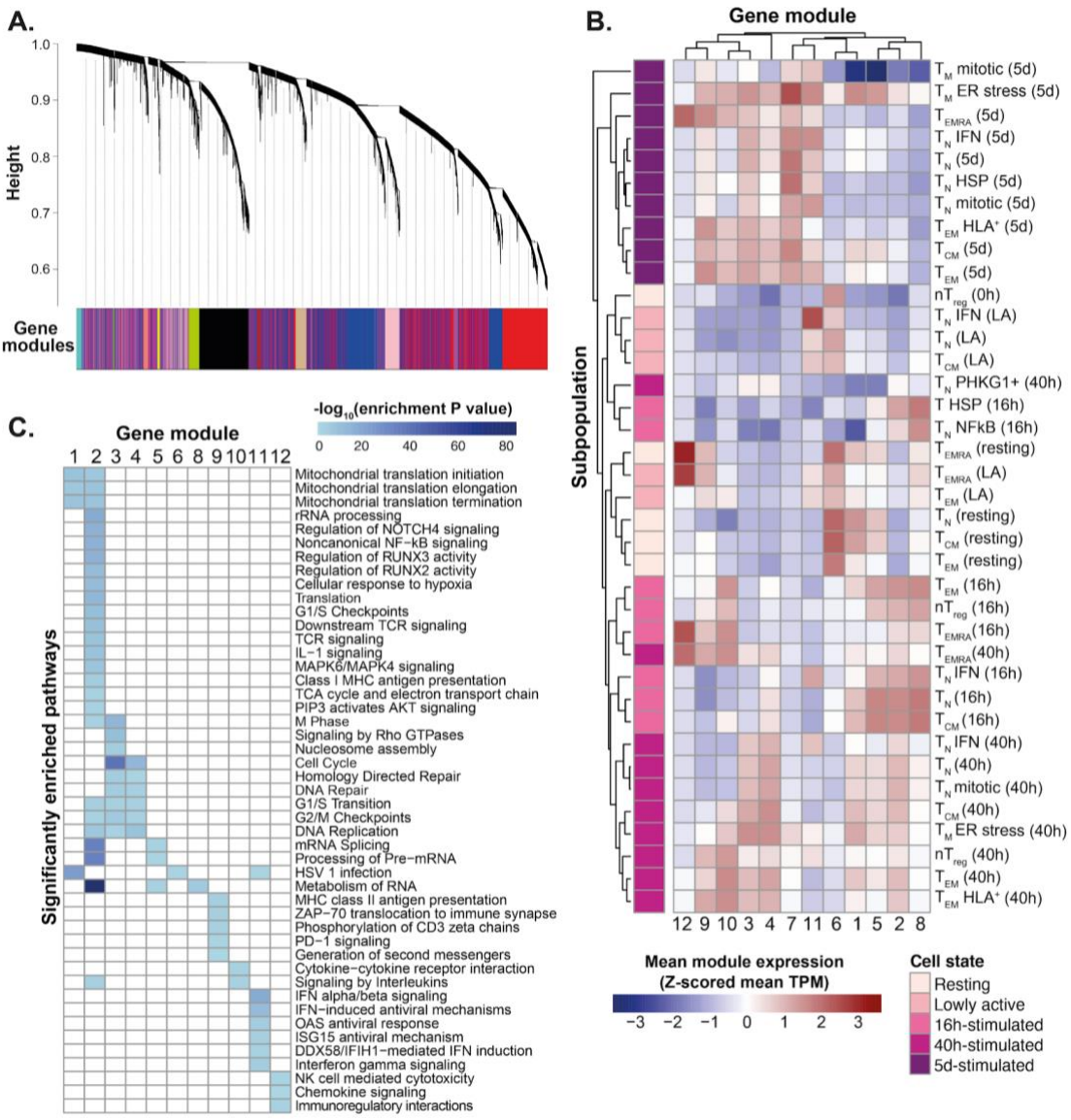



**Figure 4.2 Identification of gene expression patterns active throughout CD4+ T cell activation.** **A)** Genes were ordered into a co-expression matrix using WGCNA (**Methods**). This dendrogram shows genes arranged by similarity. Modules of genes with correlated expression are indicated by blocks of colour at the bottom. **B)** Heatmap showing the expression pattern of 12 identified gene modules. Rows correspond to cell subpopulations. Colours represent the average expression of all genes belonging to a module in a given subpopulation. **C)** Pathways enriched in each gene module. Shades of blue represent the log-transformed estimated enrichment P value.

In addition to separating genes by their temporal dynamics, our co-expression network approach also distinguished between ubiquitous and subpopulation-specific gene modules. In particular, genes involved in cytokine secretion and interleukin signaling were highly expressed in $T_{EM}$ and $T_{EMRA}$, but not $T_{CM}$ or $T_N$ cells (**Figure 4.2B-C**). This agrees with $T_{EM}$ and

$T_{EMRA}$ cells having a higher potential to initiate a fast and robust effector response (F. Sallusto et al., 1999). Interestingly, these cells also upregulated genes directly induced by TCR signaling (i.e. targets of ZAP-70 and downstream of CD3 zeta chain phosphorylation) at an earlier stage than other subpopulations. While $T_{EM}$ and $T_{EMRA}$ already showed high expression of this TCR-induced module in the lowly active and 16 hour states, other subpopulations only started expressing these genes at 40 hours (**Figure 4.2B-C**). This reinforces the notion that $T_{EM}$ and $T_{EMRA}$ cells can initiate a strong response to stimulation faster than $T_{CM}$ or $T_N$ cells. Furthermore, $T_{EMRA}$ also showed highly specific expression of a set of genes involved in cytotoxicity and chemokine signalling (**Figure 4.2B-C**), which sets them apart from any other T cell subpopulation.

In conclusion, we built a co-expression gene network which captures different expression patterns during CD4+ T cell activation and used it to identify 12 modules of genes active in different subpopulations and at different time points.

## eQTL mapping in resting and activated CD4+ T cells

We performed cis-eQTL mapping to study the genetic regulation of gene expression in resting and activated naive and memory T cells. We first did so at the level of naive and memory CD4+ T cells as a whole per time point, which would roughly correspond to the results from a bulk RNA-seq analysis of T cell activation. In total, we detected between 1,545 and 3,006 genes with a significant eQTL (eGenes) in resting and activated naive and memory T cells (**Figure 4.3A**).

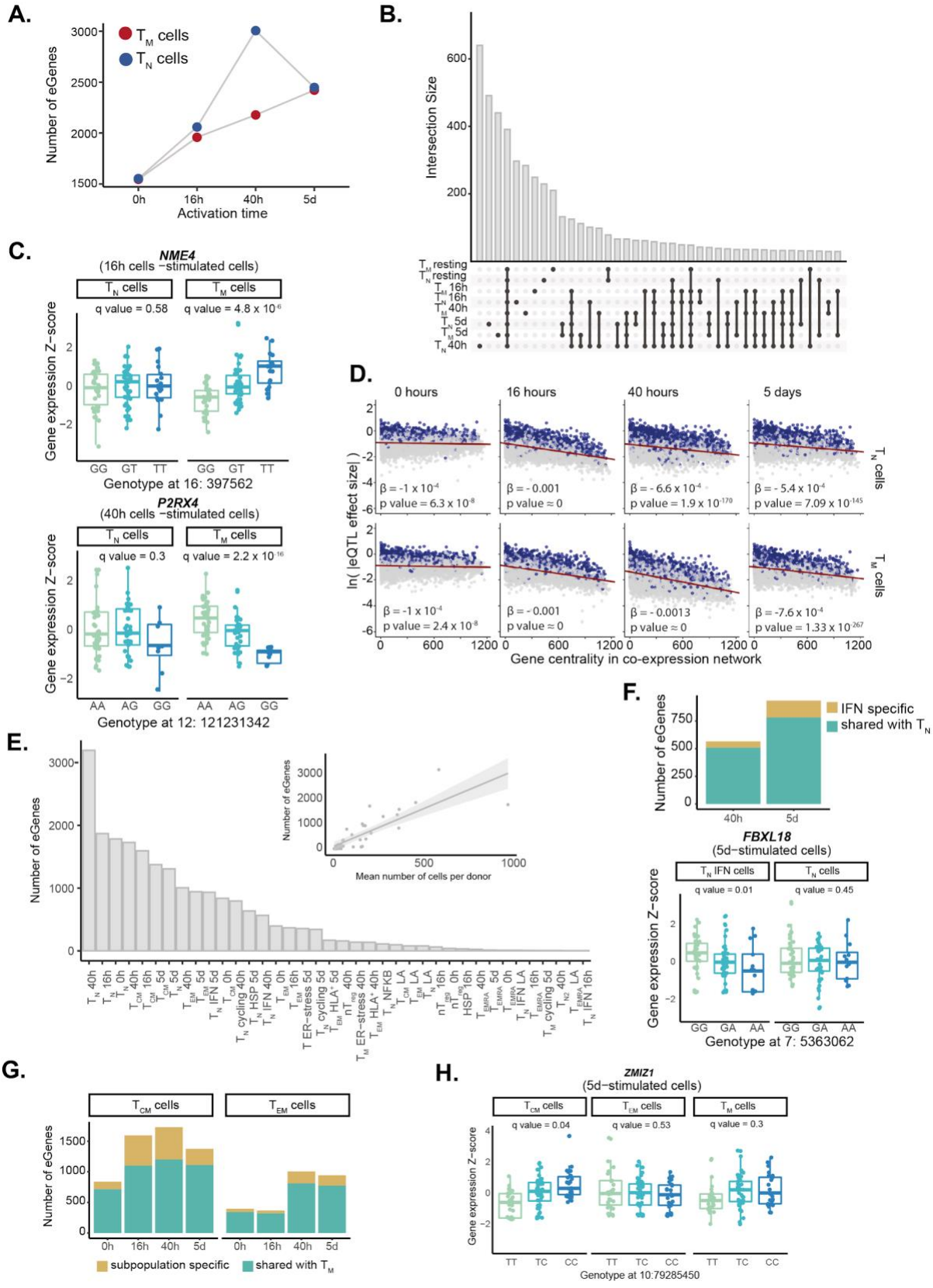



**Figure 4.3 eQTLs mapping at four time points following naive and memory CD4+ T cell activation. A)** Number of significant eGenes detected at each activation time point. Colors represent cell types ($T_N$ and $T_M$). **B)** Number of eGenes shared between cell types and time points. Bar plots represent the number of eGenes present in each set. **C)** Example of two memory cell specific eQTLs

detected at 16h (top panel) and 40h (bottom panel), respectively. Box plots show the mean expression value of the gene in each sample (Z-scored), stratified by genotype. Each dot represents a measurement obtained from a separate individual. **D)** Relationship between a gene's connectivity (as inferred from co-expression network analysis) and the effect size of its lead eQTL signal. All eQTL effect sizes were log-transformed. Blue dots represent significant eGenes, while gray dots represent genes which do not pass the eQTL multiple test correction. **E)** Relationship between eQTL detection and cluster size. The bar plot (bottom panel) shows the number of eGenes detected in each T cell subpopulation, with bars sorted in decreasing order. The scatter plot (top panel) shows the correlation between number of cells and number of detected eGenes in each cluster. **F)** Subpopulation specific eQTLs detected in naive T cells. The bar plot (top panel) indicates how many of the eGenes detected in the TN IFN subpopulation are shared with naive T cells as a whole. The boxplots (bottom panel) show an example eQTL specific to the TN IFN subpopulation. Each dot represents a measurement obtained from a separate individual. **G-H)** Subpopulation specific eQTLs detected in memory T cells. The bar plot (**G**) indicates how many of the eGenes detected in the TCM and TEM subpopulations are shared with memory T cells as a whole. The boxplots (**H**) show an example eQTL specific to the TNM subpopulation. Each dot represents a measurement obtained from a separate individual.

Interestingly, the number of detected eGenes increased with cell activation (**Figure 4.3A)**, likely because activation induces the expression of thousands of new genes per cell (**Supplementary Figure 4.2C-D**). Furthermore, when we compared the eGenes identified in different activation states, we observed that many were only detected in one cell state (**Figure 4.3B**). For example, although kinase *NME4* and purinoceptor *P2RX4* are expressed in both naive and memory T cells, they are only detected as eGenes in memory T cells at 16h and 40h of activation, respectively (**Figure 4.3C**). These examples emphasize the need to study the right cellular context when following up genetic effects on gene expression. However, such simple pairwise comparisons may be influenced by the power to detect eQTLs in each activation state. Therefore, to assess the sharing between eQTLs in a more robust manner, we performed a meta-analysis across cell types and cell states using the multivariate adaptive shrinkage (mashR) method (Urbut et al., 2019). MashR estimates the pairwise level of sharing between cell states, where an eQTL is defined as shared if it has the same effect size (within a factor of 0.5) and direction in two cell states. As expected, we observed a high level of sharing of eQTLs between naive and memory T cells within each of the time points (**Supplementary Figure 4.7**). In contrast, sharing of eQTLs across time points was significantly lower, suggesting that eQTL effect sizes change throughout T cell activation.

Given that we detected more eGenes in activated than in resting T cell states, we next asked how these eGenes related to the activation process itself. To do so, we integrated our eQTLs with the co-expression network defined above (**Figure 4.2**). Surprisingly, the effect size of an eQTL negatively correlated with the gene's centrality in the coexpression network (**Figure 4.3D**). This was particularly evident in activated T cell states, suggesting that genes which are more central in the T cell activation gene network tolerate less variation in gene expression.

Having mapped eQTLs at the level of whole cell types, we leveraged the granularity of our single cell data to map eQTLs in each of the 38 identified subpopulations. In total, we detected between 3 and 3,197 eGenes in each T cell subpopulation (**Figure 4.3E**). We noticed that in certain subpopulations such as $T_{EMRA}$, we consistently detected low numbers of eGenes (3-23). To understand why we observed such high variability in eQTL discovery, we assessed the relationship between the number of eGenes detected and the average number of cells per donor in a given cluster. This revealed that the statistical power to detect eQTLs at single-cell level is highly correlated with the number of cells profiled per individual for the subpopulation of interest ($R^2$ = 0.82, p = $4.8x10^{-10}$ ) (**Figure 4.3E**). We confirmed this by subsampling increasingly lower numbers of cells from a population before eQTL mapping. This confirmed that eGenes discovery linearly increases with cluster size (**Supplementary Figure 4.7**). These simulations suggest that gene expression and variance estimation from small numbers of cells is too noisy for reliable eQTL mapping. This observation will be crucial when designing future single-cell eQTL studies, since it shows that one needs to carefully estimate the size of the smallest cluster of interest and assess how many cells need to be profiled per donor. Due to this limitation, we could not directly compare the eQTLs identified in subpopulations of widely different size. However, for large subpopulations, where eQTL mapping is more reliable, we could still assess which eQTLs would have been missed by studying T cells in bulk. In particular, we identified 56-153 (10-16.4%) eGenes which were present in a subpopulation of naive T cells with high levels of IFN-induced genes, but absent in the whole activated naive T cell population (**Figure 4.3F**) (e.g. *Fbxl18)*. Similarly we identified 47-528 (12.8-31%) eGenes which were detected in either of the two largest memory cell subpopulations ($T_{CM}$ and $T_{EM}$) but not in bulk memory T cells (**Figure 4.3G**). One example of such genes is the STAT-like coactivator *Zmiz1* which is only detected as an eGene in $T_{CM}$ but not in $T_{EM}$ or total memory cells (**Figure 4.3H**).

Taken together, we showed that genetic regulation of gene expression varies during different stages of T cell activation, and that profiling eQTLs during this immunologically critical process can shed light on novel mechanisms of gene regulation. Furthermore, although we are limited to assess eQTLs specific to small cell populations, we demonstrated that large T cell subpopulations don't always have the same genetic control of gene expression as the whole T cell population. This emphasizes the power of single cell transcriptomics to identify eQTLs that would likely be missed in studies of T cells in bulk.

## Modelling of time-dependent eQTL effects throughout T cell activation

The vast majority of studies have so far only profiled a single time point for eQTL analysis. However, given that T cells undergo profound transcriptional changes during activation, it is important to assess the role of genetic variation on gene expression across different stages of T cell activation. Therefore, to generate a high-resolution T cell activation map, we performed trajectory inference analysis (Cao et al., 2019) (**Methods**). The trajectory obtained largely agrees with the order of time points as profiled experimentally, with the lowest pseudotime values being assigned to resting cells and the highest to cells activated for five days (**Figure 4.4A**). Furthermore, we observed that cells localizing to regions containing memory T cells had lower pseudotime values than naive cells at the same time points. This suggests that memory T cells have a shorter activation path, which is broadly consistent with the well accepted notion that activation of memory T cells is faster. Importantly, the temporal dynamics of well defined T cell activation markers such as *IL7R*, *IL2RA* and *CD69* (**Figure 4.4B**) followed their expected patterns, suggesting that the inferred trajectory agrees with previous immunological studies. In addition, we also used an unbiased method to identify genes whose expression changes as a function of pseudotime (**Supplementary Table 4.4**). This revealed instances like *IRF1* and *TOP2A*, which are active at different stages of activation (**Figure 4.4B**).

Since our trajectory contained in the order of hundreds of thousands of resting and activated T cells, we devised an approach to identify changes in eQTL strength across the T cell activation trajectory that retains high resolution while also being computationally tractable. In short, we divided the trajectory into 10 windows containing roughly equal numbers of cells (i.e. pseudotime deciles) and averaged the expression of each gene in each window (**Methods**). We chose to split the trajectory in this manner because, as shown above, the number of eGenes detected correlates with cluster size (**Figure 4.4E**). Using pseudotime deciles enabled us to retain enough cells per donor to reliably estimate gene expression means. We then applied linear mixed models to test for a significant interaction between genotypes and pseudotime using the average pseudotime value per window (**Figure 4.4C and Methods**). In accordance with previous literature, we refer to these eQTLs as *dynamic eQTLs* (Cuomo et al., 2020; Strober et al., 2019).

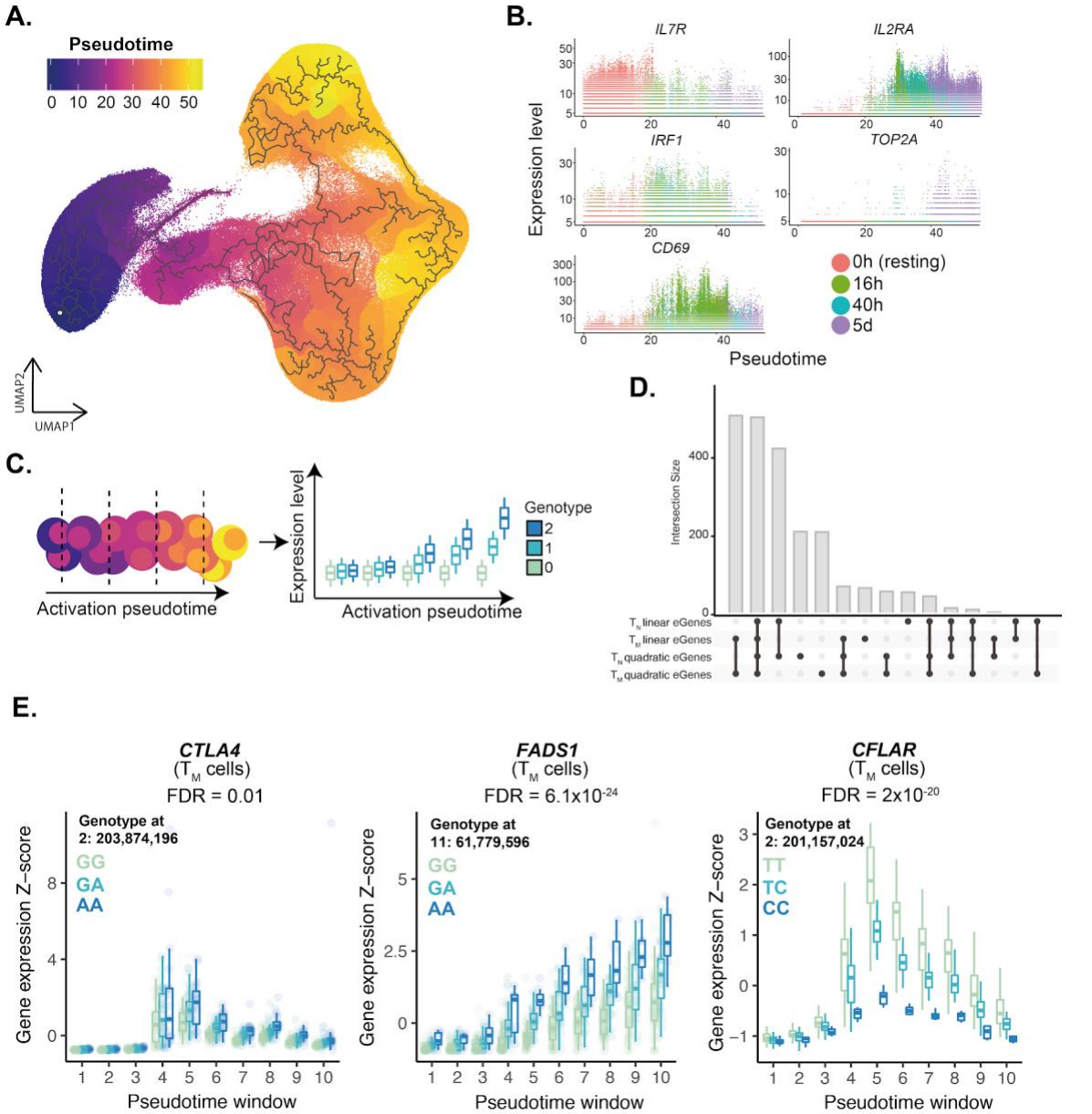



**Figure 4.4 Identification of eQTLs with dynamic effects during CD4+ T cell activation. A)** Cells were ordered into a branched pseudotime trajectory using monocle3. The UMAP embedding shows all cells in the study alongside the inferred trajectory. Cells are coloured by their estimated pseudotime values. **B)** Example of genes that significantly change as a function of activation pseudotime. Each dot corresponds to a cell, and colours represent experimental time points. **C)** Schematic of our analysis approach. Cells were split into ten windows of equal size according to their estimated pseudotime values. Next, linear and quadratic mixed models were applied to each previously identified eGene so as to test for an interaction between genotypes and T cell activation pseudotime. **D)** Number of eGenes with evidence of a significant genotype-pseudotime interaction (i.e. dynamic eQTLs) in a linear or quadratic mixed model. **E)** Examples of one linear (*FADS1*) and two quadratic (*CTLA4* and *CFLAR*) dynamic eQTLs detected in memory T cells. Boxplots show the average expression of the gene within each pseudotime window, stratified by genotype. Each dot corresponds to expression measurements obtained from a different individual.

Our analysis revealed that 2,265 eGenes were regulated by loci whose effect sizes change as cells progress through activation (**Supplementary Table 4.5**). By leveraging both linear and quadratic models, we were able to show that most eQTLs follow linear dynamics across the activation trajectory. However, some showed a non-linear interaction with T cell activation, where the variant had a stronger effect during mid stages of activation (**Figure 4.4D**). For example, we identified dynamic eQTLs for *CTLA4* and *CFLAR* whose magnitude peaked at mid stages of the T cell activation trajectory, significantly diminishing thereafter. In contrast, an eQTL for the metabolic gene *FADS1* had an effect size which linearly increased along the trajectory. The effect of this variant was more pronounced in cells profiled at later stages of activation.

In conclusion, our experimental design allowed us to identify 2,265 dynamic eQTLs, some of which are transient and would have been missed in studies profiling only one activation time point. This illustrates the power gained by profiling cells at single-cell resolution and demonstrates the existence of intricate gene-by-environment interactions, where a genetic regulation mechanism is only activated in response to specific environmental stimuli.

## Colocalization at GWAS loci identifies candidate immune disease genes

The eQTL maps described above provide a unique opportunity to study GWAS loci. In particular, they enable us to assess whether disease-associated variants alter gene expression in $CD4^+$ T cells, and to identify the genes involved. We did this systematically by testing for colocalization between eQTLs and GWAS signals at loci associated with 14 immune-mediated diseases from the GWAS catalog (Buniello et al., 2019) (at a suggestive GWAS threshold of $p < 1x10^{-5}$). The resolution of our study enabled us to do this in naive and memory $CD4^+$ T cells as a whole, as well as in T cell subpopulations and cell states. To do this, we used the Bayesian method *coloc* (Giambartolomei et al., 2018) and leveraged its masking functionality (Wallace, 2020), which removes the assumption of a single causal variant per locus (**Methods**). We identified 615 unique colocalization events (PP4 > 0.8), corresponding to 287 GWAS loci and 334 eQTLs (**Supplementary Table 4.6**). Importantly, we observed more colocalizations for traits with a better powered GWAS and for larger cell populations, where we identified more eQTLs (**Figure 4.5A**). This shows that the power to detect colocalizations increases as a function of the statistical power of the GWAS and the eQTL datasets used, as previously demonstrated (Hukku et al., 2021). Taken together, this enabled us to prioritize 139 candidate disease-causal genes (**Figure 4.5A**). Importantly, only 56 of these genes (40%) showed a colocalization in resting T cells, suggesting that most causal variants might alter gene expression specifically during T cell activation.

We next asked how disease-associated loci modulated the expression of colocalizing genes, and whether any colocalizing genes were shared across traits. We observed limited sharing of genes between diseases, with most of them colocalizing with one trait only (**Supplementary Figure 4.8A**). IBD and its subphenotypes (i.e. UC and CD) were an exception, with 20 genes shared between them. When analysing the mechanisms operating at these loci, we found that 367 disease loci (91%) regulated a single gene, while 30 (7.3%) and 7 (1.7%) regulated two and three genes in the region, respectively. Finally, 57% of the colocalizations for each trait involved a variant that increases disease risk via downregulation of gene expression (**Figure 4.5B**), representing instances where disease risk might be increased by a partial reduction in gene function. Genes in this category include *PTPRC*, *UBE2L3*, and *ZMIZ1*, which colocalize with associations for CD (**Figure 4.5C**). In contrast, 43% of colocalizations involved a variant that increases disease risk via gene expression upregulation (**Figure 4.5B**). Examples of these genes included *PTGER4*, *TYK2*, *FADS1*, and *FADS2*, which also colocalize with CD associations (**Figure 4.5C**). Taken together, our observations suggest that hundreds of immune disease associations might involve changes in gene expression regulation in activated CD4+ T cells, most often via modulation of a single nearby gene.

Given that most of our candidate disease genes only showed a significant colocalization in activated T cells, we asked whether the eQTLs involved in these colocalizations tended to show temporal dynamics. We observed that dynamic eQTLs were significantly enriched in colocalizations in both naive and memory T cells (Fisher's test p values $7.85 \times 10^{-5}$ and $2.6 \times 10^{-7}$, respectively), with 47% of colocalizations involving a dynamic eQTL. Two examples are the dynamic eQTLs for *FADS1* and *CTLA4* detected in memory T cells (**Figure 4.4E**), which colocalize with CD and T1D associations, respectively (**Figure 4.5D**). The colocalization for *CTLA4* is particularly striking, given that the same eQTL variant colocalizes with another two suggestive associations ($1 \times 10^{-8} < p < 1 \times 10^{-5}$) for RA and CeD. This suggests that the gene could play a widespread role in several traits. Individuals carrying the risk allele at this locus showed lower expression of *CTLA4* during the early stages of T cell activation, with this difference diminishing at late activation (**Figure 4.4E**). This agrees with the known role of *CTLA4* in regulating T cell activation shortly after TCR engagement (Qureshi et al., 2011).

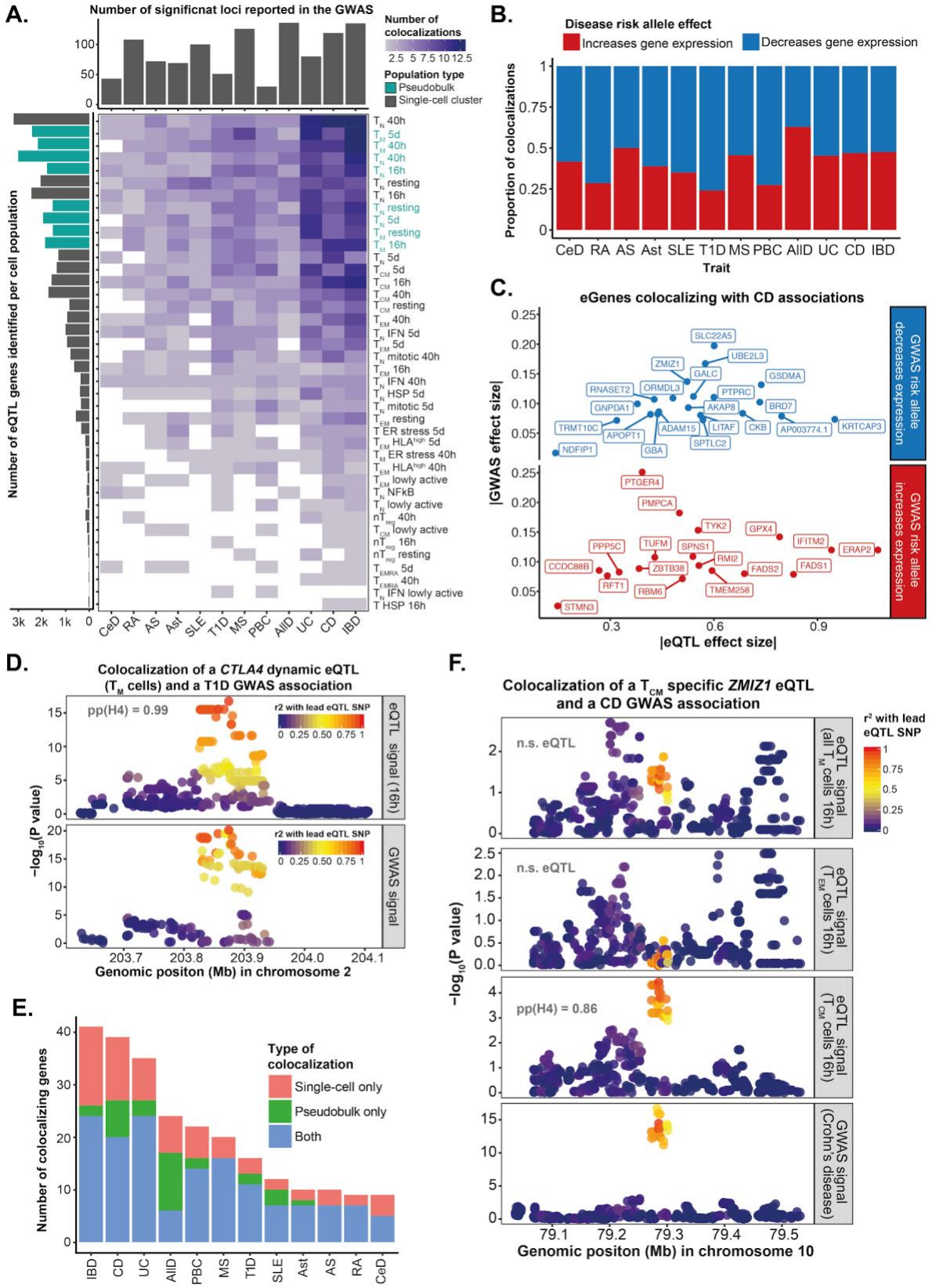



**Figure 4.5 Colocalization of CD4+ T cell eQTLs with GWAS associations for immune diseases.** **A)** Number of significant colocalizations between an eQTL and a GWAS signal identified for each cell type-trait combination. Marginal bar plots represent the number of independent associations reported in the GWAS (X axis) and the number of eGenes detected per subpopulation (Y axis). Light and dark bars indicate whole (pseudobulk) cell populations and subpopulations, respectively. **B)** Proportion of

colocalizing GWAS loci per trait which increase and decrease the expression of a nearby gene. **C)** GWAS and eQTL effect sizes for all colocalizations involving a Crohn's disease association, stratified by effect direction. **D)** Locus plot for a colocalization between a *CTLA4* dynamic eQTL and a GWAS association for type 1 diabetes. Each dot represents a variant, with colors indicating their lead with the lead eQTL variant. **E)** Number of colocalizing genes observed in whole cell populations (pseudobulk), subpopulations or both. **F)** Locus plot for a colocalization between a *ZMIZ1* eQTL specific for TCM cells and a GWAS association for Crohn's disease. Each dot represents a variant, with colors indicating their LD with the lead eQTL variant. The top three panels represent eQTL signals from whole memory cells, the $T_{EM}$ subpopulation, and the $T_{CM}$ subpopulation, respectively. Trait abbreviations: Celiac disease (CeD), Rheumatoid arthritis (RA), Ankylosing spondylitis (AS), Asthma (Ast), Systemic Lupus Erythematosus (SLE), Type 1 diabetes (T1D), Multiple sclerosis (MS), Primary biliary cirrhosis (PBC), Allergic disease (AllD), Ulcerative colitis (UC), Crohn's disease (CD), Inflammatory bowel disease (IBD).

Finally, we quantified the proportion of prioritised genes gained by profiling subpopulations as opposed to bulk cell types. To do so, we compared the number of prioritised genes coming from naive and memory CD4+ T cells as a whole (i.e. pseudobulk) with those observed only at the subpopulation level (i.e. single-cell specific) (**Figure 4.5E**). While most colocalizing genes were identified in bulk cell types (median per trait = 66%), we observed between 2 and 15 new candidate genes per trait (median = 25%) which could only be detected in individual subpopulations. For example, an eQTL for *TYK2* specifically detected in 16h-stimulated $T_{EM}$ cells colocalized with an IBD and a CD association (**Supplementary Figure 4.8B**). Similarly, we identified a GWAS association for CD which colocalized with an eQTL for *ZMIZ1* specific to 16h-stimulated $T_{CM}$ cells (**Figure 4.5F**). A closer analysis of *ZMIZ1* revealed that this eQTL is absent in other memory T cell populations such as $T_{EM}$ cells. This causes the signal to be masked in bulk memory cells, where it is no longer detectable (**Figure 4.5F**). Both of these colocalizations are subpopulation and time point-specific, thus highlighting the added value of our data set for interpreting disease associations.

In conclusion, we used colocalization analysis to nominate 139 genes likely to be causally involved in 14 immune diseases. Dysregulation of the expression dynamics of these genes during CD4+ T cell activation, often in a time point and subpopulation specific manner, could interfere with mechanisms for immune regulation, ultimately enhancing autoimmune inflammation.

## Discussion

Immune disease variants are thought to act by altering gene expression regulation during CD4+ T cell activation (Calderon et al., 2019). However, the exact molecular mechanisms that mediate these effects have not been discovered, and the target genes of most disease loci remain unknown. In this study, we mapped eQTLs throughout CD4+ T cell activation at

unprecedented resolution using scRNA-seq and integrated them with disease associations from GWAS studies.

Our study represents the largest single-cell map of CD4+ T cell activation to date, with over 655,000 high quality transcriptomes spanning four time points and 38 cell populations. This makes our data a unique resource for the design and interpretation of future studies of T cell function and genetic regulation of the immune system. Importantly, by using scRNA-seq we provide an unbiased view of the T cell response, detecting a variety of cell states that change throughout time, as well as modules of genes that share a correlated pattern of expression. This provides a unique opportunity to reinterpret previous observations in the literature. For example, a previous study of CD4+ T cell activation concluded that T cells upregulate a module of IFN-related genes early upon TCR engagement (Ye et al., 2014). Here we recapitulate this observation and conclude that it is driven by one specific subpopulation of naive cells, rather than by the whole T cell pool. Similarly, we confirm previously described changes in the expression of class II MHC molecules in the HLA locus upon T cell activation (Gutierrez-Arcelus et al., 2020), which we find are driven by $T_{EM}$ cells.

We used this T cell activation map to identify over 6,000 genes regulated by eQTLs. Importantly, we did not detect the majority of these eQTLs in resting cells, which highlights the value of profiling T cell activation as opposed to *ex vivo* cells from peripheral blood. This is particularly relevant for interpreting immune-disease associations, as only 40% of our prioritised genes would have been detected using resting cells. These observations strongly support the idea that immune disease variants regulate gene expression during T cell activation. Moreover, they broadly agree with previous colocalization efforts in the context of psychiatric traits, where profiling neuronal differentiation substantially increased the number of colocalizations compared to analysing *ex vivo* cells from the brain (Jerber et al., 2021). This suggests that profiling the right cellular context is key for identifying disease mechanisms. In addition, we showed that genes with higher network connectivity tolerate less genetic variation throughout cell activation. This lack of large *cis* eQTL effects for central genes in our network agrees with theoretical frameworks like the omnigenic model of inheritance, which hypothesise that hub genes in networks could impact disease via *trans* rather than *cis* effects (E. A. Boyle et al., 2017; X. Liu et al., 2019). However, future work is needed in this area to clarify if alterations in hub genes have a larger impact in disease risk.

Measuring expression at the level of single-cells also allowed us to order cells in a pseudotime trajectory and to directly model the interaction between genetics and T cell activation. This resulted in over 2,000 dynamic eQTLs, whose effect size changes with activation time. Similar approaches have been used to study stem cell differentiation (Cuomo et al., 2020) and

cardiomyocyte development (Strober et al., 2019) However, our study is the first to document such gene-by-environment interactions in CD4+ T cells. Colocalization analysis further highlighted the relevance of these effects, with over half of the colocalizations in our study involving a dynamic eQTL. A striking example is the *CTLA4* gene. The expression of *CTLA4* peaks during early activation, showing a noticeable difference between individuals carrying two alleles at a nearby locus. This eQTL colocalizes with GWAS associations for three traits, where the risk allele decreases the expression of *CTLA4*. This agrees with the known role of *CTLA4* in removing costimulatory molecules from the surface of antigen presenting cells early after TCR and CD28 engagement, ultimately initiating a feedback loop that prevents more T cells from activating (Qureshi et al., 2011). Thus, a partial reduction of CTLA4 function in individuals carrying this risk allele could result in impaired immune regulation and higher odds of initiating an autoimmune response.

Finally, we demonstrate that scRNA-seq is able to identify subpopulation-specific eQTLs. For example, we identify eQTLs present in IFN-responding cells (a subset of naive CD4+ T cells), but not in bulk naive cells. Similarly, we identify eQTLs specific to $T_{CM}$ and $T_{EM}$, the two main subsets of memory cells, but which would have been missed if profiling memory T cells in bulk. This confirms previous observations that eQTLs in immune cells can arise in response to a specific stimulus (Alasoo et al., 2018; Fairfax et al., 2014), and that eQTLs obtained from bulk RNA-seq are sometimes explained by a single subpopulation at the single-cell level (van der Wijst et al., 2018). However, we also demonstrate that the statistical power to map eQTLs decreases dramatically for small subpopulations due to uncertainty in estimating average expression values, which limits our study to the analysis of large subpopulations. This should be taken into consideration by future single-cell eQTL studies, which could boost their power by increasing the number of cells collected or using cell enrichment strategies if focusing on rare cell types.

Finally, we show that some subpopulation-specific eQTLs are disease relevant. Both *ZMIZ1* and *TYK2*, for which we detected a $T_{CM}$ and a $T_{EM}$ specific eQTL at 16 hours of activation respectively, colocalize with GWAS associations for Crohn's disease. *TYK2* is particularly interesting, since it is located at a locus previously associated with 10 different autoimmunities and for which at least three independent signals have been reported (Dendrou et al., 2016; Diogo et al., 2015). Previous studies have shown that a missense *TYK2* variant protects against immune disease by partially reducing signalling downstream several cytokine receptors (Dendrou et al., 2016). Here, we show a similar effect, where individuals carrying a protective allele for CD and IBD have lower expression of *TYK2* in $T_{EM}$ cells at 16 hours of activation.

In conclusion, we used scRNA-seq to identify the targets of GWAS loci for immune diseases throughout CD4+ T cell activation. Our results expand our understanding of the genetic susceptibility to immune-mediated diseases and provide a valuable resource for future drug target identification and validation.

## Methods

### Cell isolation and *in vitro* stimulation

Blood samples were obtained from 119 healthy individuals of British ancestry. Of these, 67 were male (53.7%) and 52 female (56.3%), and the mean age of the cohort was 47 years (sd = 15.61 years) (**Supplementary Figure 4.1A**). Human biological samples were sourced ethically and their research use was in accord with the terms of informed consent under an IRB/EC approved protocol (15/NW/0282).

Peripheral blood mononuclear cells (PBMCs) were isolated using Ficoll-Paque PLUS (GE healthcare, Buckingham, UK) density gradient centrifugation. Naïve (CD25- CD45RA+ CD45RO-) and memory (CD25- CD45RA- CD45RO+) CD4+ T cells were isolated from the PBMC fraction using EasySep® naïve CD4+ T cell isolation kits and memory CD4+ T cell enrichment kits (StemCell Technologies, Meylan, France) according to the manufacturer's instructions. Naive and memory T cells were then stimulated with anti-CD3/anti-CD28 human T-Activator Dynabeads® (Invitrogen) at a 1:2 beads-to-cells ratio. Cells were harvested after 16 hours, 40 hours, and 5 days of stimulation. In addition, unstimulated cells kept in culture without any beads for 16 hours were used as a negative control (i.e. zero hours of activation).

### Single-cell RNA-sequencing

Upon harvesting, cells were resuspended in RPMI media to obtain a single-cell suspension. Next, cells were stained with the live/dead dye 4',6-diamidino-2-phenylindole (DAPI) and dead cells were removed from the suspension using fluorescence-activated cell sorting (FACS). Live cells were resuspended in phosphate buffer saline (PBS), at which point cells obtained from different individuals but belonging to the same experimental condition were mixed together at equal ratios to form a single cell suspension (i.e. pool). Each pool corresponded to a mix of cells from four to six different individuals (median = 6), and we processed a total of 172 pools.

Cells were next processed for single-cell RNA-sequencing using the 10X-Genomics 3' v2 kit (G. X. Y. Zheng et al., 2017), as specified by the manufacturer's instructions. Namely, 1 x $10^4$ cells were loaded into each inlet of a 10X-Genomics Chromium controller in order to create GEM emulsions. Each experimental condition was loaded in a separate inlet. The targeted

recovery was 6,000 cells per pool. Reverse transcription was performed on the emulsion, after which cDNA was purified, amplified, and used to construct RNA-sequencing libraries. These libraries were sequenced using the Illumina HiSeq 4000 platform, with 75 bp paired-end reads and one cell pool per sequencing lane.

### Genotyping

Genomic DNA was isolated from a suspension of 1 x $10^6$ PBMCs from each individual in the study using a DNA isolation kit (Qiagen). Genotyping was then performed using the Infinium CoreExome-24 (v1.3) chip (Illumina).

### Genotyping data analysis and imputation

Quality control was performed by removing samples with <95% called genotypes, as well as keeping only variants with MAF > 5%, SNP call rate > 95%, and in Hardy-Weinberg equilibrium (HWE; p-value < 0.001). This resulted in 250,711 variants kept in the dataset.

Imputation of untyped variants was performed as described in Bossini-Castillo et al (Bossini-Castillo et al., 2019). Briefly we used BEAGLE 4.1 (Browning & Browning, 2016) with a reference panel consisting of the 1000 Genomes Phase 3 (1000 Genomes Project Consortium et al., 2015) and the UK10K samples (UK10K Consortium et al., 2015). Variants derived from imputation were quality filtered using the following parameters: allelic R-squared (AR2) >= 0.8, HWE p-value < 0.001, and MAF > 10%. In total we retained 4,641,778 variants after imputation, which were used for eQTL analysis. We confirmed that all but one of the individuals in our cohort clustered with the British in England and Scotland (GBR) population in PCA space (**Supplementary Figure 4.1B**). This individual clustered with the Finnish population and was removed from any further analyses. Finally, we identified four pairs of related individuals. These were kept for every analysis except eQTL mapping, where one random individual from each pair was kept.

### Single-cell RNA-sequencing data analysis

#### Data processing and quality control

Raw scRNA-seq data were processed using the Cell Ranger Single-Cell Software Suite (G. X. Y. Zheng et al., 2017) (v3.0.0, 10X-Genomics). In brief, reads were first assigned to cells and then aligned to the human genome using STAR (Dobin et al., 2013), with the hg38 build of the human genome (GRCh38) as a reference for alignment. Ensembl (v93) was used as a reference for gene annotation, and gene expression was quantified using reads assigned to cells and confidently mapped to the genome.

Results from RNA quantification in Cell Ranger were imported into Python (v3.8.1) and analysed using scanpy (v1.4.4) (Wolf et al., 2018). Samples with less than 70% of reads mapping to cells were discarded. This resulted in 142 (82%) cell pools and 106 (89%) individuals being kept after quality filters. In addition, any cells with less than 200 detected genes, an unusually high number of genes (defined as over four standard deviations above the mean number of detected genes), or more than 10% of reads mapping to mitochondrial genes were removed from the data set. Finally, any genes detected in less than 10 cells were discarded. This resulted in 713,403 cells (96.77% of total) and 23,360 genes passing quality filters.

### Deconvolution of single cells by genotype

Each scRNA-seq sample comprised a mix of cells from unrelated individuals. Thus, natural genetic variation was used to assign cells to their respective individuals. First, a list of common exonic variants was compiled from the 1000 genomes project phase 3 exome-sequencing data (1000 Genomes Project Consortium et al., 2015). This list included any variants with a minor allele frequency of at least 5% in the European population. Next, cellSNP (v0.99) (Y. Huang et al., 2019) was used to generate pileups at the genomic location of these variants. These pileups, in combination with the variants called from genotyping in each individual, was used as an input for Vireo (v1), (Y. Huang et al., 2019). Vireo uses a Bayesian approach to infer which cells belong to the same individual based on the genetic variants detected within scRNA-seq reads. Any cells labeled as “unassigned” (less than 0.9 posterior probability of belonging to any individual) or “doublets” (containing mixed genotypes) by Vireo were discarded. On average, 92% of the cells in each pool were unambiguously assigned to a single individual in the cohort (**Supplementary Figure 4.2A**).

### Cell cycle scoring

After quality control, the number of unique molecular identifiers (UMIs) mapping to each gene in each single cell were normalized for library size and log-transformed using scanpy’s default normalization parameters (Wolf et al., 2018). Next, a publicly available list of cell cycle genes (Kowalczyk et al., 2015) was used in combination with scanpy to perform cell cycle scoring and assign cells to their respective stage of the cell cycle.

### Exploratory data analysis and removal of cellular contaminations

We performed exploratory analysis at each experimental time point independently. Cells collected at the same time point were first loaded into scanpy, where normalised log-

transformed UMI counts were used to identify highly variable genes. Between 701 and 1,668 highly variable genes were detected at each time point (mean = 1301). Only highly variable genes were used as a basis for the remaining analyses in this section.

Having identified highly variable genes, technical covariates (cell culture batch) and unwanted sources of biological variation (i.e. number of UMIs per cell, proportion of reads mapping to mitochondrial genes, cell cycle scores, and reported sex) were regressed out using scanpy's regress_out() function. Next, log-UMI counts were scaled (setting 10 as the maximum value) and used as an input for principal component analysis. The first 40 principal components were used to build a k-nearest neighbours (kNN) graph (with k=15), which was used as an input for embedding and visualization with the uniform manifold approximation and projection (UMAP) algorithm (L. McInnes et al., 2018). This kNN graph was further used for unsupervised clustering using the Leiden algorithm (Traag et al., 2019).

At this stage, cell clustering revealed a low proportion of three contaminating cell types which were consistently detected at each time point: B cells, CD8+ T cells, and antigen-presenting cells (APCs). Furthermore, two additional sources of contamination (SOX4+ precursor cells, and cells expressing hallmarks of cell culture stress) were detected at zero hours of activation (**Supplementary Figure 4.3**). Cell contaminations were removed from the data set, resulting in 655,349 (91.86% of total) high quality cells kept and successfully annotated as CD4+ T cells.

### Identification of a lowly active T cell state

Having removed cellular contaminations, highly variable genes were recalculated and the analysis described in the previous section (i.e. batch regression, scaling, PCA, graph construction, embedding, and clustering) was repeated using CD4+ T cells only. Cells sampled at 16h and 40h showed a clear separation into two groups, one of which expressed a significantly lower number of genes and showed comparatively lower levels of the T cell activation markers described in Supplementary Table 3.2 (**Supplementary Figure 4.4**). This population of lowly active cells was separated from its original time point and treated as an independent group for clustering.

### Cell clustering and cluster annotation

Unsupervised clustering was applied independently to the five cell groups of cells identified in the study (resting, lowly active, 16 hours, 40 hours, and five days) based on their respective kNN graphs and using the Leiden algorithm (Traag et al., 2019). This resulted in 51 cell

clusters. The similarity of these clusters to each other was assessed by performing PCA on the full data set (i.e. all cells) and estimating the Euclidean distance between pairs of clusters (from cluster centre to cluster centre) based on the first 100 principal components. Clusters with high levels of similarity or overlapping biological characteristics were merged together (**Supplementary Figure 4.5**). This resulted in 38 distinct groups of cells. Gene markers for each of these groups were identified using scanpy's built-in function for gene ranking, which uses a T test to compare the average expression of a gene in a cluster versus its expression outside the cluster. Each cell group was annotated by comparing its inferred marker genes with known cell type markers reported in the literature.

### Ordering of cells in a pseudotime trajectory

To perform trajectory inference, raw gene expression measurements for all CD4+ T cells in the study (i.e. 655,349 cells spanning all time points) were imported into R (v3.6.1) and analysed using monocle3 (v0.2.0) (Cao et al., 2019). As opposed to other analyses, where cells from each time point were treated independently, here some unwanted sources of variation such as cell cycle scores correlated with the biological process of interest (i.e. T cell activation). Thus, we implemented a hierarchical batch regression approach, where cell cycle scores were first regressed within each time point, followed by batch regression in the full data set. In brief, principal component analysis was performed based on all cells using monocle3's PCA implementation. Next, a matrix containing the first 100 principal component coordinates for each cell was split by time point. Cell cycle effects were then regressed from each sub-matrix independently using limma's lmFit function (Ritchie et al., 2015). Finally, these cell cycle-corrected matrices were merged back into a full PCA matrix and cell culture batch effects were regressed based on the full data set using the mutual nearest neighbours (MNN) algorithm (Haghverdi et al., 2018).

After batch correction, the first 100 principal components were used to build a kNN graph and this graph was embedded into a two-dimensional space using UMAP. Finally, UMAP coordinates were used to infer a branched pseudotime trajectory using monocle3's learn_graph function. To identify genes that changed as a function of pseudotime, monocle3's graph test was applied to all genes. This test assesses whether cells adjacent in the trajectory show more correlated expression of a gene than cells which are far apart (i.e. autocorrelation). Correction for multiple testing was performed using the Q value procedure (John D. Storey, 2003). A gene was considered as significantly associated with pseudotime if it had a Q value ≤ 0.05 and a Moran's I (a measurement of the magnitude of autocorrelation) larger than 0.05 (Moran, 1950).

### Co-expression network analysis

In order to preserve most of the resolution provided by scRNA-seq while also minimising zero-inflation and technical noise, gene expression was aggregated per individual and per cluster into a pseudobulk count matrix. More specifically, raw UMI counts per gene were summed across all cells belonging to the same cluster and to the same individual. Summed expression values were normalised by library size and gene length, resulting in a single expression matrix with transcript per million (TPM) units (i.e. a pseudobulk matrix).

The resulting pseudobulk matrix was imported into R (v3.6.1) and analysed using the weighted gene co-expression network analysis (WGCNA) package (v1.69) (Langfelder & Horvath, 2008). Only genes with ≥ 1 TPM in at least 30 samples were used for this analysis. TPM values were first log-transformed, after which unwanted sources of variation (i.e. cell culture batch, reported sex, and age) were regressed out using limma's (v3.40.6) removeBatchEffect() function (Ritchie et al., 2015). Next, genes were filtered by their level of variability, with only genes showing a standard deviation ≥ 0.1 across samples being kept. This resulted in 11,130 genes taken forward for network construction.

The functions available in WGCNA were used to calculate network properties for these genes, such as their level of connectivity, as well as to build an adjacency matrix. The soft power parameter was set to 4 in these analyses, based on an evaluation of the [0,20] parameter space. The resulting adjacency matrix was transformed into Topological Overlap Matrices (TOM) of similarity and dissimilarity. Finally, hierarchical clustering was applied to the TOM dissimilarity matrix in order to build a dendrogram of genes. Gene modules were inferred from this dendrogram using R's dynamicTreeCut package (v1.63.1). Mean module expression values were calculated as the average expression of all genes belonging to that module in a given sample.

### Pathway enrichment analysis

All pathway enrichment analyses were performed using gprofiler2 (v 0.1.9), setting the gene list of interest as an unordered query and using all genes detected in the study as the background. Only enriched terms derived from the Kyoto Encyclopedia of Genes and Genomes (KEGG) or the REACTOME database were kept. gprofiler2's built-in approach (gSCS) was used for multiple testing correction. Enriched pathways were visualized in R using the pheatmap package (v1.0.12).

### eQTL mapping

### Mapping of eQTLs at four T cell activation time points

For each gene we calculated mean expression per cluster per donor. To ensure the high quality eQTL mapping, we only kept genes with non-zero expression in at least 10% of donors and mean count per million (cmp) higher than one. We retained between 8,940 and 11,516 genes. To identify cis-eQTLs we used tensorQTL (v1.0.3) (Taylor-Weiner et al., 2019) to run a linear regression for each SNP-gene pair, using a 500 kilobase window within the transcription start site (TSS) of each gene (i.e. cis_nominal mode). We regressed the first 15 gene expression principal components from this analysis, so as to capture the confounders within our dataset. To correct for the number of association tests performed per gene, we used a cis permutation pass per gene with 1000 permutations. Finally, to correct for the number of genes tested and identify significant eGenes we performed a q-value correction (J. D. Storey & Tibshirani, 2003) for the top associated SNP-gene pair, setting a q-value threshold of 0.1.

### Analysis of eQTL sharing across time points and cell types

To assess the sharing between eQTLs, we performed a meta-analysis across cell types and cell states using the multivariate adaptive shrinkage (mashR) method (Urbut et al., 2019). MashR is a Bayesian method which estimates the pairwise level of sharing between cell states, where an eQTL is defined as shared if it has the same effect size (within a factor of 0.5) and direction in two cell states.

### Modelling of eQTL effect sizes as a function of network centrality

To assess the relationship between a gene's genetic regulation and network centrality, the effect size of each gene's lead eQTL variant was modelled as a function of the gene's centrality value in the co-expression network described above. This was first done assuming a linear relationship. However, substantial heteroskedasticity was observed, which suggested a non-linear relationship, as confirmed using a Breusch-Pagan heteroskedasticity test (Breusch & Pagan, 1979). Thus, we log-transformed the eQTL effect sizes, which resulted in homoskedastic data and a strong linear relationship between the variables. All linear models were built and tested using base R's lm() function.

### Modelling of time-dependent (dynamic) eQTL effects

To identify pseudotime-dependent eQTL effects, we divided the activation trajectory into 10 windows containing roughly equal numbers of cells (i.e. pseudotime deciles) and averaged the expression of each gene per individual within each window. To facilitate the interpretation of coefficients, pseudotime windows were scaled from 0 to 1 prior to this analysis. In order to account for the higher correlation in expression values derived from the same individual at

multiple pseudotime windows, we applied linear (1) and quadratic (2) mixed models, with individuals modelled as random intercepts. We used these models to test for a significant interaction between genotypes (i.e. the genetic dosage carried by each individual at the lead eQTL variant for that gene) and pseudotime, as follows:

1. Z_score ~ Genotype + Pseudotime + Cell_culture_batch + Sex + Age + Genotype*Pseudotime + (1 | Donor)
2. Z_score ~ Genotype + Pseudotime + $Pseudotime^2$ + Cell_culture_batch + Sex + Age + Genotype*Pseudotime + Genotype*$Pseudotime^2$ + (1 | Donor)

In both cases, the null model was computed using the same parameters while excluding the Genotype*Pseudotime and Genotype*$Pseudotime^2$ terms. P-values were calculated by comparing each model to its respective null model using analysis of variance (ANOVA). Multiple testing correction for the number of genes tested was performed using the q-value approach (at a q-value threshold < 0.1). All models were implemented in R using the *lmer()* function. In order to reduce the burden imposed by multiple testing, we only applied this approach to variants previously identified as significant lead eQTL variants for a gene by tensorQTL in at least one time point. This was done separately for naive and memory T cells.

### Estimation of pairwise LD

We performed LD calculations based on the individual-level genotype information for the individuals in this study obtained from genotyping. Namely, we used PLINK (v1.90b4) (Purcell et al., 2007) to calculate the correlation between pairs of genetic variants across all individuals in the cohort, using either the --r or the --r2 flags. We restricted this calculation to variants with MAF > 0.01, computing correlations for any pairs of variants located within 1 Mb of each other and at most 500 variants away. The command used was: *plink --r/--r2 --ld-window 500 --ld-window-kb 1000 --maf 0.01 --vcf-half-call h*

### Integration of eQTLs with GWAS signals

#### Pre-processing of GWAS summary statistics

Full summary statistics files from previous GWAS studies were downloaded from the GWAS catalogue (MacArthur et al., 2017). These files corresponded to a recent release containing harmonised summary statistics, which were lifted over to build GRCh38 of the genome and for which the effect size and effect direction of each variant was aligned to the reported alternative allele (Buniello et al., 2019). Any signals coming from the X or Y chromosomes, as well as from the MHC region (ch6:28,510,120 – chr6:33,480,577) were discarded. These

summary statistics corresponded to 14 immune-mediated diseases: CeD (Trynka et al., 2011), RA (Eyre et al., 2012), SLE (Bentham et al., 2015), T1D (Onengut-Gumuscu et al., 2015), MS (Beecham et al., 2013), ankylosing spondylitis (International Genetics of Ankylosing Spondylitis Consortium (IGAS) et al., 2013), asthma (Demenais et al., 2018), primary biliary cirrhosis (Cordell et al., 2015), allergic disease (Ferreira et al., 2017), juvenile idiopathic arthritis (Hinks et al., 2013), psoriasis (Tsoi et al., 2012), UC, CD, and IBD (de Lange et al., 2017).

### Colocalization analysis

Genomic loci of interest were identified by intersecting eQTL signals in each cell type with GWAS loci for 14 immune-mediated diseases. For each trait-cell type pair, we applied colocalization to any locus where a lead variant for a significant eQTL (q value < 0.1) was located within 100 kb and in high LD ($r^2 > 0.5$) with a significant GWAS variant (i.e. any GWAS variant with nominal p value < $1 \times 10^{-5}$, which enabled us to capture suggestive association signals).

At each of these loci, coloc (v4.0.4) was used to test for colocalization between the eQTL and the GWAS signals. Importantly, these analyses were based on the recently developed masking approach, which relaxes coloc's previous assumption of a single causal variant per locus (Wallace, 2020). This is similar to performing conditional analyses at each locus. In brief, we defined a 500 kb window centred at the lead eQTL variant and tested for colocalization using all common variants located in the window and present in both the eQTL and the GWAS summary statistics. We used the pairwise LD calculations from our cohort as a basis for masking, setting an $r^2$ threshold of 0.01 to separate independent signals. Coloc's prior parameters were set to their recommended values in the most recent publication (Wallace, 2020) ($p1=1\times10^{-4}$, $p2=1\times10^{-4}$, $p12=5\times10^{-6}$). Significant colocalizations were defined as any instances where the estimated posterior probability of a shared causal variant (PP4) was ≥ 0.8. In order to discard potential false positives due to noisy association signals, we only kept for further analysis traits with more than one significant colocalization (12 out of 14 traits).

To infer the relationship between gene expression and disease risk at each locus, we estimated the GWAS and eQTL effect sizes (i.e. $\log_e$OR and gene expression Z-score) for the GWAS variant in highest LD with the lead eQTL variant at the locus. We concluded that a variant increased disease risk via an increase in gene expression if the variant had the same direction of effects in both studies. In the opposite case, we concluded that the variant increased disease risk via a decrease in gene expression. If the same variant had different

estimates of eQTL effect size in different T cell populations, we required that all effect sizes had the same direction.

## Code and data availability

The raw single-cell RNA-sequencing data described in this study will be deposited in the European Genome-Phenome Archive (EGA) upon publication. In addition, eQTL lead signals and summary statistics will be deposited in the eQTL catalogue (https://www.ebi.ac.uk/eqtl/). Moreover, a count matrix with processed scRNA-seq UMI counts for all cells in the study, as well as lower dimensional representations of the data set, will be made available via the Open Targets website. Finally, all the codes used for processing and analyzing the data in this study are compiled in a GitHub repository, which will be made publicly available upon publication.

# Conclusions and future perspectives

The work presented in this dissertation illustrates the challenges in understanding the impact that common genetic variants have in immune function. Moreover, it highlights recent progress in linking variants associated with immune diseases to the cell types and contexts where they act, ultimately identifying some of their target genes. At the time this work started, it was clear that immune-mediated diseases had a highly polygenic architecture involving many loci (Visscher et al., 2017), and that disease loci were enriched in regulatory elements specifically active in CD4+ T cells, particularly in memory cells and $T_{regs}$ (Hu et al., 2011; Trynka et al., 2013, 2015). However, it was not known how these loci modified T cell function. In particular, no studies had systematically assessed if immune-disease GWAS loci were enriched in specific CD4+ T cell subsets, such as known T helper phenotypes (e.g. Th1, Th2, Th17, and iTreg) (Federica Sallusto, 2016; Zhu et al., 2010). This was partly because the chromatin landscape and gene expression profiles of these cell subsets had not been systematically mapped. In addition, even though memory cells were thought to play a central role in immune diseases, their plasticity had not been characterised at the level of their transcriptome, and it was unclear whether and how they responded to certain cytokines compared to naive cells, as well as how they adapted to changes in their microenvironment. Finally, previous efforts to identify the target genes of GWAS loci had been based on eQTLs obtained from immune cells isolated from blood (Chen et al., 2016; Onengut-Gumuscu et al., 2015; Wallace et al., 2012). These mostly comprise cells in resting state and may not be the most relevant to study disease mechanisms, given that immune cells exert their function by responding to antigen in tissues and secondary lymphoid organs. This dissertation attempted to bridge some of these gaps by building a set of resources that could enable a better understanding of the role of CD4+ T cells in complex immune diseases. In this final chapter, I briefly summarise the findings from this endeavour, discuss their limitations, and highlight their contribution to our broader understanding of common immune diseases.

## Findings and limitations from this dissertation

### Most immune disease GWAS loci function during early T cell activation

We first asked which of the many T helper cell states documented in the literature underlied the SNP enrichment signals observed in previous studies. Given the lack of epigenomic datasets which could be used for this purpose, we started by building detailed maps of regions of open chromatin (using ATAC-seq), as well as active enhancers and promoters (as tagged by the histone mark H3K27ac), present across 55 *in vitro* generated immune cell states. These

cell states encompassed relevant macrophage (Martinez & Gordon, 2014) (e.g. M1 and M2) and T helper cell phenotypes (e.g. Th1, Th2, Th17, iTreg) which we profiled at two time-points of activation.

Our dataset revealed that, unlike broad immune cell types like B cells and T cells, which tend to be different from one another in terms of their active regulatory elements (Chen et al., 2016; ENCODE Project Consortium, 2012), T cell states showed similar and correlated chromatin landscapes. Even loci known to harbour lineage-defining genes, such as the hallmark Th2 transcription factor *GATA3* or the Th1 gene *TBX21 (S. J. Szabo et al., 2000; W. Zheng & Flavell, 1997)*, were active across many cell states. However, we observed differences in the quantitative levels of H3K27ac deposited at hallmark loci, which tended to be more acetylated in the lineages in which they were active. This highly correlated structure violated the assumptions of previous SNP enrichment methods and prompted us to develop a new statistical framework (*CHEERS*) to assess if disease-associated variants were overrepresented within regulatory elements with higher levels of activity in one cell state as compared to the others. This method is now publicly available and will become more relevant as chromatin datasets grow in size and scope.

By applying this method across all the cell states in our study, we discovered that Alzheimer's disease loci are enriched in macrophages. Conversely, autoimmunity and autoinflammation loci are enriched in T cells. Furthermore, we found that Th1 cells are central in inflammatory bowel disease and its subtypes, in agreement with the known involvement of hallmark Th1 genes in this disease (Neurath et al., 2002). However, in contrast to our expectations of finding a clear disease-causal T helper phenotype, the highest enrichment signal for immune disease loci came from all memory cell states profiled after 16 hours of stimulation, a time point when fully defined T helper phenotypes have not yet appeared. This strongly suggested to us that most disease variants do not affect any specific Th cell type, but rather the processes occurring early after T cells interact with antigen presenting cells, before differentiation into effector phenotypes begins. This was supported by the presence of T cell activation genes nearby many disease associated loci. A similar conclusion was also reached in a study published at the same time as ours, which sorted individual cell subsets and profiled resting and activated immune cells using ATAC-seq (Calderon et al., 2019). These findings challenged our original assumptions and convinced us that the best approach to identify disease genes was to focus on memory T cells and T cell activation as a process, a topic to which I return further below.

Despite the important advances enabled by this study, certain limitations should be acknowledged and could be addressed in the future. Firstly, our method to quantify GWAS SNP enrichment is based on variants associated with disease below the genome-wide

significance threshold (p-value < 5 x $10^{-8}$). The number of variants that reach this threshold depends strongly on the sample size and power of the GWAS study in question. Thus, our method is not able to reliably identify enrichments for traits with a lowly-powered GWAS. To address this, one could adopt methods that leverage all common variants in the genome, such as the partitioning heritability approaches (Bulik-Sullivan et al., 2015; Finucane et al., 2015, 2018). However, since these approaches work by segmenting variants into different categories according to their overlap with functional annotations, in their current state it is unclear how they could incorporate quantitative levels of chromatin activity in the way done by our method. A more promising approach would be to explicitly model the dependency of GWAS variant effect sizes on the chromatin activity levels at each locus (Calderon et al., 2017). Nonetheless, this is currently computationally intensive and methodological advances are needed to achieve it. Another area of research which could offer a solution is that of polygenic risk scores (Abraham et al., 2016; Khera et al., 2018, 2019) and their integration with molecular phenotypes (Richardson et al., 2019; Võsa et al., 2018). For example, it could be possible to evaluate whether weighing variants by their chromatin activity in relevant cell types increases risk prediction accuracy. In the near future, a combination of these approaches could make it feasible to extend our approach to all variants in the genome.

A further limitation of our study is that the results obtained using *CHEERS* represent the added contribution of many GWAS variants spread across the genome, which in our data tend to localise to enhancers and promoters involved in T cell activation. Nonetheless, we cannot discard the possibility that a subset of disease associated loci may act via a completely different cell type. The signal coming from these loci, especially if they represent a small proportion of the total, would be masked by the most prominent enrichment signals. In order to circumvent this limitation, one could combine our approach with methods designed to identify the most likely tissue of action for each locus (Corradin et al., 2016; Factor et al., 2020; Torres et al., 2020). This would open the possibility of not only following up GWAS variants that act during T cell activation, as done in this dissertation, but also dissecting the contribution of less prominent secondary cell types, which could reveal unexplored therapeutic avenues.

### Naive and memory T cells form a continuum

The observations just described demonstrated that memory T cells are particularly important in immune diseases. However, it was difficult to understand what set them apart from naive T cells without looking at their function in closer detail. Thus, we performed a comprehensive comparison of the gene expression profiles of naive and memory cells, focusing particularly

on how both cell types respond to the cytokines present in their microenvironment. We assessed this phenotype because cytokine blockade is a successful therapeutic strategy for many immune-diseases (I. B. McInnes et al., 2016; Schett et al., 2013), and we hypothesised that memory cells could be responding aberrantly to these signals in immune-mediated diseases. Surprisingly, while there was a vast body of literature documenting the response of naive T cells to cytokines (Harrington et al., 2005; Kanhere et al., 2012; Mosmann et al., 1986; S. J. Szabo et al., 2000), very few studies focused on memory T cells. The few which did, however, showed that memory cells retained a certain degree of plasticity (Putheti et al., 2010) and could re-differentiate in disease settings (Komatsu et al., 2014; Nistala et al., 2010). Thus, we exposed naive and memory T cells to 11 different cytokine combinations and used a multi-omic approach (combining RNA-seq and proteomics) to study their responses.

Our results confirmed that memory cells retain phenotypic plasticity, however we found that this was greatly diminished compared to naive cells. For example, RNA expression did not significantly change in memory cells in response to IL-4. Moreover the cytokine TGFβ, known to induce an immunosuppressive phenotype (iTreg) in naive T cells, instead biased memory cells to a pro-inflammatory phenotype characterised by high levels of IL-17 and IL-9, similar to the Th17 phenotype. This could help explain the high levels of IL-17 observed in some disease tissues (Babaloo et al., 2013), where memory cells could be differentiating into this pro-inflammatory cell state, thus exacerbating inflammation.

While this demonstrated important differences between the function of naive and memory cells, our results were limited by the fact that we did not know the composition of the memory T cell population. Since by definition memory cells must have arisen in response to a previous infection, it is possible for them to retain a memory of their previous phenotype in the form of epigenetic marks or residual gene expression (Kanno et al., 2012; O'Shea & Paul, 2010), which would confound any observations at the level of the whole cell population. Moreover, while a long tradition of immunological studies had demonstrated the existence of two different subtypes of memory T cells in blood ($T_{CM}$ and $T_{EM}$) (F. Sallusto et al., 1999; Federica Sallusto et al., 2004), these cells had not been profiled at high resolution using single-cell approaches. To address this, we applied droplet-based scRNA-seq to profile naive and memory cells at single-cell level, both in the resting state and as they responded to cytokines. This constitutes the largest single-cell map of T cell responses to cytokines to date. Our analysis found no evidence of distinct and separate clusters of T helper cells, but rather showed that CD4+ T cells form a single population which resembled a continuum. This observation at the transcriptome level contrasted starkly with previous studies at the protein level using flow cytometry and mass cytometry-based approaches. However, others have since confirmed similar observations to ours (Kiner et al., 2021; P. A. Szabo et al., 2019; Trzupek et al., 2020),

thus challenging the previous T helper paradigm and giving space for a model where CD4+ T cells form a spectrum of cell states with high plasticity (van Beek et al., 2021). In particular, we showed that the subpopulations which constitute CD4+ T cells reflect the existence of a naive-to-memory progression. This progression starts with naive cells, in turn giving rise to $T_{CM}$, $T_{EM}$ and $T_{EMRA}$ cells. We named this phenomenon “T cell effectorness”, so as to reflect its continuous nature and the fact that as cells progress from naive to memory they tend to upregulate effector molecules like cytokines and chemokines. A similar progression has since been observed in CD8+ T cells, where higher effectorness seems to correlate with better CD8+ T cell responses to melanoma (Fairfax et al., 2020).

Our results contributed to changing the previous perception of memory T cells as fully differentiated and committed to a lineage, as well as to show that CD4+ T cells are a continuum of cell states. However, there are still some limitations which need to be addressed. Perhaps the most important caveat of our analysis is polyclonality. Because memory cells are a mixture of many different T cell clones, each of which arose in response to different antigens, it is not possible to directly connect our observations with the specific context in which these cells would function. For example, studies in the context of chronic infections like cytomegalovirus (CMV) and human immunodeficiency virus (HIV) infection have shown that chronic infection can induce a CD8+ T cell phenotype similar to that observed as highly effector cells in our study (Appay et al., 2000; Champagne et al., 2001). While we addressed this indirectly using publicly available TCR sequences, future studies should leverage scRNA-seq technologies compatible with TCR-seq protocols, so as to obtain matched TCR sequences and transcriptomes for every cell. This would give a more detailed picture of which clones contract and expand and how this relates to the T cell states observed in scRNA-seq. However, studying TCR repertoires only scratches the surface of a more challenging problem, since it is not known which antigens the majority of TCR sequences recognise, and it is estimated that each TCR could recognise up to thousands of different peptides (Sewell, 2012; Wooldridge et al., 2012). A promising approach is to build systems for assessing the ability of T cells to recognise large libraries of antigens, ideally with high enough resolution to study polyclonal T cell populations (Sharma et al., 2019). However, coupling these techniques with scRNA-seq is not currently possible, so advances in this area will require a combination of several independent experimental methods, as well as further technological developments.

### Immune disease loci modify gene expression dynamics in activated T cells

Based on the observations discussed above, we concluded that most immune disease variants act during early memory T cell activation, and that memory cells differ from naive cells in several functional aspects, such as their restricted plasticy in response to cytokines, and in

forming a continuum of cell states. This suggested that in order to map the target genes for immune disease loci we needed to study memory T cell activation at a fine temporal scale and with single-cell resolution. Thus, we set out to map single-cell eQTLs throughout CD4+ T cell activation using cells from 119 individuals. This is the largest single-cell map of resting and activated T cells to date, with over half a million transcriptomes and 38 distinct subpopulations and cell states. Using it, we prioritised 139 genes as targets of immune disease GWAS loci that exert their function during CD4+ T cell activation.

One particular advantage of using single-cell as opposed to bulk RNA-seq was that it enabled us to identify eQTLs which were present in specific T cell subpopulations but absent in naive or memory T cells as a whole. These represent mechanisms of gene expression regulation, many of which would likely have been missed by bulk sequencing. This is the case of *ZMIZ1* and *TYK2*, which we detected as eGenes after 16 hours of activation in $T_{CM}$ and $T_{EM}$ cells, respectively. Both of these genes colocalize with immune disease associations, but in both cases the effects are absent in other subpopulations, which results in these effects being diluted when looking at whole memory T cells. While single-cell eQTL mapping frameworks are only beginning to emerge, our results already echo the observations from other studies. In particular, previous studies have shown that eQTL effects measured in one cell type can be masked by other cell types (van der Wijst et al., 2018) and that using scRNA-seq to study cell differentiation can dramatically increase the number of identified disease-relevant genes when compared to *ex vivo* tissues (Jerber et al., 2021). Both of these findings are recapitulated here.

A particularly exciting aspect of single-cell eQTL studies is that they enable reliable mapping of gene-by-environment interactions. While these interactions have been extensively documented in bulk, where it is known that cell stimulation can induce many new eQTLs that were not present in resting cells (Alasoo et al., 2018; Fairfax et al., 2014), scRNA-seq opens new possibilities such as investigating how eQTL effect sizes change throughout cell differentiation or cellular responses to a stimulus (Cuomo et al., 2020). Because pseudotime ordering and trajectory inference methods can reconstruct these processes at unprecedented resolution using fewer time points (Cao et al., 2019; Saelens et al., 2019), this type of analysis is greatly facilitated by scRNA-seq. We took advantage of this possibility to model the dependence of eQTL effect sizes on T cell activation time in both naive and memory T cells, ultimately identifying more than 2,000 genes with dynamic genetic regulation mechanisms. We not only showed that these effects exist, but further demonstrated their relevance for human disease by showing that over half of the colocalizations we detect between eQTLs and GWAS associations involved this type of effects, a significant enrichment over the expected null distribution. *CTLA4*, *FADS1*, and *FADS2* all provide examples of GWAS loci that regulate gene expression only during early or late activation, with reduced effects at other time points.

Taken together, these observations prove that dozens or even hundreds of loci associated with immune-mediated diseases increase disease risk by altering the expression level of nearby genes in CD4+ T cells, often during specific stages of T cell activation. This proves that fine-tuned regulation of T cell expression dynamics is crucial for eliciting a healthy immune response. However, being amongst the first single-cell eQTL studies, our analysis suffers from few caveats. The most important is the power limitation imposed by low cell numbers across many rare cell types and low frequency subpopulations. Our analysis shows that as cell populations decrease in size, the accuracy in estimating the expression level and variance of each gene drops, thus limiting eQTL detection. Future studies should take this into account by including cell numbers and estimates of subpopulation size into their power calculations. New eQTL mapping methods specifically designed for single-cell data and which take this relationship into their modelling assumptions are already on their way (Cuomo et al., 2020), but remain computationally intensive and further developments on this area will be transformative for the field.

Another limitation lies in interpreting the results from our colocalization analysis. Firstly, recent simulation studies have suggested that at current sample sizes many eQTL datasets are underpowered to detect most of the existing colocalizations (Hukku et al., 2021). This will be especially pronounced for single-cell data sets, possibly resulting in most disease genes not being captured in our study. Moreover, discrepancies have been documented between colocalization-based and Mendelian Randomization-based approaches, showing that not all significant colocalizations are expected to be causal (J. Zheng et al., 2020). This highlights the need to functionally validate the hits identified by colocalization either using CRISPR-based gene editing tools or by carrying out follow up experiments carefully designed for each gene. Due to the time limitations of this study, it was not feasible to thoroughly validate the colocalizing genes. However, future work should focus on gathering evidence that these genes are indeed regulated by the hypothesised GWAS loci and that their dysregulation causes disease.

Finally, because the scRNA-seq approaches used in this dissertation are based on the 3’ end capture of RNA fragments, they do not contain enough information to reconstruct full transcripts. This prevented us from using orthogonal approaches like allelic specific expression (Castel et al., 2015) or from assessing other molecular mechanisms independent of gene expression like changes in isoform abundance (splicing QTLs) or promoter usage (Alasoo et al., 2015; Y. I. Li et al., 2018). Technological advances enabling the reconstruction of full transcripts in large numbers of single-cells would greatly propel the discovery of non-eQTL mediated disease mechanisms.

## Concluding remarks

Immune-mediated diseases are complex and poorly understood. However, the combination of novel genomics and immunological approaches is proving invaluable in disentangling their molecular causes. In this dissertation, which summarises four years of experimental and computational work, I presented the results from three studies which attempted to understand the role of CD4+ T cells in immune disease, borrowing concepts from functional genomics, epigenetics, statistical genetics, and single-cell transcriptomics. These results show that most of the genetic loci associated with immune-mediated diseases are expected to affect the way memory CD4+ T cells respond to activation, and that many of them might alter gene expression dynamics in ways that are exquisitely cell type and time-point specific.

# References


1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.

Abraham, G., Havulinna, A. S., Bhalala, O. G., Byars, S. G., De Livera, A. M., Yetukuri, L., Tikkanen, E., Perola, M., Schunkert, H., Sijbrands, E. J., Palotie, A., Samani, N. J., Salomaa, V., Ripatti, S., & Inouye, M. (2016). Genomic prediction of coronary heart disease. *European Heart Journal*, *37*(43), 3267–3278.

Aijö, T., Edelman, S. M., Lönnberg, T., Larjo, A., Kallionpää, H., Tuomela, S., Engström, E., Lahesmaa, R., & Lähdesmäki, H. (2012). An integrative computational systems biology approach identifies differentially regulated dynamic transcriptome signatures which drive the initiation of human T helper cell differentiation. *BMC Genomics*, *13*, 572.

Alasoo, K., Martinez, F. O., Hale, C., Gordon, S., Powrie, F., Dougan, G., Mukhopadhyay, S., & Gaffney, D. J. (2015). Transcriptional profiling of macrophages derived from monocytes and iPS cells identifies a conserved response to LPS and novel alternative transcription. *Scientific Reports*, *5*, 12524.

Alasoo, K., Rodrigues, J., Danesh, J., Freitag, D. F., Paul, D. S., & Gaffney, D. J. (2019). Genetic effects on promoter usage are highly context-specific and contribute to complex traits. *eLife*, *8*, e41673.

Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A. J., Mann, A. L., Kundu, K., HIPSCI Consortium, Hale, C., Dougan, G., & Gaffney, D. J. (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nature Genetics*, *50*(3), 424–431.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 30.

Anzalone, A. V., Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., Chen, P. J., Wilson, C., Newby, G. A., Raguram, A., & Liu, D. R. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, *576*(7785), 149–157.

Appay, V., Nixon, D. F., Donahoe, S. M., Gillespie, G. M., Dong, T., King, A., Ogg, G. S., Spiegel, H. M., Conlon, C., Spina, C. A., Havlir, D. V., Richman, D. D., Waters, A., Easterbrook, P., McMichael, A. J., & Rowland-Jones, S. L. (2000). HIV-specific CD8(+) T cells produce antiviral cytokines but are impaired in cytolytic function. *The Journal of Experimental Medicine*, *192*(1), 63–75.

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., & Stark, A. (2013). Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science*, *339*(6123), 1074–1077.

Arsenio, J., Metz, P. J., & Chang, J. T. (2015). Asymmetric Cell Division in T Lymphocyte Fate Diversification. *Trends in Immunology*, *36*(11), 670–683.

Babaloo, Z., Yeganeh, R. K., Farhoodi, M., Baradaran, B., Bonyadi, M., & Aghebati, L. (2013). Increased IL-17A but decreased IL-27 serum levels in patients with multiple sclerosis. *Iranian Journal of Immunology: IJI*, *10*(1), 47–54.

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, *21*, 381.

Banovich, N. E., Lan, X., McVicker, G., van de Geijn, B., Degner, J. F., Blischak, J. D., Roux, J., Pritchard, J. K., & Gilad, Y. (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genetics*, *10*(9), e1004663.

Barbeira, A. N., Bonazzola, R., Gamazon, E. R., Liang, Y., Park, Y., Kim-Hellmuth, S., Wang, G., Jiang, Z., Zhou, D., Hormozdiari, F., Liu, B., Rao, A., Hamel, A. R., Pividori, M. D., Aguet, F., Bastarache, L., Jordan, D. M., Verbanck, M., Do, R., … Im, H. K. (2021). Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biology*, *22*(1), 1–24.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., & Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, *129*(4), 823–837.

Beatrix Bartok, G. S. F. (2010). Fibroblast-like synoviocytes: key effector cells in rheumatoid arthritis. *Immunological Reviews*, *233*(1), 233.

Beecham, A. H., Patsopoulos, N. A., Xifara, D. K., Davis, M. F., Kemppinen, A., Cotsapas, C., Shah, T. S., Spencer, C., Booth, D., Goris, A., Oturai, A., Saarela, J., Fontaine, B., Hemmer, B., Martin, C., Zipp, F., D'Alfonso, S., Martinelli-Boneschi, F., Taylor, B., … International IBD Genetics Consortium (IIBDGC). (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genetics*, *45*(11), 1353–1360.

Behan, F. M., Iorio, F., Picco, G., Gonçalves, E., Beaver, C. M., Migliardi, G., Santos, R., Rao, Y., Sassi, F., Pinnelli, M., Ansari, R., Harper, S., Jackson, D. A., McRae, R., Pooley, R., Wilkinson, P., van der Meer, D., Dow, D., Buser-Doepner, C., … Garnett, M. J. (2019). Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature*, *568*(7753), 511–516.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B,*

*Statistical Methodology*, *57*(1), 289–300.

Bentham, J., Morris, D. L., Cunninghame Graham, D. S., Pinder, C. L., Tombleson, P., Behrens, T. W., Martín, J., Fairfax, B. P., Knight, J. C., Chen, L., Replogle, J., Syvänen, A.-C., Rönnblom, L., Graham, R. R., Wither, J. E., Rioux, J. D., Alarcón-Riquelme, M. E., & Vyse, T. J. (2015). Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature Genetics*, *47*(12), 1457–1464.

Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F., & Marraffini, L. A. (2013). Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Research*, *41*(15), 7429–7437.

Borsa, M., Barnstorf, I., Baumann, N. S., Pallmer, K., Yermanos, A., Gräbnitz, F., Barandun, N., Hausmann, A., Sandu, I., Barral, Y., & Oxenius, A. (2019). Modulation of asymmetric cell division as a mechanism to boost CD8+ T cell memory. *Science Immunology*, *4*(34). https://doi.org/10.1126/sciimmunol.aav1730

Bossini-Castillo, L., Glinos, D. A., Kunowska, N., Golda, G., Lamikanra, A., Spitzer, M., Soskic, B., Cano-Gamez, E., Smyth, D. J., Cattermole, C., Alasoo, K., Mann, A., Kundu, K., Soranzo, N., Dunham, I., Roberts, D., & Trynka, G. (2019). Immune disease variants modulate gene expression in regulatory CD4+ T cells and inform drug targets. In *bioRxiv* (p. 654632). https://doi.org/10.1101/654632

Bourges, C., Groff, A. F., Burren, O. S., Gerhardinger, C., Mattioli, K., Hutchinson, A., Hu, T., Anand, T., Epping, M. W., Wallace, C., Smith, K. G. C., Rinn, J. L., & Lee, J. C. (2020). Resolving mechanisms of immune-mediated disease in primary CD4 T cells. *EMBO Molecular Medicine*, *12*(5), e12112.

Bousso, P. (2008). T-cell activation by dendritic cells in the lymph node: lessons from the movies. *Nature Reviews. Immunology*, *8*(9), 675–684.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., & Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, *132*(2), 311–322.

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, *169*(7), 1177–1186.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527.

Breusch, T. S., & Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica: Journal of the Econometric Society*, *47*(5), 1287–1294.

Browning, B. L., & Browning, S. R. (2016). Genotype Imputation with Millions of Reference Samples. *American Journal of Human Genetics*, *98*(1), 116–126.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–1218.

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., & Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, *523*(7561), 486–490.

Bugatti, S., Vitolo, B., Caporali, R., Montecucco, C., & Manzo, A. (2014). B Cells in Rheumatoid Arthritis: From Pathogenic Players to Disease Biomarkers. *BioMed Research International*, *2014*. https://doi.org/10.1155/2014/681678

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M. J., Price, A. L., & Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47*(3), 291–295.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, *36*(5), 411–420.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209.

Calderon, D., Bhaskar, A., Knowles, D. A., Golan, D., Raj, T., Fu, A. Q., & Pritchard, J. K. (2017). Inferring Relevant Cell Types for Complex Traits by Using Single-Cell Gene Expression. *American Journal of Human Genetics*, *101*(5), 686–699.

Calderon, D., Nguyen, M. L. T., Mezger, A., Kathiria, A., Müller, F., Nguyen, V., Lescano, N., Wu, B., Trombetta, J., Ribado, J. V., Knowles, D. A., Gao, Z., Blaeschke, F., Parent, A. V., Burt, T. D., Anderson, M. S., Criswell, L. A., Greenleaf, W. J., Marson, A., & Pritchard, J. K. (2019). Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nature Genetics*. https://doi.org/10.1038/s41588-019-0505-9

Calderon, D., Nguyen, M. L. T., Mezger, A., Kathiria, A., Nguyen, V., Lescano, N., Wu, B., Trombetta, J., Ribado, J. V., Knowles, D. A., Gao, Z., Parent, A. V., Burt, T. D., Anderson,

M. S., Criswell, L. A., Greenleaf, W. J., Marson, A., & Pritchard, J. K. (2018). Landscape of stimulation-responsive chromatin across diverse human immune cells. In *bioRxiv* (p. 409722). https://doi.org/10.1101/409722

Cano-Gamez, E. (2017). *Characterisation of the transcriptional response to cytokine-induced polarisation in human CD4+ T cells and monocyte-derived macrophages* (G. Trynka (ed.)) [MPhil in Biomedical Science]. University of Cambridge.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C., & Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, *566*(7745), 496–502.

Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., & Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biology*, *16*(1), 195.

Castro, M., van Santen, H. M., Férez, M., Alarcón, B., Lythe, G., & Molina-París, C. (2014). Receptor pre-clustering and T cell responses: Insights into molecular mechanisms. *Frontiers in Immunology*, *5*, 132.

Champagne, P., Ogg, G. S., King, A. S., Knabenhans, C., Ellefsen, K., Nobile, M., Appay, V., Rizzardi, G. P., Fleury, S., Lipp, M., Förster, R., Rowland-Jones, S., Sékaly, R. P., McMichael, A. J., & Pantaleo, G. (2001). Skewed maturation of memory HIV-specific CD8 T lymphocytes. *Nature*, *410*(6824), 106–111.

Chatterjee, N., Shi, J., & García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews. Genetics*, *17*(7), 392–406.

Chen, L., Ge, B., Casale, F. P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., Datta, A., Richardson, D., Burden, F., Mead, D., Mann, A. L., Fernandez, J. M., Rowlston, S., Wilder, S. P., Farrow, S., … Soranzo, N. (2016). Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell*, *167*(5), 1398–1414.e24.

Chun, S., Casparino, A., Patsopoulos, N. A., Croteau-Chonka, D. C., Raby, B. A., De Jager, P. L., Sunyaev, S. R., & Cotsapas, C. (2017). Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nature Genetics*, *49*(4), 600–605.

Cohen, C. J., Crome, S. Q., MacDonald, K. G., Dai, E. L., Mager, D. L., & Levings, M. K. (2011). Human Th1 and Th17 cells exhibit epigenetic stability at signature cytokine and transcription factor loci. *Journal of Immunology* , *187*(11), 5615–5626.

Cooper, G. S., Bynum, M. L. K., & Somers, E. C. (2009). Recent insights in the epidemiology of autoimmune diseases: improved prevalence estimates and understanding of clustering of diseases. *Journal of Autoimmunity*, *33*(3-4), 197–207.

Cope, A. P., Schulze-Koops, H., & Aringer, M. (2007). The central role of T cells in rheumatoid arthritis. *Clinical and Experimental Rheumatology*, *25*(5 Suppl 46), S4–S11.

Corces, M. R., Buenrostro, J. D., Wu, B., Greenside, P. G., Chan, S. M., Koenig, J. L., Snyder, M. P., Pritchard, J. K., Kundaje, A., Greenleaf, W. J., Majeti, R., & Chang, H. Y. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics*, *48*(10), 1193–1203.

Cordell, H. J., Han, Y., Mells, G. F., Li, Y., Hirschfield, G. M., Greene, C. S., Xie, G., Juran, B. D., Zhu, D., Qian, D. C., Floyd, J. A. B., Morley, K. I., Prati, D., Lleo, A., Cusi, D., Canadian-US PBC Consortium, Italian PBC Genetics Study Group, UK-PBC Consortium, Gershwin, M. E., … Siminovitch, K. A. (2015). International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nature Communications*, *6*, 8019.

Corradin, O., Cohen, A. J., Luppino, J. M., Bayles, I. M., Schumacher, F. R., & Scacheri, P. C. (2016). Modeling disease risk through analysis of physical interactions between genetic variants within chromatin regulatory circuitry. *Nature Genetics*, *48*(11), 1313–1320.

Cox, J., & Mann, M. (2012). 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics*, *13 Suppl 16*, S12.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., & Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(50), 21931–21936.

Csárdi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, *1695*(5). https://pdfs.semanticscholar.org/1d27/44b83519657f5f2610698a8ddd177ced4f5c.pdf

Cuomo, A. S. E., Seaton, D. D., McCarthy, D. J., Martinez, I., Bonder, M. J., Garcia-Bernardo, J., Amatya, S., Madrigal, P., Isaacson, A., Buettner, F., Knights, A., Natarajan, K. N., HipSci Consortium, Vallier, L., Marioni, J. C., Chhatriwala, M., & Stegle, O. (2020). Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nature Communications*, *11*(1), 810.

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., Hayden Gephart, M. G., Barres, B. A., & Quake, S. R. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(23), 7285–7290.

Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D., & Bock, C. (2017). Pooled CRISPR screening with

single-cell transcriptome readout. *Nature Methods*, *14*(3), 297–301.

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G. E., Stephens, M., Gilad, Y., & Pritchard, J. K. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, *482*(7385), 390–394.

de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S.-G., Heap, G., Nimmo, E. R., Edwards, C., Henderson, P., Mowat, C., Sanderson, J., Satsangi, J., Simmons, A., Wilson, D. C., … Barrett, J. C. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics*, *49*(2), 256–261.

Demenais, F., Margaritte-Jeannin, P., Barnes, K. C., Cookson, W. O. C., Altmüller, J., Ang, W., Barr, R. G., Beaty, T. H., Becker, A. B., Beilby, J., Bisgaard, H., Bjornsdottir, U. S., Bleecker, E., Bønnelykke, K., Boomsma, D. I., Bouzigon, E., Brightling, C. E., Brossard, M., Brusselle, G. G., … Australian Asthma Genetics Consortium (AAGC) collaborators. (2018). Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nature Genetics*, *50*(1), 42–53.

Dendrou, C. A., Cortes, A., Shipman, L., Evans, H. G., Attfield, K. E., Jostins, L., Barber, T., Kaur, G., Kuttikkatte, S. B., Leach, O. A., Desel, C., Faergeman, S. L., Cheeseman, J., Neville, M. J., Sawcer, S., Compston, A., Johnson, A. R., Everett, C., Bell, J. I., … Fugger, L. (2016). Resolving TYK2 locus genotype-to-phenotype differences in autoimmunity. *Science Translational Medicine*, *8*(363), 363ra149.

DeWitt, M. A., Corn, J. E., & Carroll, D. (2017). Genome editing via delivery of Cas9 ribonucleoprotein. *Methods* , *121-122*, 9–15.

Diogo, D., Bastarache, L., Liao, K. P., Graham, R. R., Fulton, R. S., Greenberg, J. D., Eyre, S., Bowes, J., Cui, J., Lee, A., Pappas, D. A., Kremer, J. M., Barton, A., Coenen, M. J. H., Franke, B., Kiemeney, L. A., Mariette, X., Richard-Miceli, C., Canhão, H., … Plenge, R. M. (2015). TYK2 protein-coding variants protect against rheumatoid arthritis and autoimmunity, with no evidence of major pleiotropic effects on non-autoimmune complex traits. *PloS One*, *10*(4), e0122271.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., & Regev, A. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, *167*(7), 1853–1866.e17.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* , *29*(1), 15–21.

Dubois, P. C. A., Trynka, G., Franke, L., Hunt, K. A., Romanos, J., Curtotti, A., Zhernakova, A., Heap, G. A. R., Adány, R., Aromaa, A., Bardella, M. T., van den Berg, L. H., Bockett, N. A., de la Concha, E. G., Dema, B., Fehrmann, R. S. N., Fernández-Arquero, M., Fiatal, S., Grandone, E., … van Heel, D. A. (2010). Multiple common variants for celiac disease influencing immune gene expression. *Nature Genetics*, *42*(4), 295–302.

Ebers, G. C. (1998). Randomised double-blind placebo-controlled study of interferon β-1a in relapsing/remitting multiple sclerosis. *The Lancet*, *352*(9139), 1498–1504.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74.

Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, *9*(3), 215–216.

Eyre, S., Bowes, J., Diogo, D., Lee, A., Barton, A., Martin, P., Zhernakova, A., Stahl, E., Viatte, S., McAllister, K., Amos, C. I., Padyukov, L., Toes, R. E. M., Huizinga, T. W. J., Wijmenga, C., Trynka, G., Franke, L., Westra, H.-J., Alfredsson, L., … Worthington, J. (2012). High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature Genetics*, *44*(12), 1336–1340.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., Matthews, L., May, B., Milacic, M., Rothfels, K., Shamovsky, V., Webber, M., Weiser, J., Williams, M., Wu, G., … D'Eustachio, P. (2016). The Reactome pathway Knowledgebase. *Nucleic Acids Research*, *44*(D1), D481–D487.

Factor, D. C., Barbeau, A. M., Allan, K. C., Hu, L. R., Madhavan, M., Hoang, A. T., Hazel, K. E. A., Hall, P. A., Nisraiyya, S., Najm, F. J., Miller, T. E., Nevin, Z. S., Karl, R. T., Lima, B. R., Song, Y., Sibert, A. G., Dhillon, G. K., Volsko, C., Bartels, C. F., … Corradin, O. (2020). Cell Type-Specific Intralocus Interactions Reveal Oligodendrocyte Mechanisms in MS. *Cell*, *181*(2), 382–395.e21.

Fairfax, B. P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., & Knight, J. C. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*, *343*(6175), 1246949.

Fairfax, B. P., Taylor, C. A., Watson, R. A., Nassiri, I., Danielli, S., Fang, H., Mahé, E. A., Cooper, R., Woodcock, V., Traill, Z., Al-Mossawi, M. H., Knight, J. C., Klenerman, P., Payne, M., & Middleton, M. R. (2020). Peripheral CD8+ T cell characteristics associated with durable responses to immune checkpoint blockade in patients with metastatic melanoma. *Nature Medicine*, *26*(2), 193–199.

Fantini, M. C., Becker, C., Monteleone, G., Pallone, F., Galle, P. R., & Neurath, M. F. (2004). Cutting edge: TGF-beta induces a regulatory phenotype in CD4+CD25- T cells through Foxp3 induction and down-regulation of Smad7. *Journal of Immunology* , *172*(9), 5149–

5153.

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., De Jager, P. L., Kuchroo, V. K., Epstein, C. B., Daly, M. J., … Bernstein, B. E. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, *518*(7539), 337–343.

Feldmann, M. (2002). Development of anti-TNF therapy for rheumatoid arthritis. *Nature Reviews. Immunology*, *2*(5), 364–371.

Ferreira, M. A., Vonk, J. M., Baurecht, H., Marenholz, I., Tian, C., Hoffman, J. D., Helmer, Q., Tillander, A., Ullemar, V., van Dongen, J., Lu, Y., Rüschendorf, F., Esparza-Gordillo, J., Medway, C. W., Mountjoy, E., Burrows, K., Hummel, O., Grosche, S., Brumpton, B. M., … Paternoster, L. (2017). Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nature Genetics*, *49*(12), 1752–1757.

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., Ripke, S., Day, F. R., ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, RACI Consortium, Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R. B., … Price, A. L. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, *47*(11), 1228–1235.

Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., Shoresh, N., Genovese, G., Saunders, A., Macosko, E., Pollack, S., Brainstorm Consortium, Perry, J. R. B., Buenrostro, J. D., Bernstein, B. E., Raychaudhuri, S., … Price, A. L. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature Genetics*, *50*(4), 621–629.

Fortune, M. D., Guo, H., Burren, O., Schofield, E., Walker, N. M., Ban, M., Sawcer, S. J., Bowes, J., Worthington, J., Barton, A., Eyre, S., Todd, J. A., & Wallace, C. (2015). Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nature Genetics*, *47*(7), 839–846.

Franke, A., McGovern, D. P. B., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., Anderson, C. A., Bis, J. C., Bumpstead, S., Ellinghaus, D., Festen, E. M., Georges, M., Green, T., Haritunians, T., Jostins, L., … Parkes, M. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, *42*(12), 1118–1125.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., & Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National*

*Academy of Sciences of the United States of America*, *89*(5), 1827–1831.

Garcia-Closas, M., Rothman, N., Figueroa, J. D., Prokunina-Olsson, L., Han, S. S., Baris, D., Jacobs, E. J., Malats, N., De Vivo, I., Albanes, D., Purdue, M. P., Sharma, S., Fu, Y.-P., Kogevinas, M., Wang, Z., Tang, W., Tardón, A., Serra, C., Carrato, A., … Chatterjee, N. (2013). Common genetic polymorphisms modify the effect of smoking on absolute risk of bladder cancer. *Cancer Research*, *73*(7), 2211–2220.

Gasper, D. J., Tejera, M. M., & Suresh, M. (2014). CD4 T-cell memory generation and maintenance. *Critical Reviews in Immunology*, *34*(2), 121–146.

Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I., & Liu, D. R. (2017). Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature*, *551*(7681), 464–471.

Gaziano, J. M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., Guarino, P., Aslan, M., Anderson, D., LaFleur, R., Hammond, T., Schaa, K., Moser, J., Huang, G., Muralidhar, S., … O'Leary, T. J. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology*, *70*, 214–223.

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, *10*(5), e1004383.

Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., Pickrell, J., Jaffe, A. E., CommonMind Consortium, Pasaniuc, B., & Roussos, P. (2018). A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* , *34*(15), 2538–2545.

Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L. E., Rubelt, F., Ji, X., Han, A., Krams, S. M., Pettus, C., Haas, N., Arlehamn, C. S. L., Sette, A., Boyd, S. D., Scriba, T. J., Martinez, O. M., & Davis, M. M. (2017). Identifying specificity groups in the T cell receptor repertoire. *Nature*, *547*(7661), 94–98.

Glinos, D. A., Garborcauskas, G., Hoffman, P., Ehsan, N., Jiang, L., Gokden, A., Dai, X., Aguet, F., Brown, K. L., Garimella, K., Bowers, T., Costello, M., Ardlie, K., Jian, R., Tucker, N. R., Ellinor, P. T., Harrington, E. D., Tang, H., Snyder, M., … Cummings, B. (2021). Transcriptome variation in human tissues revealed by long-read sequencing. In *bioRxiv* (p. 2021.01.22.427687). https://doi.org/10.1101/2021.01.22.427687

Glinos, D. A., Soskic, B., Jostins, L., Sansom, D. M., & Trynka, G. (2018). Genomic profiling of T cell activation reveals dependency of memory T cells on CD28 costimulation. In *bioRxiv* (p. 421099). https://doi.org/10.1101/421099

Global Lipids Genetics Consortium, Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., Beckmann,

J. S., Bragg-Gresham, J. L., Chang, H.-Y., Demirkan, A., Den Hertog, H. M., Do, R., Donnelly, L. A., Ehret, G. B., … Abecasis, G. R. (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, *45*, 1274.

Gosselin, D., Skola, D., Coufal, N. G., Holtman, I. R., Schlachetzki, J. C. M., Sajti, E., Jaeger, B. N., O'Connor, C., Fitzpatrick, C., Pasillas, M. P., Pena, M., Adair, A., Gonda, D. D., Levy, M. L., Ransohoff, R. M., Gage, F. H., & Glass, C. K. (2017). An environment-dependent transcriptional network specifies human microglia identity. *Science*, *356*(6344). https://doi.org/10.1126/science.aal3222

Grosselin, K., Durand, A., Marsolier, J., Poitou, A., Marangoni, E., Nemati, F., Dahmani, A., Lameiras, S., Reyal, F., Frenoy, O., Pousse, Y., Reichen, M., Woolfe, A., Brenan, C., Griffiths, A. D., Vallot, C., & Gérard, A. (2019). High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nature Genetics*, *51*(6), 1060–1066.

GTEx Consortium. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585.

Guermonprez, P., Valladeau, J., Zitvogel, L., Théry, C., & Amigorena, S. (2002). Antigen presentation and T cell stimulation by dendritic cells. *Annual Review of Immunology*, *20*, 621–667.

Gutierrez-Arcelus, M., Baglaenko, Y., Arora, J., Hannes, S., Luo, Y., Amariuta, T., Teslovich, N., Rao, D. A., Ermann, J., Jonsson, A. H., NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Navarrete, C., Rich, S. S., Taylor, K. D., Rotter, J. I., Gregersen, P. K., Esko, T., Brenner, M. B., & Raychaudhuri, S. (2020). Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci. *Nature Genetics*, *52*(3), 247–253.

Gutierrez-Arcelus, M., Teslovich, N., Mola, A. R., Polidoro, R. B., Nathan, A., Kim, H., Hannes, S., Slowikowski, K., Watts, G. F. M., Korsunsky, I., Brenner, M. B., Raychaudhuri, S., & Brennan, P. J. (2019). Lymphocyte innateness defined by transcriptional states reflects a balance between proliferation and effector functions. *Nature Communications*, *10*(1), 687.

Haghverdi, L., Lun, A. T. L., Morgan, M. D., & Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, *36*(5), 421–427.

Hall, J. C., & Rosen, A. (2010). Type I interferons: crucial participants in disease amplification in autoimmunity. *Nature Reviews. Rheumatology*, *6*(1), 40–49.

Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T. M., Troakes, C., Turecki, G., O'Donovan, M. C., Schalkwyk, L. C., Bray, N. J., & Mill, J. (2016). Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nature Neuroscience*, *19*(1), 48–54.

Harrington, L. E., Hatton, R. D., Mangan, P. R., Turner, H., Murphy, T. L., Murphy, K. M., & Weaver, C. T. (2005). Interleukin 17-producing CD4+ effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages. *Nature Immunology*, *6*(11), 1123–1132.

Hawkins, E. D., Hommel, M., Turner, M. L., Battye, F. L., Markham, J. F., & Hodgkin, P. D. (2007). Measuring lymphocyte proliferation, survival and differentiation using CFSE time-series data. *Nature Protocols*, *2*(9), 2057–2067.

Henriksson, J., Chen, X., Gomes, T., Ullah, U., Meyer, K. B., Miragaia, R., Duddy, G., Pramanik, J., Yusa, K., Lahesmaa, R., & Teichmann, S. A. (2019). Genome-wide CRISPR Screens in T Helper Cells Reveal Pervasive Crosstalk between Activation and Differentiation. *Cell*, *176*(4), 882–896.e18.

He, X., Fuller, C. K., Song, Y., Meng, Q., Zhang, B., Yang, X., & Li, H. (2013). Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *American Journal of Human Genetics*, *92*(5), 667–680.

Hindorff, L. A., Gillanders, E. M., & Manolio, T. A. (2011). Genetic architecture of cancer and other complex diseases: lessons learned and future directions. *Carcinogenesis*, *32*(7), 945–954.

Hinks, A., Cobb, J., Marion, M. C., Prahalad, S., Sudman, M., Bowes, J., Martin, P., Comeau, M. E., Sajuthi, S., Andrews, R., Brown, M., Chen, W.-M., Concannon, P., Deloukas, P., Edkins, S., Eyre, S., Gaffney, P. M., Guthery, S. L., Guthridge, J. M., … United Kingdom Juvenile Idiopathic Arthritis Genetics Consortium (UKJIAGC). (2013). Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nature Genetics*, *45*(6), 664–669.

Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., & Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *American Journal of Human Genetics*, *99*(6), 1245–1260.

Hormozdiari, F., Zhu, A., Kichaev, G., -T. Ju, C. J., Segre, A. V., Joo, J. W. J., Won, H., Sankararaman, S., Pasaniuc, B., Shifman, S., & Eskin, E. (2017). Widespread Allelic Heterogeneity in Complex Traits | Elsevier Enhanced Reader. *American Journal of Human Genetics*, *100*(5), 789–802.

Huang, H., Fang, M., Jostins, L., Umićević Mirkov, M., Boucher, G., Anderson, C. A., Andersen, V., Cleynen, I., Cortes, A., Crins, F., D'Amato, M., Deffontaine, V., Dmitrieva, J., Docampo, E., Elansary, M., Farh, K. K.-H., Franke, A., Gori, A.-S., Goyette, P., … Barrett, J. C. (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature*, *547*(7662), 173–178.

Huang, Y., McCarthy, D. J., & Stegle, O. (2019). Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biology*, *20*(1), 273.

Hukku, A., Pividori, M., Luca, F., Pique-Regi, R., Im, H. K., & Wen, X. (2021). Probabilistic

colocalization of genetic variants from complex and molecular traits: promise and limitations. *American Journal of Human Genetics*, *108*(1), 25–35.

Hu, X., Kim, H., Stahl, E., Plenge, R., Daly, M., & Raychaudhuri, S. (2011). Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *American Journal of Human Genetics*, *89*(4), 496–506.

Ihry, R. J., Worringer, K. A., Salick, M. R., Frias, E., Ho, D., Theriault, K., Kommineni, S., Chen, J., Sondey, M., Ye, C., Randhawa, R., Kulkarni, T., Yang, Z., McAllister, G., Russ, C., Reece-Hoyes, J., Forrester, W., Hoffman, G. R., Dolmetsch, R., & Kaykas, A. (2018). p53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. *Nature Medicine*, *24*(7), 939–946.

Inoue, F., & Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. *Genomics*, *106*(3), 159–164.

International Genetics of Ankylosing Spondylitis Consortium (IGAS), Cortes, A., Hadler, J., Pointon, J. P., Robinson, P. C., Karaderi, T., Leo, P., Cremin, K., Pryce, K., Harris, J., Lee, S., Joo, K. B., Shim, S.-C., Weisman, M., Ward, M., Zhou, X., Garchon, H.-J., Chiocchia, G., Nossent, J., … Brown, M. A. (2013). Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nature Genetics*, *45*(7), 730–738.

Iotchkova, V., Ritchie, G. R. S., Geihs, M., Morganella, S., Min, J. L., Walter, K., Timpson, N. J., UK10K Consortium, Dunham, I., Birney, E., & Soranzo, N. (2019). GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nature Genetics*, *51*(2), 343–353.

Jamshidian, A., Shaygannejad, V., Pourazar, A., Zarkesh-Esfahani, S.-H., & Gharagozloo, M. (2013). Biased Treg/Th17 balance away from regulatory toward inflammatory phenotype in relapsed multiple sclerosis and its correlation with severity of symptoms. *Journal of Neuroimmunology*, *262*(1-2), 106–112.

Jerber, J., Seaton, D. D., Cuomo, A. S. E., Kumasaka, N., Haldane, J., Steer, J., Patel, M., Pearce, D., Andersson, M., Bonder, M. J., Mountjoy, E., Ghoussaini, M., Lancaster, M. A., HipSci Consortium, Marioni, J. C., Merkle, F. T., Stegle, O., & Gaffney, D. J. (2021). Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nature Genetics*, *53*, 304–312.

Jiang, H., Lei, R., Ding, S.-W., & Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, *15*, 182.

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Schumm, L. P., Sharma, Y., Anderson, C. A., Essers, J., Mitrovic, M., Ning, K., Cleynen, I., Theatre, E., Spain, S. L., Raychaudhuri, S., Goyette, P., Wei, Z., … Cho, J. H. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel

disease. *Nature*, *491*(7422), 119–124.

Kaech, S. M., John Wherry, E., & Ahmed, R. (2002). Effector and memory T-cell differentiation: implications for vaccine development. *Nature Reviews. Immunology*, *2*(4), 251–262.

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30.

Kanhere, A., Hertweck, A., Bhatia, U., Gökmen, M. R., Perucha, E., Jackson, I., Lord, G. M., & Jenner, R. G. (2012). T-bet and GATA3 orchestrate Th1 and Th2 differentiation through lineage-specific targeting of distal regulatory elements. *Nature Communications*, *3*, 1268.

Kanno, Y., Vahedi, G., Hirahara, K., Singleton, K., & O'Shea, J. J. (2012). Transcriptional and Epigenetic Control of T Helper Cell Specification: Molecular Mechanisms Underlying Commitment and Plasticity. *Annual Review of Immunology*, *30*(1), 707–731.

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., & Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, *50*(9), 1219–1224.

Khera, A. V., Chaffin, M., Wade, K. H., Zahid, S., Brancale, J., Xia, R., Distefano, M., Senol-Cosar, O., Haas, M. E., Bick, A., Aragam, K. G., Lander, E. S., Smith, G. D., Mason-Suares, H., Fornage, M., Lebo, M., Timpson, N. J., Kaplan, L. M., & Kathiresan, S. (2019). Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell*, *177*(3), 587–596.e9.

Khera, A. V., Emdin, C. A., Drake, I., Natarajan, P., Bick, A. G., Cook, N. R., Chasman, D. I., Baber, U., Mehran, R., Rader, D. J., Fuster, V., Boerwinkle, E., Melander, O., Orho-Melander, M., Ridker, P. M., & Kathiresan, S. (2016). Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *The New England Journal of Medicine*, *375*(24), 2349–2358.

Khor, B., Gardet, A., & Xavier, R. J. (2011). Genetics and pathogenesis of inflammatory bowel disease. *Nature*, *474*(7351), 307–317.

Kim-Hellmuth, S., Aguet, F., Oliva, M., Muñoz-Aguirre, M., Kasela, S., Wucher, V., Castel, S. E., Hamel, A. R., Viñuela, A., Roberts, A. L., Mangul, S., Wen, X., Wang, G., Barbeira, A. N., Garrido-Martín, D., Nadel, B. B., Zou, Y., Bonazzola, R., Quan, J., … Lappalainen, T. (2020). Cell type–specific genetic regulation of gene expression across human tissues. *Science*, *369*(6509). https://doi.org/10.1126/science.aaz8528

Kimmerling, R. J., Lee Szeto, G., Li, J. W., Genshaft, A. S., Kazer, S. W., Payer, K. R., de Riba Borrajo, J., Blainey, P. C., Irvine, D. J., Shalek, A. K., & Manalis, S. R. (2016). A microfluidic platform enabling single-cell RNA-seq of multigenerational lineages. *Nature Communications*, *7*, 10220.

Kiner, E., Willie, E., Vijaykumar, B., Chowdhary, K., Schmutz, H., Chandler, J., Schnell, A.,

Thakore, P. I., LeGros, G., Mostafavi, S., Mathis, D., Benoist, C., & Immunological Genome Project Consortium. (2021). Gut CD4+ T cell phenotypes are a continuum molded by microbes, not by TH archetypes. *Nature Immunology*, *22*(2), 216–228.

Kinney, J. B., & McCandlish, D. M. (2019). Massively Parallel Assays and Quantitative Sequence-Function Relationships. *Annual Review of Genomics and Human Genetics*, *20*, 99–127.

Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R. J. A., Costello, J. F., Shendure, J., & Ahituv, N. (2019). Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nature Communications*, *10*(1), 3583.

Koguchi, Y., Buenafe, A. C., Thauland, T. J., Gardell, J. L., Bivins-Smith, E. R., Jacoby, D. B., Slifka, M. K., & Parker, D. C. (2012). Preformed CD40L is stored in Th1, Th2, Th17, and T follicular helper cells as well as CD4+ 8- thymocytes and invariant NKT cells but not in Treg cells. *PloS One*, *7*(2), e31296.

Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M. D. C., & Yusa, K. (2014). Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature Biotechnology*, *32*(3), 267–273.

Komatsu, N., Okamoto, K., Sawa, S., Nakashima, T., Oh-hora, M., Kodama, T., Tanaka, S., Bluestone, J. A., & Takayanagi, H. (2014). Pathogenic conversion of Foxp3+ T cells into TH17 cells in autoimmune arthritis. *Nature Medicine*, *20*(1), 62–68.

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., & Liu, D. R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, *533*(7603), 420–424.

Kowalczyk, M. S., Tirosh, I., Heckl, D., Rao, T. N., Dixit, A., Haas, B. J., Schneider, R. K., Wagers, A. J., Ebert, B. L., & Regev, A. (2015). Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Research*, *25*(12), 1860–1872.

Kreins, A. Y., Ciancanelli, M. J., Okada, S., Kong, X.-F., Ramírez-Alejo, N., Kilic, S. S., El Baghdadi, J., Nonoyama, S., Mahdaviani, S. A., Ailal, F., Bousfiha, A., Mansouri, D., Nievas, E., Ma, C. S., Rao, G., Bernasconi, A., Sun Kuehn, H., Niemela, J., Stoddard, J., … Boisson-Dupuis, S. (2015). Human TYK2 deficiency: Mycobacterial and viral infections without hyper-IgE syndrome. *The Journal of Experimental Medicine*, *212*(10), 1641–1662.

Kuehn, H. S., Ouyang, W., Lo, B., Deenick, E. K., Niemela, J. E., Avery, D. T., Schickel, J.-N., Tran, D. Q., Stoddard, J., Zhang, Y., Frucht, D. M., Dumitriu, B., Scheinberg, P., Folio, L. R., Frein, C. A., Price, S., Koh, C., Heller, T., Seroogy, C. M., … Uzel, G. (2014). Immune dysregulation in human subjects with heterozygous germline mutations in CTLA4. *Science*, *345*(6204), 1623–1627.

Kumar, R., Ferez, M., Swamy, M., Arechaga, I., Rejas, M. T., Valpuesta, J. M., Schamel, W.

W. A., Alarcon, B., & van Santen, H. M. (2011). Increased sensitivity of antigen-experienced T cells through the enrichment of oligomeric T cell receptor complexes. *Immunity*, *35*(3), 375–387.

Kumasaka, N., Knights, A. J., & Gaffney, D. J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature Genetics*, *48*(2), 206–213.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. de, Cappuccio, A., … Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, *21*(1), 31.

Lalive, P. H., Kreutzfeldt, M., Devergne, O., Metz, I., Bruck, W., Merkler, D., & Pot, C. (2017). Increased interleukin-27 cytokine expression in the central nervous system of multiple sclerosis patients. *Journal of Neuroinflammation*, *14*(1), 144.

Landry, J.-R., Mager, D. L., & Wilhelm, B. T. (2003). Complex controls: the role of alternative promoters in mammalian genomes. *Trends in Genetics: TIG*, *19*(11), 640–648.

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

Lappalainen, T., Sammeth, M., Friedländer, M. R., ’t Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., … Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, *501*(7468), 506–511.

Le Gros, G., Ben-Sasson, S. Z., Seder, R., Finkelman, F. D., & Paul, W. E. (2008). Generation of interleukin 4 (IL-4)-producing cells in vivo and in vitro: IL-2 and IL-4 are required for in vitro generation of IL-4-producing cells. *Journal of Immunology* , *181*(5), 2943–2951.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O’Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., … Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291.

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* , *30*(7), 923–930.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* , *26*(5), 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* , *25*(16), 2078–2079.

Li, L., Huang, K.-L., Gao, Y., Cui, Y., Wang, G., Elrod, N. D., Li, Y., Chen, Y. E., Ji, P., Peng, F., Russell, W. K., Wagner, E. J., & Li, W. (2021). An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nature Genetics*, 1–12.

Lin, C.-C., & Edelson, B. T. (2017). New Insights into the Role of IL-1β in Experimental Autoimmune Encephalomyelitis and Multiple Sclerosis. *Journal of Immunology* , *198*(12), 4553–4560.

Li, N., van Unen, V., Abdelaal, T., Guo, N., Kasatskaya, S. A., Ladell, K., McLaren, J. E., Egorov, E. S., Izraelson, M., Chuva de Sousa Lopes, S. M., Höllt, T., Britanova, O. V., Eggermont, J., de Miranda, N. F. C. C., Chudakov, D. M., Price, D. A., Lelieveldt, B. P. F., & Koning, F. (2019). Memory CD4+ T cells are generated in the human fetal intestine. *Nature Immunology*, *20*(3), 301–312.

Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E., & Montgomery, S. B. (2019). Abundant associations with gene expression complicate GWAS follow-up. *Nature Genetics*, *51*(5), 768–769.

Liu, Q., Herring, C. A., Sheng, Q., Ping, J., Simmons, A. J., Chen, B., Banerjee, A., Li, W., Gu, G., Coffey, R. J., Shyr, Y., & Lau, K. S. (2018). Quantitative assessment of cell population diversity in single-cell landscapes. *PLoS Biology*, *16*(10), e2006687.

Liu, X., Li, Y. I., & Pritchard, J. K. (2019). Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell*, *177*(4), 1022–1034.e6.

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., & Pritchard, J. K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*, *50*(1), 151–158.

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y., & Pritchard, J. K. (2016). RNA splicing is a primary link between genetic variation and disease. *Science*, *352*(6285), 600–604.

Li, Y., Oosting, M., Smeekens, S. P., Jaeger, M., Aguirre-Gamboa, R., Le, K. T. T., Deelen, P., Ricaño-Ponce, I., Schoffelen, T., Jansen, A. F. M., Swertz, M. A., Withoff, S., van de Vosse, E., van Deuren, M., van de Veerdonk, F., Zhernakova, A., van der Meer, J. W. M., Xavier, R. J., Franke, L., … Netea, M. G. (2016). A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans. *Cell*, *167*(4), 1099–1110.e14.

Lo, B., Zhang, K., Lu, W., Zheng, L., Zhang, Q., Kanellopoulou, C., Zhang, Y., Liu, Z., Fritz, J. M., Marsh, R., Husami, A., Kissell, D., Nortman, S., Chaturvedi, V., Haines, H., Young, L. R., Mo, J., Filipovich, A. H., Bleesing, J. J., … Jordan, M. B. (2015). AUTOIMMUNE DISEASE. Patients with LRBA deficiency show CTLA4 loss and immune dysregulation responsive to abatacept therapy. *Science*, *349*(6246), 436–440.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and

dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Lozupone, C., & Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, *71*(12), 8228–8235.

Luckheeram, R. V., Zhou, R., Verma, A. D., & Xia, B. (2012). $CD4^+$T cells: differentiation and functions. *Clinical & Developmental Immunology*, *2012*, 925135.

Maas, P., Barrdahl, M., Joshi, A. D., Auer, P. L., Gaudet, M. M., Milne, R. L., Schumacher, F. R., Anderson, W. F., Check, D., Chattopadhyay, S., Baglietto, L., Berg, C. D., Chanock, S. J., Cox, D. G., Figueroa, J. D., Gail, M. H., Graubard, B. I., Haiman, C. A., Hankinson, S. E., … Chatterjee, N. (2016). Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. *JAMA Oncology*, *2*(10), 1295–1302.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., & Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, *45*(D1), D896–D901.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214.

Martinez, F. O., & Gordon, S. (2014). The M1 and M2 paradigm of macrophage activation: time for reassessment. *F1000prime Reports*, *6*, 13.

Mascanfroni, I. D., Yeste, A., Vieira, S. M., Burns, E. J., Patel, B., Sloma, I., Wu, Y., Mayo, L., Ben-Hamo, R., Efroni, S., Kuchroo, V. K., Robson, S. C., & Quintana, F. J. (2013). IL-27 acts on DCs to suppress the T cell response and autoimmunity by inducing expression of the immunoregulatory molecule CD39. *Nature Immunology*, *14*(10), 1054–1063.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., … Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, *337*(6099), 1190–1195.

Mavaddat, N., Pharoah, P. D. P., Michailidou, K., Tyrer, J., Brook, M. N., Bolla, M. K., Wang, Q., Dennis, J., Dunning, A. M., Shah, M., Luben, R., Brown, J., Bojesen, S. E., Nordestgaard, B. G., Nielsen, S. F., Flyger, H., Czene, K., Darabi, H., Eriksson, M., … Garcia-Closas, M. (2015). Prediction of breast cancer risk based on profiling with common genetic variants. *Journal of the National Cancer Institute*, *107*(5). https://doi.org/10.1093/jnci/djv036

McCullum, E. O., Williams, B. A. R., Zhang, J., & Chaput, J. C. (2010). Random Mutagenesis by Error-Prone PCR. In J. Braman (Ed.), *In Vitro Mutagenesis Protocols: Third Edition* (pp. 103–109). Humana Press.

McInnes, I. B., Buckley, C. D., & Isaacs, J. D. (2016). Cytokines in rheumatoid arthritis - shaping the immunological landscape. *Nature Reviews. Rheumatology*, *12*(1), 63–68.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. In *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1802.03426

Mega, J. L., Stitziel, N. O., Smith, J. G., Chasman, D. I., Caulfield, M., Devlin, J. J., Nordio, F., Hyde, C., Cannon, C. P., Sacks, F., Poulter, N., Sever, P., Ridker, P. M., Braunwald, E., Melander, O., Kathiresan, S., & Sabatine, M. S. (2015). Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *The Lancet*, *385*(9984), 2264–2271.

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segrè, A. V., Djebali, S., Niarchou, A., GTEx Consortium, Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., … Guigó, R. (2015). Human genomics. The human transcriptome across tissues and individuals. *Science*, *348*(6235), 660–665.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C. G., Jr, Kinney, J. B., Kellis, M., Lander, E. S., & Mikkelsen, T. S. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, *30*(3), 271–277.

Melzer, D., Perry, J. R. B., Hernandez, D., Corsi, A.-M., Stevens, K., Rafferty, I., Lauretani, F., Murray, A., Gibbs, J. R., Paolisso, G., Rafiq, S., Simon-Sanchez, J., Lango, H., Scholz, S., Weedon, M. N., Arepalli, S., Rice, N., Washecka, N., Hurst, A., … Ferrucci, L. (2008). A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genetics*, *4*(5), e1000072.

Menk, A. V., Scharping, N. E., Moreci, R. S., Zeng, X., Guy, C., Salvatore, S., Bae, H., Xie, J., Young, H. A., Wendell, S. G., & Delgoffe, G. M. (2018). Early TCR Signaling Induces Rapid Aerobic Glycolysis Enabling Distinct Acute T Cell Effector Functions. *Cell Reports*, *22*(6), 1509–1521.

Mestas, J., & Hughes, C. C. W. (2004). Of mice and not men: differences between mouse and human immunology. *Journal of Immunology* , *172*(5), 2731–2738.

Minoux, M., Holwerda, S., Vitobello, A., Kitazawa, T., Kohler, H., Stadler, M. B., & Rijli, F. M. (2017). Gene bivalency at Polycomb domains regulates cranial neural crest positional identity. *Science*, *355*(6332). https://doi.org/10.1126/science.aal2913

Mittleman, B. E., Pott, S., Warland, S., Zeng, T., Mu, Z., Kaur, M., Gilad, Y., & Li, Y. (2020).

Alternative polyadenylation mediates genetic regulation of gene expression. *eLife*, *9*, e57492.

Monlong, J., Calvo, M., Ferreira, P. G., & Guigó, R. (2014). Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nature Communications*, *5*, 4698.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1-2), 17–23.

Mosmann, T. R., Cherwinski, H., Bond, M. W., Giedlin, M. A., & Coffman, R. L. (1986). Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins. *Journal of Immunology* , *136*(7), 2348–2357.

Mousavi, S. F., Soroosh, P., Takahashi, T., Yoshikai, Y., Shen, H., Lefrançois, L., Borst, J., Sugamura, K., & Ishii, N. (2008). OX40 costimulatory signals potentiate the memory commitment of effector CD8+ T cells. *Journal of Immunology* , *181*(9), 5990–6001.

Mueller, B., Mieczkowski, J., Kundu, S., Wang, P., Sadreyev, R., Tolstorukov, M. Y., & Kingston, R. E. (2017). Widespread changes in nucleosome accessibility without changes in nucleosome occupancy during a rapid transcriptional induction. *Genes & Development*, *31*(5), 451–462.

Mueller, S. N., Gebhardt, T., Carbone, F. R., & Heath, W. R. (2013). Memory T cell subsets, migration patterns, and tissue residence. *Annual Review of Immunology*, *31*, 137–161.

Muerdter, F., Boryń, Ł. M., & Arnold, C. D. (2015). STARR-seq - principles and applications. *Genomics*, *106*(3), 145–150.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., Li, X., Li, H., Kuperwasser, N., Ruda, V. M., Pirruccello, J. P., Muchmore, B., Prokunina-Olsson, L., Hall, J. L., Schadt, E. E., Morales, C. R., Lund-Katz, S., Phillips, M. C., Wong, J., … Rader, D. J. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, *466*(7307), 714–719.

Myocardial Infarction Genetics Consortium, Kathiresan, S., Voight, B. F., Purcell, S., Musunuru, K., Ardissino, D., Mannucci, P. M., Anand, S., Engert, J. C., Samani, N. J., Schunkert, H., Erdmann, J., Reilly, M. P., Rader, D. J., Morgan, T., Spertus, J. A., Stoll, M., Girelli, D., McKeown, P. P., … Altshuler, D. (2009). Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature Genetics*, *41*(3), 334–341.

Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., Murakami, Y., Yuji, K., Furukawa, Y., Zembutsu, H., Tanaka, T., Ohnishi, Y., Nakamura, Y., BioBank Japan Cooperative Hospital Group, & Kubo, M. (2017). Overview of the BioBank Japan Project: Study design and profile. *Journal of Epidemiology / Japan Epidemiological Association*, *27*(3S), S2–S8.

Neurath, M. F., Weigmann, B., Finotto, S., Glickman, J., Nieuwenhuis, E., Iijima, H.,

Mizoguchi, A., Mizoguchi, E., Mudter, J., Galle, P. R., Bhan, A., Autschbach, F., Sullivan, B. M., Szabo, S. J., Glimcher, L. H., & Blumberg, R. S. (2002). The transcription factor T-bet regulates mucosal T cell activation in experimental colitis and Crohn's disease. *The Journal of Experimental Medicine*, *195*(9), 1129–1143.

Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *368*(1620), 20120362.

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., & Dermitzakis, E. T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics*, *6*(4), e1000895.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genetics*, *6*(4), e1000888.

Nistala, K., Adams, S., Cambrook, H., Ursu, S., Olivito, B., de Jager, W., Evans, J. G., Cimaz, R., Bajaj-Elliott, M., & Wedderburn, L. R. (2010). Th17 plasticity in human autoimmune arthritis is driven by the inflammatory environment. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(33), 14751–14756.

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., Graham, R. R., Manoharan, A., Ortmann, W., Bhangale, T., Denny, J. C., Carroll, R. J., Eyler, A. E., Greenberg, J. D., Kremer, J. M., … Plenge, R. M. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, *506*(7488), 376–381.

Omenetti, S., Bussi, C., Metidji, A., Iseppon, A., Lee, S., Tolaini, M., Li, Y., Kelly, G., Chakravarty, P., Shoaie, S., Gutierrez, M. G., & Stockinger, B. (2019). The Intestine Harbors Functionally Distinct Homeostatic Tissue-Resident and Inflammatory Th17 Cells. *Immunity*, *51*(1), 77–89.e6.

Onengut-Gumuscu, S., Chen, W.-M., Burren, O., Cooper, N. J., Quinlan, A. R., Mychaleckyj, J. C., Farber, E., Bonnie, J. K., Szpak, M., Schofield, E., Achuthan, P., Guo, H., Fortune, M. D., Stevens, H., Walker, N. M., Ward, L. D., Kundaje, A., Kellis, M., Daly, M. J., … Rich, S. S. (2015). Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nature Genetics*, *47*(4), 381–386.

Ongen, H., & Dermitzakis, E. T. (2015). Alternative Splicing QTLs in European and African Populations. *American Journal of Human Genetics*, *97*(4), 567–575.

O'Shea, J. J., & Paul, W. E. (2010). Mechanisms underlying lineage commitment and plasticity of helper CD4+ T cells. *Science*, *327*(5969), 1098–1102.

Pai, A. A., Cain, C. E., Mizrahi-Man, O., De Leon, S., Lewellen, N., Veyrieras, J.-B., Degner,

J. F., Gaffney, D. J., Pickrell, J. K., Stephens, M., Pritchard, J. K., & Gilad, Y. (2012). The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLoS Genetics*, *8*(10), e1003000.

Park, H., Li, Z., Yang, X. O., Chang, S. H., Nurieva, R., Wang, Y.-H., Wang, Y., Hood, L., Zhu, Z., Tian, Q., & Dong, C. (2005). A distinct lineage of CD4 T cells regulates tissue inflammation by producing interleukin 17. *Nature Immunology*, *6*(11), 1133–1141.

Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D. P., Patterson, N., & Price, A. L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* , *30*(20), 2906.

Pasquali, L., Gaulton, K. J., Rodríguez-Seguí, S. A., Mularoni, L., Miguel-Escalada, I., Akerman, İ., Tena, J. J., Morán, I., Gómez-Marín, C., van de Bunt, M., Ponsa-Cobas, J., Castro, N., Nammo, T., Cebola, I., García-Hurtado, J., Maestro, M. A., Pattou, F., Piemonti, L., Berney, T., … Ferrer, J. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature Genetics*, *46*(2), 136–143.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419.

Paty, D. W., & Li, D. K. (1993). Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. II. MRI analysis results of a multicenter, randomized, double-blind, placebo-controlled trial. UBC MS/MRI Study Group and the IFNB Multiple Sclerosis Study Group. *Neurology*, *43*(4), 662–667.

Pelikan, R. C., Kelly, J. A., Fu, Y., Lareau, C. A., Tessneer, K. L., Wiley, G. B., Wiley, M. M., Glenn, S. B., Harley, J. B., Guthridge, J. M., James, J. A., Aryee, M. J., Montgomery, C., & Gaffney, P. M. (2018). Enhancer histone-QTLs are enriched on autoimmune risk haplotypes and influence gene expression within chromatin networks. *Nature Communications*, *9*(1), 2905.

Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, *10*(11), 1096–1098.

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American Journal of Human Genetics*, *94*(4), 559–573.

Plagnol, V., Smyth, D. J., Todd, J. A., & Clayton, D. G. (2009). Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. *Biostatistics* , *10*(2), 327–334.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575.

Purvis, H. A., Stoop, J. N., Mann, J., Woods, S., Kozijn, A. E., Hambleton, S., Robinson, J. H., Isaacs, J. D., Anderson, A. E., & Hilkens, C. M. U. (2010). Low-strength T-cell activation promotes Th17 responses. *Blood*, *116*(23), 4829–4837.

Putheti, P., Awasthi, A., Popoola, J., Gao, W., & Strom, T. B. (2010). Human CD4 memory T cells can become CD4+IL-9+ T cells. *PloS One*, *5*(1), e8706.

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., & Lim, W. A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, *152*(5), 1173–1183.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., & Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, *14*(10), 979–982.

Qureshi, O. S., Zheng, Y., Nakamura, K., Attridge, K., Manzotti, C., Schmidt, E. M., Baker, J., Jeffery, L. E., Kaur, S., Briggs, Z., Hou, T. Z., Futter, C. E., Anderson, G., Walker, L. S. K., & Sansom, D. M. (2011). Trans-endocytosis of CD80 and CD86: a molecular basis for the cell-extrinsic function of CTLA-4. *Science*, *332*(6029), 600–603.

Raj, B., & Blencowe, B. J. (2015). Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. *Neuron*, *87*(1), 14–27.

Randall, R. E., & Goodbourn, S. (2008). Interferons and viruses: an interplay between induction, signalling, antiviral responses and virus countermeasures. *The Journal of General Virology*, *89*(Pt 1), 1–47.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., … Human Cell Atlas Meeting Participants. (2017). The Human Cell Atlas. *eLife*, *6*. https://doi.org/10.7554/eLife.27041

Richardson, T. G., Harrison, S., Hemani, G., & Davey Smith, G. (2019). An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *eLife*, *8*. https://doi.org/10.7554/eLife.43657

Rieckmann, J. C., Geiger, R., Hornburg, D., Wolf, T., Kveler, K., Jarrossay, D., Sallusto, F., Shen-Orr, S. S., Lanzavecchia, A., Mann, M., & Meissner, F. (2017). Social network architecture of human immune cells unveiled by quantitative proteomics. *Nature Immunology*, *18*(5), 583–593.

Ripatti, S., Tikkanen, E., Orho-Melander, M., Havulinna, A. S., Silander, K., Sharma, A., Guiducci, C., Perola, M., Jula, A., Sinisalo, J., Lokki, M.-L., Nieminen, M. S., Melander, O., Salomaa, V., Peltonen, L., & Kathiresan, S. (2010). A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *The Lancet*, *376*(9750), 1393–1400.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330.

Rosati, E., Dowds, C. M., Liaskou, E., Henriksen, E. K. K., Karlsen, T. H., & Franke, A. (2017). Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnology*, *17*(1), 61.

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., & Bernstein, B. E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, *33*(11), 1165–1172.

Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O. N., Stümpflen, V., & Mewes, H. W. (2008). CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Research*, *36*(Database issue), D646–D650.

Sadler, A. J., & Williams, B. R. G. (2008). Interferon-inducible antiviral effectors. *Nature Reviews. Immunology*, *8*(7), 559–568.

Saelens, W., Cannoodt, R., Todorov, H., & Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, *37*(5), 547–554.

Sallusto, F. (2016). Heterogeneity of Human CD4(+) T Cells Against Microbes. *Annual Review of Immunology*, *34*, 317–334.

Sallusto, F., Geginat, J., & Lanzavecchia, A. (2004). Central memory and effector memory T cell subsets: function, generation, and maintenance. *Annual Review of Immunology*, *22*, 745–763.

Sallusto, F., Lenig, D., Förster, R., Lipp, M., & Lanzavecchia, A. (1999). Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. *Nature*, *401*(6754), 708–712.

Samstein, R. M., Arvey, A., Josefowicz, S. Z., Peng, X., Reynolds, A., Sandstrom, R., Neph, S., Sabo, P., Kim, J. M., Liao, W., Li, M. O., Leslie, C., Stamatoyannopoulos, J. A., & Rudensky, A. Y. (2012). Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell*, *151*(1), 153–166.

Saule, P., Trauet, J., Dutriez, V., Lekeux, V., Dessaint, J.-P., & Labalette, M. (2006). Accumulation of memory T cells from childhood to old age: central and effector memory cells in CD4(+) versus effector memory and terminally differentiated memory cells in

CD8(+) compartment. *Mechanisms of Ageing and Development*, *127*(3), 274–281.

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., Zhu, J., Millstein, J., Sieberts, S., Lamb, J., GuhaThakurta, D., Derry, J., Storey, J. D., Avila-Campillo, I., Kruger, M. J., … Ulrich, R. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biology*, *6*(5), e107.

Schett, G., Elewaut, D., McInnes, I. B., Dayer, J.-M., & Neurath, M. F. (2013). How cytokine networks fuel inflammation: Toward a cytokine-based disease taxonomy. *Nature Medicine*, *19*(7), 822–824.

Schmidl, C., Rendeiro, A. F., Sheffield, N. C., & Bock, C. (2015). ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nature Methods*, *12*(10), 963–965.

Schmidt, E. M., Zhang, J., Zhou, W., Chen, J., Mohlke, K. L., Chen, Y. E., & Willer, C. J. (2015). GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* , *31*(16), 2601–2606.

Schmiedel, B. J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A. G., White, B. M., Zapardiel-Gonzalo, J., Ha, B., Altay, G., Greenbaum, J. A., McVicker, G., Seumois, G., Rao, A., Kronenberg, M., Peters, B., & Vijayanand, P. (2018). Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell*, *175*(6), 1701–1715.e16.

Schubert, D., Bode, C., Kenefeck, R., Hou, T. Z., Wing, J. B., Kennedy, A., Bulashevska, A., Petersen, B.-S., Schäffer, A. A., Grüning, B. A., Unger, S., Frede, N., Baumann, U., Witte, T., Schmidt, R. E., Dueckers, G., Niehues, T., Seneviratne, S., Kanariou, M., … Grimbacher, B. (2014). Autosomal dominant immune dysregulation syndrome in humans with CTLA4 mutations. *Nature Medicine*, *20*(12), 1410–1416.

Sénécal, V., Deblois, G., Beauseigle, D., Schneider, R., Brandenburg, J., Newcombe, J., Moore, C. S., Prat, A., Antel, J., & Arbour, N. (2016). Production of IL-27 in multiple sclerosis lesions by astrocytes and myeloid cells: Modulation of local immune responses. *Glia*, *64*(4), 553–569.

Sewell, A. K. (2012). Why must T cells be cross-reactive? *Nature Reviews. Immunology*, *12*(9), 669–677.

Sharma, G., Rive, C. M., & Holt, R. A. (2019). Rapid selection and identification of functional CD8+ T cell epitopes from large peptide-coding libraries. *Nature Communications*, *10*(1), 4553.

Shifrut, E., Carnevale, J., Tobin, V., Roth, T. L., Woo, J. M., Bui, C. T., Li, P. J., Diolaiti, M. E., Ashworth, A., & Marson, A. (2018). Genome-wide CRISPR Screens in Primary Human T Cells Reveal Key Regulators of Immune Function. *Cell*, *175*(7), 1958–1971.e15.

Sinnott-Armstrong, N., Naqvi, S., Rivas, M., & Pritchard, J. K. (2020). GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic

background. In *bioRxiv* (p. 2020.04.20.051631). https://doi.org/10.1101/2020.04.20.051631

Sinnott-Armstrong, N., Sousa, I. S., Laber, S., Rendina-Ruedy, E., Nitter Dankel, S. E., Ferreira, T., Mellgren, G., Karasik, D., Rivas, M., Pritchard, J., Guntur, A. R., Cox, R. D., Lindgren, C. M., Hauner, H., Sallari, R., Rosen, C. J., Hsu, Y.-H., Lander, E. S., Kiel, D. P., & Claussnitzer, M. (2021). A regulatory variant at 3q21.1 confers an increased pleiotropic risk for hyperglycemia and altered bone mineral density. *Cell Metabolism*. https://doi.org/10.1016/j.cmet.2021.01.001

Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nature Reviews. Genetics*, *9*(6), 477–485.

Slowikowski, K., Hu, X., & Raychaudhuri, S. (2014). SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. *Bioinformatics* , *30*(17), 2496–2497.

Smith-Garvin, J. E., Koretzky, G. A., & Jordan, M. S. (2009). T cell activation. *Annual Review of Immunology*, *27*, 591–619.

Smith, G. D., Ebrahim, S., Lewis, S., Hansell, A. L., Palmer, L. J., & Burton, P. R. (2005). Genetic epidemiology and public health: hope, hype, and future prospects. *The Lancet*, *366*(9495), 1484–1498.

Soon, M. S. F., Lee, H. J., Engel, J. A., Straube, J., Thomas, B. S., Pernold, C. P. S., Clarke, L. S., Laohamonthonkul, P., Haldar, R. N., Williams, C. G., Lansink, L. I. M., Moreira, M. L., Bramhall, M., Koufariotis, L. T., Wood, S., Chen, X., James, K. R., Lönnberg, T., Lane, S. W., … Haque, A. (2020). Transcriptome dynamics of CD4 + T cells during malaria maps gradual transit from effector to memory. *Nature Immunology*, *21*(12), 1597–1610.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics*, *31*(6), 2013–2035.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. In *Proceedings of the National Academy of Sciences* (Vol. 100, Issue 16, pp. 9440–9445). https://doi.org/10.1073/pnas.1530509100

Stray-Pedersen, A., Sorte, H. S., Samarakoon, P., Gambin, T., Chinn, I. K., Coban Akdemir, Z. H., Erichsen, H. C., Forbes, L. R., Gu, S., Yuan, B., Jhangiani, S. N., Muzny, D. M., Rødningen, O. K., Sheng, Y., Nicholas, S. K., Noroski, L. M., Seeborg, F. O., Davis, C. M., Canter, D. L., … Lupski, J. R. (2017). Primary immunodeficiency diseases: Genomic approaches delineate heterogeneous Mendelian disorders. *The Journal of Allergy and Clinical Immunology*, *139*(1), 232–245.

Strober, B. J., Elorbany, R., Rhodes, K., Krishnan, N., Tayeb, K., Battle, A., & Gilad, Y. (2019). Dynamic genetic regulation of gene expression during cellular differentiation. *Science*, *364*(6447), 1287–1290.

Stubbington, M. J. T., Lönnberg, T., Proserpio, V., Clare, S., Speak, A. O., Dougan, G., &

Teichmann, S. A. (2016). T cell fate and clonality inference from single-cell transcriptomes. *Nature Methods*, *13*(4), 329–332.

Sun, B. B., Maranville, J. C., Peters, J. E., Stacey, D., Staley, J. R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., Oliver-Williams, C., Kamat, M. A., Prins, B. P., Wilcox, S. K., Zimmerman, E. S., Chi, A., Bansal, N., Spain, S. L., Wood, A. M., … Butterworth, A. S. (2018). Genomic atlas of the human plasma proteome. *Nature*, *558*(7708), 73–79.

Sun, W., Poschmann, J., Cruz-Herrera Del Rosario, R., Parikshak, N. N., Hajan, H. S., Kumar, V., Ramasamy, R., Belgard, T. G., Elanggovan, B., Wong, C. C. Y., Mill, J., Geschwind, D. H., & Prabhakar, S. (2016). Histone Acetylome-wide Association Study of Autism Spectrum Disorder. *Cell*, *167*(5), 1385–1397.e11.

Szabo, P. A., Levitin, H. M., Miron, M., Snyder, M. E., Senda, T., Yuan, J., Cheng, Y. L., Bush, E. C., Dogra, P., Thapa, P., Farber, D. L., & Sims, P. A. (2019). Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nature Communications*, *10*(1), 4706.

Szabo, S. J., Kim, S. T., Costa, G. L., Zhang, X., Fathman, C. G., & Glimcher, L. H. (2000). A novel transcription factor, T-bet, directs Th1 lineage commitment. *Cell*, *100*(6), 655–669.

Takata, A., Matsumoto, N., & Kato, T. (2017). Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nature Communications*, *8*(1), 1–11.

Tang, S., Hemberg, M., Cansizoglu, E., Belin, S., Kosik, K., Kreiman, G., Steen, H., & Steen, J. (2016). f-divergence cutoff index to simultaneously identify differential expression in the integrated transcriptome and proteome. *Nucleic Acids Research*, *44*(10), e97.

Tansey, K. E., Cameron, D., & Hill, M. J. (2018). Genetic risk for Alzheimer's disease is concentrated in specific macrophage and microglial transcriptional networks. *Genome Medicine*, *10*(1), 14.

Taylor-Weiner, A., Aguet, F., Haradhvala, N. J., Gosai, S., Anand, S., Kim, J., Ardlie, K., Van Allen, E. M., & Getz, G. (2019). Scaling computational genomics to millions of individuals with GPUs. *Genome Biology*, *20*(1), 228.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., Johansen, C. T., Fouchier, S. W., Isaacs, A., Peloso, G. M., Barbalic, M., Ricketts, S. L., Bis, J. C., Aulchenko, Y. S., Thorleifsson, G., … Kathiresan, S. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, *466*(7307), 707–713.

Thaventhiran, J. E. D., Lango Allen, H., Burren, O. S., Rae, W., Greene, D., Staples, E., Zhang, Z., Farmery, J. H. R., Simeoni, I., Rivers, E., Maimaris, J., Penkett, C. J., Stephens, J., Deevi, S. V. V., Sanchis-Juan, A., Gleadall, N. S., Thomas, M. J., Sargur, R. B., Gordins,

P., … Smith, K. G. C. (2020). Whole-genome sequencing of a sporadic primary immunodeficiency cohort. *Nature*. https://doi.org/10.1038/s41586-020-2265-1

The PsychENCODE Consortium, Akbarian, S., Liu, C., Knowles, J. A., Vaccarino, F. M., Farnham, P. J., Crawford, G. E., Jaffe, A. E., Pinto, D., Dracheva, S., Geschwind, D. H., Mill, J., Nairn, A. C., Abyzov, A., Pochareddy, S., Prabhakar, S., Weissman, S., Sullivan, P. F., Matthew W State, … Sestan, N. (2015). The PsychENCODE project. *Nature Neuroscience*, *18*(12), 1707.

The UniProt Consortium. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158–D169.

Tian, B., Pan, Z., & Lee, J. Y. (2007). Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Research*, *17*(2), 156–165.

Tian, Y., Babor, M., Lane, J., Schulten, V., Patil, V. S., Seumois, G., Rosales, S. L., Fu, Z., Picarda, G., Burel, J., Zapardiel-Gonzalo, J., Tennekoon, R. N., De Silva, A. D., Premawansa, S., Premawansa, G., Wijewickrama, A., Greenbaum, J. A., Vijayanand, P., Weiskopf, D., … Peters, B. (2017). Unique phenotypes and clonal expansions of human CD4 effector memory T cells re-expressing CD45RA. *Nature Communications*, *8*(1), 1473.

Tivol, E. A., Borriello, F., Schweitzer, A. N., Lynch, W. P., Bluestone, J. A., & Sharpe, A. H. (1995). Loss of CTLA-4 leads to massive lymphoproliferation and fatal multiorgan tissue destruction, revealing a critical negative regulatory role of CTLA-4. *Immunity*, *3*(5), 541–547.

Torkamani, A., Wineinger, N. E., & Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews. Genetics*, *19*(9), 581–590.

Torres, J. M., Abdalla, M., Payne, A., Fernandez-Tajes, J., Thurner, M., Nylander, V., Gloyn, A. L., Mahajan, A., & McCarthy, M. I. (2020). A Multi-omic Integrative Scheme Characterizes Tissues of Action at Loci Associated with Type 2 Diabetes. *American Journal of Human Genetics*, *107*(6), 1011–1028.

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 5233.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Research*, *25*(10), 1491–1498.

Trynka, G., Hunt, K. A., Bockett, N. A., Romanos, J., Mistry, V., Szperl, A., Bakker, S. F., Bardella, M. T., Bhaw-Rosun, L., Castillejo, G., de la Concha, E. G., de Almeida, R. C., Dias, K.-R. M., van Diemen, C. C., Dubois, P. C. A., Duerr, R. H., Edkins, S., Franke, L., Fransen, K., … van Heel, D. A. (2011). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature Genetics*, *43*(12),

1193–1201.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., & Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics*, *45*(2), 124–130.

Trynka, G., Westra, H.-J., Slowikowski, K., Hu, X., Xu, H., Stranger, B. E., Klein, R. J., Han, B., & Raychaudhuri, S. (2015). Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *American Journal of Human Genetics*, *97*(1), 139–152.

Trzupek, D., Dunstan, M., Cutler, A. J., Lee, M., Godfrey, L., Jarvis, L., Rainbow, D. B., Aschenbrenner, D., Jones, J. L., Uhlig, H. H., Wicker, L. S., Todd, J. A., & Ferreira, R. C. (2020). Discovery of CD80 and CD86 as recent activation markers on regulatory T cells by protein-RNA single-cell analysis. *Genome Medicine*, *12*(1), 55.

Tsoi, L. C., Spain, S. L., Knight, J., Ellinghaus, E., Stuart, P. E., Capon, F., Ding, J., Li, Y., Tejasvi, T., Gudjonsson, J. E., Kang, H. M., Allen, M. H., McManus, R., Novelli, G., Samuelsson, L., Schalkwijk, J., Ståhle, M., Burden, A. D., Smith, C. H., … Wellcome Trust Case Control Consortium 2. (2012). Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nature Genetics*, *44*(12), 1341–1348.

Turner, M. D., Nedjai, B., Hurst, T., & Pennington, D. J. (2014). Cytokines and chemokines: At the crossroads of cell signalling and inflammatory disease. *Biochimica et Biophysica Acta*, *1843*(11), 2563–2582.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., & Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods*, *13*(9), 731–740.

Ubaid Ullah, Andrabi, S. B. A., Tripathi, S. K., Dirasantha, O., Kanduri, K., Rautio, S., Gross, C. C., Lehtimäki, S., Bala, K., Tuomisto, J., Bhatia, U., Chakroborty, D., Elo, L. L., Lähdesmäki, H., Wiendl, H., Rasool, O., & Lahesmaa, R. (2018). Transcriptional Repressor HIC1 Contributes to Suppressive Function of Human Induced Regulatory T Cells. *Cell Reports*, *22*(8), 2094–2106.

Udalova, I. A., Mantovani, A., & Feldmann, M. (2016). Macrophage heterogeneity in the context of rheumatoid arthritis. *Nature Reviews. Rheumatology*, *12*, 472.

UK10K Consortium, Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J. R. B., Xu, C., Futema, M., Lawson, D., Iotchkova, V., Schiffels, S., Hendricks, A. E., Danecek, P., Li, R., Floyd, J., Wain, L. V., Barroso, I., … Soranzo, N. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, *526*(7571), 82–90.

Ulirsch, J. C., Lareau, C. A., Bao, E. L., Ludwig, L. S., Guo, M. H., Benner, C., Satpathy, A. T., Kartha, V. K., Salem, R. M., Hirschhorn, J. N., Finucane, H. K., Aryee, M. J., Buenrostro, J. D., & Sankaran, V. G. (2019). Interrogation of human hematopoiesis at

single-cell and single-variant resolution. *Nature Genetics*, *51*(4), 683–693.

Ulirsch, J. C., Nandakumar, S. K., Wang, L., Giani, F. C., Zhang, X., Rogov, P., Melnikov, A., McDonel, P., Do, R., Mikkelsen, T. S., & Sankaran, V. G. (2016). Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell*, *165*(6), 1530–1545.

Urbut, S. M., Wang, G., Carbonetto, P., & Stephens, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics*, *51*(1), 187–195.

Vahedi, G., Takahashi, H., Nakayamada, S., Sun, H.-W., Sartorelli, V., Kanno, Y., & O'Shea, J. J. (2012). STATs shape the active enhancer landscape of T cell populations. *Cell*, *151*(5), 981–993.

van Beek, J. J. P., Rescigno, M., & Lugli, E. (2021). A fresh look at the T helper subset dogma. *Nature Immunology*, *22*(2), 104–105.

van der Wijst, M. G. P., Brugge, H., de Vries, D. H., Deelen, P., Swertz, M. A., LifeLines Cohort Study, BIOS Consortium, & Franke, L. (2018). Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nature Genetics*, *50*(4), 493–497.

Veldhoen, M., Uyttenhove, C., van Snick, J., Helmby, H., Westendorf, A., Buer, J., Martin, B., Wilhelm, C., & Stockinger, B. (2008). Transforming growth factor-beta "reprograms" the differentiation of T helper 2 cells and promotes an interleukin 9-producing subset. *Nature Immunology*, *9*(12), 1341–1346.

Verhelst, J., Parthoens, E., Schepens, B., Fiers, W., & Saelens, X. (2012). Interferon-inducible protein Mx1 inhibits influenza virus by interfering with functional viral ribonucleoprotein complex assembly. *Journal of Virology*, *86*(24), 13445–13455.

Visscher, P. M., & Goddard, M. E. (2019). From R.A. Fisher's 1918 Paper to GWAS a Century Later. *Genetics*, *211*(4), 1125–1130.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, *101*(1), 5–22.

Viswanath, B., Jose, S. P., Squassina, A., Thirthalli, J., Purushottam, M., Mukherjee, O., Vladimirov, V., Patrinos, G. P., Del Zompo, M., & Jain, S. (2015). Cellular models to study bipolar disorder: A systematic review. *Journal of Affective Disorders*, *184*, 36–50.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., Pervjakova, N., Alvaes, I., Fave, M.-J., Agbessi, M., Christiansen, M., Jansen, R., Seppälä, I., Tong, L., Teumer, A., … Franke, L. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. In *bioRxiv* (p. 447367). https://doi.org/10.1101/447367

Wallace, C. (2013). Statistical testing of shared genetic control for potentially related traits.

*Genetic Epidemiology*, *37*(8), 802–813.

Wallace, C. (2020). Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genetics*, *16*(4), e1008720.

Wallace, C., Rotival, M., Cooper, J. D., Rice, C. M., Yang, J. H. M., McNeill, M., Smyth, D. J., Niblett, D., Cambien, F., Cardiogenics Consortium, Tiret, L., Todd, J. A., Clayton, D. G., & Blankenberg, S. (2012). Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Human Molecular Genetics*, *21*(12), 2815–2824.

Wang, H., La Russa, M., & Qi, L. S. (2016). CRISPR/Cas9 in Genome Editing and Beyond. *Annual Review of Biochemistry*, *85*, 227–264.

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–678.

Wen, X., Pique-Regi, R., & Luca, F. (2017). Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genetics*, *13*(3), e1006646.

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M. W., Fairfax, B. P., Schramm, K., Powell, J. E., Zhernakova, A., Zhernakova, D. V., Veldink, J. H., Van den Berg, L. H., Karjalainen, J., Withoff, S., Uitterlinden, A. G., Hofman, A., Rivadeneira, F., … Franke, L. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, *45*(10), 1238–1243.

Wherry, E. J., Teichgräber, V., Becker, T. C., Masopust, D., Kaech, S. M., Antia, R., von Andrian, U. H., & Ahmed, R. (2003). Lineage relationship and protective immunity of memory CD8 T cell subsets. *Nature Immunology*, *4*(3), 225–234.

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1), 15.

Wooldridge, L., Ekeruche-Makinde, J., van den Berg, H. A., Skowera, A., Miles, J. J., Tan, M. P., Dolton, G., Clement, M., Llewellyn-Lacey, S., Price, D. A., Peakman, M., & Sewell, A. K. (2012). A single autoimmune T cell receptor recognizes more than a million different peptides. *The Journal of Biological Chemistry*, *287*(2), 1168–1177.

Xu, Y., Vuckovic, D., Ritchie, S. C., Akbari, P., Jiang, T., Grealey, J., Butterworth, A. S., Ouwehand, W. H., Roberts, D. J., Di Angelantonio, E., Danesh, J., Soranzo, N., & Inouye, M. (2020). Learning polygenic scores for human blood cell traits. In *bioRxiv* (p. 2020.02.17.952788). https://doi.org/10.1101/2020.02.17.952788

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height.

*Nature Genetics*, *42*(7), 565–569.

Yang, J., Zeng, J., Goddard, M. E., Wray, N. R., & Visscher, P. M. (2017). Concepts, estimation and interpretation of SNP-based heritability. *Nature Genetics*, *49*(9), 1304–1310.

Yang, X. O., Pappu, B. P., Nurieva, R., Akimzhanov, A., Kang, H. S., Chung, Y., Ma, L., Shah, B., Panopoulos, A. D., Schluns, K. S., Watowich, S. S., Tian, Q., Jetten, A. M., & Dong, C. (2008). T helper 17 lineage differentiation is programmed by orphan nuclear receptors ROR alpha and ROR gamma. *Immunity*, *28*(1), 29–39.

Yao, C., Chen, G., Song, C., Keefe, J., Mendelson, M., Huan, T., Sun, B. B., Laser, A., Maranville, J. C., Wu, H., Ho, J. E., Courchesne, P., Lyass, A., Larson, M. G., Gieger, C., Graumann, J., Johnson, A. D., Danesh, J., Runz, H., … Levy, D. (2018). Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature Communications*, *9*(1), 3268.

Ye, C. J., Feng, T., Kwon, H.-K., Raj, T., Wilson, M. T., Asinovski, N., McCabe, C., Lee, M. H., Frohlich, I., Paik, H.-I., Zaitlen, N., Hacohen, N., Stranger, B., De Jager, P., Mathis, D., Regev, A., & Benoist, C. (2014). Intersection of population variation and autoimmunity genetics in human T cell activation. *Science*, *345*(6202), 1254665.

Yoshizaki, A., Miyagaki, T., DiLillo, D. J., Matsushita, T., Horikawa, M., Kountikov, E. I., Spolski, R., Poe, J. C., Leonard, W. J., & Tedder, T. F. (2012). Regulatory B cells control T-cell autoimmunity through IL-21-dependent cognate interactions. *Nature*, *491*(7423), 264–268.

Zeileis, A., & Grothendieck, G. (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software, Articles*, *14*(6), 1–27.

Zhang, L., Yu, X., Zheng, L., Zhang, Y., Li, Y., Fang, Q., Gao, R., Kang, B., Zhang, Q., Huang, J. Y., Konno, H., Guo, X., Ye, Y., Gao, S., Wang, S., Hu, X., Ren, X., Shen, Z., Ouyang, W., & Zhang, Z. (2018). Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature*, *564*(7735), 268–272.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, *9*(9), R137.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*, 14049.

Zheng, J., Haberland, V., Baird, D., Walker, V., Haycock, P. C., Hurle, M. R., Gutteridge, A., Erola, P., Liu, Y., Luo, S., Robinson, J., Richardson, T. G., Staley, J. R., Elsworth, B.,

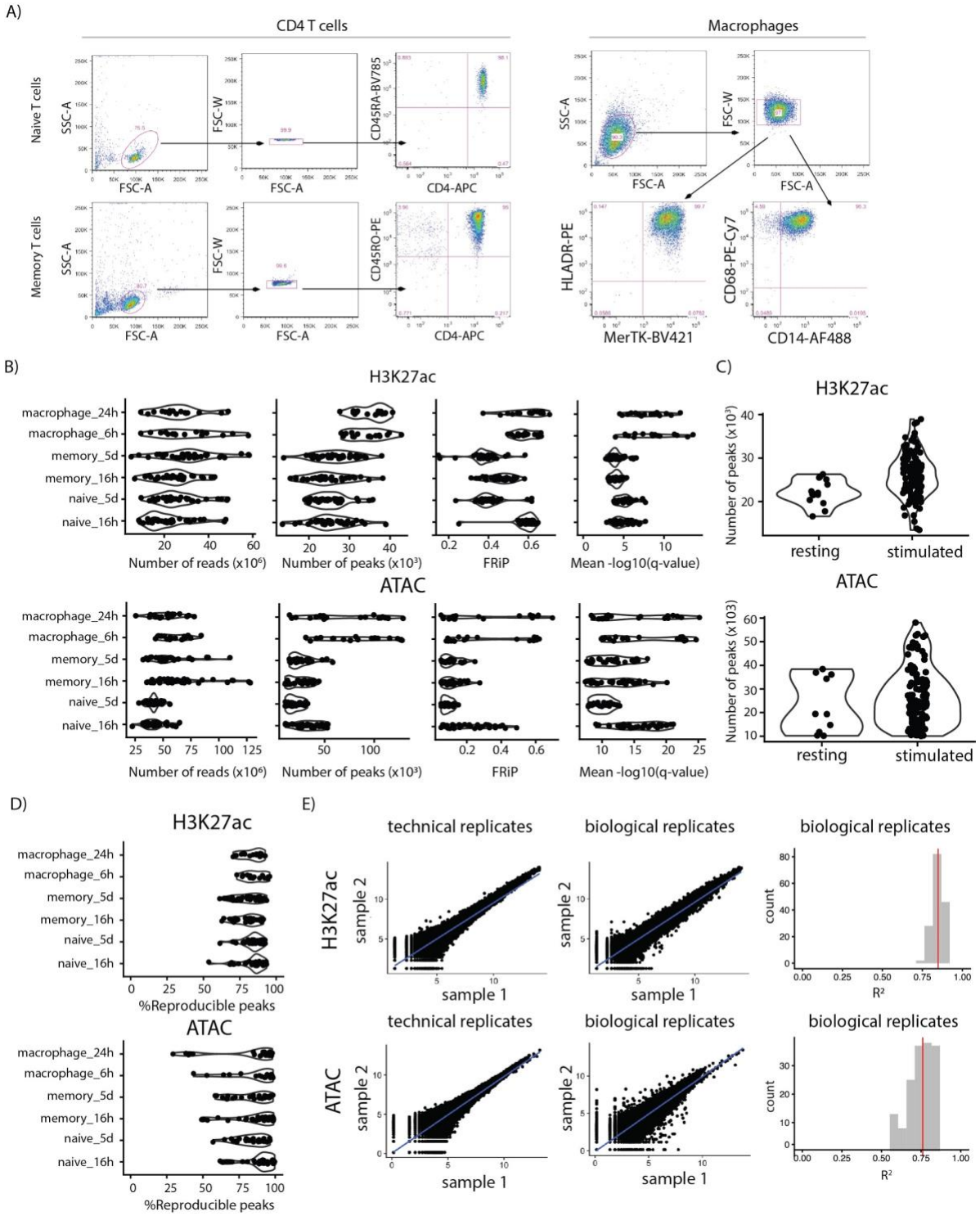
Burgess, S., Sun, B. B., Danesh, J., Runz, H., Maranville, J. C., … Gaunt, T. R. (2020). Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nature Genetics*, *52*(10), 1122–1131.

Zheng, W., & Flavell, R. A. (1997). The transcription factor GATA-3 is necessary and sufficient for Th2 cytokine gene expression in CD4 T cells. *Cell*, *89*(4), 587–596.

Zhu, J., Yamane, H., & Paul, W. E. (2010). Differentiation of effector CD4 T cell populations (*). *Annual Review of Immunology*, *28*, 445–489.

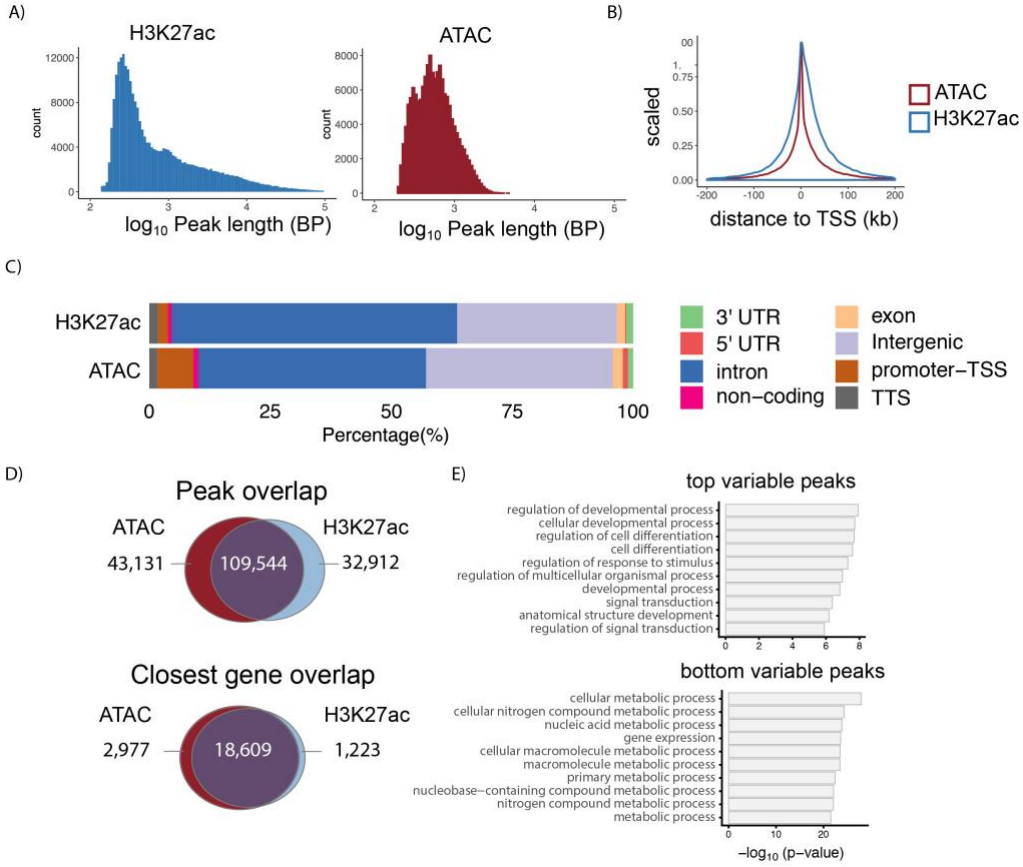
# Appendix A: Supplementary figures

## Chapter 2

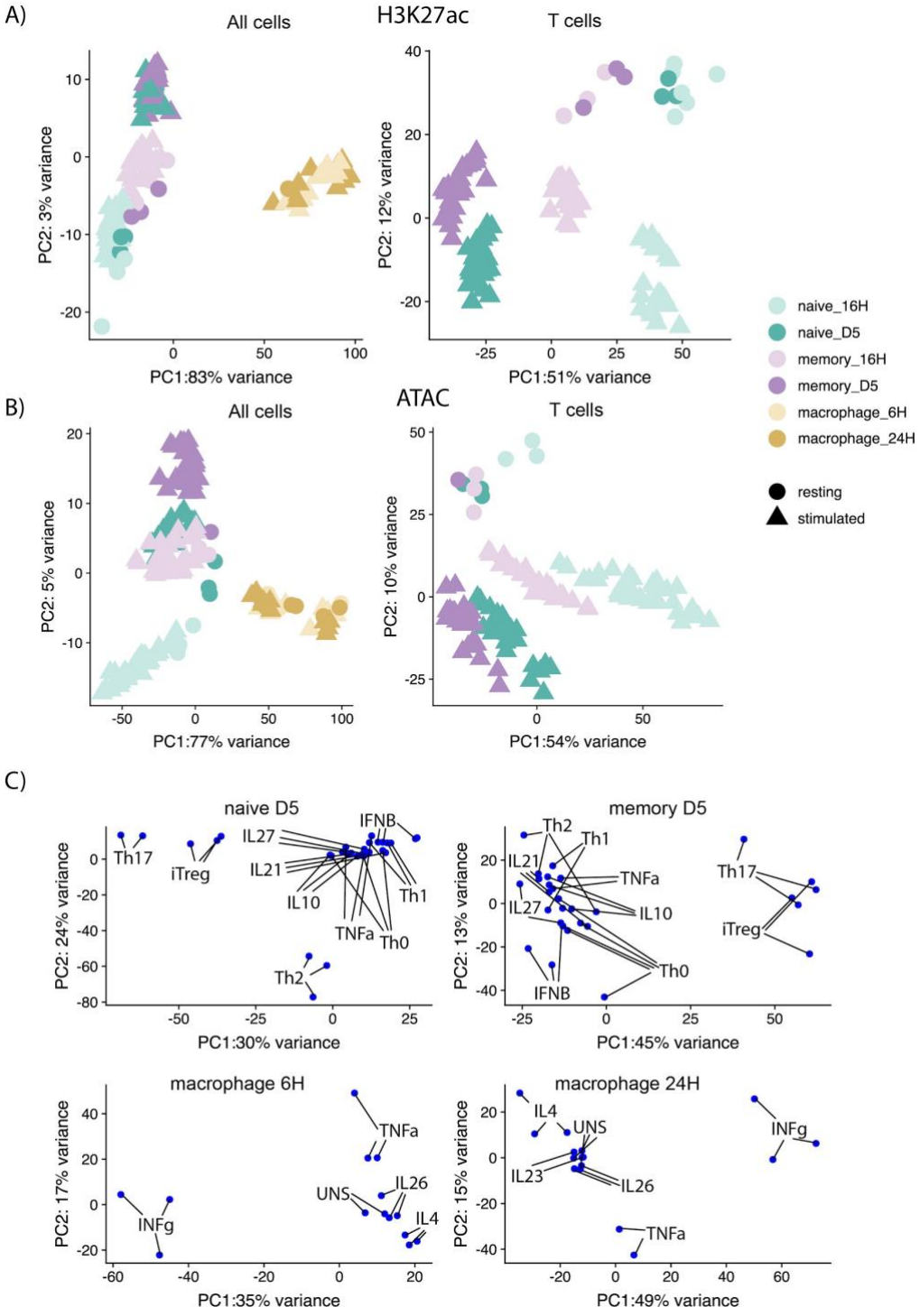



**Supplementary Figure 2.1. Quality control of ATAC and H3K27ac data**. **A)** Purity of naive T cells (CD4+CD45RA+), memory T cells (CD4+CD45RO+) and macrophages. Representative plots were

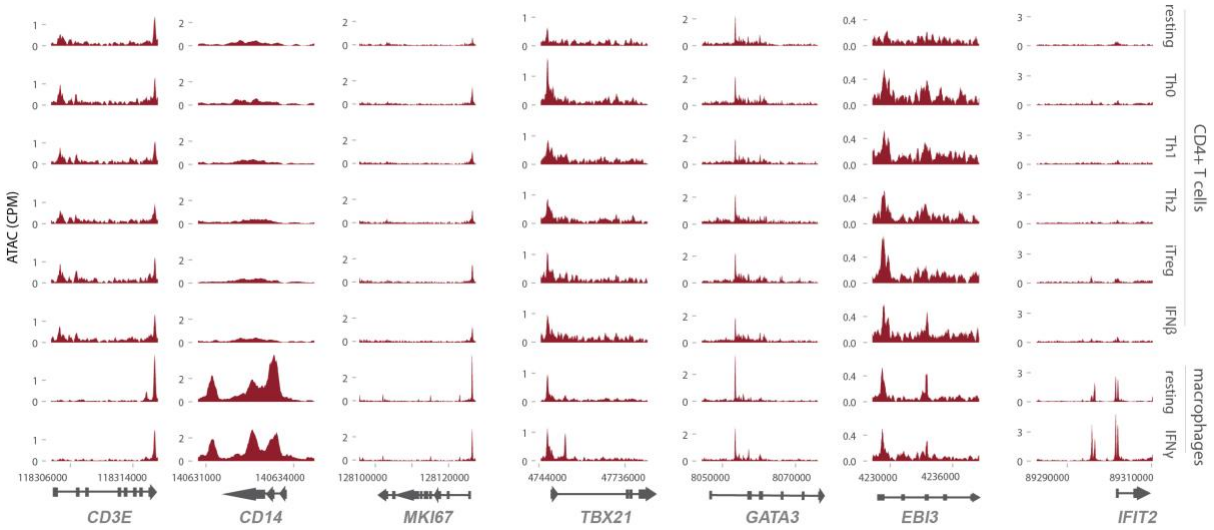
from three biologically independent samples in T cells and two biologically independent samples in macrophages. **B)** Number of reads, number of peaks, fraction of reads in peaks (FRiP) and mean q-value of H3K27ac and ATAC peaks. Peak q-values were determined using MACS2. Number of samples in H3K27ac is: 38 (naive_16h), 35 (naive_5d), 32 (memory_16h), 35 (memory_5d), 16 (macrophage_6h) and 17 (macrophage_24h). Number of samples in ATAC is: 36 (naive_16h), 31 (naive_5d), 34 (memory_16h), 28 (memory_5d), 19 (macrophage_6h) and 20 (macrophage_24h). **C)** Number of peaks in resting and stimulated naive and memory T cells. Stimulated cells were cultured with anti-CD3/anti-CD28 beads. Number of samples in H3K27ac is: 15 (resting) and 125 (stimulated). Number of samples in ATAC is: 10 (resting) and 119 (stimulated). We performed two tailed t-test (ChM: $t = 4.58$, $df = 25.1$, p-value = 0.0001; ATAC: $t = 0.87$, $df = 10.55$, p-value = 0.41). **D)** Percentage of reproducible peaks per condition. Reproducible peaks are peaks that were observed in at least two biological replicates of the same condition. Number of samples in H3K27ac and ATAC is the same as in panel B. E) Comparison of read counts in peaks between biologically independent samples (between individuals) and 15 technical replicates (within an individual) included in our study. The red line represents the mean R2.



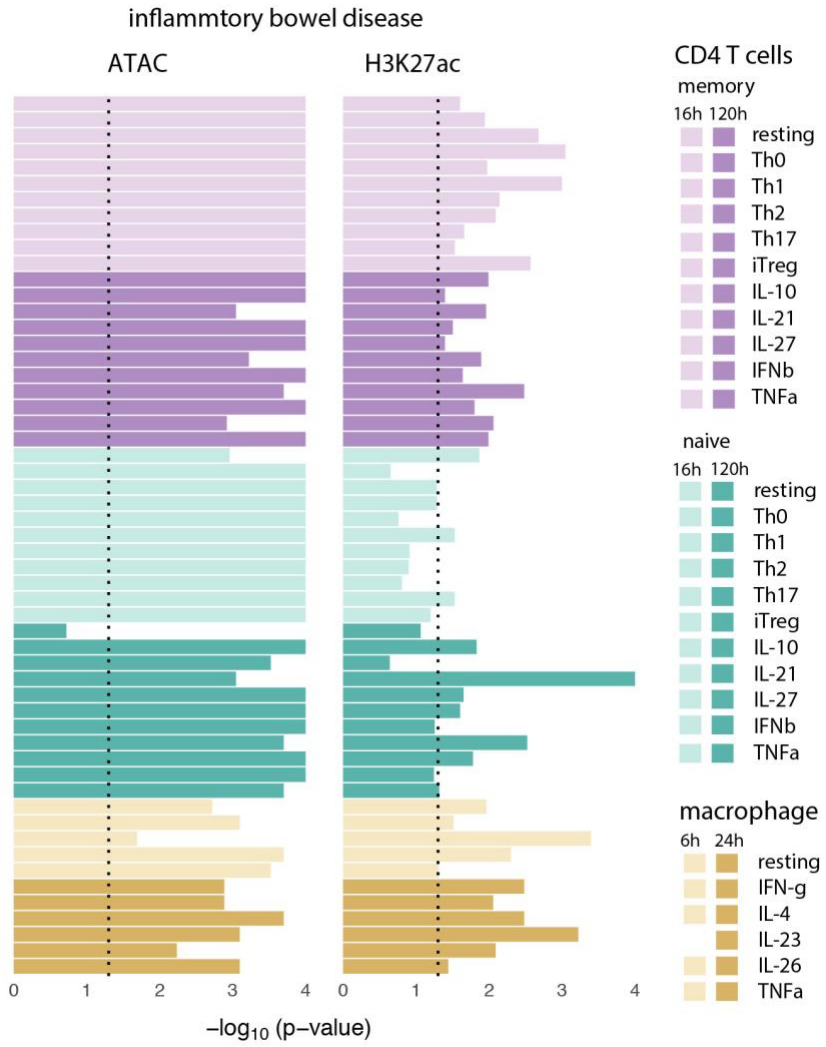

**Supplementary Figure 2.2. Chromatin regulatory landscape in cytokine induced immune cell states captured by H3K27ac ChIPmentation-seq and ATAC-seq. A)** Distribution of ATAC and H3K27ac peak length. **B)** Distribution of ATAC and H3K27ac peaks relative to TSS. **C)** Proportion of ATAC and H3K27ac in different genomic annotations. **D)** Overlap between ATAC and H3K27ac peaks (upper panel) and of the genes in proximity to peaks (bottom panel). **E)** Pathway enrichment analysis using genes in proximity to the top and bottom 1% variable H3K27ac peaks. P-values were calculated using g:ProfileR with default parameters. The number of genes used for the enrichment analysis was 1,156 (nearby top variable peaks) and 1,253 (nearby bottom variable peaks)



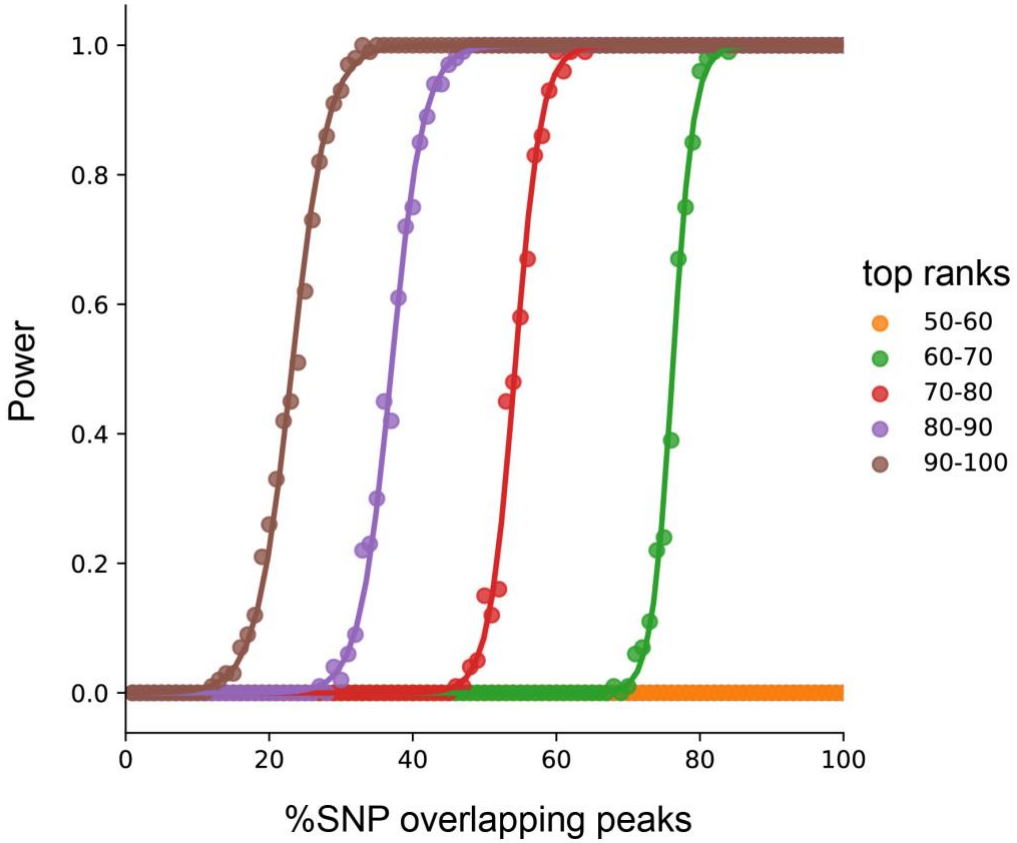

**Supplementary Figure 2.3. PCA of H3K27ac and ATAC data. A)** PCA using H3K27ac and **B)** ATAC-seq data across all cell types and cell states within the study (left), and T cells only (right). Each sample is colored by broad cell lineage, while the color shade represents duration of activation. Circles represent resting cells and triangles correspond to stimulated cells. PCA plots were derived using 73 naive T cells, 67 memory T cells and 33 macrophage states for H3K27ac (A) and 67 naive T cells, 62 memory T cells and 39 macrophage states for ATAC (B). **C)** PCA of T cells using reads in H3K27ac peaks within cell type and stimulation time point. PCA plots were derived using 31 naive D5 T cells, 29 memory D5 T cells, 14 macrophages stimulated for 6h and 15 macrophages stimulated for 24h cell states.



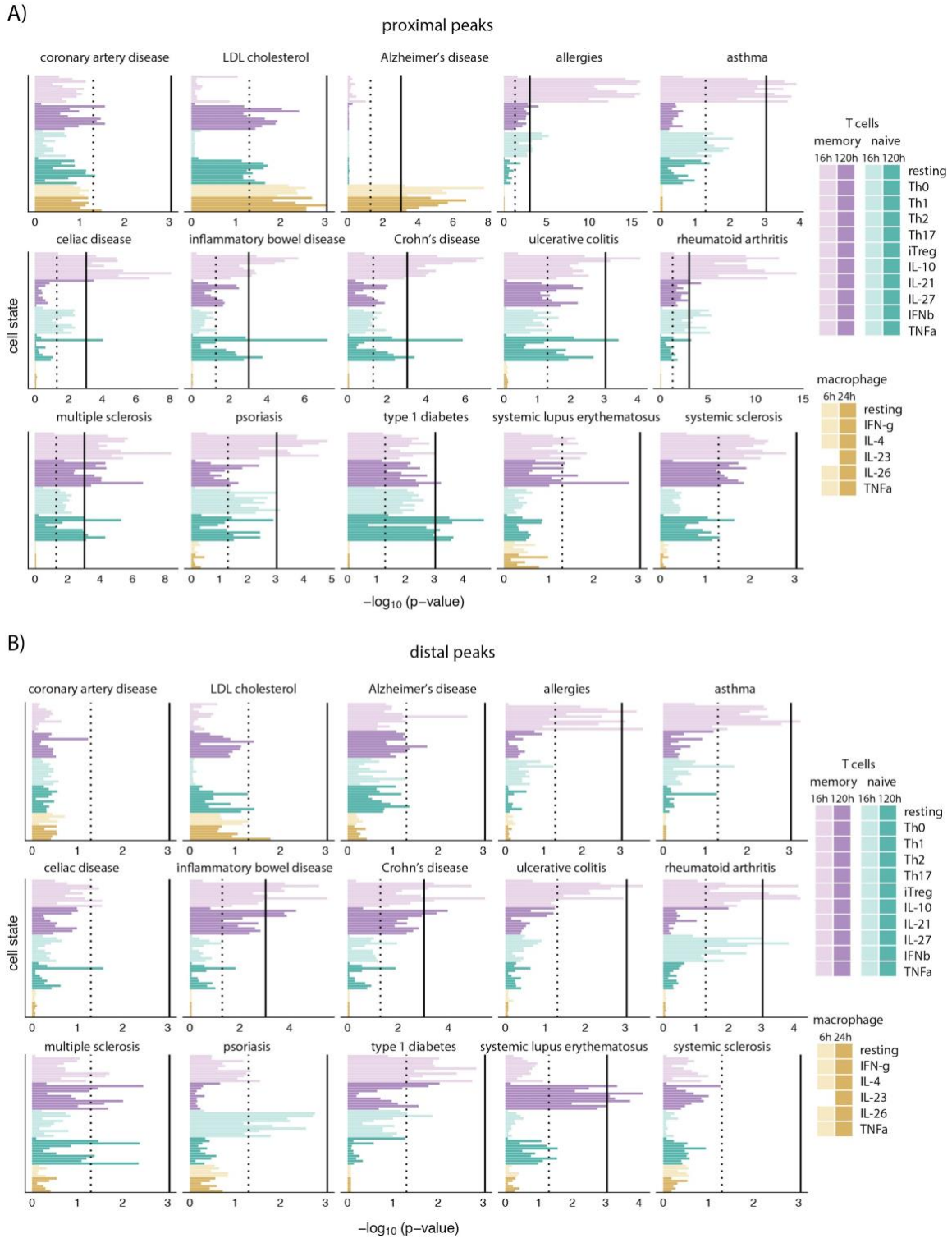

**Supplementary Figure 2.4. ATAC-seq read pile-ups at hallmark loci for cytokine induced cell states.** ATAC-seq read pile ups in regions proximal to condition specific hallmark genes: CD3 (T cells), CD14 (macrophages), KI67 (T cell activation), TBX21 (Th1), GATA3 (Th2), IL35B (i.e. EBI3, iTreg), IFIT2 (IFN-induced). Genomic coordinates (GRCh38) for each gene are labeled. ATAC-seq tracks were generated after merging three biologically independent samples per cell state.



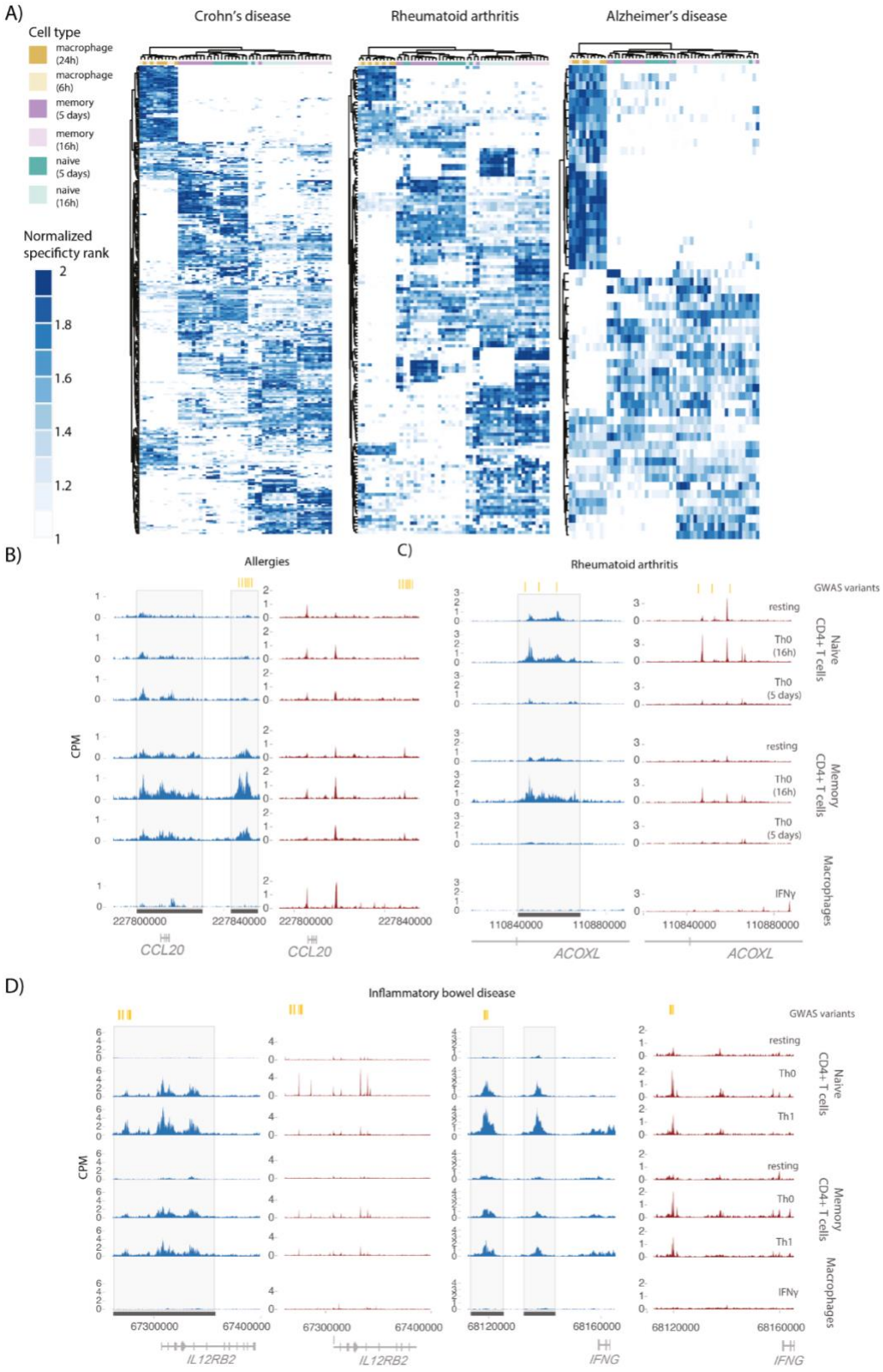

**Supplementary Figure 2.5. Binary SNP-peak overlap does not discriminate between closely related cell states.** We used GoShifter to test for SNP enrichment across cytokine induced cell states. GoShifter was performed after merging three biologically independent samples per cell state. P-values were calculated using 10,000 permutations implemented in GoShifter.




**Supplementary Figure 2.6. Power calculations for CHEERS.** Colors represent percentile of top cell state peak specificity ranks which were sampled to compute power of detecting significant enrichment under different percentages (x-axis) of SNPs mapping in these top peaks.

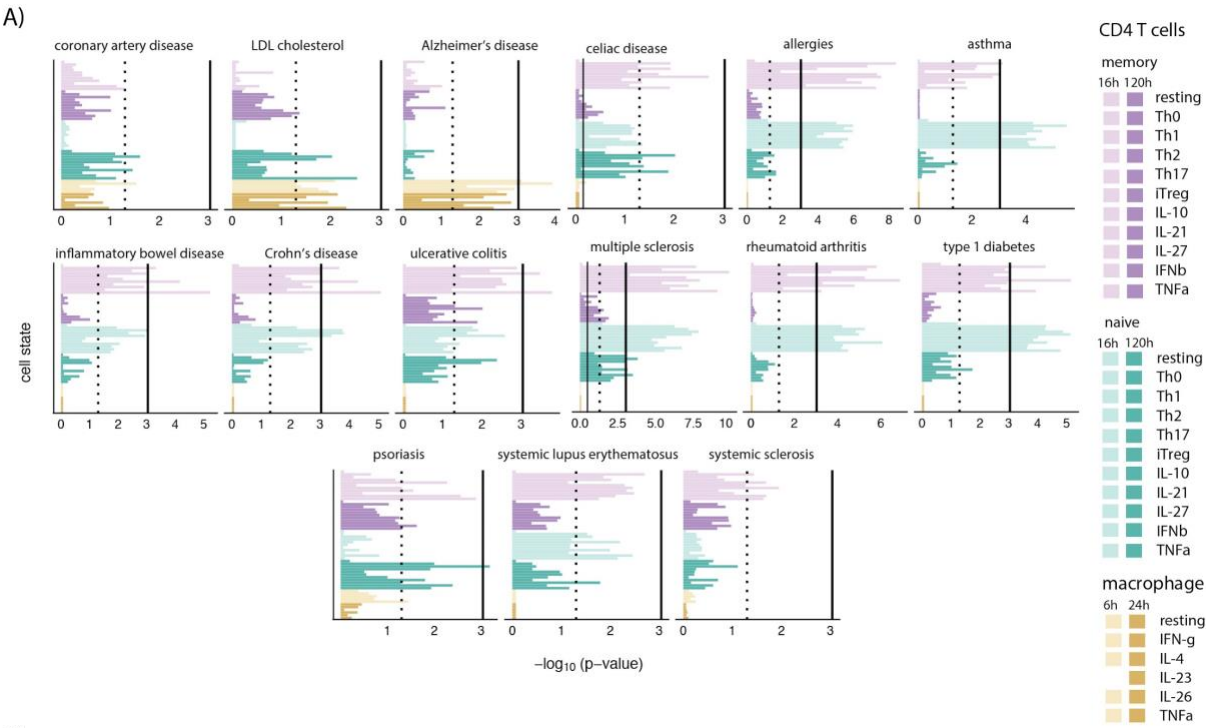


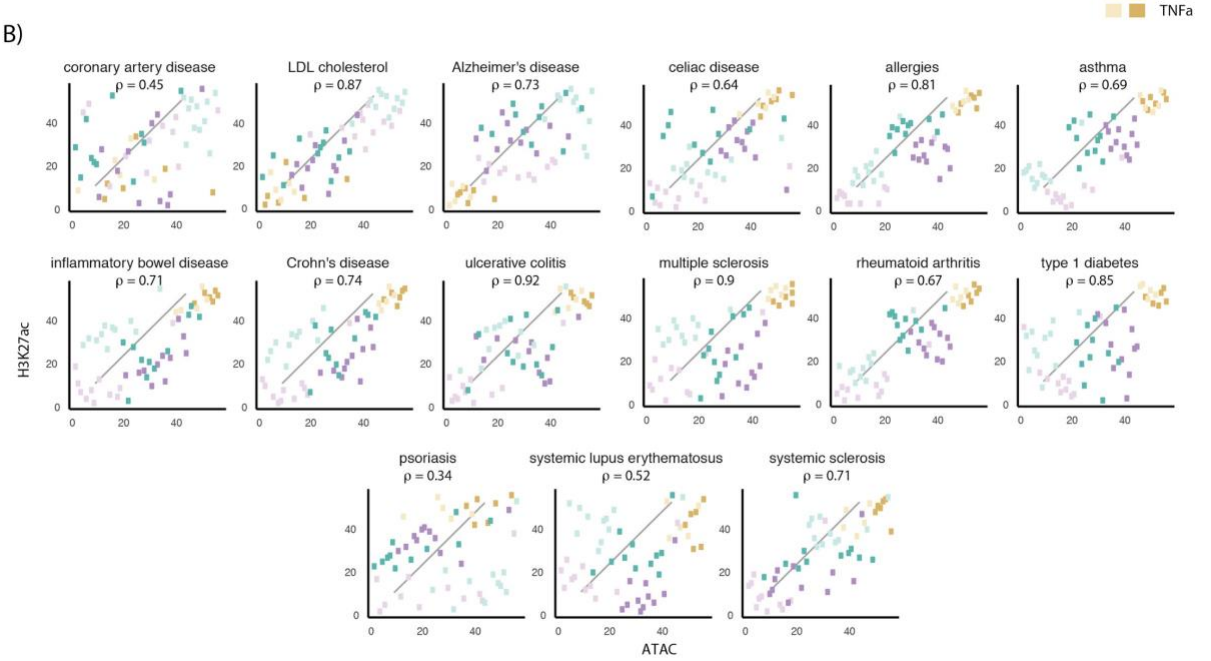

**Supplementary Figure 2.7. Disease SNP enrichment in TSS proximal and distal H3K27ac regions.** One-sided p-values are reported from a discrete uniform distribution. CHEERS was performed after merging three biologically independent samples. GWAS data and the statistical method are described in the "GWAS data processing" and “CHEERS” sections of the Methods. The dotted gray line marks the nominal p-value threshold of 0.05, while the solid gray line represents the Bonferronicorrected significance threshold (p-value ≤ 9 x 10-4). A) Enrichment using proximal (≤5kb from TSS, 79,419 peaks) and B) distal peaks (>5kb from TSS, 48,305 peaks).



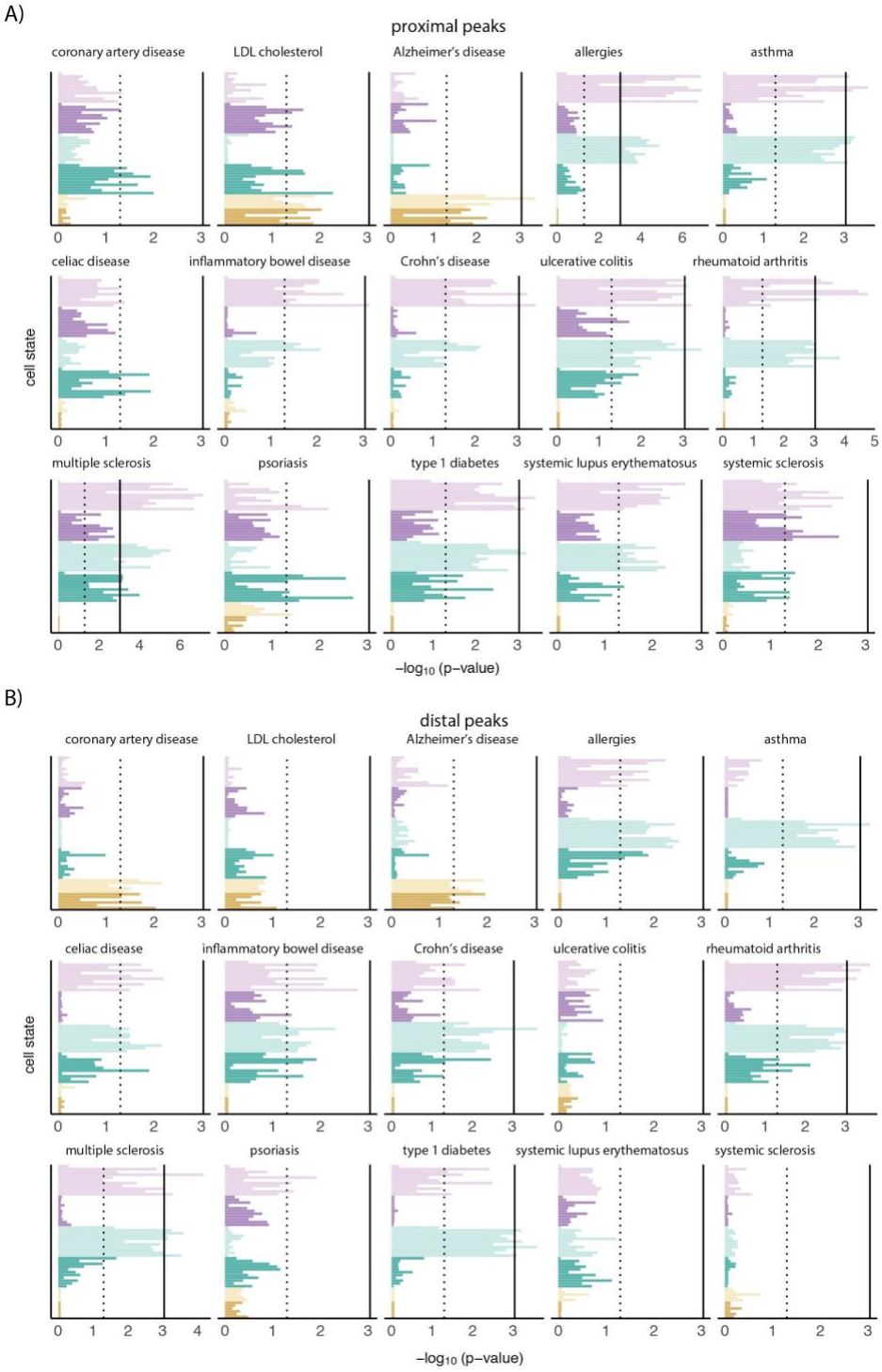

**Supplementary Figure 2.8. Clustering by cell state specificity score of peaks overlapping SNPs associated with Crohn's disease, rheumatoid arthritis and Alzheimer's disease. A)** Each row corresponds to a H3K27ac peak, while each column corresponds to a cytokine induced cell state. Shades of blue represent specificity of a peak (specificity rank of a peak normalized to the mean specificity rank of all peaks). B,C,D) Genomic coordinates (GRCh38) for each gene are labeled.

H3K27ac tracks are generated after merging three biologically independent samples per cell state. **B)** Read pileups of H3K27ac (blue) and ATAC-seq (red) peaks at loci driving the enrichment of allergy variants in T cell activation. **C)** Read pileups of H3K27ac (blue) and ATAC-seq (red) peaks at loci driving the enrichment of RA variants in T cell activation. **D)** Read pileups of H3K27ac (blue) and ATAC-seq (red) peaks at loci driving the enrichment of IBD variants in Th1 cells.

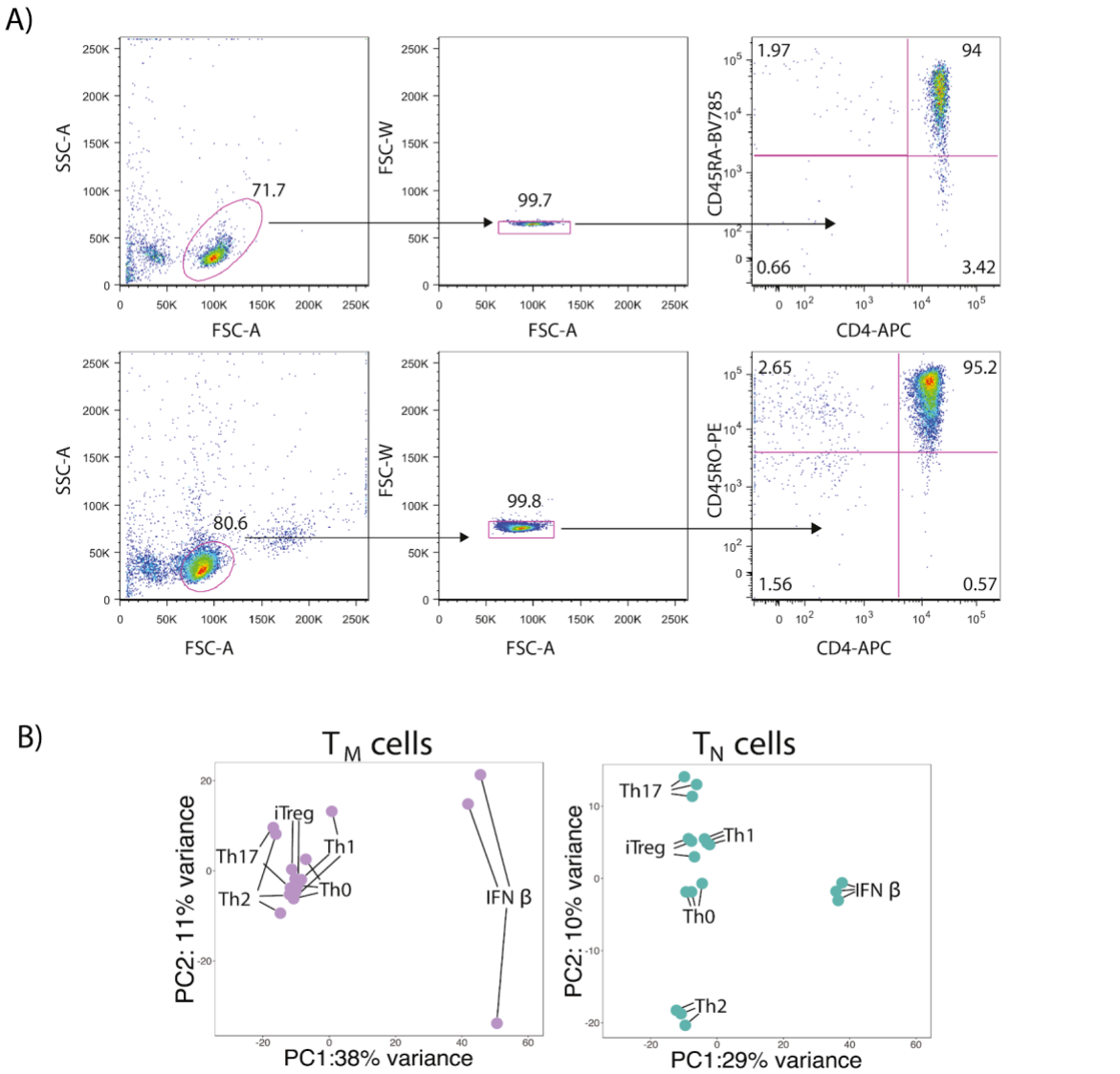


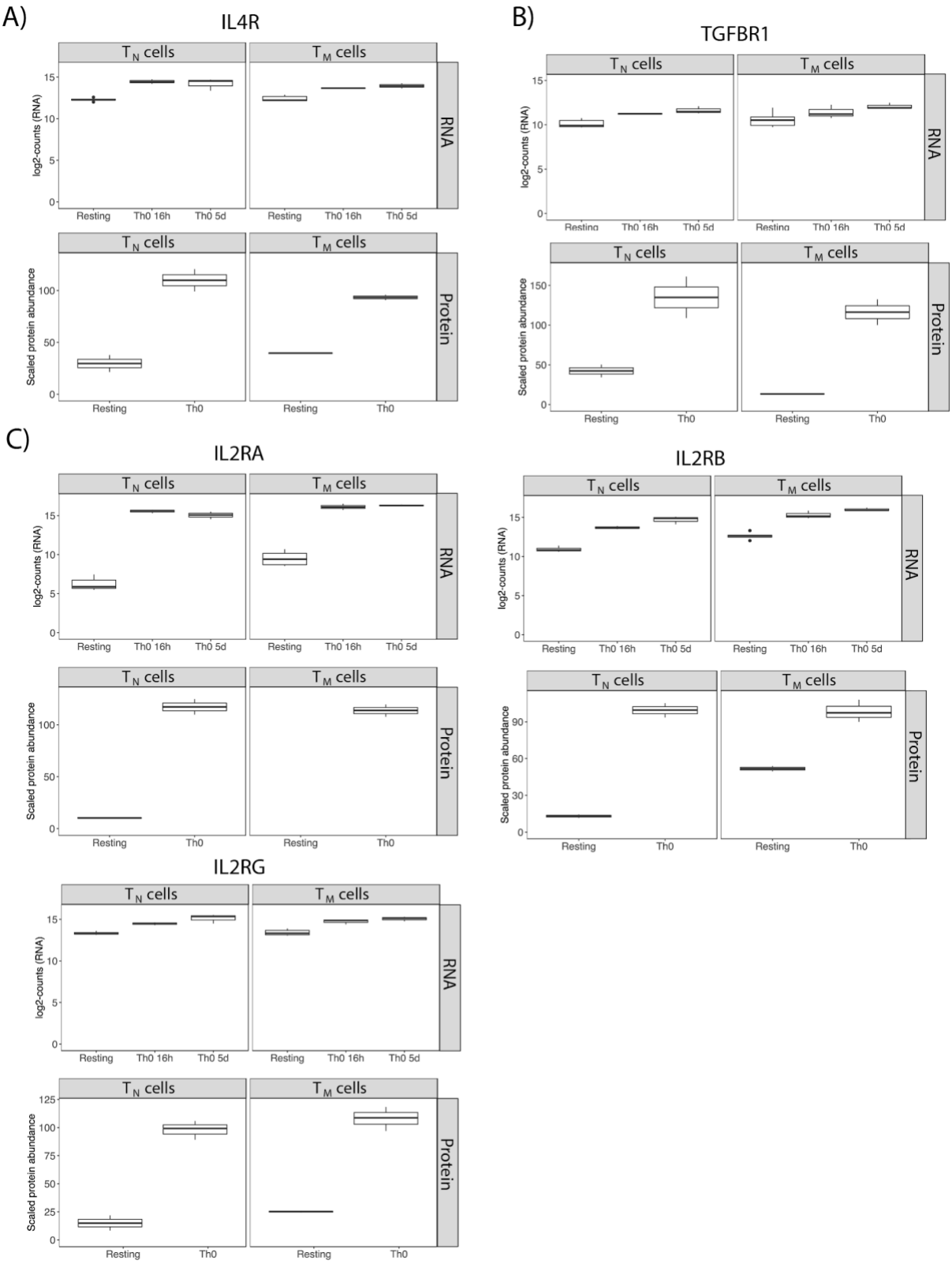
**Supplementary Figure 2.9. CHEERS SNP enrichment in ATAC-seq defined regions in cytokine induced cell states. A)** One-sided p-values are reported from a discrete uniform distribution. CHEERS was performed after merging three biologically independent samples resulting in 136,692 peaks. GWAS data and the statistical method are described in the "GWAS data processing" and “CHEERS” sections of the Methods. The dotted gray line marks the nominal p-value threshold of 0.05, while the solid gray line represents the Bonferroni-corrected significance threshold (p-value ≤ 9 x 10-4). **B)** Correlation of 55 p-value ranks from CHEERS SNP enrichments obtained using H3K27ac and ATAC-seq. The grey line depicts the identity line. Pearson’s correlation coefficient (PCC) is provided below each disease label.



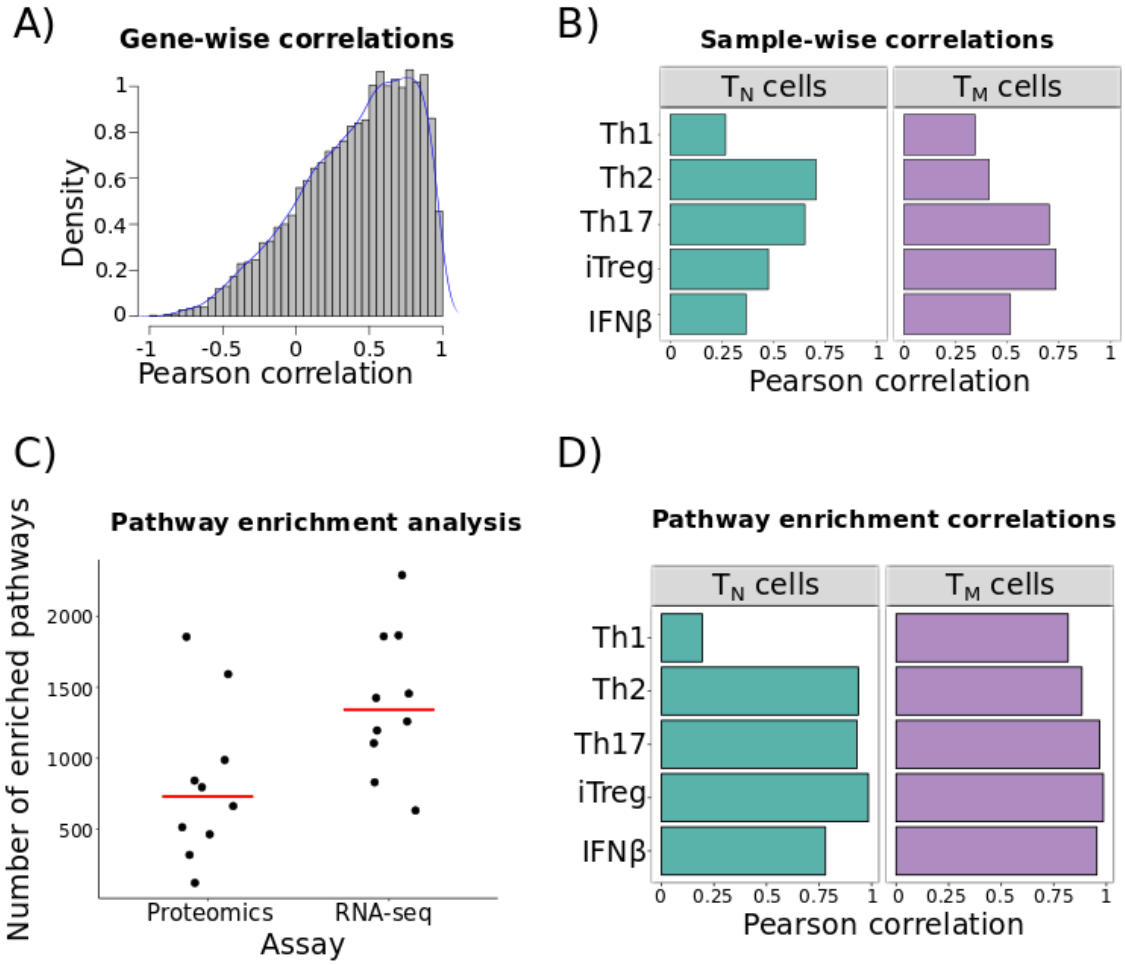

**Supplementary Figure 2.10. Disease SNP enrichment in TSS proximal and distal ATAC peaks** One-sided p-values are reported from a discrete uniform distribution. CHEERS was performed after merging three biologically independent samples. GWAS data and the statistical method are described in the "GWAS data processing" and “CHEERS” sections of the Methods. The dotted gray line marks the nominal p-value threshold of 0.05, while the solid gray line represents the Bonferronicorrected significance threshold (p-value ≤ 9 x 10-4). A) CHEERS was performed on proximal ATAC peaks (≤ 5kb from TSS, 88,150 peaks). B) CHEERS was performed on distal peaks (> 5kb from TSS, 48,543 peaks).



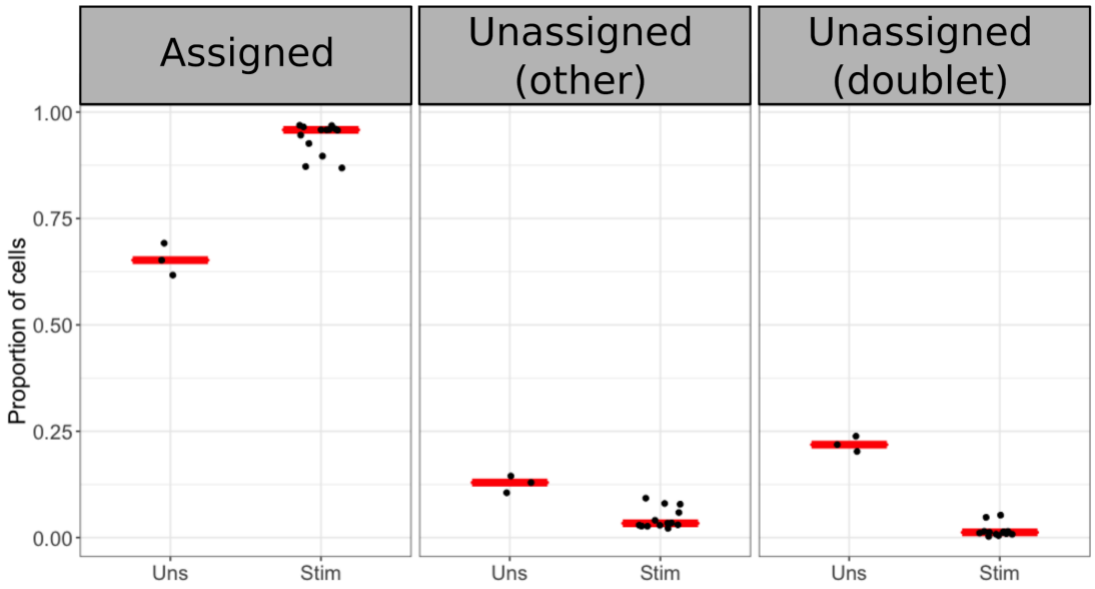

**Supplementary Figure 3.1. Purity and quality of samples. A)** Representative flow cytometry plots from three biologically independent samples. These profiles refer to the cells depicted in the scematic in Figure 1A. **B)** PCA plots from the full transcriptome of $T_N$ and $T_M$ cells following 16 hours of cytokine stimulations. Only stimulated cells were included in this analysis. PCA plots were derived using 21 naive and 20 memory T cells.



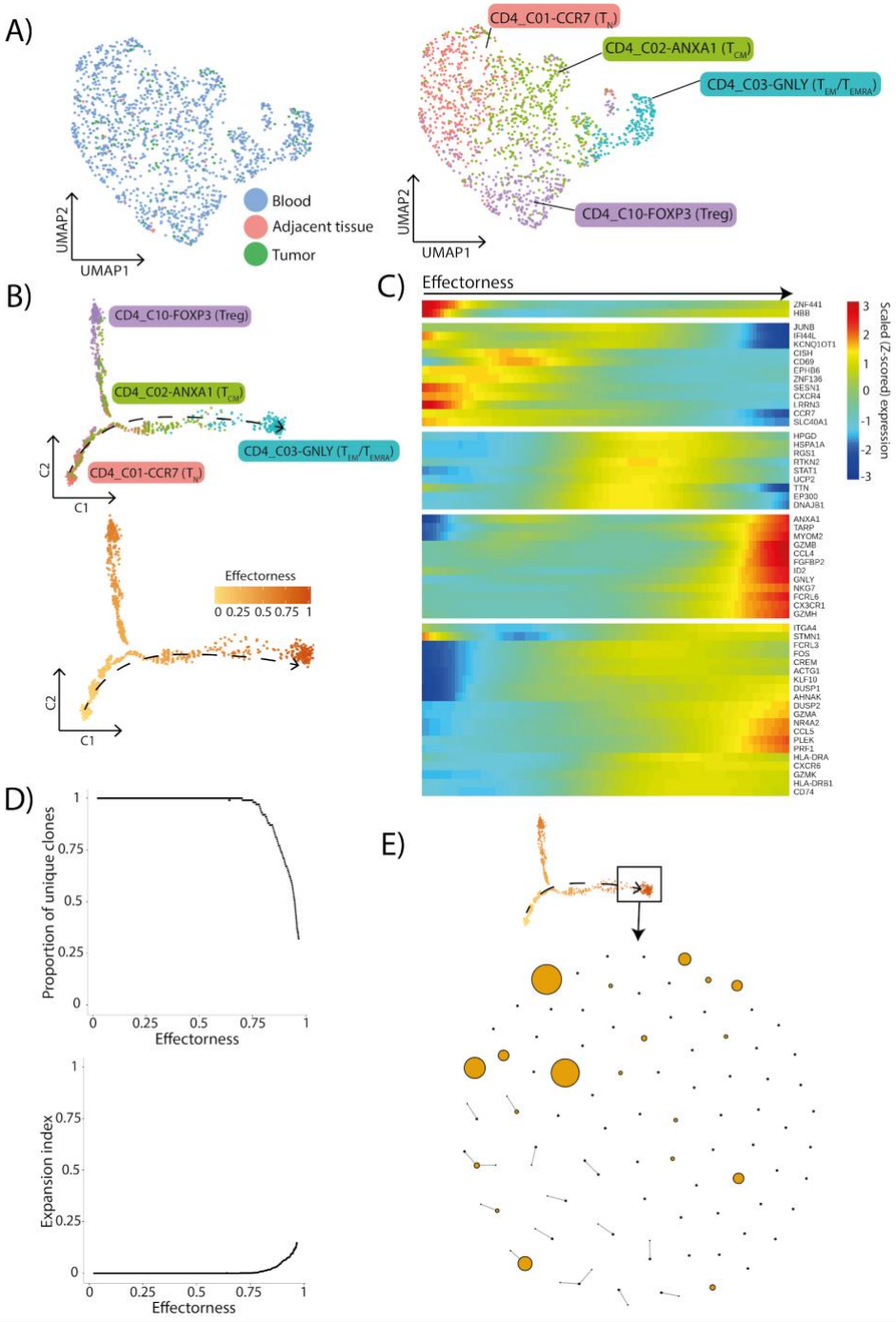

**Supplementary Figure 3.2. Expression of receptors for cytokines with differential effects between $T_N$ and $T_M$ cells.** RNA levels and protein abundances of cytokine receptor components in resting and stimulated $T_N$ and $T_M$ cells. Subunits of the **A)** IL-4, **B)** TGF-β and **C)** IL-2 receptors. Each boxplot was generated using three biologically independent samples. Boxplot centers represent median values, while the boxplot bounds represent the 25% and 75% quantiles. Boxplot whiskers represent the 25% quantile − 1.5*interquartile range (IQR) and the 75% quantile + 1.5*IQR, respectively.



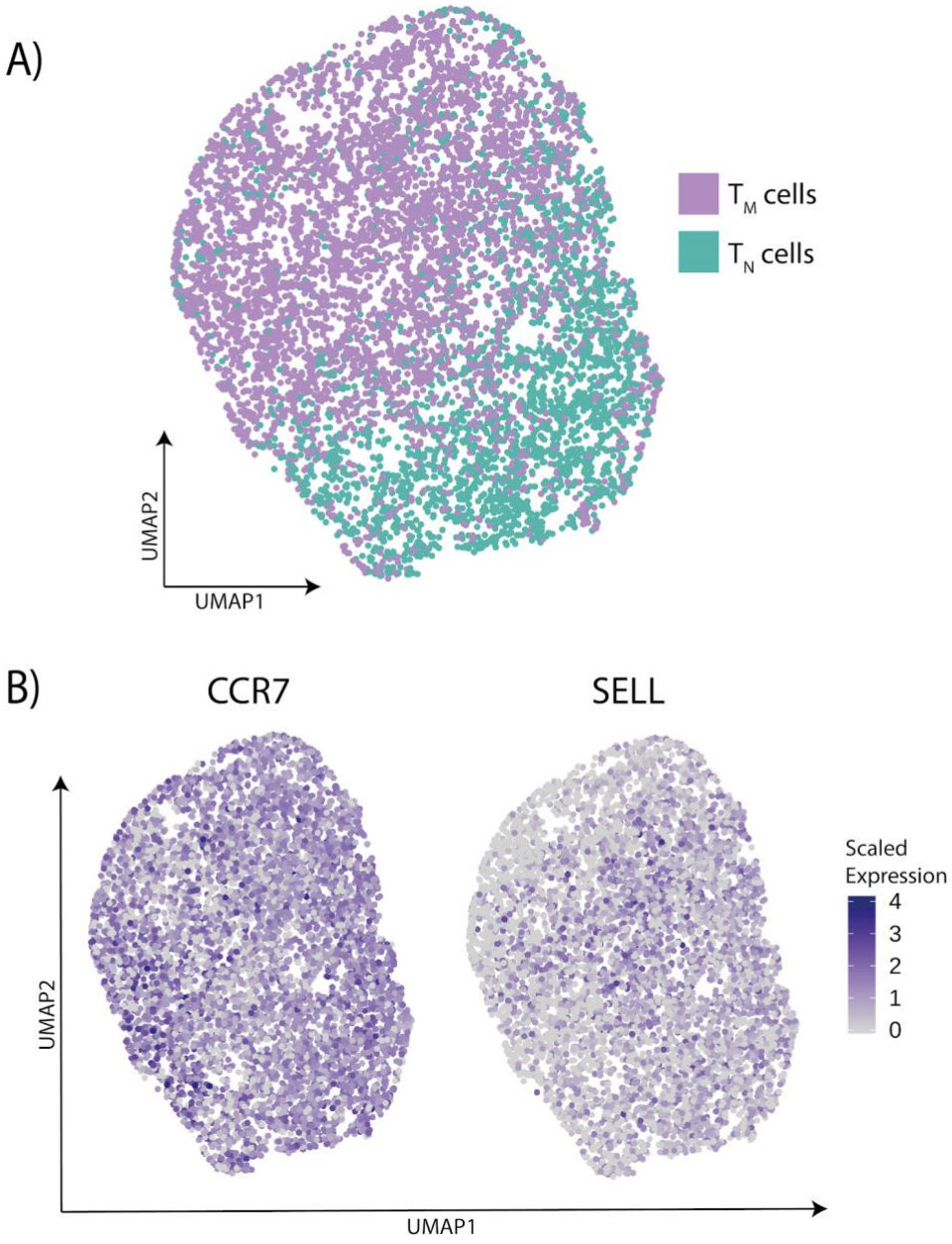

**Supplementary Figure 3.3. Correlation between RNA and protein expression. A)** Pearson correlation between RNA and protein log-fold changes for each gene across all cytokine conditions and cell types. Resting cells were excluded from this analysis. **B)** Pearson correlations between RNA and protein log-fold changes for all genes within each cytokine condition in $T_N$ and $T_M$ cells. **C)** Number of pathways detected as significantly enriched in differentially expressed genes and proteins from each cytokine condition. Each dot represents a cytokine condition in naive or memory T cell (N = 10). **D)** Pearson correlation between the pathway enrichment estimates derived from RNA and protein expression in each cytokine condition. Only pathways with an absolute enrichment score > 0.25 in both data sets were used for this analysis.



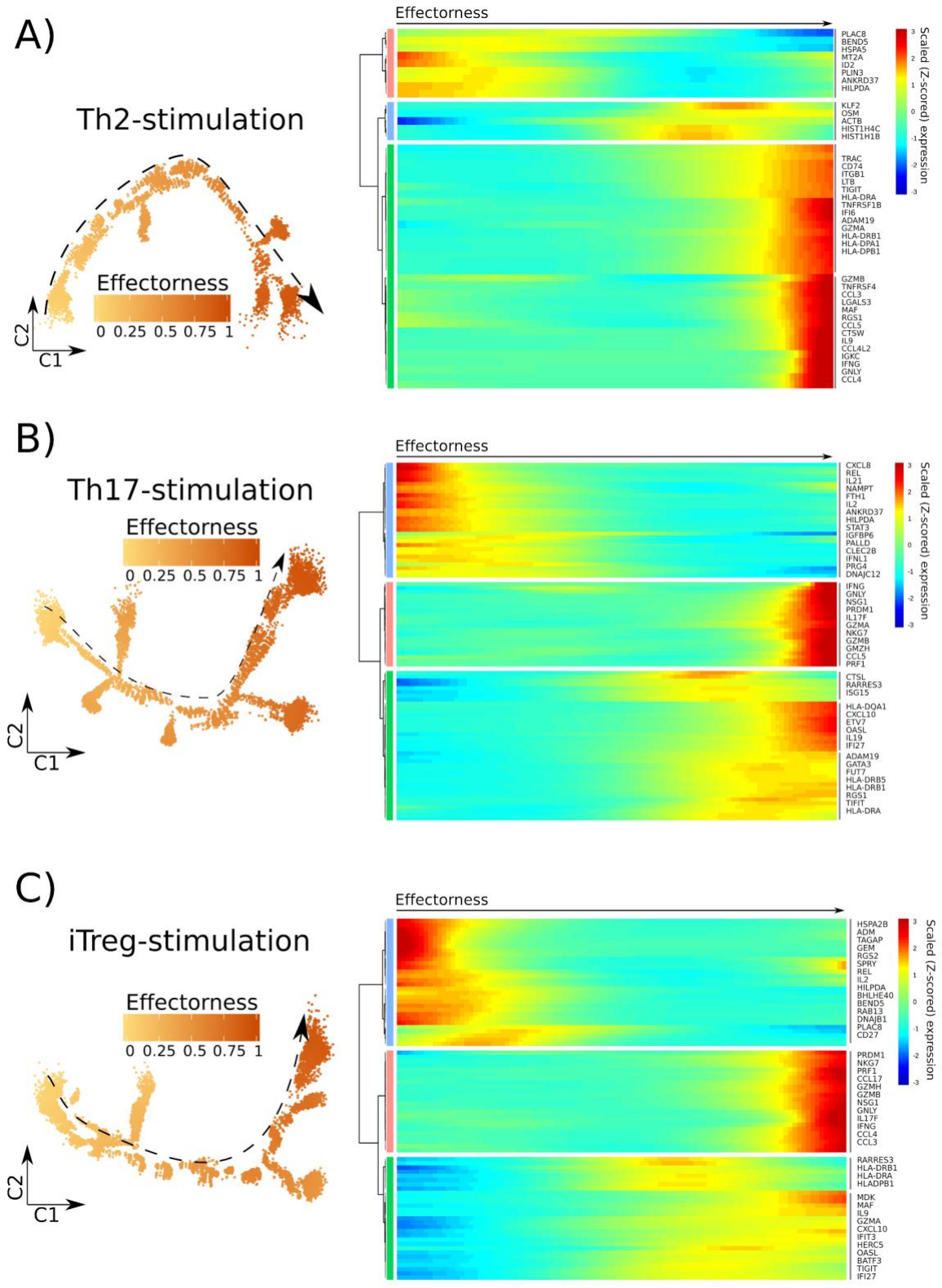

**Supplementary Figure 3.4**. **Single-cell expression of signature genes and markers in stimulated CD4+ T cells.** Percentage of cells uniquely assigned to one individual (left panel), unassigned because of low posterior probability (central panel) or containing DNA from more than one genotype (doublets, right panel). Each dot is one sample corresponding to a pool of four biologically independent samples.



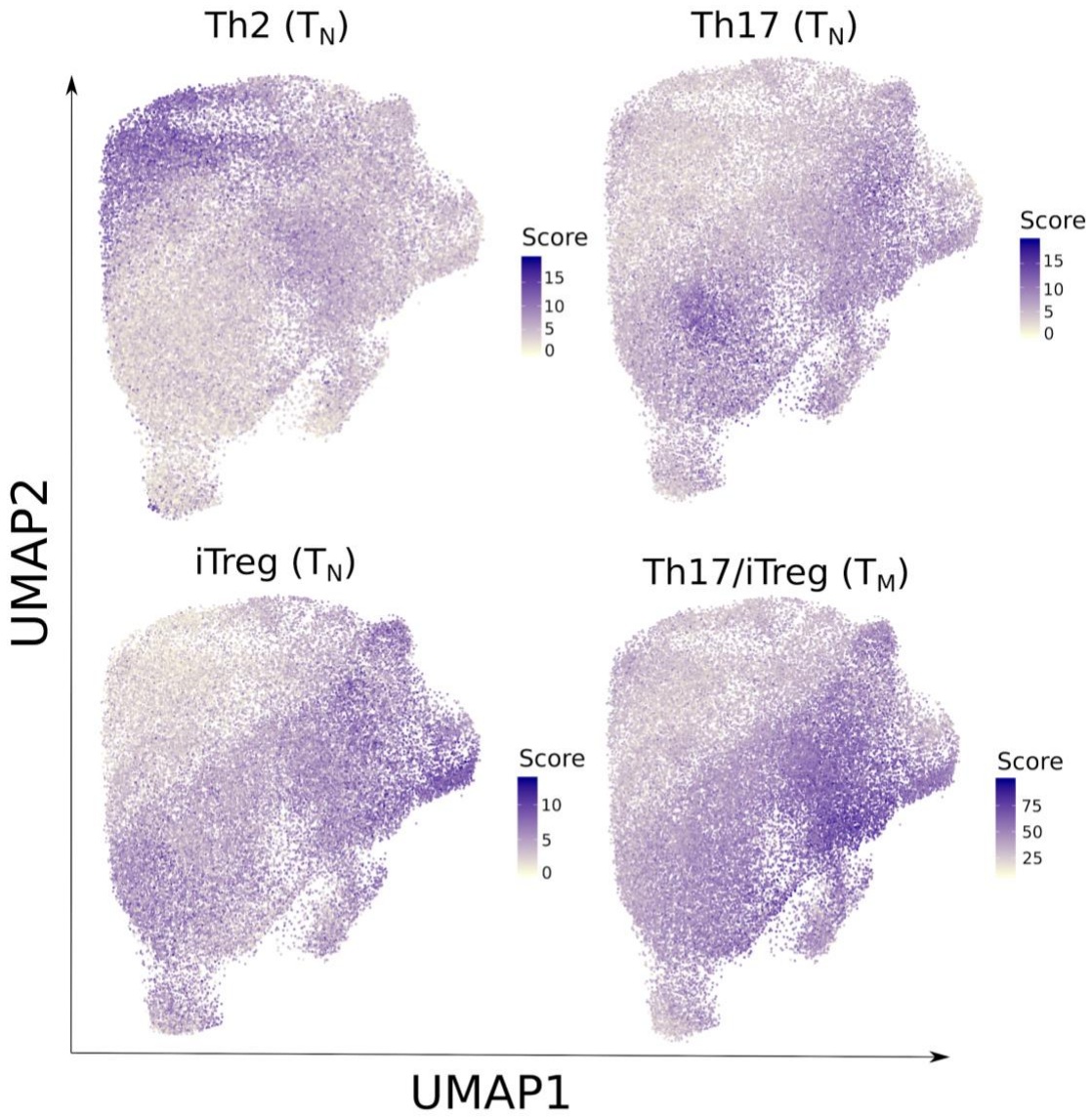

**Supplementary Figure 3.5**. **Integration of effectorness with T cell clonality** Re-analysis of public scRNA-seq data of CD4+ T cells obtained from colorectal cancer patients (**Methods**) **A)** UMAP embedding of CD4+ T cells color coded by tissue of origin (left panel) or cluster label (right panel). **B)** Cells ordered in a branched pseudotime trajectory and color coded by cluster label (top panel) or effectorness (bottom panel). **C)** Heatmap of the most variable genes along the pseudotime trajectory. Colors correspond to the scaled (Z-scored) expression of each gene in each cell. **D)** Two measurements of TCR clonotype diversity (fraction of unique clones in the left panel and expansion index in the right panel) were calculated for T cells at different ranges of effectorness using a sliding window approach (**Methods**). **E)** TCR sequences clustered by the predicted pMHC specificity of the CDR3 regions using GLIPH. Nodes represent independent TCR sequences, node size is proportional to the number of cells containing the sequence. Edges are drawn between TCR sequences predicted to recognize the same pMHC complex. Only cells with effectorness > 0.7 were included in this graph.



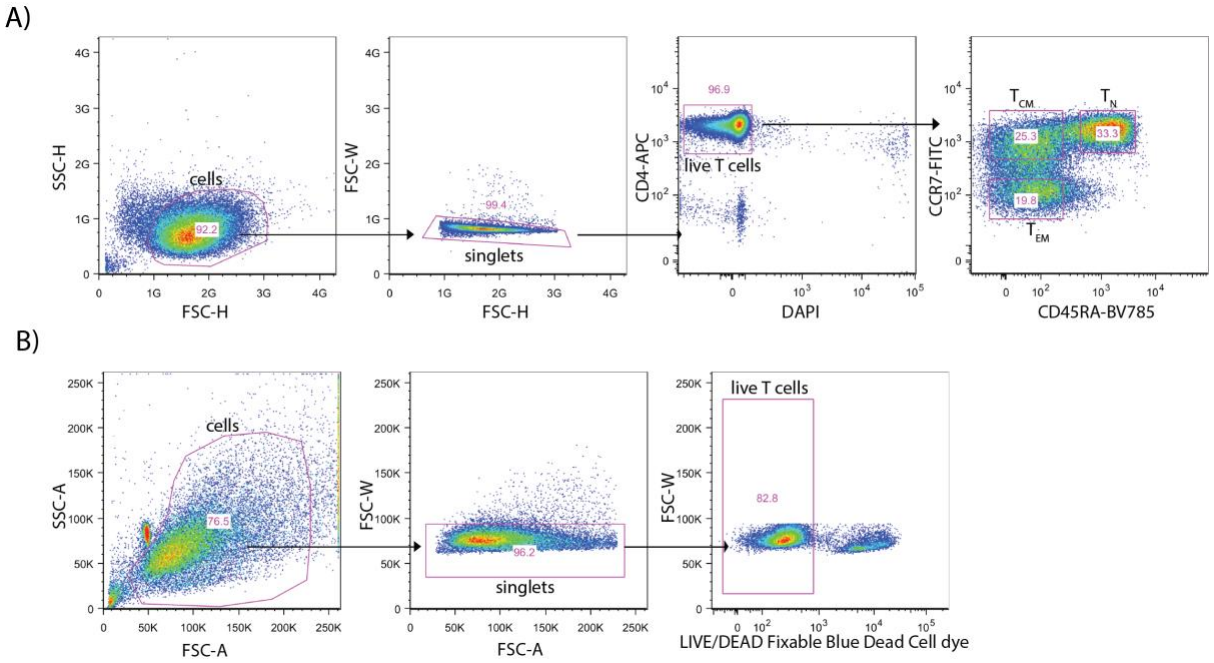

**Supplementary Figure 3.6**. **Single-cell expression of $T_{CM}$ markers in Th0-stimulated CD4+ T cells.** **A)** UMAP embedding of Th0-stimulated $T_N$ and $T_M$ cells. Cells were color coded by either **A)** their cell type ($T_N$ and $T_M$ cells) or **B)** their expression levels of *SELL* (left panel) and *CCR7* (right panel).



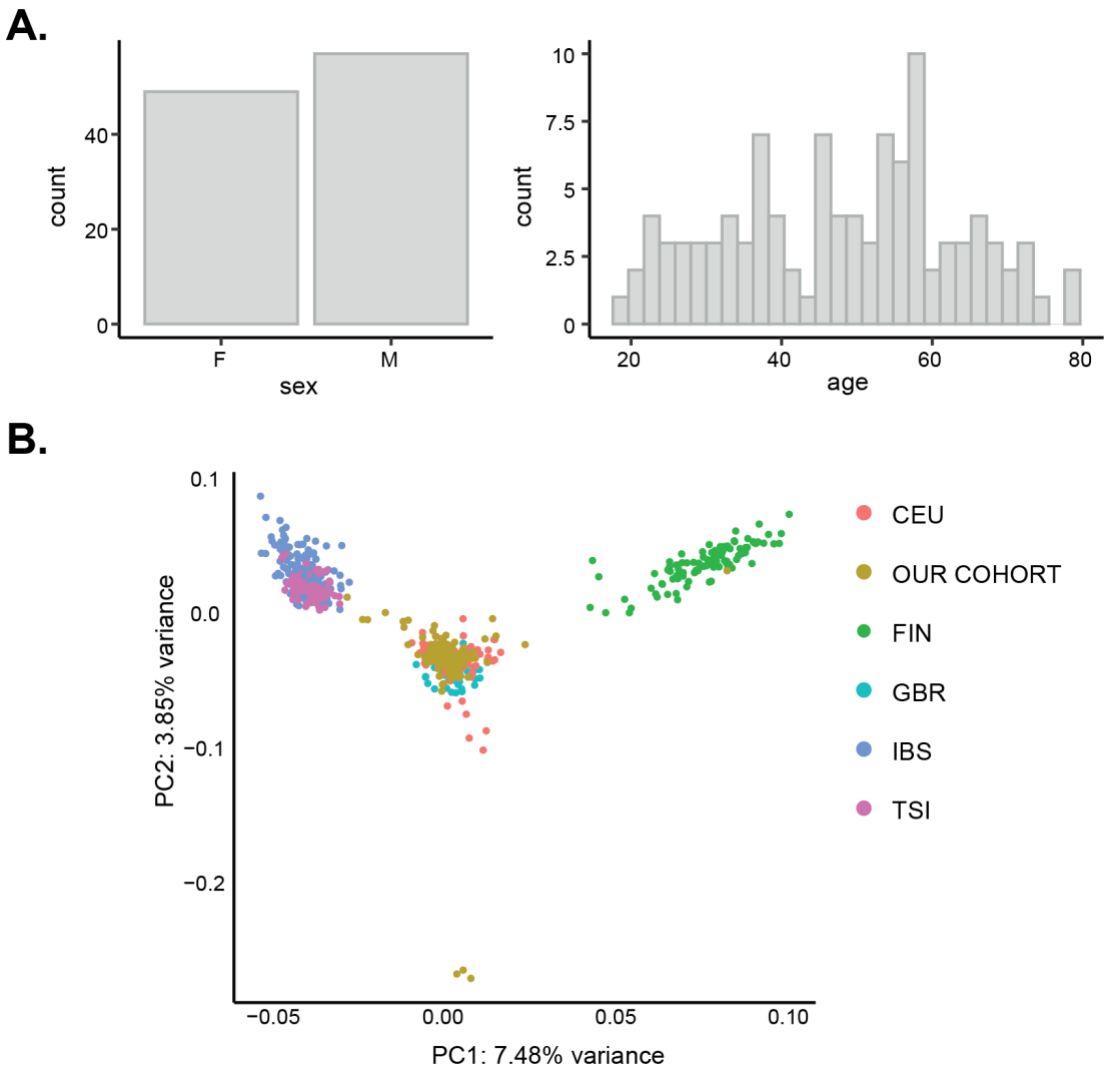

**Supplementary Figure 3.7**. **Cytokine-specific pseudotime trajectories.** $T_N$ and $T_M$ cells ordered in branched pseudotime trajectories. Plots are colored by pseudotime value (right panels) and accompanied by a heatmap of the most variable genes along the pseudotime trajectories (right panels). Heatmap colors correspond to the scaled (Z-scored) expression of each gene in each cell. Panels correspond to four independent trajectories from **A)** Th2, **B)** Th17 and **C)** iTreg-stimulation. Labels were added to a number of example genes.



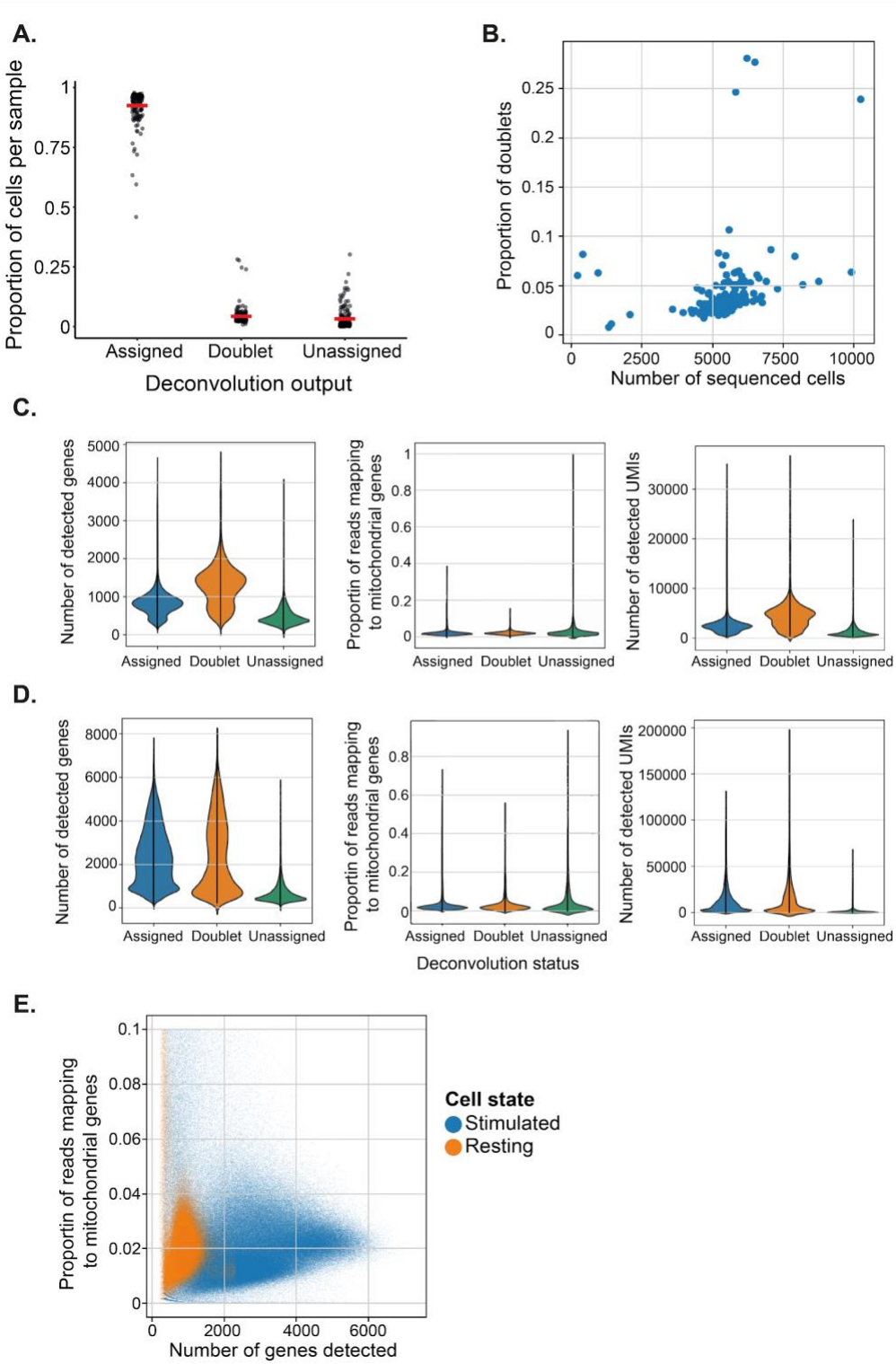

**Supplementary Figure 3.8**. **Single-cell expression of cytokine-specific signature genes in stimulated CD4+ T cells.** UMAP embedding of stimulated $T_N$ and $T_M$ cells from all cytokine conditions. Colors represent the average expression of all the genes in the cell state specific signatures defined from RNA-seq and proteomics data in each cell.



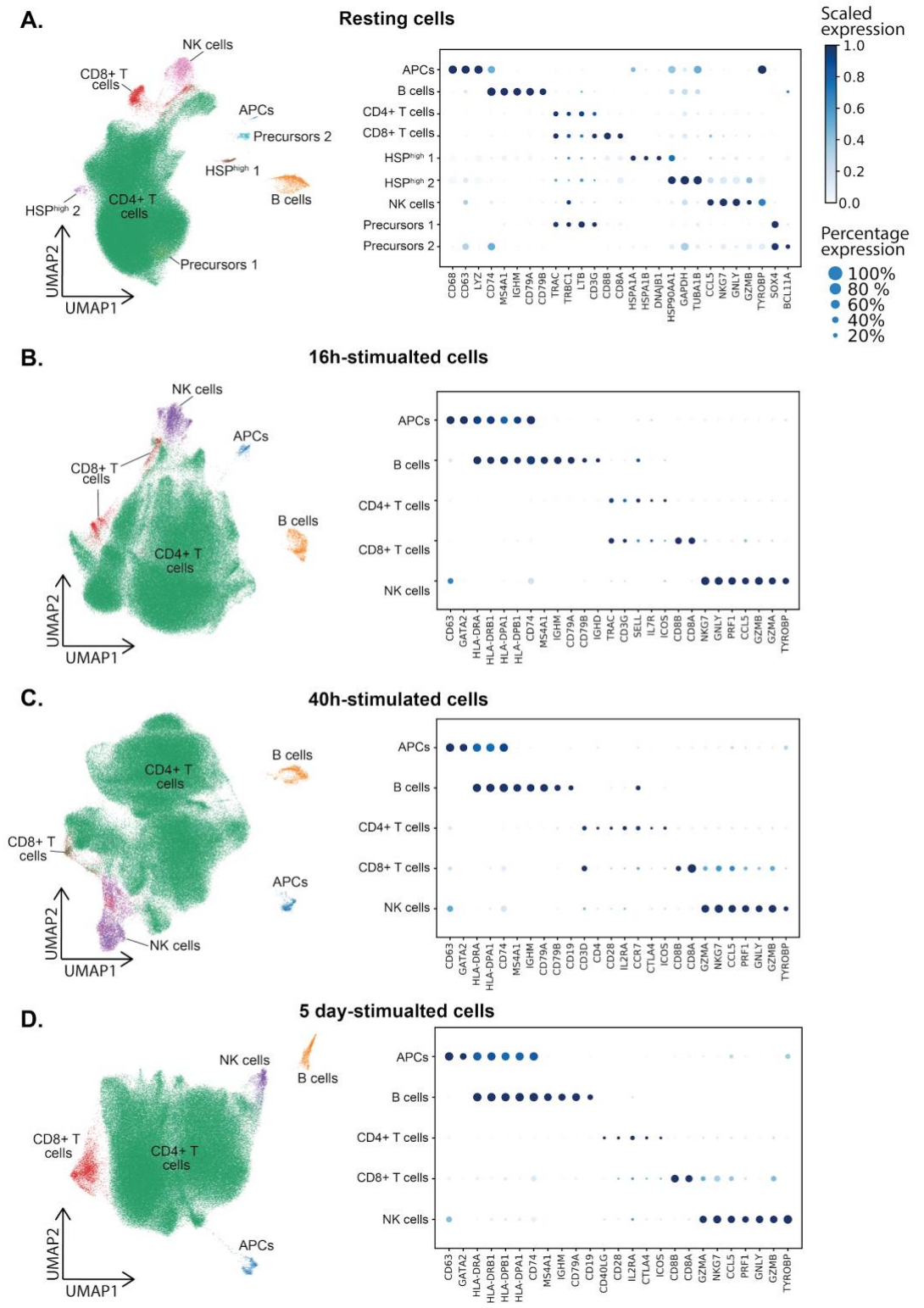

**Supplementary Figure 3.9**. **Sorting and gating strategy for cytokine staining. A)** CD4+CCR7+CD45RA+ ($T_N$), CD4+CCR7+CD45RA- ($T_{CM}$) and CD4+CCR7-CD45RA- ($T_{EM}$) cells were isolated from CD4+ T cells via FACS using a MoFlo XDP cell sorter. Representative flow cytometry plots of six biologically independent samples. Sorting strategy for the data presented Figure 6D. **B)** Representative flow cytometry plots of six biologically independent samples show the gating strategy for live cell isolation after restimulation. Gating strategy for the cytokine staining results presented in Figure 6D.



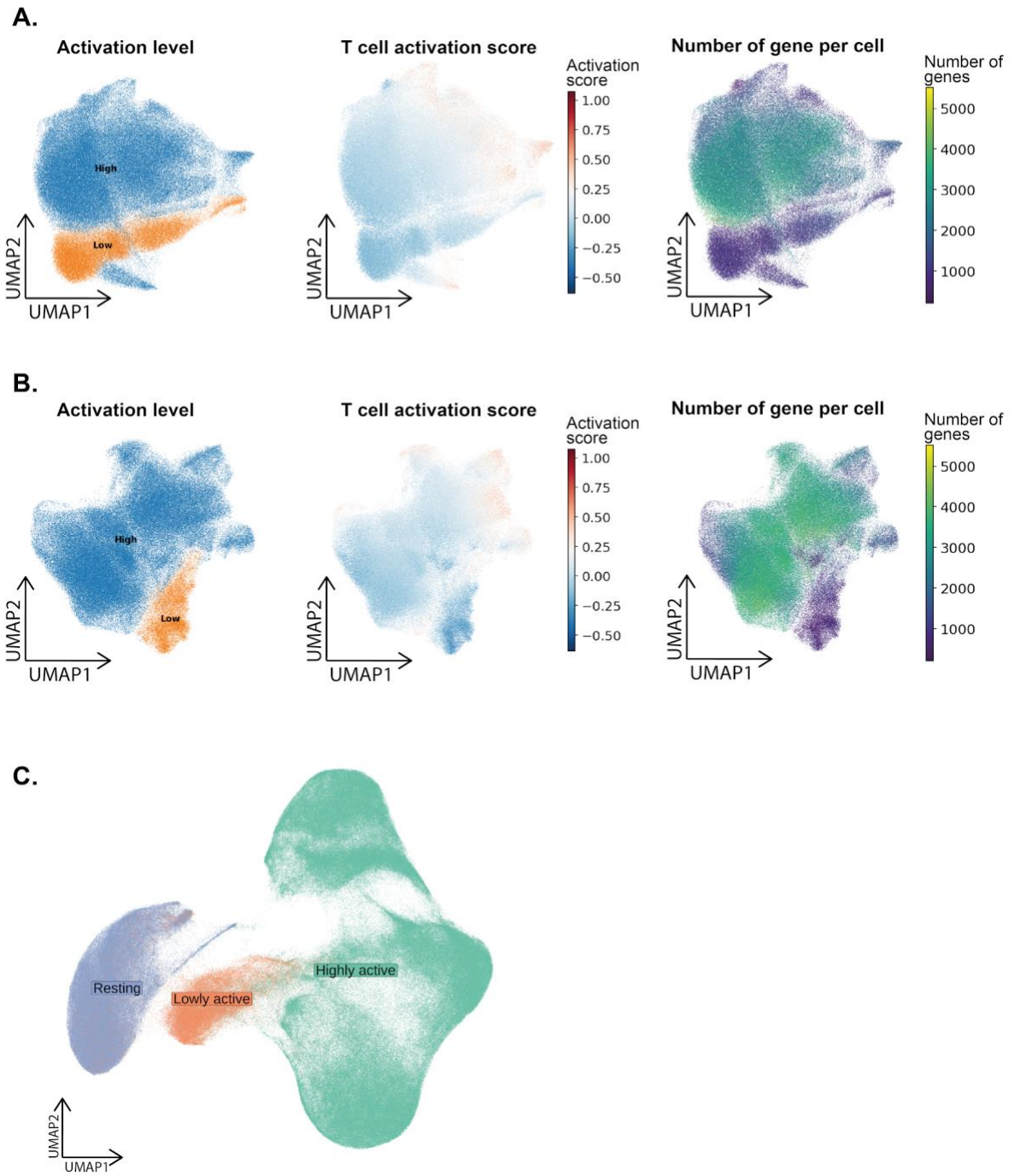

**Supplementary Figure 4.1. Demographic characteristics of our cohort. A)** Distriution of reported sex and age for all individuals in the cohort. **B)** Genetic ancestry was inferred for each individual in our cohort by comparing their genotypes (common genetic variants obtained from genotyping and imputation) with a panel of five well-defined European ancestries in PCA space. The populations in this panel were derived from the 1000G project (**Methods**) and corresponded to Utah residents with Northern and Western European ancestry (CEU), Finnish in Finland (FIN), British in England and Scortland (GBR), Iberian Population in Spain (IBS), and Toscani in Italy (TSI).



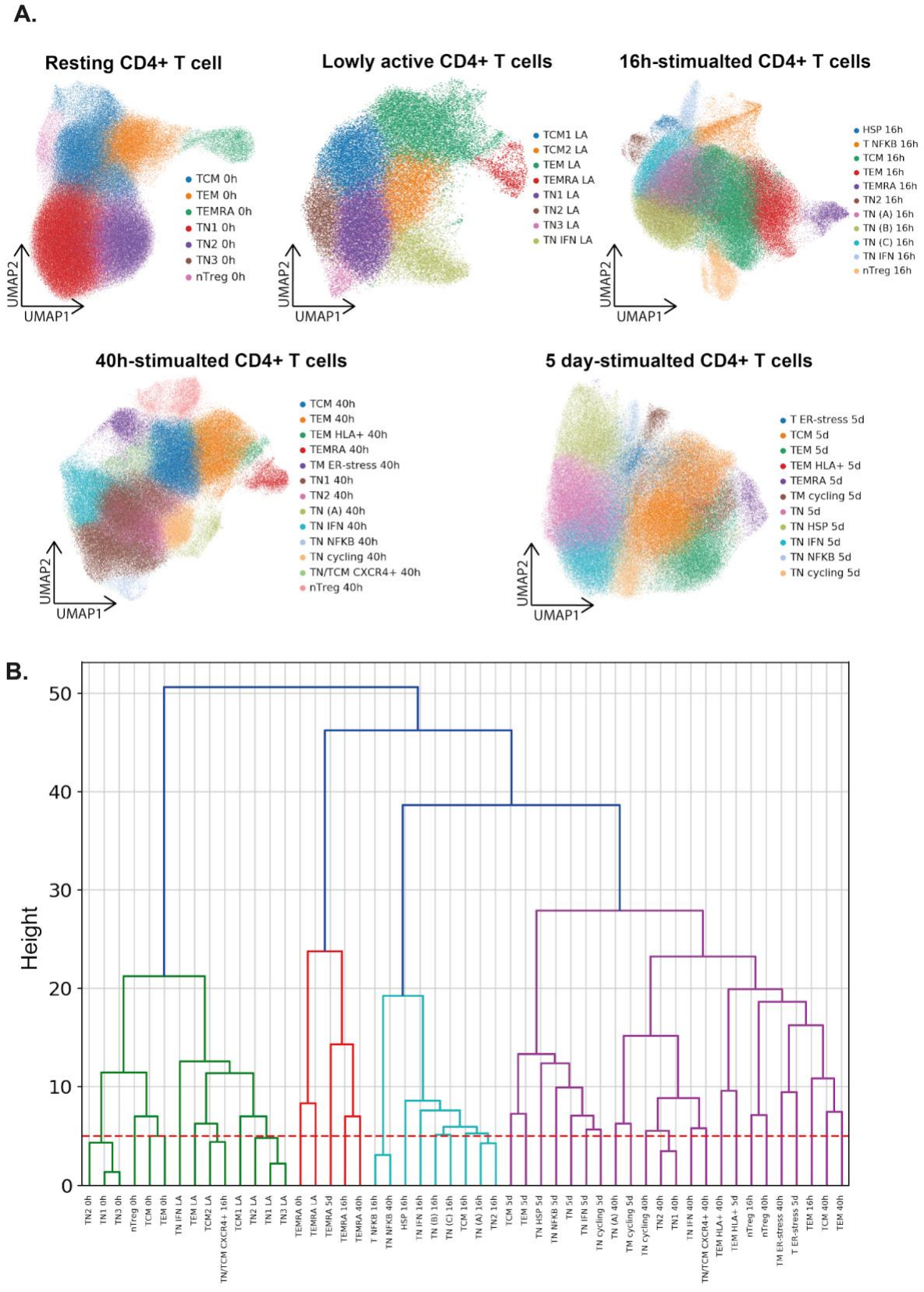

**Supplementary Figure 4.2. Quality metrics of the scRNA-seq data. A)** Proportion of cells per sample which are either confidently assigned to one individual (assigned), contain RNA from more than one individual (doublets) or are unassigned because of a low posterior probability (unassigned). **B)** Dot plot showing the relationship between number of cells per inlet and number of doublets. **C-D)** Violin plots showing the number of detected genes (left), proportion of reads mapping to mitochondrial genes (central) and number of detected UMIs (right) in assigned cells, doublets and unsigned cells. These metrics are shown separately for resting (**C**) and stimulated (**D**) samples. **E)** Dot plot showing the number of detected genes and proportion of reads mapping to mitochondrial genes in each cell. Colours indicate resting (orange) and stimulated (blue) cells.




**Supplementary Figure 4.3. Identification and removal of cellular contaminations in the scRNA-seq data set. A)** UMAP embeddings for cells obtained at each of the four time points. Colours represent cell populations derived from unsupervised clustering. Green cells are CD4+ T cells. Other colours represent contaminations that were removed from subsequent analyses. **A)** Resting cells. **B)** 16h-stimulated cells. **C)** 40h-stimulated cells. **D)** 5 day-stimulated cells.

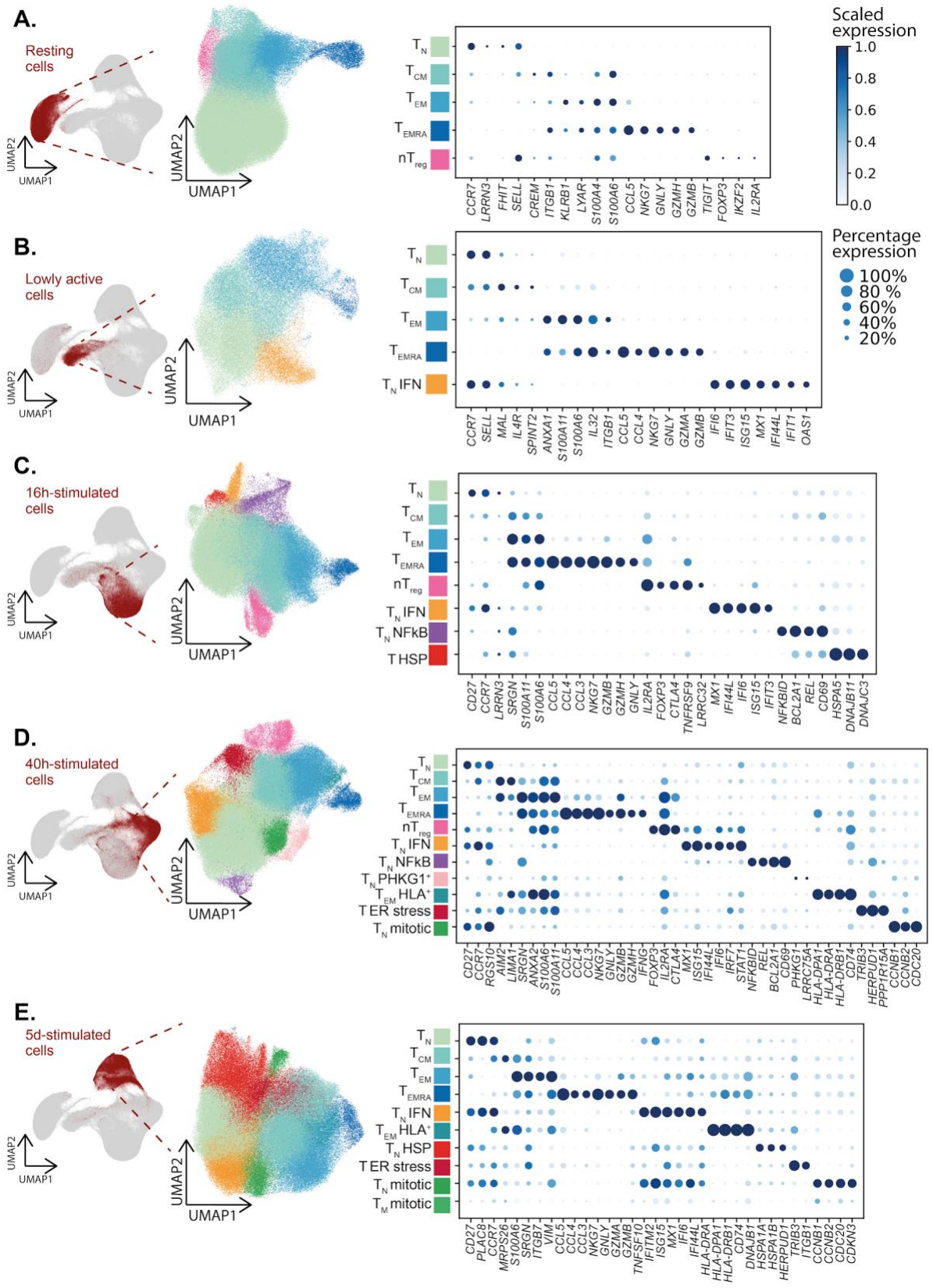



**Supplementary Figure 4.4. Identification of a population of lowly active CD4+ T cells. A-B)** UMAP embedding of all CD4+ T cells sampled at 16h (**A**) and 40h (**B**) after activation. Each dot is a cell, with colours indicating either the lowly active population (orange; left panel), the scaled average expression of a set of T cell activation markers from bulk RNA-seq (central panel), or the number of genes detected per cell (right panel). **C)** Location of the lowly active population of cells in a UMAP embedding containing CD4+ T cells from all time points in the study.



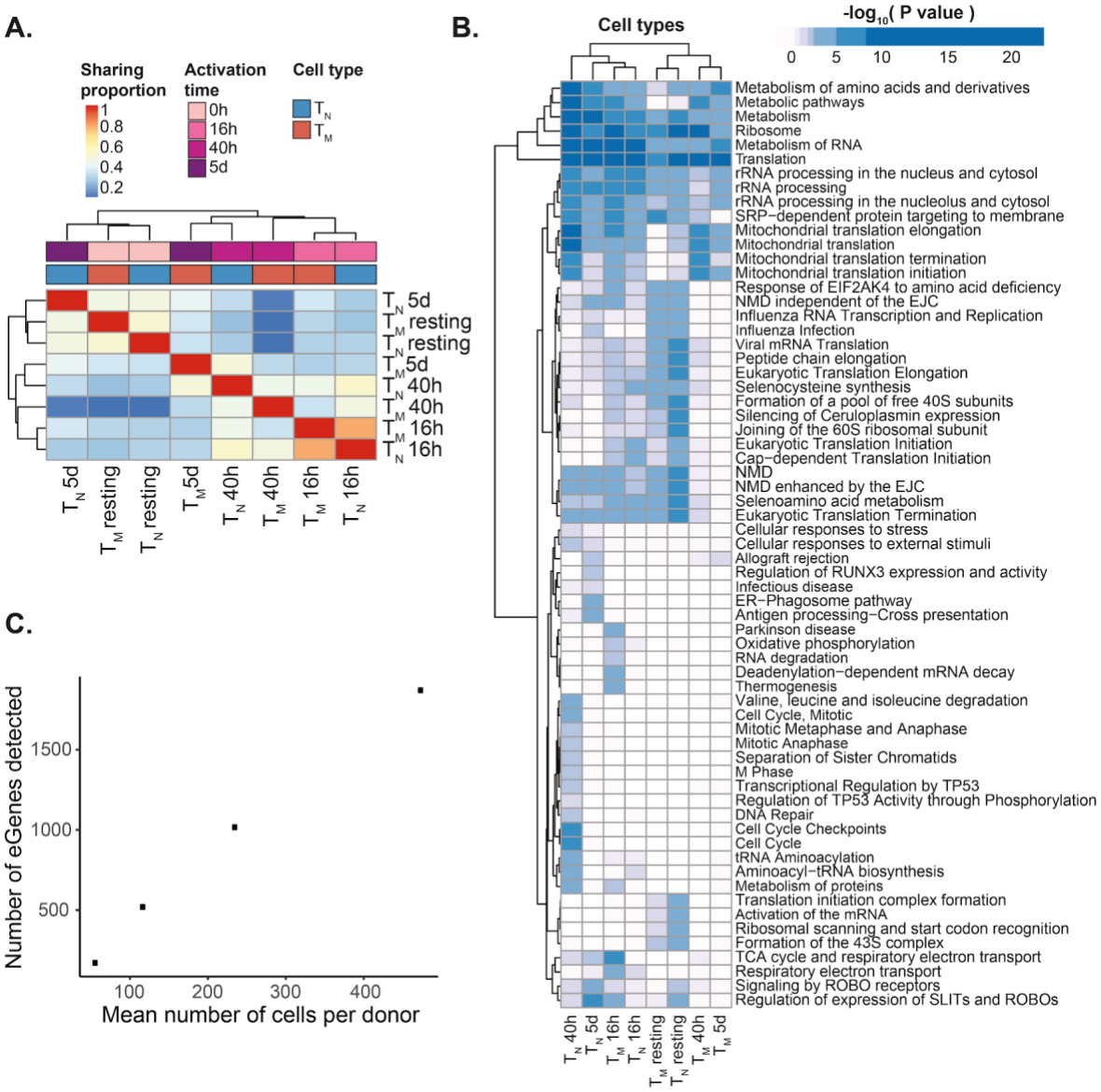

**Supplementary Figure 4.5. Cluster merging strategy. A)** Unsupervised clustering was applied independently to the five cell groups of cells identified in the study (resting, lowly active, 16 hours, 40

hours, and five days). Colours represent cell populations derived from unsupervised clustering. This resulted in 51 cell clusters. **B)** Cluster similarity was assessed by performing PCA and estimating the Euclidean distance between pairs of clusters based on the first 100 principal components. Below the dotted red line are clusters with high levels of similarity that were merged together This resulted in 38 distinct groups of cells.

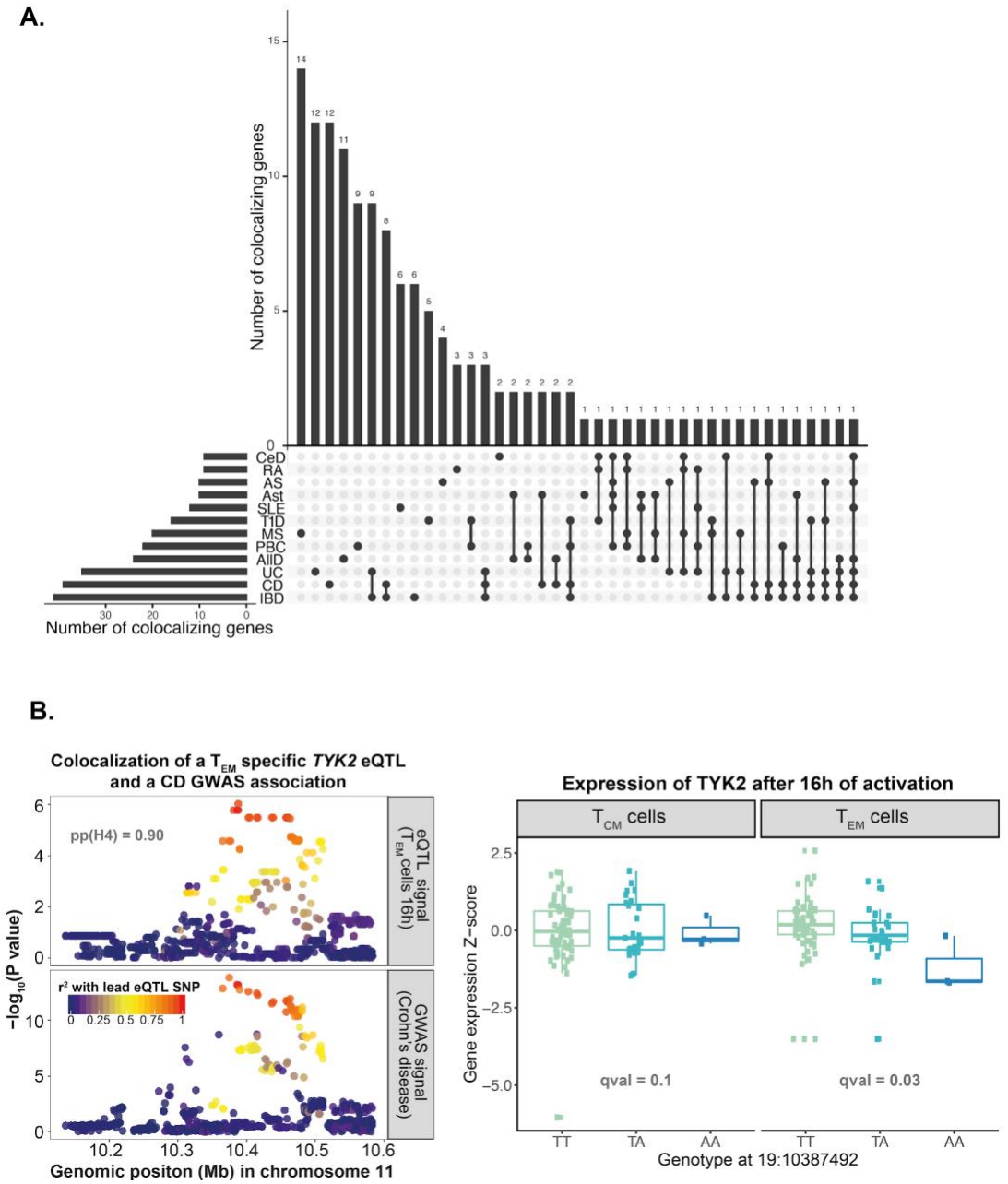



**Supplementary Figure 4.6. CD4+ T cell subpopulation markers.** Cell subpopulations and marker genes detected independently for CD4+ T cells belonging to each of five broad cell groups: resting (**A**), lowly active (**B**), 16h-stimulated (**C**), 40h-stimulated (**D**), and 5 day-stimulated (**E**) cells. UMAP

embeddings (left panels) show each cell as one dot, with colours indicating cell subpopulations derived from unsupervised clustering and cluster merging (38 subpopulations in total). Equivalent subpopulations detected at more than one time point are indicated in the same colour. Dot plots (right panels) show the top marker genes identified for each subpopulation. Shades of blue indicate the scaled mean expression of a gene in each subpopulation, while dot sizes correspond to the proportion of cells in the subpopulation that express the gene.




**Supplementary Figure 4.7. Characteristics of CD4+ T cell eQTLs. A)** eQTL sharing was assessed by comparing the effect sizes and directions of eQTLs across cell types (naive and memory cells) and time points using mashR (**Methods**). This heatmap indicates the proportion of eQTLs shared by sign and magnitude between cell type and time point combinations. **B)** Heatmap of biological pathways significantly enriched in eGenes from different cell types and time points. Shades of blue indicate log-transformed estimated enrichment P values. **C)** The relationship between cell numbers and statistical power was assessed by randomly subsampling increasingly lower numbers of cells from the 40h-stimulated $T_{CM}$ subpopulation and repeating the eQTL mapping step. At each iteration, the number of significant eGenes detected was tabulated. This plot shows the number of significant eGenes detected when different numbers of cells per individual are used to estimate the mean gene expression.




**Supplementary Figure 4.8. eGenes colocalizing at immune disease loci. A)** Number of colocalizing genes shared between different immune traits. Bar plots indicate the number of genes in each set. **B)** Locus plots for a colocalization between a *TYK2* eQTL specific to 16h-stimulated $T_{EM}$ cells (top left panel) and a GWAS signal for Crohn's disease (bottom left panel). Each dot corresponds to a genetic variant, with colours indicating their LD with the lead eQTL variant. Box plots (right panel) indicate the expression level of *TYK2* in 16h-stimulated $T_{EM}$ cells isolated from different individuals, stratified by genotype. Each dot corresponds to a measurement from a different individual. Trait abbreviations: Celiac disease (CeD), Rheumatoid arthritis (RA), Ankylosing spondylitis (AS), Asthma (Ast), Systemic Lupus Erythematosus (SLE), Type 1 diabetes (T1D), Multiple sclerosis (MS), Primary biliary cirrhosis (PBC), Allergic disease (AllD), Ulcerative colitis (UC), Crohn's disease (CD), Inflammatory bowel disease (IBD).

# Appendix B: Supplementary tables

All the supplementary tables discussed in this dissertation are publicly available online at:

https://drive.google.com/drive/folders/1muYng_OSIU3n1IuFXjI1zOSFj-gSMRWF?usp=sharing

Hyperlinks to each individual table are provided throughout the text whenever these tables are referenced.