

Recessive and rare variant effects on common diseases and the immune cell transcriptome



Heng Teng Hiang

Peterhouse

University of Cambridge

September 2024

This thesis is submitted for the degree of Doctor of Philosophy

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted, or, is being concurrently submitted, for any degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Heng Teng Hiang

September 2024

Abstract

Studying genetic variation provides insight into the biology of phenotypes, improves risk prediction, and contributes to developing better therapeutics. Different types of variant effects contribute to the heritability of a trait, and in this thesis I focused on recessive effects and rare, loss-of-function variant (LoF) effects. Genes & Health (G&H) is a cohort of ~50,000 British South Asians with elevated rates of consanguinity and consequently, increased homozygosity, therefore making it a suitable cohort to study both recessive effects and LoFs.

Little is known about recessive effects on common diseases. Therefore, we searched for common and rare variants that had a recessive effect on common diseases in G&H. We imputed variants into 44,190 genotyped individuals, using two imputation panels: a set of 4,982 exomes from within the cohort, and the TOPMed-r2 panel. We then performed association testing with 898 common diseases, identifying 207 loci that reached standard genome-wide significance ($p\text{-value} \leq 5 \times 10^{-8}$) under the recessive model and were more significant than under the additive model. Of these, about 70% demonstrated a nominally-significant ($p\text{-value} \leq 0.05$) dominance deviation $p\text{-value}$. I discuss several findings in detail in my thesis.

Among subtypes of variants with recessive effects, homozygous LoFs are of interest as they can function like naturally-occurring gene ‘knockouts’. We performed a recall-by-genotype of 114 individuals in G&H with rare, putative LoFs affecting immune genes, for single-cell RNA-sequencing (scRNAseq) of peripheral blood mononuclear cells (PBMCs). Cell-type-specific LoF effects on transcription have yet to be studied at scale, and in my thesis I discuss the three phases to building this unique dataset: firstly, the curation of true LoFs from bioinformatically-called putative LoFs (which are enriched for errors), secondly, LoF effects on the immune transcriptome at rest, and thirdly, LoF effects on the immune transcriptome after stimulation (by interferon alpha for innate immune processes, and by anti-CD3 for adaptive immune processes). We showed that direct knockdown effects of LoFs on the genes they affect can be detected in RNA expression, and that LoFs were associated with pathway derangements. These effects were cell-type- and context-specific.

In summary, during my PhD, I leveraged the increased homozygosity within G&H to first demonstrate widespread recessive effects of common and rare variants on common diseases, and secondly, to profile the direct and indirect effects of rare LoFs, including homozygous LoFs, on the immune transcriptome.

Acknowledgements

Completing this work has been a joy. I would first like to express my gratitude to my supervisors, Hilary C. Martin and Gosia Trynka. If the PhD is the beginning of my scientific career, it has made me excited for the road ahead, and I have the both of you to thank for that.

To Hilary, thank you for taking me under your wing four years ago and teaching me nearly everything I know about statistical genetics. Your guidance has been formative and your leadership inspiring. Thank you for challenging me to be rigorous, but also creative. I will always cherish the weekly meetings we had in your office talking about science, but more so, I think I will remember the lab lunches, the picnics and the barbecues where we talked about anything.

To Gosia, your enthusiasm about immunology is infectious (pun intended). Thank you for coordinating the many moving parts of my second PhD project, from costings to wet lab logistics and dry lab analysis. But more importantly, thank you for pushing me to ask bigger questions, and for the invaluable guidance you provided in the process of answering those questions.

Thank you both for the wonderful opportunity to work with you, for your dedication to developing me into a better scientist, and for your time and patience to listen to my problems, no matter how big or small they were.

I would also like to thank the members of the Martin and Trynka labs for creating such a wonderful environment to work in. Thank you all for the stimulating conversations and for always being generous with your knowledge. I have made wonderful friendships along the way. In particular, I thank Qinqin Huang, Anna Lorenc, Klaudia Walter, Tarran Rupall, Bess Chau, Georgios Kalantzis, Daniel Malawsky and Carla Jones, for their guidance during key periods of my PhD and important contributions to my research.

I would like to thank Carl Anderson and my university supervisor, William Astle, for being a part of my thesis committee and helping to shape the direction of my PhD. I would also like to thank my rotation supervisors, Matthew Hurles and Sam Behjati, for the great start to my PhD journey by introducing me to the exciting science at the Sanger Institute. I am also grateful to Annabel Smith, Christina Hedberg-Delouka and Catherine Ingle at the Sanger Graduate Office for their support. I also would like to thank Human Genetics Informatics and the administrative team in the Human Genetics Department for their help in getting both my research projects off the ground and running.

The research in this thesis would not be possible without the participants of Genes & Health, and I thank them all for their time and contribution. I would also like to thank David A. van Heel and the Genes & Health Research Team for all the important work they do to build this amazing cohort. I also thank Nicole Soranzo and the Cardinal Team for graciously accommodating me and my recall-by-genotype study alongside the Cardinal project. I also would like to express my gratitude to my collaborators at FinnGen and the Garvan Institute of Medical Research.

I would also like to thank the Agency for Science, Technology and Research for awarding me the National Science Scholarship to pursue this PhD. I thank my mentors in Singapore, Ren Ee Chee and Isaac Liu, for motivating me to pursue a career in biomedical research.

I would like to thank Goh Zhang-He for his patience and support throughout my PhD. Lastly, I thank my parents Tan Lay Ngoh and Heng Siyau Teck, and my sister Heng Teng Wei, for being there from the beginning. You have always encouraged me to pursue my dreams, whatever they may be. Your hard work and dedication to others have always been an inspiration. I would not be where I am today without your unconditional love and support.

Contents

DECLARATION.....	2
ABSTRACT	4
ACKNOWLEDGEMENTS.....	6
CONTENTS	8
INTRODUCTION.....	12
THE POTENTIAL OF GENETIC VARIATION FOR UNDERSTANDING THE BIOLOGY OF TRAITS AND DISEASES	12
DIFFERENT TYPES OF GENETIC EFFECTS CONTRIBUTE TO HERITABILITY	15
<i>Figure-1.1</i>	19
THE TECHNOLOGIES AND DATASETS AVAILABLE IN THE FIELD	20
GENES & HEALTH AS THE COHORT OF CHOICE IN THIS THESIS	24
OUTLINE OF THE THESIS	26
RECESSIVE EFFECTS IN A COHORT OF 44,000 BRITISH PAKISTANIS AND BANGLADESHIS.....	28
OVERVIEW	28
INTRODUCTION.....	29
<i>Figure-2.1</i>	31
<i>Figure-2.2</i>	33
<i>Figure-2.3</i>	39
METHODS	42
<i>Table-2.1</i>	44
<i>Figure-2.4</i>	48
<i>Figure-2.5</i>	51
<i>Figure-2.6</i>	53
<i>Table-2.2</i>	58
<i>Figure-2.7</i>	60
<i>Table-2.3</i>	63
RESULTS.....	66
<i>Table-2.4</i>	67
<i>Figure-2.8</i>	68
<i>Figure-2.9</i>	70
<i>Figure-2.10</i>	72
<i>Table-2.5</i>	73
<i>Figure-2.11</i>	75
<i>Figure-2.12</i>	77
<i>Figure-2.13</i>	79

<i>Table-2.6</i>	80
<i>Figure-2.14</i>	82
<i>Figure-2.15</i>	84
<i>Figure-2.16</i>	86
<i>Figure-2.17</i>	90
<i>Table-2.7</i>	92
DISCUSSION	93
<i>Figure-2.18</i>	97
<i>Table-2.8</i>	100
CONCLUSION	105
ACKNOWLEDGEMENTS FOR THIS CHAPTER	106

IDENTIFYING HUMAN GENE KNOCKOUTS AND DETECTING THEIR EFFECTS ON THE BLOOD

TRANSCRIPTOME	108
OVERVIEW	108
INTRODUCTION.....	109
ACKNOWLEDGEMENTS FOR THIS CHAPTER	118
METHODS	119
<i>Figure-3.1</i>	120
<i>Figure-3.2</i>	125
<i>Figure-3.3</i>	127
RESULTS.....	133
<i>Figure-3.4</i>	135
<i>Figure-3.5</i>	137
<i>Figure-3.6</i>	141
<i>Figure-3.7</i>	144
<i>Table-3.1</i>	148
<i>Table-3.2</i>	150
<i>Figure-3.8</i>	151
<i>Figure-3.9</i>	154
<i>Figure-3.10</i>	157
<i>Figure-3.11</i>	159
<i>Table-3.3</i>	162
<i>Figure-3.12</i>	164
<i>Figure-3.13</i>	168
<i>Figure-3.14</i>	172
DISCUSSION	173
CONCLUSION	180

EFFECTS OF NATURALLY-OCCURRING IMMUNE GENE KNOCKOUTS ON THE RESTING AND STIMULATED IMMUNE TRANSCRIPTOME.....	182
OVERVIEW.....	182
INTRODUCTION.....	183
<i>Figure-4.1</i>	184
<i>Table-4.1</i>	188
METHODS	193
<i>Figure-4.2</i>	194
<i>Figure-4.3</i>	196
<i>Figure-4.4</i>	201
<i>Figure-4.5</i>	204
<i>Figure-4.6</i>	208
<i>Figure-4.7</i>	210
<i>Figure-4.8</i>	212
<i>Figure-4.9</i>	213
<i>Figure-4.10</i>	214
<i>Figure-4.11</i>	216
<i>Figure-4.12</i>	219
<i>Figure-4.13</i>	222
RESULTS.....	224
<i>Figure-4.14</i>	225
<i>Figure-4.15</i>	227
<i>Figure-4.16</i>	230
<i>Figure-4.17</i>	234
<i>Figure-4.18</i>	237
<i>Figure-4.19A</i>	239
<i>Figure-4.19B</i>	240
<i>Figure-4.19C</i>	241
<i>Figure-4.20</i>	242
<i>Figure-4.21</i>	243
<i>Figure-4.22</i>	245
<i>Figure-4.23</i>	246
<i>Figure-4.24</i>	248
<i>Figure-4.25</i>	249
<i>Figure-4.26</i>	251
<i>Figure-4.27</i>	252
<i>Figure-4.28</i>	254
<i>Figure-4.29</i>	256
<i>Figure-4.30</i>	258

DISCUSSION	259
CONCLUSION	264
ACKNOWLEDGEMENTS FOR THIS CHAPTER	265
DISCUSSION.....	266
SUMMARY OF FINDINGS	266
WIDER IMPLICATIONS OF THESE FINDINGS	269
LIMITATIONS OF CURRENT APPROACHES	271
FUTURE DIRECTIONS	272
<i>Figure-5.1</i>	273
APPENDICES.....	276
APPENDIX A - SUPPLEMENTARY DATA FOR CHAPTER 2	277
<i>Supplementary-figure-2.1</i>	277
<i>Supplementary-figure-2.2</i>	278
<i>Supplementary-table-2.1</i>	279
<i>Supplementary-table-2.2</i>	289
APPENDIX B - SUPPLEMENTARY DATA FOR CHAPTER 2	293
<i>Supplementary-figure-3.1</i>	293
<i>Supplementary-figure-3.2</i>	294
<i>Supplementary-figure-3.3</i>	295
<i>Supplementary-figure-3.4</i>	296
<i>Supplementary-figure-3.5</i>	297
<i>Supplementary-figure-3.6</i>	298
<i>Supplementary-figure-3.7</i>	299
<i>Supplementary-table-3.1</i>	300
<i>Supplementary-table-3.2</i>	301
<i>Supplementary-table-3.3</i>	304
APPENDIX C - SUPPLEMENTARY DATA FOR CHAPTER 4	318
<i>Supplementary-figure-4.1</i>	318
APPENDIX D - GENERAL SUPPLEMENTARY DATA	319
BIBLIOGRAPHY.....	320

Introduction

The potential of genetic variation for understanding the biology of traits and diseases

The first references of the human genome were published in 2001 (Lander et al., 2001; Venter et al., 2001), with the goal of understanding human biology through interpreting the genetic code. At that point, both references were composed of genetic data from a small number of individuals, and information on individual variation was not a focus. However, within an individual, there can be about three to four million variants relative to a reference (Frazer et al., 2009; Jackson et al., 2018). This enormous genetic diversity contributes to the phenotypic diversity we see in humans. One of the ways to interpret the genetic code is to look for such variations among individuals, and relate them to differences in observable phenotypes. Therefore, since then, many large-scale cohorts have been developed, where linked genetic and phenotypic data of hundreds or thousands of individuals are collected together (Bastarache et al., 2023).

Studying genetic variation in this manner has allowed us to observe the natural effects of different alleles on phenotypes, providing insight into the biological processes underpinning traits and diseases (Frazer et al., 2009), and allowing the discovery of potential drug targets (Barrett et al., 2015; Plenge et al., 2013).

Developing a genetic basis for disease

A combination of genetic and environmental factors dictate the presentation of most human traits and diseases (Hunter, 2005). A major aspect of understanding the genetic basis for disease is interpreting the consequence of genetic variation, such as finding out whether a variant is benign or pathogenic when considering Mendelian disorders, and quantifying the effect size of a variant on a phenotype.

Rare diseases often have a large genetic contribution, with approximately eighty percent having a genetic cause (Jackson et al., 2018; C. F. Wright et al., 2018). Many of these conditions are monogenic, or affected by pathogenic variants in a single gene with a large effect, and follow Mendelian inheritance patterns (Hamosh et al., 2005). Therefore, the interpretation of variant consequence and gene function is often quite direct, and was a focus of human genetics research even before the development of large-scale sequencing projects (Antonarakis & Beckmann, 2006).

In contrast, the analysis of the genetic basis of complex traits and common diseases is less straightforward. Two hypotheses in human genetics present an approach to studying the role of genetic variation: the common disease-common variant hypothesis (Pritchard & Cox, 2002), which suggests that a collection of common variants with small effect sizes contribute to the development of complex traits, and the rare variant hypothesis (Fearnhead et al., 2005), which suggests that variants with large effects become rare due to natural selection. Together the two hypotheses provide a framework for the classification and analysis of genetic variants in the context of complex traits and diseases (Bodmer & Bonilla, 2008; Frazer et al., 2009; Pritchard & Cox, 2002).

The effects of genetic variation on complex traits and common diseases have mainly been studied through genome wide association studies (GWAS) (Uffelmann et al., 2021; Visscher et al., 2012, 2017). (An introduction to GWAS algorithms is provided in Chapter 2.) GWAS, typically focusing on common (defined to have a minor allele frequency (MAF) > 5%) single-nucleotide polymorphisms (SNPs) (Cirulli & Goldstein, 2010), have successfully identified many loci with significant associations across numerous traits (Visscher et al., 2012, 2017). Supporting the common disease-common variant hypothesis, indeed most of these common variants have modest effect sizes (Cirulli & Goldstein, 2010; Gibson, 2012). In addition, the majority are noncoding, thus making it challenging to identify the target gene and mechanism of action. In contrast, rare variants (MAF <1%) tend to have larger effect sizes, particularly those that alter proteins, which are mechanistically easier to interpret (Gibson, 2012).

Both common and rare variants play a role in improving our management of diseases. Genetic testing and polygenic risk scores help in the screening, risk profiling and diagnosis of patients (Cirulli & Goldstein, 2010; Lambert et al., 2019; Truong et al., 2024). In addition, interpretable variant consequences can lead to a better understanding of gene function and regulation, potentially leading to new therapeutics (Trajanoska et al., 2023).

The role of genetics in drug discovery

About ninety percent of drug targets fail at clinical trials, particularly when targets are not causally linked to disease (Barrett et al., 2015; Plenge et al., 2013). Another contributor to the failure rate are the limits of preclinical models in predicting the safety and efficacy of a drug target in humans, particularly in the long term (Minikel et al., 2020; Paul et al., 2010). Naturally-occurring genetic variation can provide information about both the likely consequences of, and the likely ability of the body to accommodate, a drug target perturbation. Therefore, it is a powerful resource for drug target identification, particularly

loss-of-function variants (LoFs, more details below) (Minikel et al., 2020; Plenge et al., 2013). Genetic studies have led to many successful therapeutic approaches in rare, Mendelian disorders (Heilbron et al., 2021), such as Nusinersen, an oligonucleotide targeting the survival motor neuron 2 (*SMN2*) gene in the management of autosomal recessive infantile-onset spinal muscular atrophy (Finkel et al., 2017).

Rare variant discovery has also led to the discovery of therapeutic targets in common diseases. The most notable example is the observation of loss-of-function and gain-of-function mutations in proprotein convertase subtilisin/kexin type 9 (*PCSK9*) being associated with reduced and increased levels of low-density lipoprotein (LDL) respectively, leading to the development of *PCSK9* inhibitors such as evolocumab (Cohen et al., 2006; Kaddoura et al., 2020).

As previously mentioned, common variants have modest effect sizes and are less interpretable since many of them are noncoding. Therefore, in contrast to rare coding variants, they have had less contribution to drug target identification. However, increasing evidence from more deeply-curated genotype and phenotype datasets are suggesting that a distinct separation between common and rare variants may not be appropriate, and treating variant effects as a spectrum may be the way forward (Ghoussaini et al., 2023). With richer cataloguing of variant effects, allelic series can be developed - where the set of rare and common variant effects on a phenotype can be organised together to aid interpretation of a gene's function (Astle et al., 2016; Plenge et al., 2013).

Different types of genetic effects contribute to heritability

When studying the effects of genetic variation, one should also have an idea of the extent to which genetic variation contributes to phenotypic variation. This is loosely referred to as the heritability of a trait. Formally, “broad-sense-heritability” (H^2) is the proportion of variance in a measured phenotype that can be attributed to total genetic variation (Visscher et al., 2008). Classically, this can be derived from family and twin studies (Sahu & Prasuna, 2016). The subset of H^2 attributed specifically to the additive genetic effects of known variants is the “narrow-sense-heritability” (h^2) (Zuk et al., 2012). With the advent of GWAS, many common variants have been found to be associated with phenotypic variation, and aggregating their effect sizes together results in an estimation for a subset of h^2 , the SNP heritability (Zaitlen & Kraft, 2012). However, h^2 is often significantly smaller than H^2 , implying that many more genetic effects that contribute to phenotypic variation have not been captured by existing additive GWAS. This is sometimes referred to as the “missing heritability problem” (Maher, 2008).

There have been several reasons proposed to explain this missing heritability:

- GWAS often focus on common single nucleotide variants tested with an additive model, and it is possible that many other variants that contribute to heritability remain to be discovered. This includes rare variants, structural variants, or variants with non-additive effects that are not captured in traditional additive GWAS (Zaitlen & Kraft, 2012).
- Most GWAS do not capture the interactions between the variant or gene and the environment (gene-environment interactions), or the interactions between variants (epistatic effects); H^2 may be overestimated in twin studies when these interactions are not accounted for appropriately (Young, 2019; Zuk et al., 2012).

In this thesis we will focus on two variant effects that are usually not investigated in conventional GWAS: recessive variant effects and rare variant effects. I have discussed these variants more extensively in the individual project chapters, but here I provide a brief introduction.

Recessive variant effects

Recessive Mendelian conditions constitute a non-negligible public health burden, with about two to five neonates affected out of every thousand born, with the incidence estimated to be higher in populations with consanguinity, endogamy or founder effects (Baird et al., 1988; Xiao & Lauschke, 2021). This statistic reflects known recessive disorders, but it has been

frequently proposed that many more remain to be found (Guindo-Martínez et al., 2021; H. C. Martin et al., 2018; Vissers et al., 2015).

In the context of rare diseases, several studies have attempted to quantify the contribution of recessive effects, unbiased by whether variants are known to be pathogenic. From a dataset of about 6,000 families, Martin et al., 2018 estimated that 3.6% of European-ancestry patients with developmental disorders were attributable to recessive coding variants. Notably, for individuals with Pakistani ancestry, due to elevated homozygosity, recessive coding variants contributed to 31% of the developmental disorder disease burden (H. C. Martin et al., 2018). From analysing a combined dataset of almost 30,000 trios, Chundru et al., 2023 expanded on this approach, and determined that the recessive coding contribution to developmental disorders ranged from 2%-18% across ancestries, and that this attributable fraction increased with homozygosity (Kartik Chundru et al., 2023).

For common diseases and complex traits, the burden of recessive effects is not clear. Dominance heritability refers to the deviations from the additive / dosage effect quantified in GWAS, i.e. the non-additive contribution to heritability, and this includes recessive effects (Palmer et al., 2023). Several efforts to quantify the non-additive contribution to complex trait heritability have resulted in conflicting reports, depending on the method applied: Studies that calculate dominance deviations from biobank datasets estimate dominance heritability to be small, from 0.1% (Hivert et al., 2021; Pazokitoroudi et al., 2021) to 3% (Zhu et al., 2015). However, estimates from twin studies place this around 25% (X. Chen et al., 2015). We discuss the differences between additive and non-additive tests, as well as dominance deviation testing, further in Chapter 2.

There are several limitations of current GWAS in accounting for recessive effects: (a) common variant recessive effects might be small, (b) rare variants are often excluded from GWAS, and (c) GWAS often assume an additive pattern of inheritance. All of these factors relate to power. The power of a GWAS is optimised when the correct model of inheritance is used during testing, and additive models, often the choice in GWAS, might not detect recessive variants with large enough power (Lettre et al., 2007). Power for recessive testing is driven by the presence of sufficient numbers of individuals homozygous for the variants of interest.

$$(1 - q)^2 + 2(1 - q)q + q^2 = 1$$

q : *Frequency of the alternate allele*

$(1 - q)^2$: *Frequency of the wildtype (homozygous reference)*

$2(1 - q)q$: *Frequency of the heterozygous genotype*

q^2 : *Frequency of the homozygous alternate genotype*

The Hardy-Weinberg Equilibrium (HWE) (above) states that the frequency of the homozygous alternate is the square of the frequency of the alternate allele. Therefore, in populations with random mating and obeying the HWE, exponentially larger sample sizes are required to test for recessive effects, compared to testing for additive effects with the same power. This is particularly the case with small-effect variants and rare variants. We perform some simulations in Chapter 2 to demonstrate this.

GWAS have traditionally been performed in cohorts that closely follow the HWE, and also performed with the additive model, and together they may have potentially underestimated the contribution of recessive variants due to lack of power. However, as in the case of rare diseases, it is likely that the burden of recessive effects on common diseases and complex traits is dependent on the population in which the recessive effects are studied. Indeed heritability for traits also vary across populations, specifically, it may vary depending on consanguinity due to differences in the contribution of non-additive effects (Robertson, 1952; Visscher et al., 2008).

Together, these studies highlight the need for identifying more variants that act recessively on common diseases. This is the focus of Chapter 2 of my thesis.

Rare variant effects

Other than applying an additive model of testing, traditional GWAS have also often restricted their analysis to common variants, defined as those with a MAF of 5% or more (S. Lee et al., 2014). One of the reasons for this is because genetic datasets large enough for GWAS have, until recently, been generated by microarray-based genotyping technologies (S. Lee et al., 2014). These technologies do not genotype rare variants well (Weedon et al., 2021), due to the clustering algorithms applied by these technologies to call genotypes (Lamy et al., 2011; Mizrahi-Man et al., 2022). To supplement this, GWAS often include an imputation step, where millions of variants from a sequenced reference panel is imputed into the array

data, allowing rare variants (for example up to a MAF of 0.5%) to be tested (Marchini & Howie, 2010; Taliun et al., 2021).

Yet, rare variants were predicted to present a significant contribution to “missing heritability” (Zuk et al., 2014), because as previously described, they often have large effect sizes. Variants that fall into the coding regions of the genome, or the exome, are particularly interesting to study, also, because they are mechanistically easier to interpret. (Due to selection pressure, these are often rare.) With the development of next generation sequencing technologies, we are now able to sequence and profile more rare variants. Sequenced sample sizes have now finally become large enough to formally estimate the total contribution of rare variants to narrow-sense heritability using whole-genome sequence data. Results for height and body mass index suggest that indeed rarer variants contributed more to heritability than common variants (Wainschtein et al., 2022).

Rare variants are a part of all the projects covered in this thesis (Chapters 2-4).

Loss-of-function variant effects

A subset of rare variants that are particularly interesting are loss-of-function variants (LoFs), the focus of Chapters 3 and 4 of the thesis. LoFs are nonsense, frameshift and splice site mutations in exonic regions that are predicted to cause truncation of the translated protein, preventing the protein from carrying out its function (Figure-1.1, inspired by MacArthur & Tyler-Smith, 2010) (MacArthur & Tyler-Smith, 2010). Some of these LoFs may also act recessively, as heterozygous expression of the normal allele may be able to compensate for the variant. A survey of different variant effects on protein structure by Gerasimavicius et al., 2022 suggested that autosomal recessive effects are enriched for LoFs (Gerasimavicius et al., 2022). Homozygous LoFs, where variants on both copies of the gene inactivate the protein, are particularly valuable as they act as naturally-occurring gene knockouts that allow for the *in vivo* observation of the consequences of gene knockout (Burgess, 2016). They can be very useful both in the functional evaluation of genes, as well as for drug target identification (Minikel et al., 2020). However, the study of LoFs is challenging because true LoFs are rare, and the process of calling LoFs is enriched for errors (K. J. Karczewski et al., 2020). I discuss this in detail in Chapter 3.

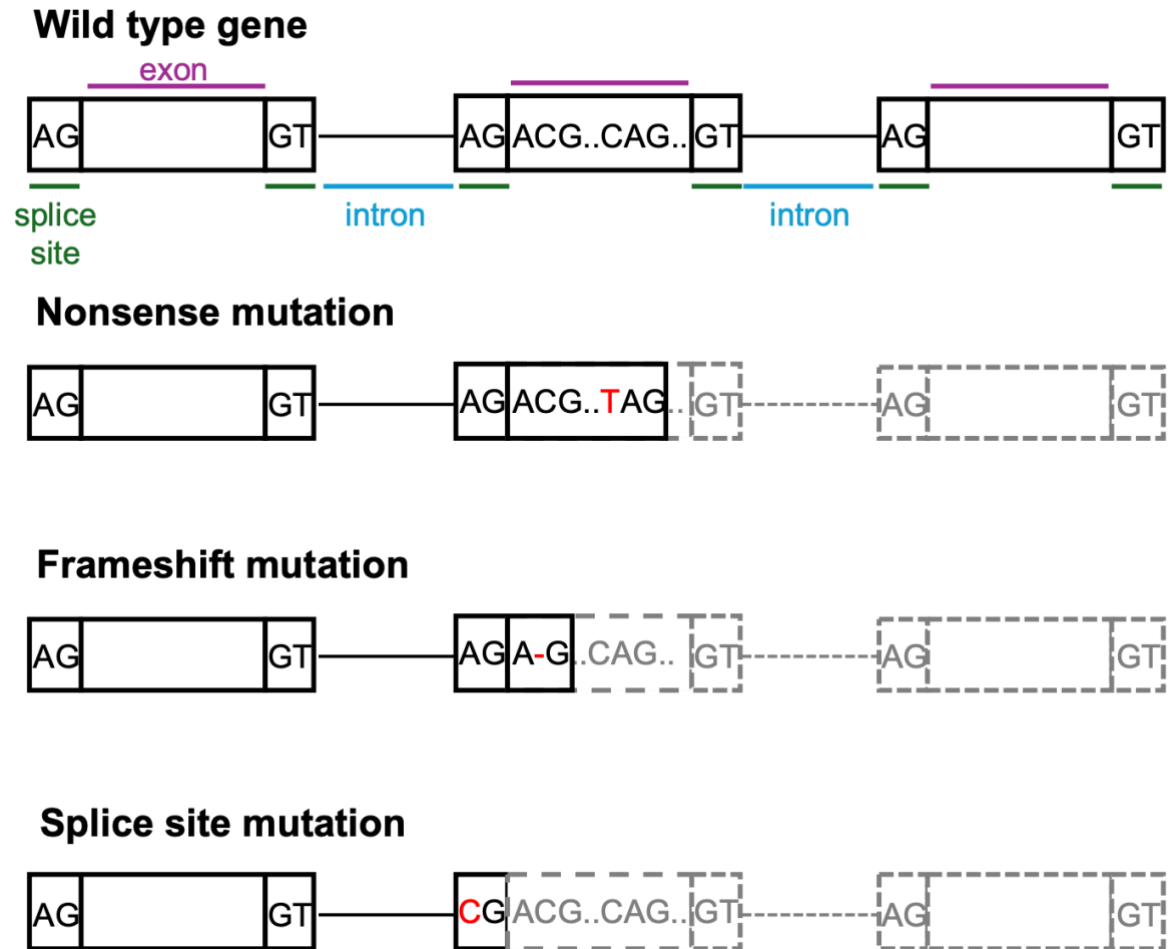


Figure-1.1

Types of loss-of-function variants. Graphic generated with reference to MacArthur & Tyler-Smith, 2010 (MacArthur & Tyler-Smith, 2010).

The technologies and datasets available in the field

The development of next generation sequencing technologies, genetic imputation algorithms, and single-cell RNA sequencing (scRNAseq) technologies allow us to generate the relevant datasets required to study variant effects. Here I provide a brief introduction to the methods that are key to my projects. In addition, large-scale datasets are integral to the study of genetic variation as well, and here I discuss the state of the available datasets in the field, contextualising them to the questions I aim to address in this thesis.

Technologies

Genotyping and sequencing technologies

The first human genomes were sequenced by Sanger sequencing (Sanger et al., 1992), the first generation of sequencing technology developed. Sequencing was performed by binding various labelled dideoxynucleotides to one base pair at a time and detecting the nucleotide match (T. Hu et al., 2021). This was effective, but very costly and time consuming. The Human Genome Project that sequenced the first human genomes used this technology and took more than ten years and almost three billion dollars to complete (Kisak, 2016).

Since then, genotyping methodologies have improved. As Sanger sequencing was very costly, the first large-scale genetic datasets built for GWAS were genotyped by microarray technologies (D. G. Wang et al., 1998). These methods focused on common single nucleotide polymorphism (SNPs), and are able to profile hundreds of thousands of SNPs at a low cost, from tens to hundreds of dollars per sample, depending on the complexity of the design (Iowa Institute of Human Genetics, 2024; Smith, 2015). However, the technology used on SNP chips is often not very accurate for calling rare variants (even if such variants are included on the array) (Auer & Lettre, 2015; C. F. Wright et al., 2019).

“Next-generation” sequencing (NGS) technologies generate the sequence from single molecules of DNA, allowing massively parallel processing of multiple fragments of DNA in a high throughput fashion (T. Hu et al., 2021). This has simultaneously brought down the cost and time required to sequence the human genome, and allowed an exponential scaling up of the size of sequencing datasets. Now a human genome can be sequenced within a few weeks (even within a few hours (Gorzynski et al., 2022)), and it costs around a thousand dollars (Goodwin et al., 2016). As only about one percent of the genome is coding (Richard Boland, 2017), an alternative approach is whole exome sequencing (WES) which specifically targets protein-coding regions to use resources more efficiently. (WES cost five hundred to five thousand dollars, compared to a thousand to twenty-five thousand dollars for whole genome sequencing (WGS) (Schwarze et al., 2018).) Large-scale WES projects are

emerging and have enabled the detection of many rare coding variants for further study (Backman et al., 2021).

Together, genotyping arrays, WGS and WES have driven the generation of multiple large-scale datasets of human genetic variation. While the costs of sequencing have gone down dramatically, the genotyping array is still the more cost-effective approach to genotype common SNPs. This is due to genetic imputation which helps bridge this gap, by allowing statistical ‘guessing’ of the genotypes at a larger number of variants (usually millions of SNPs) from a reference catalogue of sequenced variants into a backbone of genotyped SNPs. We performed imputation as part of the project in Chapter 2, therefore I provide a more comprehensive introduction there.

The majority of the NGS contributing to these datasets have been built from short series of sequences that have been generated in parallel and assembled back together bioinformatically. These paired-end short-reads cover about a maximum length of 75-150 base pairs per read. Therefore, they do not capture large structural variants, or regions with multiple repeats or low complexity well. “Long-read” technologies can overcome these limitations and present the next frontier to genomic sequencing (Goodwin et al., 2016).

Multi-omics data generation and the single-cell approach

Association analyses are performed between genotype and phenotype to study variant effects, and these phenotypic readouts can be collected across multiple levels: They can be traits such as quantitative measures related to the whole organism, such as height, or binary classifications such as the presence or absence of disease phenotypes. They can also be RNA expression and protein expression levels, or a whole host of measurements from various biomarker assays.

With the advent of GWAS, many variant-trait or variant-disease associations have been discovered (Visscher et al., 2017). Naturally, there have also been great efforts to functionally evaluate these variant effects (Cano-Gamez & Trynka, 2020). Due to the number of variants screened at once, many of these efforts have to be performed at scale. This has been enabled by the development of high-throughput “multi-omics” data generation technologies, which allow the collection of RNA and protein expression data on top of genetic data at scale (Hasin et al., 2017).

Lichou and Trynka 2020 outlined several of these functional approaches to evaluate genetic variants, two of which are particularly relevant to this thesis: scRNAseq, and *in vitro* cellular models (Lichou & Trynka, 2020). Many variant effects can be highly cell-type- or tissue-

specific, and identifying this specificity is an important aspect of effectively interrogating variant functions (GTEx Consortium et al., 2017).

Evaluating cell-type-specific variant effects can be achieved at increasingly high resolutions of cell-types due to scRNAseq. There have been many methods developed to perform high-throughput scRNAseq, but we applied the droplet-based approach in this thesis, as it is widely-used with many algorithms that we found relevant to our aims. This briefly includes the following steps: single-cell isolation into droplets that can be barcoded, reverse transcription of RNA into DNA for amplification, and library preparation for sequencing (Jovic et al., 2022). scRNAseq has revolutionised the study of cell-type-specific variant effects, driven by the generation of large-scale scRNAseq datasets and cell atlases (Ding et al., 2023; Eraslan et al., 2022).

Cells are also dynamic, and undergo changes into different cell states as they carry out their functions. Therefore, the study of variant effects *in vitro* would have to account for this context-specificity as well, by creating the appropriate cellular environment or designing the appropriate model (Fairfax et al., 2014).

An example of combining a single-cell approach with designing appropriate *in vitro* cellular models to interrogate variant effects has been described in Soskic et al., 2019: by profiling chromatin activity across T cells in various contexts (generated by cytokine stimulations), the authors identified cell states that were associated with the enrichment of GWAS loci in their open chromatin region, providing functional insight into several immune-mediated diseases (Soskic et al., 2019).

LoFs allow for a direct interpretation of variant consequences, which makes them attractive targets for functional evaluation. In chapter 4, I present our approach to functionally evaluate LoFs at the single-cell level and across several *in vitro* conditions, with the immune system as the model of choice.

Datasets

The development of the technologies described above have enabled the generation of large-scale genotype-phenotype datasets well powered to study genetic variation.

However, many of these studies have focused on populations with White European ancestries. In 2009, only 4% of GWAS participants were not of European origin (Need & Goldstein, 2009), with this proportion increasing to 20% in reports from 2016 (Popejoy & Fullerton, 2016) and 2019 (A. R. Martin et al., 2019). Still, more needs to be done to ensure that genetic studies are inclusive.

Performing genetic research across diverse ancestries makes human genetics, and its applications, more equitable (Fatumo et al., 2022; Ruan et al., 2022). Diverse datasets also give us a better insight into heritability (Visscher et al., 2008), genetic diversity, and population structure (Habel et al., 2015). These are some of the motivations for focusing this thesis on a cohort that is not predominantly of White European ancestry.

Genes & Health as the cohort of choice in this thesis

All the discovery work in this thesis was performed using Genes & Health (G&H) data. G&H is a community-based cohort of ~50,000 (as of 2024) British South Asians, predominantly of Bangladeshi and Pakistani ancestry. Genetic data is collected and linked to electronic health records, and it is set up to enable the design of recall-by-genotype studies within the cohort. The main goal of the cohort is to improve the health of this understudied community; and through the study of genetic variation, we can gain a better understanding of the biology and landscape of disease in British South Asians (Finer et al., 2020; Genes & Health, 2024).

As of 2019 (data freeze at ~40,000 individuals), the median year of birth is 1977 (interquartile range 1967-1985), with 56% female, and 67% British Bangladeshi and 33% British Pakistani (Finer et al., 2020). (More up to date demographic information relevant to specific analyses are described in subsequent chapters.) The most prevalent conditions in the cohort are type 2 diabetes mellitus, hypertension and ischaemic heart disease, which are important conditions to study in this cohort, given that British South Asians are at two to four times higher risk for these conditions compared to the White British population (Barnett et al., 2006; Sattar & Gill, 2015).

Due to differing practices of endogamy and consanguinity, with about 18.8% of individuals self-reporting parental relatedness (Finer et al., 2020; Minikel et al., 2020), the cohort is enriched for homozygosity. From the distribution of runs of homozygosity (ROHs) in the genomes of participants, Malawsky et al., 2023 inferred that that actually about 33% of the cohort have parents who are second cousins or closer, compared to 2% in UK BioBank Europeans (Malawsky et al., 2023). There is also strong population structure within the cohort, particularly among the individuals with Pakistani ancestry; in a study including British individuals with Pakistani ancestry from the Born in Bradford study, Arciero et al., 2021 observed that this structure was driven by the *biraderi* social stratification system (similar to the Indian caste system). Within some of these Pakistani subgroups, there was evidence of bottlenecks and increased recent identity-by-descent (IBD) sharing (Arciero et al., 2021).

Increased homozygosity is associated with increased risks of rare, recessive, Mendelian disorders (H. C. Martin et al., 2018). It has also been shown to be associated with increased risks for common diseases and complex traits, with the fraction of the genome in runs of homozygosity (FROH) shown in this cohort to be associated with 12 common diseases such as type 2 diabetes mellitus, and immune-mediated conditions such as asthma and dermatitis (Malawsky et al., 2023). A hypothesis for these associations is that the increased homozygosity results in more recessive-acting variants to be present in a homozygous state,

resulting in disease. This motivates the search for the underlying variants in the cohort that may have recessive effects on common diseases. Similarly, increased homozygosity and bottlenecking may increase the frequencies of certain variants that are very rare in other populations, including LoFs. Therefore, there would be more power to investigate LoF effects in this cohort. Collectively, these reasons justify G&H as the cohort of choice for our analysis of recessive and rare LoF effects.

Outline of the thesis

In this thesis, we profiled recessive effects and LoF effects in G&H, a British South Asian cohort with elevated homozygosity.

In Chapter 2, we first characterised the recessive effects of both common and rare variants on common diseases in G&H. I start by laying out the theoretical and empirical justifications for applying a recessive model of GWAS testing, and then present power calculations that demonstrate the power of G&H in detecting recessive effects. Next, we imputed variants into 44,190 genotyped individuals, using two imputation panels: a set of 4,982 exomes from within the cohort (to enrich for rare, coding variants), and the TOPMed-r2 panel (to include noncoding variants). After evaluating the imputation accuracy, we then performed association testing between the imputed variants and 898 common diseases, identifying 207 loci that reached standard genome-wide significance ($p\text{-value} \leq 5 \times 10^{-8}$) under the recessive model and were more significant than under the additive model. Of these, about 70% demonstrated a nominally-significant ($p\text{-value} \leq 0.05$) dominance deviation $p\text{-value}$, supporting a recessive pattern of association. I discuss several subcategories of these findings in detail, including disease associations with coding variants that provided insight in gene function.

In Chapters 3 and 4, I describe the recall-by-genotype of 114 individuals with putative LoFs affecting immune genes, for single-cell RNA-sequencing (scRNAseq) of peripheral blood mononuclear cells (PBMCs). I discuss the three phases to building this unique dataset: firstly, the curation of true LoFs from bioinformatically-called putative LoFs (which are enriched for errors), secondly, LoF effects on the immune transcriptome at rest, and thirdly, LoF effects on the immune transcriptome in stimulation (by interferon alpha for innate immune processes, and by anti-CD3 for adaptive immune processes).

I first introduce LoFs in detail in the introduction to Chapter 3, discussing the errors related to bioinformatically-called putative LoFs, and the cell-type- and context-specificity of variant effects. Then, I present the prioritisation approach we undertook to enrich the recall for true LoFs that can be studied in immune cells. I then provide an account of the recall process and data generation. From the scRNAseq data, we observed that LoFs were indeed associated with knockdown effects on the affected gene, and this can be observed at the RNA level.

In Chapter 4, we further investigated if these LoF effects differed across cell-types and contexts (stimulation conditions). I provide an introduction to the immune system, our system

of choice for this study, as well as its dynamic nature in the presence of relevant stimuli. I then described the steps we took to profile these LoF effects at rest and in the presence of two stimuli: interferon alpha to study innate immune processes, and anti-CD3 to study adaptive immune processes. First, we saw that direct LoF effects on the genes carrying the LoFs were cell-type- and context-specific. Next, we detected pathway derangements associated with these LoFs, that were again cell-type- and context-specific. Notably, LoF effects on cell-type proportions, if any, were subtle.

In the final chapter, I give a summary of the findings from Chapters 2 to 4, discuss the wider implications of these findings, and several limitations that applied across these projects. I end by outlining several future directions that could be undertaken in light of the findings from this thesis.

In summary, in this thesis, I leveraged the increased homozygosity within G&H, first, to demonstrate widespread recessive effects on common diseases, and second, to profile the direct and indirect effects of rare LoFs, including homozygous LoFs, on the immune transcriptome. Together, these findings provide insight into the genetic basis of multiple traits and diseases, highlighting the potential of studying genetic variation. In particular, I demonstrated the importance of including non-additive effects and rare variants in genetic analyses.

Recessive effects in a cohort of 44,000 British Pakistanis and Bangladeshis

Overview

Genetic association studies have largely focused on testing additive models, and the best-studied cohorts, which are typically of European ancestry, are underpowered to test recessive associations. Genes & Health (G&H) is a cohort of British Pakistanis and Bangladeshis with genetic data and linked electronic health records (EHRs). The cohort has a high rate of consanguinity (32% offspring of second cousins or closer) as well as endogamy, with multiple bottlenecked subpopulations. We leveraged these properties to firstly, impute variants from 4,982 whole-exome-sequenced (WES) individuals into 44,190 genotyped individuals, then, perform recessive association testing with phenotypes from the EHRs. We also performed whole-genome imputation from the Trans-Omics for Precision Medicine (TOPMED-r2) reference panel to expand the search for recessive associations to noncoding regions. We tested 607,981 whole-exome imputed variants and 10,045,406 whole-genome imputed variants (≥ 3 homozygous individuals) for association with 898 disease phenotypes (≥ 30 cases), using REGENIE's whole-genome-regression model. We identified 207 independent loci that reached standard genome-wide significance ($p\text{-value} \leq 5 \times 10^{-8}$) under the recessive model and had p -values more significant than under the additive model. Of these, 154 loci demonstrated a nominally-significant ($p\text{-value} \leq 0.05$) dominance deviation p -value, confirming a recessive pattern of association. After correcting for testing multiple phenotypes, 167 tests were significant at a 5% false-discovery rate threshold (corresponding to $p\text{-value} \leq 3.7 \times 10^{-8}$) and 32 tests at a Bonferroni-corrected threshold $p\text{-value} \leq 5.5 \times 10^{-12}$. In the independent cohort FinnGen, we replicated 71% of the expected number of Bonferroni-significant loci we would be powered to replicate given the case rates, allele frequencies and sample size of FinnGen. We replicate recessive associations between regions encoding for haemoglobin genes and hereditary anaemias, demonstrate a recessive association between folate deficiency and a *MTHFR* missense variant that has known recessive associations to other phenotypes related to folate metabolism, and discuss novel findings including a *SGLT4* missense variant protective for hypertension.

Introduction

Recessive effects in common diseases and complex traits

Recessive effects in humans have been primarily studied in the context of rare, monogenic disorders, and little is known about recessiveness in common diseases and complex traits (Palmer et al., 2023). With the construction of large-scale biobanks, it is now possible to study the recessive contribution to polygenic traits (Bycroft et al., 2018; Heyne et al., 2023). Identifying variants with recessive associations to diseases could improve polygenic risk scoring, provide better insight into gene and variant function, improve understanding of disease pathophysiology, and allow for the identification of novel drug targets (Gebreyesus et al., 2020; Hinney et al., 2010; O'Connor et al., 2022).

Performing genome-wide association studies to look for recessive effects

Theoretical basis for fitting a recessive model

The effects of genetic variation on common complex phenotypes are typically discovered through genome wide association studies (GWAS) (Klein et al., 2005; Lewis, 2002; Visscher et al., 2017). Researchers predominantly test an additive model in GWAS, as it has been demonstrated with theoretical and empirical support that this should capture much of the variant effect regardless of the underlying genetic architecture (Hill et al., 2008).

The additive model assumes that the effect of the variant is linear; i.e. in a quantitative trait, if a heterozygous genotype is associated with an effect on the trait of size β , being homozygous for the alternate allele should have a proportionately stronger association (double the heterozygous effect, 2β). However, one can therefore expect that, when the relationship between an allele's effect and the trait is non-linear, the linear model would become less powerful. We expand on this in Figure-2.1, with reference to models described in Palmer et al., 2023 and Heyne et al., 2023.

We simulated the relationship between genotype and a quantitative trait in three ways: additive, purely recessive, and predominantly recessive with a mild heterozygous effect. In all scenarios we scaled the effect sizes such that the homozygous effect size will be at 100% and the wild-type at 0%. An additive effect is therefore represented in this simulation by having effect sizes of 0%, 50% and 100% for the wild-type, heterozygous and homozygous individuals respectively. If the variant follows an additive model, the linear regression is a perfect fit for the data ($r^2=1$), regardless of the allele frequency (AF) (Figures-2.1A,D). With a purely recessive effect, the heterozygous population would have a 0% effect size, just like the wild-type individuals, and only the homozygous population would express a change in

the phenotype. Some recessive variants can have a mild heterozygous effect, but are still described as “recessive” because the difference between the homozygous effect and the heterozygous effect exceeds the difference between the heterozygous effect and the wild-type. Here, for recessive with mild heterozygous effect, we simulate the heterozygotes to have an effect size of 10%. As expected, the linear fit worsened for both recessive patterns of inheritance (r^2 in Figures-2.1B,C,E,F). Secondly, with a rarer AF, which is more realistic for disease-causing variants, the model performed even worse. This demonstrates the weight of the homozygotes on the effect detected by a linear model, as with a rarer frequency of homozygotes, the magnitude of the effect (represented by the “slope” in Figures-2.1) and fit (r^2) of an additive model both weakened.

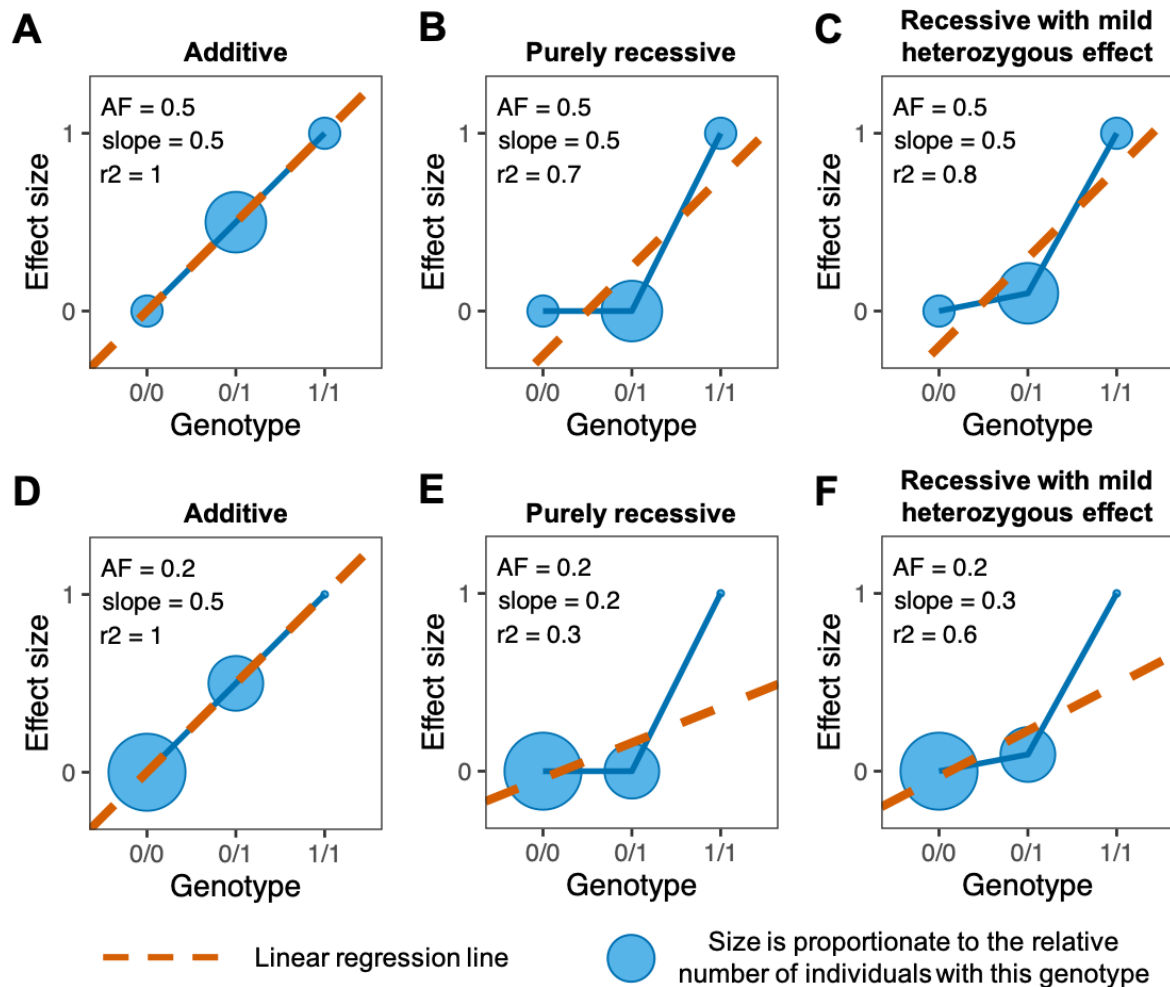


Figure-2.1

Fitting an additive model to additive (A,D), purely recessive (B,E), and recessive with mild heterozygous effect (C,F) inheritance patterns. The allele frequency (AF) is used to calculate genotype frequencies based on the Hardy-Weinberg Equilibrium (HWE). The effect sizes for each genotype were then weighted by the genotype frequencies, and this was regressed against the genotype to generate the linear regression line, slope and adjusted- r^2 in R (version 4.0.2). Two AFs were simulated, 0.5 (**A-C**), and 0.2 (**D-F**).

From this we can see that nonlinear effects, while they may still be detected under the additive model, will be better studied with a model that accurately assumes the underlying pattern of inheritance (Bagos, 2013; Lettre et al., 2007). We demonstrated this in the next simulation (Figure-2.2), along with an explanation of how each test was performed.

We considered a hypothetical example of a recessive relationship with a mild heterozygous effect, where the case rates (CRs) were 10%, 20% and 90% for wild-type, heterozygous and homozygous individuals respectively. We calculated genotype numbers for an AF of 0.3 and sample size of 4000, and assumed the genotypes would follow Hardy-Weinberg Equilibrium (HWE). For visualisation we plotted a linear regression, although a logistic regression would be performed in practice. If an additive model were fitted, the three genotypes would be considered as three distinct groups to be regressed with the phenotype CR (Figure-2.2A), giving a slope of 0.56 and r^2 of 0.67. If a recessive model were fitted, the wild-type and heterozygous individuals would be collapsed together to be compared to the homozygotes (Figure-2.2B), and the slope and r^2 would increase (slope = 0.75; $r^2=0.95$). Removing the homozygous individuals from the additive model greatly diminished the detected effect (slope = 0.20, $r^2 = 1$), demonstrating the weight of the homozygotes on the additive test (Figure-2.3C).

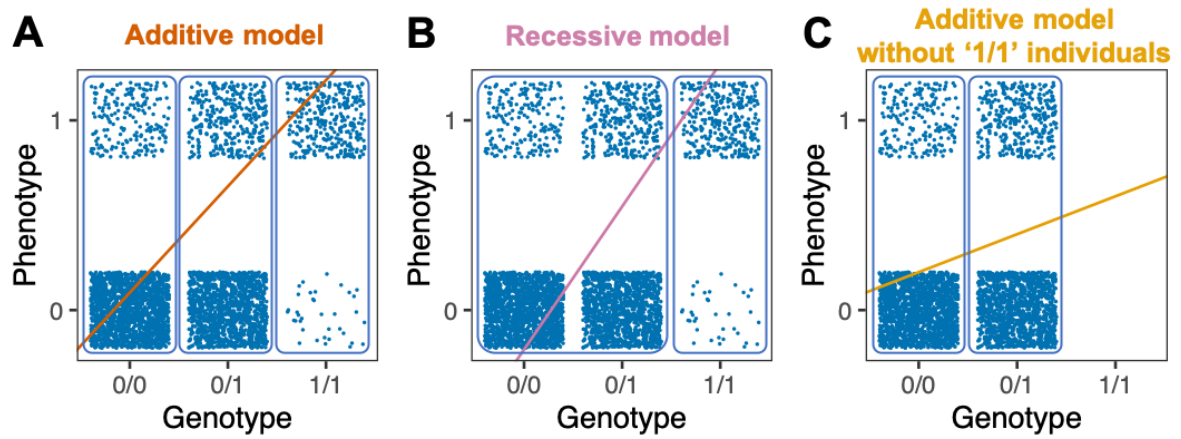


Figure-2.2

Simulation of a recessive pattern of inheritance, picked up by different models of testing. Each point represents an individual ($n = 4000$). The round-cornered boxes represent the way the genotypes are grouped for testing, with the lines showing the line of best fit from a linear regression.

One way to test for non-additive effects is to calculate the dominance deviation, which quantifies the extent to which the relationship between genotype and phenotype deviates from an additive pattern. This is done by adding an additional variable to the association test, re-coding the genotypes from '0' (homozygous wild-type), '1' (heterozygous), '2' (homozygous alternate), to '0', '1', '0' respectively (Fisher, 1919; Palmer et al., 2023; Purcell et al., 2007). We formalise this mathematically in the Methods. This is not normally performed to detect new recessive associations, but serves as an additional step to characterise non-additive effects (Guindo-Martínez et al., 2021; O'Connor et al., 2022).

Empirical results from other studies

Applying a recessive model has allowed for the discovery of associations that would have been otherwise missed under the conventional additive testing. Here we summarise several of the key previous papers.

The FinnGen project (FG) is a biobank in Finland with a goal of studying genetic variation in this isolated population (Kurki et al., 2023). They first imputed coding variants from a population specific reference panel comprising of 3,775 whole-genome sequenced Finnish individuals into 176,899 individuals with array data, taking advantage of the greater haplotype sharing in this bottlenecked population to impute lower-frequency variants (Kals et al., 2019). Next, by performing recessive tests on 44,370 variants and 2,444 diseases, they identified 31 loci at genome-wide significance (GWS, $p\text{-value} \leq 5 \times 10^{-8}$) where the recessive model fitted better than the additive model (recessive $p\text{-value}$ two orders of magnitude more significant than the additive $p\text{-value}$). Twenty of these findings were further validated (in the later release of FinnGen or in the UK Biobank), and notably thirteen loci would have been missed with the additive model alone. This approach allowed them to replicate known associations (e.g. with retinal dystrophy), and identify novel associations, such as a *CASP7* missense variant with adult-onset cataract. (Heyne et al., 2023)

O'Connor et al., 2022 performed the largest GWAS meta-analysis using a recessive model in type 2 diabetes mellitus (T2DM): 33,139 cases, 279,507 controls, 11,634,328 variants. They compared their results to the largest additive GWAS performed (Mahajan et al., 2018; Vujkovic et al., 2020), and classified novel recessive associations as findings that were not in linkage disequilibrium (LD, defined here as $r^2 < 0.3$) with known additive findings and a significant dominance deviation $p\text{-value}$ (< 0.05). Of the 51 findings that reached GWS, 5 were distinct from known additive hits.

To expand the search for non-additive associations with complex traits, Guindo-Martínez et al., 2021 developed GUIDANCE, which firstly improved genotype imputation by applying

multiple reference panels (we discuss multiple reference panels for imputation in subsequent sections), included X chromosome analysis, and incorporated non-additive association testing. This allowed them to identify 94 GWS loci from 62,281 individuals across 22 age-related diseases. Amongst 26 novel loci, 4 were identified only with the recessive model.

Palmer et al., 2023 expanded beyond recessive effects to study dominance, as defined in classical statistical genetics to be any deviation from the additive pattern of inheritance. They systematically quantified the contribution of dominance deviations to heritability across 1,060 common traits in UK Biobank ($n = 361,194$), by applying a “dominance scan” across the genome, quantifying the additive and dominance deviation at each variant. Next, they observed that linkage disequilibrium between dominance associations were less correlated than pairs of additive associations, and applied the concept to finemap causal variants more accurately. While not strictly devoted to identifying novel recessive associations, the concepts in the paper pertaining to characterising deviation from the additive model have been applied in our study.

The evidence in the literature outlined above motivated us to perform recessive testing in Genes & Health (G&H).

GWAS algorithms

There are multiple algorithms that have been developed to perform GWAS. Early methods include SNPTEST (Wellcome Trust Case Control Consortium, 2007) and PLINK (Purcell et al., 2007), which were applied within smaller cohorts of a few thousand individuals (Mbatchou et al., 2021). However, they were not adequate to account for the complexities of larger and more structured cohorts. When population-based cohorts were being used for GWAS testing across phenotypes, challenges arose including (a) population stratification, (b) relatedness between individuals, (c) increased Type 1 error rates due to case-control imbalances (since a population-based cohort is not recruited to study any particular disease, so rare diseases will have a low fraction of cases), and (d) high computational cost from the large number of individuals and phenotypes.

To address these limitations, several new methods have been developed in recent years, including linear mixed models and whole-genome regression models to account for the population structure and relatedness. First, principal component analysis (PCA) was introduced into genetic association studies as a way to control for population stratification (Price et al., 2006). However, PCA fails to account adequately for relatedness, leading to the development of linear mixed model (LMM)-based methods (H. M. Kang et al., 2010). LMM also provides more power than simple linear or logistic regressions, as they are able to

account for polygenic effects (Yang et al., 2014). Examples of these algorithms are BOLT-LMM (Loh et al., 2015) and fastGWA (L. Jiang et al., 2019). SAIGE (Zhou et al., 2018) also developed a saddle-point approximation to control for Type 1 error rates, making it popular for performing “pheWAS”, or GWAS of multiple phenotypes in a population-based cohort (Zhou et al., 2018).

REGENIE (Mbatchou et al., 2021) was also developed as a two-step approach to address the challenges described above. In the first step of the algorithm, whole-genome regression is performed via two levels of ridge regressions, to decompose common variant markers into twenty-three chromosome predictors. These predictors would provide more power by representing the polygenic background, and as they would be correlated with population structure and relatedness, they can be accounted for in the subsequent association analysis. In the leave-one-chromosome-out (LOCO) approach of the algorithm (which is the option selected in our study), each chromosome predictor has left out the chromosome containing the test variant. The assumption is that confounding structure would influence all the autosomes universally, and so the LOCO approach prevents the variant of interest from becoming fitted into this regression. The relevant LOCO variable (i.e. the variable excluding the chromosome of the variant to be tested) is then fed into the second step of the algorithm, where the variant of interest will be tested with the phenotype, while the “genetic background” is controlled for by the LOCO variable. Next, REGENIE developed an approximate Firth correction for situations of case-control imbalances. The Firth correction is a penalised likelihood approach to correct for bias from small sampling: a penalty term that approaches 0 as sample size approaches infinity is applied (Firth, 1993). In the case of logistic regression testing, this penalises the inflation that would have resulted in Type 1 error when case counts get very low (X. Wang, 2014). The actual calculation would be computationally intensive for each test and REGENIE developed an approximate Firth correction by estimating the correction required from testing covariate effects in the null model during step 1. Runtime and computation cost are reduced, parallelising tests performed, and decoupling the whole-genome regression from the association testing. This made it the algorithm of choice for our study.

The statistical power needed for recessive analyses

As previously described, a recessive effect is detected when individuals homozygous for the variant demonstrate a statistically-significant difference in disease risk compared to the combined group of homozygous wild-type and heterozygous individuals. Therefore, power for recessive tests is driven by the presence of sufficient numbers of individuals homozygous for the variants of interest. In outbred populations following the HWE, the frequency of the

homozygous alternate is squared of the frequency of the alternate allele. Therefore, in outbred populations, large sample sizes, more than what would be powered for additive effects, are required for adequate power to test recessive effects, particularly for rare variants. Power to detect recessive effects is expected to be increased in bottlenecked populations like Finland, where recessive variants may rise in frequency due to founder effects (Heyne et al., 2023; Kurki et al., 2023), or in populations enriched for consanguinity, such as G&H, where genotype frequencies deviate from the HWE and there is increased homozygosity.

We ran simulations to show that power can be boosted firstly by fitting a recessive rather than an additive model to test for a recessive effect, and secondly by increased homozygosity in the cohort (Figure 2.3). Setting the sample size to 44,000 (the size of G&H), using an odds ratio (OR) of 2, a phenotype CR of 1% and p-value $\leq 5 \times 10^{-8}$ for the p-value threshold, we calculated the power of a recessive versus an additive test across the AF spectrum, simulating various levels of inbreeding, ranging from a completely outbred cohort (inbreeding coefficient $F = 0$), to a consanguineous cohort with an average inbreeding coefficient of 10% (which corresponds to the average fraction of individuals homozygous at any position in the genome). In practice, the average inbreeding coefficient of G&H is ~2.2%. We first see that the additive model had less power for detecting a truly recessive effect than the recessive model, particularly at lower AFs (Figures-2.3A,B). Next, we see that power to detect this recessive effect increased with the inbreeding coefficient, and for this set of parameters, this was most noticeable around the AF 0.2 - 0.5 range (Figure-2.3A). This sample size and levels of consanguinity simulated were not powered to detect an effect size of OR = 2 in rare variants, therefore we increased the OR to 5 and reran the simulations for the rare and low frequency spectrum, and we still saw that power is higher with increasing F (Figure-2.3B).

We next wanted to evaluate the minimum OR we would be powered (80% power) to detect in G&H. To do that we simulated varying CRs from 0.1% to 25%. We see that as expected, the rarer the variant and the phenotype, the higher the OR would need to be for us to have power to detect it. We projected the recessive findings reported in FinnGen by Heyne et al., 2023 onto our simulations, and found that the majority of the findings had ORs higher than the minimum that we are powered to detect at their corresponding CRs and AFs (Figure-2.3C). This suggested that G&H will be well powered to detect very large ORs at rare variants i.e. effectively Mendelian associations, and smaller recessive effects at common variants and common traits. We were therefore hopeful that our study would have adequate power to detect novel recessive effects, particularly since many of the sub-populations

amongst the Pakistanis are very bottlenecked, and so may carry recessive founder variants that have risen in frequency (Arciero et al., 2021).

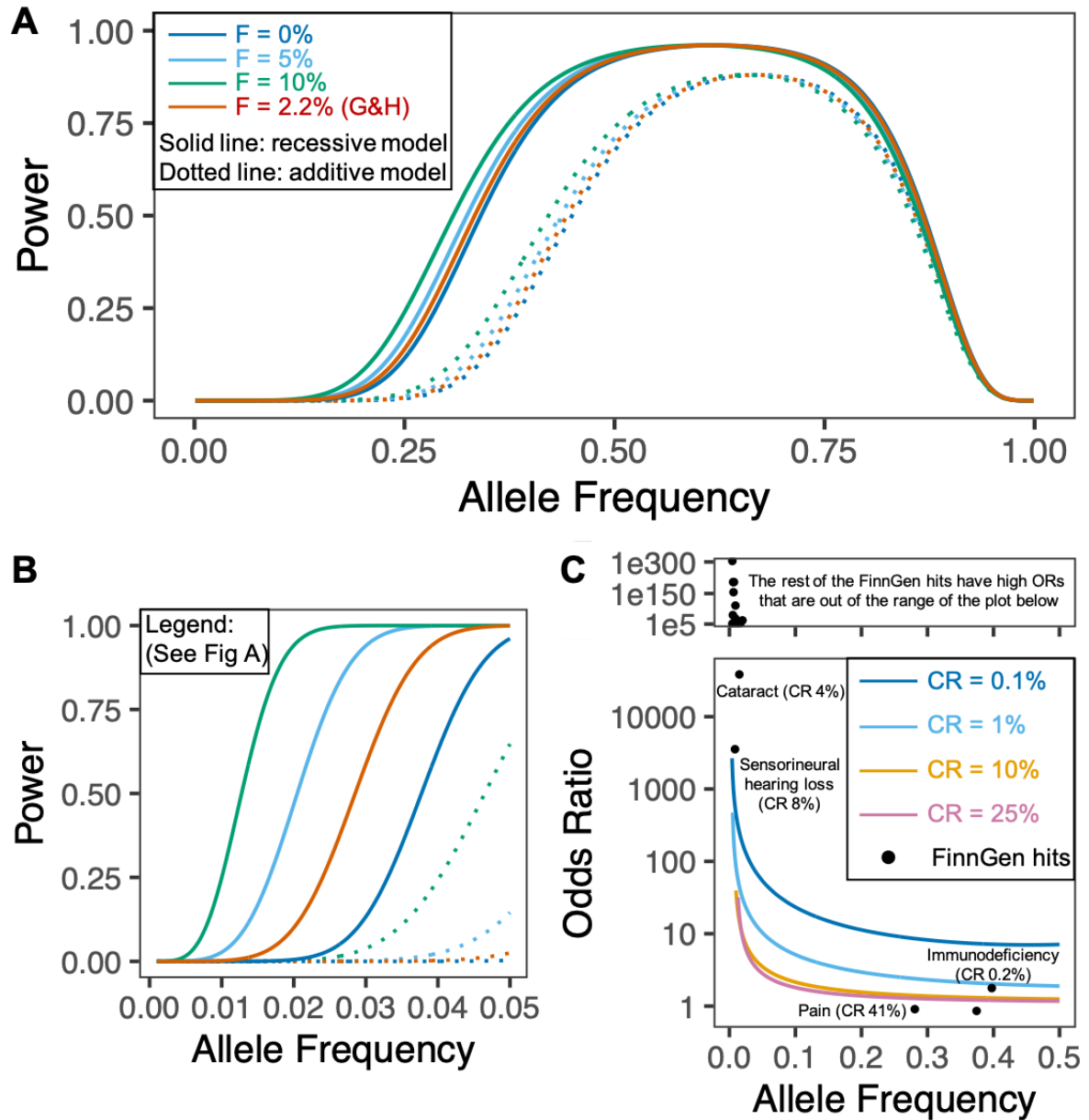


Figure-2.3

Simulations to evaluate power to detect a recessive effect in G&H when using a recessive versus additive model. **A:** Calculating power to detect an association at $p\text{-value} \leq 5 \times 10^{-8}$ for a range of allele frequencies and inbreeding coefficients ("F"s). Sample size = 44,000, OR = 2, CR = 1%. **B:** Calculating power for a range of rarer allele frequencies and Fs. Parameters as per **A**, except OR = 5. **C:** Calculating the minimum odds ratio G&H would be powered to detect at 80% power, and $F = 2.2\%$. "CR" : Case rate. Recessive hits reported in FinnGen (Heyne et al., 2023) were then projected onto the plot, showing their OR and AF estimated in FinnGen.

The value of imputation from population-specific reference panels for improving power in GWAS

The power calculations shown above assume that the genotypes are all measured with perfect accuracy. However, in reality, most biobanks still employ array-based genotyping (which is more cost-effective than sequencing (Giani et al., 2020; VanRaden et al., 2022)) and imputation to a reference panel for GWAS, so most of the variants tested are imputed with imperfect accuracy. (Y. Li et al., 2009; Marchini & Howie, 2010).

The use of ancestry-matched reference panels has been shown to improve the accuracy of imputation (Kals et al., 2019), leading to the development of multiple large-scale, publicly available, reference panels of diverse ancestries, such as the genotyping-based International HapMap Project (International HapMap Consortium et al., 2007) and the 1000 Genome Project (1000 Genomes Project Consortium et al., 2015). The current 'gold standard' imputation panel comes from the Trans-Omics for Precision Medicine (TOPMED) programme, which sequenced 53,831 whole genomes from diverse ancestry and demonstrated increased gains in the number of rare variants available for imputation and testing, as well as improved accuracy in imputation (Taliun et al., 2021).

Population-specific reference panels have been developed, and depending on their relative size and the characteristics of the population, may allow for higher-quality imputation at common variants as well as imputation of population-specific rare variants. For example, this was demonstrated in improved accuracy in human leukocyte antigen (HLA) variants imputation in Japanese (Okada et al., 2015) and Finnish (Ritari et al., 2020) cohorts. Studies constructing population-specific reference panels for Estonian (Mitt et al., 2017), Chinese (L. Li et al., 2021), and African (O'Connell et al., 2021) cohorts have also allowed for the imputation of low-frequency and rare variants enriched or specific to those populations. Notably, Barton et al., 2021 applied this concept to impute from 49,960 whole-exome sequencing (WES) samples in UK Biobank to the larger genotyped cohort ($n \sim 500,000$), boosting their power for association testing of rare coding variants. This paper was a key inspiration for the work of this chapter.

Genes & Health as a resource for detecting recessive effects

Genes & Health (G&H) was introduced in Chapter 1. At the time of beginning this project, genotype chip data were available for ~44k individuals from the project, of whom ~5.5k also had low-coverage WES data, thus providing an opportunity to build a population-specific reference panel, with a focus on rare, exonic variants. (In parallel, high-coverage WES was being performed on the entire cohort, funded by an industry consortium, but it did not

become available until this work was largely completed, and an industry priority period gave limitations for its use for academic research.)

Arciero et al., 2021 characterised the extent of IBD sharing, or regions of the genome that are identical-by-descent, in G&H. They demonstrated that subpopulations of G&H, particularly those that corresponded to certain Biraderi (kinship) groups among British Pakistanis, had elevated levels of IBD sharing (due to both consanguinity and endogamy), more than 10-20 times that found in the Finnish. We therefore hypothesised that the increased IBD-sharing within these subgroups in G&H would allow a reasonable (or better) quality imputation of variants, particularly with a reference panel built from individuals in the same cohort.

From this imputation, which would include rare, protein-coding variants from exomes, we would be able to take advantage of increased homozygosity to test for recessive effects.

The aims of this chapter

We aimed to firstly, impute exonic variants from the exome-sequenced individuals into the larger genotyped G&H cohort. Secondly, as the TOPMED reference panel is already a resource for whole-genome variation with some South Asian contribution, we also performed a whole-genome imputation to study recessive effects in the noncoding region. At this point, as a secondary aim, we also compared our imputations based on a population-specific reference panel and an external reference to describe the benefits and disadvantages of both. Thirdly, with the imputations, we performed association testing with binary phenotypes curated from the electronic health records, focusing on detecting recessive effects. We then sought to find replication for any novel recessive findings in a distinct cohort (FinnGen). The overall goal of the project is to provide a profile of recessive effects of variants in this cohort of British South Asian individuals.

Methods

Calculating power for this study

The results for this are described in the introduction (Figure-2.3). We performed simulations using the genpwr R package (Moore et al., 2019). We applied the following parameters for all calculations: sample size = 44,000 (to simulate G&H), p-value $\leq 5 \times 10^{-8}$, model = recessive (or additive), regression = logistic, and assumed there was no gene-environment interaction. The power, AF, CR and OR were modified depending on what was being simulated.

Since the genpwr package does not accept genotype frequencies, but rather, takes the AF as input and assumes HWE to determine the genotype frequencies, to account for increased autozygosity in G&H, we calculated an “ $AF_{autozyg}$ ” such that the expected number of homozygotes would be correctly calculated. Specifically, we calculated:

$$AF_{autozyg} = \sqrt{(1 - F) * AF_{out}^2 + F * AF_{out}}$$

$AF_{autozyg}$: AF of a variant taking into account the cohort's autozygosity

AF_{out} : AF of the variant in a hypothetically completely outbred cohort

F = The average inbreeding coefficient in the sample

The inbreeding coefficient F corresponds to the average fraction of the genome homozygous in a given sample. The frequency of a homozygous genotype is therefore the sum of:

- A. The “inbred” contribution: the fraction of the population already homozygous at this variant (represented by F) multiplied by the AF [$F * AF_{out}$]
- B. The “outbred” contribution: the remaining fraction of the population (1 - F), multiplied by the homozygous frequency following HWE (AF^2) [$(1 - F) * AF_{out}^2$]

(Futuyma, 1986)

The FinnGen recessive hits were collated from Table 1 of (Heyne et al., 2023). We filtered the 31 findings to the 18 findings that were validated in release 6 (Supplementary Table 3) (Kurki et al., 2023). The ORs, AFs and CRs were then projected into the power calculations we performed.

Generation and initial processing of genetic data

Genotyping was performed on 44,396 G&H individuals with an Illumina single nucleotide polymorphism (SNP)-array chip (GSAv3EAMD, build 38). Initial quality control (QC) was performed by the G&H team in Queen Mary University of London (QMUL) using Illumina's GenomeStudio, and involved the following steps: Structural variants and variants with cluster separation scores <0.57 , GenTrain score <0.7 , excess of heterozygotes >0.03 or ChiTest $100 <0.6$ were excluded. Low call rate samples ($<99.5\%$ for male, $<99.2\%$ for female) and samples that failed sex checks were excluded, resulting in 44,396 individuals. For duplicate variants or samples, the variant/sample with the higher call rate was retained. (Q. Q. Huang et al., 2022)

Exome sequencing was performed between 2016-2018 on a subset of 5,236 individuals who self-declared as having consanguineous parents, using Agilent V5 capture kits. The distribution of average sequencing coverage is bimodal, with a subset sequenced to $\sim 20X$ and the rest to $\sim 40X$ (Figure-2.6A). The lower coverage was chosen since the primary interest in these samples was the ability to identify rare homozygous variants, which was possible even with low-coverage data. The exome data were mapped with bwa-mem (H. Li, 2013) and variants called with GATK HaplotypeCaller (McKenna et al., 2010) by the Human Genetics Informatics team at Sanger.

QC of genetic data

Before preparing the imputation backbone and reference panel, QC for the SNP-array and the WES data was performed, as summarised in Table-2.1 and described below.

	SNP-Array Data	WES Data
Initial QC	44,396 individuals after QC by QMUL: Variant QC in GenomeStudio Sample QC (exclude samples with low call rate or failed sex checks)	5,236 individuals after light sample QC (exclude samples with low coverage or sex discrepancies)
Genetically infer ancestry	44,190 individuals assigned as Pakistani or Bangladeshi	5,073 individuals with array data and assigned as Pakistani or Bangladeshi
Genotype QC		Set genotype as missing when: GQ <20 binomAD <10 ⁻² DP ≤7
Variant QC	534,806 autosomal SNPs	2,245,547 autosomal SNPs and indels, after excluding: SNPs: "QD < 2.0 FS > 30 MQ < 40.0 MQRankSum < -12.5 ReadPosRankSum < -8.0" Indels: "QD < 2.0 FS > 30 ReadPosRankSum < -20.0"
	534,531 variants with a call rate ≥99%	1,645,161 variants with a post-genotype QC call rate ≥70%
	533,362 variants with HWE p-value ≥10 ⁻⁶ in the Bangladeshi subgroup	
	533,166 biallelic variants	
<i>Evaluate Concordance</i>		
MAF cut-offs	Keep 469,678 variants with MAF >0.1%	Remove singletons
Resolve Overlapping Positions	Exclude 25 common palindromic variants with MAF >40%	Keep, as no problem of allele-switching
	Exclude 515 variants with overlapping positions but unmatched alleles	Keep, as some are indels that have been genotyped incorrectly on the array
	Keep 38,248 variants with matched alleles and MAF >0.1%	Exclude, as the array has a higher overall call rate (Note: rare variants with MAF ≤0.1% that overlapped with the array are retained)
Final Sample QC		4,982 individuals after excluding: Pre-genotype QC call rate <4SD from mean Post-genotype QC call rate <2SD from mean Post-genotype QC NRD >4SD from mean

Table-2.1

Summary of QC steps for the SNP-Array and WES Data

Array data

From the genotyped 44,396 individuals, we genetically inferred 44,190 individuals to be either of Bangladeshi or Pakistani ancestry (next section). Another round of variant filters was applied to include only autosomal, bi-allelic SNPs with $\geq 99\%$ call rate. We noticed that 27% of variants failed HWE testing at a p-value threshold of 10^{-6} in the Pakistani subgroup, compared to 0.2% in the Bangladeshi subgroup. The Pakistani subgroup has high autozygosity and strong population structure, while the Bangladeshi subgroup has minimal structure and much less autozygosity, so it is likely that many more “true” variants in Pakistanis would not obey the HWE (Malawsky, 2021). Therefore, the HWE test was performed (using PLINK1.9) (C. C. Chang et al., 2015) only in the Bangladeshi subgroup to limit ourselves to excluding variants that do not obey the HWE testing due to genotyping errors, and the variants that failed this p-value threshold of 10^{-6} in Bangladeshis were then excluded from the entire dataset.

WES data

After excluding individuals with sex discrepancies or $<10\times$ average on-target coverage, 5,236 individuals in G&H also have WES data. Variants with the following criteria were excluded through bcftools (Danecek et al., 2021) :

- SNPs: "QD < 2.0 || FS > 30 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0"
- Indels: "QD < 2.0 || FS > 30 || ReadPosRankSum < -20.0"

(The above WES QC was performed by Hilary C. Martin) (Genes & Health, 2019)

Within this, 5,073 individuals were genetically inferred to be of Bangladeshi or Pakistani ancestry from their array data (next section).

QC at the genotype level was performed by using the bcftools setGT plugin to set to missing genotypes that had genotype quality (GQ) <20 , a p-value from a binomial test for allele depth at heterozygous sites (binomAD) $<10^{-2}$, and depth ≤ 7 . Variants were excluded if they had a post-genotype QC call rate $<70\%$.

Genetic inference of ancestry

We genetically inferred ancestry for the 44,396 individuals by merging the SNP-array data with reference cohorts.

From the SNP-array that only went through initial QC (Table-2.1), we filtered to variants that were autosomal, common (defined as a minor allele frequency, MAF > 0.01), had a call rate of $\geq 99\%$, and passed the HWE exact test ($p > 10^{-6}$) in self-declared Bangladeshi individuals. This was merged with reference sequences of 3,433 individuals from the 1000 Genome Project (1000G) (1000 Genomes Project Consortium et al., 2015) and Central and South Asian individuals from the Human Genome Diversity Project (HGDP) (Bergström et al., 2020). We excluded palindromic variants, and variants with AF differences between G&H and 676 reference South Asians curated from 1000G and HGDP. We defined variants with significant AF differences in the following manner: Firstly, we calculated the residuals from the linear regression between the AFs in both datasets. Next, we binned the variants into bins by frequency (in intervals of 0.01), and selected variants for which the residual was >5 standard deviations (5SD) away from the mean of the residuals in that frequency bin. (This choice seemed reasonable after testing various SD thresholds, Supplementary-figure-2.1A). Lastly, we performed Fisher's exact tests to compare the genotype counts between the G&H data and the 676 reference South Asians at the variants selected above and excluded those with a $p\text{-value} < 10^{-5}$. (Again, we tested $p\text{-value}$ thresholds ranging from 0.05 to a multiple testing correction of $< 0.05/349,632$ variants, and felt that 10^{-5} was reasonable, Supplementary-figure-2.1B). The various thresholds and the distribution of the outlier variants excluded are graphically represented in Supplementary-figure-2.1C-D.

After merging of G&H with the reference samples, LD pruning was performed (window size 1000 kilobases (kb), step size 50, LD r^2 0.1) with PLINK1.9 and long LD regions were excluded. (Meyer, 2020)

Principal component analysis (PCA) was performed with PLINK1.9 on the reference individuals, then the G&H individuals were projected into the reference PC space (PCA1). We calculated uniform manifold approximation and projection (UMAP) coordinates (umap R package) (McInnes et al., 2018). We found that the UMAP with 7 PCs was optimal to separate the reference individuals into superpopulations. 44,320 out of 44,396 G&H individuals were inferred to be South Asian at this stage (Figure-2.4A).

KING (Manichaikul et al., 2010) was run to estimate pairwise relationships with the PropIBD algorithm in the G&H cohort up to 4 degrees, and we removed a minimal set of 14,727 individuals who had at least one relative (3rd degree and closer) in the dataset, leaving 29,668 unrelated individuals.

We performed a second PCA (PCA2) on the unrelated G&H individuals, projecting the related G&H individuals into the PC space. The UMAP with 4 PCs identified distinct clusters

that corresponded well to self-declared Bangladeshi/Pakistani ancestry, and this was used to genetically classify individuals as genetically Bangladeshi or Pakistani (Figure-2.4B).

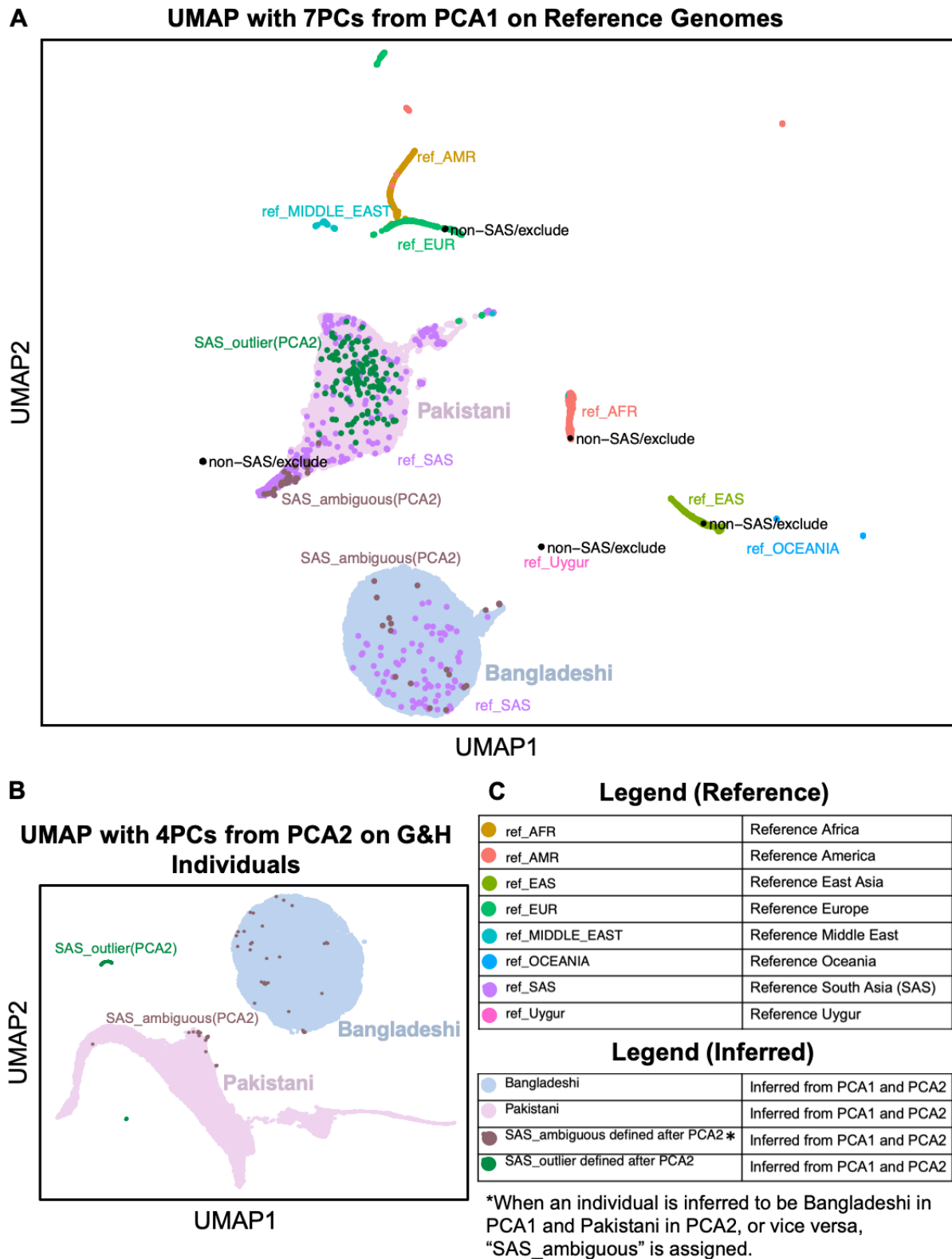


Figure-2.4

Genetic inference of ancestry using the G&H SNP-array data. **A:** UMAP with 7 PCs from the PCA of reference individuals (PCA1), with G&H individuals projected into the PC space.

B: UMAP with 4 PCs from the PCA of unrelated G&H individuals (PCA2), with related G&H individuals projected into the PC space. **C:** Legends for both UMAPs.

Preparing genetic data for imputation

Building and phasing of the imputation backbone from the array data

Based on the concordance analysis (see Results), we decided to include only array variants with $MAF > 0.1\%$ in the imputation backbone. Overlapping positions with the WES were resolved as follows:

- For SNPs at overlapping positions with matched alleles and with $MAF > 0.1\%$, we retained these in the GSA data but removed them from the WES.
- We excluded overlapping common palindromic variants ($MAF > 0.4$) where the reference allele could not be confidently assigned.
- We excluded overlapping positions with unmatched alleles (including all indels).

This resulted in 469,678 variants, which were then phased with EAGLE2 (Kpbwt=20,000) (Loh et al., 2016).

Building and phasing of the within-cohort whole-exome reference panel

Before building the reference panel, we first carried out quality control on the N samples with the WES data. Based on the concordance of the SNP-array and WES, 91 samples that had high missingness and/or high non-reference discordance (NRD; described below) values were excluded (Figure-2.5B, Figure-2.5C), leaving 4,982 samples. Specifically, we excluded individuals who had

- a raw WES call rate $< 4SD$ from the mean raw WES call rate,
- a post-genotype QC WES call rate $< 2SD$ from the mean post-genotype QC WES call rate,
- or an NRD based on the post-genotype QC WES and post-QC SNP-array data of $> 4SD$ from the mean

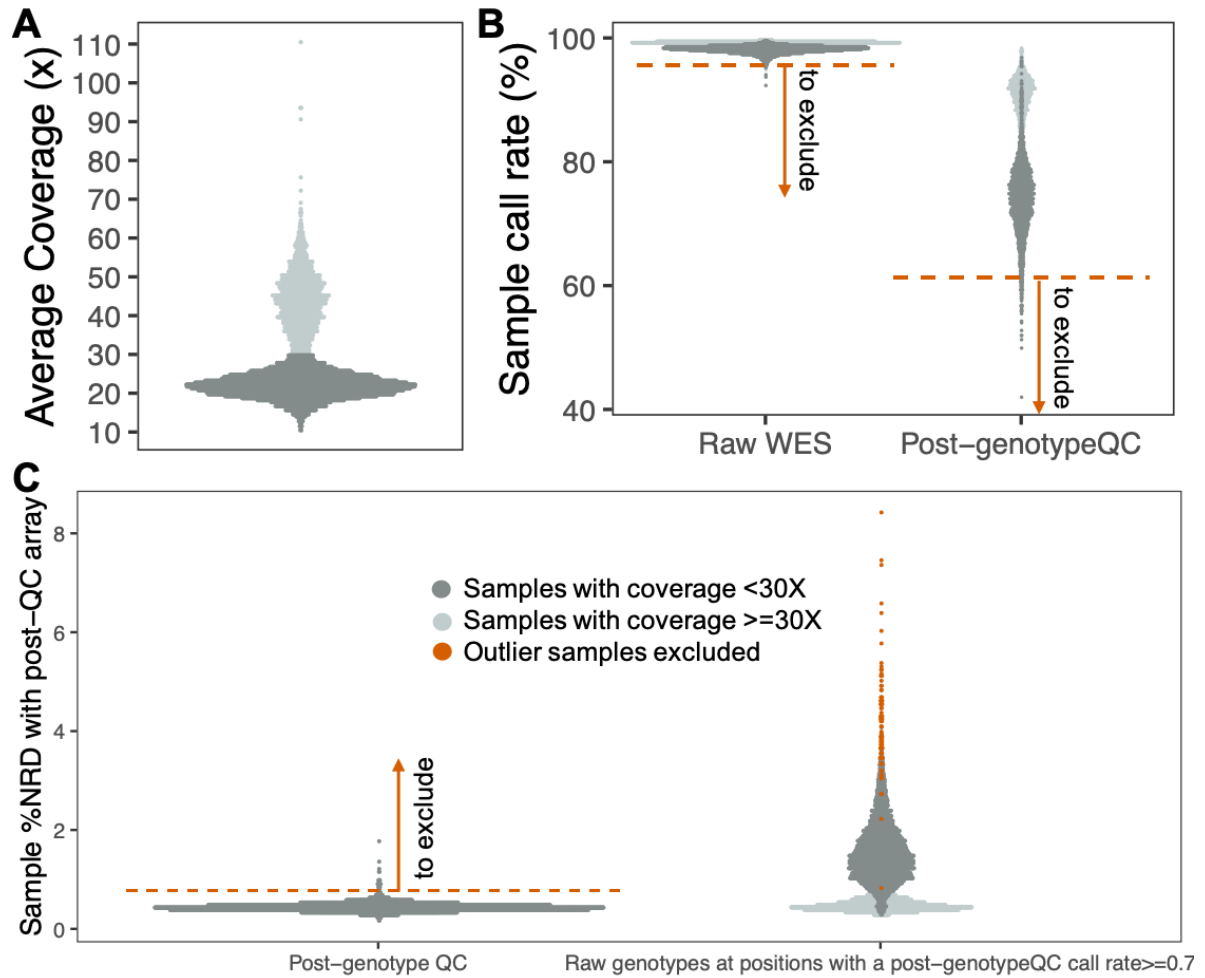


Figure-2.5

Per-sample coverage, call rate, and NRD before and after QC of WES data. **A:** Distribution of average on-target coverage across samples. The bimodality observed is because different batches of samples were sequenced to either ~20X or ~40X **B:** The distribution of WES sample call rates pre-genotype QC (left) and post-genotype QC (right). **C:** The distribution of NRDs across samples in the post-genotype QC WES. In B and C, the bimodality is again likely due to the different sequencing coverage.

To build the reference panel, we filtered the WES data to a subset of high-quality variants that had $\geq 70\%$ call rate after genotype-level QC, and removed singletons. Since Minimac3 does not cope with missing genotypes, we used the genotypes at those sites from the raw data (i.e. pre-genotype QC), and replaced missing genotypes ($\sim 0.02\%$ of the total genotypes, Figure-2.6A) with 0/0 if the reference allele was the major allele and 1/1 otherwise. The distribution of raw per-variant missingness at sites that pass a post-genotype QC call rate $\geq 70\%$ is shown in Figure-2.6B, with most variants having a raw missingness of $< 1\%$.

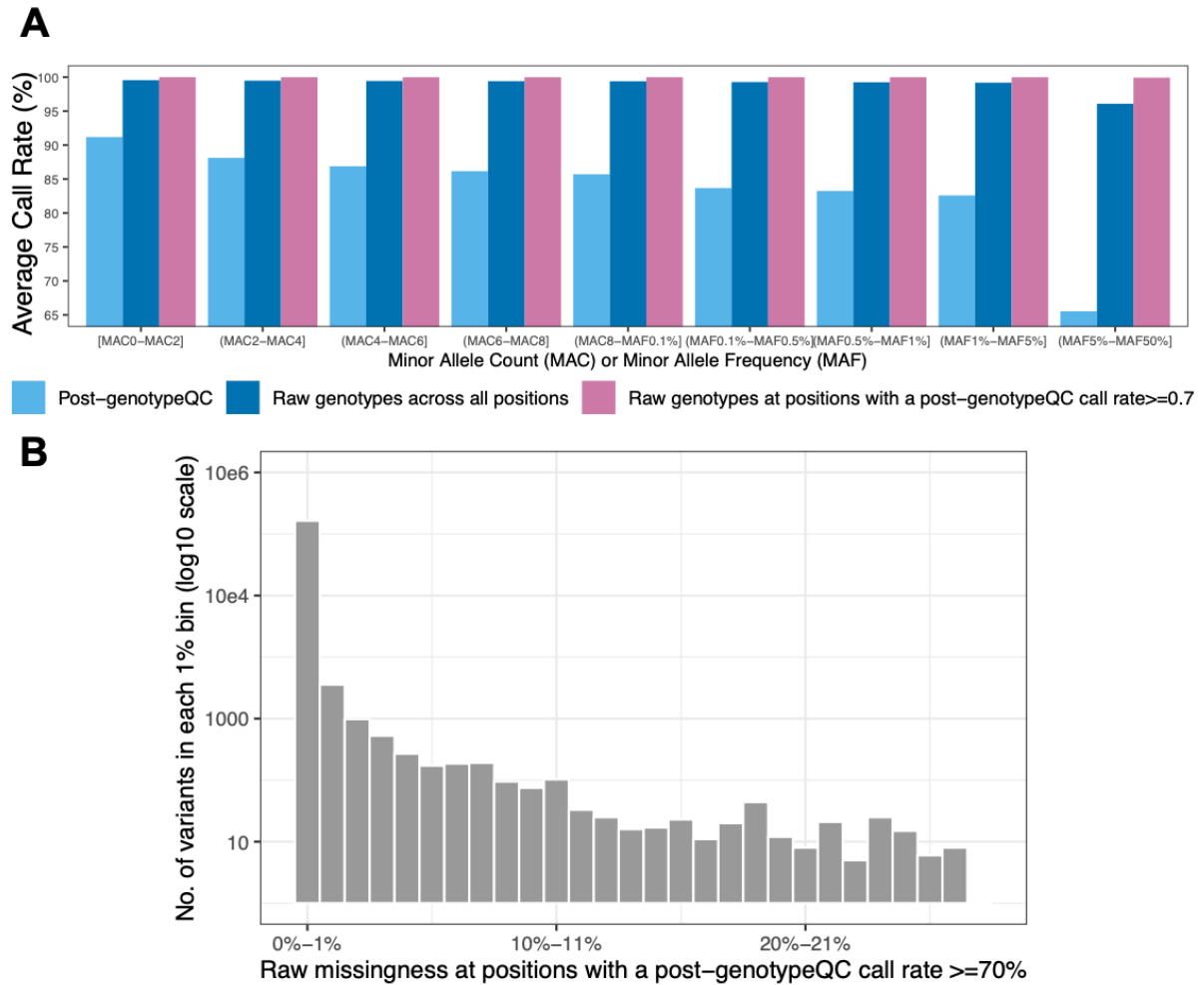


Figure-2.6

Distribution of call rates pre and post QC. A: Average call rate stratified across MAF bins. As the missingness is low in the raw WES, when filtering to positions with a post-genotypeQC call rate $\geq 70\%$, the call rate of the raw genotypes is $>99\%$ across the frequency spectrum. **B:** Distribution of per-variant raw missingness at positions with a post-genotype QC call rate $\geq 70\%$. The y-axis has been transformed to a log 10 scale as an overwhelmingly large number of variants have a raw missingness of 0-1%.

Variants with MAF>0.1% that were also present on the GSA array and had been included in the the imputation backbone were removed from the reference panel, since they tended to have higher call rates in the array data. The cleaned WES and cleaned SNP-array data from the 4,982 samples were then merged to form the reference panel and phased with EAGLE2 (Kpbwt=20,000). The reference panel consisted of 1,385,942 variants.

Performing the imputations

The within-cohort whole-exome imputation

Variants were imputed from the within-cohort whole-exome reference panel into the individuals without WES data with Minimac4 (Das et al., 2016). 5,297 spanning or overlapping deletions (“* alleles”) were not imputed from the reference panel (0.38% of the reference panel), resulting in 1,380,645 variants in the imputed dataset.

To assess the imputation accuracy, ten “leave-10%-out” trials were performed, with each trial leaving out 10% of the exome-sequenced individuals from the reference panel and imputing them into their GSA imputation backbone. Genotypes at variants imputed into those 10% were compared to their genotypes from the WES data.

After these trials, the following filters were applied to the full imputed dataset: imputed $R^2 \geq 0.5$, and as recessive tests would be performed, number of homozygous individuals ($N_{Hom} \geq 3$). This left 605,263 variants (referred to as the “WES5Kimputation” dataset), although the exact number of variants subsequently tested varied depending on the number of individuals with phenotype data, as shown in Table-2.2.

The TOPMed whole-genome imputation

The SNP-array data (after initial QC) was also submitted to the TOPMed-r2 Minimac4 1.5.7 Imputation Server (Das et al., 2016; Fuchsberger et al., 2015; Taliun et al., 2021) for whole-genome imputation against the TOPMED-r2 panel (Performed by David van Heel and Qinqin Huang).

The same filters (imputed $R^2 \geq 0.5$, $N_{Hom} \geq 3$) were applied to extract variants for association testing. The largest filtered dataset, with 10,045,406 variants, will be referred to as the “TOPMEDimputation”.

Evaluating concordance

To assess the imputation accuracy, the imputed genotypes for the 4,982 individuals with WES data were compared to their sequenced genotypes. Genotype concordance was

evaluated between the SNP-array and the WES data at overlapping sites, and between the sequenced and imputed variants.

Concordance was primarily evaluated by the non-reference discordance rate (NRD) calculated using the following formula:

$$NRD = \frac{xRR + xRA + xAA}{xRR + xRA + xAA + mRA + mAA}$$

*R : reference allele**

*A : alternate allele**

x : nMismatches

m : nMatches

*Note that the ‘truth’ was defined by the array genotype when comparing concordance between array and WES data, and defined by the sequenced genotype when comparing concordance between sequenced and imputed data. For example, “xRR” would mean that the truth dataset had a homozygous reference genotype while the other dataset did not, and “mRA” means that both datasets had a heterozygous genotype.

When examining NRD stratified by allele frequency, the NRD was modified to calculate a “minor allele discordance” (MAD) rate:

$$MAD = \frac{xMajMaj + xMajMin + xMinMin}{xMajMaj + xMajMin + xMinMin + mMajMin + mMinMin}$$

*Maj : major allele**

*Min : minor allele**

Where relevant, Pearson correlation r^2 between genotypes / dosages were calculated as a secondary measure of concordance.

Variant annotation

Variants were annotated with Ensembl’s Variant Effect Predictor (VEP) v107, performed by the Wellcome Sanger Institute Human Genetics Informatics team. For each variant, the worst consequence for any transcript was extracted for subsequent analyses.

Phenotype curation

Two lists of phenotypes were curated from participants' electronic health records, and phenotype information was encoded as a binary with '1' coding for a case and '0' coding for a control.

A list of 237 custom phenotypes were compiled manually, with reference to codelists from multiple sources.

A second set of 1281 phenotypes were defined based on International Classification of Disease (ICD10) codes. To derive these, SNOMED codes from primary care data were first converted to ICD-10 codes (for those with 1:1 mapping) using the Interactive Map-Assisted Generation of ICD-10 Codes algorithm. These were then combined with ICD-10 codes from secondary care data and reduced to 3-digit ICD-10 binary encodings from A00 to Q99. This was similar to UK Biobank's 3 digit ICD10 methods (*UK Biobank: Resource 593*, n.d.).

Phenotype curation was performed by the G&H Team at QMUL led by David van Heel, and is described in further detail in the G&H documentation here:

https://docs.google.com/spreadsheets/d/1ipwdF2j_owfr_QbkDYk1rk0TW3KtdfQYVQn-Vf-o38s/edit?usp=sharing

For this study, a phenotype cutoff of ≥ 30 cases was applied. The phenotypes were also classified into phenotypes that affected both sexes, or were sex-specific (i.e. occurred only in females or males). For sex-specific phenotypes, the cohort was filtered to the relevant sex for testing. The number of phenotypes tested is summarised in Table-2.2.

We then calculated a Pearson correlation matrix between phenotype pairs to evaluate the independence of the phenotypes tested. The 898 phenotypes tested are not completely independent of each other, and there may be significant correlation between phenotypes, particularly as some diseases are repeated in the custom list and the ICD10 codes. We sought to calculate Pearson correlation r^2 between phenotype pairs to quantify the degree of correlation in our phenotypic data. This serves two purposes - to evaluate if phenotypes we would expect to be highly correlated indeed have a high r^2 (meaning that they have been curated correctly), and secondly, to try to eliminate highly-correlated phenotypes to reduce the multiple-testing burden.

Of the 402,753 pairs of phenotypes generated, 153 pairs had a correlation r^2 of ≥ 0.5 .

Manually inspecting these pairs, the majority were sensible correlations; for example, the ICD10 encoding for sarcoidosis and for multiple sclerosis fully correlated with the respective

custom encodings for these conditions. There were also correlations between biologically similar phenotypes, such as pulmonary heart disease and pulmonary hypertension ($r^2 = 0.89$), and correlations between pairs for which one trait was a subset of the other, such as acute pancreatitis and pancreatitis ($r^2 = 0.90$). By reviewing these highly-correlated pairs, we estimated that only about 80-90 phenotypes could be excluded due to being highly correlated, as the rest of the pairs had differences in their definitions that warranted the inclusion of both phenotypes.

We therefore tested all 898 phenotypes available, and to be stringent, the Bonferonni cutoff we used accounted for all tests as if they were independent. The correlation r^2 values were still useful, and were used to determine whether we could collapse findings into the gene-to-broad-phenotype pairs that we described above.

Association testing

The tests were performed on REGENIE (Mbatchou et al., 2021) through its two-step pipeline. In step 1, the model was fitted onto variants filtered from SNP-array, starting after initial QC and with the following cut-offs: call rate $\geq 99\%$, MAF $\geq 1\%$ and passing a HWE p-value threshold of 10^{-15} in the inferred Bangladeshi subset. Other specifications included selecting the leave-one-out cross-validation (LOOCV) scheme, a genotype block size of 1000 and using the low memory flag.

Step 2 was performed twice for each phenotype, once under the additive model (default setting), and once under the recessive model (defined under the “--test” flag). As per REGENIE defaults, in the additive tests, the dosages in the PGEN file were used, while in the recessive tests dosages above 1.5 were encoded as 1 while the rest were encoded as 0. The genotype block size was 1000 again and we set a p-value threshold of 0.05, below which the approximate Firth correction was applied.

The number of phenotypes, individuals, and variants tested is summarised in Table-2.2.

Phenotypes	n phenotypes	n individuals	n variants (WES5KImputation)	n variants (TOPMEDImputation)
Custom phenotypes (In both sexes)	199	44,186	605,263	10,045,406
Custom phenotypes (Female only)	15	24,387	553,716	9,215,675
Custom phenotypes (Male only)	5	19,799	529,480	8,855,245
ICD10 codes (In both sexes)	565	42,027	599,758	9,948,708
ICD10 codes (Female only)	99	23,710	550,988	9,171,960
ICD10 codes (Male only)	15	18,317	522,117	8,741,181

Table-2.2

The number of phenotypes, individuals, and variants being tested.

Covariates controlled for in the association testing

The covariates included were age (at year of phenotype curation, 2022), sex, age², age x sex, age² x sex and the first ten principal components (PCs) from the principal component analysis on unrelated G&H individuals described above.

We controlled for ten genetic PCs derived from common variants. These top ten PCs explained more than 85% of the variance explained by the top 50 PCs (Figure-2.7A). (B. B. Sun et al., 2022) suggested that the common variant PCs may not sufficiently control for the population structure captured within rare variants. Therefore, for the recessive loci that involved rare variants (AF <1%), we further investigated if the common PCs were adequately controlling for population structure by checking that the results held when additionally controlling for PCs based on an IBD sharing matrix that should capture more subtle, recent population structure. We used KING to call IBD segments and calculate the length of IBD sharing across pairs of individuals in the cohort, and performed a PCA on the matrix of IBD sharing. We then repeated the association analyses for the recessive loci involving rare variants, adding the first 15 PCs from the IBD PCA to the existing covariates (since these explained most of the variance in the PCA; Figure-2.7B). Reassuringly, the betas (i.e. the effect sizes, linear regression slope 1.03, r^2 0.98, p-value $<2 \times 10^{-16}$, Figure-2.7C), and the p-values (converted to $-\log_{10}p$, linear regression slope 0.67, r^2 0.25, p-value 5.4×10^{-5} , Figure-2.7D), of the associations correlated well between the association tests performed controlling for IBD PCs and the association tests performed controlling for common PCs, with the exception of a single outlier.

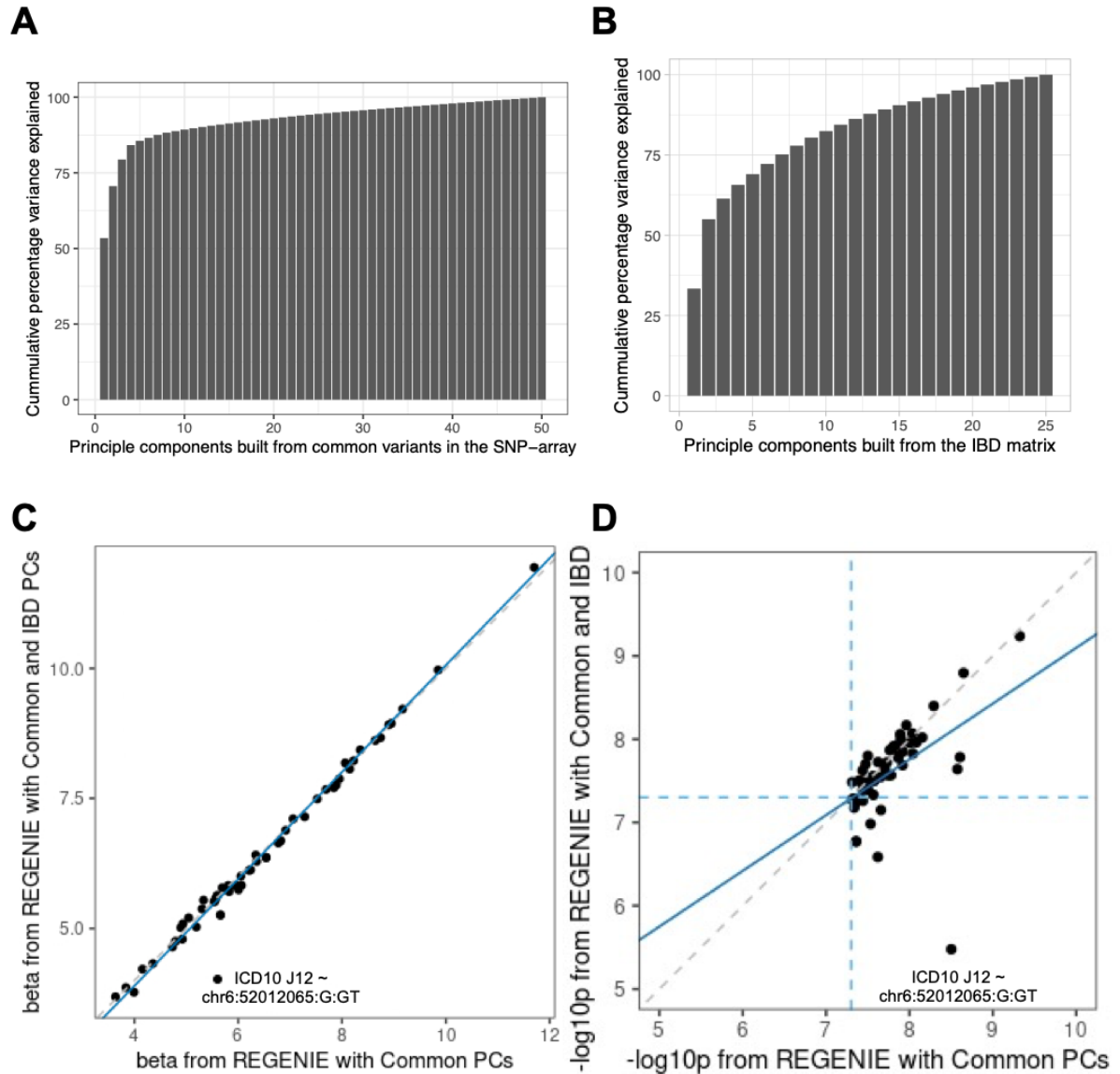


Figure-2.7

Principal component analyses of population structure. **A:** The cumulative percentage variance explained across the 50PCs generated from common variants in the SNP-array, with the variance explained by all 50PCs scaled to 100%. **B:** The cumulative percentage variance explained across the 25PCs generated from the IBD matrix, with the variance explained by all 25PCs scaled to 100%. **C&D:** For the 56 recessive loci associated with rare (AF 1% and lower) variants, the association analysis was repeated controlling for 10 common variant PCs and 15 IBD PCs, and the betas (C) p-values (D) were compared. The blue solid line represents the line of best fit, the grey dashed line represents the $x=y$ line, the blue dashed line represents the $p=5 \times 10^{-8}$ cutoff. The outlier finding has been labelled.

Investigating F_{ROH} as a possible confounder

To investigate if the fraction of the genome in runs of homozygosity (F_{ROH}) was a confounder, selected tests were rerun including F_{ROH} as a covariate. F_{RO} per individual was calculated by PLINK1.9 on the SNP-array data with the following specifications: maximum inverse density 50kb/SNP, maximum internal gap 1000kb, minimum SNP count 50, maximum 1 heterozygous in scanning window hit, maximum 4 missing calls in scanning window hit and a scanning window size of 50. The total length (in kb) was then divided by the length of the autosome (approximately 2700000kb) to obtain F_{ROH} .

Identifying independent loci for significant recessive effects

We defined significant recessive associations as tests with p-values $\leq 5 \times 10^{-8}$ (GWS), and a p-value lower under the recessive model than under the additive model. We then excluded the human leukocyte antigen (HLA) region due to complex LD (chr6:25mb-35mb) and defined independent loci with the following steps:

- For each phenotype with significant tests, identify the test with the most significant recessive p-value. This is assigned as the lead variant.
- Calculate the LD r^2 using PLINK1.9 between the lead variant and variants within a +/- 50kb window of it. Variants with an LD $r^2 \geq 0.8$ with the lead variant are defined as being part of the same locus.
- Identify the next most significant variant in the remaining variants that are not part of the locus defined in the previous steps, and repeat the loop.

These cutoffs to define a locus are taken from Huang et al., 2022, to align with the power-adjusted transferability ratio calculation we perform later to assess replication of our findings in an independent cohort (see section below). These definitions are relatively stringent, in order to be conservative when assessing replication, with the trade-off that our approach may have split up a single recessive locus into multiple separate loci that are probably not fully independent. Indeed, the results in Table 2.6 suggest that this has happened, as we see three clusters of Bonferroni-significant loci that likely all represent a single locus.

Evaluating the fit of the recessive model at loci with significant recessive associations

For the lead variants of the significant recessive associations, we reran the additive test (step 2 of REGENIE) after removing homozygous individuals, to explore the influence of heterozygous and homozygous individuals on the association with phenotypes.

Additionally, we reran the logistic regression in R, controlling for the same covariates: age, sex, age², age x sex, age² x sex and the first ten PCs. We re-coded the genotypes to perform additive, recessive and genotypic tests with 2 degrees of freedom, as shown in Table-2.3 and the equations below. The tests were performed on the full cohort (i.e. the set of individuals used in REGENIE, sample sizes in Table-2.2), as well as on the subset of individuals genetically-inferred to be unrelated by KING (26,579 individuals in the largest cohort, subsetted accordingly depending on the imputation and phenotype tested).

Genotype	Additive	Recessive	Dominance deviation
RR	0	0	0
RA	1	0	1
AA	2	1	0

Table-2.3

Genotype encodings for the additive, recessive and dominance deviation tests performed as logistic regression tests in R. “RR” refers to the homozygous wild-type, “RA” the heterozygous genotype, and “AA” the homozygous alternate genotype.

The models

Additive test

$$phenotype \sim \beta_{add}G_{additive} + BC$$

where $G_{additive}$ is the genotype of the SNP using the additive encoding (0/1/2), β_{add} is the effect size under an additive model, C is a matrix of covariates (defined below) and B is a vector of effect sizes for those covariates.

Recessive test

$$phenotype \sim \beta_{rec}G_{recessive} + BC$$

where $G_{recessive}$ is the genotype of the SNP using the recessive encoding (0/0/1), and β_{rec} is the effect size under a recessive model.

Genotypic (2 degrees of freedom) test, for extracting dominance deviation

$$phenotype \sim \beta_{add}G_{additive} + \beta_{domdev}G_{domdev} + BC$$

where G_{domdev} is the genotype of the SNP using the dominance deviation encoding (0/1/0), and β_{domdev} is the effect size of the dominance deviation under the genotypic model.

Covariates

The covariates included were age, sex, age², age x sex, age² x sex and the first ten PCs.

Finding replication in FinnGen

We sought to replicate our findings in another cohort, FinnGen from Finland. Phenotypes were matched manually to FinnGen binary phenotypes curated from Finnish health registers.

The variants from FinnGen release 10 were then tested with the phenotypes using REGENIE under the recessive model, controlling for the covariates sex, age, 10 PCs and genotyping batch, as described for release 10 on <https://www.finnngen.fi/>. This was performed by our collaborators Juha Karjalainen and Mark J Daly.

Next, we quantified replication. Firstly, for lead variants that could be found in FinnGen, we performed an exact binomial test to investigate if more tests have the same direction of effect than what would be expected by chance (i.e. 50% of the time).

Secondly, we calculated a power-adjusted transferability ratio, a measure described in (Q. Huang et al., 2022). Specifically, for each locus, we first identified proxy variants as

variants that are within a +/- 50kb window of the lead variant with LD $r^2 \geq 0.8$, and a p-value ≤ 100 times the p-value of the lead variant.

Proxy variant defined as:

(a) +/- 50kb from the lead variant

AND

(b) LD $r^2 \geq 0.8$ with the lead variant in G&H

AND

(c) $p_{proxy} \leq 100 * p_{lead}$

A locus was defined as ‘transferable’ if any of the lead or proxy variants had the same direction of effect as observed in G&H at a nominally-significant p-value ≤ 0.05 , when tested in a recessive model with a similar trait.

We then estimated the expected number of loci that we would be powered to detect in FinnGen. For each locus, we used the effect size of the lead variant in G&H, the AF of the lead variant in FinnGen, and the CR and sample size in FinnGen to estimate the power for the test with the *genpwr* R package. The expected number of transferable loci was calculated as the sum of the power estimates across all the loci. If the lead variant is not found in FinnGen, the locus is excluded from the calculation.

$$E_{transferred} = \sum_i^n Power_i$$

Where $E_{transferred}$ is the expected number of loci transferred and $Power_i$ is the power calculated at locus i in n loci.

The power-adjusted transferability (PAT) ratio was then calculated by dividing the observed number of transferred loci over the expected.

$$PAT = \frac{O_{transferred}}{E_{transferred}}$$

Where $O_{transferred}$ is the observed number of loci transferred.

Results

Evaluating the concordance between the SNP-array and the WES data

Before merging the SNP-array and WES data to build the reference panel, we wanted to evaluate concordance of genotypes between them at overlapping sites. QC on both the SNP-Array and WES datasets improved the overall NRD from 5.02% to 0.40%, with the WES GQ filter making the biggest difference (Table-2.4). We anticipated that the genotyping accuracy would be low for rare variants in the SNP-array, but surprisingly the array data continued to demonstrate good MADs of ~0.7-0.9% with the WES for variants with array minor allele counts (MACs) 3-6 (equivalent to MAF 0.03-0.06%, Figure-2.8). We decided to use variants with MAF>0.1% from the SNP array for the imputation backbone.

Since minimac3 does not tolerate missingness in the imputation reference panel, we needed to use the WES data without any genotype-level QC to minimise missingness. To select for variants with a majority of high-quality genotypes, we retained those that had <30% missing genotypes after applying the genotype-level QC. Using the raw genotypes at those sites increased the NRD to 1.26% (Table-2.4). Note that this value is inflated by the discordance at sites with MAC 0-2 (Figure-2.8), but variants this rare were filtered out of the SNP-array when building the imputation backbone, and instead retained in the WES data, since we assume these are likely to be more accurate than the array genotypes given the difficulties of genotyping uber-rare variants on arrays (C. F. Wright et al., 2019).

WES Filters Array Filters	Raw Genotypes (GTs)	GQ ≥ 20	Call rate ≥ 70% after: GQ ≥ 20 binomAD ≥ 10 ⁻² DP > 7	With raw GT at sites with call rate ≥ 70% after: GQ ≥ 20 binomAD ≥ 10 ⁻² DP > 7
After QMUL QC	5.018% (46,778 SNPs)			
Call rate ≥ 99% HWE pval ≥ 10 ⁻⁶	5.000% (46,656 SNPs)	0.500 % (46,401 SNPs)	0.398% (38,249 SNPs)	1.258 % (38,249 SNPs)

Table-2.4

Overall NRD between the array and WES data with different QC filters applied to the SNP-array data and the WES data.

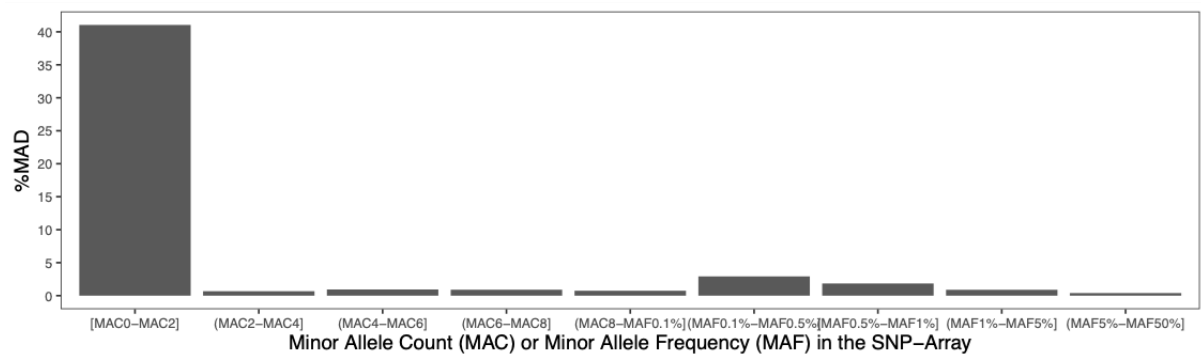


Figure-2.8

The Minor Allele Discordance (MAD) between the SNP-Array and the WES at overlapping variants, stratified by the frequency on the array. Array filters: After QMUL QC + Call rate $\geq 99\%$ + HWE $p\text{val} \geq 10^{-6}$. WES filters: Call rate $\geq 70\%$ after GQ ≥ 20 + binomAD $\geq 10^{-2}$ + DP > 7 . Here, the array genotypes were treated as “truth”.

Evaluating accuracy of imputation with the WES5K panel

After QC, ten trials of imputation were performed using the WES5K panel, with each trial leaving 10% of the WES samples out of the reference panel.

The total number of variants in each imputed dataset ranged from 1,325,855-1,327,697 across the 10 trials. Although the genotyped backbone is imputed by Minimac4 as well, to evaluate imputation accuracy purely at positions with no prior information, we excluded backbone SNPs resulting in 855,697-857,537 variants to compare in each trial. The overall NRD of the imputed genotypes compared to the sequenced genotypes ranged from 7.44-7.69% across 10 trials (Figure-2.9A).

The confidence of the imputation is quantified by Minimac4's imputed R^2 . By convention, a minimum cutoff of ≥ 0.3 is applied to QC imputed data, which should sufficiently filter out poorly-imputed variants based on the distribution of imputed R^2 scores (Figure-2.9B). However the overall NRD doesn't improve yet at this cutoff (Figure-2.9A) and it was decided that a more stringent cutoff of ≥ 0.5 should be applied. As expected, the rarer MAF bins had more variants with a lower imputed R^2 (Figure-2.9C), and it is likely that the inclusion of rarer variants has driven the need to apply a more stringent cutoff, as the conventional cutoff is based on common variants.

MAD also increases with decreasing allele frequencies, as expected. Many of these uber-rare variants will not be submitted for association testing as they are both poorly imputed and not powered enough for recessive tests.

After applying imputed $R^2 \geq 0.5$ and $N_{\text{Hom}} \geq 3$ to prepare for association tests, and including the positions of the SNP-array backbone (which will be included in association tests), the overall NRD improved to 1.78-1.84% across the trials (Figure-2.10).

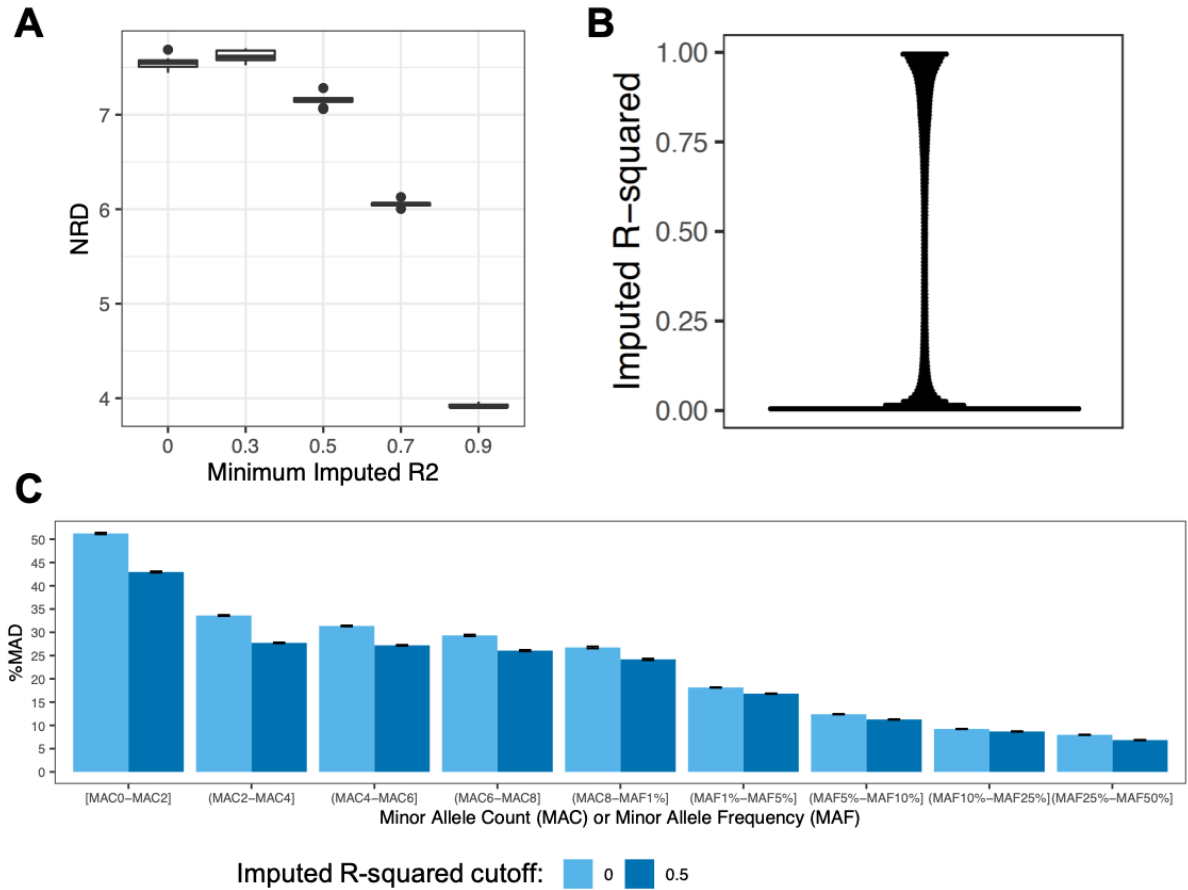


Figure-2.9

Results of the leave-10%-out trials. A: The distribution of NRD across 10 trials when different minimum imputed r^2 cutoffs are applied. **B:** The distribution of imputed r^2 scores for a representative trial. **C:** The MAD before and after applying an imputed $r^2 \geq 0.5$ cut-off, stratified by MAF. (Error bars are the standard errors in %MAD across the 10 trials.)

Comparing imputation accuracy with TOPMED versus WES5K

The TOPMEDimputation genotypes of the 4,982 samples with WES were compared to their sequenced genotypes. The same cutoffs (imputed $R^2 \geq 0.5$ and $N_{\text{Hom}} \geq 3$) were applied to result in 10,045,406 variants. Of these variants, 523,018 variants overlapped with the WES and at these, the overall NRD was 1.19%.

We then considered NRD at variants stratified by the number of homozygotes (as defined in the WES5Kimputation; the rationale being that the WES should be considered “truth” in this case, and that the power of the recessive association testing depends on the number of homozygotes rather than directly on the allele frequency). First, we observed that the within-cohort WES5K reference panel allowed for more variants to be imputed at the chosen level of accuracy, especially at lower MAFs (Table-2.5). Second, for $N_{\text{Hom}} \geq 9$ up to a homozygous frequency of 5%, the MAD is lower for the WES5KImputation compared to the TOPMEDImputation, but the opposite is true for the lowest and highest MAF bins. Thirdly, the lower overall NRD of the TOPMEDimputation (1.19% versus ~1.7% for the WES5K Panel) is driven by the improved accuracy of common variant imputation with the TOPMED reference panel. (Figure-2.10)

As both imputation sets carried their own strengths and weaknesses across the frequency spectrum, it was decided to bring both sets forward to association testing.

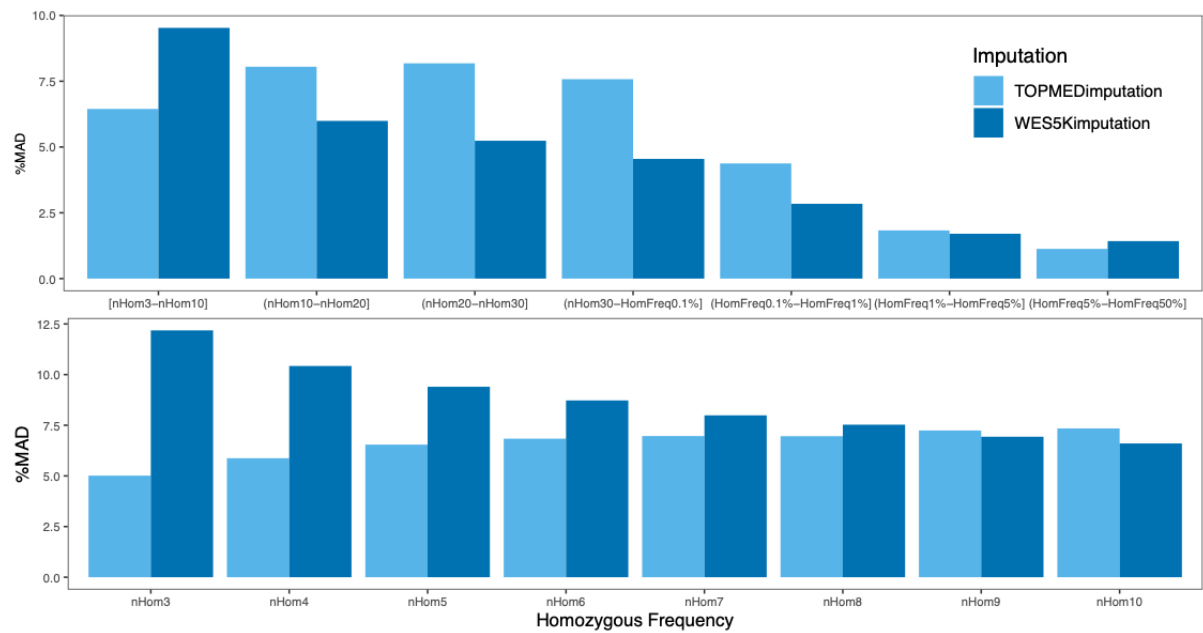


Figure-2.10

MAD of the TOPMEDimputation compared to the WES, and the mean MAD across the leave 10% out trials compared to the WES (represented as the WES5Kimputation for simplicity), stratified by homozygous frequency. The variant counts contributing to each bin is tabulated below.

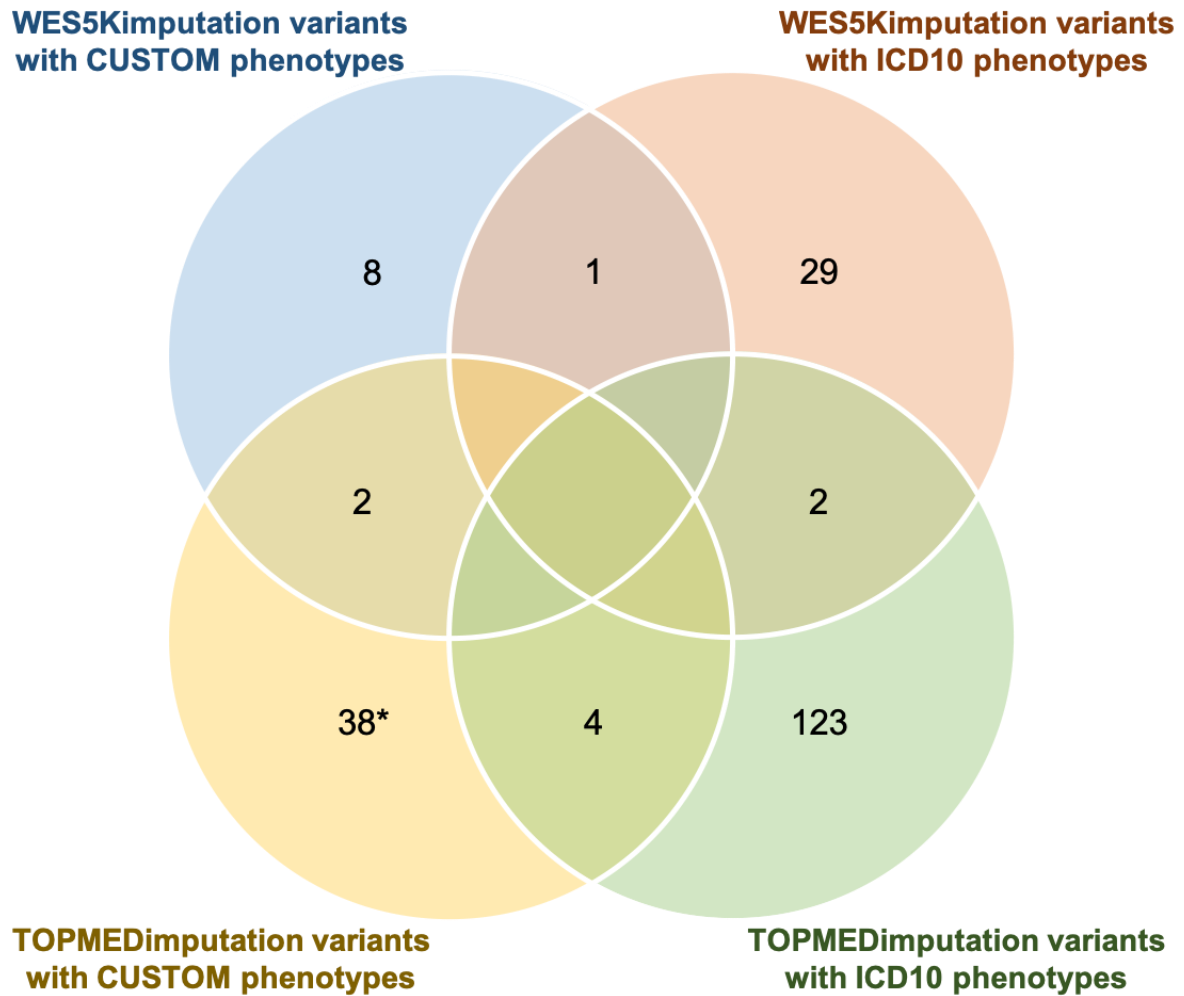
	n variants (TOPMEDimputation)	n variants (WES5Kimputation)	WES5Kimputation / TOPMEDimputation
nHom 3	10,806	26,001	2.41
nHom 4	10,855	19,588	1.80
nHom 5	9,771	15,532	1.59
nHom 6	8,725	12,615	1.45
nHom 7	7,701	10,546	1.37
nHom 8	6,875	9,044	1.32
nHom 9	6,233	7,864	1.26
nHom 10	5,519	6,774	1.23
(nHom 10 – nHom 20]	35,924	41,259	1.15
(nHom 20 – nHom 30]	18,717	20,156	1.08
(nHom 30 – HomFreq 0.1%]	17,065	17,907	1.05
(HomFreq 0.1% - HomFreq 1%]	109,350	111,916	1.02
(HomFreq 1% - HomFreq 5%]	106,049	109,146	1.03
(HomFreq 5% - HomFreq 10%]	151,338	173,978	1.15

Table-2.5

Number of imputed variants (that can be compared to WES positions for concordance analyses) binned by homozygous frequency.

Recessive association testing identifies 207 loci

We identified 207 unique loci across 898 phenotypes, where the lead variant had a GWS recessive p-value that is smaller than the additive p-value. 11 lead variants were found multiple times, such as being found in both phenotype sets (custom and icd10), or through both imputations (WES5Kimputation and TOPMEDimputation) (Figure-2.11). The p-values for the set of variants significant on one or the other panel are highly correlated between the two imputation panels, as we will show in Figure-2.14. However, the lead variant at a given locus often differs between the imputation panels, resulting in fewer overlaps in Figure-2.11 when only comparing among lead variants. There is also correlation between CUSTOM and ICD10 phenotypes, but as we have decided to test these phenotypes independently (see methods), and correct for them stringently in the multiple testing correction), in Figure-2.11 we were also stringent in defining an “overlap” in a CUSTOM and ICD10 phenotype - the phenotype description had to match almost exactly.



* 2 lead variants here were found in multiple highly-correlated CUSTOM phenotypes. Specifically, one lead variant was found to be associated with 2 CUSTOM encodings for type 2 diabetes mellitus and the other lead variant was associated with 2 CUSTOM encodings for hypertension

Figure-2.11

A Venn diagram illustrating which imputation panel and phenotype set was used to discover the 207 recessive loci.

The frequency and consequence distributions of the lead variants are shown in Figure-2.12. Despite applying an exome reference panel including many rare protein-coding variants, many of the findings from the WES5K imputation were still common (Figure-2.12A) and intronic (Figure-2.12B), implicating variants near exonic regions that happened to be captured with WES, or common SNPs from the GSA backbone. Similarly, the majority of the significant hits from the whole genome TOPMEDimputation were within the noncoding regions (Figure-2.12B). This is expected as common variants would have better power and 99% of the human genome is noncoding. However, when evaluating the lead SNPs as a proportion of the variants tested, we find that significant SNPs were enriched for coding variants (Figure-2.12C). To test this statistically, we evaluated the 10,045,406 variants of the TOPMED imputation, among which 91,930 were coding variants. We observed significant enrichment of recessive associations in coding variants (Chi-square test P-value = 8.3×10^{-20}), with 0.044% having significant associations compared to 0.011% in non-coding variants. We observed significant enrichment for coding variants amongst common variants (MAF > 5%; p-value = 2.3×10^{-16}) and low-frequency variants (MAF 1–5%; p-value = 2.4×10^{-8}) but not in rare (MAF 0.1–1%; p-value = 0.66) or ultra-rare variants (MAF < 0.1%; no coding variants reached significance in recessive testing). The enrichment analysis for the rare and ultra-rare groups is likely underpowered, since only 94 and 9 variants, respectively, showed significant recessive associations. These findings are consistent with the expectation that protein-coding variants are more likely to impact gene function and health outcomes, and that they have larger effect sizes leading to better power for detection.

56 loci had lead variants in the noncoding region without a gene annotation. The remaining 151 loci had lead variants that had been annotated with a gene by VEP (although we note that, of course, this is not necessarily the causal gene). Some of these loci have lead variants annotated to have their worst consequences on the same gene. Furthermore, some phenotypes are highly correlated and can be classified into a more general term. As a result, these loci can be collapsed into 117 gene-to-broad-phenotype pairs.

Across imputation sets and phenotype sets, a total of 9,197,933,046 tests were performed. Because of the high number of tests performed and the inclusion of rare variants, the genome-wide significant cutoff (which is calculated for common variants) arguably may include false positives. At a stringent Bonferroni cutoff ($0.05/9,197,933,046 = 5.5 \times 10^{-12}$), 32 loci remain, corresponding to 14 gene-to-broad-phenotype pairs and 2 noncoding loci. The more lenient Benjamini–Hochberg cutoff for a false discovery rate of 5% (FDR 5%) was $10^{-7.43}$ (which is close to the genome-wide significance cutoff) resulted in 167 loci corresponding to 91 gene-to-broad-phenotype pairs and 45 noncoding loci.

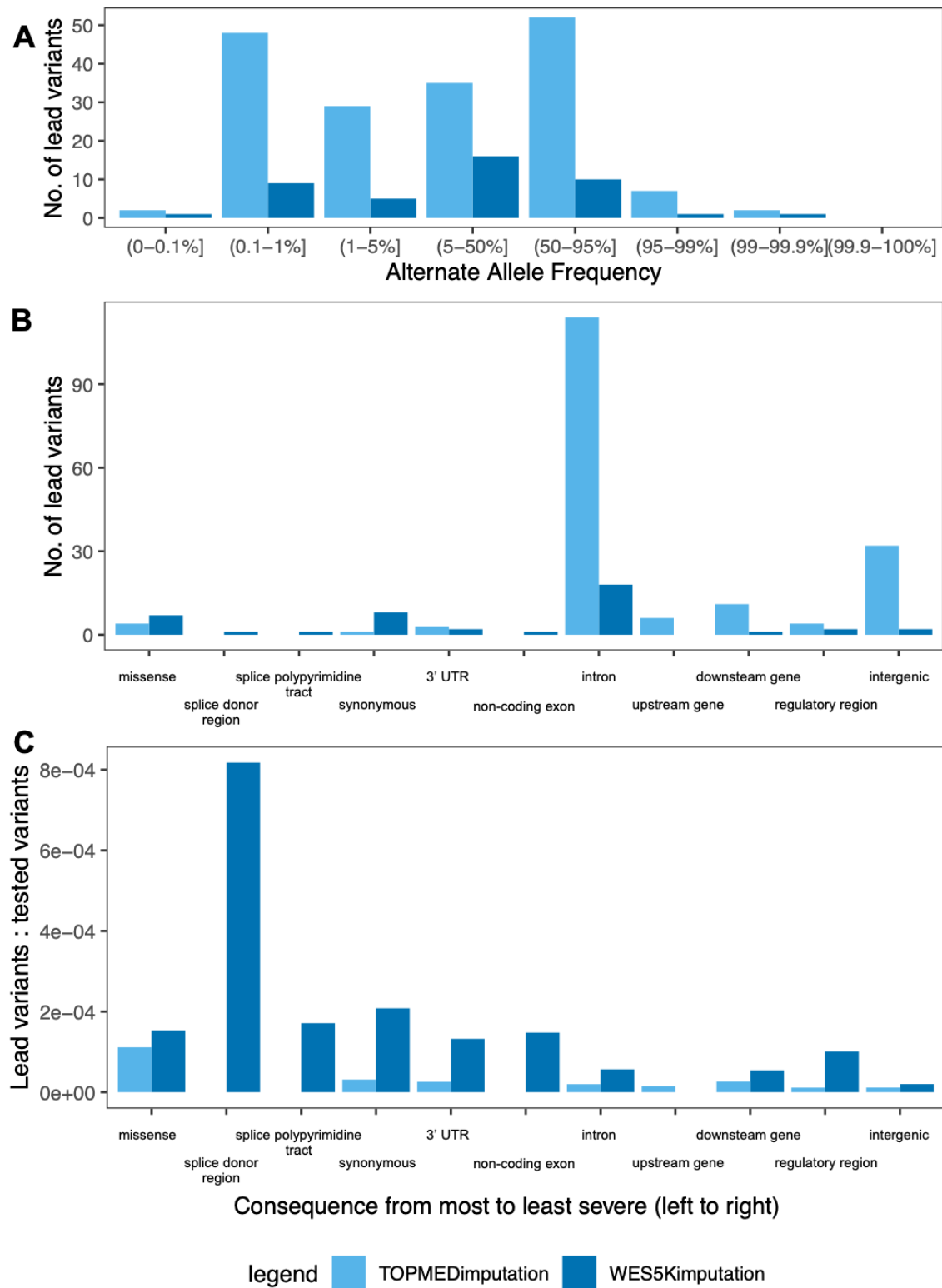


Figure-2.12

Characteristics of lead SNPs in the recessive findings. **A:** Distribution of the allele frequencies of the lead SNPs. **B:** Distribution of the variant consequences of the lead SNPs. **C:** The ratio of the number of lead SNPs over the number of variants tested for each variant consequence.

The Bonferroni-significant findings have been summarised in Figure-2.13 and Table-2.6. The Manhattan plot (Figure-2.13A) illustrates that the Bonferroni-significant hits cluster around 3 regions of the genome, with each region associated with the same or similar phenotypes. It is likely that given our strict cutoffs to define loci, that the loci in each region are not entirely independent. Nonetheless, we stuck with the stricter definition to be conservative in assessing replication in the independent cohort, FinnGen. Also, several of the findings have an AF > 0.5, and we hypothesise that these findings were detected likely because alternate alleles with an AF > 0.5 have more power. It was decided not to switch the alleles for these tests (to test the minor allele), because there are AF differences across cohorts and populations, and in order to perform replication studies, we standardised the reference allele.

The quantile-quantile (QQ) plot ($\lambda = 0.66$, Figure-2.13B) of all the recessive tests performed for these phenotypes with Bonferroni-significant findings suggest that the tests are underpowered, as we see substantial deflation of the test statistics below what is expected under the null. When we split the variants contributing to these tests into common (AF > 5%) and low frequency variants (AF $\leq$ 5%), indeed we see that the λ of 0.97 for the common variants is close to 1 (Figure-2.13C), while the λ for the rare variants is low at 0.31 (Figure-2.13D), indicating that the deflation is likely due to reduced power for rare variants.

In Table-2.6, for variants with AF > 0.5, the alternate allele has still been tested to standardise the tested allele across replication cohorts. Findings with variants of AF $\sim$ 1 have been retained and reported, though it is worth noting that a non-additive test on the reference allele may be more valuable in the future. However, many of these variants with AF $\sim$ 1 are in the chromosome 11 genomic region associated with hereditary anaemias, and it is possible with wider locus definitions, they may collapse into one locus.

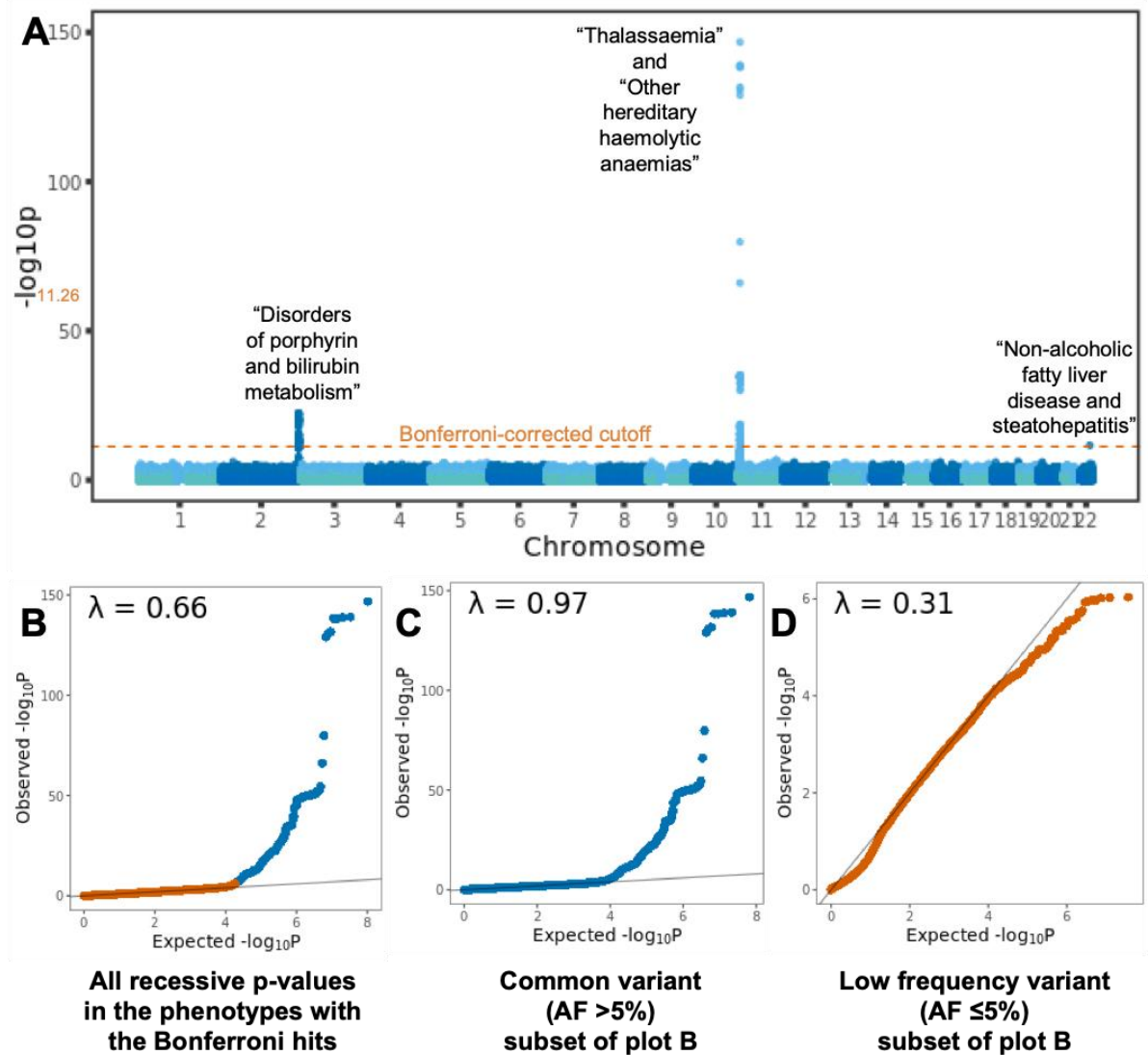


Figure-2.13

Manhattan plot (A) and quantile-quantile (QQ) plots (B-D) for the phenotypes that have Bonferroni-significant findings. In **A**, recessive p-values that are smaller than additive p-values have been plotted. In **B**, all the recessive p-values in the phenotypes with the Bonferroni-significant findings have been plotted, with low frequency variants in red and common variants in blue. We then split the QQ-plot into common variants (AF > 5%, **C**) and low frequency variants (**D**).

Gene	Lead variant	AF	CSQ	Phenotype	OR	p-value	delta log10p	DD p-value	FG OR	FG p-value
MROH2A	2:233780897:C:G	0.44	intron	E80[Disorders of porphyrin and bilirubin metabolism]	5.5	1.6E-16	1.4	8.4E-03	3.7	4.5E-31
	2:233790144:A:G	0.46	intron		5.0	2.7E-14	0.8	6.7E-02	3.3	4.1E-20
UGT1A5	2:233713783:T:C	0.78	synonymous	E80[Disorders of porphyrin and bilirubin metabolism]	3.9	2.0E-12	0.4	3.7E-01	4.1	9.0E-52
UGT1A6	2:233763993:C:A	0.53	intron	E80[Disorders of porphyrin and bilirubin metabolism]	6.4	5.3E-23	2.9	3.8E-03	6.0	2.3E-74
	2:233763993:G:T	0.53	intron		6.4	5.3E-23	2.9	3.8E-03	6.0	2.3E-74
	2:233738671:G:A	0.50	intron		6.1	1.8E-20	2.0	2.0E-02	5.5	1.1E-59
	2:233715640:T:C	0.53	intron		5.8	1.5E-19	2.6	4.2E-03	5.0	7.5E-63
	2:233716651:G:A	0.51	intron		5.8	3.2E-19	2.2	1.2E-02	5.0	1.8E-55
	2:233703893:C:A	0.45	intron		5.5	4.7E-16	1.9	6.5E-03	5.5	1.7E-57
	2:233702448:A:C	0.46	intron		5.5	5.6E-16	1.8	9.7E-03	5.5	2.0E-57
	2:233757013:T:G	0.75	intron		4.4	1.1E-14	0.3	8.0E-01	4.5	3.5E-58
	2:233750066:G:A	0.37	intron		5.8	2.7E-14	0.8	4.6E-02	4.5	3.3E-36
	2:233743701:C:G	0.37	intron		5.7	5.1E-14	0.6	4.6E-02	4.5	2.6E-36
	2:233693319:A:G	0.58	synonymous		4.4	1.3E-13	0.2	2.9E-01	5.0	1.1E-59
	2:233693556:A:C	0.58	missense		4.4	1.4E-13	0.2	2.9E-01	5.0	1.1E-59
.	2:233755003:C:G	0.76	splice donor region	E80[Disorders of porphyrin and bilirubin metabolism]	4.3	3.9E-14	0.3	8.0E-01		
HBE1	11:5478732:G:C	0.81	intron	D56[Thalassaemia]	0.6	1.9E-16	1.7	6.3E-04		
	11:5330439:C:T	0.71	intron	Thalassaemia	0.4	1.7E-12	0.6	1.1E-02		
	11:5367606:A:C	0.98	intron	D58[Other hereditary haemolytic anaemias]	0.3	6.2E-36	0.5	1.6E-03		
	11:5401996:G:T	0.93	intron		0.5	2.5E-12	0.6	2.0E-02		proxy replicated
HBG1	11:5247392:T:G	0.61	downstream gene	D58[Other hereditary haemolytic anaemias]	0.4	2.3E-19	5.1	9.0E-05		proxy replicated
MMP26	11:4931228:C:A	0.79	intron	D58[Other hereditary haemolytic anaemias]	0.4	6.2E-18	0.7	3.5E-03		proxy replicated
	11:4930728:C:T	0.73	intron		0.4	4.4E-15	0.2	6.3E-03		proxy replicated
OR51A7	11:4908482:G:A	0.95	3 prime utr	D58[Other hereditary haemolytic anaemias]	0.3	9.6E-67	0.5	1.9E-08		
OR51L1	11:5005537:T:C	0.97	downstream gene	D58[Other hereditary haemolytic anaemias]	0.2	1.6E-80	0.3	2.1E-12		
OR51V1	11:5197578:AAT:A	0.55	downstream gene	D58[Other hereditary haemolytic anaemias]	0.2	1.0E-30	1.6	6.5E-07		proxy replicated
OR52D1	11:5489111:C:T	0.81	synonymous	D56[Thalassaemia]	0.6	4.0E-13	1.5	4.9E-03		
OR52E1	11:5067264:A:G	0.99	upstream gene	D58[Other hereditary haemolytic anaemias]	0.1	8.7E-140	5.4	1.5E-14		
OR52E2	11:5058716:C:T	0.99	synonymous	D58[Other hereditary haemolytic anaemias]	0.1	1.7E-147	7.2	7.9E-19		
OR52H1	11:5544800:A:G	0.99	missense	D58[Other hereditary haemolytic anaemias]	0.3	3.6E-33	2.9	3.7E-04		proxy replicated
.	11:5216054:A:G	0.76	intergenic	D56[Thalassaemia]	0.6	5.4E-16	1.4	3.1E-03		
PNPLA3	22:43939790:TGG:T	0.61	intron	Non-alcoholic fatty liver disease and steatohepatitis	1.3	2.4E-12	0.9	3.0E-02	1.3	6.4E-12

Table-2.6

Table of recessive findings that were Bonferroni-significant. “CSQ”: VEP consequence annotation of the variant. “OR”: Odds ratio, converted from beta in the test output. “delta log10p”: The difference between the -log10p in the recessive test and the -log10p in the additive test. “DD p-value”: Dominance deviation p-value (see section below). If the lead variant was replicated in FinnGen (see section below), the FinnGen OR and p-value are listed here. “proxy replicated”: the lead variant is not replicated, but a proxy variant in the locus is replicated.

Correlation between the findings from the WES5K imputation versus TOPMED imputation

Among the combined lead variants, 39 were successfully tested in both imputation sets (found at adequate numbers in both datasets, with the same reference and alternate allele). Effect sizes and allele frequencies correlated well (Figure-2.14), though interestingly, a small subset of findings significant with the WES5K imputation had much less significant p-values when their TOPMED imputation genotypes were tested. These findings corresponded to variants with lower imputed R-squared values in the TOPMED imputation compared to the WES5K imputation (Figure 2.10 bottom right), suggesting that higher confidence in the imputation had improved the sensitivity of the testing through more accurate prediction of the genotypes.

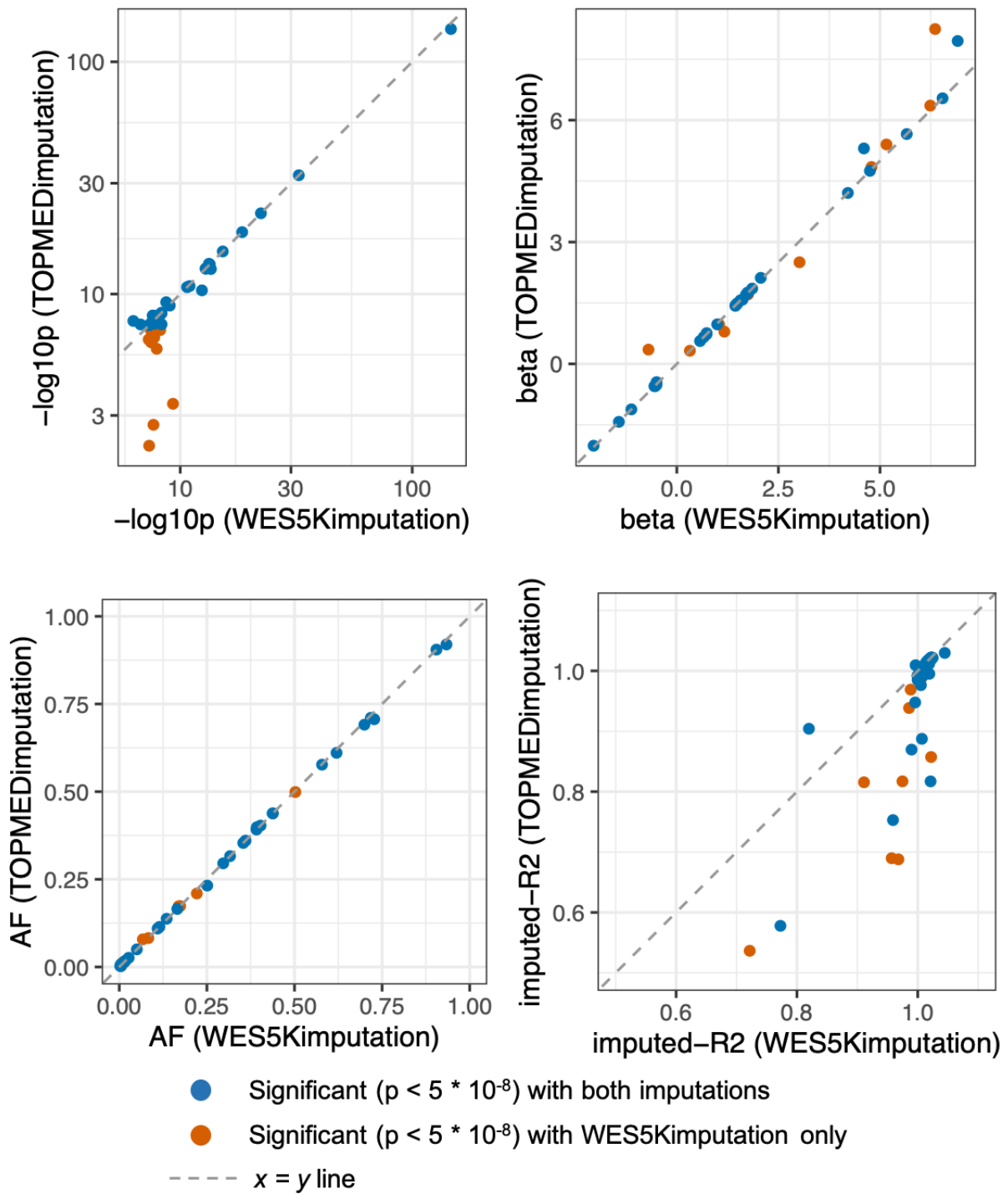


Figure-2.14

Scatter plots of the p-values ($-\log_{10}p$), betas, AFs, imputed-Rsquared values of lead variants that can be found in both imputations.

Investigating the influence of F_{ROH} on the recessive findings

Previous work in the lab by Malawsky et al., 2023 (Malawsky et al., 2023) demonstrated that increased runs of homozygosity (ROH) in the genome was associated with several conditions such as anxiety and type 2 diabetes mellitus. We sought to explore if the fraction of the genome in ROH (F_{ROH}) is a possible confounder for these recessive findings by adding it as a covariate and rerunning the association tests for the 42 leadSNPs in the WES5Kimputation dataset. The recessive p-values correlate well with or without controlling for F_{ROH} suggesting that it is not a confounder that needed to be controlled for (Figure-2.15).

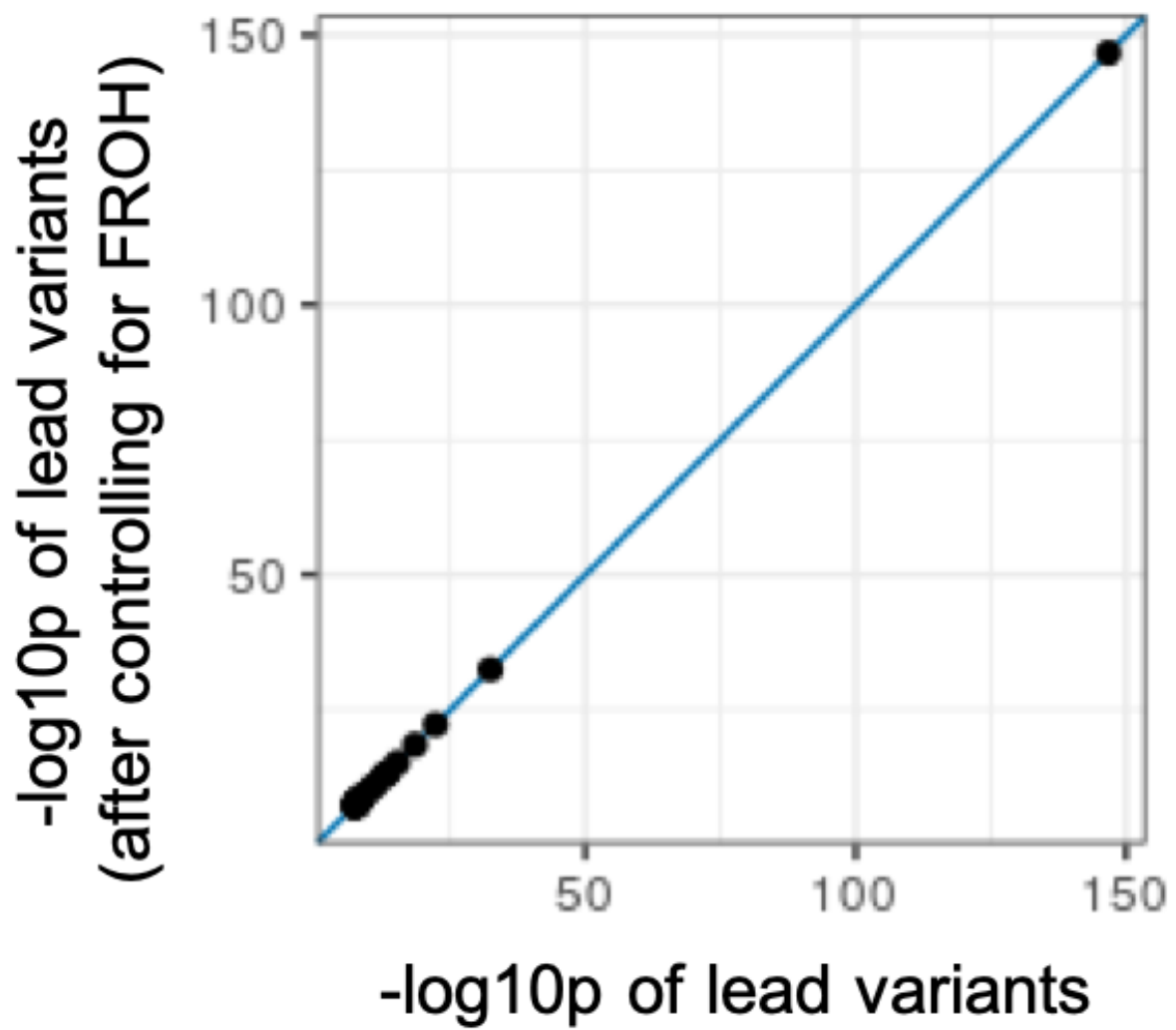


Figure-2.15

Scatter plot of the WES5Kimputation lead variants p-values ($-\log_{10}p$) compared to their p-values when the recessive test is rerun to control for FROH.

Exploring the recessive associations with other models

For the lead variants of the 207 recessive loci, we first explored the correlation between the p-values in the recessive and additive tests. 146 lead variants were not GWS in the additive test, suggesting that they would have been missed in conventional additive GWAS-es (Figure-2.16D). Looking at the distribution of AFs, recessive hits that were GWS in the additive test had higher AFs than hits that were insignificant in the additive test, suggesting that common variants simply had more power to be detected under the additive model even if the underlying pattern of inheritance might be recessive. (Figure-2.16A).

In the additive test without individuals homozygous alternate genotypes, all but two lead variants had p-values below GWS, demonstrating both the weight of these homozygotes on the significant results in the original additive tests and the loss of power by reducing the sample size of the tests. The remaining two lead variants likely have strong heterozygous effects that can be detected even at reduced power. (Figure-2.16E)

It is possible that the tests that dropped below GWS after homozygotes were removed may still have heterozygous effects. We compared the betas in the different models of testing. As expected, additive tests estimated betas that were smaller in magnitude compared to the recessive tests (Figure-2.16F). After dropping the homozygotes, 113 of these tests became insignificant ($p\text{-value} > 0.05$), but for tests that remained nominally significant, their betas correlated well with the full additive tests, though again, slightly underestimated in terms of their magnitudes (Figure-2.16G). These nominally significant additive tests performed without homozygous individuals demonstrate heterozygous effects despite reduced power.

When we removed the homozygous individuals and reran the additive tests, a small minority of tests (26/207 lead variants) could not be run in REGENIE when the homozygotes were excluded. These tests tend to have higher AFs (suggesting that the removal of homozygotes resulted in the test being too underpowered) (Figure-2.16B). All tests were run in R (described further below), and the results from these gave similar conclusions, and are plotted in Supplementary-figure-2.2.

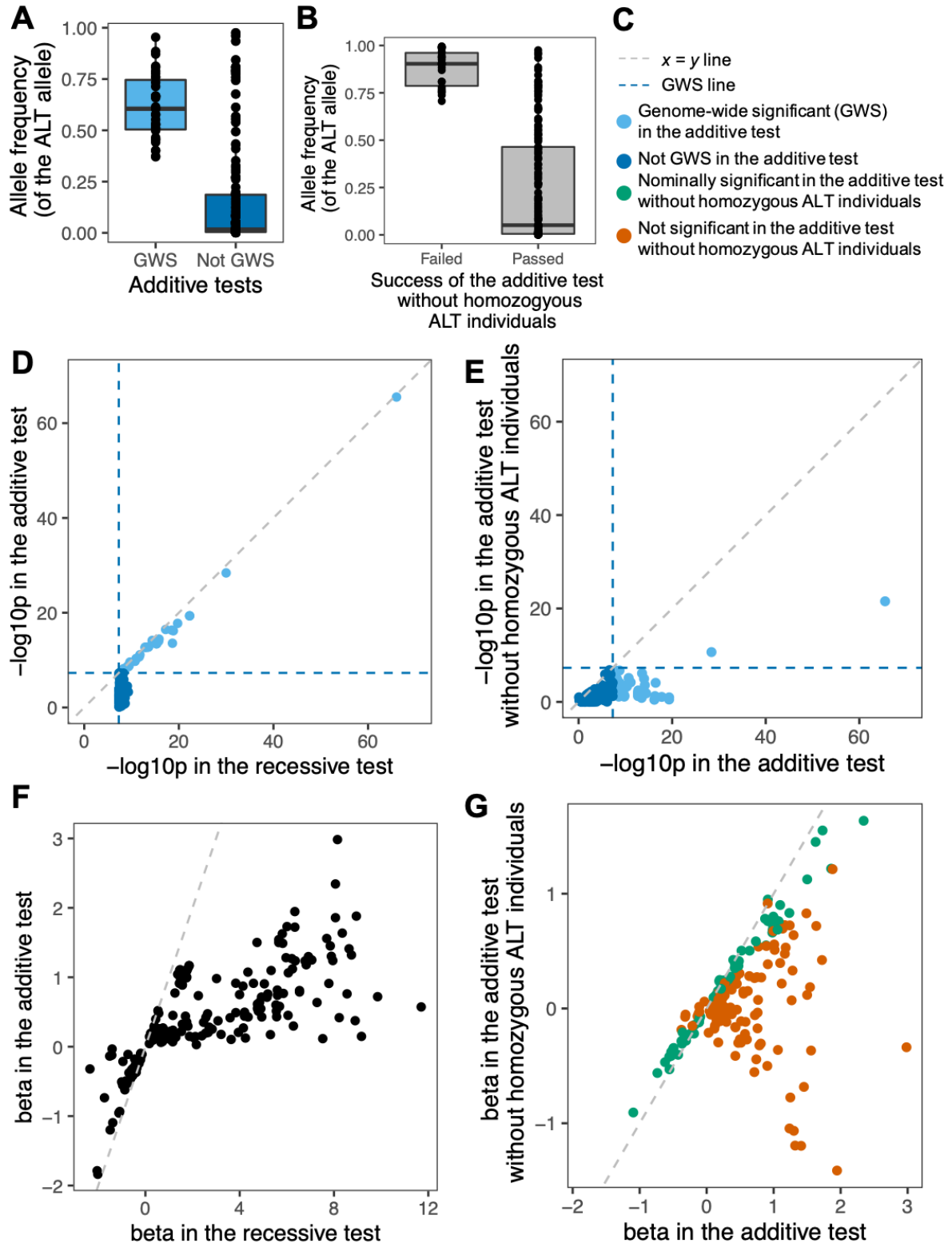


Figure-2.16

Exploring the recessive lead variants with different models of testing in REGENIE. A: AF distribution of recessive hits that were GWS and not GWS in additive tests. **B:** AF distribution of recessive hits that could and could not run under the additive model without

homozygous individuals. **C**: legend for panels D-G. **D,F**: P-values ($-\log_{10}p$, **D**) and betas (**F**) in the recessive tests compared to the additive tests. **E,G**: P-values ($-\log_{10}p$, **E**) and betas (**G**) in the additive tests compared to the additive tests without homozygotes.

To further explore whether the recessive model was indeed the best fit for the variants that were significant on the recessive test in REGENIE, we performed logistic regression testing in R for these findings controlling for the same covariates, using a genotypic model that included a dominance deviation encoding (Table-2.3 and Methods). For comparison, we also fitted a standard additive and recessive model in R. We found that the $-\log_{10}(\text{p-values})$ correlated well between the logistic regression tests in R and REGENIE (Figure-2.17A, linear regression slope = 1.1, $r^2 = 0.95$). This correlation still held when restricting to a set of 26,579 unrelated individuals (in the largest set) in R, though unsurprisingly there was less power (linear regression slope = 0.6, $r^2 = 0.92$). Still, there were some REGENIE tests that were not significant ($\text{p-value} > 0.05$) in the R logistic regression, and these outliers had much larger effect size estimates in R (Figure-2.17B). Otherwise, for tests that were nominally significant in R, their betas correlated well with REGENIE (linear regression slope = 0.98, $r^2 = 0.97$). The outlier R tests that had insignificant p-values tended to involve rarer variants (Figure-2.17C, Wilcoxon two-sided $\text{p-value} = 10^{-5}$), although the distribution of phenotype case counts were similar (Figure-2.17C, Wilcoxon two-sided $\text{p-value} = 0.17$). They were excluded from subsequent analyses described below.

74% (or 154) of these lead variants had nominally significant dominance deviation p-values. These variants tended to be at least nominally significant in the R recessive test as well (chi-squared $\text{p-value} = 2 \times 10^{-11}$, Figure-2.17D). For the tests that had an insignificant dominance deviation p-value, we cannot rule out that the underlying inheritance pattern may still be recessive, especially as some also had recessive tests with insignificant p-values in R, suggesting that R is an imperfect model to replicate the results from the more complex model fitted by REGENIE. We did not find any difference in the distribution of AFs and case counts between the hits with nominally significant and insignificant dominance deviation p-values (Wilcoxon two-sided $\text{p-value} = 0.4$ and $= 0.9$ respectively, Figure-2.17I).

We explored the relationship between the recessive and dominance deviation p-values further. As expected, we find that the $-\log_{10}p$ of both tests are correlated (linear regression slope = 3.7, $r^2 = 0.3$, $\text{p-value} < 2.2 \times 10^{-16}$, Figure-2.17E), with hits having a nominally significant dominance deviation tending to have lower recessive p-values (Wilcoxon two-sided $\text{p-value} = 10^{-3}$, Figure-2.17F). Next, we also saw that hits having a nominally significant dominance deviation also tended to have a larger difference between their recessive and additive p-values (Wilcoxon two-sided $\text{p-value} = 5 \times 10^{-11}$, Figure-2.17H). This suggested that taking hits with a more significant p-value in a recessive model compared to an additive model is indeed an effective method to identify recessive associations, and a potential approach to filtering to more “strictly-recessive” variants would be to set a cutoff on the

difference between recessive and additive p-values (e.g. at least two orders of magnitude). Notably, the recessive betas do not differ between hits with and without a nominally significant dominance deviation (Wilcoxon two-sided p-value =0.4, Figure-2.17G).

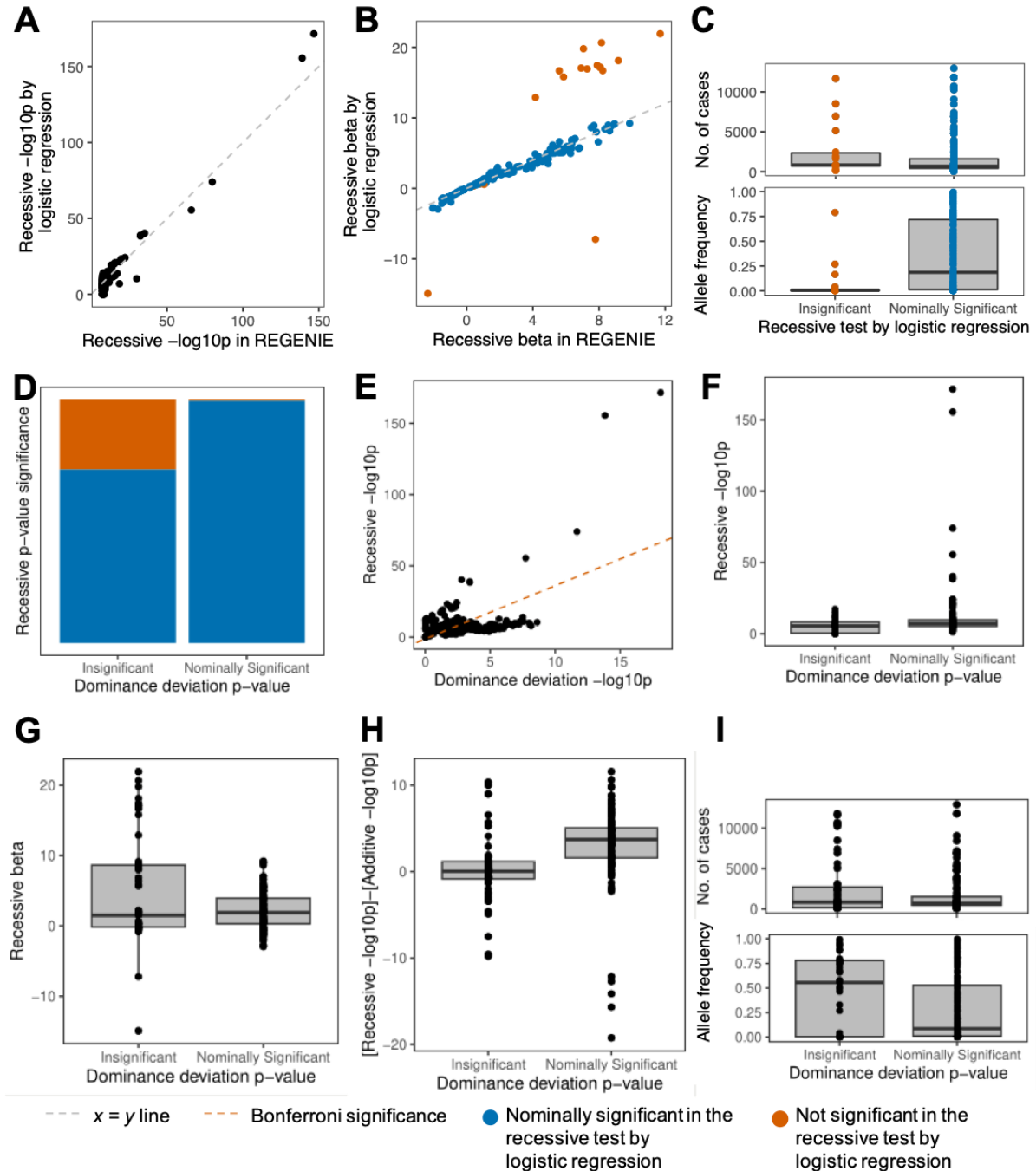


Figure-2.17

Logistic regression testing in R for the recessive findings. A,B: P-values ($-\log_{10}p$, **A**) and betas (**B**) in REGENIE compared to logistic regression testing. **C:** Distributions of AF and case counts between nominally significant and insignificant ($p\text{-value} > 0.05$) tests in the recessive logistic regression. **D-F:** Distribution of recessive logistic regression p-values ($-\log_{10}p$, **D**), betas (**E**), and AF (**F**) across the dominance deviation tests.

Finding replication in FinnGen, an independent cohort

We sought to identify how many of the loci could be replicated in another cohort, FinnGen. 153 loci could be found in FinnGen, meaning the lead variant of the locus could be tested with a comparable phenotype in FinnGen.

Of these 153 lead variants, 102 had the same direction of effect in the recessive tests of both cohorts. This deviates significantly from the number expected by chance (exact binomial test, p-value 4.6×10^{-5}).

Next, we calculated the power-adjusted transferability score, a quantitative measure of replicability, measuring how many loci can be replicated in FinnGen relative to the power in this independent cohort, taking into account AFs, CRs and sample size have changed. Detailed description of the calculation can be found in the Methods section. We expected to be able to replicate 146 of the loci that passed p-value $\leq 5 \times 10^{-8}$ on the recessive test in G&H; of these, 41 loci replicated in FinnGen, resulting in a power-adjusted transferability ratio of 28%.

We hypothesised that a reason for the low replication rate is that the genome-wide significance threshold may be too lenient given the large number of phenotypes tested. Restricting to the FDR 5% findings, the power-adjusted transferability ratio increased to 38%, and for the Bonferroni-significant findings, the power-adjusted transferability ratio was a reassuring 71% (Table-2.7). Notably, the 9 remaining Bonferroni-significant lead variants that did not replicate tested with the same direction of effect in both cohorts. Furthermore, from the three clusters of Bonferroni-significant loci in Table-2.6 and Figure-2.13A, at least one of the loci in each cluster replicated in FinnGen, suggesting that with more lenient locus definitions, all Bonferroni-significant loci would have replicated.

p-value cutoff	N loci	N lead variants found (for power calculation)	Expected N loci replicated	Observed N loci replicated	PAT
GWS	207	158	146	41	28%
FDR 5%	167	122	108	41	38%
Bonferroni	32	31	31	22	71%

Table-2.7

Power-adjusted transferability ratio of the recessive loci, stratified by p-value cutoffs.

Discussion

Summary of findings

Using the exome-sequences of 4,982 individuals, we generated a reference panel for the imputation of 605,263 variants, including rare exonic variants, into the larger cohort of 44,186 individuals in G&H. Simultaneously, we carried out whole-genome imputation to the TOPMED panel, on which we tested 10,045,406 variants for recessive associations with 898 phenotypes. From both sets of imputation, 207 unique loci were identified at genome-wide-significance, with 32 loci in three clusters passing a stringent Bonferroni multiple testing correction. 154 of the 207 loci could be confirmed as recessive with a nominally-significant dominance deviation p-value. After adjusting for changes in case rates, allele frequencies and sample size in the independent cohort FinnGen, we replicated 71% of the expected number of Bonferroni-significant loci we would be powered to replicate.

Strengths and limitations of the different imputation panels used for association testing

Within this project we demonstrated several of the advantages and disadvantages of using a population-specific, within-cohort reference panel in contrast to a larger, publicly-available reference panel composed of diverse ancestries (TOPMED individuals are 41% European, 31% African, 15% Hispanic, 9% Asian, 4% Other ancestries) (Taliun et al., 2021). For common variants, the TOPMEDimputation still proved to be more accurate (refer to relevant plot). This is likely due to the larger sample size of individuals contributing to the reference panel. However, for low-frequency variants, the within-cohort reference panel was more accurate, and also allowed for the imputation of more variants, some which may be unique to the cohort. Despite its large size, the TOPMED panel only includes 644 South Asians (Taliun et al., 2021). Undoubtedly having a larger number of South Asians individuals in the reference panel would improve the accuracy of rare variant imputation. Secondly, we can see that for the variants that have been imputed in both the WES5Kimputation and the TOPMEDimputation, there is a subset of recessive findings with p-values that are only significant with the WES5Kimputation and not the TOPMEDimputation. We hypothesise that this might be due to more accurate imputation of these variants with the within-population reference panel. This suggests that an additional benefit of using a within-cohort reference panel is increased sensitivity for discovering associations from more confidently imputed variants.

With the WES5Kimputation, despite applying an exonic reference panel with the aim of finding recessive associations for rare, exonic variants, the majority of the findings were still common, intronic or noncoding, implicating variants along the SNP-array backbone, or

picked up from the edges of exons during the exome-sequencing. This is expected, since common variants have more power. The TOPMEDimputation was therefore an important analysis to perform, and indeed resulted in more recessive findings found for common and noncoding variants.

From the above two points, the natural conclusion is that a reference panel composed of whole-genome sequencing variants from a subset of G&H, imputed into the larger cohort would bring together the advantages of the WES5Kimputation and the TOPMED imputation, and it is worth considering for future work.

The Bonferroni-significant recessive associations

From the 207 recessive loci we detected, we first focus on the 32 Bonferroni-significant findings. They were found in three clusters, corresponding to (a) Non-alcoholic fatty liver disease and steatohepatitis (NAFLD) (chr22:43939790), (b) disorders of porphyrin and bilirubin metabolism (chr2:233693319-2337901144), and (c) thalassaemia and other hereditary haemolytic anaemias (chr11:4908482-5544800).

We report the recessive association between *rs66812091* (chr22:43939790:TGG:T) and NAFLD, with an odds ratio of 1.3 at a p-value of 2×10^{-12} . This p-value is an order of magnitude more significant in the recessive test than the additive test. The relationship also had a significant dominance deviation p-value of 3×10^{-2} and was replicated in FinnGen with a recessive test OR of 1.3 and p-value of 6×10^{-12} . The variant is an intron variant of Patatin-like Phospholipase Domain-containing Protein 3 (*PNPLA3*), which is a gene involved in lipid and fatty acid metabolism and has high expression in adipose and liver tissues (Baulande et al., 2001; Xiang et al., 2021). Variants in *PNPLA3* have been associated with NAFLD, demonstrated in an additive GWAS (Romeo et al., 2008). The role of *PNPLA3* in fatty acid metabolism and liver function has been further supported by functional knockout studies in mice (Basu Ray, 2019; Xiang et al., 2021). This variant has also been shown in additive GWAS to be associated with NAFLD (FinnGen, 2022) and deranged liver enzymes (a marker of hepatitis) (Pazoki et al., 2021). However, to our knowledge, there are no recessive studies relating this variant to NAFLD. Our results suggest, at least for one or some variants in or near *PNPLA3*, the association with NAFLD is more likely recessive than additive.

All of the Bonferroni-significant variants associated with disorders of porphyrin and bilirubin metabolism cluster around the 233693319-2337901144 base pair positions of chromosome 2, and 13 of them are in uridine 5'-diphospho-glucuronosyltransferase (UDP-glucuronosyltransferase, UGT) genes: UDP-glucuronosyltransferase 1-5 (*UGT1A5*) and UDP-glucuronosyltransferase 1-6 (*UGT1A6*). The genes *UGT1A5* and *UGT1A6* have been

reported in numerous studies to be associated with bilirubin levels (e.g. (Barton et al., 2021; Mulder et al., 2020; Seo et al., 2020; Sinnott-Armstrong et al., 2021)). The variants (*rs6742078*, *rs6722076*, *rs6749496*, *rs4124874*, *rs12987866*, *rs13009407*, *rs1105880*, *rs1105879*) also have multiple additive associations to bilirubin-related phenotypes according to Open Target Genetics (Ghoussaini et al., 2021) and the GWAS Catalog (Sollis et al., 2023). UGTs are enzymes responsible for the glucuronidation of lipophilic molecules, including bilirubin, into the conjugated, hydrophilic form that can be excreted in the urine (King et al., 2000). Mutations in UDP-glucuronosyltransferase 1-5 (*UGT1A1*), another UGT, cause the well-known autosomal recessive conditions Gilbert's syndrome (Scriver, 2001) and Crigler-Najjar syndrome (Crigler & Najjar, 1952; Scriver, 2001), in which patients present with excess unconjugated bilirubin. The findings presented here demonstrate recessive associations between other UGTs and dysfunctional bilirubin clearance.

Thirdly, we found multiple Bonferroni-significant recessive hits at the genomic region chr11:4908482-5544800 in association with thalassaemia and other hereditary haemolytic anaemias. These variants are also associated with multiple anaemia traits in additive GWAS (Ghoussaini et al., 2021; Sollis et al., 2023). The Bonferroni-significant recessive hits in our analysis span two haemoglobin genes, haemoglobin epsilon locus (*HBE1*) and haemoglobin gamma a (*HGB1*), as well as the gene *MMP26* and a cluster of olfactory receptor genes. The two haemoglobin genes seem most likely to be the causal genes (i.e. involved in the pathophysiology), although formal fine mapping and colocalisation would be required to demonstrate this definitively. They encode for embryonic and foetal haemoglobin subunits respectively. In normal human development, embryonic and foetal haemoglobin (HbF) are replaced by adult haemoglobin (HbA) in adulthood (Higgs et al., 1989). Beta-thalassaemia is an inherited anaemia caused by dysfunctional production of the beta globin subunit of HbA, and heterozygous, compound heterozygous and homozygous genotypes of beta globin mutations result in clinical manifestations of varying severity, from thalassemia minor, to thalassaemia intermedia and thalassaemia major (Cao & Galanello, 2010). In some cases of beta-thalassaemia, elevated HbF can persist past infancy, possibly as a compensatory mechanism to anaemia from decreased HbA (Lim et al., 2015). Embryonic haemoglobin is normally only present in the first trimester of pregnancy, though chromosomal abnormalities have been associated with delayed transformation to HbF (Al-Mufti et al., 2000). Genes encoding for embryonic haemoglobin and HbF, including *HBE1* and *HGB1*, have been targeted as part of possible therapies in thalassaemias, where reactivating these alternative forms of haemoglobin expression may help to supplement low HbA production (He & Russell, 2002; Métais et al., 2019; Russell & Liebhaber, 1998). Therefore, the reported associations involving variants in these genes could provide further insight into the regulation

of haemoglobin expression and inactivation. Some of the other variants are near olfactory genes and there are reports of their involvement in anaemia traits (Y. Hu et al., 2021; Menzel et al., 2013; van Rooij et al., 2017; Wojcik et al., 2019). It is worth recognising that the beta globin gene is sandwiched between olfactory receptor gene clusters (Bulger et al., 2000), so these variants may be in LD with variation within the beta globin gene. It is therefore more likely that the haemoglobin genes are involved in thalassaemia pathophysiology than the olfactory genes, but fine-mapping these associations is challenging, as we will discuss below.

From this example we see that G&H provided a unique opportunity to evaluate variants and diseases that are more prevalent in South Asians relative to White Europeans, the ancestry which most GWAS are typically performed on. The prevalence of some hereditary haemoglobinopathies such as beta-thalassaemia is much higher in the Mediterranean, and south, central and southeast Asian regions compared to Northern Europe (De Sanctis et al., 2017), providing a possible reason as to why we were powered to detect these recessive associations in G&H. For the 15 variants that reached Bonferroni-significance in G&H, 12 had lower allele frequencies in FinnGen (Figure-2.18). It is also worth noting that thalassaemia is not encoded specifically in FinnGen, and we had to find replication within the more general phenotype of “Other and unspecified anaemias” (D3_ANAEMIANAS, Supplementary-table-2.2), which would reduce the power for detecting the same association in the Finnish cohort. The case rate for thalassaemia in G&H is 2.3%, while the prevalence of beta-thalassaemia in Finland has been reported at 1 in 100,000 persons (Kattamis et al., 2020).

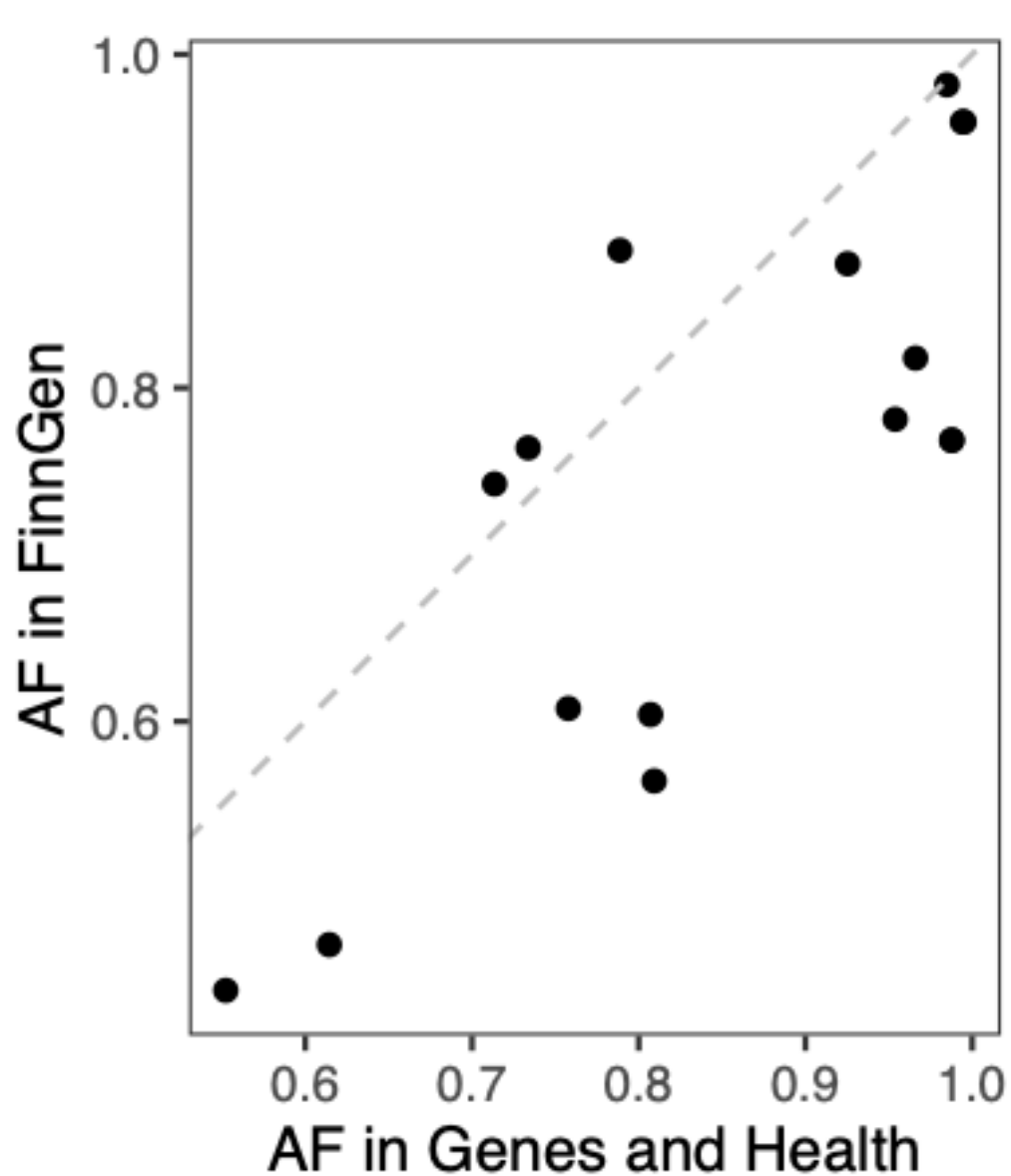


Figure-2.18

Comparison of the AFs in G&H and FinnGen for the Bonferroni-significant lead variants that were associated with thalassaemia and hereditary anaemias.

In modelling binary traits, one usually assumes a liability threshold model, in which an individual develops the disease once they pass a threshold of a continuous, quantitative trait (the 'liability') that follows a normal distribution in the population (S. Wright, 1934). Under this model, it is possible that a variant may have an additive effect on the underlying liability but a recessive effect on disease status. This may be relevant to our three clusters of Bonferroni-significant findings. It may be that the variant in *PNPLA3* has an additive effect on liver enzyme levels, leading to the accumulation of fatty acids that results in hepatic inflammation (i.e. the underlying quantitative trait), but a recessive effect on NAFLD. The variants near the UGT genes may have an additive effect on bilirubin levels but a recessive effect on disorders of porphyrin and bilirubin metabolism. The variants near the haemoglobin genes may have an additive effect on adult haemoglobin levels but a recessive effect on anaemia risk.

The recessive model of testing in a cohort enriched for homozygosity increased the power for detecting recessive associations.

The Bonferroni-significant findings reflect very strong associations, and most had very significant additive p-values as well. However, the majority of the GWS findings (146 / 207) were not GWS under the additive model, illustrating the importance of applying the recessive model of testing for finding recessive effects. We discuss an example in detail to demonstrate this: We showed that an intronic variant (chr13:109179501:G:C, *rs2038707*) in myosin 16 (*MYO16*), a gene involved in the musculoskeletal system (Telek et al., 2020), had a recessive association with the ICD10 code M85 "Other disorders of bone density and structure" (OR = 3.2, p-value = 2×10^{-8} , dominance deviation p-value = 10^{-5}). When tested under the additive model, the effect size is smaller and the result is 5 orders of magnitude less significant (OR = 1.3, p-value = 9×10^{-3}), and likely would have been missed. Furthermore, when the homozygous individuals are excluded from the additive test, the test becomes insignificant (p-value = 0.26), demonstrating the influence of the homozygotes on the signal detected via the additive test.

This example also illustrates the value of applying recessive testing to a cohort enriched for homozygosity. The AF for this variant in G&H is 0.17, which is similar to the AF in non-finnish europeans (NFE) of 0.16 (S. Chen et al., 2022). The NFE population has low levels of consanguinity, therefore at a sample size of 44,000, one could estimate based on the HWE that the number of homozygous individuals would be 1,126 and the power to perform this recessive test in this population would be 57%. At an AF of 0.17, the number of homozygous individuals in 44,000 is expected to be 1,204 by the HWE, but the actual number of homozygotes in G&H is 1,408, due to the increased autozygosity in this cohort. This resulted in an increased power for the test to 82%. Therefore, it is possible that additive tests in larger

outbred cohorts have missed this recessive association. An exome-based study of 394,841 UK Biobank individuals and 4,529 phenotypes detected a gene-based additive hit for the set of putative Loss-of-Function variants in *MYO16* with the same ICD10 code (OR = 1.1, p-value = 5×10^{-4}) (K. J. Karczewski et al., 2022). Furthermore, additive tests with FinnGen (release 6) report associations between our lead variant and related phenotypes such as fibroblastic disorders (OR = 1.1, p-value = 10^{-3}) and benign neoplasms in the scapula and long bones of the upper limb (OR = 0.7, p-value = 10^{-3}) (FinnGen, 2022). Although, to our knowledge, there are no known recessive reports between this variant and phenotypes relevant to disorders of bone density and structure.

Variants with recessive effects may contribute to the association between increased homozygosity and common disease risk

From another study within this cohort, Malawsky et al., 2023 (Malawsky et al., 2023) demonstrated that increased homozygosity (higher F_{ROH}) was associated with multiple common diseases. The primary hypothesis for these associations is that homozygous regions of the genome contained causal variants exercising recessive effects on these phenotypes (Charlesworth & Willis, 2009). From the meta-analysis of highly consanguineous cohorts in the study, F_{ROH} was found to be associated with 12 ICD10 subchapters (passing the FDR 5% multiple testing threshold). We found significant (p-value $\leq 5 \times 10^{-8}$) recessive associations with eight of these phenotypes, four with the exact ICD10 codes, and four with closely-related phenotypes (Table-2.8), although none passed our more stringent Bonferroni correction threshold.

Gene	Lead variant	AF	CSQ	Phenotype	OR	p-value	DD p-value	FG OR	FG p-value
FROH-Associated Phenotypes:									
A01[Infectious gastroenteritis and colitis, unspecified]									
.	2:171680912:G:A	0.3279	intergenic	A09[Other gastroenteritis and colitis of infectious and unspecified origin]	0.6	2.7E-08	6.0E-02		
E11[Type 2 diabetes mellitus]									
.	6:104459545:G:A	0.0066	regulatory region	E11[Type 2 diabetes mellitus]	38.0	3.6E-08	7.2E-06	not tested	
Y_RNA	10:92706738:G:C	0.7379	downstream gene	GNH0242 Type 2 Diabetes narrow	0.8	8.8E-09	1.0E-01	0.9	2.0E-18
Y_RNA	10:92706682:A:G	0.7559	downstream gene	Type 2 Diabetes	0.8	4.3E-09	5.6E-02	0.9	6.7E-18
ADAMTS16	5:5261658:G:A	0.0052	intron	E14[Unspecified diabetes mellitus]	296.0	5.1E-09	6.9E-06	not tested	
UBE2E2	3:23530638:C:A	0.0008	intron	E14[Unspecified diabetes mellitus]	2664.1	4.6E-08	9.6E-01	not tested	
SNX5	20:17949555:G:T	0.0103	intron	GNH0244 Unspecified or Rare Diabetes narrow	62.7	1.4E-08	3.5E-03		
E78[Disorders of lipoprotein metabolism and other lipidaemias]									
APOA5	11:116796367:A:G	0.9373	upstream gene	E78[Disorders of lipoprotein metabolism and other lipidaemias]	0.8	3.3E-11	7.1E-03	0.9	1.5E-16
F41[Other anxiety disorders]									
.	2:236834949:C:T	0.0071	regulatory region	F40[Phobic anxiety disorders]	137.6	2.7E-08	2.6E-03		
.	7:131782270:G:A	0.0775	intergenic	Anxiety and phobia	1.8	3.5E-08	4.8E-03		
.	7:131780486:G:A	0.0800	intergenic	F41[Other anxiety disorders]	1.9	7.7E-10	2.7E-03		
.	7:131773247:A:G	0.1270	regulatory region	F41[Other anxiety disorders]	1.8	1.8E-08	7.7E-05		
H61[Other disorders of external ear]									
ESRRG	1:216911690:AG:A	0.0053	intron	H60[Otitis externa]	54.3	2.5E-09	4.7E-04	not tested	
J11[Influenza, virus not identified]									
PKHD1	6:52012065:G:GT	0.0039	intron	J12[Viral pneumonia, not elsewhere classified]	271.1	3.1E-09	6.6E-06	not tested	
PKHD1	6:52012065:G:GT	0.0039	intron	B97[Viral agents as the cause of diseases classified to other chapters]	179.9	2.7E-09	1.9E-05		
.	6:51529735:A:G	0.0020	intergenic	B97[Viral agents as the cause of diseases classified to other chapters]	429.9	2.4E-08	8.0E-04	not tested	
J34[Other disorders of nose and nasal sinuses]									
ADAMTS9-AS2	3:64802433:A:G	0.0273	intron	J34[Other disorders of nose and nasal sinuses]	7.9	1.5E-08	1.8E-07	not tested	
.	7:48916122:T:C	0.0159	intergenic	J34[Other disorders of nose and nasal sinuses]	12.9	5.0E-09	1.5E-03	not tested	
.	5:91760313:C:G	0.0109	intron	J34[Other disorders of nose and nasal sinuses]	10.9	2.4E-08	1.5E-03	not tested	
L30[Other dermatitis]									
PTPRN2	7:157970755:T:G	0.0053	intron	L20[Atopic dermatitis]	63.8	1.1E-08	8.6E-01	not tested	
NKAIN1	1:31215264:T:C	0.8847	intron	L21[Seborrhoeic dermatitis]	1.3	6.2E-09	8.4E-01		
BCAR3	1:93847011:TCGG GCGCGGCGG:*	0.3519	intron	Dermatitis (atopic, contact, other, unspecified)	1.2	1.6E-08	6.3E-08	not tested	

Table-2.8

Single recessive associations involving phenotypes found to be significantly associated with genome-wide homozygosity in Malawsky et al., 2023). Only FinnGen (FG) replication with nominally-significant p-values are listed here; associations that have been tested but not replicated are left blank, and there are also several loci that have not been tested in FG because the lead variant, proxy variants or a matching phenotype could not be found in FG (see methods).

The prevalence of T2DM is high in South Asians, with the incidence being two to five times higher than in White Europeans and North Americans (Gujral et al., 2013; Narayan & Kanaya, 2020). The results of Malawsky et al., 2023 suggest that this increased incidence is partly due to the high rates of consanguinity in certain South Asian groups, which lead to an increased burden of homozygosity at recessive variants. Two of the lead variants we found with recessive associations to T2D were replicated in FinnGen as well (Table-2.8). Interestingly, their ORs suggest they may be protective for T2DM, whereas, at the other four recessive loci, homozygosity for the alternate allele is associated with increased risk of disease. One of these lead variants (chr5:5261658:G:A, *rs79878016*, OR = 296, p-value = 5×10^{-9} , dominance deviation p-value = 7×10^{-6}) is an intron in *ADAMTS16*, a member of the *ADAMTS* (a disintegrin and metalloproteinase with thrombospondin motifs) superfamily of metalloproteinases (Dubail & Apte, 2015). The role of *ADAMTS* genes in T2DM pathogenesis is unclear, but multiple studies report associations between these genes and T2DM (Boesgaard et al., 2009; Zeggini et al., 2008) and related cardiometabolic conditions (Heid et al., 2010; Reilly et al., 2011). In particular, a recessive association between an intron (*rs140453320*) in *ADAMTS6* and T2DM (OR = 6.8, p-value = 5×10^{-9}) has been reported (O'Connor et al., 2022). Although these loci are in noncoding regulatory regions, which makes functional interpretation challenging, these findings support the hypothesis that the association between increased autozygosity and the prevalence of some common diseases are due to underlying genetic variants with recessive effects.

Recessive findings implicating coding variants provided insight into gene functions

We also find significant recessive associations for coding variants, which seem more likely to be the causal variants at those loci, and are easier to interpret. For example, we see a recessive association between a missense variant (chr1:11796321:G:A, *rs1801133*) in methylenetetrahydrofolate reductase (*MTHFR*) and folate deficiency (OR = 2.1, p-value = 5×10^{-9} , dominance deviation p-value = 10^{-3}). *MTHFR* is an enzyme involved in homocysteine metabolism (Goyette et al., 1998). Previous functional studies have shown that the missense variant we report causes instability in the enzyme, resulting in the accumulation of homocysteine, though the enzyme can be stabilised with additional folate, resulting in low folate levels (Frosst et al., 1995; Guttormsen et al., 1996; Jacques et al., 1996; S. S. Kang et al., 1988). There are reports that this mutation has a recessive effect on homocysteine levels (Frosst et al., 1995; Hustad, S., 2007; Liew & Gupta, 2015), and our recessive association is with folate deficiency. Our finding is also replicated in FinnGen (OR = 2.1 p-value = 0.02), and multiple GWAS, although additive, have reported associations

between this variant and folate deficiency anaemia (FinnGen, 2022) or being on folate supplements (Neale, 2018).

As a second example, we have found a protective recessive association between a missense variant in *SLC5A9* (chr1:48228922:G:A, *rs61746559*) hypertension (OR = 0.2, p-value = 3×10^{-8} , dominance deviation p-value = 7×10^{-6}). *SLC5A9*, solute carrier family 5 member 9, is also known as sodium-glucose transporter 4 (*SGLT4*), and is a member of the solute carrier (SLC) superfamily. Specifically, it is a sodium-dependent glucose transporter of mannose, 1,5-anhydro-D-glucitol, and fructose (Tazawa et al., 2005). Another member of this family is the well-studied sodium-glucose transporter 2 (*SGLT2*), a glucose transporter largely expressed in the kidney, and is already a drug target for lowering serum glucose levels in patients with T2DM; Glifozins, or *SGLT2* inhibitors, are anti-diabetic drugs that target the *SGLT2* glucose transporters along the proximal tubules of the kidney to reduce the reabsorption of glucose from filtered urine (Shubrook et al., 2015). It has been shown that a concomitant benefit of inhibiting *SGLT2* is lowering blood pressure (Mazidi et al., 2017; Thomas & Cherney, 2018; Zaccardi et al., 2016), possibly from haemodynamic changes in kidney glomeruli (Dharia et al., 2023) and regulating the renin–angiotensin–aldosterone system (Schork et al., 2019). *SGLT4* may function in a similar manner, as, like *SGLT2*, it is expressed in the kidneys (Tazawa et al., 2005). Missense mutations may inhibit its function, explaining the relationship we detect with hypertension. While there is limited information about this gene's function currently (Gyimesi et al., 2020) to support this hypothesis, there is an additive association between this variant and renin levels (Folkersen et al., 2020), suggesting it may indeed be involved in the renin–angiotensin–aldosterone system.

Limitations of this study

There are several limitations to the project. Firstly, errors during imputation undoubtedly limit the power of the association tests. In particular, accuracy worsens with decreasing allele frequency, which could contribute to the paucity of associations with rare exonic variants, on top of the reduced power due to lower allele frequency. This could be ameliorated in future by increasing the number of individuals in the reference panel, or applying even more stringent filtering during the QC steps. Exome-sequencing of the entire G&H cohort is now underway, although these data were not available when I carried out the work in this chapter. Validation of the imputed findings from this study within exonic regions could be performed in those data, as well as association analyses for more variants discovered during exome-sequencing.

Secondly, despite performing imputation into a cohort of 44,000 in an attempt to increase sample size, it is likely that we were still underpowered to identify many recessive associations, particularly with rare exonic variants. With the ongoing exome-sequencing project, power for the discovery of new recessive variants should improve. Members in the Martin group are working on implementing recessive gene-based burden tests, aggregating multiple protein-altering variants in the same gene, and this may further boost power.

Thirdly, we have not carried out fine-mapping of these recessive associations, so we have not identified the causal variants within the loci discovered. To our knowledge, there are no established methods for fine-mapping which would help disentangle, for example, a recessive hit that is in LD with a strongly additive hit. It is worth noting also that LD diminishes with distance between non-additive variants at a rate that is squared of the rate between additive variants (Palmer et al., 2023), which should, in theory, make it easier to pinpoint causal recessive variants due to the lower LD.

Next, there is limited replication of our findings in FinnGen for findings slightly above genome-wide significance. Beyond the possibility that the GWS cutoff is not stringent enough given the number of tests we have performed, there could be many other reasons for this. Firstly, power calculations were performed on a simple logistic regression model, which may not be reflective of what SAIGE and REGENIE applied, as they both adopt a more complex model accounting for multiple other factors such as large case-control imbalances, population structure and relatedness within the cohort, and include covariates. Thus, our true power may be lower than what we have estimated. Secondly, the Finnish are an ancestry group very different from British South Asians, and despite accounting for AF and CR differences in our power calculations, and considering a set of proxy variants that are in LD with the lead SNP within G&H, differences in effect sizes and linkage disequilibrium patterns. In the future, it would be interesting to do a comprehensive comparison of the extent to which additive versus recessive associations transport between different ancestry groups. We could also carry out this same replication analysis in the BangladEsh Longitudinal Investigation of Emerging Vascular Events (BELIEVE), to assess whether we see higher rates of replication than in FinnGen.

Lastly, we focused only on single-variant association tests, specifically homozygous variants (two copies of the same variant). Aggregating different variants in the same gene may boost power to find genes with a recessive effects, just as gene burden and variance component tests are used to test the additive model (add refs). These gene-based tests would ideally include compound heterozygous genotypes, in which an individual has two different variants affecting the same gene. This would require accurate phase information, which is

challenging since most coding variants are rare, and phasing rare variants without having the parental data is difficult, although it has recently been shown to work reasonably well in UK Biobank using statistical phasing approaches, and to help with identifying recessive associations at the gene level (Lassen et al., 2023). Expanding the study of recessive effects to compound heterozygotes is the natural extension to this study, when more of the cohort is exome-sequenced and phasing can then be more accurately performed. A colleague in the lab is working on this.

Conclusion

There are two main conclusions to this chapter. First, we demonstrated that a within-cohort reference panel allowed for the imputation of more population-specific rare variants, and at higher accuracy. The higher accuracy, in some cases, resulted in higher sensitivity for detecting recessive associations. Secondly, with our whole-exome and whole-genome imputation sets, we profiled the recessive landscape at single variants in this cohort of 44,000 British South Asians across a broad spectrum of clinical phenotypes, identifying 207 recessive associations. It is likely that many recessive findings remain to be found, and this project provides a sound argument to expand the search, especially with the new exome-sequencing of the entire cohort.

Acknowledgements for this chapter

Hilary C. Martin, Qinqin Huang, David A. van Heel and the Genes & Health Team performed some of the initial QC of the genetic data (described in detail in the Methods). Qinqin Huang, David A. van Heel and the Genes & Health Team performed the TOPMED-r2 imputation. David A. van Heel and Benjamin M. Jacobs curated the phenotypes. The human genetics informatics team at the Wellcome Sanger Institute performed the VEP variant annotation. Juha Karjalainen and Mark J. Daly performed the recessive association testing in FinnGen. Qinqin Huang, Klaudia Walter, Georgios Kalantzis, Daniel Malawsky and Henrike Heyne provided guidance on the analyses. Hilary C. Martin led and supervised the project.

Identifying human gene knockouts and detecting their effects on the blood transcriptome

Overview

Loss-of-function variants (LoFs) are naturally-occurring gene 'knockouts' that allow *in vivo* human observations on the function and expression of genes. Cell-type-specific LoF effects on transcription have yet to be studied at scale, and the rarity of true LoFs makes it difficult to study them. Genes & Health (G&H) is enriched for consanguinity, and therefore enriched for LoFs in a homozygous state, providing more power. We identified individuals with putative LoFs (pLoFs) in immune genes, recalling them for single-cell RNA sequencing (scRNAseq) of peripheral blood mononuclear cells (PBMCs), to study the transcriptomic patterns associated with these variants. We took several steps to prioritise participants for recall. First, in the exome sequences of G&H participants who self-declared as having related parents, we identified homozygous pLoFs and applied additional filters to reduce the risk of prioritising spurious LoFs. Next, we filtered to LoFs located in genes expressed in blood and involved in the immune system by studying various publicly available datasets and considering immune trait association studies. Lastly, we performed power calculations to estimate which knockout effects could be detected, based on gene expression levels and cell-type proportions. This resulted in 22 genes affected by 32 pLoFs, with 488 individuals who are heterozygous or homozygous for these variants. From these, 88 individuals with pLoFs, 6 first-degree relatives of homozygous individuals, and 20 controls were successfully recruited. Additional curation of the shortlisted pLoFs was performed by collaborators, accounting for technical and biological factors to assess the likelihood of these pLoFs resulting in true loss-of-function. The expression analyses from the PBMC scRNAseq data confirmed the presence of direct knockdown effects on affected genes in most cases, and largely concurred with the verdicts from the additional curation. We showed how a splice donor variant (rs571047357) in the *CAT* (catalase) gene resulted in decreased expression of the gene transcript specifically in monocytes. We detected a possible compensatory mechanism in individuals with this pLoF, where the production of nitric oxide, a catalase inhibitor, is downregulated. Lastly, we discuss instances of non-additivity of LoF effects on RNA expression. These findings justify deeper scRNAseq, described in the next thesis chapter, to interrogate LoF effects on the immune gene network at rest and in stimulation.

Introduction

Loss-of-function variants as naturally-occurring gene knockouts

Loss-of-function variants (LoFs) are nonsense, essential splice site, and frameshift variants that are predicted to result in the truncation or inactivation of the encoded protein (MacArthur et al., 2012). In individuals heterozygous for the LoF, the wild-type allele may allow for some expression of the gene. Homozygous individuals however, present with inactivation of both copies of the gene, with the LoF operating as a naturally-occurring gene “knockout” (Lek et al., 2016).

Knockout experiments are important studies in molecular biology to interrogate gene function, as they provide a direct link between gene inactivation and an observable phenotype (Michael Conn, 2011). This is usually accomplished *in vitro* by introducing artificial means of suppressing gene expression, such as small interfering RNA (Elbashir et al., 2001) or CRISPR-cas9 gene editing (Adli, 2018), in animal or cellular models. As an adjunct to such studies, LoFs naturally result in decreased gene expression *in vivo*, allowing for observations of the function and expression of affected genes directly in humans (Minikel et al., 2020).

The severity of LoF effects is variable

Variants that completely disrupt a protein’s function have been thought to cause severe, Mendelian disorders, or be highly penetrant. (Forrest et al., 2022; MacArthur et al., 2012; MacArthur & Tyler-Smith, 2010; Maddirevula et al., 2020). Notable examples include cystic fibrosis (Ong & Ramsey, 2023) and Duchenne muscular dystrophy (Duan et al., 2021). Inference from experimental knockout studies in mouse models suggested that 30% of gene knockouts are not compatible with life, while most of the remaining gene knockouts that allowed survival at birth displayed observable phenotypes (Ayadi et al., 2012). Chong et al., 2015 estimated based on the conservation of the genome between species that a similar percentage of genes would result in human embryonic lethality should they contain LoFs, and that a further three thousand genes have had variants, including LoFs, identified to be associated with Mendelian diseases (Chong et al., 2015). In addition, the penetrance of many deleterious variants have been estimated from family studies or case-control studies, leading to inflation of the severity interpretation (Forrest et al., 2022).

Yet, benign LoFs also exist, such as homozygous deletions in the *ABO* gene resulting in the O blood type (Calafell et al., 2008). With the increasing availability of large-scale sequenced datasets, it has been observed that healthy humans all harbour a surprisingly large collection of LoFs, with estimates ranging from 200 to 800 LoFs per individual, heterozygous and

homozygous (1000 Genomes Project Consortium et al., 2010; Ng et al., 2008; Pelak et al., 2010). The extent to which these LoFs contribute to common diseases and complex traits, or the qualifiers that would allow healthy individuals to carry LoFs, are unclear. In an effort to systematically characterise the spectrum of LoF effects, MacArthur et al., 2012 proposed that they broadly fell into the following overlapping classifications (MacArthur et al., 2012):

- LoFs that cause severe disease in the recessive state but are carried in the population by healthy heterozygous individuals
- LoFs with small effect sizes on phenotypes or influence on disease risk
- LoFs in redundant genes that are therefore benign
- LoFs that do not diminish gene function
- LoFs that have been called erroneously

The first three points relate to gene essentiality and function, the fourth relates to gene rescue, and the last point relates to errors in LoF calling. We discuss all these in the next sections of this introduction, as they inform the aims of this part of the study.

Key studies surveying the population prevalence of LoFs and estimating constraint

A decrease or loss of an essential gene is likely to result in severe disease or is incompatible with life, and therefore variants that knock out these genes would be subjected to natural selection. Several large genetic datasets have allowed surveys of the extent of LoF constraint across the genome:

The Exome Aggregation Consortium (ExAC) is a joint variant call set of 60,706 exomes from which Lek et al., 2016 calculated the gene constraint against LoFs. The density of LoFs across the exome was variable, and the LoF frequency in a gene is assumed to be a function of the degree of negative selection against LoFs in the gene and the mutation rate. If the gene was more essential to survival (prior to reproductive age) or reproduction, the stronger the selection pressure against LoFs in that gene and the sparser the LoFs. However, this has to be balanced with the mutation rate (which in turn is a function of gene length), with higher mutation rates resulting in more LoFs forming. From this concept, Lek et al., 2016 calculated an expected LoF frequency in a gene based on the mutation rate, and compared the observed frequency in the cohort to report the metric 'pLI' - the probability of a gene being LoF-intolerant. While the pLI ranges from 0 to 1, it is often used dichotomously, i.e. it serves to differentiate LoF-intolerant genes from others (Seaby et al., 2021). Through this they identified 3,270 LoF-intolerant genes (defined as $pLI \geq 0.9$) and observed that these genes were involved in core biological processes. Conversely, olfactory receptor genes were

some of the least constraint genes. It is worth noting that a related metric, pRec, or the probability of being intolerant to biallelic LoFs had also been measured by Lek et al, though it is less predictive of recessive Mendelian genes than pLI for dominant Mendelian genes. We tested pRec in our project as well. (Lek et al., 2016)

The pLI was updated in the Genome Aggregation Database (gnomAD) project comprising 125,748 exomes and 15,708 genomes (in version 2), doubling the sample size from ExAC. In order to annotate a large number of LoFs with increased accuracy, they further developed the Loss-Of-Function Transcript Effect Estimator (LOFTEE, first introduced in Lek et al., 2016), which compiles nonsense, splice site and frameshift variants into putative LoFs (pLoFs), assigns a confidence level and accounts for annotation errors. LOFTEE is a variant annotation plugin that applies a set of filters to assign a confidence level to stop-gain, splice site disrupting, and frameshift variants on their likelihood of causing true loss-of-function. Filters include whether the variant was at the end of the transcript, surrounded by non-canonical splice sites, or whether there were rescue splice sites nearby, among others. A detailed description of the filters can be found with their plugin installation (K. Karczewski, 2020), though we discuss relevant filters in this chapter, and many overlap with the Centre for Population Genomics Loss of Function Variant Curation Standard Operating Procedure Version 2, which we also discuss later.

LOFTEE, combined with the larger sample size and improved model for estimating variation, allowed the authors to calculate the loss-of-function observed/expected upper bound fraction (LOEUF) metric, which allowed for a higher-resolution, more accurate measure of gene tolerance to LoF across the spectrum of the continuous metric as compared to pLI. (K. J. Karczewski et al., 2020) In our project we focused on pLI to annotate genes, as it was the more commonly used metric at that time. For completeness, we describe another relevant metric, S_{het} , defined as the selection pressure on heterozygous LoFs, with the first estimates also derived from ExAC. Unlike pLI and LOEUF, which relate more to gene essentiality, or the “survivability” of LoFs in a gene, S_{het} has a direct population genetic interpretation corresponding to the decrease in fitness (reproductive success) due to heterozygous LoFs in the gene. (Cassa et al., 2017)

LoF effects provide insight into gene function and identify possible drug targets

Since LoFs exhibit phenotypic effects across the spectrum of severity, this can be exploited to understand gene function and identify possible drug targets (Minikel et al., 2020).

Using naturally-occurring gene knockouts to identify potential drug targets can supplement other costly and limited studies such as randomised controlled trials (Barrett et al., 2015). For example, nonsense mutations in proprotein convertase subtilisin/kexin type 9 (*PCSK9*) are associated with low serum low-density lipoprotein (Kotowski et al., 2006), and consequently, the gene has become a successful target for monoclonal antibody drug development (Horton et al., 2007; Kaddoura et al., 2020).

Drug target identification is a complex process which requires building a comprehensive knowledge, beyond understanding the direct effect of a gene on one disease phenotype. Studying naturally-occurring knockouts already present in humans can reveal how gene inhibition is tolerated physiologically across systems. In another example, genetic studies show that gain-of-function variants in leucine-rich repeat kinase 2 (*LRRK2*) is associated with increased risk for Parkinson's disease, suggesting that it could be a possible target for drug inhibition (West et al., 2005). However, *LRRK2* knockout studies in animal models raised toxicity concerns with multisystemic consequences arising from homozygous knockout of the gene (Baptista et al., 2013). Yet, by observing for the presence of naturally-occurring *LRRK2* LoFs in human population cohorts, it was shown that while individuals with heterozygous LoFs demonstrated decreased *LRRK2* protein levels, they did not have significant associations with any disease state (Whiffin et al., 2020).

Finally, in a study to systematically assess the value of LoF effects in drug discovery, Minikel et al., 2020 demonstrated that approved drug targets were slightly more constrained (against LoFs) than other genes (Minikel et al., 2020). Taken together these studies provide an argument for complementing knockout experiments in animal models with observational studies of LoF effects in humans for drug discovery.

Biological mechanisms that relate to LoF effects

However, studying naturally-occurring knockouts is complicated by the fact that complete gene inactivation is rare. Firstly, in heterozygous individuals, the wild-type allele can allow for some expression of the gene, and variation in the degree that this compensates for the LoF allele could explain incomplete penetrance in heterozygous carriers of recessive variants (Barton et al., 2022).

In addition, alternative splicing results in genes being able to produce multiple transcripts, and LoFs that affect a fraction of a gene's transcripts, but are skipped in other transcripts due to different splicing, may allow the other transcript isoforms to rescue the gene expression and function (MacArthur & Tyler-Smith, 2010; Yngvadottir et al., 2009).

Also, even when a variant results in a premature stop codon, not all will result in complete non-expression of the transcript - some may indeed trigger the nonsense-mediated decay (NMD) of the whole transcript (Y.-F. Chang et al., 2007), but others may still allow for the truncated product to persist, and this truncated transcript may still exhibit some function (MacArthur & Tyler-Smith, 2010). This is particularly the case if the variation occurs at the end of the open reading frame (Kaiser et al., 2015; Lindeboom et al., 2016).

In addition, there are several biological mechanisms that can diminish LoF effects through gene regulation. Firstly, expression quantitative trait loci (eQTL) are variants that are associated with differential gene expression among individuals, by modifying regulatory and splicing elements (Kerimov et al., 2021). It has been proposed that heterozygous carriers of LoFs may have different expressions of their wild-type allele due to *cis*-eQTL (Castel et al., 2018).

Next, there are also feedback loops that regulate gene expression to maintain homeostasis (Krishna et al., 2006), and these have been known to rescue expression via upregulation of the wild-type allele (Lindeboom et al., 2016). More generally, genes with overlapping functions, such as duplicated genes known as paralogs, may also have changes in their expression to compensate for knockout in the gene (Bailon-Zambrano et al., 2022; Zande et al., 2023).

Together, these biological mechanisms complicate the quantification and analysis of LoF effects. Depending on the mode of analysis, some of these mechanisms may affect the analysis more than others. In this project, we are aiming to quantify LoF effects on gene expression at the RNA transcription level. Therefore, transcripts that have not undergone NMD (and may result in a truncated protein expression), may reflect as expression within the normal ranges.

LoF variant calling is enriched for errors

Another major challenge with studying LoFs is that many bioinformatically-predicted pLoFs are artefacts and do not result in true protein inactivation (Singer-Berk et al., 2023). These artefacts include errors in sequencing, mapping, variant calling and annotation (MacArthur et al., 2012; Minikel et al., 2020). While some annotation errors can be flagged by high-throughput automated filtering such as that implemented in the LOFTEE package (K. J. Karczewski et al., 2020), other errors need manual curation to identify and remove. In the gnomAD project, approximately forty percent of pLoFs identified with high-confidence by

LOFTEE would likely be eliminated upon further manual curation (K. J. Karczewski et al., 2020).

True homozygous LoFs are rare, but could be enriched in Genes & Health

True LoFs that cause a decrease in gene expression might be rare, especially in healthy population cohorts, particularly if they would result in severe phenotypes. Furthermore, it is expected that they would occur even more infrequently in a homozygous state, the state at which a “two-hit”, or a complete “knockout”, could be expected, particularly in outbred populations (Kaiser et al., 2015; MacArthur et al., 2012). In outbred populations, the frequency of the homozygous variant is the square of the frequency of the heterozygous variant. In addition, should the variant act recessively, there would be further selection pressure on homozygous variants. This rarity has served as a major challenge in the study of naturally-occurring gene knockouts.

As discussed in the previous chapter, individuals with consanguineous parents have been shown to have more homozygous LoFs (per individual) due to the increased autozygosity in their genome (Alsalem et al., 2013). Bottlenecked populations also demonstrate increased frequencies of rare deleterious variants, further increasing the frequency of LoFs (Lohmueller et al., 2008). This makes the Genes & Health (G&H) cohort particularly suited to the study of homozygous LoF effects, both for its increased autozygosity, as well as the presence of multiple bottlenecked subpopulations amongst the individuals of Pakistani ancestry (Arciero et al., 2021).

G&H is enriched for consanguinity, with 19% of participants self-reporting parental relatedness. In an earlier study based on the exomes of 2,912 individuals who have parents who were reported to be second cousins or closer, it was estimated that about 5.8% of the genomes in these individuals would be autozygous. By calculating the expected frequency of homozygous LoFs based on autozygosity and without natural selection, it was predicted that the number of homozygous LoFs in G&H would be much higher than in an outbred cohort, at five homozygous LoFs per million individuals for the median gene (Minikel et al., 2020). While industry-funded whole exome sequencing (WES) of the whole cohort is ongoing, the individuals with self-reported parental relatedness were previously prioritised for WES using Wellcome Sanger Institute funds (n~5,000 in 2019), allowing for the identification of individuals with homozygous LoFs. By design of the cohort, these individuals can be recalled, and with them, further studies of LoF effects can be planned.

A recall-by-genotype was therefore planned in collaboration with the Cardinal project, a study involving the UK Biobank and G&H which recruited approximately 7,000 individuals

from both cohorts (including 1,500 from G&H) for scRNAseq of peripheral blood mononuclear cells (PBMCs). The aims of the study include deeper profiling of the blood transcriptome, expression quantitative trait loci (eQTL) mapping, and increasing the representation of diverse populations in PBMC scRNAseq datasets. Within the G&H recruitment, there was an opportunity to include individuals with pLoFs predicted to affect immune genes.

LoF effects can be tissue- and cell-type-specific

Presently, there is a lack of deep profiling of the gene expression consequences of LoFs, (Minikel et al., 2020). The majority of the study on LoF effects have been focused on phenotypic presentations, protein quantification, and bulk RNA sequencing analyses. It is well established that variant effects on gene expression can be tissue or even cell-type-specific (GTEx Consortium et al., 2017). The magnitude of gene expression, alternative splicing, eQTLs and regulatory elements vary across tissues and cell-types, and can therefore also modulate a variant's influence on gene expression (Kim-Hellmuth et al., 2020). LoFs are no different (Rivas et al., 2015), and while the direct effect of a LoF on the exon it resides in can be predicted based on its annotation, tissue or cell-type-specific eQTLs and splicing differences may also influence the overall expression of LoF-containing genes.

The Genotype-Tissue Expression (GTEx) project catalogued whole genome sequences together with the gene expression in 44 human tissue types, with the goal of characterising tissue-specific variant effects (GTEx Consortium et al., 2017). From analysing GTEx data, Cummings et al., 2020 demonstrated that pLoFs had variable effects on expression across tissue types, due to alternative splicing determining the specific exons expressed in different tissues, as alluded to earlier in this introduction. The authors took into account the different transcript isoforms across tissues, and the exon conservation across these isoforms to develop a transcript expression-aware annotation for pLoFs, called the pEXT (proportion of express transcripts) score. Where there are multiple transcript isoforms of a gene, the pEXT score (PEXT) evaluates the magnitude of transcript expression across tissues, with the hypothesis that lowly expressed transcripts across tissues are unlikely to cause loss-of-function when knocked out. This enabled the authors to add another layer to LoF effect interpretation, and with the pEXT metric they demonstrated how in patients with neurodevelopmental disorders, pLoFs affecting poorly conserved exons had similar effects to synonymous variants, while pLoFs in conserved exons were significantly enriched in cases. (Cummings et al., 2020)

Extending from these conclusions, we hypothesise that the specificity of LoF effects could apply at the cellular level, and by studying LoF effects at single-cell resolution, it would allow even more specific profiling of LoF effects across different cell-types.

Designing a recall-by-genotype scRNAseq study to profile LoF effects in immune cells

The immune system plays a critical role in a range of different biological processes, from defending against infectious diseases, to allergic and hypersensitivity responses, to autoimmunity, and cancer. These processes cause disease across multiple organs and systems, and reflect the importance of developing a better understanding of the immune system. (Marshall et al., 2018)

Studying the effects of LoFs can help us better understand gene modulation and human traits. Especially in the context of immune traits, where the main relevant tissue (i.e. blood) is relatively more easily accessible compared to other tissue types, and would be suitable for a recall study built within a healthy community-based cohort.

Therefore, we set out to conduct a recall-by-genotype study focused on LoFs affecting genes likely to be involved and expressed in the immune system, and study the effects of these LoFs on the immune transcriptome.

Given the challenges related to LoFs outlined in the introduction, we describe our approach to designing and conducting the recall-by-genotype, taking into account the following considerations during recall prioritisation: Firstly, bioinformatically-called pLoFs would require further manual curation, to avoid wasting resources on erroneously called variants. Secondly, we needed to perform some power calculations based on existing PBMC scRNAseq datasets to ensure that we had power to study the selected LoF effects. Lastly, in order to study the effect of these LoFs on wider gene networks, such as genes upstream and downstream to the gene affected by the naturally-occurring knockout, we would first have to characterise the direct LoF effects on the gene transcript. These considerations are the focus of this chapter.

The aims of this chapter

The first aim was to prioritise participants for recall based on the presence of pLoFs, with an emphasis on homozygous pLoFs (hom-pLoFs), that are likely to result in true protein inactivation and observable transcriptomic effects in immune cells (PBMCs). To this end, the hom-pLoFs and the genes they affect were characterised, and as blood was the tissue of choice in this study, these characteristics were then contextualised to the immune system.

Secondly, in this chapter, we also describe the recall-by-genotype process. We cover the study design, the recall, and the demographics of the recalled individuals.

Thirdly, we present an initial analysis of LoF effects based on the scRNAseq data from the Cardinal project (approximately 2,000 cells per individuals), with the aim of confirming the knockdown effects of our prioritised pLoFs in the immune transcriptome. We leveraged Cardinal's larger dataset to quantify the LoF effects relative to the population distribution. Finally, we evaluate whether there is any evidence for non-additive effects of LoF variants on gene expression.

The conclusions from this chapter justify and inform the detailed scRNAseq and *in vitro* stimulation study described in the next chapter.

Acknowledgements for this chapter

This project was a collaborative effort. I have put the acknowledgements here so that the methods section can be read with the relevant context.

Hilary C. Martin and Gosia Trynka supervised the overall project. Quality control of the exome sequences and LoF calling used for the recall prioritisation was performed prior to the start of my PhD by Hilary C. Martin.

The Cardinal project is overseen by Nicole Soranzo, with sample collection efforts managed by Nicole Staudt, wet lab sample processing led by Stephen Watt and conducted by his team, and the Human Genetics Informatics Team at the Sanger Institute (Matiss Ozols as the primary contributor) conducted the computational quality control of the scRNAseq data.

The Genes & Health cohort is overseen by David A. van Heel, with Robin Lerner and Jessry Russell coordinating the recall-by-genotype. Samiha Rahman and Sidra Malik (primary contributors, along with other members of the Genes & Health Team), conducted the recall, consent taking, and sample collection. Genetic and phenotypic data curation of the cohort is described in further detail in Chapter 2.

Daniel Macarthur led the efforts to perform further manual curation of the LoFs recalled, and the analysis was primarily performed by Samantha Bryen.

In addition, Joanna Kaplanis supervised the prioritisation of individuals for recall, Carla Jones Bell supervised the wet lab study design, and Anna Lorenc supervised the computational analysis. I also analysed some scRNAseq data of PBMCs processed by Niek de Klein and Elena Lukyanova to inform the prioritisation of recall individuals, but they are not elaborated on in this thesis.

Methods

This recall-by-genotype study is the subject of Chapters 3 and 4. An overview to the key steps of data generation for the whole project is in Figure-4.2 of Chapter 4.

Prioritisation of individuals for recall

We first had to prioritise individuals for the recall. This section discusses the steps taken, with the results section justifying the cutoffs applied. Figure-3.1 presents a summary of the prioritisation.

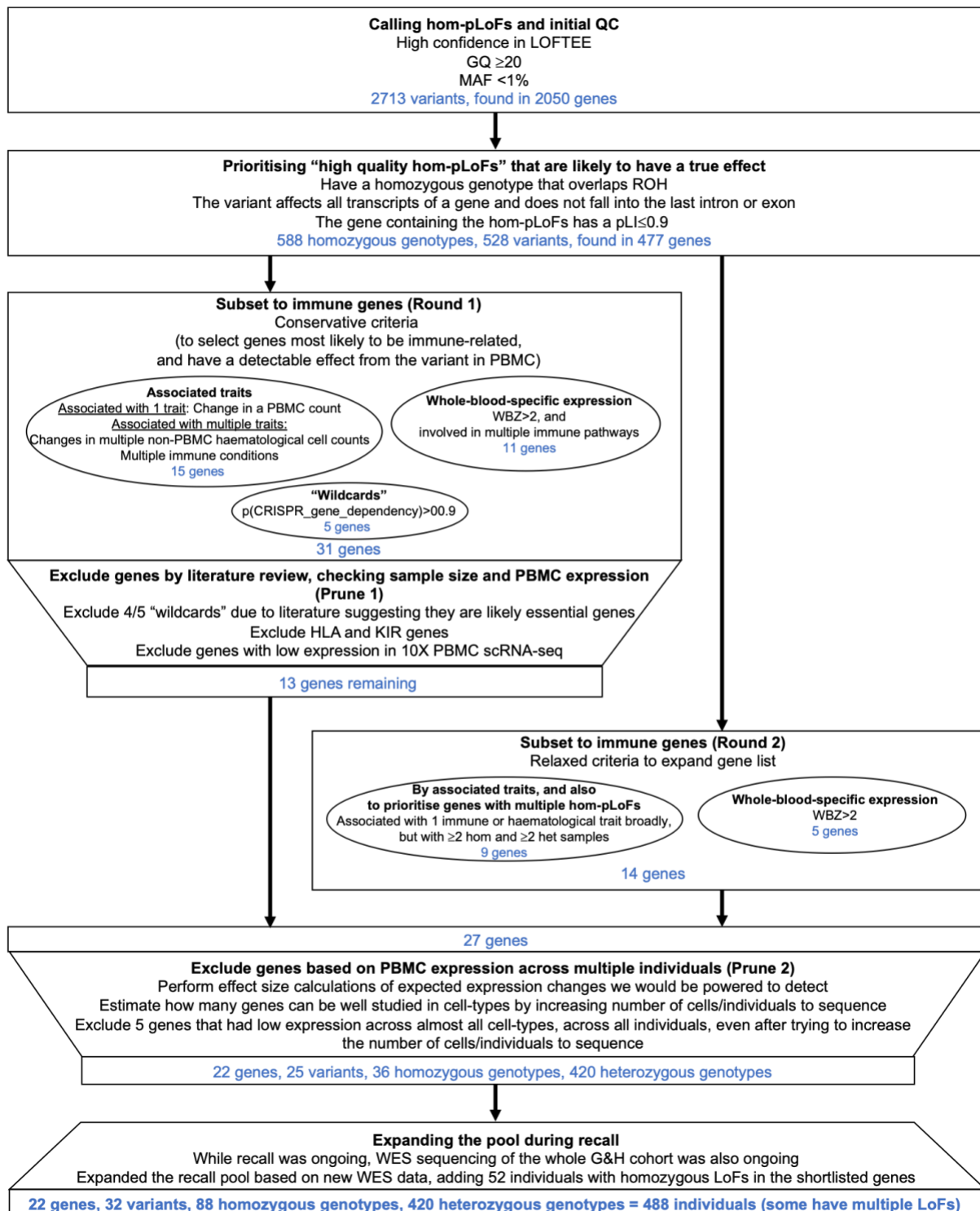


Figure-3.1

Flowchart of the steps taken to prioritise genes, hom-pLoFs and individuals for recall.

Calling and annotation of loss-of-function variants

Exome sequencing was performed on 5,236 Pakistani/Bangladeshi adults in G&H who self-declared as having consanguineous parents (This subset of the cohort will be called the WES5K). Variants were called with GATK HaplotypeCaller. These were performed by Hilary C. Martin before the start of my PhD.

Mapping, quality control and filtering steps have been summarised in Chapter 2. Variants were annotated with the Variant Effect Predictor (McLaren et al., 2016) and those with stop-gain, splice donor/acceptor, or frameshift consequence annotations were defined as pLoFs. Using the LOFTEE plugin (K. J. Karczewski et al., 2020) to annotate pLoFs as high or low confidence, only high confidence pLoFs with at least one homozygous genotype (hom-pLoFs) were retained.

Manual characterization of hom-pLoFs

Hom-pLoFs were compiled with the following characteristics:

- For each sample:
 - The genotype quality (GQ)
 - Whether the variant overlapped a region of homozygosity (ROH)
- For each variant:
 - The gene(s) and transcript(s) affected
 - Whether the variant affected all protein-coding transcripts of a gene
 - Whether the variant was in the last intron or exon of a gene
 - The number of individuals homozygous for the variant
- For each gene affected by a hom-pLoF*:
 - The LoF intolerance of the gene
 - The probability of a real depletion effect from CRISPR knockout in blood cell lines
 - “Locus-to-gene” (L2G) scores for known variants in the gene
 - Gene expression patterns

*See subsequent sections for descriptions of how these characteristics were compiled.

LoF Intolerance

The probability of LoF intolerance (pLI) and the probability of biallelic (recessive) LoF intolerance (pRec) of each gene were collated from the Exome Aggregation Consortium (ExAC) and gnomAD (K. J. Karczewski et al., 2020; Lek et al., 2016).

Prediction of the effects of CRISPR gene knockout in haematological cancer cell lines

From the Dependency Map Portal (DepMap) (DepMap, 2021), for each gene, we collated the median probability of a real depletion effect from CRISPR knockout across 48 blood lineage cell lines (defined as $p(\text{CRISPR\_gene\_dependency})$)

Locus-to-gene scoring of GWAS performed in genes affected, with categorisation

From the Open Targets portal, “Locus-to-gene” (L2G) scores that grade the association of a locus to a trait / disease in a GWAS were extracted for genes likely affected by hom-pLoFs (Ghoussaini et al., 2021). Traits from studies with an L2G score ≥ 0.5 were classified into categories based on the organ system affected, and subsequently further grouped into “immunologic” and “other” traits. The distribution of associated traits was then compared across different subsets of genes. Specifically, we performed chi-squared tests to determine whether different gene sets had a different proportion of associations to immunologic traits.

Gene expression analyses

The goals of gene expression analysis were firstly to evaluate if prioritising genes that were highly-expressed or differentially-expressed in blood or blood cell subtypes would select for genes with immune functions, and secondly, to evaluate if genes of interest had expression profiles that were appropriate for study with PBMC scRNAseq.

Analysing gene expression across tissues for tissue-specific (Whole blood) genes

Median transcript-per-million (TPM) expression values were downloaded from the Gene-Tissue Expression Project (GTEx) (GTEx Consortium et al., 2017). The top 100 genes containing hom-pLoFs expressed in whole blood were determined. Whole-blood specific genes were defined as genes with z-score, or “Whole Blood Z-score” (WBZ), > 2 after the median TPMs were scaled across all tissue types.

$$Z_i = \frac{x_i - u_i}{\sigma_i}$$

z_i : z – score of gene i

x_i : median TPM of gene i in whole blood across all individuals

u_i : mean of the median TPMs of gene i across all tissue types in GTEx

σ_i : standard deviation of median TPMs of gene i across all tissue types in GTEx

Analysing gene expression in venous blood and PBMC for cell-type-specific genes

A scRNAseq dataset of 3,000 PBMCs from a healthy donor from 10x Genomics (10xGenomics, n.d.) was annotated using canonical marker genes (satijalab.org, 2020).

The average expression of each gene in each cell-type was extracted (Seurat R package) (Hao et al., 2021). We determined the top 100 highly-expressed genes containing hom-pLoFs. Cell-type-specific genes were defined as genes with a z-score >2, with the z-score defined as follows:

$$Z_{ij} = \frac{x_{ij} - u_i}{\sigma_i}$$

j: cell – type cluster in PBMC scRNAseq

z_{ij}: z – score of gene *i* in cluster *j*

x_{ij}: average expression of gene *i* in cluster *j*

u_i: mean of the average expression (of each cluster) of gene *i* across all clusters

σ_i: standard deviation of the average expression (of each cluster) of gene *i* across all clusters

Cell-type-specificity of genes in PBMC scRNAseq data was validated by comparing to RNA-expression in venous blood samples from the BLUEPRINT project (Fernández et al., 2016), in which cell-types were separated by sorting before bulk RNA-sequencing. Cell-types in BLUEPRINT were matched to the identified clusters from the PBMC scRNAseq data, and the median TPMs were scaled. The gene expression values and z-scores were correlated between the two datasets.

Power calculations with PBMC scRNAseq

In order to estimate which genes we would have the statistical power to detect knockdown effects for in PBMC scRNAseq, and to calculate the number of cells per individual to sequence, we analysed scRNAseq from immune cells of 32 healthy Dutch individuals (henceforth referred to as the ‘1M-scBloodNL’ dataset) (Oelen et al., 2021). We used the 10x v3 chemistry samples. With the distribution (across all individuals) of mean normalised (SCTransform, Seurat) counts of gene transcripts for each cell-type, we performed one-tailed T-tests (n=3, p≤0.05, power 0.8) to estimate the magnitude of the decrease in expression required for a significant result.

Additional prioritisation during the recall to increase sample size

While the recall was ongoing, increasing numbers of the G&H cohort were being whole exome sequenced, funded by an industry consortium. From this dataset, 52 newly-identified individuals homozygous for pLoFs affecting genes with pLoFs already recruited were added to the recall pool to boost power.

Recall-by-genotype

The recall workflow

The recall was performed by the G&H clinic based in East London over the span of 1 year from 8 March 2022 to 12 March 2023. Having previously consented during recruitment into G&H to be approached for recall-by-genotype, participants prioritised in the recall and control pools were approached by a phone call. Individuals with hom-pLoFs were requested to invite first-degree relatives to participate in the study as well. It was likely these relatives would be at least heterozygous for the same pLoF, and this was done to increase power.

Consenting individuals were scheduled for a weekday clinic session where 10-20 millilitres (mL) of venous blood was drawn by a trained phlebotomist. This consisted of one 2mL tube, and one to two 10mL tubes per individual (EDTA-coated phlebotomy tubes). The samples were stored in 4°C refrigerators for the clinic day, and then transferred to a pre-chilled cold box and transported to the Wellcome Sanger Institute overnight or within 72 hours (when collected during Friday clinic). The 2mL tube of blood was processed with other Cardinal samples; The 10mL tubes were collected specifically for PBMC banking in this project, and were cryogenically stored for further studies (Chapter 4). Blood samples were processed by staff that were blinded to their pLoF status (including me), preventing bias during PBMC extraction. At several intervals during the recall, I was unblinded to the identities of the already-processed samples, in order to rearrange the recall pool, prioritising other individuals carrying pLoFs that were present amongst the individuals already recalled, to boost power in the project.

Figure-3.2 summarises the recall workflow, with emphasis on the differences between the samples for Cardinal and the samples for deeper scRNAseq (Chapter 4) in this project.

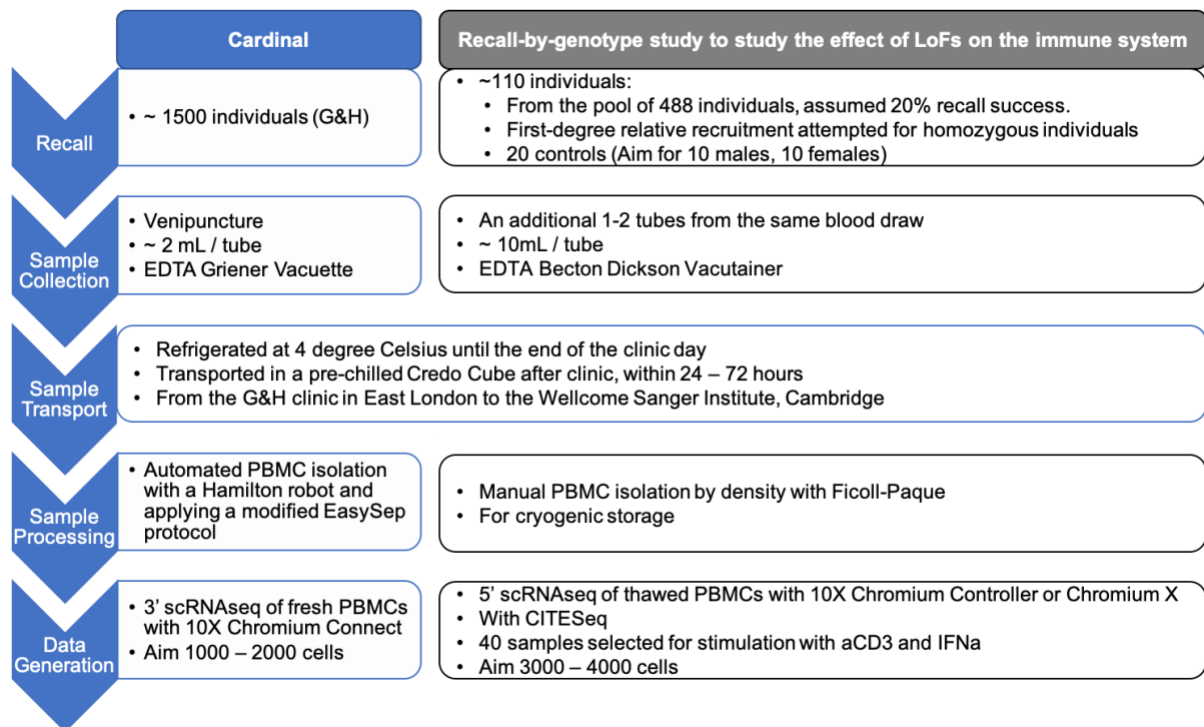


Figure-3.2

A summary of the recall workflow from sample collection to data generation. The values reflected here were the planned estimates. Sample processing and data generation of the cryogenically-banked PBMCs is described in further detail in Chapter 4.

Further manual curation of prioritised pLoFs

While the recall was ongoing, the prioritised pLoFs were subjected to further curation, using both an improved LoF assessment pipeline (CPG LoF Curation SOP V2: Centre for Population Genomics Loss of Function Variant Curation Standard Operating Procedure Version 2) and manual annotation of individual samples. This was performed by Sam Bryen and other collaborators in Daniel MacArthur's team at the Garvan Institute of Medical Research. This additional curation was performed after the recall had begun, and we included in the recall some pLoFs that would be given "Likely_Not_LoF" and "Not_LoF" verdicts, to validate the curation pipeline.

The CPG LoF Curation SOP V2 is summarised in Figure-3.3. Using information related to the pLoF and individual samples, possible technical errors or impact errors were indicated by the pipeline. Based on the flags, the pLoF was assigned to one of five classifications based on the likelihood of the pLoF generating a true effect. Individual flags that are relevant will be elaborated further in the results and discussion.

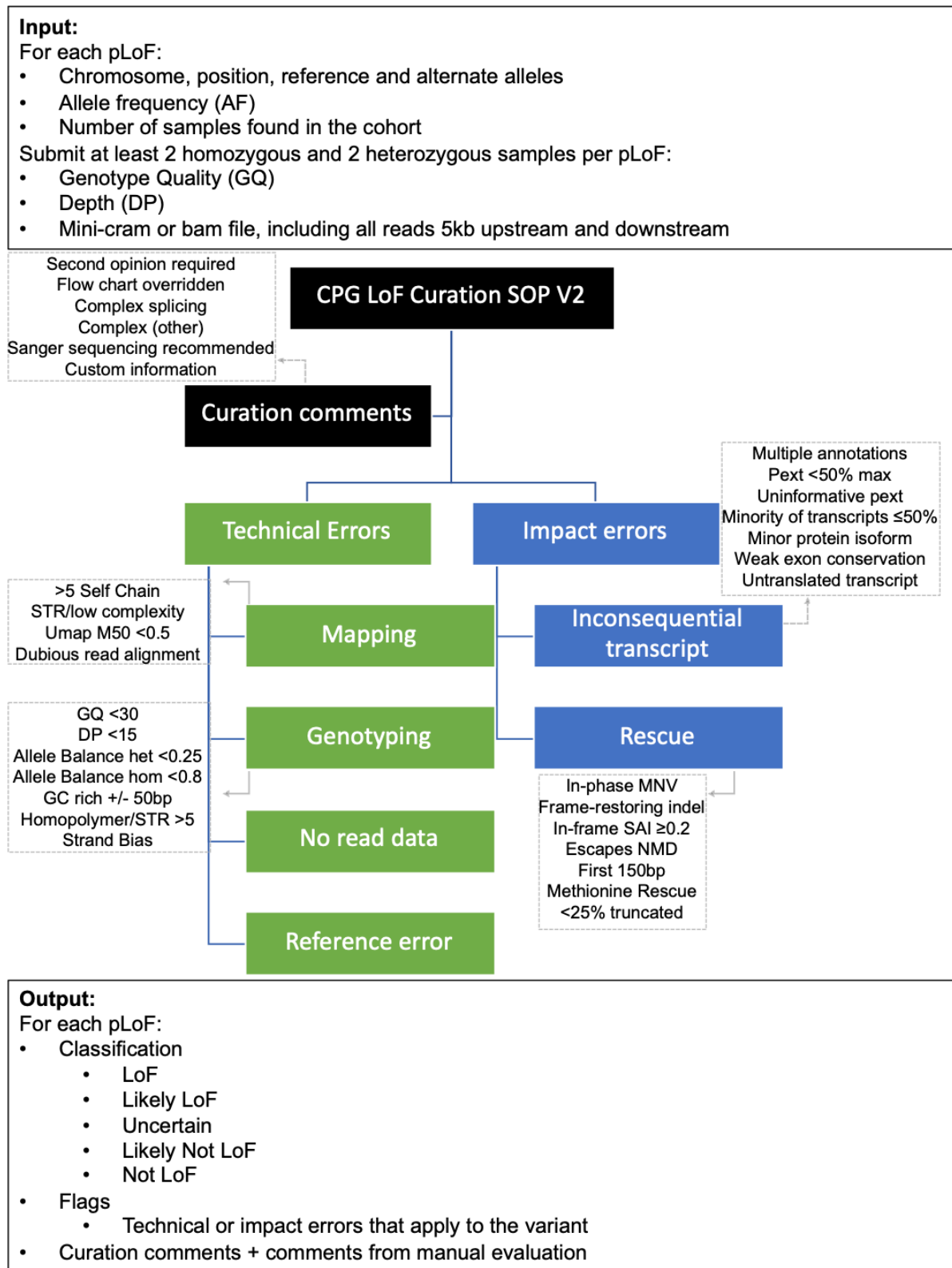


Figure-3.3

Summary of the CPG LoF Curation SOP V2 workflow.

The scRNAseq data generation, processing and quality control (by Cardinal)

The work in this section was performed by the Cardinal team. It is therefore only briefly summarised in this chapter. I performed my own version of these steps, following the same general principles, from PBMC extraction to quality control (QC), but on the new data that formed the extension of this project (Chapter 4). Therefore to avoid repetition, detailed information of these steps, and any relevant theory and results, will be covered in the next chapter. Emphasis in this section will be on the parameters that are unique to Cardinal and differ from my experimental design in Chapter 4.

PBMC isolation and scRNAseq

From the 2mL EDTA tubes, 150uL of blood per sample was plated into each well of a 96-well plate. Subsequent steps were programmed and executed by a Hamilton Star liquid handler ("Hamilton Company," 1972). PBMCs were then extracted with the Stem Cell Technologies PBMC EasySep Direct Kit following a modified protocol (STEMCELL Technologies, n.d.).

The Cardinal project received samples from multiple clinics across two cohorts (UK Biobank and G&H) on any given day. Therefore, samples were pooled in a manner to ensure that every clinic was approximately equally represented across the pools. Twenty thousand cells per donor were combined together into pools of ~11 donors (mean) and concentrated to ~6400 cells/uL, aiming to load 25,000 cells per channel, to achieve ~1,000-2,000 cells per donor after quality control filtering.

scRNAseq was performed with the following kits and parameters: (a) For single-cell encapsulation and library preparation, the 3' Library and Gel Bead Kit and Automated Chip G, on the Chromium Connect, followed by the dual index kit, all from 10X Genomics; (b) Twelve polymerase chain reaction (PCR) cycles and ten PCR cycles for cDNA and Illumina index amplifications respectively; and (c) for sequencing, the NovaSeq 6000 S4 flow cell with read parameters 26 base pairs (bp) (Read 1), 8bp (Index), and 98bp (Read 2).

A small subset of samples were rerun on thawed cryopreserved PBMCs or using the 5' library, in some cases because of the low sample quality of the initial run. Samples that were from thawed cells or resulted in a 5' library were annotated accordingly in my analysis.

scRNAseq quality control

Read alignment of the sequencing output was performed with Cellranger to generate feature matrices. The YASCP pipeline ("Yet Another Single Cell Pipeline") developed by the Sanger Institute's Human Genetics Informatics Team was then deployed to perform automated

processing of the scRNAseq data (Ozols, 2023). Steps (relevant to this project) included ambient RNA removal, donor deconvolution, cell-type annotation, and quality control at the cell-level, pool-level and donor-level. I employed a similar approach with manual adjustments tailored to this project in Chapter 4, therefore further discussion is carried out there.

Data analysis

After processing by the Cardinal pipeline, deconvoluted and annotated cells assigned to donors from my recall study were identified for further analyses specific to this project. A total of 132 samples from 82 pools were suitable for downstream analysis, corresponding to 108 out of the 114 donors in this study, with some donors having repeated samples, mostly due to the need of reruns because of poor sample quality. All six donors that were excluded were pLoF individuals, one homozygous and five heterozygous, and the reasons for dropout were as follows: missing sample, possible sample swap, two samples without cells assigned after pool deconvolution, and two failed at earlier stages before the data freeze.

Dimensionality reduction and identification of covariates for expression analysis

Cells were “pseudobulked” by their Azimuth level 1 cell-type annotation (Hao et al., 2021), i.e. RNA counts from cells of the same type in the same donor were aggregated (summed) together. DESeq2 was run to perform a variance stabilised transformation (VST) on the aggregated counts (Love et al., 2014). Principal component analysis (PCA) with DESeq2 and pcatools (Blighe et al., 2022) was performed to identify factors that influenced gene expression, and determine covariates for linear modelling downstream.

At the single-cell level, the Seurat SCTransform function was used to perform the normalisation and VST of individual cell counts. Dimensionality reduction with PCA, uniform manifold approximation and projection (UMAP) and reciprocal PCA (RPCA) integration was performed for visualisation of the data, using the Seurat R package.

Technical batch-correction and statistical modelling of gene expression

From the data visualisation in the previous step, I found that technical batch correction was required to correct for effects on expression that correlated with technical factors during sample processing such as the pool the sample belonged to and whether the expression data came from fresh or frozen PBMCs. However, as these samples were pooled with samples from the wider Cardinal project, some samples were the only sample from this project in their pool. Therefore it was necessary to incorporate the other samples in the pool

into the model when performing batch correction, to avoid regressing out LoF effects erroneously attributed to pool-level batch effects.

The wider Cardinal data contained UK biobank samples, which were firstly of different ancestry to G&H and secondly, at the time of processing, contained incomplete biological information such as age and sex. This made testing for LoF effects while controlling for all covariates in one step impossible. A two-step model was thus fitted, where control individuals from this project, together with Cardinal samples (from both G&H and UK Biobank) that came from pools containing this project's samples were used to fit the first model:

$$Expression \sim Pool + State + Cohort + (1 | Donor)$$

Expression: VST of aggregated counts from pseudobulked data

Pool: The pool the sample is multiplexed in during scRNAseq

State: Whether the sample was sequenced 24, 48, or 72 hours after collection,

or whether the sample was processed from frozen, or with 5' technology

(These are all treated as categorical variables)

Cohort: Whether the sample is from G&H or UK Biobank

A variable was included to control for donor-associated random effects, as several donors were seeded across multiple pools, or sequenced multiple times in different states (e.g. one fresh, one frozen). With the coefficients generated from this first model, expression of samples from this project (pLoF carriers and controls) were corrected and the residuals were fitted into the second step:

$$Residual \sim Genotype\ Encoding + Age + Sex$$

Residual: The residuals from regressing the VST aggregated expression with

the coefficients calculated from Step 1.

*Genotype Encoding: ([0,1,2], [0,0,1], [0,1,1]) * for the pLoF or any pLoF in a gene*

** Depending on the inheritance model tested (respectively: additive, recessive, dominant)*

The second step only contained the individuals for this recall, for which we had all the necessary covariate information. For each pLoF, we combined the controls with the individuals containing that pLoF to run the model, rather than including all individuals in the project. This was to avoid LoF effects from other samples affecting one another, as these are pLoFs affecting immune genes that may operate in related pathways. This was also repeated at the gene-level, where individuals with different pLoFs but affecting the same gene were run together, in these cases, the encoding variable was changed to whether the sample was heterozygous or homozygous for any pLoF in that gene. Further discussion about this has been made in Chapter 2.

The two-step model also brought with it the added advantage that it circumvented the unlikely (but not impossible) problem that these UK Biobank samples, treated as controls, might contain the same pLoF as the G&H sample. We felt that LoF calling in that cohort was beyond the scope of this project.

For most of the pLoFs recalled, there was only one homozygous individual, and therefore there was insufficient power to perform formal testing of recessive effects on gene expression. Thus, taking advantage of the rest of the Cardinal samples used in step 1 of the model, the distribution of the residual expression was plotted for each gene of interest, and as most followed a normal distribution, the Z-score of the homozygous individual's gene expression was calculated based on the cohort distribution.

Lastly, at the single-cell level, SCTransformed expression was modelled against genotype with a random effect for donor (to account for multiple cells from the same donor), to supplement the results from the pseudobulked analysis. This was run for each cell-type at the resolution used for pseudobulking, to allow for comparisons.

$$Expression \sim Genotype + (1 | Donor)$$

Linear mixed modelling was performed with the lmerSeq R package (Vestal et al., 2022), and R linear modelling was used in situations where mixed modelling was not required.

Pathway enrichment analysis

In order to identify patterns in the genes that were differentially expressed between pLoF carriers and controls, where relevant, we performed pathway enrichment analysis on the differential gene expression results obtained in the previous step. We filtered to genes with a log2 fold change difference of +/- 0.01 at nominal significance (p-value ≤ 0.05), and performed over-representation analysis with the enrichR package (E. Y. Chen et al., 2013),

looking to see if these genes were enriched in any pathways from the Reactome 2022 database (Gillespie et al., 2022).

Results

Prioritisation of individuals for recall

Identifying characteristics that would increase the likelihood of selecting for hom-pLoFs with true effects

Since we know that pLoFs are enriched for artefacts and that there are many compensatory effects that would diminish pLoF knockdown, we began this project by evaluating properties of the hom-pLoFs and the genes they affect, identifying metrics that might be informative about whether these genotypes were likely to result in true loss-of-function.

From the exome sequences of 5,236 Pakistani/Bangladeshi adults in G&H who self-declared as having consanguineous parents, 2,713 hom-pLoFs were called. The variants were rare; ~80% had ≤ 10 individuals who were homozygous for the variant ($n=2,143$, Figure-3.4A) and ~70% had an allele frequency of <0.01 ($n=1,864$, Figure-3.4B).

As these variants were rare, they were thus unlikely to be truly homozygous if outside a region of homozygosity (ROH) (where their probability is p^2 , versus p if in a ROH, where p is the allele frequency). Therefore, hom-pLoFs that were called outside ROHs were hypothesised to be more likely spurious. Indeed, of the variants with minor allele frequency (MAF) <0.01 , ~90% ($n=2,099/2,367$) of the samples had a genotype quality (GQ) ≥ 30 , and ~90% ($n=2,138/2,367$) overlapped ROHs. Homozygous genotypes that overlapped regions of homozygosity (ROHs) were shifted towards significantly higher GQ than homozygous genotypes that did not overlap ROHs (Figure-3.4C, Wilcoxon one-sided p -value $<2.2 \times 10^{-16}$).

Next, genes that are highly constrained against loss-of-function variation are unlikely to have true hom-LoFs (K. J. Karczewski et al., 2020; Lek et al., 2016; MacArthur et al., 2012). Indeed, genes with hom-pLoFs had lower pLI scores than genes without hom-pLoFs in G&H (Figure-3.4D, Wilcoxon one-sided p -value $<2.2 \times 10^{-16}$). In the ExAC and gnomAD studies, an additional metric, pRec, which is the probability of biallelic (recessive) LoF intolerance for a gene was also calculated, and we evaluated this in this study as well, but there was no significant difference in the distribution of pRec between genes affected by hom-pLoFs and the rest of the protein-coding genes (Supplementary-figure-3.1). We therefore proceeded with pLI in our prioritisation, as it was a more effective measure of LoF constraint than pRec.

Similarly, we hypothesised that *in vitro*, it is unlikely that genes that caused a depletion effect in blood cell lines with CRISPR knockout would tolerate true hom-pLoFs. As expected, the distribution of the probability of a real depletion effect from CRISPR knockout $p(\text{CRISPR\_gene\_dependency})$ in genes predicted to be affected by hom-pLoFs were shifted

towards lower values compared to the remainder of protein-coding genes not affected by hom-pLoFs (Wilcoxon one-sided p-value $=1.1 \times 10^{-9}$, Figure-3.4E).

Of the 2,713 hom-pLoFs predicted to affect 2,050 genes, 938 variants would affect all annotated protein-coding transcripts of 768 genes (~35-37% of variants and genes) and did not fall into the last intron or exon of the gene. This subset of variants were less likely affected by alternative splicing and hence more likely to result in true protein inactivation. Indeed, when we considered genes for which all transcripts were affected by hom-pLoFs, the distributions of pLI, p(CRISPR_gene_dependency) and pRec were on average lower than genes with hom-pLoFs that only affected a subset of transcripts (Wilcoxon one-sided p-values $<2.2 \times 10^{-16}$, $=5.0 \times 10^{-5}$, $=1.7 \times 10^{-3}$, Figure-3.4G-Figure-3.4I).

Lastly, we investigated whether hom-pLoFs in high pLI genes would be more likely to be spurious and have lower GQ, but surprisingly, pLI of genes affected by hom-pLoFs poorly correlated with the GQ of the variants affecting these genes (Figure-3.4F). This suggested that the other indicator, of whether a hom-pLoF overlapped a ROH, was more informative of genotype quality than pLI (particularly when pLI < 0.9).

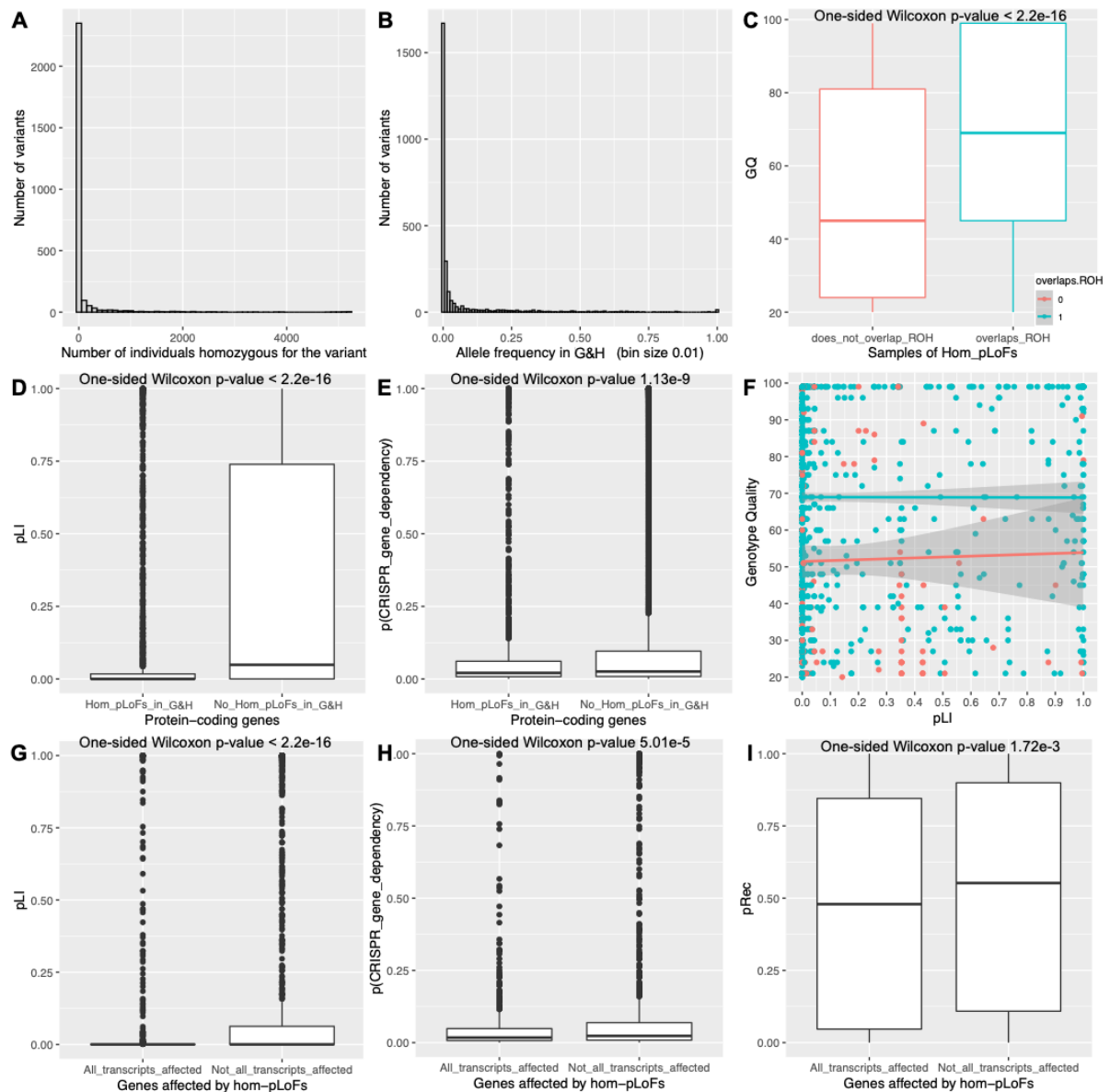


Figure-3.4

Characteristics of hom-pLoFs and the genes they affect. **A:** Distribution of the number of individuals with the same sequence of hom-pLoF. **B:** Distribution of allele frequencies in G&H of the hom-pLoFs. **C:** The distribution of GQs for hom-pLoFs overlapping RoHs trend higher than for hom-pLoFs not overlapping RoHs. **D & E:** Genes that contain hom-pLoFs in G&H are more likely to have a lower pLI (**D**) and p(CRISPR_gene_dependency) (**E**) than the remainder of protein-coding genes. **F** - The GQ of hom-pLoFs is not correlated with the pLI of the genes affected. **G, H & I:** Genes where all transcripts would be affected by hom-pLoFs are more likely to have lower pLI (**G**), p(CRISPR_gene_dependency) (**H**), and pRec (**I**) than genes where not all transcripts would be affected by hom-pLoFs.

Identifying characteristics that would select for immune-related genes

Next, as the tissue of choice for this study is blood, we aimed to select hom-pLoF-containing genes that are related to the immune and haematologic system. In order to do so we considered gene expression patterns in blood and genetic associations with haematological and immune traits.

Gene expression analyses of publicly available datasets

We hypothesised that genes with hom-pLoFs that were highly expressed in blood or demonstrated cell-type-specific expression (in blood subtypes) would more likely show transcriptomic variation as a result of the hom-pLoF in PBMCs. We evaluated the expression of the genes likely affected by hom-pLoFs in three blood datasets: (a) scRNAseq data of 3000 PBMCs from a healthy donor provided by 10x Genomics, (b) RNA-Seq data from venous blood samples in BLUEPRINT, and (c) global RNA expression data from whole blood in GTEx.

The PBMC scRNAseq sample was clustered into 9 cell-types with canonical markers (Figure-3.5A). Marker gene expressions in each cluster are demonstrated in Supplementary-figure-3.2 and Supplementary-figure-3.3. 470 genes with hom-pLoFs were defined as being specific to at least one cell-type (Figure-3.5C), having z-scores >2 in a cell-type cluster. They were prioritised for further analyses, together with the top 100 genes with hom-pLoFs with the highest average expression throughout all PBMCs.

Five of the cell-type clusters could be matched to cell-types found in the venous blood samples of BLUEPRINT. For genes affected by hom-pLoFs in each cell-type cluster, the median TPMs of the samples in BLUEPRINT correlated well with the cluster average expression in PBMC scRNAseq (Figure-3.5B, Supplementary-figure-3.4). When scaled among the five cell-type clusters in BLUEPRINT or in the PBMC dataset, there were 321 genes with z-score >1.645 (a 95th-percentile cutoff). Of these, 122 also demonstrated similar cell-type-specificity in BLUEPRINT when considering the same five cell-types (Supplementary-figure-3.5).

From GTEx, 115 genes with hom-pLoFs had z-scores >2 in whole blood when their median TPMs were scaled across all tissue types (Figure-3.5D) and they were collated along with the 100 affected-genes with the highest median TPM in whole blood (Supplementary-figure-3.6) for further analysis. As expected, clustering of these expression profiles revealed that spleen, EBV-transformed-lymphocytes and lung (likely enriched for immune functions) have the most similar expression profiles to whole blood.

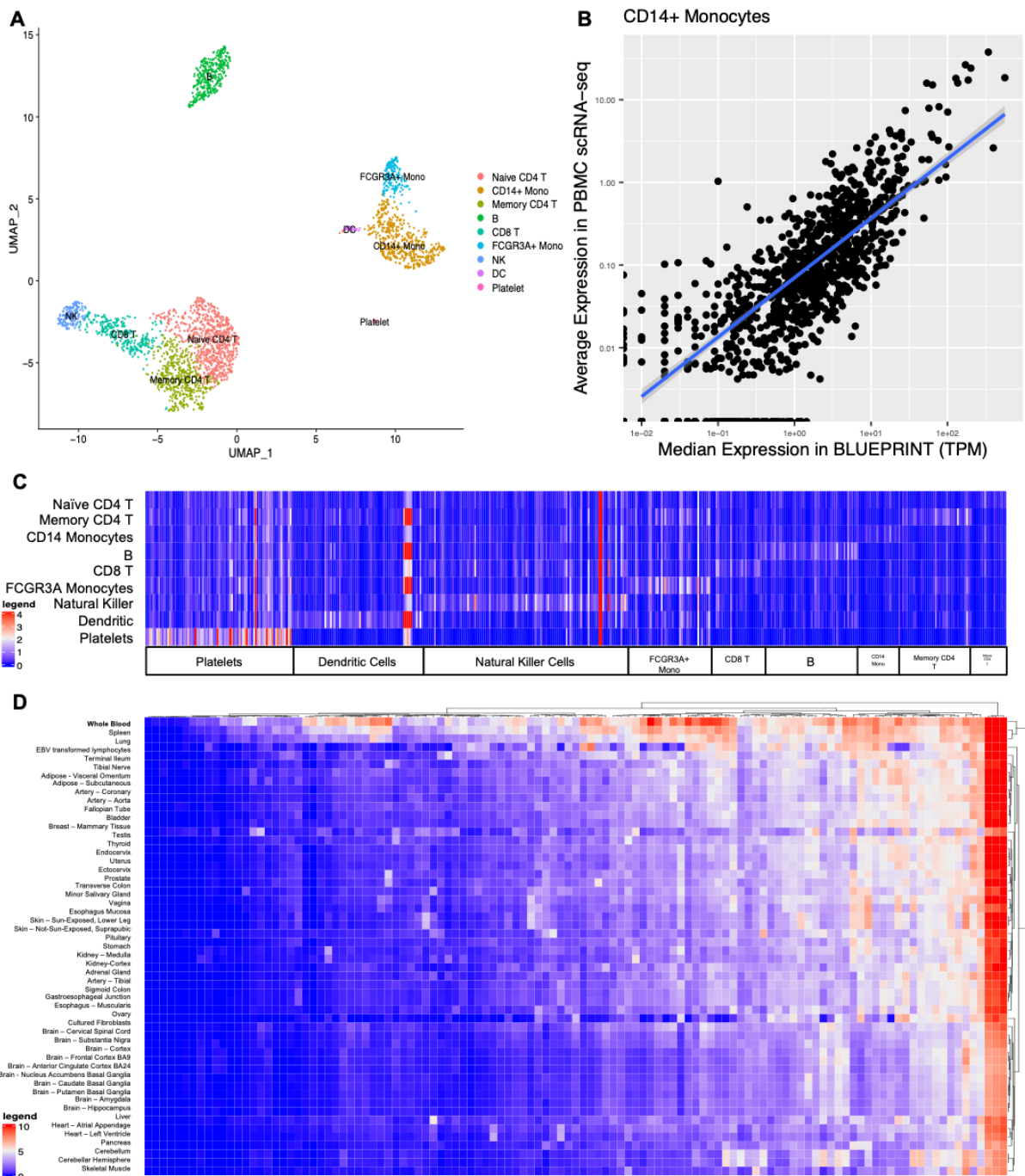


Figure-3.5

Gene expression in blood. A: Uniform Manifold Approximation and Projection (UMAP) of the PBMC scRNAseq data, with cell-type cluster annotations. **B:** Correlation between the average expression of genes in CD14+ monocytes in PMBC scRNAseq and the median TPM of genes across CD14+ monocytes samples in BLUEPRINT. **C:** Heatmap of the $\log_2(\text{average expression} + 1)$ of cell-type-specific genes with hom-pLoFs in PBMC scRNAseq. The y-axis represents the cell-type clusters, while the x-axis represents the genes with z-score > 2 in one of the cell-type cluster labelled in the black-border boxes. **D:**

Heatmap of the $\log_2(\text{MedianTPM} + 1)$ of genes with hom-pLoFs that are specific to whole blood (z-score >2) across all tissue types in GTEx.

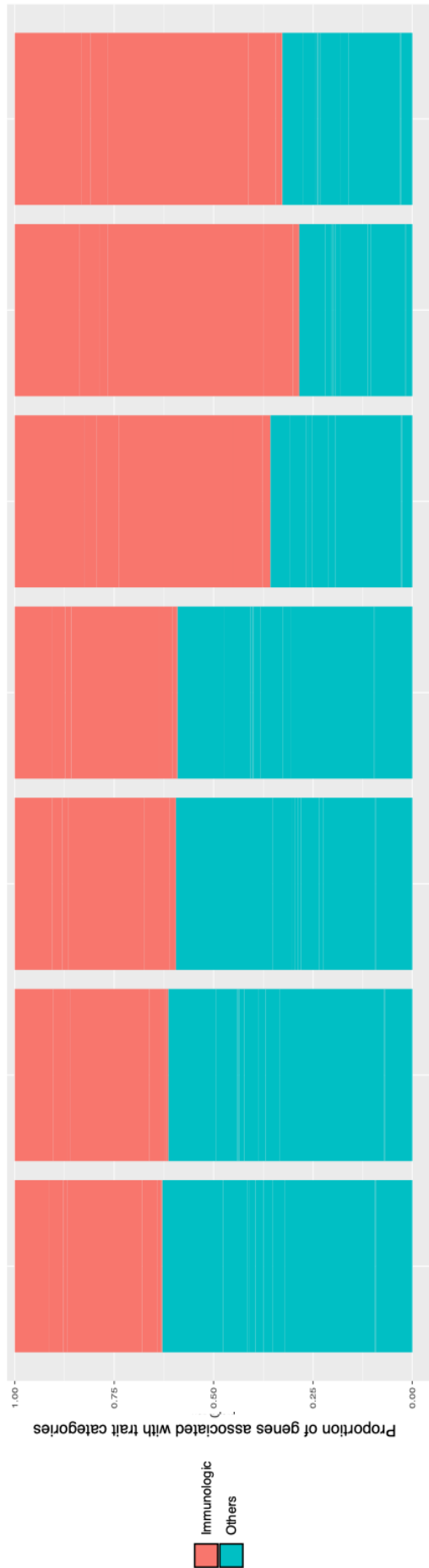
L2G scoring and traits associated with affected genes

969 (47%) of the 2050 affected-genes had an L2G score ≥ 0.5 in at least one GWAS, versus 41% of all genes. Of these, trait information could be extracted for 877 genes. The traits were grouped into 18 categories by system affected, and the categories were also divided into “Immunologic” and “Others” according to the involvement of the immune system.

The various gene subsets defined according to expression patterns (described above) were assigned trait information based on their L2G scoring. Throughout, about 38-53% of genes in each subset have L2G scores > 0.5 and associated trait information.

The proportion of genes associated with immunologic traits and diseases increased when selecting cell-type-specific genes in PBMC scRNAseq, genes specific to whole blood, and high-expressing genes in both PBMC scRNAseq and whole blood (chi-squared test with all affected-genes, p-values < 0.01). This is expected since genes expressed in blood are already known to be enriched for immune traits (Mohr & Liew, 2007).

These results are summarised in Figure-3.6. Based on these findings, we used whole-blood-specific expression, defined by the “whole blood z-score” (WBZ) as a prioritisation criterion to prioritise genes that may not have known immune trait associations in the OpenTargets Catalogue.



Gene Subset	Affected-genes		PBMC subsets			Whole Blood subsets	
	All Affected-genes	Affected-genes (with all transcripts affected)	Cell-type-specific (BLUEPRINT and PBMC scRNA-seq)	Cell-type-specific (PBMC scRNA-seq)	Top 100 Average Expression in PBMC scRNA-seq	Tissue-specific (Whole blood vs other tissues)	Top-100-Whole-Blood
Criteria for subsetting	All genes that would be affected by homozygous LOF variants	All transcripts affected by homozygous LOF variants	z-score > 1.645 in the same cell-type in BLUEPRINT and PBMC scRNA-seq	z-score > 2 in one cell-type in PBMC scRNA-seq	Top 100 Average Expression in PBMC scRNA-seq	z-score > 2 in whole blood	Top 100 median TPM in whole blood
No. of genes with $L2G > 0.5$ and trait identified/ No. of genes in subset (%)	877 / 2050 (42.8%)	324 / 756 (43.3%)	64 / 122 (52.5%)	209 / 470 (44.5%)	38 / 100 (38%)	60 / 115 (52.2%)	54 / 100 (54%)
chi-squared test with Affected-genes (Proportion of immunologic vs other traits)		p-value = 0.212	p-value = 0.141	p-value = 0.007	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16

Figure-3.6

Genes with associations to immunologic and other traits (View in landscape). Top:

Distribution of trait associations from the OpenTargets catalogue for genes affected by hom-pLoFs, stratified by gene subsets determined by gene expression. **Bottom:** Details of the gene subsets and chi-squared p-values between the gene subset and all genes likely affected by hom-pLoFs.

Deciding the criteria for recall prioritisation

Based on the observations above, the following criteria was decided upon to prioritise samples, hom-pLoFs and genes affected:

- A. High-confidence in LOFTEE
- B. Criteria for genotype prioritisation:
 - a. $GQ \geq 20$
 - b. Overlapped with a ROH in that sample
- C. Criteria for hom-pLoF prioritisation:
 - a. $AF < 0.01$
 - b. Affected all transcripts of a gene
 - c. Did not fall into the last intron or exon of the gene
- D. Criteria for gene prioritisation
 - a. $pLI \leq 0.9$

This left 588 genotypes at 528 high-quality hom-pLoFs variants in 477 genes. Next, we prioritised genes related to the haematological and immune system. We aimed to curate a small gene set in order to increase the number of individuals with interesting genotypes per gene and allow sensible statistical analyses downstream. In round 1 of prioritisation, a conservative approach was applied. Specifically, genes were prioritised if they were:

- directly associated in Open Targets ($L2G \geq 0.5$) with changes in PBMC cell count, implying that there would be a detectable effect in PBMC scRNAseq
- or were associated with multiple immune-related traits in Open Targets ($L2G \geq 0.5$):
 - associated to changes in multiple non-PBMC blood counts, suggesting the gene is implicated in general haematological pathways and unlikely to spare PBMC,
 - or associated with multiple immune conditions.

Next, we explored the possibility of including some genes based on their probability that a knockdown would have a depletion effect on blood cell lines (a “p(CRISPR_gene_dependency)” criterion). While p(CRISPR_gene_dependency) and pLI are positively correlated (linear regression p-value $< 2.2e-16$), there were genes with high p(CRISPR_gene_dependency) and low pLI which we hypothesised were essential for the survival of blood cell lines *in vitro*, but may not be essential for reproductive fitness or survival of the whole organism. It is possible that individuals carrying hom-pLoFs in genes with low pLI and high p(CRISPR_gene_dependency) have blood disorders which would be interesting to investigate by scRNAseq. Only five genes with hom-pLoFs had

p(CRISPR_gene_dependency) >0.9 and pLI <0.9, and of these, four have literature supporting their role as essential genes (*ATR*, *CENPE*, *SBDS*, *TIMELESS*), so we removed them, retaining just one (*ZC3H8*) that may have a possible immune role in thymocyte homeostasis (Hwang & Ho, 2002).

At this stage, one round of filtering was performed to exclude genes with high levels of diversity that are error-prone in sequencing data (HLA and KIR regions), genes with inadequate sample sizes and genes with low expression in PBMCs.

As many genes were removed, the gene set was subsequently expanded with a second round of prioritisation with relaxed criteria for trait associations and whole-blood-specificity. This resulted in 27 shortlisted genes.

Note: Several other approaches to prioritisation were attempted with different criteria, tabulated in Supplementary-table-3.1, before the systematic approach described here was decided upon.

Power calculations were performed during prioritisation

We next performed power calculations to identify if there was statistical power to detect knockdown effects on the expression of these elected genes. We used the 1M-scBloodNL dataset to determine the magnitude of expression changes we would have power to see in pLoF-affected samples. An example with *CAT* gene is demonstrated in Figure-3.7. In myeloid cell-types (dendritic cells (DCs), megakaryocytes, monocytes), the variance in expression across individuals is low enough and the magnitude of expression high enough that we would have reasonable power to detect a significant decrease of expression (21.9% below the mean for DCs, 71.9% for megakaryocytes, 20.5% for monocytes). In contrast, in lymphoid cell-types (where the average expression is low), we would need to see an expression value below zero (change has to be 162.9% below the mean) to see significant depletion, which is impossible.

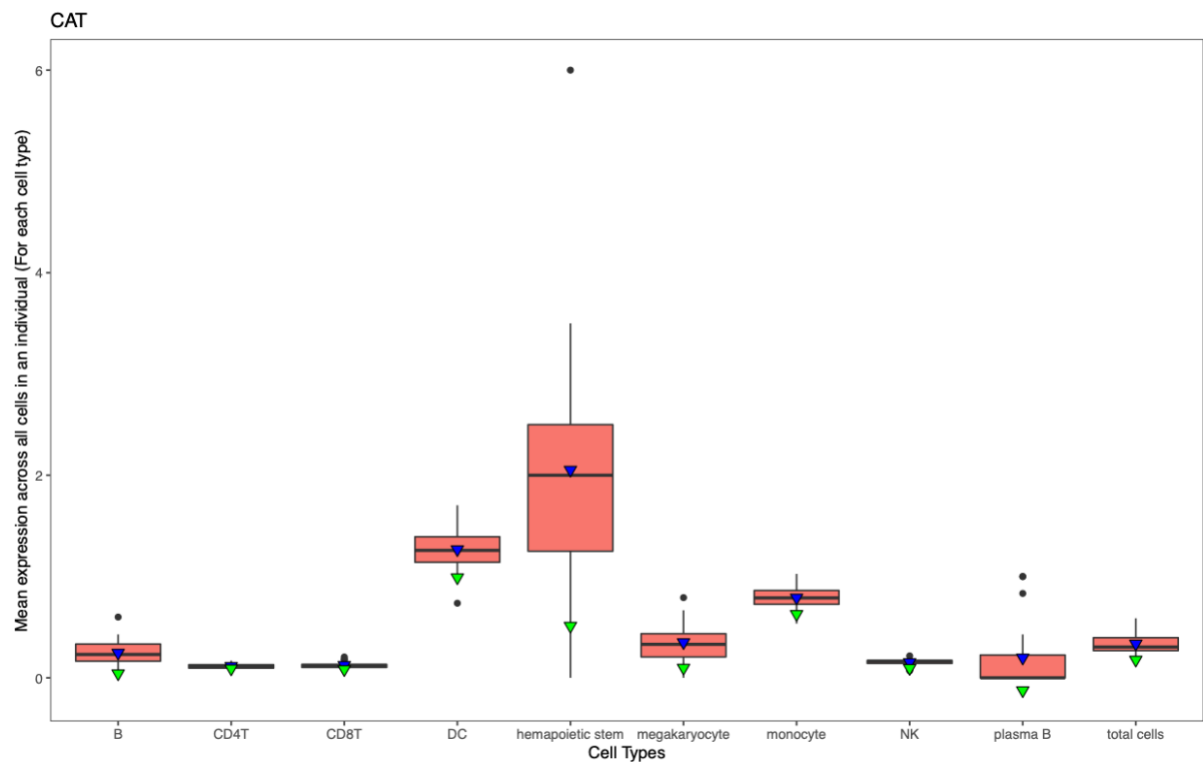


Figure-3.7

Distribution of mean normalised counts of *CAT* transcripts in each cell-type across 32 healthy Dutch individuals. The mean expression (across the n=32 cohort) is annotated with a blue triangle, and the green triangle annotates the minimum decrease in expression that would be required to see significant depletion at $p \leq 0.05$.

It is well-established that scRNAseq generates sparse expression matrices, with many cells presenting zero transcript counts for lowly-expressed genes (R. Jiang et al., 2022). This may pose challenges to estimating gene expression, especially when assuming standard count distribution models such as a normal distribution or a Poisson distribution (Bouland et al., 2023). Therefore, when estimating gene expression in a given cell-type, one would get a better estimate when more cells have detectable expression (i.e. non-zero transcript count), and this can be achieved by sequencing more cells.

We therefore performed a second power calculation to determine the number of cells we would have to sequence to accurately quantify gene expression in various cell-types. In each cell-type, we first calculated the median number of cells with detectable expression of our gene of interest. From previous PBMC scRNAseq datasets generated in the Trynka lab, we estimated that 50 cells with detectable expression would allow us to accurately quantify the magnitude of gene expression for that cell-type. We then determined, by extrapolation, how many additional cells would need to be sequenced for our gene of interest to pass this 50-cell threshold in each cell-type.

Hypothetically, assume cell-type “X” makes up 2.5% of PBMCs, and our target gene has detectable expression in 80% of the cell-type. At 1,000 cells sequenced, we would only have 20 cells with non-zero counts to estimate the magnitude of gene expression. At 3,000 PBMCs sequenced, we would expect 60 cells with detectable expression, and therefore would be able to calculate a more accurate gene expression for cell-type “X”. (Note: We acknowledge that this cut-off is slightly arbitrary, but we did not perform extensive optimisation at this point in time as this was an initial power calculation intended to aid in designing our study. In Chapter 4, we determined the number of cells to aggregate when calculating gene expression in a cell-type based on the actual distribution of cell counts per cell-type in our dataset.)

We performed this for our prioritised genes, and determined that we should aim for ~3,000 cells/sample. This motivated the deeper sequencing with more cells per sample in Chapter 4. In the context of the Cardinal data, where we were expecting ~1,000 cells/sample, we acknowledged that we might have to group cells into lower-resolution cell-types to maintain power, or limit analysis to the cell-type with the highest expression for each gene.

In addition, 5 genes were excluded due to low expression in nearly all cell-types (Prune 2), resulting in 22 genes. 436 individuals were homozygous or heterozygous for 25 pLoFs affecting these 22 prioritised genes (some have multiple pLoFs). This is summarised in

Table-3.1. Subsequently, during the recall, 52 newly-sequenced homozygous individuals of pLoFs already recalled were added to the recall pool to improve power (Figure-3.1).

Gene	Number individuals with homozygous LoFs	Number of individuals with heterozygous LoFs	Optimal number of cells to sequence	Number of cell-types with >50 cells with detectable expression at the optimal number of cells to sequence	Number of cell-types with >50 cells with detectable expression at 3,000 cells sequenced
CAT	2	4	3,000	6	6
CD109	6	21	2,000	1	Collapse to broader cell-types
CDC25B	2	36	5,000	5	4
CST7	1	7	4,000	5	4
DDX58	2	62	3,000	6	6
GCA	1	5	4,000	4	3
HPSE	3	19	1,000	2	2
IFIH1	3	11	5,000	5	4
MNDA	1	50	5,000	4	3
MYO1F	1	3	4,000	6	5
MYO1G	1	2	5,000	6	5
NOD2	1	6	1,000	2	2
RAD51B	1	5	5,000	2	1
SIGLEC14	1	2	2,000	2	2
SIGLEC5	1	42	5,000	2	Collapse to broader cell-types
SLC9A3R1	3	30	2,000	5	5
SYTL3	1	5	3,000	5	5
TMEM71	1	19	2,000	5	5
TRAT1	1	54	2,000	3	3
TRIM58	1	25	5,000	1	Collapse to broader cell-types
ZBP1	1	8	4,000	4	3
ZC3H8	1	4	4,000	4	3

Table-3.1

The number of cell-types with >50 cells with detectable expression at the optimal number of cells to sequence, and at 3,000 cells sequenced (View in landscape).

The demographics of the recalled participants

While planning, we assumed a ~20% recall success rate (after discussion with G&H), therefore anticipated that about 90-100 individuals would be successfully recalled. We would then have multiple genes with at least 3 individuals with pLoFs affecting the gene, which was the cutoff used during the sample size power calculations. Throughout the study we also targeted to recall 10 male and 10 female controls who have been exome-sequenced and do not have any of the “high-quality” pLoFs prioritised. Consequently, a total of 114 participants were included in the recall study: 22 controls, and 92 individuals with LoFs. The recalled individuals had 24 pLoFs affecting 19 genes, of which 17 pLoFs and 14 genes had a minimum of 3 individuals with either heterozygous or homozygous genotypes. Table-3.2 compiles the number of individuals for each pLoF. The detailed breakdown of the process behind “unblinding” these pLoF genotypes is illustrated in Supplementary-figure-3.7 with further explanation in its caption.

Recalled participants ranged in age from 19-71 years old with a median age of 45 years old. The ages in both control and pLoF groups were normally distributed (Shapiro-Wilk normality test p-values 0.4 and 0.5 respectively), and there was no statistically significant difference in the variance of the age distribution between controls and pLoF individuals (F test p-value 0.2). However, the mean age of control participants (39 years old) was lower than the mean age of pLoF participants (45 years old) (two-sided t-test p-value 0.03). This could reflect that younger participants were more willing to be recalled, and as the recall clinic could not be blinded to the participants’ pLoF status by design, more effort was put in to persuade individuals with pLoFs to participate, resulting in older individuals being retained. Nevertheless, the magnitude of the age difference is not big, and we do not anticipate that this would prevent sensible comparison between controls and pLoFs. (Figure-3.8A)

Forty-eight males (42%) and sixty-six females (58%) participated in this study, reflecting the sex distribution of G&H (44% and 56% respectively) (Finer et al., 2020). There is no difference between the sex distributions of the control group (9 males (41%), 13 females (59%)) and the pLoF group (39 males (42%), 53 females (58%)) (chi-squared p-value =0.9). (Figure-3.8B)

However, due to their smaller sample sizes, age and sex distributions within individual pLoFs may differ significantly from the study population (or the cohort population), simply by chance.

Gene	Variant	N_Total	N_Hom	N_Het
control	control	22	NA	NA
<i>CAT</i>	chr11_34456818_G_C	2	0	2
<i>CD109</i>	chr6_73736509_G_A	8	2	6
<i>CD109</i>	chr6_73758966_T_TTG	3	1	2
<i>CD109</i>	chr6_73782709_C_T	1	1	0
<i>CDC25B</i>	chr20_3787140_G_GT	6	1	5
<i>CST7</i>	chr20_24957358_G_T	3	0	3
<i>DDX58</i>	chr9_32466355_C_A	7	0	7
<i>DDX58</i>	chr9_32488757_GA_G	5	2	3
<i>HPSE</i>	chr4_83302158_G_GT	4	1	3
<i>IFIH1</i>	chr2_162277515_AC_A	5	1	4
<i>MNDA</i>	chr1_158845998_C_T	13	1	12
<i>MYO1F</i>	chr19_8522497_G_A	1	0	1
<i>MYO1G</i>	chr7_44964942_TAC_T	1	0	1
<i>NOD2</i>	chr16_50716612_C_T	3	0	3
<i>SIGLEC14</i>	chr19_51645475_A_G	5	5	0
<i>SIGLEC14</i>	chr19_51643777_ACT_A	1	1	0
<i>SIGLEC5</i>	chr19_51628003_A_AG	5	1	4
<i>SLC9A3R1</i>	chr17_74749171_G_T	5	1	4
<i>TMEM71</i>	chr8_132722068_C_CAGCA	3	0	3
<i>TMEM71</i>	chr8_132758873_G_A	3	2	1
<i>TRAT1</i>	chr3_108849184_GCTATGAA_G	7	1	6
<i>TRIM58</i>	chr1_247857547_G_T	5	0	5
<i>ZBP1</i>	chr20_57615536_CA_C	1	0	1
<i>ZC3H8</i>	chr2_112230930_CTTCTT_C	2	1	1

Table-3.2

The number of individuals (N_Total), homozygous (N_Hom) and heterozygous (N_Het) recalled for each pLoF.

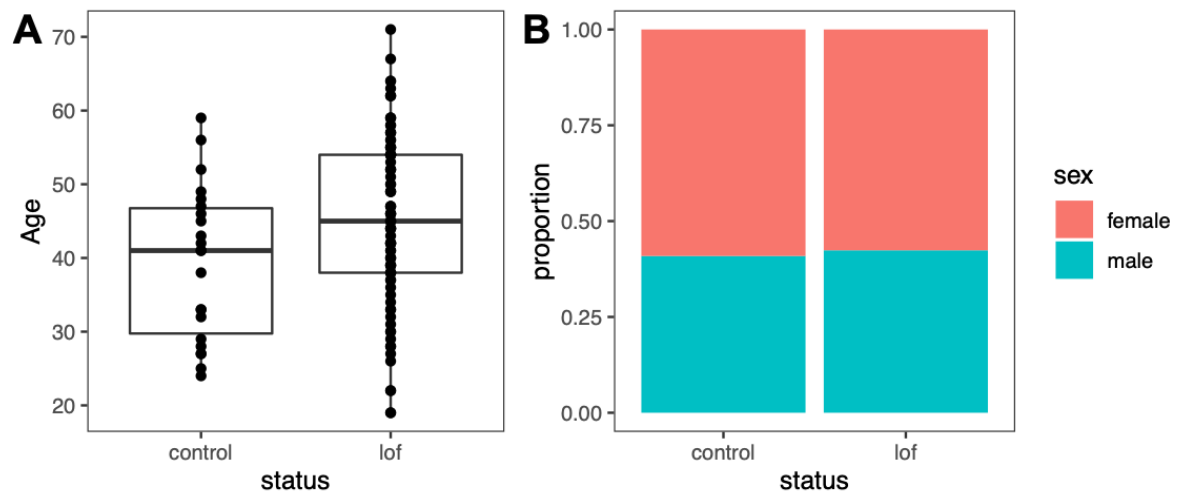


Figure-3.8

Age and sex distributions between control individuals and pLoF individuals.

Visualisation of the Cardinal scRNAseq data

Dimensionality reduction and visualisation at the single-cell level

The current analysis was performed on “freeze 2” of the Cardinal dataset, comprising a total of 132 samples corresponding to 108 donors. As Cardinal is a large-scale ongoing project with multiple goals outside of the scope of this project, the final freeze was not available at the time of writing.

For duplicated samples, we chose the sample with the most cells passing Cardinal QC as a single representative sample. Combined, the 108 samples from 73 pools contained 238,707 cells with a median of 2,043 cells/sample. (Figure-3.9A)

The goals of visualisation and integration were firstly, to identify the drivers of variation in gene expression in the data, so that relevant covariates could be identified and controlled for in differential gene expression analyses; secondly, to test if any pLoFs had effects that would drive pLoF-carrying cells to cluster away from the rest.

The first round of dimensionality reduction (performed with 2,000 most highly variable genes, PCA up to 50 PCs, then UMAP with 35 PCs) revealed that while cells were clustering based on their cell-type annotations, there were notable batch effects driven by the state at which the cells were harvested for sequencing (referred in this chapter as “state”: 24h, 48h, 72h fresh after collection, frozen and thawed, or with 5' technology). As pools contained samples all in the same state, the batch effect could be noticeable when classifying cells by pools as well. (Figure-3.9C)

Pool and state batch effects were driven by cell cycle genes, reflecting the differences in cell cycle stages between 24h samples and frozen samples in particular. Ribosomal and mitochondrial genes, a measure of cell health, were not drivers of variation between states, reflecting that cells were relatively equally healthy throughout the dataset. (Figure-3.9B)

In order to test if biological clusters due to LoF effects would emerge in the absence of technical effects, we attempted to integrate the data. Single-cell integration is an approach to match biological cell states across datasets, accounting for technical differences while allowing biological differences to be preserved (Ryu et al., 2023). T cell and B cell receptors (TCR, BCR), mitochondrial, ribosomal, and cell cycle genes were excluded from the list of highly variable genes used to perform another PCA (Note: TCRs and BCRs are very diverse and result in a high degree of structure in the data that do not reflect biologically-relevant cell-type differentiation. They are therefore excluded, except marker genes required for the identification of mucosal-associated invariant T cells and invariant natural killer cells were

retained). RPCA integration was then performed with 30 PCs, with the 24h state cells set as the reference, for cells from other states to be integrated into. Variation related to pools and states were successfully resolved in this manner, and the remaining outlier clusters corresponded to rare cell-types that have cell contributions from nearly all donors. No pLoF-containing cells clustered away from other donors, suggesting that LoF effects, if any, were probably more subtle. (Figure-3.9C).

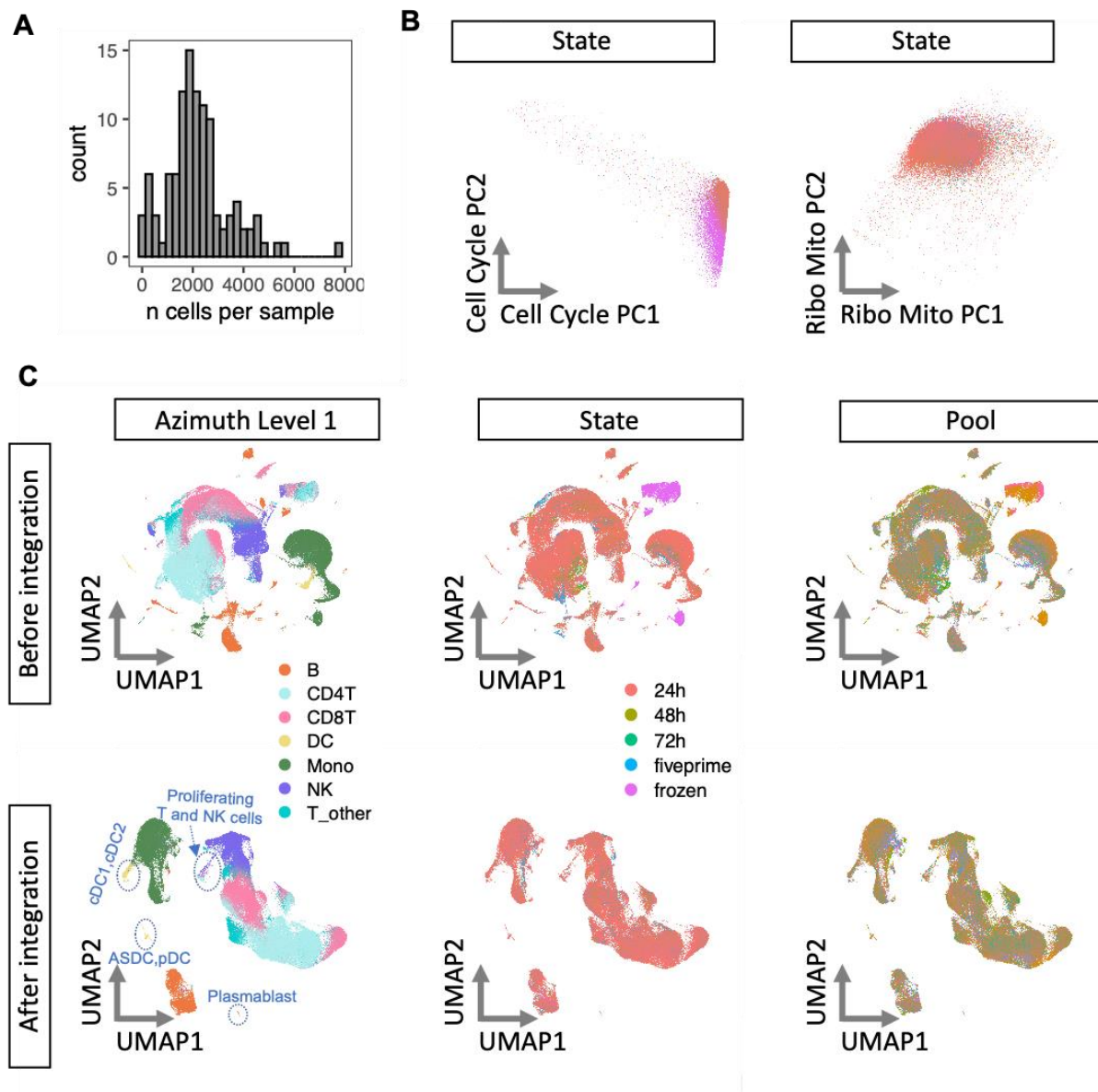


Figure-3.9

Visualising scRNAseq data from Cardinal. A: The distribution of the number of cells per sample. **B:** PCA of the following gene subsets: cell cycle genes (left), ribosomal and mitochondrial genes (right). Cells are coloured according to the legend in **C** present in plots with the same title. **C:** Top row: PCA & UMAP of the top 2000 most variable genes. Bottom row: After excluding TCR, BCR, cell cycle, mitochondrial, ribosomal genes, RPCA integration and UMAP. Outlier clusters are composed of rare cell-types, with contributions from nearly every donor. From left to right, cells are coloured by their azimuth level 1 annotations, state, and pool. There are 73 pools represented here, therefore a legend for pool colour is not provided.

Visualisation of the pseudobulked data and technical batch effect correction

For each azimuth level 1 cell-type, cell counts were aggregated within each sample. PCA was then performed on the variance-stabilised-transformed expression. Several samples were outliers based on the first two PCs, and they corresponded to samples that failed Cardinal's recommended criteria for sample-level QC: at least 400 cells per sample, 200 cells that pass cell-level QC per sample, at least 400 unique molecular identifiers per sample. (Figure-3.10A)

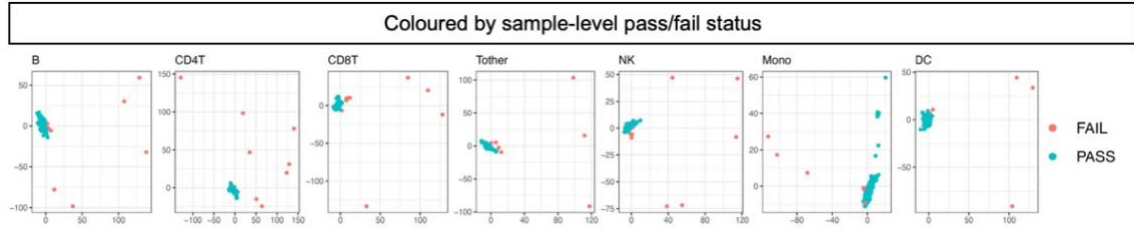
From 129 samples corresponding to 108 donors, we filtered to samples that passed sample-level QC, keeping 119 samples corresponding to 98 donors. We performed another PCA and, like in the scRNAseq visualisation, state and pool were the two variables driving the first two PCs. (Figure-3.10B)

As Cardinal is designed to profile fresh PBMCs, the samples have been pooled based on the day that they were received, and therefore pooled with other samples in the wider Cardinal project. As a result, some pools were represented in this recall project by only one sample, and sometimes that sample was a pLoF sample. Therefore, batch-correction for pool effects exploited other samples from Cardinal, to avoid correcting away LoF effects. Further explanation of the analysis of gene expression data with batch correction is provided in the Methods section. In the first step of our statistical model, we used our controls and other samples in each pool that were part of the wider Cardinal study, excluding samples with our pLoFs of interest, to model technical effects on expression. To improve the modelling, multiple samples from the same donor across multiple pools were used, and the donor was fitted as a random effect. 717 samples from 687 donors distributed across 80 pools were included in this first step.

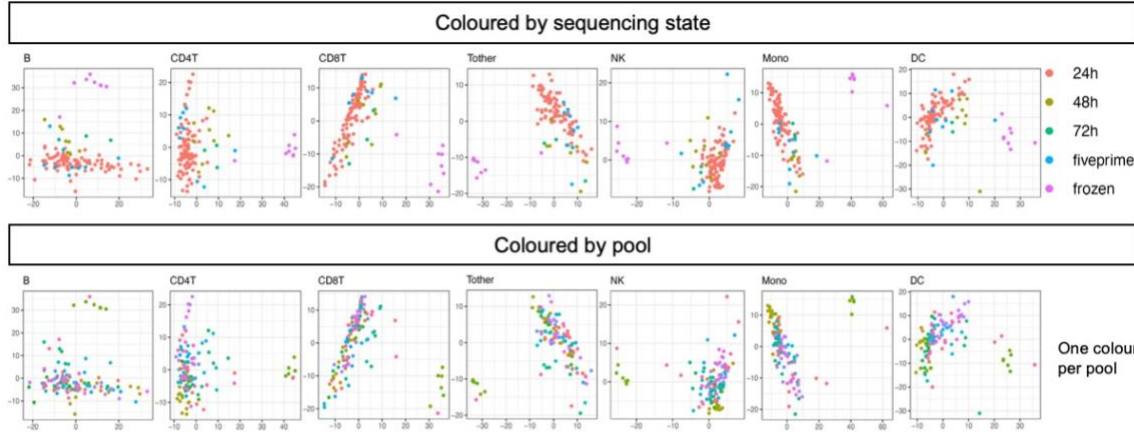
Based on the models from the first step, returning to the samples in this recall, residuals were obtained from regressing out the pool, state, and cohort effects from gene expression. PCA on the residuals demonstrated that pool and state batch effects were no longer driving PC1 and PC2 strongly, though the batch correction appeared to be more effective in some cell-types (lymphocytes) than others (monocytes). Instead, the residual PCs were now driven by biological factors such as sex and age, reflecting the importance of considering these factors as covariates in subsequent differential expression analyses. (Figure-3.10C)

A

PCA Run 1

**B**

PCA Run 2

**C**

PCA Run 3

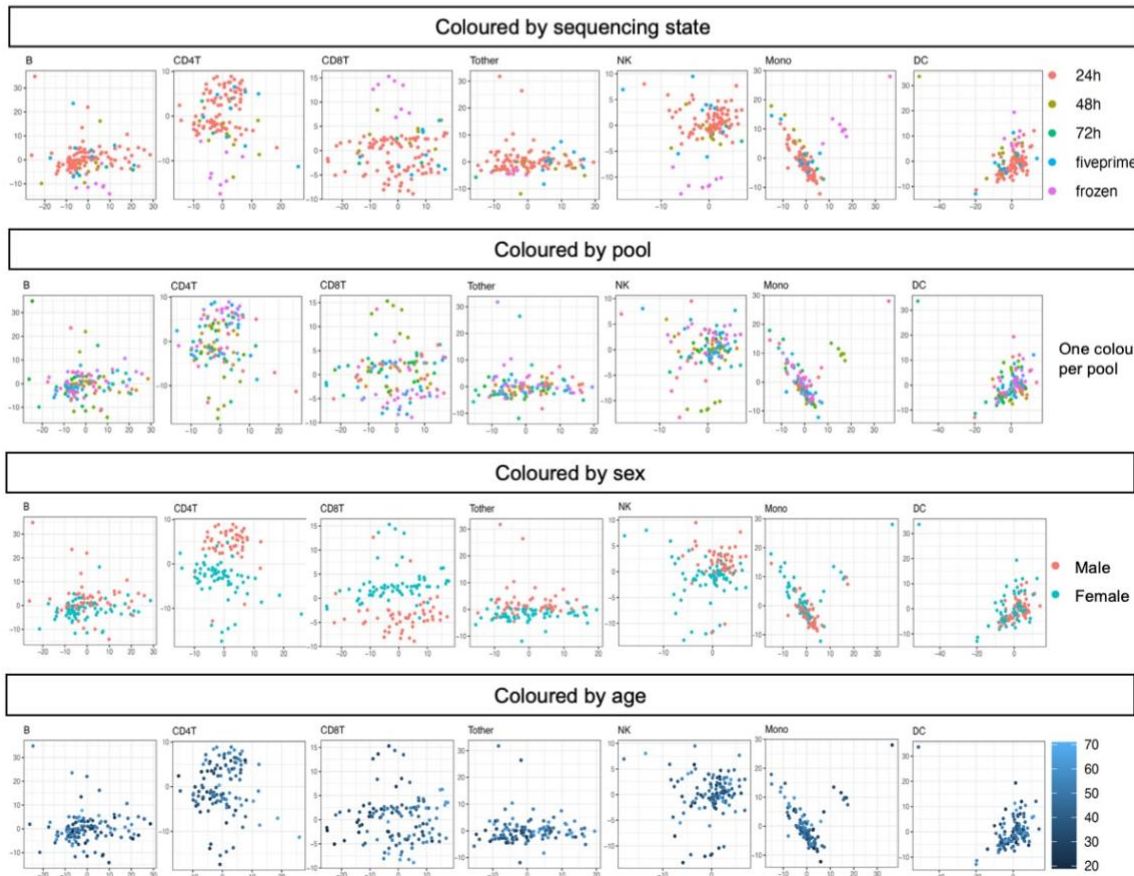


Figure-3.10

PCAs of pseudobulked expression. A: Before applying sample-level QC. **B:** After applying sample-level QC. **C** - After performing technical batch effect correction.

As further check that the technical batch effect correction performed in step 1 of the model was effective, we studied the correlation of gene expression between samples. Twenty-one donors had duplicated sequencing across different pools, and could be used to evaluate the technical batch effect correction. For each sample (defined as a donor in a pool), the gene expressions (with and without correction) were pseudobulked across azimuth level 1 cell-types, and a correlation r^2 of expression values was calculated between pairs of samples. The correlation between different donors in the same pool remained the same before and after correction, and the correlation between same donors in different pools improved after correction. Both of these findings were reassuring. (Figure-3.11)

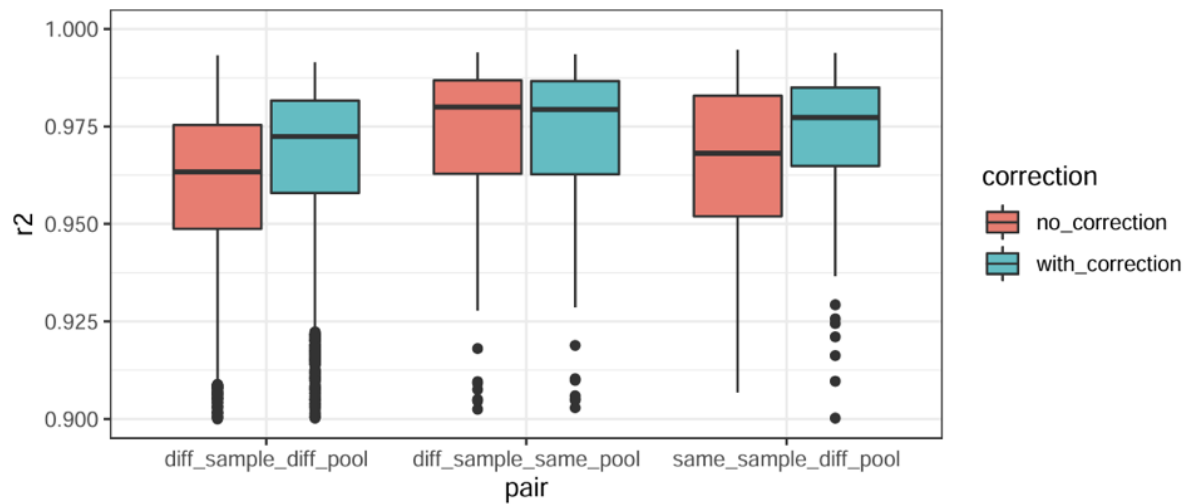


Figure-3.11

Correlation r-squared of gene expression between samples. Each data point contributing to the plotted distributions represents the r^2 calculated from correlating the gene expressions pseudobulked by one cell-type, and between a pair of samples.

Analysing the direct knockout effects of pLoFs at the RNA level

Twenty-two pLoFs remained after QC and correction for technical batch effects. We then ran linear models to assess the association between gene expression and genotype at each of the pLoFs, controlling for age and sex. For each run, only samples from the control group were used to represent the wild-type, to avoid confounding from other pLoFs in related immune genes. While these tests were run on all genes, we first investigated if the direct knockdown effect of pLoFs could be detectable on the gene they were predicted to affect. For this, the additive model was selected as it would most likely capture the variant effect regardless of the underlying genetic architecture (Hill et al., 2008).

Direct knockdown effects revealed cell-type-specificity

We performed 154 additive tests (22 pLoFs, 7 cell-types), and saw that 39 tests demonstrated reduced expression of the pLoF-affected gene in carriers compared to controls at nominal significance ($p \leq 0.05$), with 19 passing a Benjamini–Hochberg false discovery threshold of 5% when correcting for multiple-testing (FDR 5%). These results are tabulated in Supplementary-table-3.2 and visualised in Figure-3.12.

We found that pLoF effects demonstrated cell-type-specificity. The 19 FDR 5% significant knockdown effects came from 9 pLoFs, implying that for these pLoFs not all cell-types demonstrated significant knockdown effects (i.e. the mean was 2.1 cell-types per pLoF). This might be because the genes are barely expressed in other cell-types. For example *MNDA* is a gene known to have myeloid-specific expression (Goldberger et al., 1986), and indeed we saw in our data that the pLoF affecting *MNDA* only demonstrated statistically significant (FDR 5%) knockdown effects on monocytes and dendritic cells (Supplementary-table-3.2). This served as a motivation for further interrogation of cell-type-specific pLoF effects in Chapter 4. However, as a first step, we wanted to quantify the number of pLoFs that demonstrated a statistically significant direct knockdown effect, in an unbiased manner that did not require prior knowledge of a gene's cell-type-specificity.

Focusing on the cell-type where the target gene was most highly expressed

Therefore, to fairly and systematically determine whether a pLoF demonstrated a statistically significant effect, we focused on the cell-type where the target gene was most highly expressed in control individuals. Next, we required the pLoF genotype to be nominally significantly ($p \leq 0.05$) associated with reduced expression under an additive model.

Under these criteria, 8 of the 22 tested pLoFs (affecting *CAT*, *CST7*, *DDX58*, *IFIH1*, *MNDA*, *SIGLEC5*, *TMEM71*, *TRAT1*) demonstrated a statistically significant knockdown effect, and

another 12 (affecting *CD109* (two pLoFs), *DDX58*, *HPSE*, *MYO1G*, *NOD2*, *SIGLEC14* (two pLoFs), *SLC9A3R1*, *TRIM58*, *ZBP1*, *ZC3H8*) demonstrated a negative direction of effect that was not significant ($p>0.05$). These results included some pLoFs with fewer than three heterozygous or homozygous samples, therefore suffering from limited power. The remaining two pLoFs (affecting *CDC25B* and *TMEM71*) demonstrated no knockdown effect despite having more than three samples with the pLoF.

In an attempt to increase power, for four genes with multiple pLoFs, pLoFs in the same gene were grouped together and similar tests were performed. Of these, only *DDX58* had a statistically significant knockdown effect when tested at the gene level, though the association was not as strong (in beta and p-value) as one of the pLoFs, chr9_32488757_GA_G, tested alone. Effects of different pLoFs in the same gene are not necessarily likely to be equal, and should some of these pLoFs not demonstrate true effects, adding them to a gene-based test would add noise to the analysis instead of power. This is revisited later incorporating the results of the pLoF curation.

These results are summarised in Table-3.3 and visualised in Figure-3.12.

Gene	pLoF	N Hom	N Het	Highest Expressing PBMC	Group	Additive Beta	Additive p-value	Dominant Beta	Dominant p-value
CAT	chr11_34456818_G_C	0	2	Mono	pLoF	-4.5E-01	1.9E-05	-4.5E-01	1.9E-05
CD109		4	8	Mono	Gene	-7.5E-02	2.1E-01	-6.4E-02	4.7E-01
CD109	chr6_73736509_G_A	2	6	Mono	pLoF	-1.2E-01	8.2E-02	-1.2E-01	1.9E-01
CD109	chr6_73782709_C_T	1	0	Mono	pLoF	-4.7E-02	4.1E-01	-9.5E-02	4.1E-01
CDC25B	chr20_3787140_G_GT	1	5	CD4T	pLoF	7.9E-02	2.1E-01	8.8E-02	2.6E-01
CST7	chr20_24957358_G_T	0	3	NK	pLoF	-9.4E-01	9.1E-08	-9.4E-01	9.1E-08
DDX58		2	10	Mono	Gene	-3.2E-01	1.3E-03	-3.2E-01	9.8E-03
DDX58	chr9_32466355_C_A	0	7	Mono	pLoF	-1.5E-01	2.5E-01	-1.5E-01	2.5E-01
DDX58	chr9_32488757_GA_G	2	3	Mono	pLoF	-4.3E-01	8.8E-04	-6.1E-01	1.0E-03
HPSE	chr4_83302158_G_GT	1	3	Mono	pLoF	-1.4E-01	7.8E-02	-1.2E-01	2.4E-01
IFIH1	chr2_162277515_AC_A	1	4	Mono	pLoF	-2.7E-01	3.2E-02	-3.0E-01	6.4E-02
MNDA	chr1_158845998_C_T	1	12	Mono	pLoF	-7.8E-01	8.4E-11	-7.4E-01	4.0E-06
MYO1G	chr7_44964942_TAC_T	0	1	Mono	pLoF	-1.2E-01	3.0E-01	-1.2E-01	3.0E-01
NOD2	chr16_50716612_C_T	0	3	Mono	pLoF	-1.7E-01	5.1E-02	-1.7E-01	5.1E-02
SIGLEC14		2	4	Mono	Gene	-5.4E-02	4.2E-01	-1.1E-01	4.2E-01
SIGLEC14	chr19_51645475_A_G	1	4	Mono	pLoF	-2.4E-02	7.4E-01	-4.8E-02	7.4E-01
SIGLEC14	chr19_51643777_ACT_A	1	0	Mono	pLoF	-2.0E-01	1.7E-01	-3.9E-01	1.7E-01
SIGLEC5	chr19_51628003_A_AG	5	0	Mono	pLoF	-2.8E-01	2.6E-03	-3.1E-01	1.3E-02
SLC9A3R1	chr17_74749171_G_T	1	4	CD4T	pLoF	-9.3E-02	1.7E-01	-8.0E-02	3.5E-01
TMEM71		2	4	CD4T	Gene	1.3E-02	7.5E-01	-4.1E-02	5.1E-01
TMEM71	chr8_132722068_C_CAGCA	0	3	CD4T	pLoF	-1.9E-01	1.5E-02	-1.9E-01	1.5E-02
TMEM71	chr8_132758873_G_A	2	1	CD4T	pLoF	5.8E-02	1.4E-01	7.9E-02	2.6E-01
TRAT1	chr3_108849184_GCTATGAA_G	1	6	CD4T	pLoF	-2.6E-01	7.6E-05	-3.3E-01	7.2E-05
TRIM58	chr1_247857547_G_T	0	5	Mono	pLoF	-9.6E-02	5.9E-01	-9.6E-02	5.9E-01
ZBP1	chr20_57615536_CA_C	0	1	CD4T	pLoF	-2.5E-01	1.3E-01	-2.5E-01	1.3E-01
ZC3H8	chr2_112230930_CTTCTT_C	1	1	CD4T	pLoF	-2.3E-02	5.9E-01	-3.5E-02	6.2E-01

Table-3.3

Table of pLoF effects on corrected gene expression in the cell-type where the gene is most highly-expressed, tested at different levels (by gene - if no variant is listed - or by pLoF variant), and with different models of association (additive and dominant).

The additive betas tested by pLoFs have been plotted in Figure-3.12.

Relating gene expression results to manual pLoF curation results

Recalled pLoFs were curated via the CPG LoF Curation SOP V2 workflow, which looked for technical and impact factors that might cause the pLoF to be less likely to result in true loss-of-function. Of the 24 pLoFs recalled, 10 were assigned as “LoF”s, 5 as “Likely_LoF”s, 5 “Uncertain”s, 3 “Likely_Not_LoF”s, and 1 “Not_LoF”.

We compared these curation results to the results from the analysis above of whether the pLoF was associated with reduced RNA expression of the affected gene. We suspected that there was a relationship between the curation verdict and the effect of the variant on gene expression. Firstly, the nine pLoFs that demonstrated FDR 5% significant knockdown effects when tested across different cell-types all had “LoF” and “Likely_LoF” verdicts. Secondly, we see from Figure-3.12 that for variants where the verdict returned as “LoF”s or “Likely_LoF”s, in the cell-type where the affected gene had the highest expression, more of these variants were associated with a negative beta, at a larger magnitude, and more likelihood of being statistically significant.

To statistically test this observation, taking into account both the patterns in the effect size and the p-value, we used the test statistic (“t statistic”). We grouped the “LoF”s and “Likely_LoF”s together, and the “Uncertain”s, “Likely_Not_LoF”s and “Not_LoF”s variants formed the other group. Across all 154 tests performed, the t statistics of the “LoF”s/“Likely_LoF”s were significantly lower than the t statistics from the “Uncertain”s/“Likely_Not_LoF”s/“Not_LoF”s (Wilcoxon one-sided p-value = 2.3×10^{-4}). It is worth noting that when restricting to the test from the highest expressing cell-type, this was not significant (p-value 0.16), however, with only 22 tests in this subset, we may be underpowered.

Returning to the gene-based tests (Table-3.3), both of *DDX58*’s pLoFs have been curated as “LoF”s, explaining how collapsing still demonstrated a knockout effect. For *SIGLEC14*, both pLoFs were rates as “uncertain”, for *CD109* two were “LoF”s (but one of each had samples that all failed QC) and one was “uncertain”, and lastly, for *TMEM71*, one was assigned as “Likely_LoF” and the other “Likely_Not_LoF”. All of these further explained why collapsing these variants for analysis at the gene level did not improve power.

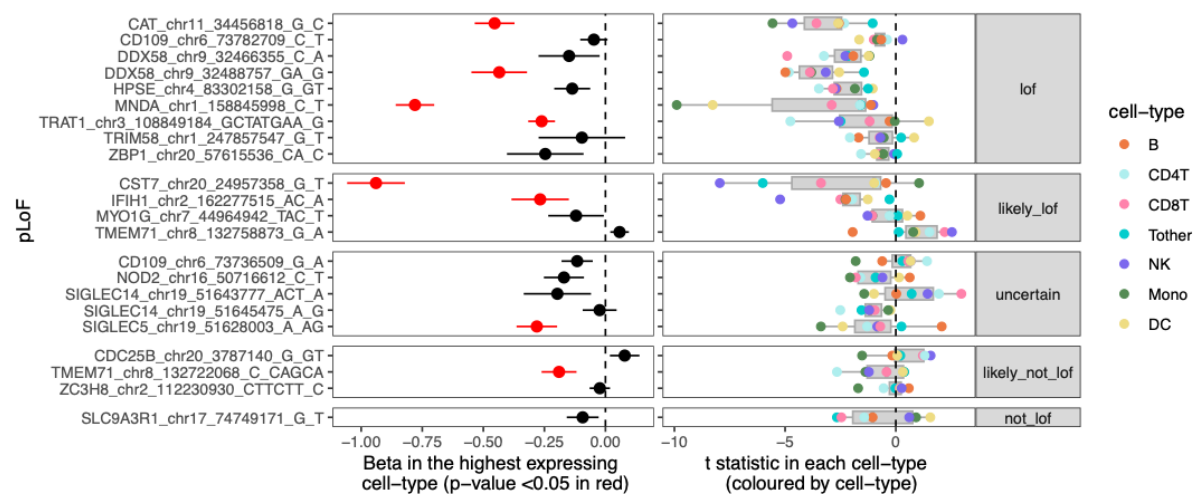


Figure-3.12

Relationship between the curation verdict and additive testing for direct pLoF

knockdown effects. Left: Effect sizes (betas) of the genotype (additive encoding) on the corrected gene expression in the highest expressing cell-type for each variant. Error bars reflect standard errors. Associations that had a p-value ≤ 0.05 are coloured in red. **Right:** test statistic (t statistic) of the genotype on the corrected gene expression in each cell-type (data points have been coloured by cell-type). Box plots have been plotted to reflect the underlying distribution across cell-types.

Evaluating non-additive pLoF effects

Testing for direct knockdown effects on the genes affected by pLoFs was also performed with non-additive models. Again, for each pLoF we focused on the tests in the cell-type with the highest expression of the affected gene.

Dominant pLoF effects

Only one pLoF, chr3_108849184_GCTATGAA_G in *TRAT1*, demonstrated a more negative beta and more significant p-value with the dominant model than the additive model (additive beta -0.26, additive p-value 7.6×10^{-5} , dominant beta -0.33, dominant p-value 7.2×10^{-5}).

Recessive pLoF effects

Most pLoFs had only one homozygous sample, and therefore did not have the sample size for formal testing of a recessive model using the pseudobulked expression values. Thus, we tackled this with a couple of different approaches leveraging a larger sample size from the Cardinal cohort.

First, since there were many more G&H donors in Cardinal, a cohort distribution of gene expression was estimated from the 303 G&H samples used to build step 1 of our model, and z-scores could be calculated for each individual. This provided a visual and quantifiable representation of the degree to which pLoF expression differed from the wild-type, and allowed us to infer if other modes of inheritance could be at play.

Next, we also hoped that multiple measurements of gene expression at the single-cell level could add power for recessive testing in cases where there was one or just a few homozygous individuals. We therefore ran a linear mixed model at the single-cell level, regressing the normalised gene expression in cells of the highest expressing cell-type against genotype, accounting for donor-level random effects. We acknowledge that for tests with only one homozygous individual, we still might not be able to differentiate fully between LoF effects and donor effects. Comparing between a recessive and additive encoding was only sensible for pLoFs for which both heterozygous and homozygous individuals were available. The results are shown in Supplementary-table-3.3.

Three pLoFs, one with a “LoF” verdict affecting *HPSE*, one with an “Uncertain” verdict affecting *SIGLEC5*, and one with a “Not_LoF” verdict affecting *SLC9A3R1*, had a more negative beta and smaller p-value in the recessive model than the additive model. However, none had a significant recessive p-value, although the pLoF affecting *SLC9A3R1* had a nominally significant dominance deviation p-value. Figure-3.13 shows the pseudobulked gene expression values of these pLoF individuals compared to the normal distribution in the

population. Heterozygous individuals had a more similar distribution of gene expression to the wild-type cohort, while homozygous individuals deviate drastically, with two of the pLoFs (chr4_83302158_G_GT in *HPSE* and chr19_51628003_A_AG in *SIGLEC5*) having homozygous individuals with gene expression levels more than two standard deviations away from the mean, i.e. these individuals have gene expression values for their pLoF-affected gene that are within the bottom two percent of the population.

Overall, we did not detect strong evidence for non-additivity in the knockout effects of pLoFs at the level of RNA expression. We suspect we may still be underpowered especially for recessive effects, given that for most pLoFs only one homozygous individual was recalled. Furthermore, two of the three pLoFs reported here to have potential recessive effects have been given curation verdicts of “Uncertain” and “Not_LoF”, suggesting that they may not result in true loss-of-function in the first place.

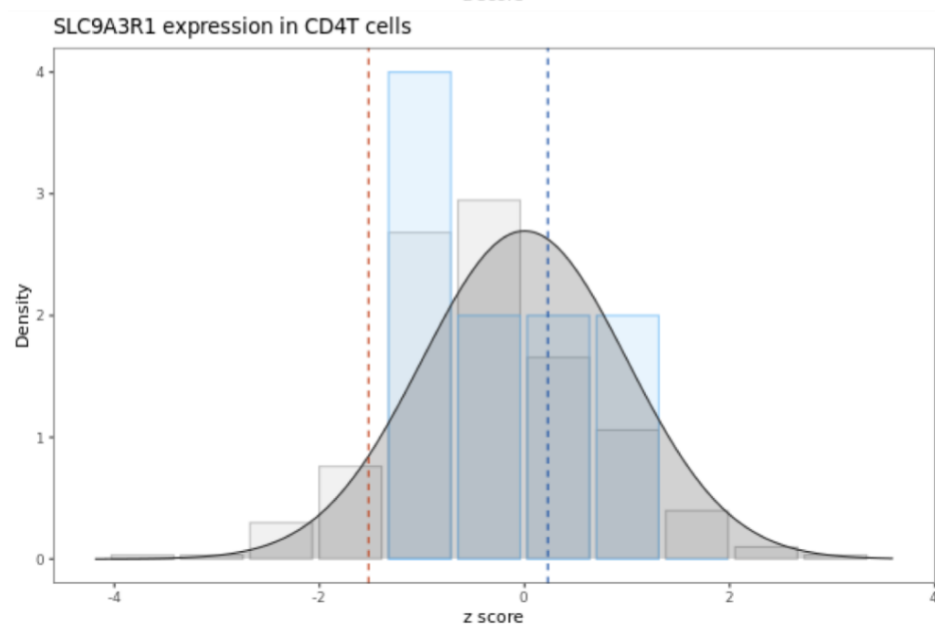
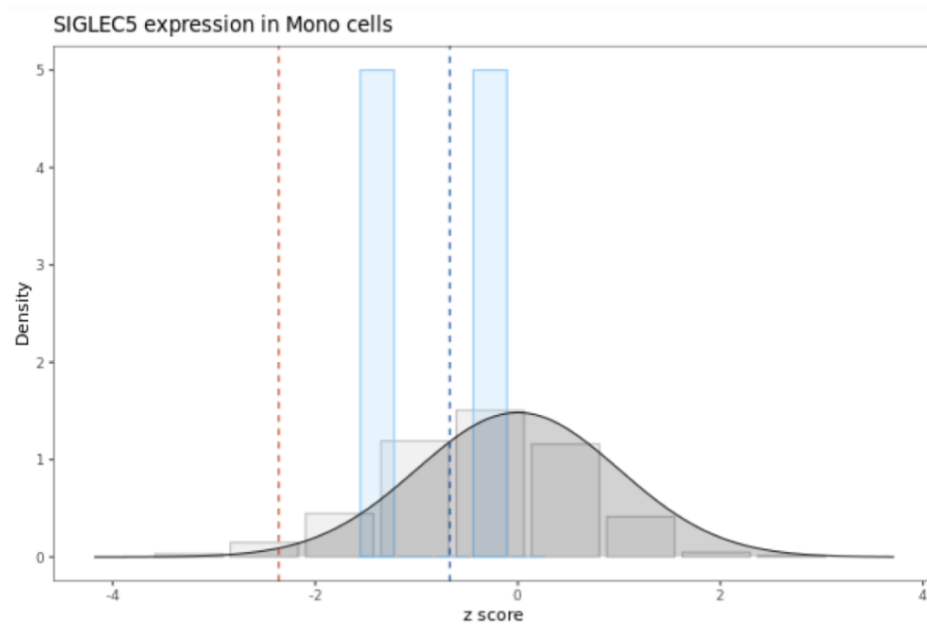
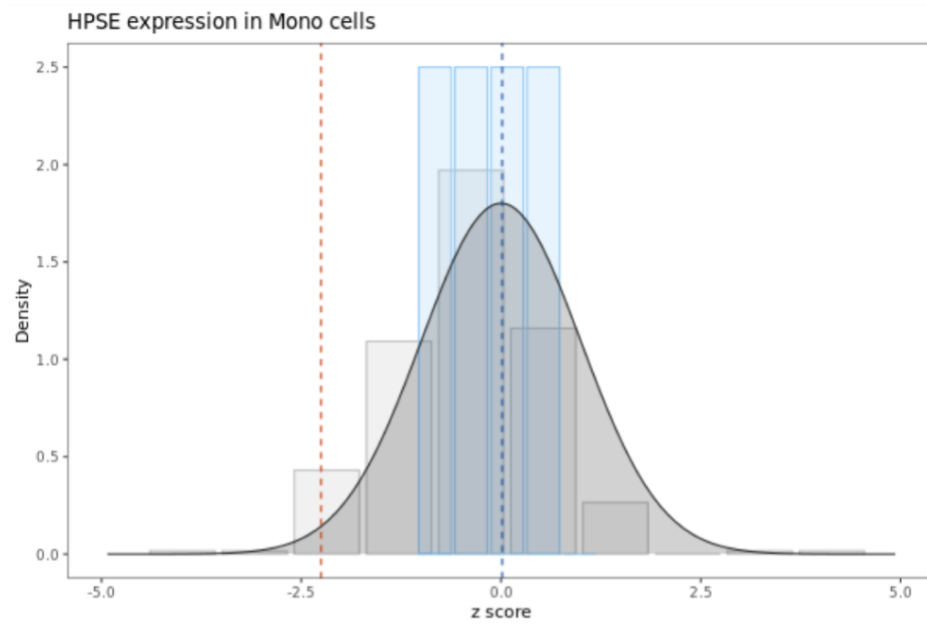


Figure-3.13

Distribution of the pseudobulked gene expression (corrected for technical effects) in the indicated cell-type, in pLoF individuals compared to other G&H individuals. Gray - G&H cohort, blue - heterozygous individuals, red - homozygous individuals. When there are multiple individuals, a histogram is plotted, with the vertical line representing the mean. When there is one individual, it is represented by a vertical line.

Indirect pLoF effects at the RNA and phenotypic level

As we found above that pLoFs curated as “LoF”s or “Likely_LoF”s were more likely to be associated with detectable direct knockdown effects on the affected gene, we decided to restrict our analysis of indirect pLoF effects to variants assigned with these two verdicts, in order to have confidence in the accuracy of our results. Of the 24 pLoFs recalled, 15 were assigned as “LoF” and “Likely_LoF” after further curation, with 10 having at least 3 donors recalled. These 10 were carried over to the deeper scRNAseq project and discussed in the next chapter. Four out of the remaining five pLoFs had only one carrier each and therefore were too underpowered to make conclusions about. The remaining pLoF is chr11_34456818_G_C in the *CAT* gene, the example that we have been using throughout this chapter, from power calculations to LoF effects analysis. It was not taken for deeper scRNAseq owing to its small sample size (two heterozygotes), however, we were still able to gain some insights on this pLoF with the current analysis. We complete the discussion for this pLoF here.

chr11_34456818_G_C (rs571047357) is a splice donor variant in catalase (*CAT*), a gene that encodes for a protein of the same name that catalyses the breakdown of hydrogen peroxide to water and oxygen, protecting cells from oxidative stress (Chelikani et al., 2004). There is some evidence that catalase has a role in the innate immune response due to its function in managing reactive oxygen species (Gayashani Sandamalika et al., 2021). Oxidative damage is a component of the pathophysiology of several autoimmune conditions such as systemic lupus erythematosus, asthma and rheumatoid arthritis (Smallwood et al., 2018). There is evidence that the gene is also involved in growth promotion of several immune cell-types, specifically T cells, B cells and myeloid leukaemia cells (Takeuchi et al., 1995). Our variant of interest has a single submission in ClinVar (Landrum et al., 2014) (Accession: SCV003822578.2, accessed on 27 May 2025), and it has been annotated to be “likely pathogenic” for acatalasia. Acatalasia is an autosomal recessive disease (OMIM entry 614097) characterised by low levels of catalase resulting in physical manifestations of oxidative stress such as ulcers and gangrene (Ogata, 1991; Szeinberg et al., 1963).

The gene was prioritised for recall (Figure-3.1) because it had multiple hom-pLoF individuals in G&H, and L2G associations (>0.5) with blood protein levels. It is highly expressed in whole blood, particularly erythrocytes, and within PBMCs, it has high expression in myeloid cells. The gene has a low pLI (gnomAD v2) of 1.6×10^{-10} but a high LOEUF (gnomAD v4.1.0) of 0.9 (S. Chen et al., 2022).

The pLoF is very rare, though enriched in G&H. The G&H AF of 0.1% in the WES5K cohort was higher than multiple publicly reported AFs. In the thousand genome project, this variant had an allele count of 1 (South Asian ancestry) (1000 Genomes Project Consortium et al., 2015), and in gnomAD v4.1.0, it has an AF of 3×10^{-5} , again with South Asian enrichment - Out of an allele count of 49, 42 are from the South Asian subset, and the only homozygous individual in the whole 807,162 cohort was South Asian (Comparing the AF in G&H and the AF in non South Asians in gnomad, the Fisher's Exact Test p-value $< 2.2 \times 10^{-16}$) (S. Chen et al., 2022). In contrast, in our G&H recall pool, there were two homozygous and four heterozygous individuals that passed QC to be included. This further demonstrates the strength of conducting this study in the G&H cohort. Unfortunately we were only able to recruit two heterozygous individuals.

The CPG LoF Curation SOP V2 workflow annotated the pLoF as a "LoF" with nothing flagged. Indeed, even with just two heterozygous individuals, we found a statistically significant association between *CAT* expression in monocytes, the highest expressing cell-type, and pLoF genotype (Table-3.3).

This additive model was run on the rest of the genes expressed in monocytes to identify differentially expressed genes associated with this *CAT* pLoF genotype, and genes with significant differential expression ($\log_2$ fold change ± 0.01 at p-value ≤ 0.05) were tested for over-representation in Reactome pathways. Sixteen pathways were significantly downregulated in *CAT* pLoF individuals, defined as pathways enriched with genes that had decreased expression in these individuals compared to controls (Figure-3.14A), and one pathway was upregulated (Figure-3.14B). Many of the pathways fell into the metabolic category, particularly related to glycogen metabolism (R-HSA-70221, R-HSA-8982491), and other pathways were related to the immune system, particularly interactions with the human immunodeficiency virus (R-HSA-167172, R-HSA-167161) and sensing of pathogenic DNA to trigger an interferon response (R-HSA-1834949).

Catalase deficiency has been associated with diabetes mellitus, particularly a subtype called type 3c diabetes mellitus which is a result of pancreatic inflammation. It has been proposed that the pancreatic inflammation is triggered by excess hydrogen peroxide (when there is insufficient catalase) resulting in dysregulation of glycogen metabolism (Nandi et al., 2019). This may provide a possible explanation for the associated pathways shown here.

The only statistically significant upregulated pathway is NOSTRIN mediated eNOS trafficking (R-HSA-203641). Endothelial nitric oxide synthase (eNOS) is the key source of nitric oxide in the cardiovascular system, and upregulation of this pathway decreases eNOS activity

(Icking et al., 2005; Zimmermann et al., 2002). Nitric oxide, involved in multiple signalling pathways such as macrophage-induced cytostasis, has been shown to be a competitive inhibitor of catalase (Purwar et al., 2011). It is possible that the upregulation of this pathway may be a compensatory mechanism in *CAT* pLoF individuals; with decreased catalase, there needs to be decreased inhibition from nitric oxide to maintain the same function, and therefore upregulation of the pathway that decreases nitric oxide production.

These are preliminary findings obtained from a small sample size of two heterozygous pLoFs), and we will be better powered to study downstream molecular effects of other pLoF genes in the next chapter.

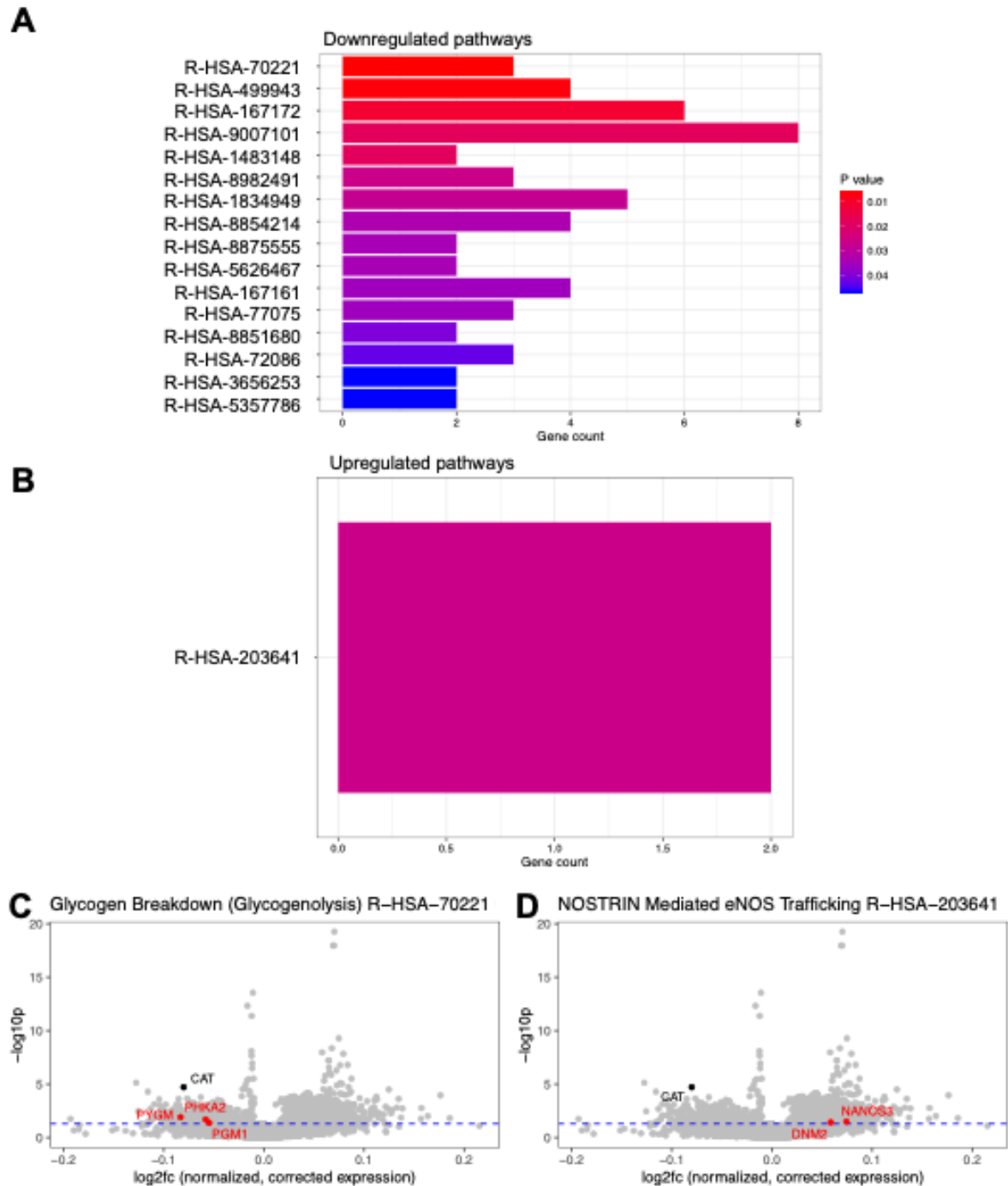


Figure-3.14

Gene set enrichment analysis on the differentially expressed genes between control individuals and individuals with CAT pLoFs. **A:** Pathways enriched for downregulated genes in CAT pLoF individuals. **B:** Pathway enriched for upregulated genes in CAT pLoF individuals. **C&D:** Volcano plots of differentially expressed genes between controls and CAT pLoF individuals, with red-labelled genes being involved in the glycogenolysis pathway (**C**) or the NOSTRIN mediated eNOS trafficking pathway (**D**). The blue dotted line represents nominal significance ($p\text{-value} \leq 0.05$).

Discussion

Summary of findings

The main aims of this chapter were to firstly prioritise individuals with pLoFs in immune genes for recall, and secondly to evaluate that the recalled LoFs demonstrated true knockdown effects on RNA expression.

From the exome sequences of 5,236 Pakistani/Bangladeshi adults in G&H who self-declared as having related parents (supplemented along the way with exomes from the ongoing WES effort of the whole cohort), we prioritised 32 pLoFs in 22 genes across 488 individuals for recall.

Recall took place over one year, with 114 individuals recalled. This comprised 18 control individuals, and 96 individuals homozygous or heterozygous for 24 pLoFs affecting 19 genes. The recalled pLoFs were put through further manual curation, where they were assessed for possible technical and impact errors that might diminish their likelihood of being loss-of-function (i.e. the CPG LoF Curation SOP V2 workflow), and given a curation verdict: ten were assigned as “LoF”s, five “Likely_LoF”s, five “Uncertain”s, three “Likely_Not_LoF”s, and one “Not_LoF”.

After Cardinal scRNAseq and QC, 108 out of the 114 participants were sequenced at a median of 2,043 cells/sample, with 22 pLoFs affecting 18 genes passing QC. From the dimensionality reduction and visualisation of the scRNAseq data, it was determined that technical batch effects, such as the state of the sample during processing (e.g. fresh or frozen) and pool, and biological factors, such as age and sex, were important covariates to control for in differential gene expression analyses.

Linear regressions were performed to identify knockdown effects of pLoFs. Twenty of the pLoFs demonstrated a negative effect on gene expression in the cell-type where the gene is most highly expressed, with eight of those effect sizes being nominally significant. When relating to the LoF curation verdicts, pLoFs curated as “LoF”/“Likely_LoF” had a more negative test statistic than pLoFs curated as “Uncertain”/“Likely_Not_LoF”/“Not_LoF” (Wilcoxon one-sided p-value = 2.3×10^{-4}).

Other models of inheritance were also explored, with one pLoF in *TRAT1* with a suggestive dominant effect on gene expression. The single-cell nature of the data, along with the additional samples in the wider Cardinal project, allowed us to evaluate for possible recessive effects in rare homozygous individuals, and three pLoFs in *SLC9A3R1*, *SIGLEC5*,

and *HPSE* showed possible evidence for a recessive pattern of association, although we were underpowered to call this more definitively.

A novel framework for profiling LoF effects on RNA at the single-cell level and at scale

To our knowledge, this is the first “genotype-led” study of LoF effects on RNA at the single-cell level. Currently, scRNAseq cohorts focus on the creation of cell-type atlases, curating normal variation or profiling disease cohorts. Notable studies that have studied LoF effects at the single-cell level are typically disease-specific studies, working from phenotype to genotype (Donnard et al., 2022; Lam et al., 2021; Nassir et al., 2021). Furthermore, there is a lack of scRNAseq datasets including samples with naturally-occurring knockouts. For example, Donnard et al., 2022 aimed to evaluate the molecular consequences of *Fmr1* LoFs at the single-cell level in the pathophysiology of Fragile X syndrome, and studied the brain transcriptome of *Fmr1* knockout mice to do so (Donnard et al., 2022). In Nassir et al., 2022, they identified a set of genes with LoFs that have been implicated in autism spectrum disorder. They then studied the spatial and single-cell transcriptomic profile of these genes (in the absence of LoFs) in neurotypical brain samples (Nassir et al., 2021). Recall-by-genotype studies of LoFs typically focus on one gene or phenotype profiling, such as a recent study in G&H recalling individuals with LoFs in *MCR3* (Lam et al., 2021). Extensive phenotyping allowed for deep insights into the role of the gene on growth and puberty, and the data was supplemented by external scRNAseq brain data (that do not contain the LoF) (Lam et al., 2021). Therefore, the effects of LoFs on RNA transcription, particularly at the single-cell level, is understudied. This can be challenging because different genes have different involvements in biological processes. Nevertheless, a systematic, genotype-led, approach is required to profile LoFs at scale and prior to this study, has yet to be developed. In these aspects, this study serves as a unique and valuable framework and dataset.

Relating pLoF effects on RNA expression to the curation verdicts

Discrepancies between RNA expression and the curation verdicts

In general, direct knockdown effects of the pLoFs corresponded to their manual curation verdicts. However, there were still discrepancies and there are several possible explanations for this. All the FDR 5% significant direct knockdown effects were associated with variants that had “LoF” and “Likely_LoF” verdicts. The six remaining pLoFs with “LoF” and “Likely_LoF” verdicts but no statistically significant knockdown effects in any cell-type were the two “LoF”s affecting *CD109*, the “Likely_LoF” affecting *MYO1G*, the chr8_132758873_G_A “Likely_LoF” affecting *TMEM71*, the “LoF” affecting *TRIM58*, and the “LoF” affecting *ZBP1*.

With only one heterozygous donor each for the pLoFs affecting *MYO1G* and *ZBP1*, these two were likely underpowered for the analysis. *TRIM58* had low expression in the dataset, potentially also affecting the power to detect a knockout effect. We discuss *TRIM58*'s low expression further in Chapter 4.

Lastly, *CD109* and *TMEM71* were the two largest genes in our recall. The median size of the genes in our dataset is 32,236bp, with the interquartile range from 17,444bp to 47,610bp. At 132,532bp (z-score=3.1) and 75,706bp (z-score=1.2) respectively, it is possible that due to their large sizes, even if nonsense-mediated decay (NMD) was happening for these transcripts, large fragments may still be present and registered as gene expression in our assay. Increased exon length has been shown to be associated with decreased NMD efficiency (Lindeboom et al., 2016).

There are also other reasons why we have not detected knockout effects at the RNA level. It is possible that knockout for some variants happens at the protein expression level instead. RNA transcripts may be present and result in the translation of a truncated protein (MacArthur & Tyler-Smith, 2010), and the protein then proceeds to be degraded by protein quality-control mechanisms (Ciechanover & Schwartz, 1998; MacArthur & Tyler-Smith, 2010). Some of the factors that contribute to whether the knockout happens at the RNA or protein level, such as the percentage of the transcript that is truncated by the variant, has been accounted for in the curation, while others, such as the length of the gene, have not.

Also, just as LoF effects may demonstrate cell-type-specificity, they may also be context-specific, meaning that they only manifest knockdown effects in the presence of specific cellular states, or in situations where the gene expression has changed in the process of the cell carrying out its functions. We discuss this further in Chapter 4.

Lastly, pLoF prioritisation and curation is a manual process and a developing field, and it is possible that many more impact and technical factors that result in pLoFs not causing true loss-of-function remains to be characterised.

Curation flags provide insight into RNA expression patterns and inform next steps

When a pLoF is given an “Uncertain”/“Likely_Not_LoF”/“Not_LoF” verdict, additional curation flags may be annotated to explain this verdict (Figure-3.3). These flags give us additional context to interpret both the verdict and the RNA expression.

For example, chr19_51628003_A_AG in *SIGLEC5*, which demonstrated both a nominally significant knockdown, and a possible recessive pattern to the knockdown, in the highest-expressing cell-type, but not in all cell-types, has been given an “Uncertain” verdict. Firstly, it

is an insertion in a homopolymer region, and homopolymer regions have a higher genotyping error rate. However, small homopolymers such as this (six nucleotides, extending to seven with the insertion) are less error prone. Secondly, this pLoF also has an “uninformative PEXT” flag. However, due to the PEXT data not being updated, the flag has been applied here to indicate that the PEXT score was not available when curating this pLoF. Furthermore, this pLoF is predicted to affect a region of the transcript with weak exon conservation. Together, these flags suggest that with the existing information, it is difficult to ascertain if this pLoF would cause a true effect, hence the “Uncertain” verdict. More detailed (or more updated) information into the transcript isoforms of *SIGLEC5* may allow us to curate this pLoF more accurately.

Lastly, chr17_74749171_G_T affecting *SLC9A3R1* is the only pLoF in our recall given a “Not_LoF” verdict. It has a “rescue” flag raised during curation, as it is part of an in-phase multi-nucleotide variant which in combination would result in a missense change instead of the nonsense change predicted when the variant was considered in isolation. This is unlikely to change mRNA expression, which would explain why we did not detect a statistically significant knockdown effect. Interestingly, the one homozygous individual had a slightly decreased *SLC9A3R1* expression, although with just one individual we are unable to rule out that this result occurred by chance.

Strengths and limitations

In the recall prioritisation

Beyond simply relying on the LOFTEE classification of pLoFs, as many pLoFs were still likely to be spurious, we used characteristics of the genotype, variant, gene, and gene expression to further prioritise individuals for this study. In addition, we ensured that the pLoFs selected affected genes that were relevant to the immune system. However, some of these pLoFs are still likely to be artefacts, and while we have taken steps to filter them out, some of the recalled variants were still given “Uncertain” to “Not_LoF” verdicts based on the CPG LoF Curation SOP V2, where a more detailed analysis was performed to screen for various technical and impact errors. Given good concordance between the curation verdicts and knockdown at the RNA level, future studies could consider including this curation pipeline as part of the prioritisation.

There are several limitations to our prioritisation, both in terms of criteria to select for true LoFs and criteria to select for LoF effects detectable in the immune system. Firstly, we evaluated various metrics that could be used to focus our study on pLoFs with true effects. In terms of evaluating gene constraint, we settled on pLI as a binary indicator to separate

highly-constrained genes from other genes, and only retained pLoFs in genes with $pLI < 0.9$, reasoning that pLoFs in high pLI genes were more likely to be spurious. In future studies, one could consider incorporating LOEUF and S_{het} as well, and evaluate if these metrics correlate with the severity of knockout effects, or allow for the selection of genes that have more important functions. We excluded pRec as a criteria since it did not seem to be as discerning (Supplementary-figure-3.1). In contrast, $p(CRISPR\_gene\_dependency)$, used together with pLI, was too strict. However, it is important to note that in terms of the $p(CRISPR\_gene\_dependency)$ values we derived, we used the median of the probabilities across blood cell lines. This might not capture the variation in knockout effect in blood subtypes. Further dividing the cell lines to higher granularity could result in more accurate $p(CRISPR\_gene\_dependency)$ values for the subset of affected genes determined above to have cell-type-specific expression.

To further narrow down to pLoFs that would be less affected by alternative splicing and would have detectable effects in multiple cell-types, we narrowed our pLoFs to those that were predicted to affect all transcripts of a gene and did not fall into the last intron or exon of a gene. However, prediction in this field has been driven by data from bulk RNA sequencing, and may not reflect cell-type-specific alternative splicing. This may be an explanation for discrepancies between the pLoF curation results and our differential expression results (Figure-3.12).

In terms of identifying LoF effects that would be detectable in the immune system, gene expression was evaluated on datasets of different ancestries to British Pakistani and Bangladeshi individuals, as well as a different age distribution, and this may not accurately reflect the gene expression profiles of these genes in recalled individuals. However, we observed similarities in our expression patterns across multiple datasets (10X, BLUEPRINT (Supplementary-figure-3.4), GTEx, 1M-scBloodNL).

In the recall-by-genotype

Collaborating with Cardinal has allowed us to take advantage of their scale when analysing our pLoF individuals. The central limit theorem suggests that a sample size of at least 30 is typically required to estimate a normal distribution for a population (Kwak & Kim, 2017), therefore the 22 controls of the original dataset would have been insufficient. However, expanding to 303 G&H Cardinal participants allowed us to visualise where individuals stood relative to a larger sample of the population. Therefore, even in the absence of statistical testing, we were able to identify outliers and quantify their deviations.

Throughout the study, despite efforts to enrich for hom-pLoFs and pLoFs (such as working within G&H and keeping the gene list small) the lack of power is still a limitation. In Chapter 4, we performed deeper scRNAseq with more cells, which we hoped would improve the precision of the expression estimates (by reducing intra-individual, inter-cell variance) and therefore improve power.

Next, to improve the power, we have recruited first-degree relatives of the individuals with hom-pLoFs. Due to the complex modelling required to account for relatedness and our small sample size, we have not accounted for relatedness in this analysis. Therefore, the current findings of LoF effects could be confounded by having included related individuals in the analysis. In the future, it could be worthwhile to add steps to account for relatedness, of which some methods have been described in Chapter 2, or increase the sample size such that an unrelated subset of the analysis would be possible.

Technical batch effects were also a noticeable challenge we had to deal with. Cardinal spanned a larger cohort, with a focus on profiling fresh PBMCs. Therefore, as our samples were a subset of Cardinal, and due to their staggered collection, some of our samples ended up being the only sample from our study present in a given pool in Cardinal. In addition, repeat samples were thawed from cryogenically banked samples and there were noticeable batch effects caused by the different states of the samples. We managed to deploy a two-step correction and testing model, along with including other cardinal samples into our study, to counter these technical batch effects. However, as described in the results section (Figure-3.12C), this correction is more effective in some cell-types than others. Also, it is possible that the correction is more effective in some genes than others, or that some genes are less involved in the biology causing these batch effects, and therefore suffer less systematic variation than others.

Future work

While the original focus of the study was on hom-pLoFs, to improve power the recall has now included a majority of heterozygous pLoFs. There is probably value in performing allele-specific expression analysis, to understand the compensatory mechanisms (or lack thereof) involved in maintaining expression of genes with heterozygous LoFs at more than half the expression level of the wild-type. Naturally, power will continue to be a significant issue in this analysis, especially if we also consider the question of variation in expression among the heterozygotes for each pLoF. Furthermore, there may be a need to consider noncoding variation and their impact on the regulation of the expression of allele-specific transcripts.

G&H does not have whole genome sequencing, but this could be done using imputed genotype data instead, since most known cis-eQTLs are common anyway.

The Cardinal dataset also included an additional 1,400 G&H individuals, many of whom have other pLoFs as well. While these pLoFs may firstly, affect genes that may not have a high expression in blood or a specific immune function, and secondly, not be recruited to large enough sample sizes, we might still be able to expand our analysis to them, as long as expressions are sufficiently detectable in PBMCs. This will give us more samples to validate the curation of pLoFs.

Lastly, to add a phenotypic layer to our map of LoF effects, we plan to test whether the G&H individuals carrying these variants are enriched for immune traits in their electronic health records (EHRs). Since G&H uses community-based recruitment which is likely depleted of very sick individuals, there may not be any signal to find here, especially for binary disease traits. However, many of these genes are associated with quantitative blood counts, and therefore analysis with quantitative traits may prove meaningful. At the time of writing this chapter the quantitative traits in G&H have just been curated, and phenotypic analysis is ongoing.

Conclusion

In this chapter, we have demonstrated a proof-of-concept for using scRNAseq to study LoF effects on RNA transcription. We described the steps we have taken to enrich our dataset for variants that seem to truly be affecting gene expression, and showed that the direct knockdown effects of these variants at the RNA level correlated with their manual curation verdicts. Still some pLoFs did not demonstrate the expected knockdown effects, and we suspect that this might be because: (a) the knockdown effects are not detectable at the level of RNA transcripts, (b) these knockdown effects may be context-specific, i.e. they only manifest in particular cellular states, and finally, (c) despite extensive manual curation some of these variants may still be erroneously called and not result in true loss-of-function. We have also shown here some initial evidence of cell-type-specific LoF effects. Single-cell resolution gives us the ability to toggle between cell-type resolutions, and this finding has motivated more detailed profiling of cell-type-specific effects in the next chapter. Together, the findings provide the support for further analyses into this particular set of recalled individuals and variants.

Leveraging the large sample size of the Cardinal study has also allowed us to explore potential non-additive LoF effects, however we would need more power (in the form of more homozygous individuals) to investigate recessive effects more definitively.

Lastly, we have demonstrated some biological consequences of one of these pLoFs, a splice donor variant affecting catalase, and these results serve as the start of gaining biological insights into the functions of the genes we have prioritised. Further work on this will be the focus of the final content chapter of this thesis.

Effects of naturally-occurring immune gene knockouts on the resting and stimulated immune transcriptome

Overview

A knockout experiment provides insight into a gene's function and its role in the wider biological system, particularly when considering the appropriate cell-type and context in which the gene would be acting. As loss-of-function variants (LoF) act like naturally-occurring knockouts, they allow for the observation of gene inactivation consequences if approached similarly. I previously described the recall of 114 individuals from Genes & Health (G&H) with 24 putative LoFs (pLoFs) in 19 genes, demonstrating that direct knockdown effects could be detected in the RNA transcripts of pLoF-affected genes. In this chapter, we performed single-cell RNA sequencing (scRNAseq) aiming to capture higher numbers of peripheral blood mononuclear cells (PBMCs, 3,000 per sample), to increase cell-type resolutions and therefore increase our power to detect LoF effects. We also stimulated 40 donor samples with aCD3 (anti cluster of differentiation 3 antibody) and IFN α (interferon alpha) to study how LoF effects change during the innate and adaptive immune response respectively. After stimulation we transcriptionally profiled cells at two timepoints of activation, early activation (16h) and late activation (64h). We identified cell-type- and context-specific LoF effects at every level of the analysis: Firstly, direct knockdown effects on pLoF-affected genes became more significant across more cell-types in the presence of stimulation. Next, we detected pathway derangements associated with pLoFs, and they too differed across cell-types and stimulation conditions. Lastly, we observed subtle but present cell-type proportion changes, with differential abundances between controls and pLoF individuals becoming convincing in two instances, both upon stimulation. We then systematically described findings for selected pLoFs. A key example highlighted is the compensatory and consequential effects of a *TRAF1* pLoF on PI3-AKT signalling during T cell activation. In summary, we performed a systematic survey of LoF effects on the immune transcriptome, via the generation of a unique dataset, providing insight into the role of these pLoF-affected genes in the immune system, and demonstrating the power of LoFs in interrogating gene function.

Introduction

The role of loss-of-function variants in studying gene function

Studying the phenotypic consequences of gene inactivation is one of the most informative ways to interpret a gene's function (Alberts, 2002). While the approach is feasible *in vitro* in cell cultures or in animal models, it is often ethically wrong to induce gene knockouts in humans. Loss-of-function variants (LoFs) in the general population can be considered as naturally-occurring gene knockouts, allowing for observations of both the direct gene expression changes as a result of the LoFs, as well as the wider *in vivo* consequences of the LoFs on other genes, pathways, and on individuals as a whole (MacArthur et al., 2012). This can serve as an adjunct to artificially-induced gene inactivation experiments, where both phenotypic readouts across multiple cell-types and the organism as a whole (such as the presence or absence of a disease) can be challenging or impossible in humans. Especially when considering a gene as a possible drug knockout target, understanding the wider consequences of gene inactivation, and whether this can be tolerated physiologically, is crucial for inferring toxicity effects (Minikel et al., 2020).

The immune system is dynamic

Immune cells undergo dramatic changes in response to infection or inflammation in order to carry out their roles, and they are regulated by complex networks of receptor signalling and transduction pathways. Therefore, gene knockout experiments in immunology often have to take into account the dynamic response of the immune system, and that gene expression patterns vary across different cell-types and contexts. (Shi et al., 2023)

Immune cells develop from the hematopoietic stem cell, which differentiates into the lymphoid cell lineage and the myeloid cell lineage. The common lymphoid progenitor differentiates into B cells, T cells and natural killer (NK) cells, while the common myeloid progenitor differentiates into monocytes and dendritic cells, which together with the lymphoid cells complete the set of cells found in a peripheral blood mononuclear cells (PBMCs). Other myeloid cells include granulocytes (basophils, eosinophils, neutrophils), erythrocytes (red blood cells) and thrombocytes (platelets). Further subclassifications of these broad cell-types exist, and may be defined by the cluster of differentiation (CD) proteins they selectively express. The hematopoietic lineage is illustrated in Figure-4.1. (Janeway et al., 2001; J. Y. Lee & Hong, 2020)

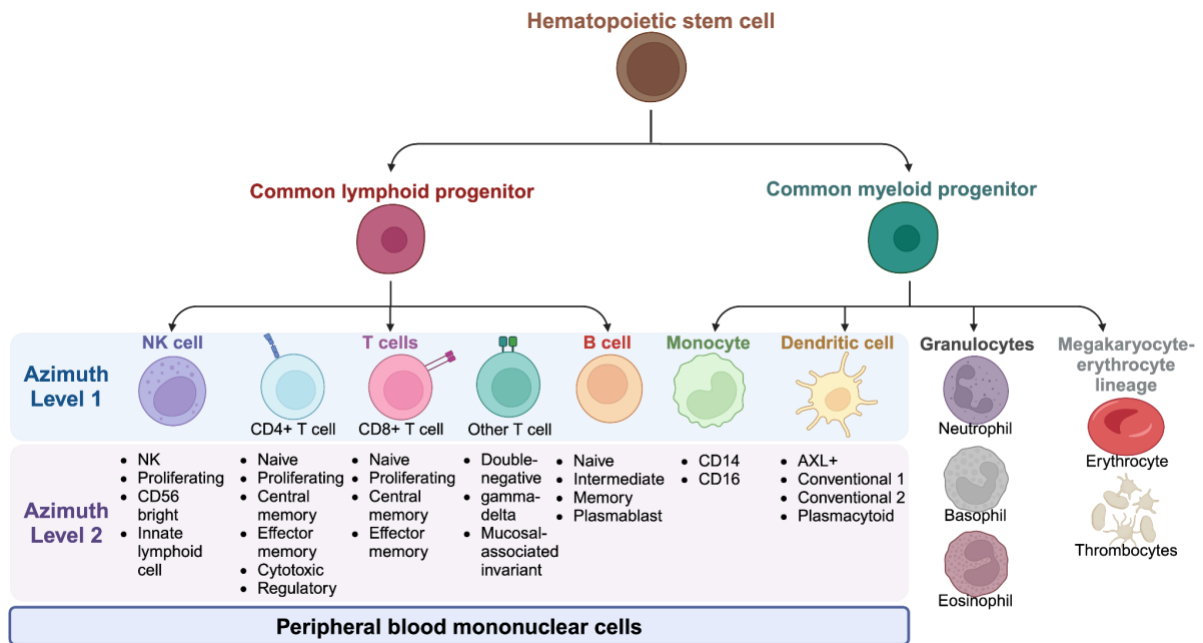


Figure-4.1

The hematopoietic stem cell lineage. The hierarchy of the azimuth level 1 and 2 labels (Hao et al., 2021), which are the cell-type annotations used in this project, is also reflected here. Created with BioRender.com.

These cell-types are present at different proportions within different contexts. In the presence of a challenge (for example, an infection by a pathogen), these cells may be activated, resulting in further differentiation, proliferation, and dramatic changes to their gene expression profiles (Chaplin, 2010). Immune responses are classically divided into two arms: the adaptive immune response and the innate immune response (Medzhitov & Iwasaki, 2024). Generally, B and T cells are involved in the adaptive immune response, with the rest of the PBMCs being involved in the innate immune response, although there is significant overlap between both systems; for example, dendritic cells and macrophages (derived from monocytes) function as antigen presenting cells that activate T cells which then orchestrate the adaptive response, and cells such as NK cells adopt multiple roles across both responses (Marshall et al., 2018).

The dynamic nature of immune cells' response to a challenge motivated the study of these pLoFs at the single-cell level and across various contexts. Single-cell profiling allowed for observations across the dynamic spectrum of cell-types and cell contexts, and deeper sequencing increased our power to capture the effects in rarer cell-types.

Key processes of the immune system

We decided to stimulate the PBMCs for this study with two stimuli that would represent the adaptive and innate responses. Here we introduce several key processes within these systems that are particularly relevant to the genes affected by the pLoFs in the individuals we have recalled.

The innate immune response

One of the several roles of the innate immune system is to identify broad categories of molecules that signify infection or damage, known as “pathogen-associated molecular patterns” (PAMPs) (Janeway, 1989) and danger/damage associated molecular patterns (DAMPs) (Matzinger, 2002). Examples of such molecular patterns include bacterial lipopolysaccharide (Pålsson-McDermott & O'Neill, 2004) and DNA or RNA that are of microbial origin (L. Sun et al., 2013). Different PAMPs are recognised by different proteins, known collectively as “pattern recognition receptors” (PRRs), with notable examples being Toll-like receptors (TLRs) (Medzhitov et al., 1997), NOD-like receptors (NLRs) (Hoffman et al., 2001), and RIG-I-like receptors (RLRs) (Yoneyama et al., 2005). Several of the recalled pLoFs affect PRRs: *DDX58* (also known as *RIG-I*), *IFIH1* (also known as *MDA-5*), and *NOD2*.

When bound to their target molecules, PRRs trigger a response via the production of interferons (IFN). Interferons are cytokines (signalling proteins) named because of their

shared historical role in interfering with viral replication in the host organism (Isaacs & Lindenmann, 1957), however current knowledge suggests that their functions are far-ranging, including implications in allergy, autoimmunity, and activating the adaptive immune response (Londe et al., 2023; Pestka et al., 2004; Schoenborn & Wilson, 2007). Interferons are divided into three classes based on their structure: type I, further divided into several subgroups of which interferon alpha (IFN α) and beta (IFN β) are the most well-described, type II, or interferon gamma (IFN γ), and type III (de Weerd & Nguyen, 2012). We focus on type I interferons here, as they are the main class that is induced by PRRs (Goubau et al., 2013). Type I interferons trigger several signal transduction pathways, a key one being the janus kinase (JAK) – signal transducer and activator of transcription (STAT) pathway (JAK-STAT) (Negishi et al., 2018). The JAK-STAT pathway is a regulator of wide-ranging cellular functions that contribute to inflammation, including apoptosis, cell cycle proliferation and differentiation (X. Hu et al., 2021).

The availability of multiple pLoFs affecting PRRs and genes involved in interferon regulation contributed to the selection of IFN α as a stimulus in this study.

The “effector” cells of the innate immune system are monocytes, dendritic cells, and NK cells. Other cells also involved in this process, such as epithelial cells, platelets and granulocytes, are excluded from this discussion. Monocytes and dendritic cells express PRRs and release cytokines to trigger inflammation upon activation. They are also phagocytic, removing and breaking down the encountered pathogens. Components of the degradation are presented as antigens to B cells and T cells, triggering the adaptive immune response. NK cells also produce immunomodulatory cytokines, and function as cytotoxic cells capable of detecting and removing damaged, stressed or cancerous cells. (Aristizábal & González, 2013; Vivier et al., 2008)

The adaptive immune response

Adaptive immunity provides protection against specific pathogens, and is capable of developing memory, such that a repeat infection from the same pathogen can incite a more efficient and rapid response. Cells involved in the adaptive immune response generally undergo three phases: First, they start as naive cells, populations of cells in circulation, surveying for antigens. Second, in the presence of an antigen, cells are activated and enter the effector phase where they undergo differentiation and clonal expansion to produce effector cells that carry out the immune response. Finally, after the acute challenge, cell populations go through contraction, leaving behind the smaller population of long-lived memory cells for surveillance. (Chi et al., 2024)

The first of the two key cell-types involved are T cells, which are further differentiated into helper T cells (CD4 positive) and cytotoxic T cells (CD8 positive) (Chopp et al., 2023). Antigen presenting cells, typically dendritic cells, present antigens to CD4 and CD8 T cells to begin the process of T cell activation (Eiz-Vesper & Schmetzer, 2020). These antigens are typically fragments of pathogens, degraded during the innate immune response, and usually occur in the presence of additional cytokines to signal inflammation and the need for activation.

Antigen presentation is performed by the major histocompatibility complex (MHC), which is recognised by the T cell antigen receptor complex (TCR). The TCR multimer comprises a component called CD3, and upon binding of the peptide antigen, CD3 triggers the intracellular cascade of signals produced to kickstart various activation processes (Menon et al., 2023). This motivated the selection of anti-CD3 as our second stimuli for this study, to allow analysis of the pLoF effects in the presence of T cell activation. Several of the pLoFs affect genes involved in T cell activation, or have increased expression in activated lymphocytes.

Finally, B cells are the second key lymphocyte that participate in the adaptive immune response. As a collective, they secrete a diverse array of antibodies and are also capable of functioning as antigen presenting cells. (Cooper, 2015)

The pLoF-affected genes and their roles in the immune system

In this chapter, out of the 24 pLoFs recalled, we focused on systematically characterising those that:

1. Were assigned with a “Likely_LoF” or “LoF” curation verdict based on the CPG LoF Curation SOP V2. This is because we wanted to explore indirect LoF effects for which we wanted to have confidence in the accuracy of the pLoF calling and definite knockdown effects on transcription.
2. Had at least three donors, which is the cutoff used for power calculations in the previous chapter. This is to avoid making extensive conclusions based on underpowered analyses.

We shortlisted ten pLoFs affecting nine genes (Table-4.1), with eight pLoFs subjected also to stimulation. Here, I summarise the existing literature regarding these genes, and where relevant, I provide justification for why the pLoFs were selected for stimulation.

Gene	Variant	Curation	N	Selected
<i>CAT</i>	chr11_34456818_G_C	LoF	2	No; Donors *
<i>CD109</i>	chr6_73736509_G_A	Uncertain	8	No; Curation
<i>CD109</i>	chr6_73758966_T_TTG	LoF	3	Yes
<i>CD109</i>	chr6_73782709_C_T	LoF	1	No; Donors
<i>CDC25B</i>	chr20_3787140_G_GT	Likely Not LoF	6	No; Curation
<i>CST7</i>	chr20_24957358_G_T	Likely LoF	3	Yes
<i>DDX58</i>	chr9_32466355_C_A	LoF	7	Yes
<i>DDX58</i>	chr9_32488757_GA_G	LoF	5	Yes
<i>HPSE</i>	chr4_83302158_G_GT	LoF	4	Yes
<i>IFIH1</i>	chr2_162277515_AC_A	Likely LoF	5	Yes
<i>MNDA</i>	chr1_158845998_C_T	LoF	13	Yes
<i>MYO1F</i>	chr19_8522497_G_A	Likely LoF	1	No; Donors
<i>MYO1G</i>	chr7_44964942_TAC_T	Likely LoF	1	No; Donors
<i>NOD2</i>	chr16_50716612_C_T	Uncertain	3	No; Curation
<i>SIGLEC14</i>	chr19_51645475_A_G	Uncertain	5	No; Curation
<i>SIGLEC14</i>	chr19_51643777_ACT_A	Uncertain	1	No; Curation
<i>SIGLEC5</i>	chr19_51628003_A_AG	Uncertain	5	No; Curation
<i>SLC9A3R1</i>	chr17_74749171_G_T	Not LoF	5	No; Curation
<i>TMEM71</i>	chr8_132722068_C_CAGCA	Likely Not LoF	3	No; Curation
<i>TMEM71</i>	chr8_132758873_G_A	Likely LoF	3	Yes
<i>TRAT1</i>	chr3_108849184_GCTATGAA_G	LoF	7	Yes
<i>TRIM58</i>	chr1_247857547_G_T	LoF	5	Yes
<i>ZBP1</i>	chr20_57615536_CA_C	LoF	1	No; Donors
<i>ZC3H8</i>	chr2_112230930_CTTCTT_C	Likely Not LoF	2	No; Donors; Curation

Table-4.1

The 24 pLoFs that have been recalled as described in the previous chapter, and the selection status of the gene for extended profiling in this chapter. In the “Selected” column, “Donors” refers to excluding pLoFs with less than three donors, and “Curation” refers to excluding pLoFs that have been annotated as “Uncertain”, “Likely_Not_LoF” and “Not_LoF”.

* This pLoF affecting *CAT* is the only one excluded that has more than one donor and has a “LoF”/“Likely_LoF” annotation. It has been extensively described in the previous chapter.

DDX58 is a PRR that senses viral RNA

DEAD (Asp-Glu-Ala-Asp) box polypeptide 58 (*DDX58*) is also known as retinoic acid-inducible gene I (*RIG-I*)* (Seal et al., 2023). It is a PRR predominantly in the cytoplasm that is activated by immunostimulatory RNA, specifically, certain single-stranded and short double-stranded RNA usually of viral origin, and is involved in the defence against multiple classes of RNA viruses, but also some DNA viruses (Kell & Gale, 2015). Upon binding with its PAMP, it activates a signalling cascade known as the RIG-I-like receptor signalling pathway that leads to the production of type-I interferons (Rehwinkel & Gack, 2020). *DDX58* is lowly expressed across many tissues and cell-types, and upregulated in the presence of type-I interferons, possibly as part of a positive feedback loop to amplify the immune response (Kell & Gale, 2015; Thoresen et al., 2023). This makes it an interesting gene to study under IFN α stimulation: increased expression could boost power to detect pLoF expression, and using IFN α is a safer approach to elucidate the downstream *DDX58* knockout consequences than a viral challenge by.

**DDX58* was the HGNC approved symbol as of 19 April 2021, when we performed our pLoF and gene selection for recall prioritisation (Tweedie et al., 2021). It has since been updated to *RIG-I* (Seal et al., 2023). However, for consistency, we have used *DDX58* throughout this thesis.

IFIH1 is a PRR that senses long double-stranded RNA

Interferon induced with helicase C domain 1 (*IFIH1*) is also known as melanoma differentiation-associated gene 5 (*MDA-5*) (Seal et al., 2023), another PRR (specifically, a RLR) that detects immunostimulatory RNA and triggers the RLR signalling pathway (Rehwinkel & Gack, 2020). It has a similar role to *DDX58*, but in contrast it senses long double-stranded RNA (Kato et al., 2008; Peisley et al., 2011), again usually of viral origin. It is also upregulated in response to IFN α stimulation (Aparici-Herraiz et al., 2023).

MNDA is a PYHIN protein involved in IFN regulation, transcription, and apoptosis

Myeloid nuclear differentiation antigen (*MNDA*) was first discovered to be specific to the nuclei of myeloid cells (Goldberger et al., 1986), however subsequent datasets showed additional expression in B cell subsets (Miranda et al., 1999). It is part of the *PYHIN* family of proteins, characterised by the presence of Pyrin and HIN domains, and two *PYHIN* proteins, *AIM2* and *IFI16*, have been reported to be DNA-sensing PRRs (Connolly & Bowie, 2014). However, *MNDA* has not been shown to be involved in PAMP/DAMP sensing, but rather it works downstream by regulating the type-I interferon cascade (Gu et al., 2022). This makes

samples with pLoF affecting *MNDA* an attractive candidate for IFN α stimulation, and its expression also increases after IFN α stimulation (J. A. Briggs et al., 1992).

MNDA has also been reported to have probable roles in regulating transcription by binding to nucleolin and the transcription factor YY1, which may have potential implications on cell growth and differentiation (Johnstone & Trapani, 1999; Xie et al., 1995, 1998). In addition, it has been suggested that it inhibits apoptosis, and expression levels have been considered as a marker for several haematological malignancies, including B-cell lymphoma, myelodysplastic syndromes and acute myeloid leukaemia (i.e. the malignancies in the two cell-types that express *MNDA*) (R. C. Briggs et al., 2006; Fotouhi-Ardakani et al., 2010; Gupta et al., 2023; Xie et al., 1997).

TRAT1 is involved in T cell activation

One of the key signalling pathways involved in T cell activation is the phosphoinositide-3-kinase–protein kinase B/Akt (PI3K-AKT) signalling pathway, which regulates multiple aspects of the immune response including glucose metabolism, proliferation and apoptosis (Hemmings & Restuccia, 2012; Koyasu, 2003). T cell receptor associated transmembrane adaptor 1 (*TRAT1*), historically known as the T-cell receptor-interacting molecule (*TRIM*), induces the PI3K-AKT pathway by binding to the 85kd regulatory subunit of PI3K upon activation of the TCR (Bruyns et al., 1998; Hübener et al., 2000; Ward & Cantrell, 2001). These reports motivate studying the pLoFs affecting *TRAT1* in aCD3 stimulation.

HPSE cleaves heparan sulfate, promoting infiltration of activated T cells

Heparanase (*HPSE*) is the enzyme that cleaves heparan sulfate, which is a proteoglycan that forms the extracellular matrix in a wide range of tissues and cell-types (Vlodavsky et al., 1999). Heparan sulfate cleaving leads to modification of the extracellular matrix which allows cell infiltration and tissue invasion, it is also a co-receptor in many processes such as inflammation and cancer; therefore, it has been suggested that *HPSE* could be an important regulator of these processes (Agelidis et al., 2021; Vlodavsky et al., 2020). (Vlodavsky et al., 2020) It has been reported that activated T cells produce heparanase to allow extravasation out of blood vessels during the adaptive immune response (Naparstek et al., 1984). Increased expression with T cell activation makes the pLoFs affecting *HPSE* an interesting candidate for aCD3 stimulation.

CD109 has expression across several cell-types that is activation-specific

CD109 is a cell surface antigen with specific expression in hematopoietic stem cells, activated T cells and activated platelets; conversely it has low expression in mature PBMCs at rest (Lin et al., 2002). It has been previously reported that *CD109* may be involved in B

cell antibody induction by CD4 T helper cells (Suciu-Foca et al., 1985), but its function is currently understudied. pLoFs affecting *CD109* are good candidates to study with aCD3 stimulation.

***TMEM71* may play a role in the immune contribution to glioma chemoresistance**

Transmembrane protein 71 (*TMEM71*) is part of the gene encoding transmembrane (*TMEM*) protein family with an unknown function. A study characterising the gene in the setting of glioma, where it was found to be highly expressed in glioma cells resistant to temozolomide (chemotherapy) treatment, and the genes that were co-expressed with *TMEM71* related to the immune response (K.-Y. Wang et al., 2019). It demonstrates tissue-specific expression in bone marrow and lymphoid tissues, along with muscle tissues (Uhlén et al., 2015).

***CST7* has high and specific expression in leukocytes, with a possible immune role**

The *CST7* gene encodes for the cysteine proteinase inhibitor known as cystatin F, which is known to have high expression in NK and CD8 T cells. Some of the cysteine proteinases, such as cathepsin C, are expressed in NK and CD8 T cells to trigger cell death (Perišić Nanut et al., 2014), therefore, because of its inhibitory function against cysteine proteinases, it has been reported that *CST7* may play a role in regulating the cytotoxicity triggered by NK and CD8 T cells (Kos et al., 2018; Perišić Nanut et al., 2017).

TRIM58* may regulate the innate immune signalling via *TLR2

Tripartite motif containing 58 (*TRIM58*) is an E3 ubiquitin ligase with multiple reports of involvement in carcinogenesis and chemotherapy resistance (Gan et al., 2021; J. Wang et al., 2022). Its role in immunity is less understood, however it has been shown to interact with *TLR2* in myeloid cells to downregulate the innate immune response in the context of inflammatory bowel disease (Eyking et al., 2019).

The aims of this chapter

In the last chapter we have demonstrated the presence of direct LoF effects on decreasing the expression of the genes these LoFs are coded within. In this chapter we aim to leverage the single-cell nature of our study design to characterise these direct effects across different cell-types and cell contexts.

Secondly, we aim to look for indirect LoF effects on the wider immune system, in the form of other differentially expressed genes, pathway differences and cell-type proportion changes associated with these LoFs. We hypothesise again that these effects would vary across cell-types.

Lastly, we hypothesise that dynamic LoF effects could be uncovered in the presence of challenges to the immune system, particularly as we have prioritised genes with possible roles in immunity. Therefore, we have selected a subset of LoFs for *in vitro* stimulations to activate the innate and adaptive immune responses. We aim to identify both direct and indirect LoF effects that are context-specific, and could possibly be absent when profiling cells only at rest.

Collectively, we believe this will provide us better insight into the function of these LoF-affected genes in the immune system.

Methods

The recall-by-genotype process has been described in the previous chapter. The data generation for both chapters (Chapter 3 and Chapter 4) is summarised in Figure-4.2.

In brief, 488 individuals with homozygous or heterozygous pLoFs in genes predicted to be active in the immune and haematological systems were prioritised for recall. We then successfully recalled 92 individuals, homozygous and heterozygous for 24 pLoFs in 19 genes, and 22 control individuals. From each of these individuals, an additional (to Cardinal's blood collection) 10-20mL sample of whole blood was collected in EDTA-coated phlebotomy tubes for this extended analysis. These samples were stored at around 4°C and transported from the clinic to the Wellcome Sanger Institute (WSI) within 24-72 hours.

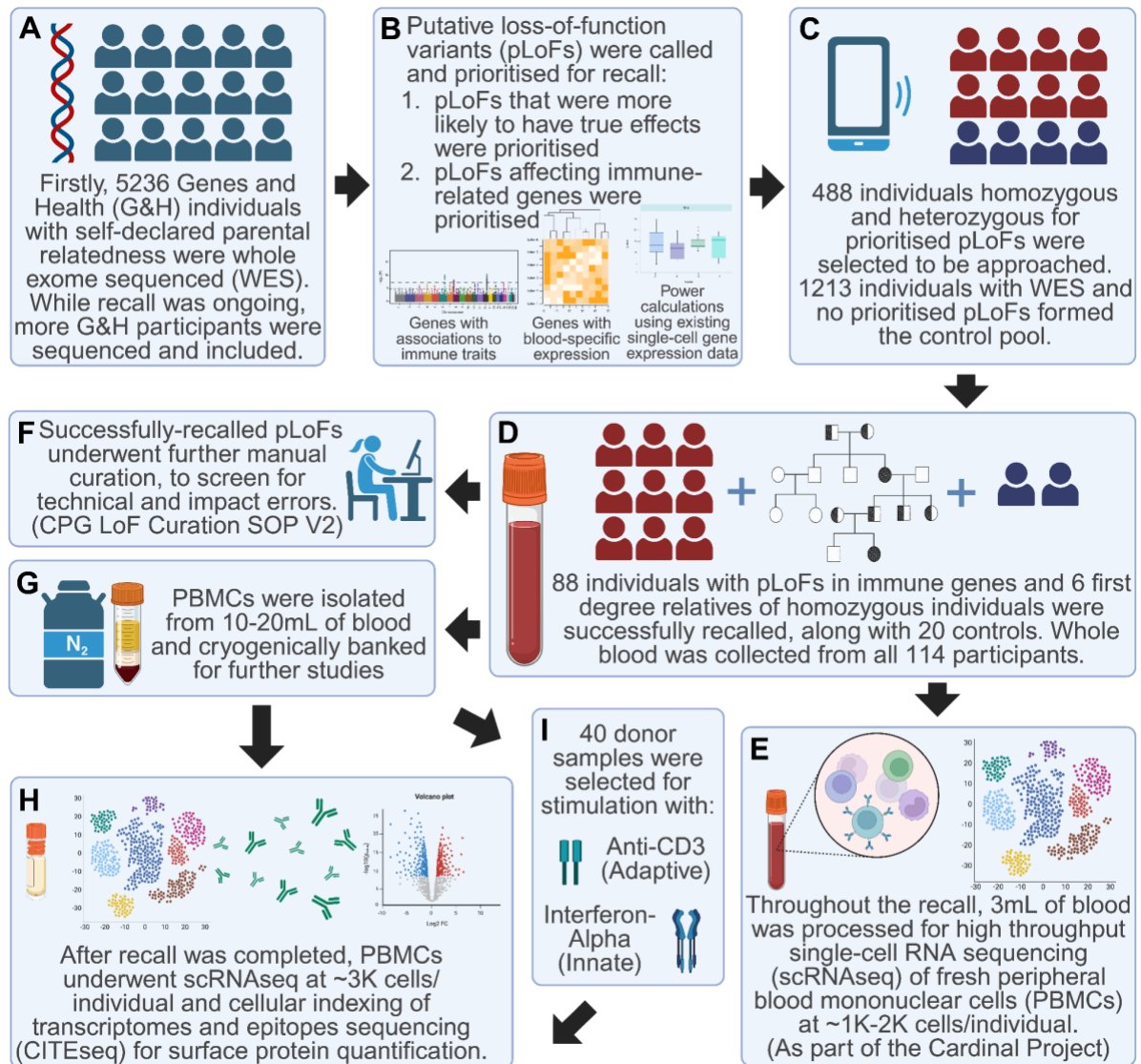


Figure-4.2

A summary of the data generation workflow in this recall-by-genotype. Steps **A** to **F** have been covered in Chapter 3, while steps **G** to **I** is covered here. Created with BioRender.com.

PBMC banking

PBMC isolation from whole blood

On the morning of receiving the samples at the WSI, peripheral blood mononuclear cells (PBMCs) were isolated from whole blood by density gradient centrifugation and banked. At room temperature, whole blood from EDTA tubes belonging to the same individual were pooled and diluted with Phosphate Buffer Solution (PBS pH 7.4) to a volume of 35mL. (Note: In nearly all cases, this was equivalent to a 1:1 dilution. When the amount of blood collected was significantly lower than 17.5mL, the volumes of the protocol were scaled down accordingly.) The blood-PBS mixture was then layered over 15mL of Ficoll-Paque™ Plus, and centrifuged at 700g for 30 minutes with the brakes off. The mononuclear cell layer was then pipetted into cell culture media cRPMI (RPMI Medium 1640, 10% Foetal Bovine Serum, 1% Penicillin-Streptomycin 10,000U/mL, 1% L-Glutamine 200mM) to a final volume of 50mL. To wash, this was centrifuged at 400g for 5 minutes (brakes on), supernatant discarded, and the resultant pellet was resuspended in 20mL of cRPMI. After a haemocytometer cell count and trypan blue viability check, the sample was washed again. The resultant pellet was resuspended in freeze media (RPMI Medium 1640, 50% Foetal Bovine Serum, 10% Dimethylsulfoxide) to a concentration of 5-10 million cells/mL and distributed into 2mL cryovials in 1mL aliquots. The cryovials were transferred into Mr. Frosty freezing containers and stored in the -80°C freezers overnight (16-24 hours), before being transferred into the liquid nitrogen storage tank.

Protocol for thawing PBMCs

This thaw protocol is used throughout the project whenever banked PBMCs were required. The cryovial containing the selected PBMCs was transferred from the liquid nitrogen storage tank onto dry ice on the day of the experiment. The cryovial is then quickly thawed in a 37°C water bath for no longer than 45 seconds before being transferred into 15mL of warmed (37°C) cRPMI. This was washed thrice: centrifuged at 400g room temperature for 5 minutes, discarded supernatant, and resuspended pellet in 15mL cRPMI, with the final resuspension volume being variable depending on downstream requirements.

Preparing resting and stimulated PBMCs for scRNAseq

The cohort was divided into donors that would be analysed at rest only (the “RESTONLY” workflow) and individuals whose cells would be analysed at rest and with stimulation (the “RESTSTIM” workflows). A summary of the workflows is illustrated in Figure-4.3.

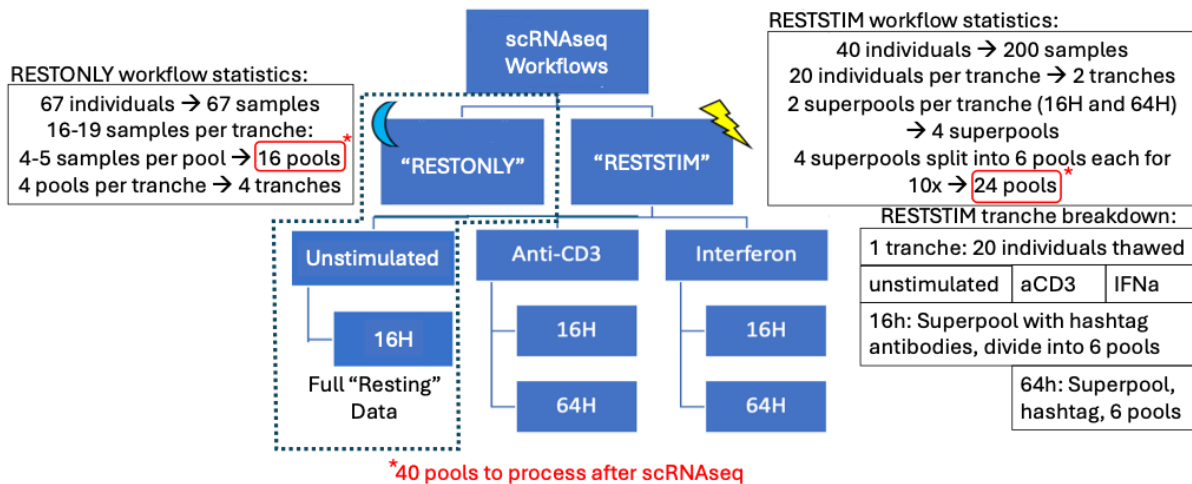


Figure-4.3

An overview of both scRNAseq workflows with a summary of the numbers of donors and pools. Two workflows were used to generate data in this study: “RESTONLY” and “RESTSTIM”. A tranche is defined as the set of donors processed through a single workflow. E.g. The “RESTONLY” workflow was performed 4 times, with 16-19 samples processed each time, resulting in 4 tranches. Within each tranche, a pool is formed when cells from different donors are mixed together for scRNAseq. This mixing is performed to avoid batch effects being correlated with donor effects. In the “RESTSTIM” workflow, where there are multiple stimulation conditions, superpools are formed when samples from different donors, and different stimulation conditions, are mixed together. Again, this is performed to avoid batch effects being correlated with donor or stimulation effects. A superpool is large, so many pools are then formed by evenly splitting these superpools, for scRNAseq.

The “RESTONLY” workflow: Preparing samples without stimulation

Thawing and resting cells

Upon thawing the selected cryovials following protocol (above), the samples were resuspended in 1mL of warmed cRPMI and counted by a Cellaca cell counter with AO/PI staining for viability. Cells were then plated onto 24 well plates (1 donor per well, 1mL each) and incubated overnight (16 hours, or 16h) at 37°C, 5% carbon dioxide. (Samples with more than 5 million cells/mL were diluted to 5 million cells/mL before resting.)

Pooling donors and CITEseq staining

Subsequent steps were performed on ice. Unless otherwise stated, other centrifuge settings were 4°C, 400g, 5 minutes, brakes on. After resting, 900uL of the cell suspension was transferred into a 15mL Falcon tube, following which 1mL of EDTA solution (calcium and magnesium deficient PBS with 1mM EDTA) was added to the well for 10 minutes to lift remaining cells off the well surfaces, before all remaining solution was transferred into the Falcon tube. The tube was spun down, supernatant discarded, and the cell pellet resuspended in 1mL of Cell Staining Buffer (Biolegend). Cells were counted (Cellaca, AO/PI) and 550,000 cells per donor were pooled together into pools of four to five donors. Each pool was concentrated by centrifugation and resuspended in 40uL of Cell Staining Buffer, and a cell count was taken (Cellaca, AO/PI). 750,000 cells per pool were extracted for antibody staining: First blocked with Fc Blocker (Human TruStain FcX™ Fc Receptor Blocking Solution) following the product recommendation of 5uL per 1 million cells for 10 minutes, and subsequently stained with the CITEseq antibody cocktail (TotalSeq™-C Human Universal Cocktail, V1.0, reconstituted following the recommended protocol) at a 1:12 dilution ratio in a final volume of 75uL for 30 minutes.

Sorting PBMCs for viability using flow cytometry

Each pool was diluted with 1mL Cell Staining Buffer for three washes, before being resuspended in 500uL of Cell Staining Buffer for viability staining with 7-aminoactinomycin D (7AAD, a fluorescent marker that binds to DNA) (10uL per million cells). The cells were then sorted for viability using a fluorescence-activated cell sorting (FACS) machine: Firstly by gating to select cells by plotting complexity (BSC-A) against cell size (FSC-A), secondly by gating to select singlets by plotting cell size width (FSC-W) against cell size (FSC-A), and lastly by gating to exclude 7AAD-stained dead/dying cells, aiming to collect 250,000 live singlet events per pool. A final cell count and viability check (Cellaca, AO/PI) was performed at this stage to prepare 25,000 cells per pool for scRNAseq.

Preparing for scRNAseq

Single-cell encapsulation and library preparation were performed following the recommended protocol from 10X Genomics, using the Single Cell 5' HT Reagent Kits v2 (Dual Index) and Chip K on the Chromium Controller. The 5' Gene Expression library and Cell Surface Protein library was pooled at a ratio of 5:1 to aim for 50,000 RNAseq reads and 10,000 CITEseq reads per pool. For sequencing, the NovaSeq 6000 S4 flow cell read parameters were 28bp (Read 1), 10bp (Index), and 90bp (Read 2).

The “RESTSTIM” workflow: Stimulating selected samples

Thawing, resting and stimulating cells

Upon thawing the selected cryovials following protocol (above), the samples were resuspended in 500uL of warmed cRPMI. They were then counted (Cellaca, AO/PI) and diluted to 4 million cells/mL. Subsequent steps were programmed and executed by a Viaflo liquid handler. Cell suspensions were plated onto 96 well plates: 100uL each well, 1 donor per well, 5 wells per donor. Cells were rested for 2 hours in an incubator (37°C, 5% carbon dioxide).

There were five wells per donor in preparation for collection at five data points. Therefore for each donor, 100uL of cRPMI or stimuli were added accordingly: cRPMI to one well (for the unstimulated 16h data collection to match the RESTONLY workflow), aCD3 (anti cluster of differentiation 3) reaching a working concentration of 500ng/mL into two wells (one each for the 16h and 64h data collection), and IFNa (interferon alpha) reaching a working concentration of 1,000 U/mL into two wells (one each for the 16h and 64h data collection). The cells were then incubated up to their designated timepoint for data collection.

Pooling donors, hashtag antibody and CITEseq staining

Subsequent steps were performed on ice. The Viaflo liquid handler was still used here. Unless otherwise stated, centrifuge settings remained at 4°C, 400g, 5 minutes, brakes on. At the timepoint for data collection, 160uL of the cell suspension from each well was transferred into each well in a 1mL round bottom deep well collection plate, following which 200uL of EDTA solution was added to the well for 10 minutes to lift remaining cells off the well surfaces, before all remaining solution was transferred. The plate was spun down, supernatant discarded, and the cell pellet resuspended in 150uL of Cell Staining Buffer. Cells were counted (Cellaca, AO/PI) and 60,000 cells per donor were pooled together into pools of 20 donors of the same stimulation condition (no longer in well plates, transferred to 2mL FACS tubes). Each pool was concentrated by centrifugation and resuspended in 45uL of Cell Staining Buffer. This resulted in 1.2 million cells per pool for antibody staining: First

blocked with 5uL of Fc Blocker for 10 minutes, and subsequently stained with the CITEseq antibody cocktail at a 1:12 dilution ratio along with 2uL of TotalSeq™ anti-human hashtag antibody in a final volume of 100uL for 30 minutes. (“Hashtag 2” for unstimulated pools, “Hashtag 3” for aCD3 pools, “Hashtag 4” for IFNα pools).

Sorting PBMCs on viability using flow cytometry

Each pool was diluted with 1mL Cell Staining Buffer for three washes, before being resuspended in 1mL of Cell Staining Buffer for a cell count and viability check (Cellaca, AO/PI). An equal number of cells from each pool (700,000 - 1 million cells per pool) were combined into a “superpool” containing cells from all stimulation conditions at that timepoint. Each superpool was spun down and resuspended in 1mL of Cell Staining Buffer with DAPI (1:20,000 dilution) for viability staining. The cells were then sorted by FACS aiming to collect 1 million live singlet events per pool. A final cell count and viability check (Cellaca, AO/PI) was performed at this stage to prepare for harvesting 115,000 cells per superpool.

Preparing for scRNAseq

This followed the RESTONLY workflow, with the main difference being that Chip N was used with Chromium X for single-cell encapsulation (following the relevant recommended protocol). Each superpool (containing cells at a timepoint, labelled by hashtag antibodies to indicate stimulation conditions) was processed in six equal aliquots.

Contributions to the data generation component of the project

I performed the isolation and banking of the PBMCs, along with the interim fluidigm genotyping and viability checks. Together with Bess Chau and Tarran Rupall, we conducted the “RESTONLY” workflow, and Bess and Tarran conducted the “RESTSTIM” workflow. After GEM RT-PCR, library preparation and sequencing (scRNAseq and CITEseq) were performed by the DNA pipelines team at the WSI, up to the point of read alignment with Cell Ranger below.

Quality control (QC) of the scRNAseq and CITEseq data

Annotation and filtering

The scRNAseq and CITEseq data was generated as FASTQ sequencing output, and read alignment was performed with the 10x Genomics Cell Ranger v7.0.0 multi pipeline to generate barcode-feature matrices of counts. For each pool of droplets sequenced, two matrices were generated: one with features (genes) detected from RNA reads, and one with features (genes/proteins) detected from reads of antibody tags (the CITEseq panel, and for the RESTSTIM workflow, hashtag antibodies). To QC the data, we first annotated the full

dataset and then performed filtering based on these annotations. This is summarised in Figure-4.4:

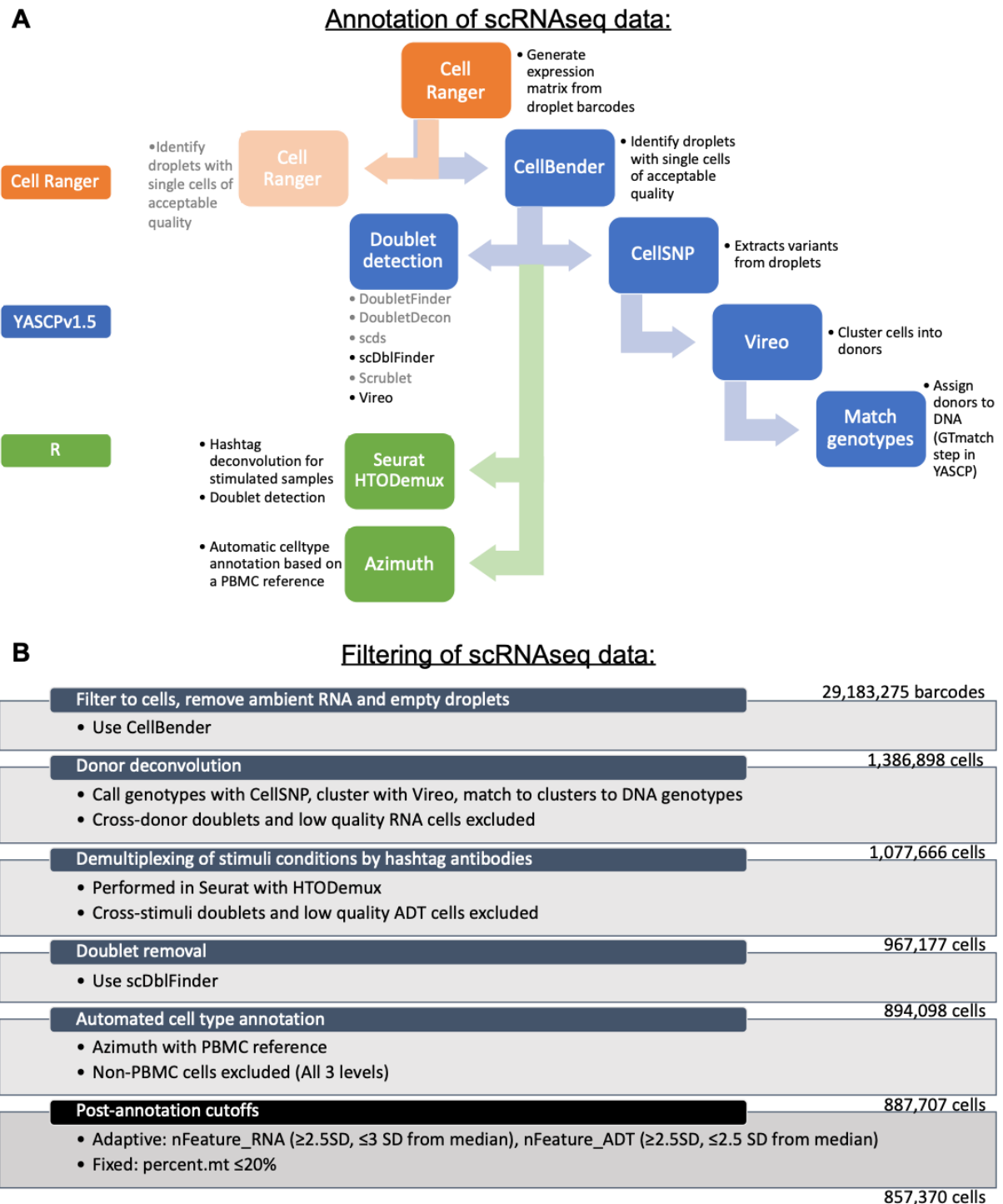


Figure-4.4

Annotation and filtering of the scRNAseq and CITEseq data. A: Annotations performed on the scRNAseq data and the softwares used, in darker font are algorithms taken forwards for filtering. **B:** Stepwise filtering of the scRNAseq data based on annotations from A.

CellBender (Fleming et al., 2023) was then used to define cells from the RNA feature matrices (differentiating cells from empty droplets, or droplets with dying/lysed cells). Donor deconvolution was performed with cellSNP (X. Huang & Huang, 2021) and Vireo (Y. Huang et al., 2019), and genotypes were matched to the TOPMED-r3 (Taliun et al., 2021) imputed genotypes from DNA samples of the same individuals. For the RESTSTIM superpool aliquots, deconvolution of hashtags to differentiate stimulation conditions at timepoints was performed with the Seurat v5 R package (Hao et al., 2024) on the ADT feature matrices. Cross-donor and cross-stimuli doublets were already removed by the two deconvolution steps, and an additional doublet removal was performed with scDblFinder (Germain et al., 2022). Cell-type annotation was performed with Azimuth, mapped with the Hao et al., 2021 PBMC reference dataset (Hao et al., 2021). The YASCP v1.5 pipeline (“Yet Another Single Cell Pipeline”) (Ozols, 2023) developed by the WSI’s Human Genetics Informatics Team was deployed with default settings to run CellBender, cellSNP, Vireo, scDblFinder and Azimuth. Cell Ranger filtering (similar to CellBender in purpose) and other doublet detection algorithms built into YASCP were also run, and we performed comparisons between the different methods before deciding on the pipeline described here. The results of these comparisons are in the next subsection “Assessment of the QC”.

It is standard practice to exclude outlier cells based on several additional QC metrics. Firstly, read counts in a droplet is an important measure to evaluate: too low and it is suggestive of an empty droplet or capture of a poor quality cell, while an abnormally high count is suggestive of a droplet with multiple cells (multiplet). We can quantify the number of counts in a droplet either by the number of unique genes detected (nFeature), or the number of unique molecular identifiers detected (UMI count, or nCount). The nCount is usually correlated with the nFeature (satijalab.org, 2020).

Another metric assessed is the fraction of reads assigned to mitochondrial genes. A high percentage of mitochondrial reads (percent.mt) may represent apoptotic or lysing cells that result in mitochondrial contamination.

As the cells have now been annotated, outliers could be excluded based on adaptive cutoffs that account for the variable distributions across different cell categories. We settled on the following cutoffs to include cells for each subset by workflow (RESTONLY or RESTSTIM), stimulation condition (stimulus and timepoint), and cell-type (Azimuth level 2):

- Number of unique genes counted from the RNA feature matrix (nFeature_RNA) ≥ 2.5 standard deviations (SDs) and ≤ 3 SDs from median

- Number of unique genes/protein counted from the antibody tags feature matrix (nFeature_ADT) ≥ 2.5 SDs and ≤ 2.5 SDs from median
- percent.mt $\leq 20\%$

This resulted in a final dataset of 857,370 cells (Figure-4.4B shows the number of cells at each step of filtering).

Assessment of the QC

Comparing Cell Ranger and CellBender for filtering ambient RNA

The raw feature-barcode matrix compiled by Cell Ranger contained barcodes assigned to empty droplets, and low quality, lysing, or dying cells. Cell Ranger and CellBender are capable of performing a similar function (with slightly different algorithms) of annotating these as “noisy” droplets and identifying barcodes assigned to droplets that likely contain single cells. The majority of the raw barcodes were indeed assigned as droplets without cells, with both algorithms being concordant. However, CellBender was able to retain about 10% more cells than Cell Ranger. (Figure-4.5A)

We found that the barcodes identified to be droplets not containing single cells if an acceptable quality by both algorithms indeed had poorer measures of nFeature count and higher percent.mt. Interestingly, the barcodes assigned as cells by CellBender only and not Cell Ranger had a nFeature and percent.mt distribution similar to barcodes assigned as cells by both algorithms. From this we inferred that the 10% gain in assigned cells from using CellBender alone would not result in keeping more low quality cells. In contrast, barcodes assigned as cells only by Cell Ranger and not by CellBender demonstrated poorer QC metrics. (Figure-4.5B). We therefore proceeded to filter barcodes for single cells by only the CellBender annotation.

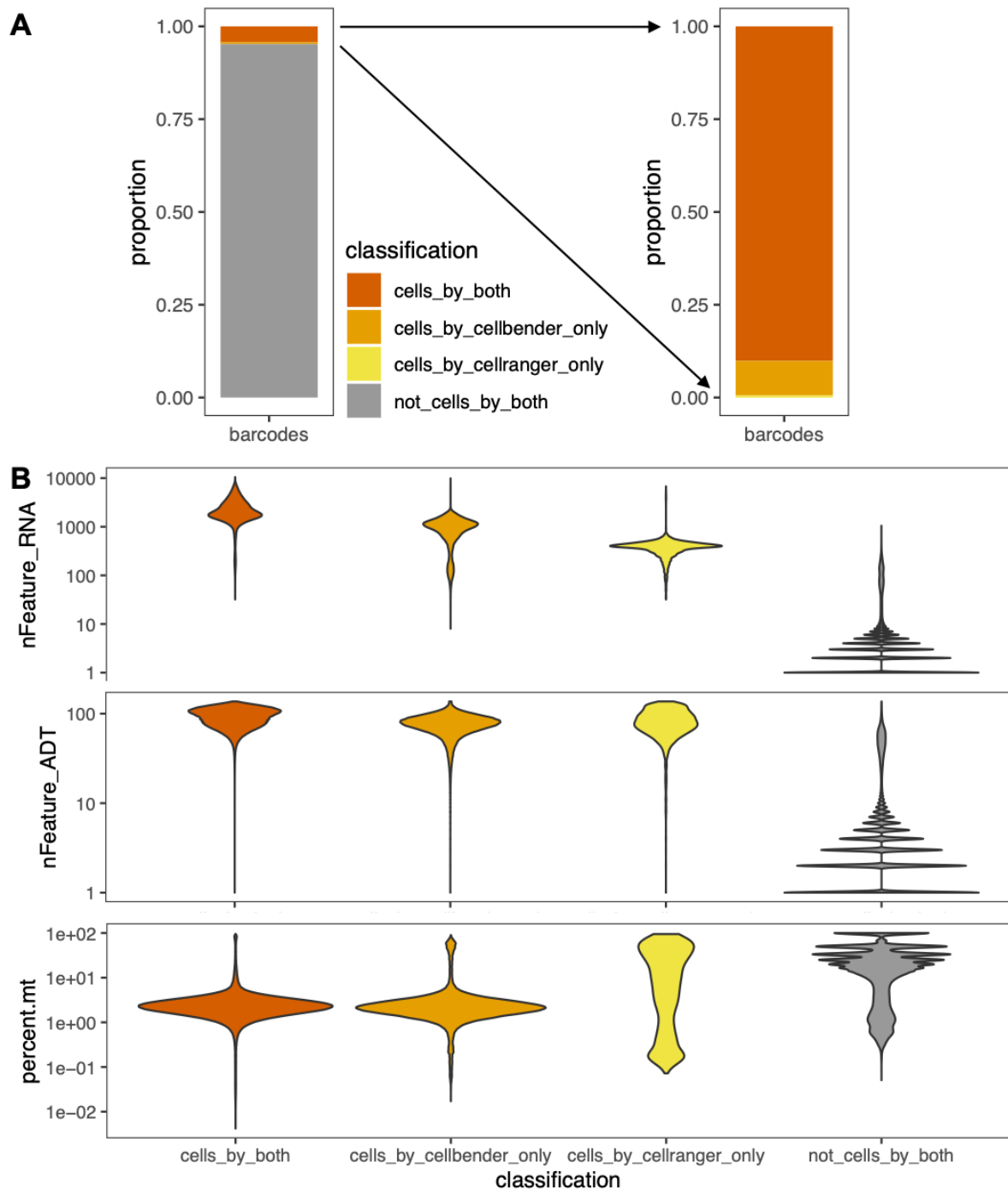


Figure-4.5

Comparing CellBender and Cell Ranger. A: The proportion of raw barcodes assigned as cells or not by both algorithms. “cells_by_both”: Barcodes that have been identified to represent single cells in both algorithms; “cells_by_cellbender_only”: Barcodes that have been identified to represent single cells in CellBender but not Cell Ranger; “cells_by_cellranger_only”: Barcodes that have been identified to represent single cells in Cell Ranger but not Cell Bender; “not_cells_by_both”: Barcodes that have been identified by both algorithms to represent droplets without viable single cells. **B:** Distribution of QC

metrics nFeature_RNA, nFeature_ADT and percent.mt across barcodes based on their classification by both algorithms.

Benchmarking doublet detection methods

The yasep v1.5 pipeline incorporated annotation of droplets as singlets or multiplets by five doublet detection methods (Bais & Kostka, 2019; DePasquale et al., 2019; Germain et al., 2022; McGinnis et al., 2019; Wolock et al., 2019). We saw good correlation between the annotations across the five algorithms (Figure-4.6A). Although a common practice is to use more than one multiplet detection method, applying all would be too stringent. On top of the doublet detection performed during deconvolution by vireo (cross-donor doublets) and with hashtag antibodies (cross-stimulus doublets), we decided to add the best performing doublet detection method to additionally filter away within-donor, within-stimulus doublets, which we hypothesised to be a minority.

Among the doublets defined by vireo and hashtag antibodies, scDbIFinder was the most concordant, i.e. it detected the most number of doublets among this set of “positive controls” (Figure-4.6B). Secondly, among the cells that have not been identified as doublets by vireo and hashtags, scDbIFinder did not find too many doublets, suggesting that it might not detect too many “false positives” as well (Figure-4.6C).

This made scDbIFinder our doublet detection method of choice for this dataset.

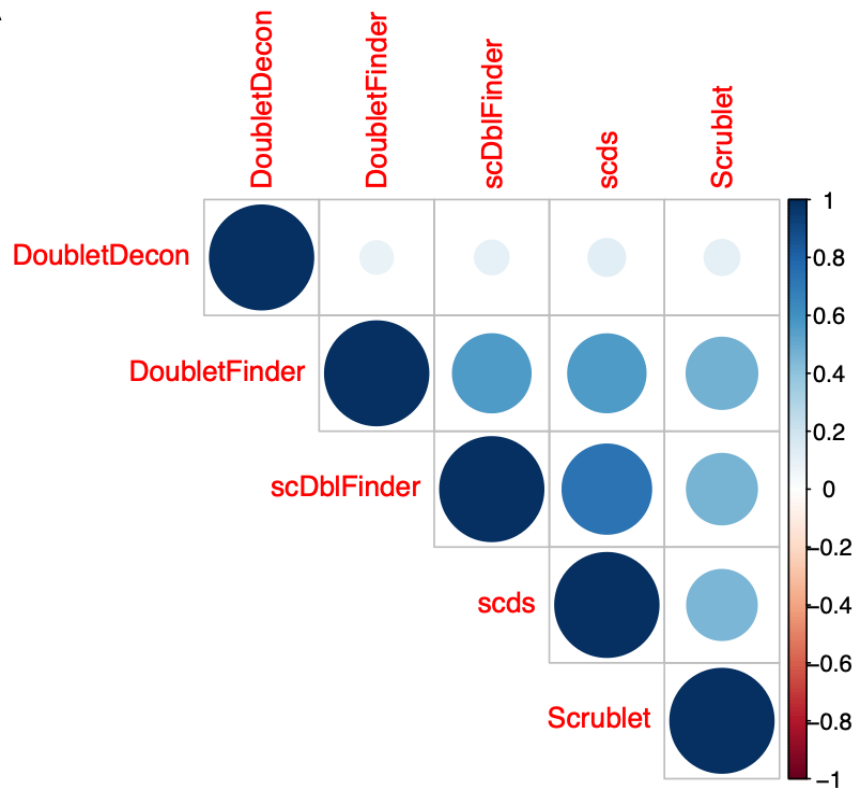
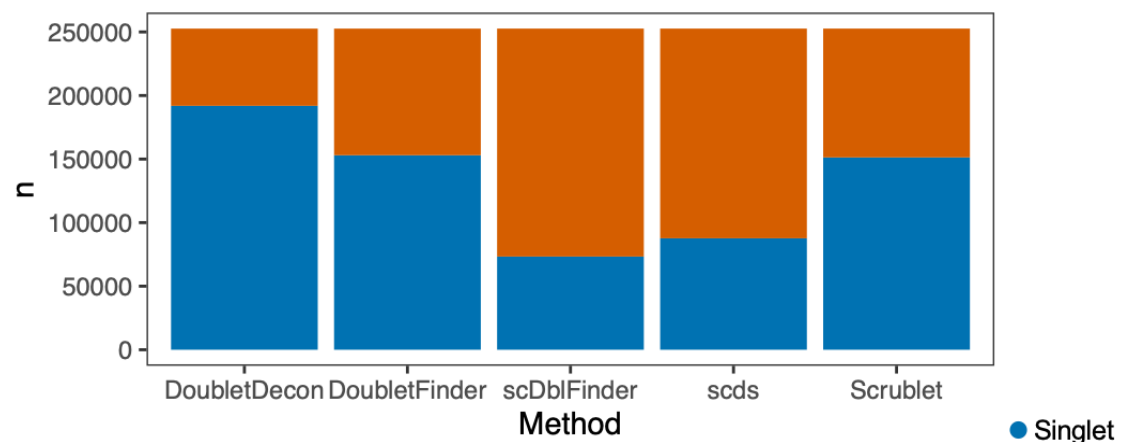
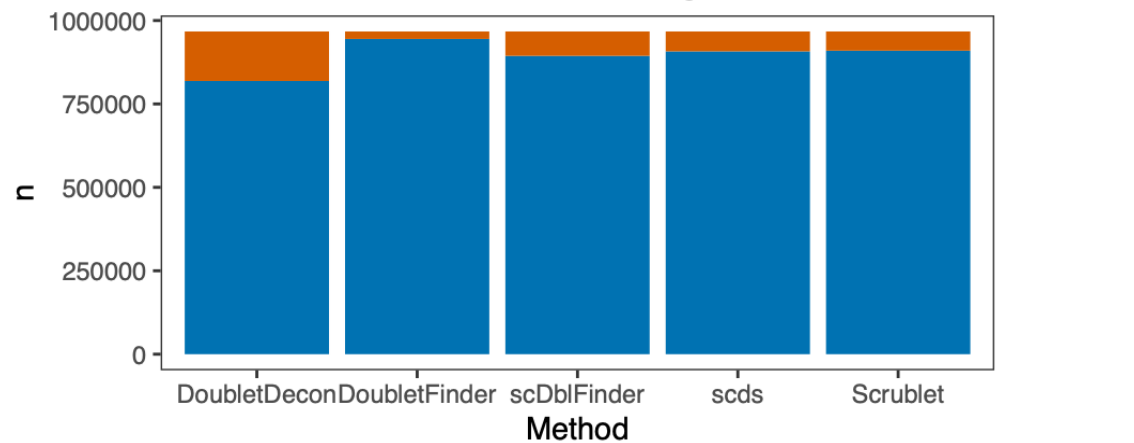
A**B** Doublets from Donor and Hashtag Deconvolution**C** Cells Passed Donor and Hashtag Deconvolution

Figure-4.6

Comparing doublet detection methods. A: Correlation plot of the doublet annotations across the doublet detection methods. **B:** Proportions of the doublet annotation performed across doublet detection methods performed on doublets defined by vireo and hashtag antibodies. **C:** Proportions of the doublet annotation performed across doublet detection methods performed on cells not defined as doublets by vireo and hashtag antibodies.

Evaluating the cell-type annotation by azimuth

The Azimuth annotation is performed to three levels of resolution, and comes with a mapping score, which rates how closely the data matches the reference, and a “prediction score”, which quantifies the confidence of assigning a cell-type at that level of resolution, based on the combined support from different anchor features in the algorithm (Hao et al., 2021).

The Hao et al., 2021 reference dataset we used is composed of unstimulated PBMCs, therefore, as expected, the unstimulated cells from both workflows mapped well. Stimulated samples had poorer mapping scores. However, even in the stimulated dataset, the prediction scores remained high for level 1 and 2 annotations. This indicates that the expression profiles of PBMCs in the stimulated dataset is different from the unstimulated reference, as expected, but cell-type annotations were still performed with good confidence. (Figure-4.7)

Level 3 had a poorer distribution of confidence scores, and upon inspection of their cell-type labels, did not provide further biological classification relevant to the aims of this project. (For example, central memory T cells were divided into “1”, “2”, and “3”.) Therefore, in this project, we focused on level 1 and level 2 cell-type annotations. The agreement between the two levels in our data was 97%.

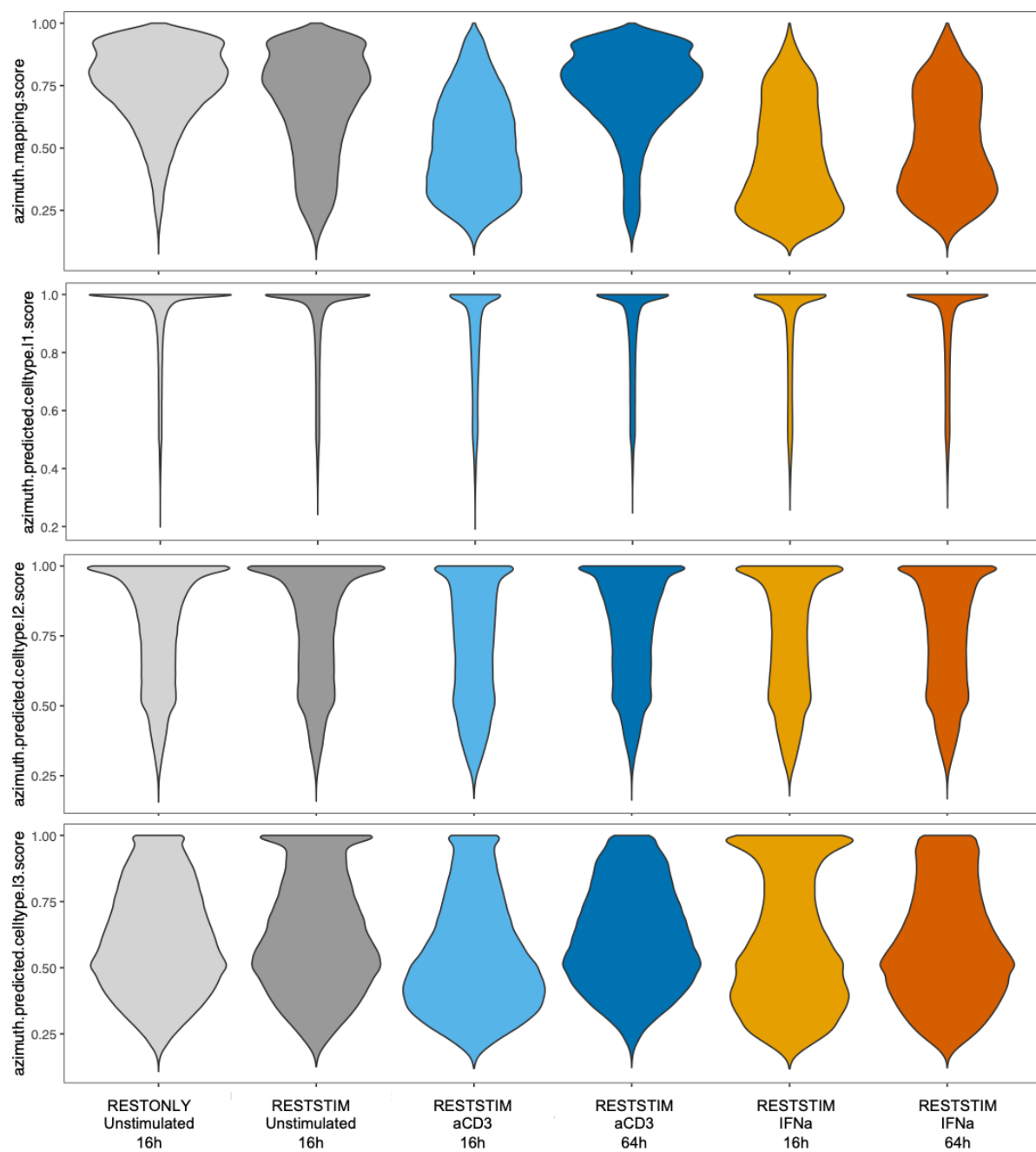


Figure-4.7

Distribution of azimuth QC scores across workflow, stimulus, and timepoint.

Optimising adaptive cutoffs

As mentioned previously, metrics such as gene counts and percentage of mitochondrial reads are useful to filter out multiplets and low quality cells. However, the distribution of these measures can vary across cell-types and contexts; different cell-types may express fewer or more genes, cells with more metabolic activity during stimulation may present with higher mitochondrial reads. We therefore decided to employ adaptive cutoffs based on subgroups of cells, and investigated which resolution of cell classification would be optimal.

We saw that the distributions of `nFeature_RNA` and `nFeature_ADT` varied across workflows, stimuli, timepoints, and azimuth level 2 cell-type annotations. Therefore we applied adaptive cutoffs that would exclude outliers of these metrics in each distribution. However, the percentage of mitochondria reads did not vary drastically across the different subsets, so a fixed cutoff was selected. (Figure-4.8 and Figure-4.9)

In Figure-4.10 we show several examples of the adaptive cutoffs applied on `nFeature_RNA` across different cell subsets.

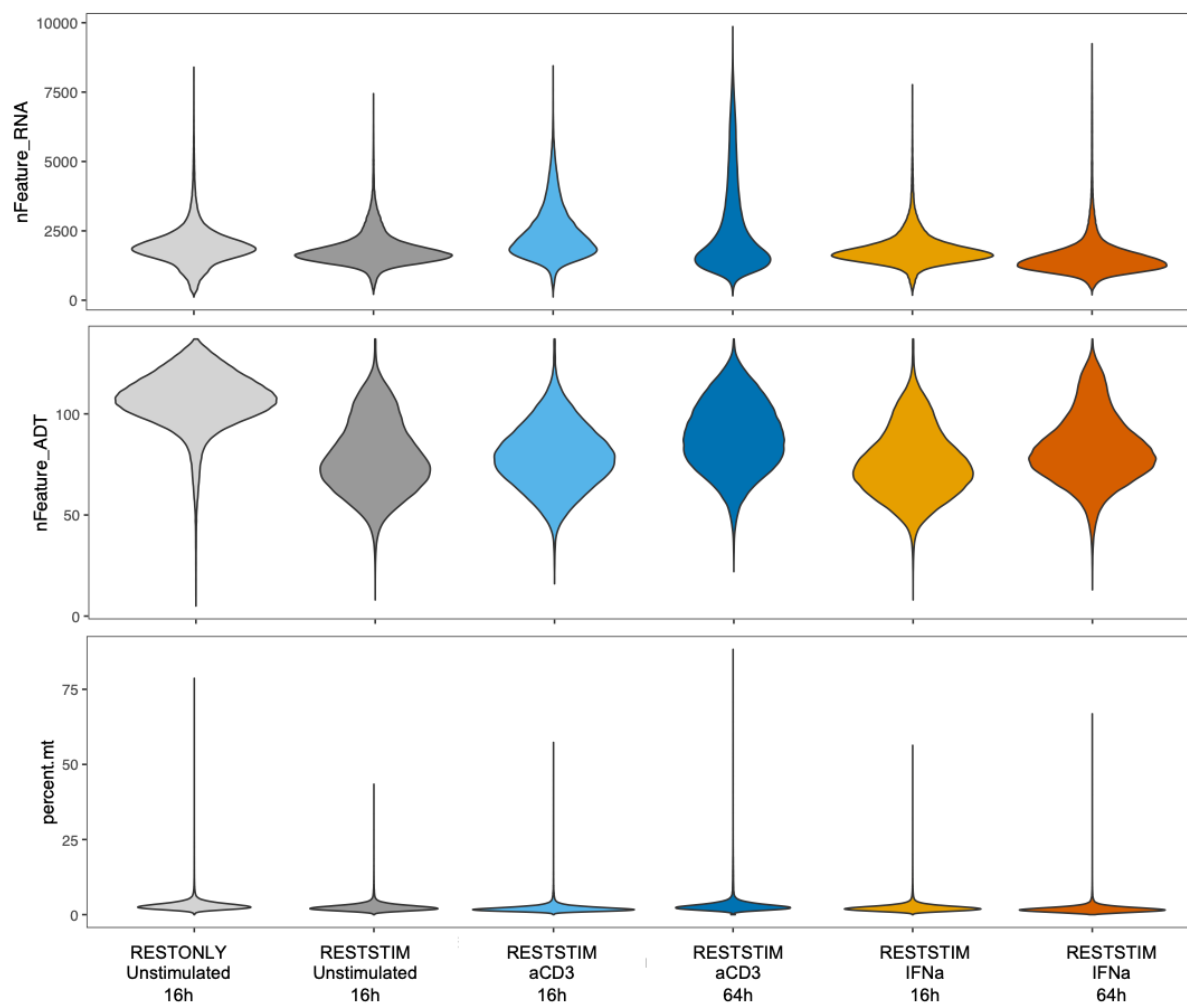


Figure-4.8

Distribution of QC metrics RNA UMI counts (nFeature_RNA), antibody tags UMI counts (nFeature_ADT) and percentage of mitochondria reads (percent.mt) across workflows, stimuli and timepoints.

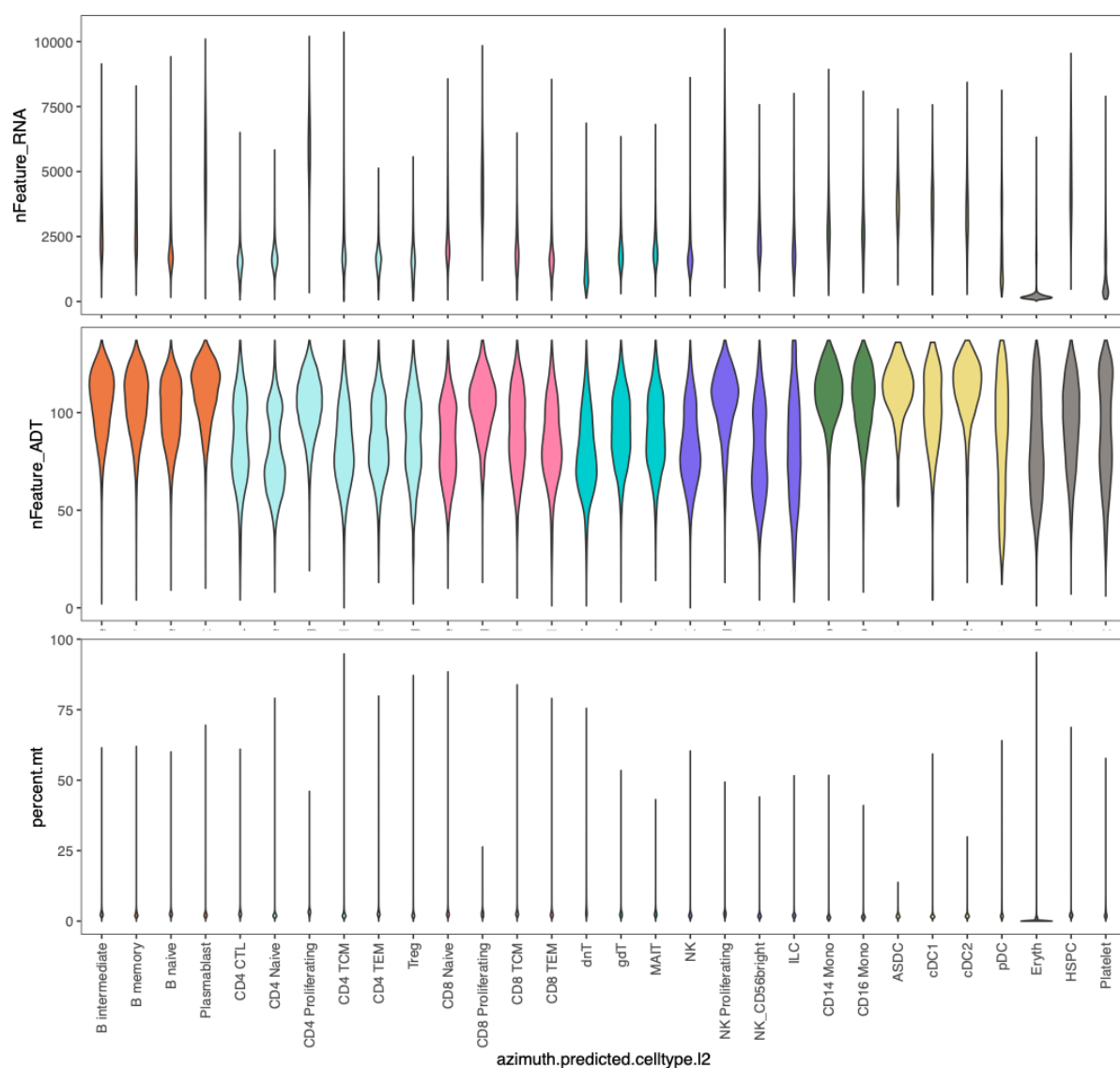


Figure-4.9

Distribution of QC metrics RNA UMI counts (nFeature_RNA), antibody tags UMI counts (nFeature_ADT) and percentage of mitochondria reads (percent.mt) across azimuth level 2 cell-type annotations.

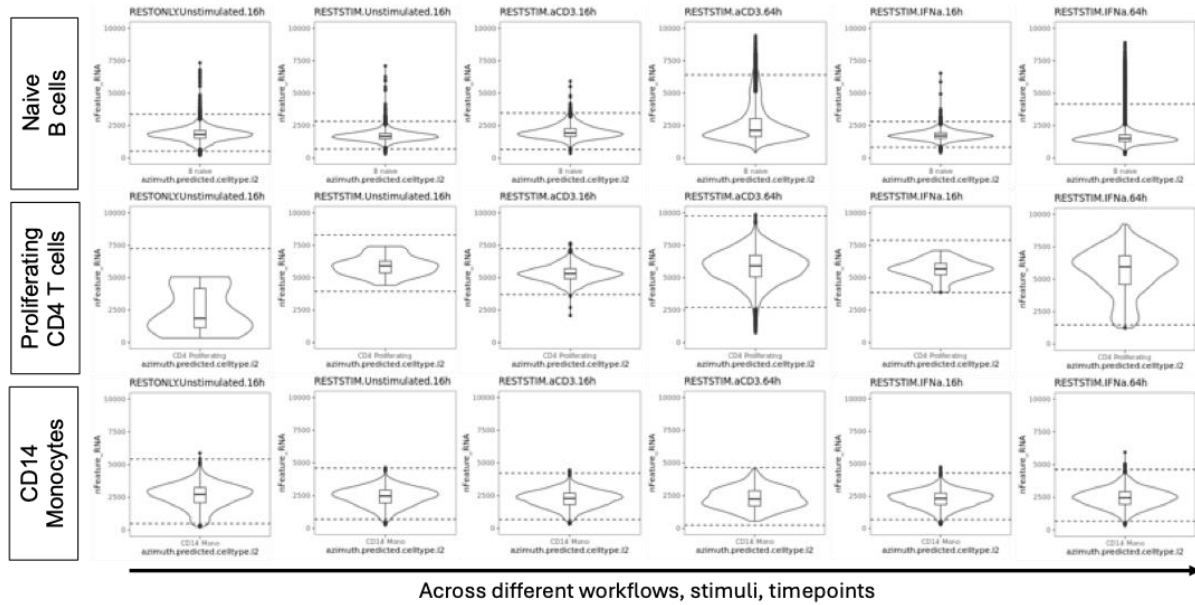


Figure-4.10

Distribution of RNA UMI counts (nFeature_RNA) across different cell-types, workflows, stimuli and timepoints, with their adaptive cutoff overlaid as horizontal dotted lines.

Dimensionality reduction and visualisation of scRNAseq data

At different steps of the QC, we performed dimensionality reduction in order to visualise the data, monitor the progress and outcome of a QC step, and identify factors that we would need to consider during downstream analysis. We used the Seurat v5 R package to perform the dimensionality reduction, and each time it was performed we followed the same approach: Firstly, after the target QC step or filtering, we extracted a 10% subsample of cells to perform the reduction on, and then projected the rest of the samples into the reduced space (Seurat sketch). For the sketch reduction, we identified 2,000 highly variable features excluding T cell and B cell receptor (TCR, BCR) genes (explained in the previous chapter), before performing a principal component analysis (PCA) and uniform manifold approximation and projection (UMAP) with 30 principal components (PCs). The rest of the cells were then projected back into the sketched reduction. Normalisation and scaling of counts were performed where appropriate, following Seurat's recommendations. The visualisation of subsets by context (a stimulation condition and timepoint) is shown in Figure-4.11, with the final combined dataset after QC shown in the results section (Figure-4.14). As expected, within each context, the key driver of variation is cell-type. For data generated via the RESTSTIM workflow (i.e. all stimulated cells), there were no observable batch effects among pools. However, for the unstimulated dataset, there were noticeable batch effects among pools, likely driven by the workflow used to generate the data.

Core data analysis

Differential expression analysis

Model design and choosing limma-voom over ImerSeq

Within each workflow, stimulation and timepoint, cells were “pseudobulked” (counts in a subgroup summed together) per donor by their Azimuth level 1 and level 2 cell-type annotations for differential gene expression (DGE) analysis with limma-voom (Law et al., 2014; Ritchie et al., 2015).

For unstimulated samples, we saw that there were notable differences between samples from the RESTONLY and RESTSTIM workflows (Figure-4.11), and therefore wanted to control for workflow in the DGE. By design, we had included several samples in both workflows, allowing us to include a random intercept to model repeated donors. Limma-voom accommodates a single random effect via its “duplicateCorrelation()” function, where a correlation variable can be calculated for the seeded samples and included during the linear modelling within “lmFit()”. We fit the following model:

$$Expression \sim Genotype\ Encoding + Age + Sex + Workflow^*$$

* Workflow as a covariate was only required for the unstimulated DGE.

In the previous chapter, ImerSeq (Vestal et al., 2022) was employed because of the complex two-step modelling performed. The output of ImerSeq is a beta coefficient that has to be interpreted relative to the normalised and variance-stabilised transformed gene expression. This was useful in the previous chapter when residuals had to be calculated after batch correction for further model-fitting in the two-step model. However, log2 fold-change is a more widely interpretable metric of DGE, and is the output of limma-voom, which is also a more widely used and established method compared to ImerSeq. Here, our model design was more straightforward, as we had complete information about the covariates for all samples to be fitted in a single step. Where possible, one-step modelling presents a lower risk of incorrect fitting especially if there is an unbalanced design (T. Li et al., 2023). In addition, with ImerSeq in the unstimulated runs, many of the mixed model tests failed due to singular fits (many more compared to limma-voom), suggestive of overfitting and that the modelling of the random effect by ImerSeq may be more complex than required (i.e. the batch effect is not as significant as assumed) (Matuschek et al., 2017). Lastly, since we did not notice any batch effects within the RESTSTIM workflow, the stimulated runs required a simple linear regression, which ImerSeq does not implement. These informed our decision to use limma-voom in this chapter.

Normalisation and filtering of lowly-expressed genes for limma-voom

Normalisation was performed following limma-voom's recommended steps: edgeR (Y. Chen et al., 2016) was first used to calculate normalisation factors, after which voom used these factors to perform a transformation to counts per million (cpm) and calculate variance weights that would be applied by limma downstream during the DGE. Like in the variance-stabilised transformation of the previous chapter, these steps are necessary to normalise the data and account for the relationship between mean and variance.

It is also recommended that filtering of lowly expressed genes should be done prior to normalisation, due to scaling bias that can be introduced by the presence of zero counts in RNA sequencing. Following the edgeR user guide (updated 29 April 2024), we aimed to use genes with a cpm value corresponding to ~5-10 reads in a minimum number of samples that matched the smallest comparison group (Y. Chen et al., 2024). This corresponded to a cutoff of at least 3cpm in 30% of samples, to filter out lowly expressed genes (Figure-4.12, with more details in the caption).

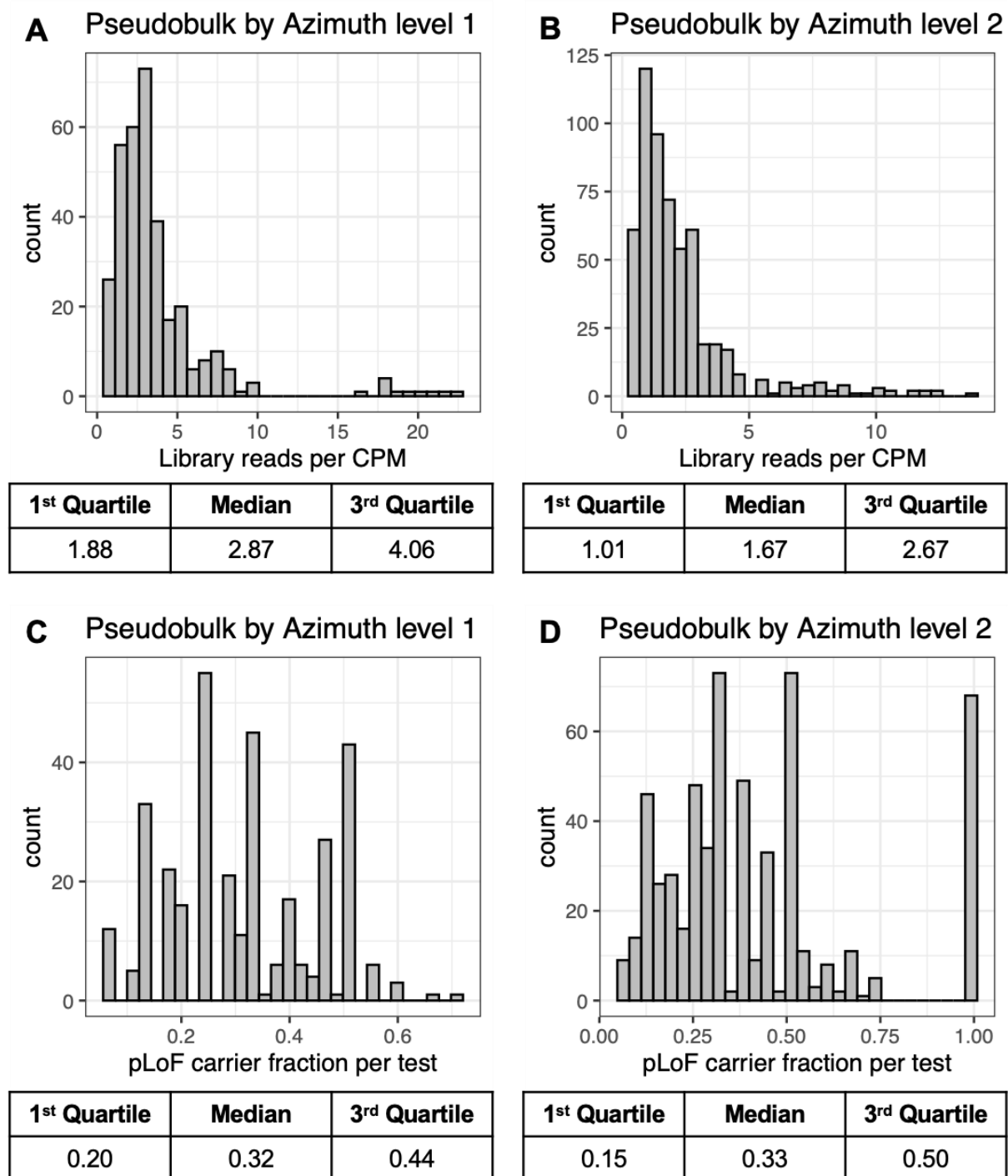


Figure-4.12

Distribution of the number of library reads per cpm and the pLoF carrier fraction for the DGE tests that would be performed. **A:** We performed the edgeR transformation from library reads to cpm within each test (i.e. involving only the samples relevant to a controls vs pLoF carriers test) at Azimuth level 1, and calculated the number of library reads corresponding to one cpm. (**B** for Azimuth level 2 DGE tests.) From the medians, we see that 3cpm corresponds to ~8.61reads at Azimuth level 1 and ~5.01reads at Azimuth level 2.

C: In each DGE tests, the pLoF carriers would be typically the smaller comparison group compared to the controls, and we calculated the pLoF carrier fraction in each test at Azimuth level 1. (**D** for Azimuth level 2 DGE tests.) From the medians, we see that the smaller comparison group is ~30%.

Assigning a cutoff for the minimum number of cells to pseudobulk

One of the aims of sequencing more cells per sample in this extended experiment was to allow for more accurate estimations of expression in rare cell-types. We therefore also implemented a minimum cell cutoff that excluded samples that had aggregated counts derived from too few cells. We set the cutoff at 50 cells for Azimuth level 1 annotations (following our power calculations in the previous chapter) and 10 cells for Azimuth level 2. This was supported by the distribution of the number of cells across the cell-types (Figure-4.13). We wanted cutoffs that allowed the capture of the majority of samples within the cell-type, while being high enough to exclude low-quality sample outliers. We therefore decided that the cutoff should fall below the interquartile range of the number of cells per sample for a cell-type and context. For example, at Azimuth level 1, a hundred-cells cutoff would be too high for monocytes, as we would lose many RESTONLY samples and samples stimulated with aCD3 for 64h, ten cells would be too lenient, so fifty is the most appropriate (Figure-4.13A). At Azimuth level 2, the cell-type resolution has increased, therefore the cell counts have decreased, and based on the distributions of cell counts per sample, we felt that a ten-cells cutoff would be most appropriate (Figure-4.13B).

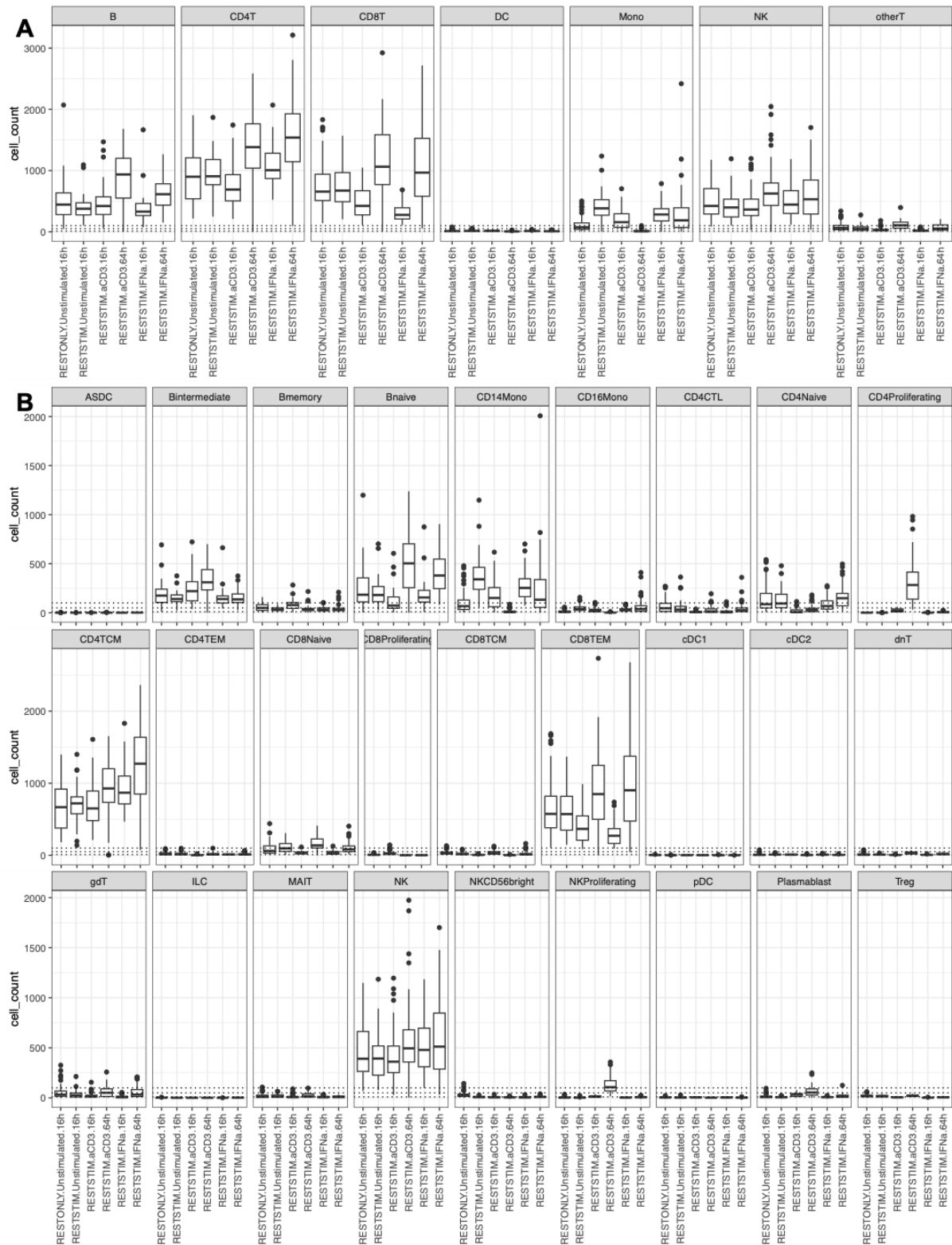


Figure-4.13

Cell counts per sample for a given workflow, stimulation condition, timepoint and cell-type (**A**: Azimuth Level 1, **B**: Azimuth Level 2). Horizontal dotted lines represent the ten-cells, fifty-cells, and hundred-cells cutoffs.

Gene set enrichment analysis

For each DGE run, we used the test statistic of all genes tested in ClusterProfiler (Yu et al., 2012) to run the gene set enrichment analysis (GSEA), looking for enrichment in the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa & Goto, 2000).

We acknowledge that GSEA using the Kolmogorov-Smirnov statistic may not account adequately for Type I error in our results, since it assumes gene independence, which is unlikely in ranked gene lists derived from gene expression data (Tamayo et al., 2017). In the future, it may be worthwhile validating our findings with other pathway enrichment approaches such as fgsea (Korotevich et al., 2021) or confirming strongly significant pathways through over-representation analysis (Boyle et al., 2004).

Differential abundance analysis

We applied the Milo algorithm (miloR implementation) to perform the differential abundance analysis (Dann et al., 2021). From the neighbourhoods defined in a k-nearest neighbour (KNN) graph of the cells, this algorithm systematically tests partially overlapping neighbourhoods for differential abundances.

For each pLoF comparison at a stimuli and timepoint, we constructed a Seurat object containing the raw counts of only the relevant control and pLoF donors. We then performed standard QC: log normalisation and scaling of raw counts, PCA with 2,000 highly variable features, and UMAP with 30 PCs. This data was converted to a milo object (via a single cell experiment object). Where specified, a k of 10 and 30 PCs/dimensions were used. First, the KNN graph was constructed by Milo, then representative neighbourhoods were defined by randomly sampling 0.1 proportion of the cells. Next, counting cells and calculating neighbourhood distances were performed following the recommended guidelines. Finally, differential abundance testing was performed, controlling for the same covariates described in DGE testing:

*Cell abundance in a neighbourhood ~ Genotype Encoding + Age + Sex + Workflow**

* Workflow as a covariate was only required for the unstimulated DGE.

Visualisation and annotation of the results followed the recommended guidelines.

Results

A summary of the methods and data generated

Of the initial 114 individuals with pLoFs and controls that were recalled, we successfully profiled 94 donors that passed viability checks after thawing and 16h of resting. A subset of samples from 40 donors have been additionally stimulated with two stimuli, aCD3 and IFN α , and profiled at two stimulation timepoints, 16 and 64h. Data generation was performed across two workflows, one with unstimulated profiling only (RESTONLY), and one with unstimulated and stimulated profiling (RESTSTIM), with 13 donors seeded across both workflows. This resulted in 267 samples (defined as a donor profile in a workflow, one stimulation condition and one timepoint) represented by 857,370 cells, with a mean of 3,211 cells per sample (Figure-4.14A). The cell counts were also evenly distributed across the workflows, stimulation conditions and timepoints. Crucially, the two workflows that generated unstimulated data resulted in similar mean cells/sample (2,902 and 3,023). This is reassuring as a balanced design is beneficial for batch correction (T. Li et al., 2023). There is on average about 1,000 more cells/sample in the 64h timepoint than in earlier timepoints. This is due to the experimental design - we initially hypothesised that at 64h, the cells would be less healthy than the cells at the 16h timepoint, therefore more cells per pool were intentionally loaded for sequencing (see the superpool design in methods). However, after QC, fortunately, we recovered more cells than anticipated. (Figure-4.14B)

After dimensionality reduction of the QC-ed dataset, we saw that the single-cell data has captured the highly heterogeneous expression profiles across these PBMCs, and the main drivers of variation were, as expected, cell-type and context (defined by a stimulation and timepoint) (Figure-4.14C).

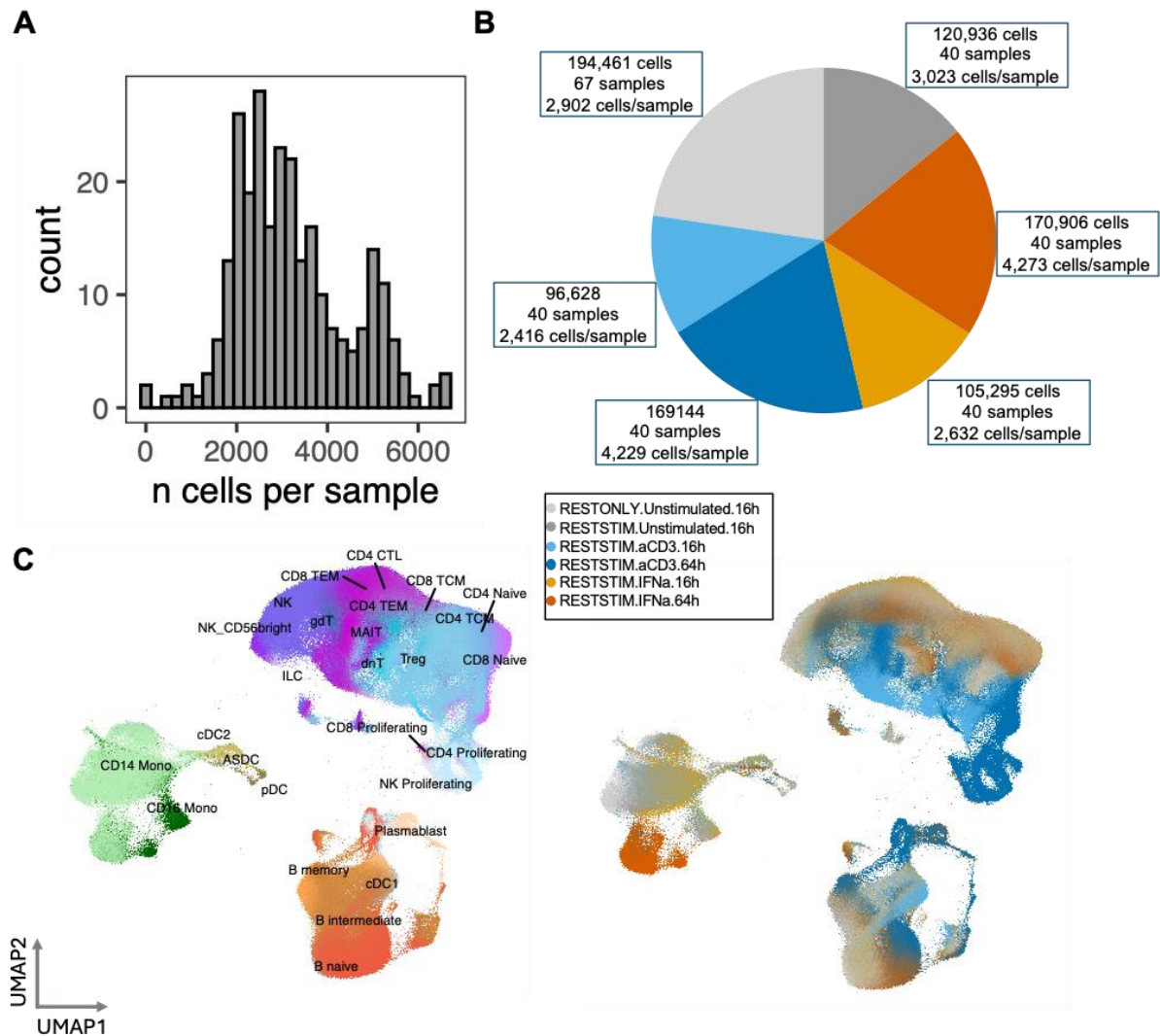


Figure-4.14

Summary statistics on the data generated. **A:** Distribution of the number of cells per sample across the 267 samples in the dataset. **B:** 857,370 cells after QC by workflow, divided by stimulation and timepoint. **C** - UMAP representation of the 857,370 cells after QC coloured by their azimuth level 2 cell-type annotations (left) and by their workflow, stimulus and timepoint (right).

Direct LoF effects demonstrate cell-type- and context-specificity

We aggregated gene expressions for each donor, context (stimulus and timepoint) and cell-type, and tested the target gene's expression against the pLoF genotype, controlling for age, sex and workflow.

At Azimuth level 1 which resulted in the assignment of seven cell-types, a total of 180 tests were successfully performed. Ten pLoFs, seven cell-types, five contexts, but not all pLoFs were stimulated, and not all cell-types had sufficient number of cells to be pseudobulked for testing. Of these 180 tests, 24 (13%) demonstrated knockdown effects (i.e. negative direction of effect) that passed a Bonferroni-corrected threshold for multiple testing, with 39 (22%) passing a Benjamini–Hochberg cutoff for a false discovery rate of 5% (FDR 5%), and 80 (44%) at nominal significance (p-value ≤ 0.05).

At a higher cell-type resolution (Azimuth level 2, 27 cell-types), we performed 499 tests, with 32 (6%) demonstrating knockdown effects passing the Bonferroni-corrected threshold, 75 (15%) passing FDR 5% and 185 (37%) at nominal significance. The lowered percentage of tests passing multiple testing correction could be attributed to more tests being performed. However, with higher cell-type resolutions, we might be able to narrow down more precisely which cell-types have the most apparent knockdown effects, and may potentially uncover effects that were missed at level 1 due to being masked when aggregating heterogeneous cells together at lower resolution. This is gene-specific, therefore in the individual gene subsections below, I dissect this cell-type-specificity further.

However, in general, at both cell-type resolutions, the proportion of significant tests increased in the presence of stimulation, highlighting the value of stimulation in profiling LoF effects. (Figure-4.15)

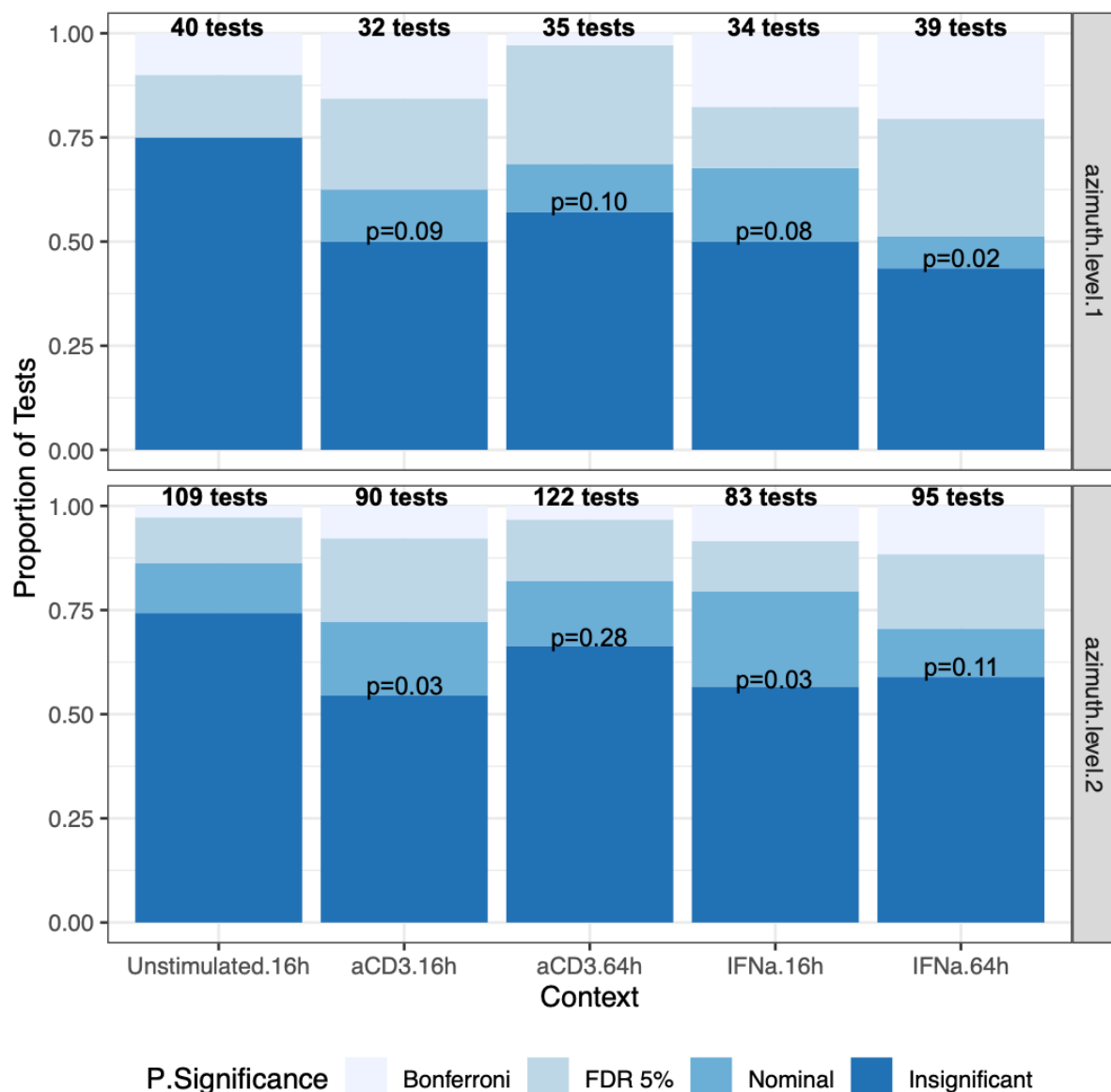


Figure-4.15

The proportion of significant tests when testing for direct knockdown effects of pLoFs on the gene they are predicted to affect. The p-values listed in the graph are from two-sided chi-squared tests comparing the proportion of nominally significant tests in a stimulated timepoint compared to the unstimulated subset.

Pathway derangements are associated with LoF effects

For each pLoF comparison (controls vs the pLoF), DGE testing was extended to the rest of the genes (after filtering out lowly-expressed genes). Significantly differentially expressed genes were defined as those that passed a p-value threshold corrected for multiple testing (Benjimini-Hochberg correction for a false discovery rate of 5%, for all the genes tested within each pLoF comparison (FDR 5%)). The distributions of the number of significantly differentially expressed genes in each pLoF comparison differed across cell-types and contexts. For example, most notably, in B cells and other T cells, more genes were differentially expressed at the 64h timepoints of stimulation compared to other contexts. Unexpectedly, fewer significantly differentially expressed genes were detected in monocytes with both stimulation. This could be driven by the comparisons performed for pLoFs affecting the PRRs, as the addition of IFN α (or in the case of aCD3 stimulation, the production of IFN α by other cell-types), a downstream product of these PRRs, may allow for compensation of the initial pathway derangements at rest. In addition, another possibility is that different selection of time points, or ideally, a proper time-course experiment, may reveal different pLoF effects. For example, IFN α may trigger a response in monocytes within the first few hours, or aCD3 may have effects on activated or proliferating T cells after a few days. (Figure-4.16A)

For each pLoF comparison, we tested a mean of ~322 KEGG pathways (minimum 319 pathways, maximum 327 pathways, tested by ClusterProfiler). We defined a significantly enriched pathway when the p-value passed the multiple testing Benjimini-Hochberg correction for a false discovery rate of 5%. Out of a total of 58,007 tests, we detected 6,078 significantly enriched pathways, highlighting pathway derangements associated with tested pLoFs. The dysregulated pathways are independent of whether the pLoF-carrier genes had a statistically significant direct knockdown or not - the rationale being that even if the RNA of the pLoF-affected gene did not show decreased expression, downstream protein degradation or a non-functional protein could lead to downstream pathway changes. We acknowledge that there may be an inflation in the number of pathways we identified, given the limitation in our chosen method in correcting for Type I error, as described in the “Gene set enrichment analysis” subsection of the Methods.

KEGG pathways are organised into categories with further subcategories, and as expected in the “Organismal System” category, the “Immune system” subcategory has the highest proportion of significantly enriched pathways. In “Human Diseases”, immune-related disease pathways (“Immune disease”, and the three “Infectious disease” subcategories) have more significantly enriched pathways. (Figure-4.16C)

Interestingly a high proportion of “Neurodegenerative disease” pathways tested are significantly enriched too (536 out of 1,260 neurodegenerative disease pathways tested were significant). This may be due to a narrow set of genes implicated in multiple overlapping neurodegenerative pathways: The “Neurodegenerative disease” category is only composed of seven disease pathways, with 559 unique genes implicated across the seven diseases, and during testing 77 (14%) of the 559 genes were enriched in at least half of the significant tests. In gene ontology (Aleksander et al., 2023), the 77 genes are involved in very general biological processes that are not neurology-specific, such as mitochondrial processes and protein degradation.

More generally, signalling pathways, energy metabolism and cell motility subcategories also have a high proportion of significant pathways, perhaps reflecting immune-related and activation-related perturbations.

The distributions of the number of significantly enriched pathways differed across cell-types and contexts. In the context of a lack of external stimulus, monocytes resulted in the highest number of significantly enriched pathways between controls and pLoF carriers compared to other cell-types (In monocytes, a median of 58.5 pathways test significant out of ~322 tested, compared to 20 to 46.5 in the other cell-types). However, this decreases with stimulation. In lymphocytes, the distribution of the number of significantly enriched pathways widens under different stimulation conditions, with some pLoF comparisons resulting in many more significantly enriched pathways in the presence of stimulation. Significantly enriched pathways by pLoFs will be elaborated in the individual gene subsections below. (Figure-4.16B)

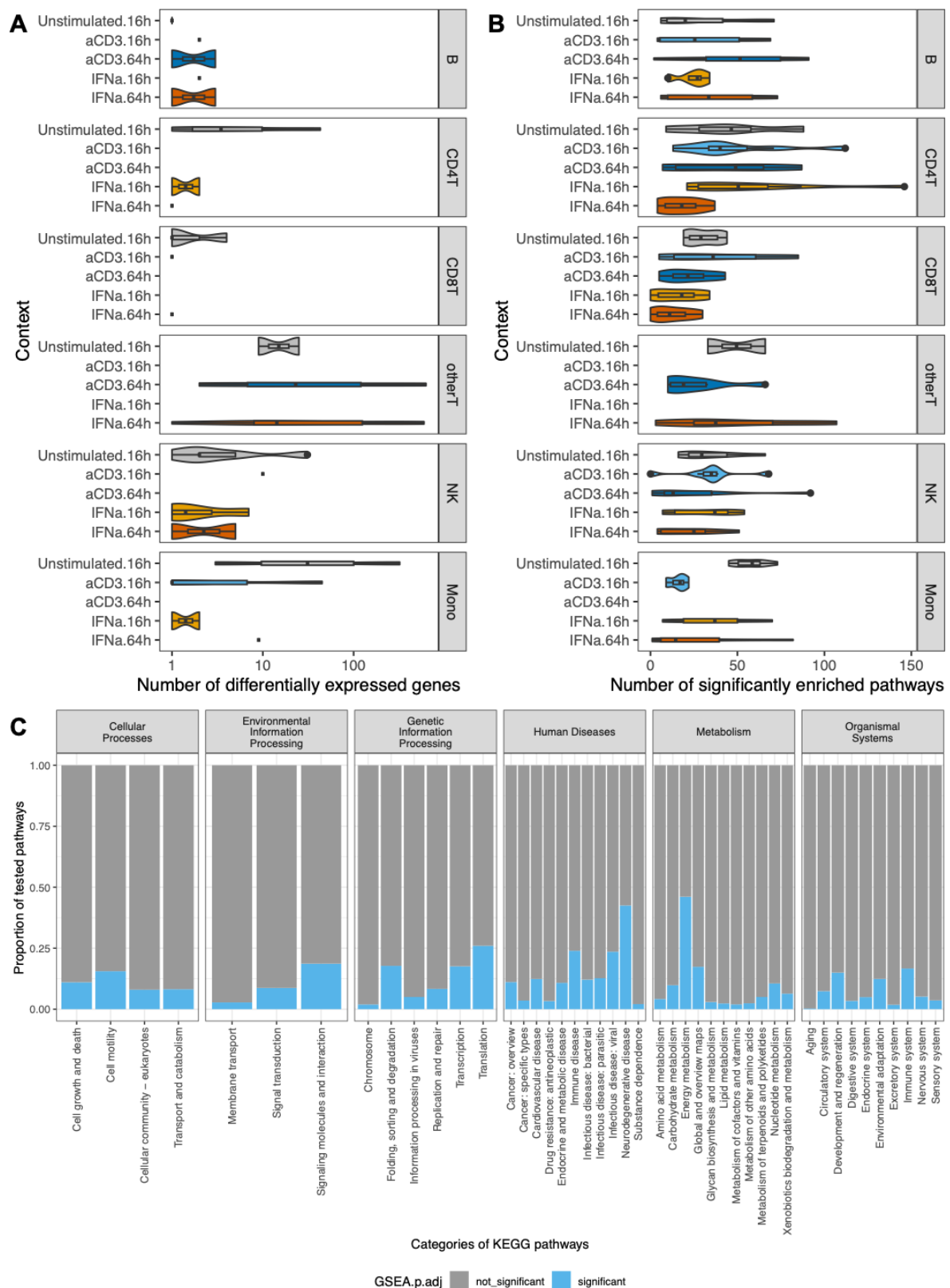


Figure-4.16

Significantly differentially expressed genes and pathways.

A, B - Distribution of significantly differentially expressed genes (**A**) or pathways (**B**) between pLoF carriers and controls, across cell-types and contexts. **C** - Proportion of significant pathways in tested pathways, sorted over KEGG categories. The actual numbers of significant pathways is visualised in Supplementary-figure-4.1.

Subtle differential abundances in cell-types are associated with pLoFs

We further hypothesised that LoF effects could present as differential cell-type proportions, particularly as several of these genes have been associated with blood cell counts. In reality, PBMCs are more likely to exist along a spectrum of cell-type differentiation, and as we had single-cell data, it was possible to test for more subtle differences at higher resolution. Milo is a statistical algorithm that leverages single cells plotted into KNN graphs to perform differential abundance testing in partially overlapping neighbourhoods of cells. For the ten pLoFs profiled in this chapter, eight underwent stimulation with two stimuli at two timepoints, together with the unstimulated data point, we had a total of 42 single-cell datasets to perform Milo analysis. This resulted in 129,488 partially overlapping neighbourhoods tested, with 521 (4/1000 tests) identified as significantly differentially abundant. Because of the lack of independence between overlapping neighbourhoods, MiloR calculates a modified multiple testing correction for the p-value taking into account the spatial overlap between tested neighbourhoods, called spatial false discovery rate, or spatial FDR. We defined a significantly different abundant neighbourhood as one with a spatial FDR lower than 5%. The neighbourhoods with significant differential abundances were then referred back to their Azimuth annotation to infer cell-types.

We calculated the proportion of tests that came back with a statistically significant differential abundance between pLoF carriers and controls, broken down by cell-type and context (Figure-4.17). Indeed, the effects are very small, with the cell-type with the largest number of significant differential abundances in neighbourhoods identified for CD14⁺ monocytes, at 0.11 proportion of tests, followed by CD16⁺ monocytes (0.07 proportion of tests), both in IFN α stimulation at the 64h timepoint. In MiloR each pLoF is tested across multiple neighbourhoods, and we find that this signal is mainly driven by the *MNDA* pLoF, which we will discuss later.

We found that CD4 T cells (specifically CD4 central memory T cells, 'CD4 TCM') and CD8 T cells (specifically effector memory T cells, 'CD8 TEM') demonstrated several neighbourhoods that tested significantly for differential abundance. This persisted in the same direction of effect (depleted in controls) across nine of the ten pLoFs (all except the pLoF in *IFIH1*) tested (Figure-4.17). It is possible that the T cell neighbourhoods may have formed similarly across multiple pLoF tests as the genes carrying the pLoFs are all relevant to T cell activation. However, as the control individuals are the same throughout the different pLoF testing, it is also possible that this is a systematic depletion of CD4 TCM and CD8 TEM in unstimulated controls, and not a feature of pLoFs. These results need to be treated with

caution, as these are very subtle differential abundances, and could be driven by technical aspects.

In light of this, we have been conservative with our interpretation of the differential abundance testing results. The first conclusion is that true LoF effects on cell-type proportions, if any, would be subtle. By taking into account other supporting evidence, we were convinced by only two differential abundance findings, which we described in the next sections. Of note, both findings were detected only in the presence of stimulation.

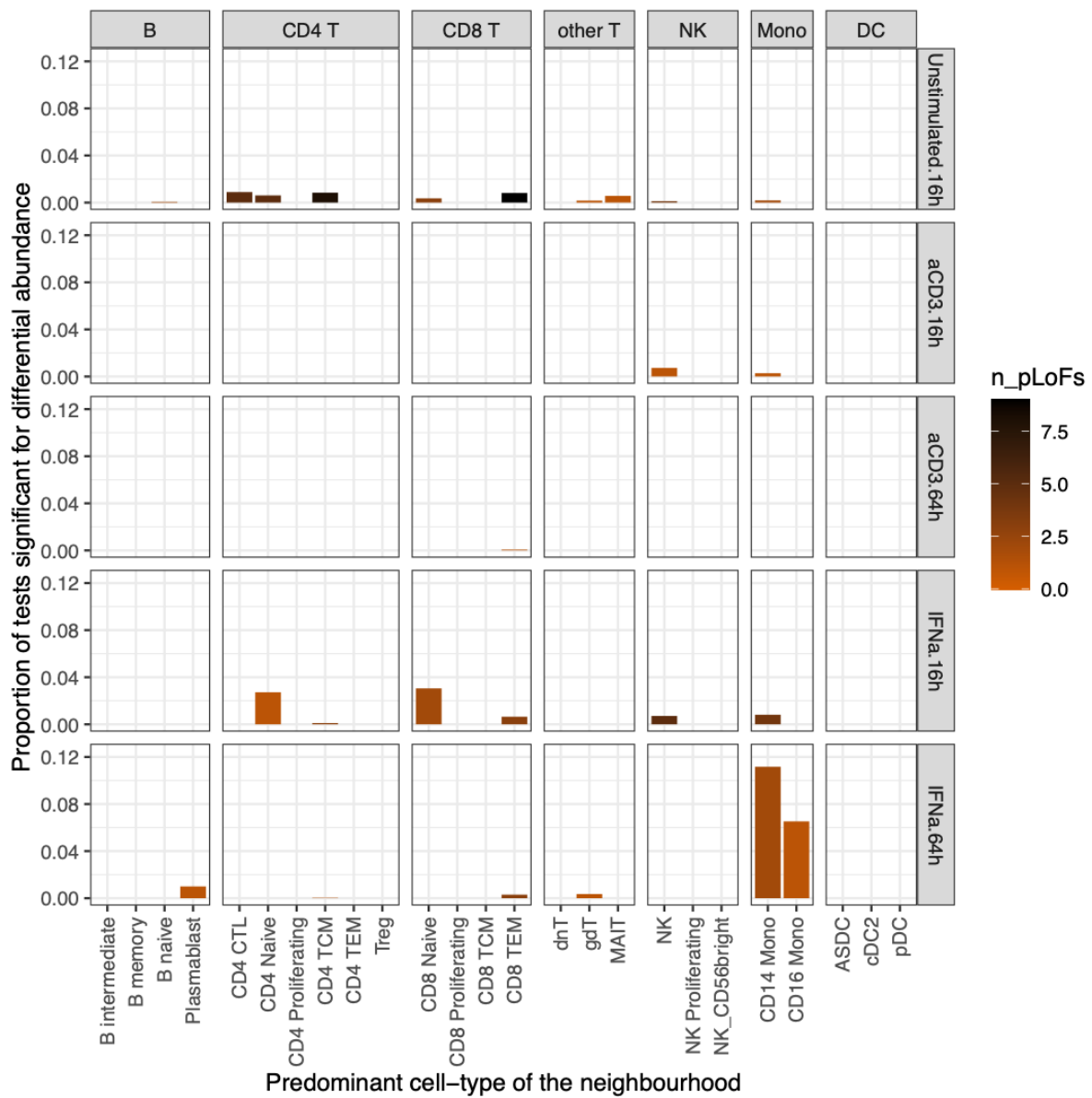


Figure-4.17

Proportion of tests significant for differential abundance, across cell-types and contexts. Within a cell-type and context, a test for differential abundance is performed between carriers of a pLoFs and controls.

Systematic characterisation of LoF effects

Combining all of the analyses together, we systematically characterise pLoF effects in nine genes tested.

TRAT1

Congruent with the literature, *TRAT1* was specifically expressed by T cells in our dataset. At rest, the direct knockdown effect of the pLoF was only significant in CD4 and CD8 T cells. Upon stimulation, *TRAT1* expression increased across multiple lymphocytes, and similarly, more cell-types demonstrated significant knockdown effects. These knockdown effects apply non-discriminately across CD4 and CD8 subtypes. However, an interesting observation at this resolution is that proliferating T cells (note: in our dataset, T cell proliferating cells primarily exist only after aCD3 stimulation, Figure-4.14C) experience significant knockdown effects only at the 64h timepoint. (Figure-4.18)

From the DGE analyses run on level 1 cell-types and all stimulation contexts, 5805 KEGG pathways were tested via GSEA, and 497 pathways were significantly enriched. We first focused on the PI3-AKT signalling pathway in T cells. As introduced earlier, it has been reported that *TRAT1* induces this pathway by binding to *PI3K* upon activation of the TCR. Specifically, the bound *TRAT1-PI3K* complex converts phosphatidylinositol (4,5)-bisphosphate (PIP2) to phosphatidylinositol (3,4,5)-trisphosphate (PIP3), PIP3 is then the substrate that triggers *AKT* expression. (Garapati et al., 2008) *AKT* expression is then involved in the inhibition of glycolysis/gluconeogenesis, apoptosis, and the NFkB and p53 signalling pathways; It is also involved in regulating cell cycle progression (Kanehisa & Goto, 2000). In our data, we see that the PI3-AKT pathway is significantly upregulated in the T cells of *TRAT1* pLoF carriers compared to controls, at rest and in the presence of aCD3 stimulation. (Figure-4.20)

As parts of the pathway result in activation and other parts inhibition, the result of overall upregulation of the pathway has to be interpreted separately for these in parts. We used the pathview R package (Luo et al., 2017) to plot the differential expression of the involved genes along the pathway, in CD4 T cells across three contexts: At rest, we see that the genes upstream of the PIP3 substrate are upregulated in pLoF carriers, perhaps as a compensatory mechanism for the *TRAT1* knockdown (Figure-4.19A). At 16h of aCD3 T cell activation, these genes before and after the PIP3 substrate have reduced expressions relative to controls. As a consequence, downstream genes inhibited by *AKT* become upregulated. This could be a result of inadequate compensation under the stress of stimulation. (Figure-4.19B) At 64h, these effects are diminished, suggesting that the *TRAT1*

knockdown effect on the PI3-AKT pathway affects the early phase of T cell activation more than the later phase. (Figure-4.16C) This highlights the importance of profiling across timepoints.

Since the PI3-AKT signalling pathway regulates multiple downstream pathways, we looked at the consequence of *TRAF1* knockdown on these pathways. Indeed we see significant enrichment (direction: upregulation) of glycolysis/gluconeogenesis, the NFkB signalling pathway and the cell cycle pathway in pLoF carriers. (Figure-4.20)

Many of the patterns of pathway derangements we saw in CD4 T cells were observed in CD8 T cells, other T cells, and interestingly, NK cells. (Figure-4.20) Furthermore, the *TRAF1* pLoF is also associated with one of the two convincing differential abundance results we detected. We found that a subset of pLoF-carrying NK cells were clustering away, and these neighbourhoods were enriched for mitotic NK cells. This further supports that cell cycle disruption is associated with *TRAF1* knockdown, possibly through disruption to the PI3K-AKT signalling pathway. (Figure-18)

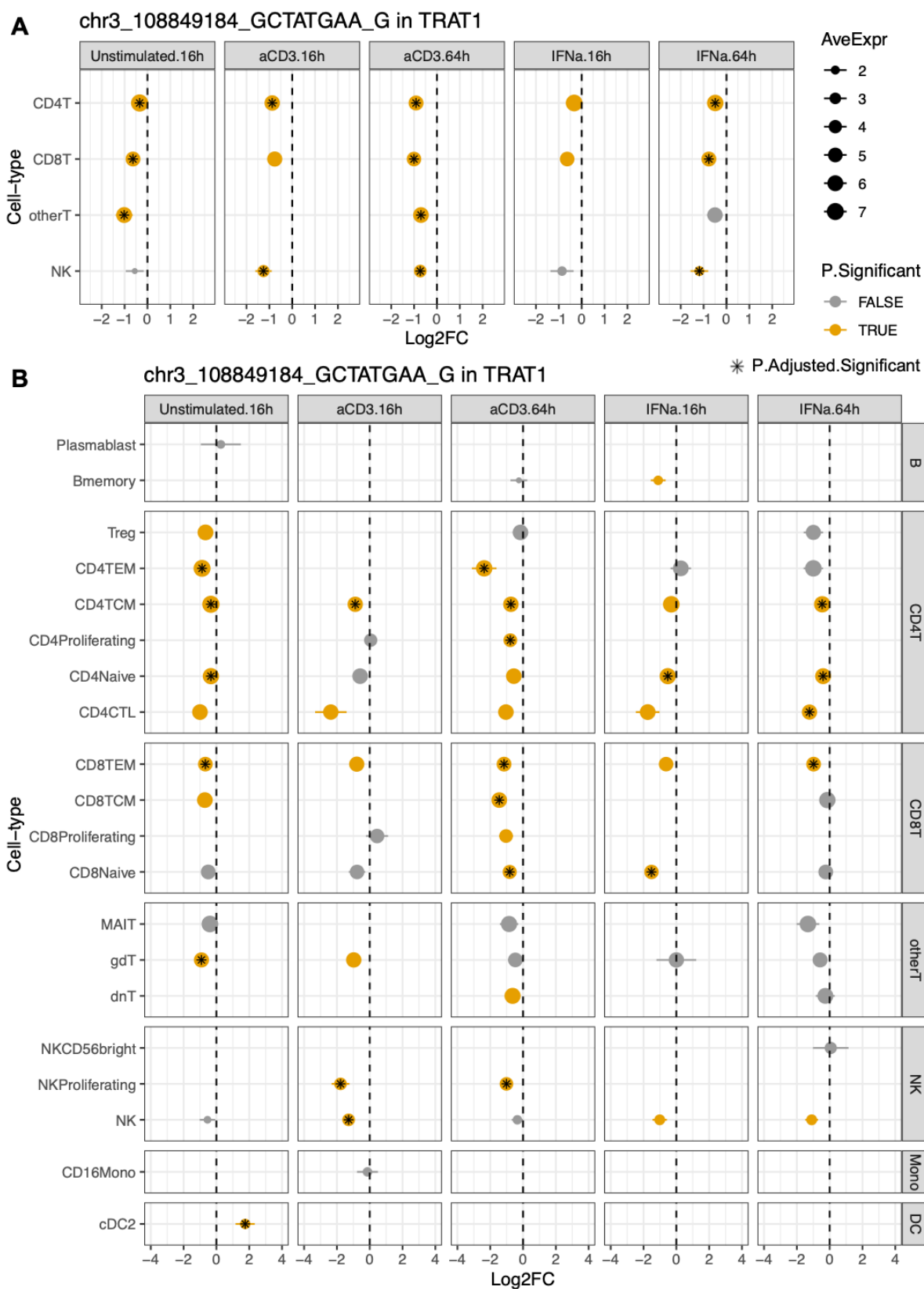


Figure-4.18

Differential expression of *TRAT1* in chr3_108849184_GCTATGAA_G carriers

compared to controls. A - Cells were pseudobulked at azimuth level 1. Only samples with at least 50 cells of that cell-type were included, and DGE was performed after filtering out lowly-expressed genes. Data points presented here represented successful tests. **B** - Cells were pseudobulked at azimuth level 2, excluding samples with less than 10 cells of that cell-type. “AveExpr” refers to the average expression calculated by limma-voom. “P.Significant” is defined as the nominal p-value ≤ 0.05 . “P.Adjusted.Significant” defined as a false discovery rate $\leq 5\%$ when correcting for multiple testing across the number of pLoFs tested, cell-types, and contexts.

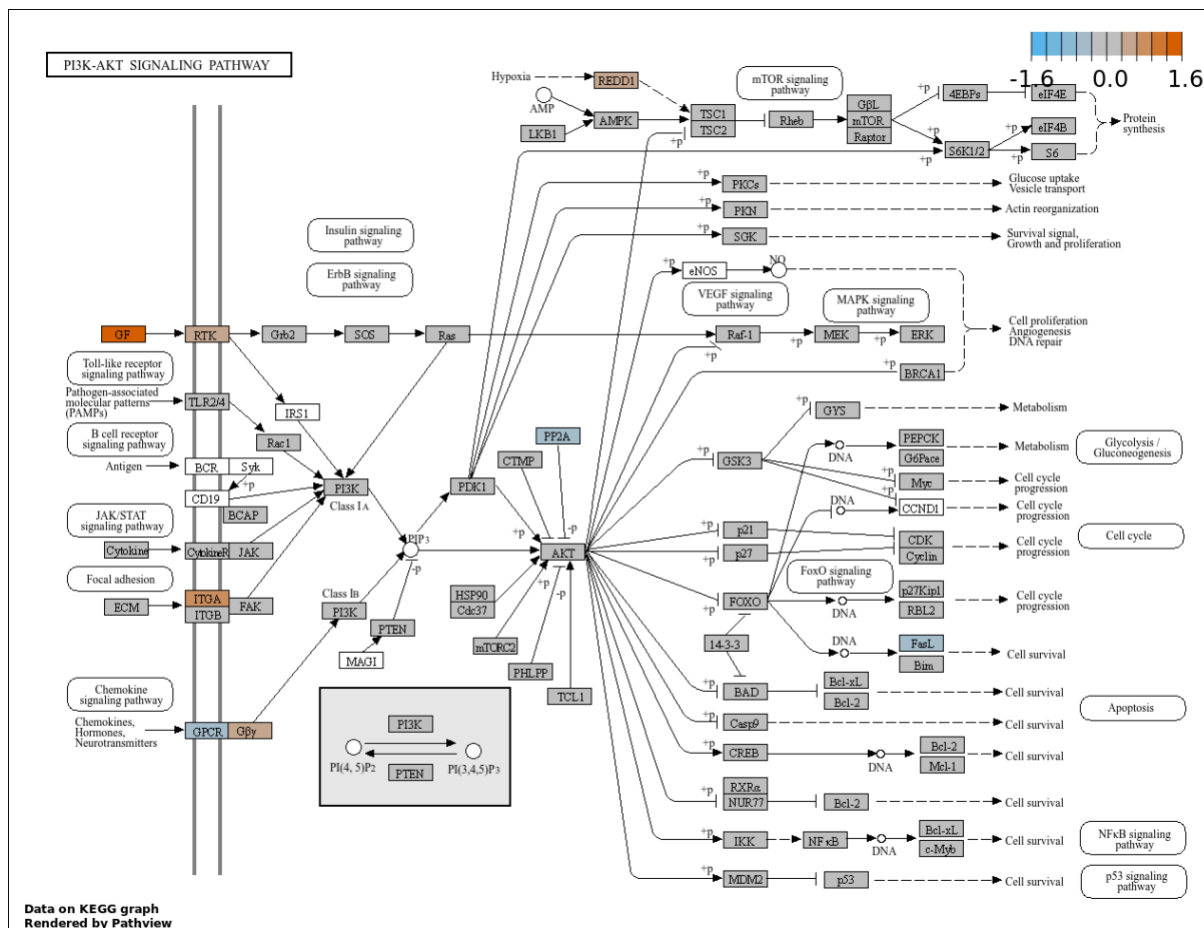


Figure-4.19A

The PI3-AKT signalling pathway, with genes coloured by their log2-fold-change in pLoF carriers compared to controls. A: In CD4 T cells, unstimulated.

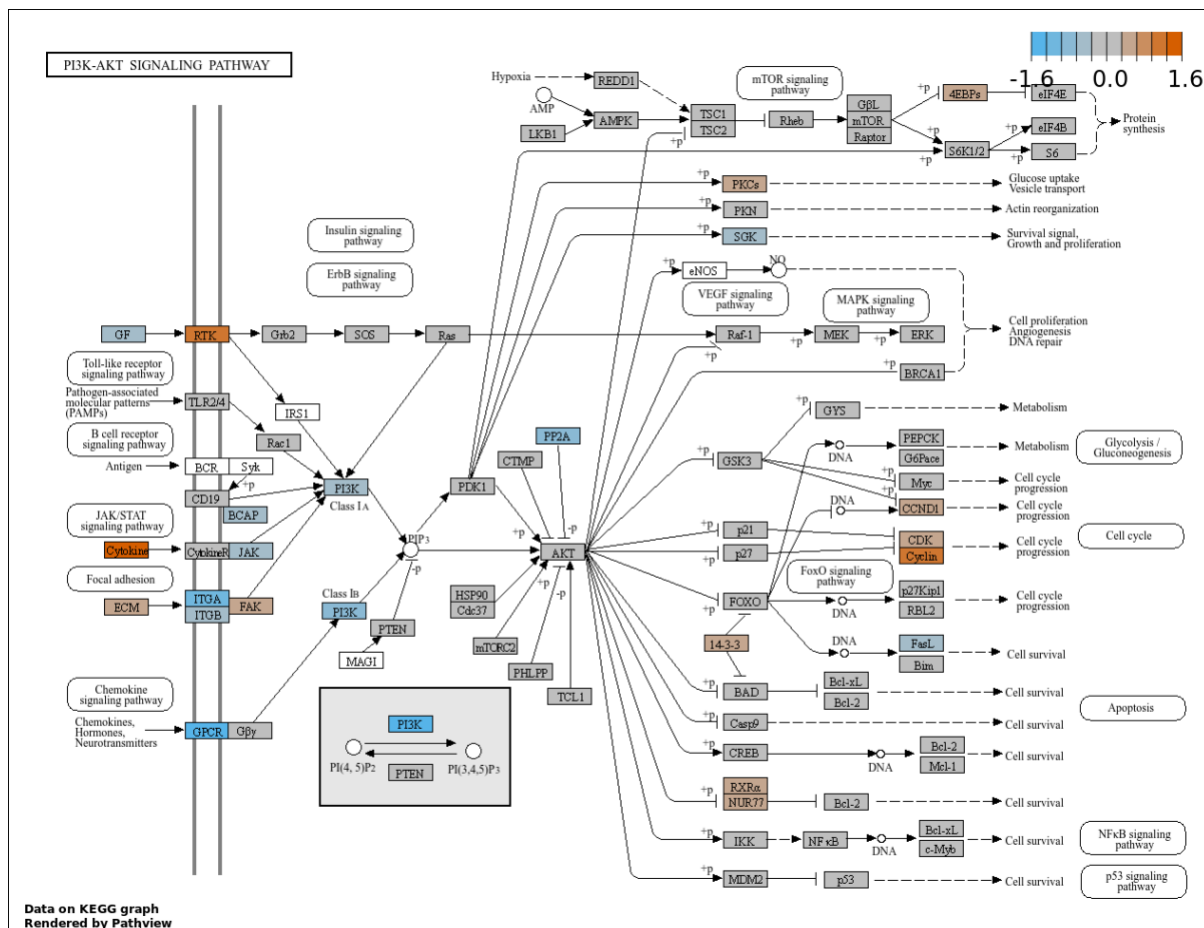


Figure-4.19B

The PI3-AKT signalling pathway, with genes coloured by their log2-fold-change in pLoF carriers compared to controls. **B:** In CD4 T cells, stimulated with aCD3 at 16h.

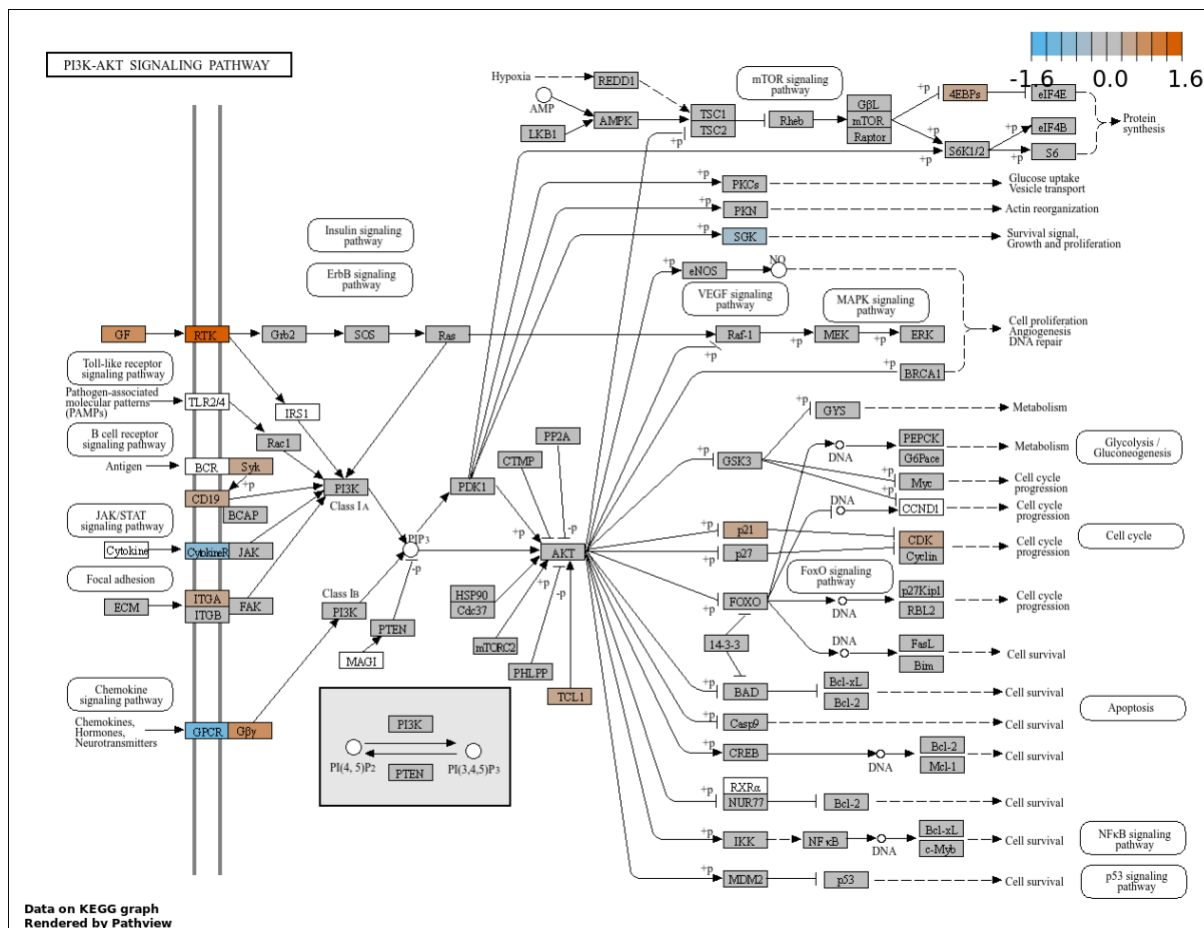


Figure-4.19C

The PI3-AKT signalling pathway, with genes coloured by their log2-fold-change in pLoF carriers compared to controls. C: In CD4 T cells, stimulated with aCD3 at 64h.

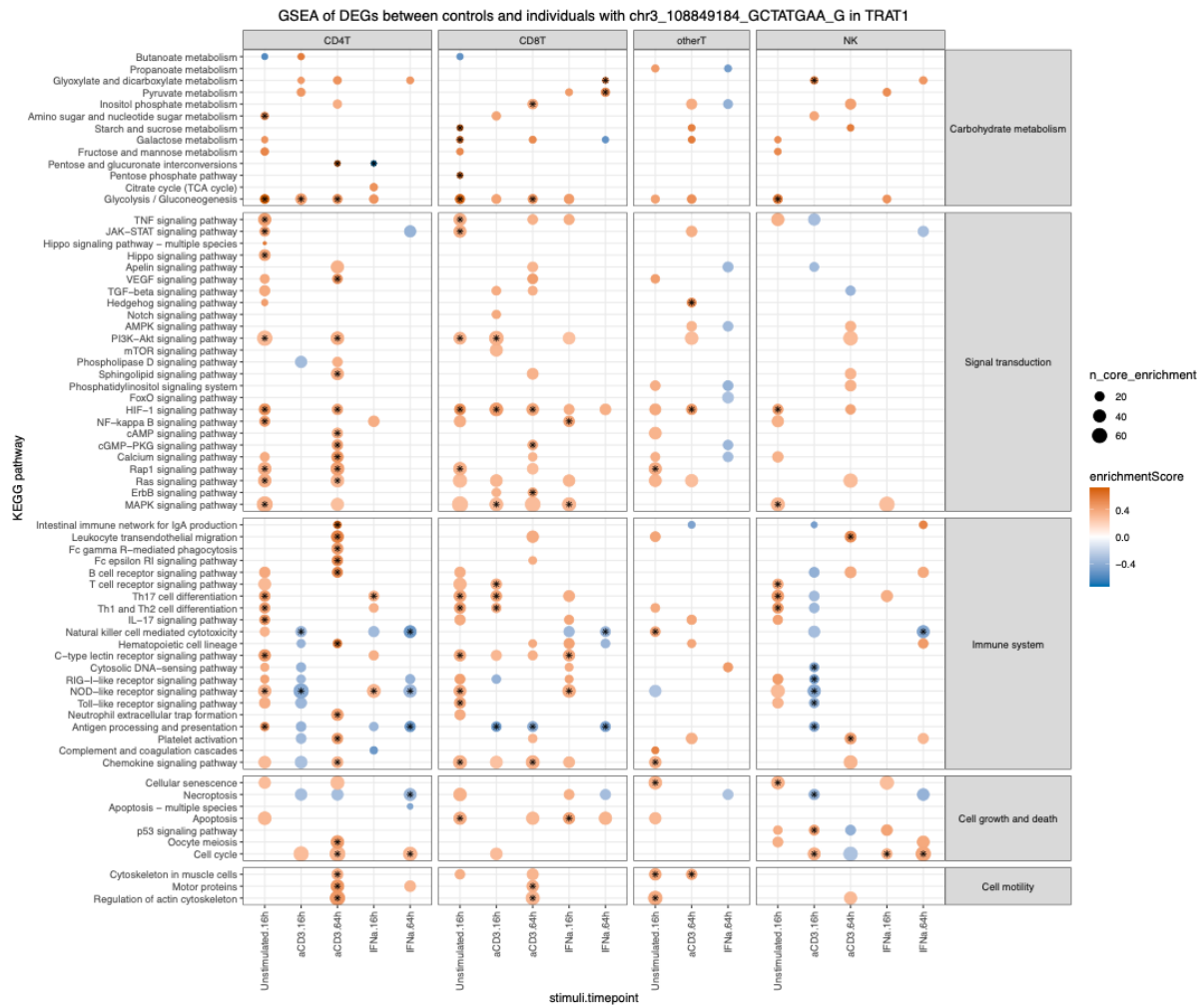


Figure-4.20

A selected subset of pathways dysregulated across cell-types and contexts in *TRAT1* pLoF carriers compared to controls.

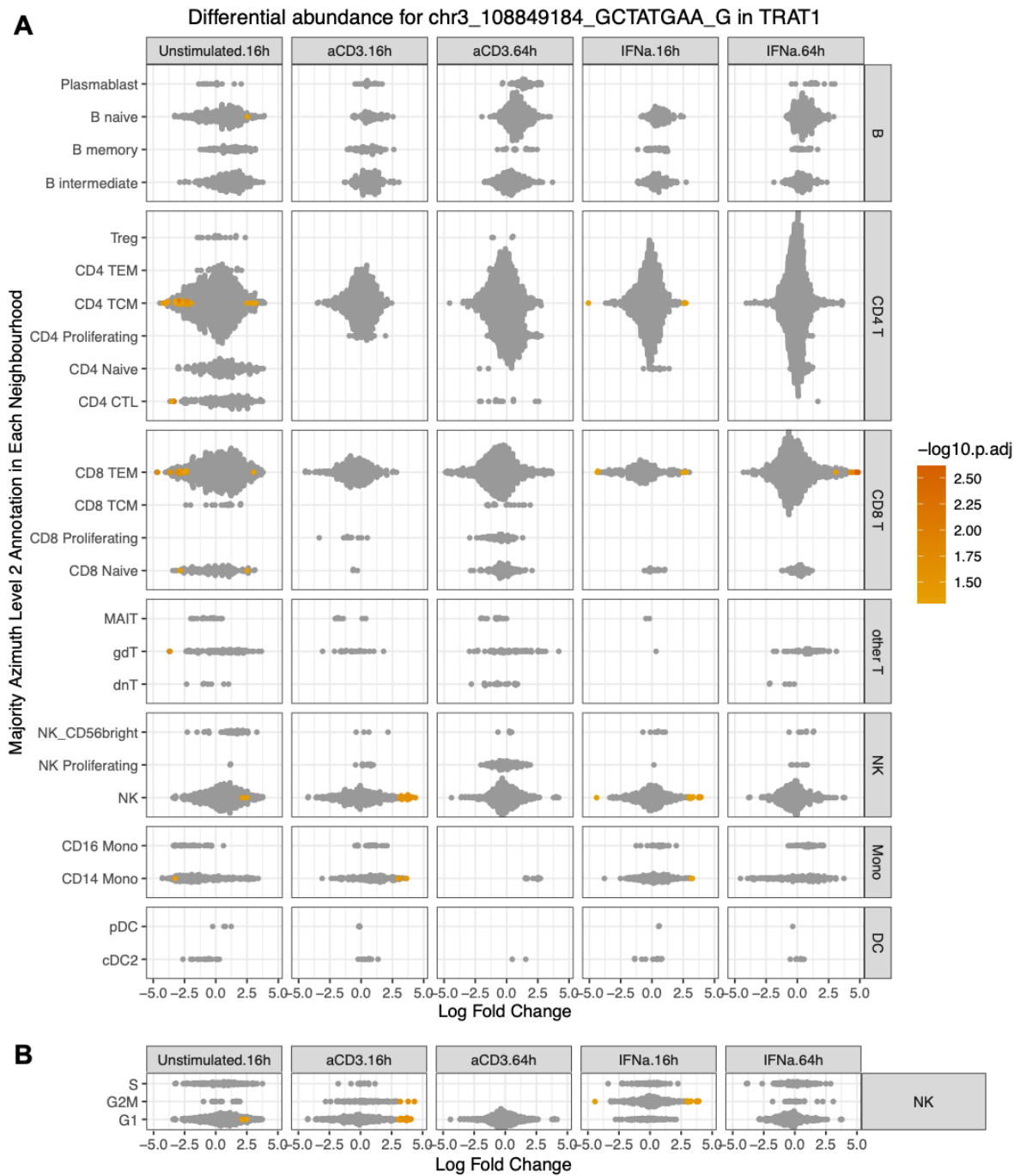


Figure-4.21

Testing for differential abundance of *TRAT1* pLoF carriers and controls across single-cell neighbourhoods, coloured by the SpatiaFDR ($-\log_{10}p.\text{adj}$). **A:** Neighbourhoods plotted by cell-type and context. **B:** The NK cell neighbourhoods are further plotted by cell cycle state.

MNDA

We see that in our dataset, as reported in the literature, *MNDA* is selectively expressed by B cells and monocytes. Expression in B cells appeared to be constant regardless of context, and similarly, the knockdown effect of the pLoF on *MNDA* was also significant across contexts. In contrast, *MNDA* expression increased in monocytes with IFN α stimulation, and the knockout effect became significant towards the later stage of stimulation. (Figure-4.22A)

Interestingly, we see upregulation of the PRR signalling pathways (RIG-I, NOD, Toll, cytosolic DNA sensing), all of which produce interferons, as well as the JAK-STAT signalling pathway, the key pathway that processes and regulates interferon. It is possible these are compensatory mechanisms to maintain interferon levels in the presence of *MNDA* knockdown. *MNDA* has also been reported to inhibit apoptosis, and indeed we see that *MNDA* pLoF containing cells tended to have more expression of the genes from the apoptosis pathway (Figure-4.22B)

In differential abundance analysis, a subset of pLoF-carrying monocytes, CD14 in particular, clustered away from control monocytes. We suspect that this is related to the difference in expression profiles of the signalling pathways described above. (Figure-4.23)

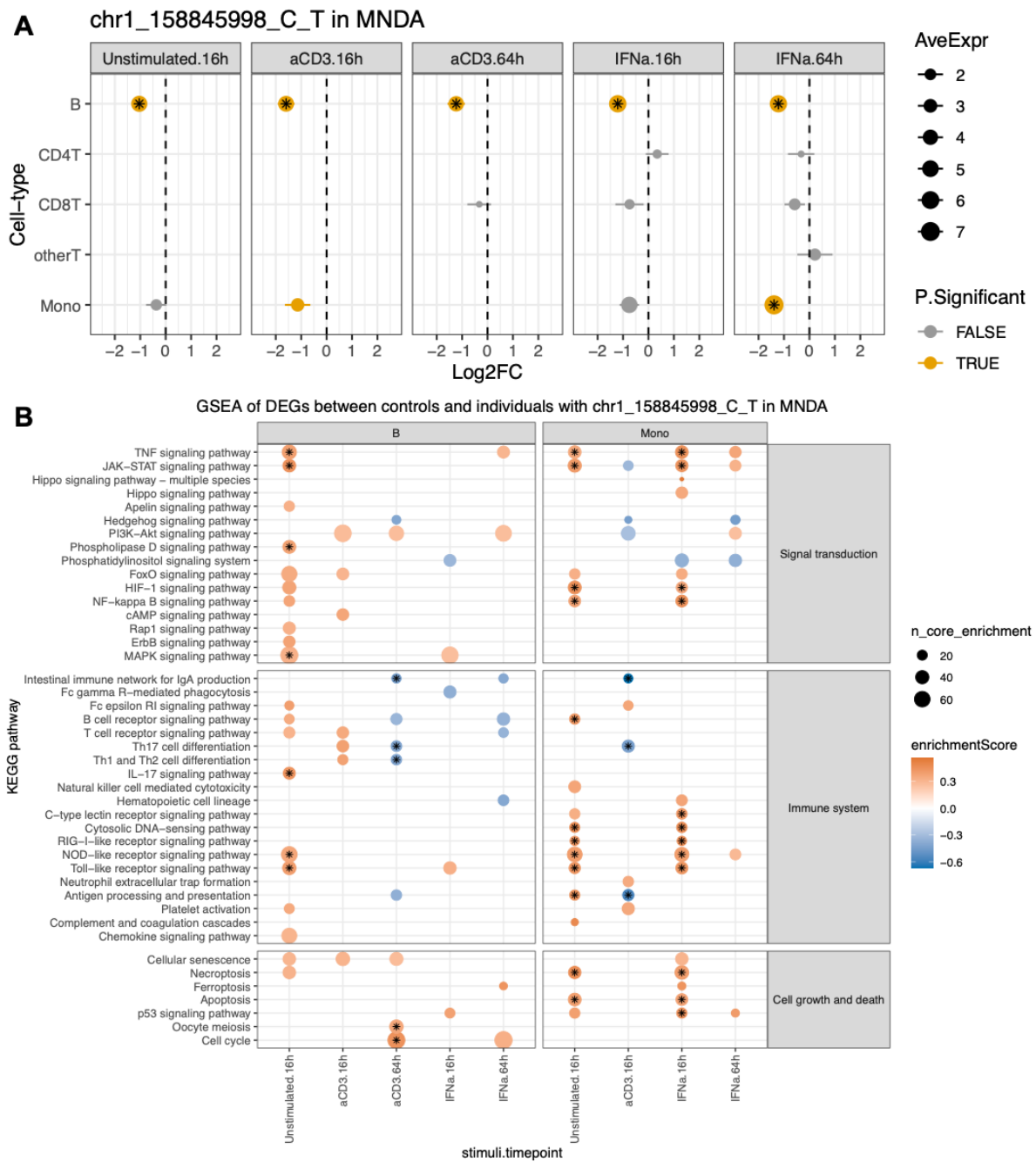


Figure-4.22

Effects of an *MNDA* pLoF. A: Differential expression of *MNDA* in *MNDA* pLoF carriers compared controls. **B:** A selected subset of pathways dysregulated across cell-types and contexts in *MNDA* pLoF carriers compared to controls.

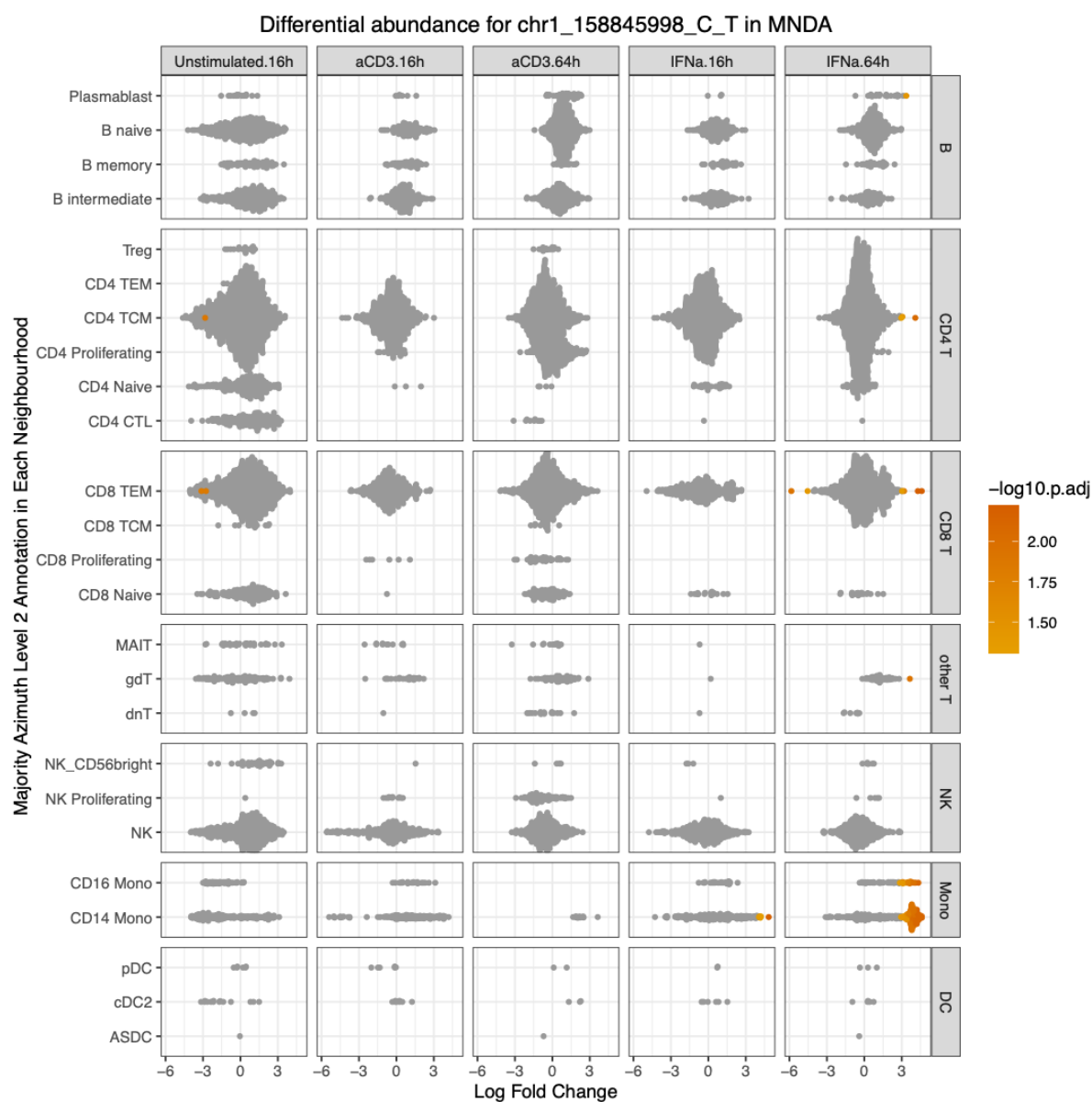


Figure-4.23

Testing for differential abundance of *MNDA* pLoF carriers and controls across single-cell neighbourhoods, coloured by the SpatiaFDR ($-\log_{10} p_{\text{adj}}$).

DDX58 and IFIH1

As *DDX58* and *IFIH1* are both PRRs that work similarly (acted within the RIG-I-like signalling pathway), we profiled them together. Both genes had good expression across multiple PBMC cell-types, but upon stimulation, especially with IFN α , expressions of both genes increased, and the knockdown effects also became more pronounced. (Figure-4.24)

The RIG-I-like signalling pathway and its downstream associated pathways have been plotted in Figure-4.25. We decided not to jointly analyse the two pLoFs affecting *DDX58* together because we wanted to use the two variants as “independent replicates” of each other, along with the pLoF affecting *IFIH1*. Indeed, we see that the three pLoFs regulate pathways in a similar fashion. Broadly speaking, the RIG-I-like signalling pathway and the associated downstream pathways seemed to be downregulated in pLoF-carrying lymphocytes, but upregulated in monocytes. This could reflect the inability of lymphocytes to compensate for the PRR knockouts, but monocytes, which have antigen-presentation as one of its main functions, may have sufficient compensatory mechanisms.

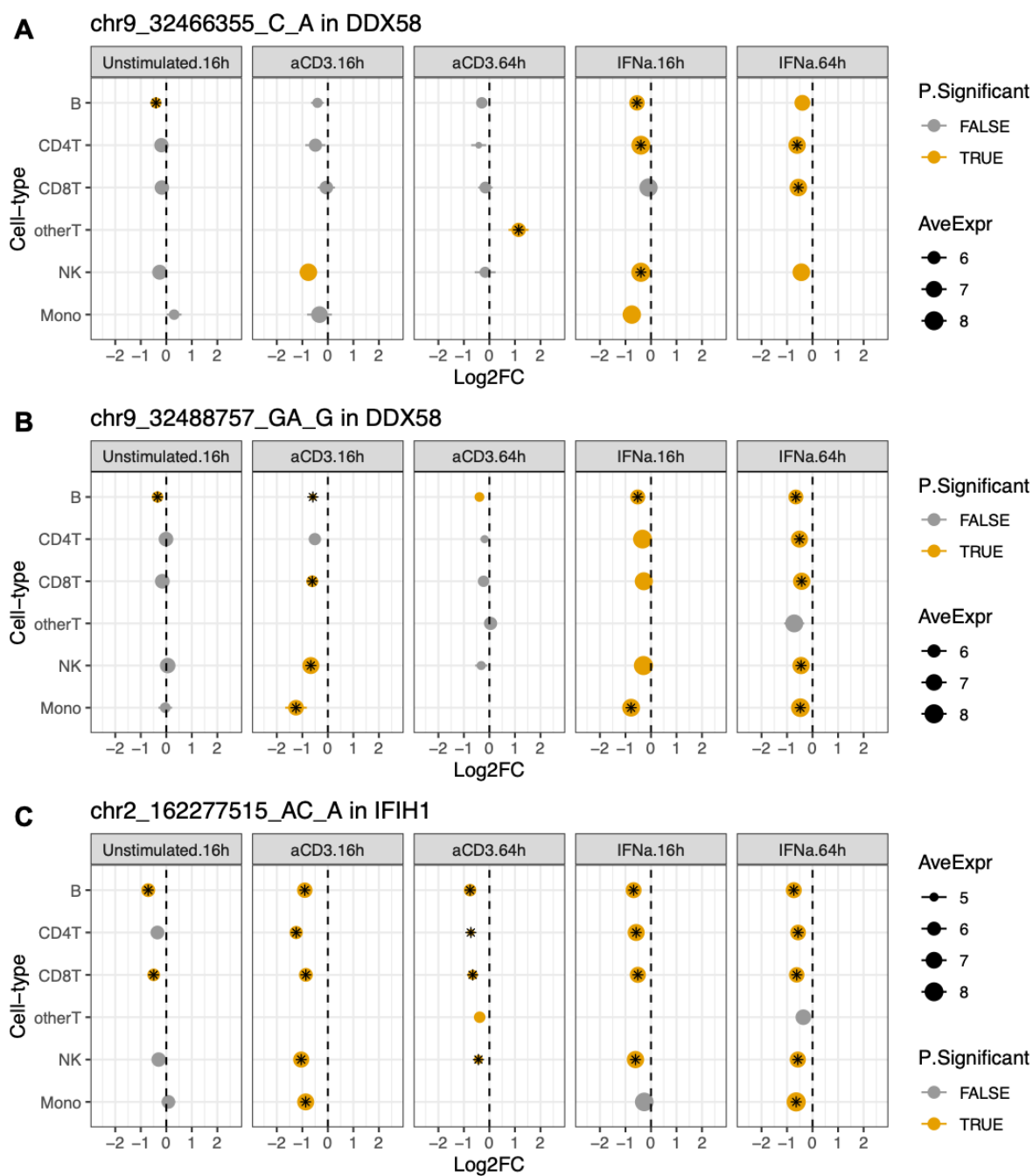


Figure-4.24

Direct knockdown effects of pLoFs in *DDX58* and *IFIH1*. **A:** Differential expression of *DDX58* in chr9_32466355_C_A carriers compared to controls. **B:** Differential expression of *DDX58* in chr9_32488757_GA_G carriers compared to controls. **C:** Differential expression of *IFIH1* in chr2_162277515_AC_A carriers compared to controls.

HPSE

HPSE knockout effects seemed to be very cell-type- and context-specific, with the most significant direct knockdowns occurring within aCD3-stimulated T cells, notably proliferating CD4 and CD8 subsets. Given its reported role in helping activated T cells extravasate into their target sites, this is not surprising (Figure-4.26). Indeed we see that *HPSE* knockdown is associated with dysregulation of T cell differentiation, transport, and cell mobility pathways (Figure-4.27).

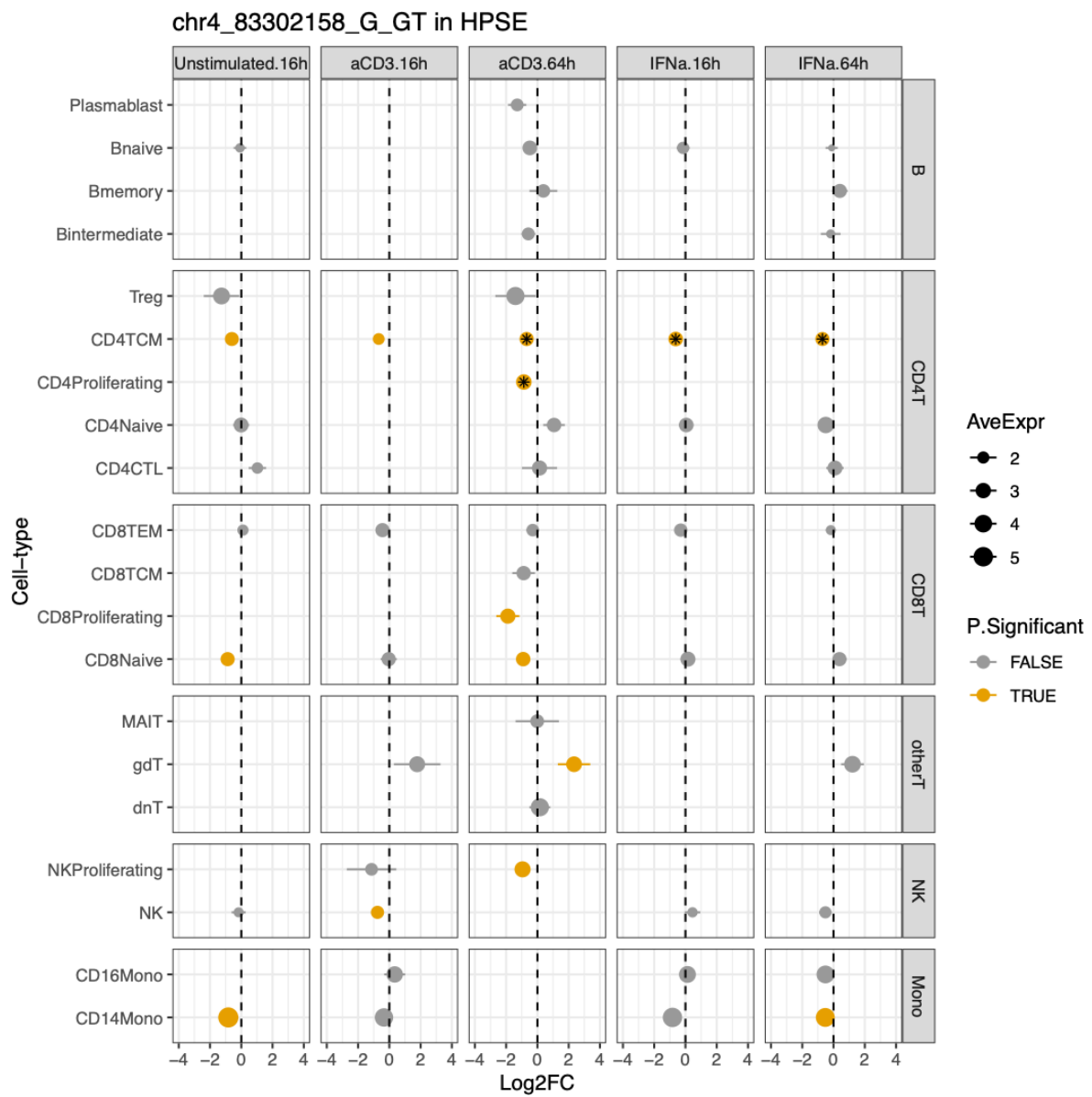


Figure-4.26

Differential expression of *HPSE* in pLoF carriers compared to controls.

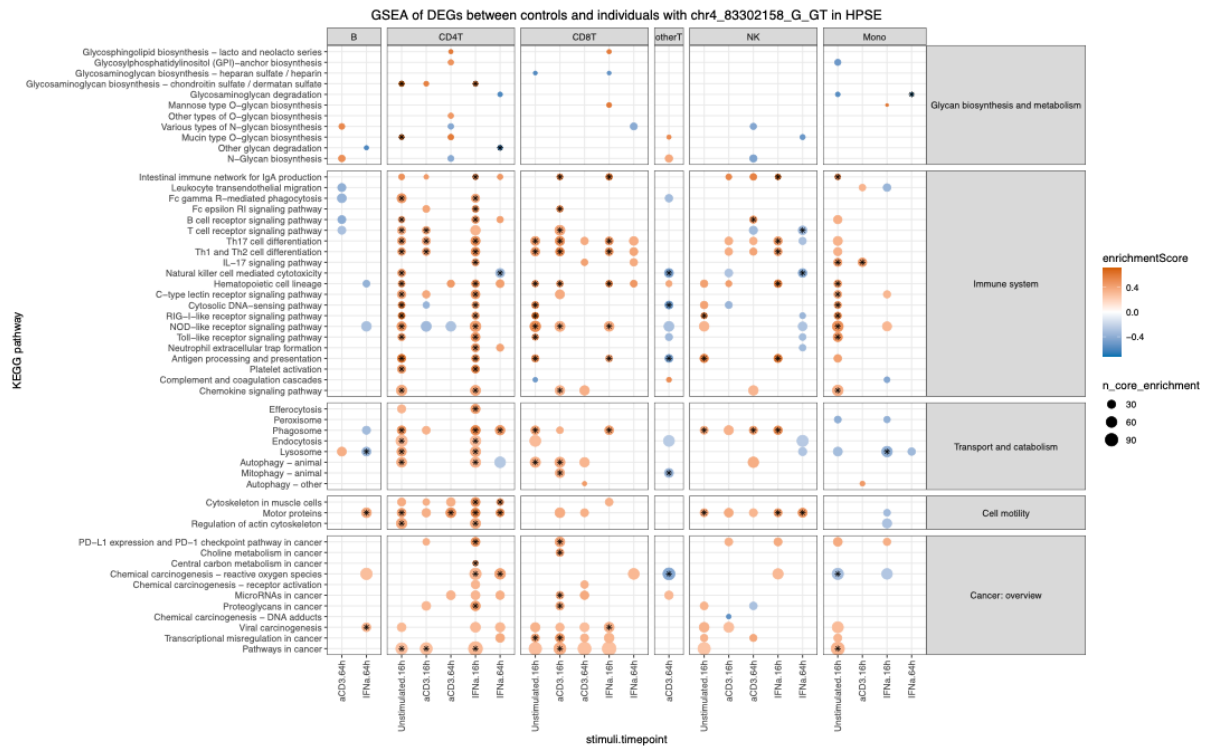


Figure-4.27

A selected subset of pathways dysregulated across cell-types and contexts in *HPSE* pLoF carriers compared to controls.

CD109 and TMEM71

Despite high expression of *CD109* and *TMEM71* across multiple cell-types and contexts, we did not detect many convincing knockout effects in pLoF carriers. This lack of direct pLoF effects was also observed in the Cardinal data in the previous chapter, and our hypothesis was that we were still detecting transcript expression of these genes due to their large sizes. We also hypothesised that LoF effects would become more apparent with stimulation. Indeed we see that *TMEM71* demonstrated significant knockout effects in stimulated B cells at 64h. However, overall, due to the limited evidence of knockouts, we did not proceed further. (Figure-4.28)

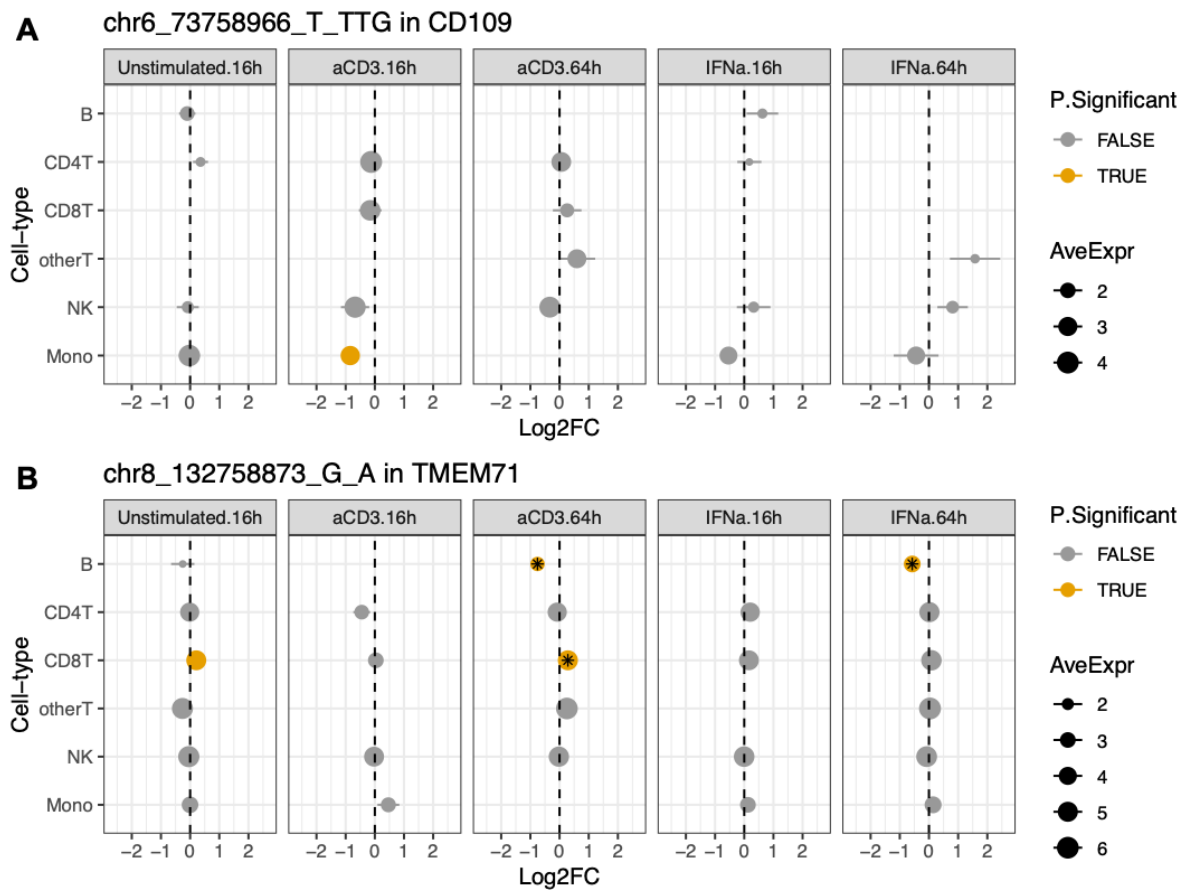


Figure-4.28

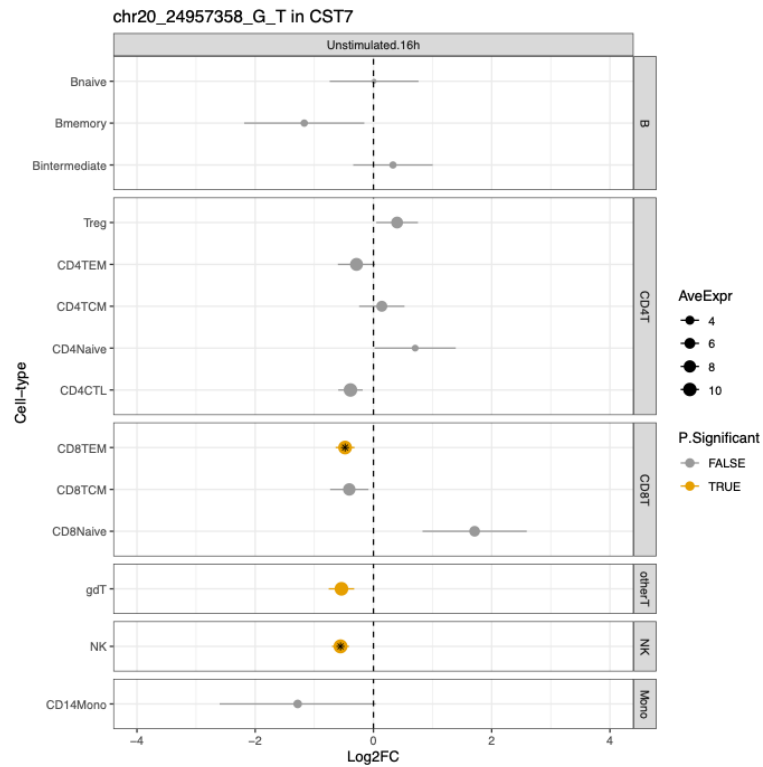
Minimal direct knockdown effects of pLoFs in *CD109* and *TMEM71*. **A:** Differential expression of *CD109* in chr6_73758966_T_TTG carriers compared to controls. **B:** Differential expression of *TMEM71* in chr8_132758873_G_A carriers compared to controls.

CST7

CST7 was not stimulated because we only had three heterozygous individuals (no homozygotes), and expression was high across cell-types at rest. In Cardinal we found that knockout effects were most noticeable in CD8 T cells, other T cells and NK cells. In this extended study, more cells per sample allowed us to profile these cell-types at higher resolutions, and we saw that these knockout effects aligned with the Cardinal data, but also further narrowed down to CD8 effector memory T cells, gamma-delta T cells and NK cells. (Figure-4.29A)

This cell-type-specificity was expected, given reports of the role of *CST7* in NK - and CD8 T cell - regulated cytotoxicity. As expected the knockout was associated with dysregulation in immune cell differentiation and upregulation of apoptosis. Interestingly, the knockout was also associated with upregulation of several PRR sensing pathways. (Figure-4.29B)

A



B

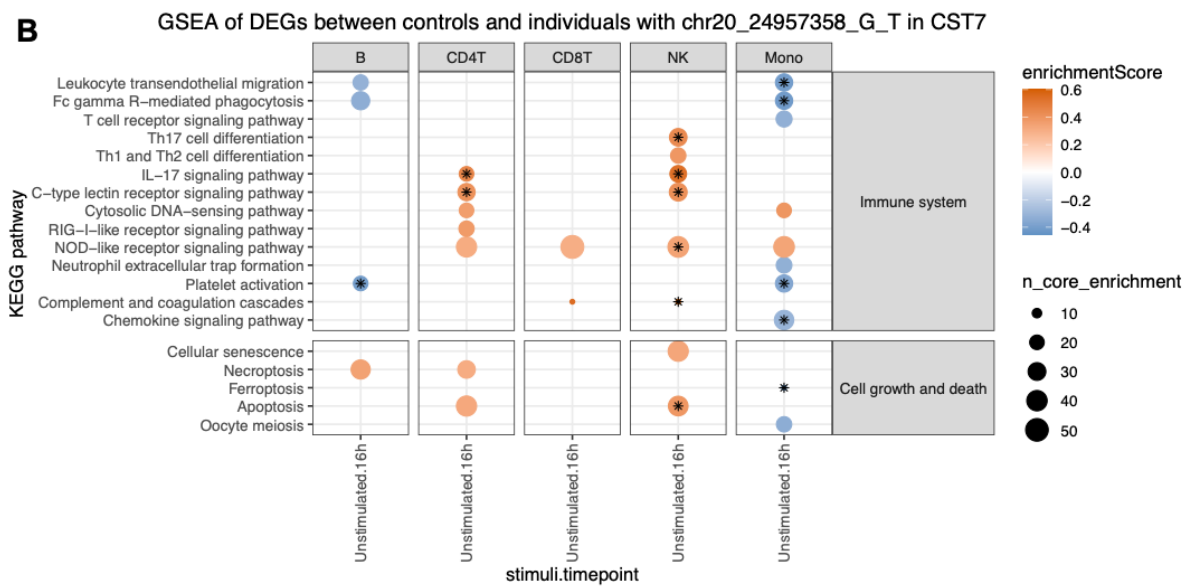


Figure-4.29

Effects of the *CST7* pLoF. A: Differential expression of *CST7* in pLoF carriers compared to controls. **B:** A selected subset of pathways dysregulated across cell-types and contexts in *CST7* pLoF carriers compared to controls.

TRIM58

Despite initial power calculations in publicly available dataset suggesting that *TRIM58* would have sufficient expression in PBMCs to be studied, unfortunately, expression was low in our dataset, with the gene being removed during filtering out lowly-expressed genes. Several possible reasons for this low expression include: the power calculations that we performed in donors from a different ancestry did not relate exactly to the British South Asian population, a technical aspect affecting banking and thawing steps of our protocol leading to poor *TRIM58* expression, low capture efficiency of *TRIM58* in our experiments or too strict filtering of gene expression. Upon plotting the distribution of the normalised counts at the single-cell level, we see that indeed, the expression of *TRIM58* is very low, with a slightly higher expression in B cells and a nominally significant knockdown effect found there. However, to be conservative, we have excluded *TRIM58* from further analysis. (Figure-4.30)

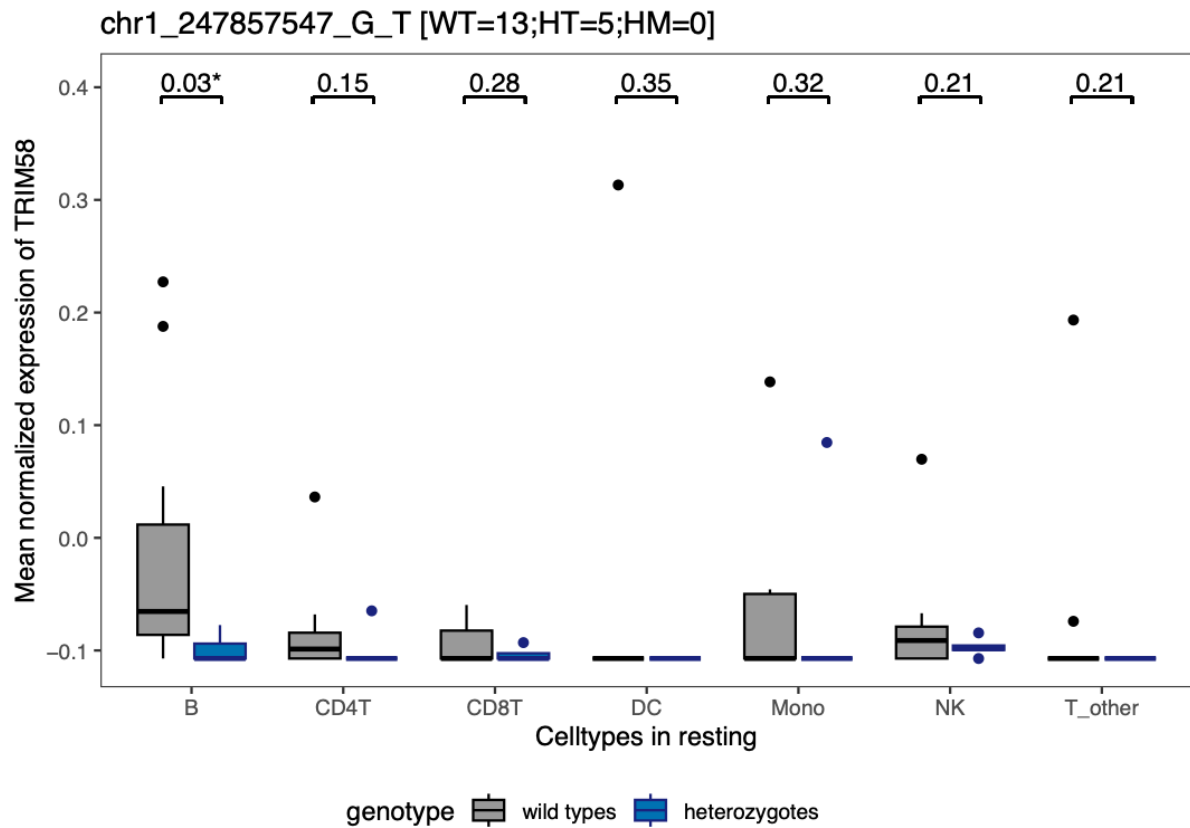


Figure-4.30

Expression of *TRIM58* in controls and chr1_247857547_G_T carriers. Single-cell expression was normalized, and the mean expression per cell-type and donor was calculated. The distribution of mean expression across donors was then plotted for each cell-type. A Wilcoxon test was performed to compare the expression between controls and carriers in each cell-type, with the results plotted on the top of the graph. The key takeaway is that expression of *TRIM58* is low across the cell-types.

Discussion

Summary of findings

In this study we profiled 94 pLoF carriers and control individuals, while stimulating a subset of 40 individuals with aCD3 and IFN α across two timepoints, generating a scRNAseq dataset of 857,370 PBMCs to study LoF effects on the immune transcriptome. We focused on ten pLoFs affecting nine immune-related genes, and demonstrated cell-type and context (defined as a stimulation condition and timepoint) specificity in both direct and indirect LoF effects. Direct LoF effects, i.e. knockdown of the gene transcript containing the variant, were detected at rest in specific cell-types, but became more significant in more cell-types in the presence of stimulation: At Azimuth level 1 cell-types, 10 out of the 40 unstimulated tests demonstrated nominally significant knockdown, while 70 out of 140 stimulated tests demonstrated nominally significant knockdown (chi-squared test p-value 8.6×10^{-3}). We also found indirect LoF effects, in the form of pathway derangements. This covered mechanisms that could be either compensatory for or a direct consequence of the pLoFs. As expected we found an enrichment in derangements affecting immune-related signalling and disease pathways. In contrast, the presence of pLoFs was associated with far more subtle changes to cell-type proportions, with only two such associations being convincing. Lastly, we systematically described what the pLoFs revealed about the function of each of the nine affected genes.

LoF effects are cell-type- and context-specific

At every level of the analysis, we observed cell-type- and context-specificity in the results. This was expected in our hypothesis, given that genetic variants in general have been known to have cell-type-specific effects (GTEx Consortium et al., 2017; Jonkers & Wijmenga, 2017; Soskic et al., 2022). Mechanisms cited to contribute to this specificity include alternative splicing, genetic elements that regulate expression vary in their functions across cell-types and contexts (Kim-Hellmuth et al., 2020).

In our study, we have chosen pLoFs that were predicted to affect all transcripts of the gene. Therefore, alternative splicing is a less likely mechanism regulating this specificity, however to definitively conclude this we would have to conduct transcript-specific expression analyses, which given the initial filtering step, is beyond the scope of this study.

In contrast, we observed that gene expression differences across cell-types and contexts are the key contributor to this specificity. We frequently observed that the pLoF-affected gene had higher expression in the control individuals in the cell-type and context in which the knockdown effect was statistically significant. Therefore, indicating that the detection of more

effects, or different effects, across cell-types and stimulation conditions is closely related to the gain in power from increased expression. An additional reason for this is that gene upregulation is often indicative of identifying the relevant cell state at which the gene exerts its function (and therefore needs to be expressed at high levels).

This demonstrates the power of our single-cell approach, with stimulation, in interrogating LoF effects and gene function.

LoF effects can be divided into compensation and consequence

The indirect LoF effects we have reported can be divided into compensatory effects and consequential effects. Broadly, compensatory effects are pathways and genes upstream of or parallel to the pLoF-affected gene that are upregulated in carriers compared to controls. And consequential effects are pathways and genes downstream that are downregulated in carriers to controls. (In inhibitory pathways, the directions of effect are reversed).

We found that many of the pLoFs were associated with both types of effects, although not always existing simultaneously in the same cell-type and context. For example, in *TRAT1*, we primarily see compensatory effects in CD4 T cells at rest, and consequential effects after 16h of CD4 T cell activation. A possible explanation for this is that consequential effects manifest when compensation becomes inadequate, such as in scenarios where the cells are stressed (during an immune response).

Again, the value of our cell-type- and context-specific approach is demonstrated here.

Studying LoF effects provide insight into gene function and supplement other functional analyses

This study also served as a proof-of-concept for leveraging on LoFs to interrogate cell function. Through the systematic profiling of these nine genes, we were able to gain further insight into their function, and the network of pathways that relate to them. We have provided a basis for deeper interrogation into these genes, and a framework for further similar profiles of other pLoFs. While our findings could be supplemented with knockout experiments, observations of the knockout's behaviour in the human being as a whole is often not possible. By studying these naturally-occurring knockouts, interestingly, we detected many compensatory pathways associated with these pLoFs. It is possible that we have enriched our data with "tolerated" pLoFs, as we recalled from within a community-based cohort, and to reduce the error rate we also filtered out pLoFs affecting constrained genes. These findings on how these pLoFs are tolerated *in vivo* are valuable in the context of drug target discovery.

Strengths, limitations and future directions

The choice of stimulation and the potential for additional bespoke assays

A main strength of the study was that we profiled LoF effects on the stimulated immune transcriptome at scale. It was unknown at the start the extent to which LoF effects would influence RNA transcription. scRNAseq is costly to perform at scale, but at the same time, benefits from pooling strategies and processing multiple samples at once, as it would reduce the confounding of batch effects compared to profiling samples one-by-one. Therefore, two rather general stimuli were selected to maximise the gain of multiple pLoFs being profiled together. We have shown here that this general approach has been informative across the multiple pLoFs we recalled and selected to stimulate. We demonstrated multiple LoF effects that were relevant to LoFs in general, as well as specific effects related to their affected genes' functions, prompting multiple threads for further investigation.

Firstly, very specific mechanisms known to involve one or two of the pLoF-affected genes could potentially benefit from more specific stimulation. For example, the pLoF carriers of the two PRRs, *DDX58* and *IFIH1*, could be presented with a viral challenge or viral nucleic acid analogues to interrogate their function directly, especially as IFN α acts downstream to their primary known function. However, viral challenges often require further laboratory safety requirements. In addition, selecting specific stimuli for each pLoF is not efficient in a multiple-pLoF study, as they may not be applicable to the other pLoFs that we aimed to profile in this study.

By extension, we could also perform specific functional analyses relevant to the findings from this study. As a first approach to studying LoF effects at scale, such bespoke studies were less resource-effective to implement from the beginning, but could now be considered in light of the findings from this study.

This study provides the foundation for larger scale surveys

Secondly, another strength of performing this study in Genes & Health (G&H) was the higher prevalence of homozygous pLoF carriers. This allowed us to profile homozygous pLoF carriers that might not have been possible even in the other larger, often majority white European, cohorts.

However with rare variants, like in the previous chapter, sample size and power was still a limitation in this study. Throughout this chapter, an additive approach was employed to represent the genotypes of the pLoFs carriers accurately and capture the most variance associated with pLoFs. Despite the increased homozygosity in G&H many of the pLoFs we

profiled had only one homozygous carrier recalled. And unlike the previous chapter, we did not have a large cohort of control individuals to plot normal distributions and relate the homozygous profiles to these distributions. Therefore, for instances such as the *TRAT1* pLoFs, which had evidence in the previous chapter that it might have a dominant effect, we could not test if this dominant effect persisted in the presence of stimulation. Given that we have identified cell-type- and context-specificity in every step of the analysis, we hypothesise that deviation from additivity could demonstrate such specificity as well.

The findings of this study therefore provide the support required for a larger, more extensive analysis of homozygous pLoFs. With funding from an industry consortium, G&H have been extending the whole-exome-sequencing (WES) to more individuals of the cohort, with the current numbers reaching ~51,000 individuals. With more exome-sequenced individuals, more pLoFs can be found and recalled. The approach presented across these two chapters could be extended to these individuals, and with more power, additional analyses such as non-additive testing could be performed without significant increase in cost (the same data, tested differently, as presented in chapters 2 and 3 of this thesis).

Given the cell-type- and context-specificity of LoF effects, if an extended list of pLoFs were to be profiled in the future, profiling more cells per individual, to allow even higher resolution analysis, and more stimuli, to cover more immune functions, could prove very informative. However, the more general approach to stimuli selection presented in this study would still be more appropriate in such a scenario compared to more bespoke analyses and more precise stimuli. In addition, the expression data provided here could be used for further power calculations, in a similar manner described in the previous chapter, to determine the number of cells required for scRNAseq.

Next, we demonstrated that indirect LoF effects on other genes and wider pathways can be very informative of the initial target gene function. Therefore, this approach could potentially be applied to a wider range of variants, such as deleterious missense mutations. While their direct knockdown effects on target genes may not be detectable at the RNA level, the indirect pathway changes could be useful for interpretation, and suggestive of the presence of target gene inactivation.

Proteomic and phenotypic analyses as further supplements to our findings

Lastly, this study has attempted to provide functional insight into genotype and phenotype associations by demonstrating how LoFs affect the transcriptome. Many of these genes have associations with blood and immune traits, and pathways related to these traits have been shown in this study to be dysregulated in the presence of gene inactivation by pLoFs.

A natural extension of this would be to link these findings directly back to the phenotype presentations of these individuals. With the extended 51,000 WES dataset, genome-wide association studies (GWAS) have been run for many of our recalled pLoFs with a range of binary and quantitative phenotypes in the electronic health records. This includes immune-related diseases and blood count results. This is significantly more powered and efficient than performing our own GWAS limited to the individuals we have recalled. However, at the time of submitting this thesis, the GWAS results from the extended 51,000 WES dataset is under a priority access period as part of the agreement with the funders.

In addition, blood protein quantification has also been performed by G&H with Olink proteomic assays. There is potential to corroborate our transcriptomic findings with the protein expression data. This work is being performed by our collaborators and presently, is also under a priority access period. Secondly, the CITEseq data collected here, which focuses on surface proteins, while not directly targeting the pLoF-affected genes in our study, would prove useful in corroborating the indirect LoF effects on surface proteins involved in the reported pathways.

All of this work is ongoing, and we do hope to include the analyses in a future publication.

Conclusion

From this systematic survey of LoF effects on the resting and stimulated immune transcriptome, we arrived at several conclusions. Firstly, direct knockdown effects of LoFs on the affected gene's expression is not universal, and appropriate stimulation made more of these knockdown effects more apparent across more cell-types. Secondly, LoFs were indirectly associated with pathway derangements, which can be compensatory or consequential. As expected, immune pathways and immune-related diseases were enriched for dysregulated pathways. In contrast to pathway derangements, LoF-associated cell-type proportion changes were subtle, with only two convincing instances presented in this study, both only in the presence of stimulation. Thirdly, a systematic characterisation of the LoF-affected genes assayed in this study has allowed us to gain further insight into their function. Lastly, across the different surveys of LoF effects performed, there was evidence of cell-type- and context-specificity, highlighting the complexity of studying LoF effects at scale, and the value of our single-cell and stimulation approach. Taken together, from generating this unique dataset profiling LoF effects on immune cells, we have gained insights into both the power of LoFs in interrogating gene function, and the role of the actual pLoF-affected genes we recalled.

Acknowledgements for this chapter

Gosia Trynka and Hilary C. Martin supervised the overall work in this chapter. Wet lab work was supervised by Carla Jones, and performed by Tarran Rupall, Bess Chau and me. The breakdown of the wet lab workload has been described above. The DNA Pipelines Team at the Sanger Institute performed the library preparation and sequencing to generate the scRNAseq and CITEseq data. Anna Lorenc supervised the computational quality control and data analysis. The main developer of the YASCP pipeline was Matiss Ozols, and it was developed together with several others in the Sanger Institute, as listed in the pipeline documentation. Within it, quality control and the TOPMED-r3 imputation of the genetic data used for deconvolution of pooled scRNAseq data was performed by the Human Genetics Informatics Team. The Genes & Health cohort is overseen by David A. van Heel, and acknowledgements for the recall process is the same as in the previous chapter.

Discussion

Summary of findings

In this thesis, I have described two projects that leveraged the increased homozygosity in Genes & Health (G&H) to profile two types of genetic effects, recessive effects and rare, loss-of-function variant (LoF) effects.

Widespread recessive effects contribute to the heritability of common disease

In the first project, I looked for single variant recessive associations with common diseases in G&H.

I first performed within-cohort whole-exome imputation, with the aim of including rare variants in my testing. I also included imputation with TOPMED-r2, to include the noncoding regions. I observed that the within-cohort reference panel allowed for more variants to be imputed, especially at lower MAFs, and that for the low frequency variants ($N_{Hom} \geq 9$ to homozygous frequency 5%), the within-cohort exome panel performed better. However, the TOPMED-r2 imputation was more accurate for common variants, and allowed us to test many common, noncoding variants.

Then, from testing 607,981 whole-exome imputed variants and 10,045,406 whole-genome imputed variants (≥ 3 homozygous individuals) for association with 898 disease phenotypes (≥ 30 cases), we identified 207 independent loci that reached standard genome-wide significance ($p\text{-value} \leq 5 \times 10^{-8}$) under the recessive model and had p -values more significant than under the additive model. Of these, 154 loci demonstrated a nominally-significant ($p\text{-value} \leq 0.05$) dominance deviation p -value, confirming a recessive pattern of association.

The 32 loci that passed a Bonferroni-corrected threshold $p\text{-value} \leq 5.5 \times 10^{-12}$ corresponded to three genomic clusters that associated with three conditions: Non-alcoholic fatty liver disease and steatohepatitis (NAFLD), disorders of porphyrin and bilirubin metabolism, and thalassaemia and other hereditary haemolytic anaemias. In the independent cohort FinnGen, we then replicated 71% of the expected number of Bonferroni-significant loci we would be powered to replicate given the case rates, allele frequencies and sample size of FinnGen.

We identified several coding variant associations that provided insight into gene function, such as the missense variant in *MTHFR* revealing the gene's role in folate metabolism, and

the protective effects of a missense variant in *SGLT* for hypertension. We showed that the recessive model of testing in a cohort enriched for homozygosity increased the power for detecting recessive associations, and that these recessive effects may contribute to the association between increased homozygosity and common disease risk.

Loss-of-function variants are associated with direct and indirect effects on RNA expression that are cell-type- and context-specific

In the second project, we aimed to study the effects of LoFs on the immune transcriptome by designing and performing a recall-by-genotype study in G&H.

To prioritise participants for recall, we called homozygous putative LoFs (pLoFs) from the exome sequences of G&H participants who self-declared as having related parents and applied additional filters to reduce the risk of prioritising spurious LoFs. Next, we filtered to genes expressed in blood and involved in the immune system by studying various publicly available datasets and considering immune trait association studies. We also performed power calculations to estimate which knockout effects could be detected. This resulted in 22 genes affected by 32 pLoFs, with 488 individuals prioritised for recall who are heterozygous or homozygous for these variants. From these, 114 individuals were recruited: 94 individuals with 24 pLoFs, and 20 controls. The recalled pLoFs then underwent further manual curation, and were given a curation verdict: 15 were assigned as “LoF” and “Likely_LoF”, 9 were assigned “Uncertain”, “Likely_Not_LoF” and “Not_LoF”.

This recall project was part of a larger peripheral blood mononuclear cell (PBMC) single-cell RNA sequencing (scRNAseq) effort, known as the Cardinal project which included the recruitment of ~1,500 G&H participants and ~5,000 UK Biobank participants. One hundred and eight individuals from our recall were found in Cardinal and sequenced to a total of 238,707 cells, with a median of 2,043 cells/sample. We found that the direct knockdown effects of the pLoFs on affected genes largely concurred with the verdicts from the additional curation, with “LoF”s and “Likely_LoF”s demonstrating knockdown effects more than the other verdicts. We also found weak evidence for a recessive pattern to the knockdown effect in three pLoFs, although more power is needed to confirm this.

Next we performed an extended scRNAseq on 94 of our recalled individuals (the cryogenically-banked samples that passed viability checks), aiming for more cells per donor, and with stimulation of 40 donors with two stimuli at two timepoints. This resulted in a dataset of 857,370 cells, with a mean of 3,211 cells per sample. We focused on the ten pLoFs with a “LoF” and “Likely_LoF” verdict, and at least three donors recruited. We observed that globally carriers of LoFs in immune genes had upregulation of mRNA

surveillance, RNA and protein degradation pathways in immune cells. We also identified cell-type- and context-specific LoF effects at every level of the analysis: Firstly, direct knockdown effects on pLoF-affected genes became more significant across more cell-types in the presence of stimulation. Next, we detected pathway derangements associated with pLoFs, and they too differed across cell-types and stimulation conditions. Lastly, in two selected instances we observed subtle cell-type proportion changes: an increase in mitotic natural killer (NK) cells in *TRAT1* pLoF samples undergoing anti-CD3 (aCD3) stimulation, and a new cluster of monocytes in *MNDA* pLoF samples undergoing interferon-alpha (IFNa) stimulation. Finally, we systematically collated and described the findings for the 10 pLoFs. A key finding was the compensatory and consequential effects of a *TRAT1* pLoF on PI3-AKT signalling during T cell activation.

Wider implications of these findings

The majority of large-scale studies of genetic variation have focused on the study of common variants with additive effects. The findings from this thesis suggest that there is a lot of potential in evaluating other types of variant effects.

More recessive effects may remain to be found

Firstly, from the findings in Chapter 2, we can see that more recessive variants may remain to be found. Prior reports have suggested that the additive model in a GWAS should capture most of the variant effect regardless of the underlying pattern of inheritance (Hill et al., 2008), and that dominance heritability, when calculated, is small (Palmer et al., 2023; Zhu et al., 2015). Indeed, we see that is the case in our dataset as well. However, it is established that heritability in general (Visscher et al., 2008), as well as the burden of recessive effects on disease (Kartik Chundru et al., 2023; H. C. Martin et al., 2018), both change depending on population factors such as the degree of homozygosity in the population. Therefore it is possible that some recessive effects still remain to be found, and may explain some of the different disease burdens in different communities. We believe that there is potential for a more extensive search for recessive effects, particularly in G&H (See future directions below).

LoF variants are a promising class of variants to interrogate gene function, and can be evaluated at a single-cell RNA level at scale

We also confirmed that LoF variants, because of the direct way their effects can be interpreted, are a promising class of variants that can supplement other functional studies of genes and molecular pathways. Healthy individuals are estimated to carry anywhere between 200 to 800 LoFs per person (1000 Genomes Project Consortium et al., 2010; Ng et al., 2008; Pelak et al., 2010), therefore as a variant class, LoFs are a good resource for observing the effects of gene inactivation in the natural environment. In this thesis we have demonstrated the potential in evaluating them at a single-cell RNA level and at scale.

The multi-omics approach to studying genetic variation adds further biological insight

In this thesis we studied LoF effects on RNA expression, uncovering one of the intermediate steps linking genotypes to phenotypes. More dimensions to this analysis could provide further insight into how LoFs contribute to traits and diseases. This includes the ongoing CITEseq analysis, as well as further proteomics and biomarker analysis that have been planned (or can be planned) in the future.

We also show that these LoFs were associated with widespread indirect effects in the form of pathway derangements, and differential gene expression in other genes. Potentially, this approach could be used to evaluate other types of variants that do not have a direct knockout effect on the genes they encode, such as deleterious missense variants, and even noncoding, regulatory variants. Further analysis into RNA transcription or protein structure could also uncover the reasons why different variants have a different pattern of association (additive, dominant, recessive) with phenotypes (Gerasimavicius et al., 2022).

Specific findings within this thesis could benefit from deeper functional evaluation

A good place to start with more functional evaluations would be the coding variants with recessive effects identified in Chapter 2. Specifically, the missense mutation of *SGLT2* was associated with being protective for hypertension, and may serve as a potential drug target. In addition, several of the genes extensively characterised in Chapters 3 and 4, *TRAT1* in particular, could be studied with other functional assays to interrogate gene function further.

The importance of diverse cohorts in genetic studies

Finally, in this thesis we have presented a strong argument for the importance of performing genetic studies in diverse cohorts. Different ancestries have different population structures, and are affected by different diseases with different underlying genetic contributions and pathophysiology. For example, the rates of coronary artery disease and type 2 diabetes mellitus are two to four times higher in British South Asians compared to White British (Finer et al., 2020), and the incidence of thalassaemia is also much higher in people originating from South Asia and the Mediterranean (De Sanctis et al., 2017). There are also significant ancestry differences in immune-mediated diseases (Slight-Webb et al., 2023). The variant effects evaluated in this thesis involved many of these conditions, and have provided further insight in the causes of these conditions in British South Asians. In summary, not only will these studies make genetics more applicable to a wider community of people, they will also add to our knowledge of biology in general.

Limitations of current approaches

There are several limitations that apply to both projects in this thesis.

Power is a recurrent limitation throughout the thesis

Repeatedly, we saw that both projects would benefit from larger sample sizes. In Chapter 2, more sequenced individuals would increase the number of variants that could be tested, particularly rare, coding variants. In Chapters 3 and 4, most of the LoFs we profiled had at most one to two homozygous individuals. It was clear that with a larger sequenced cohort, we would have better power to perform many of these analyses. Secondly, the number of participants in the recall-by-genotype study was set at around 100, with 40 selected for stimulation, at around 3,000 cells per donor. These numbers were limited by the high cost of scRNAseq. Nevertheless, the sample sizes that we obtained in both projects were still sufficiently large to address the aims of the analyses, due to the fact that we performed power calculations prior to starting both projects.

Recessive analysis needs further development

In both projects, we attempted to test for recessive effects of variants on phenotypes. There is still a lot of room for further methods development in this area, and several specific limitations of the methods we applied have already been raised in the individual chapters. Notably, fine-mapping of recessive associations is not well-established, and to my knowledge, there are no established algorithms developed for scRNAseq to evaluate non-additive effects.

Functional evaluation is challenging to perform at scale

Lastly, while we have demonstrated the potential of multi-omics and functional analysis of variant effects, these are still challenging analyses to perform at scale. Not only are functional assays usually costly, as we showed in Chapter 4, variant effects can also be very cell-type- and context-specific, and the interrogation of different variants are likely to require bespoke experimental designs.

In addition, one of the reasons why the immune system was selected as our system of choice for our recall-by-genotype was that blood was relatively more accessible compared to many other tissues. LoFs (and other variants) are implicated in diseases across all systems, and some of these tissues are difficult, or maybe impossible, to access, such as brain tissue to evaluate neurological disorders.

Future directions

Ongoing plans for the expansion of Genes & Health

As alluded to in the individual chapters, there are ongoing plans by G&H to address the power limitations raised here. With funding from an industry consortium, the rest of the G&H cohort (N ~ 50,000) have been whole-exome-sequenced and analysis on this expanded exome dataset is ongoing. There are also plans to expand the cohort to ~100,000 individuals in the future.

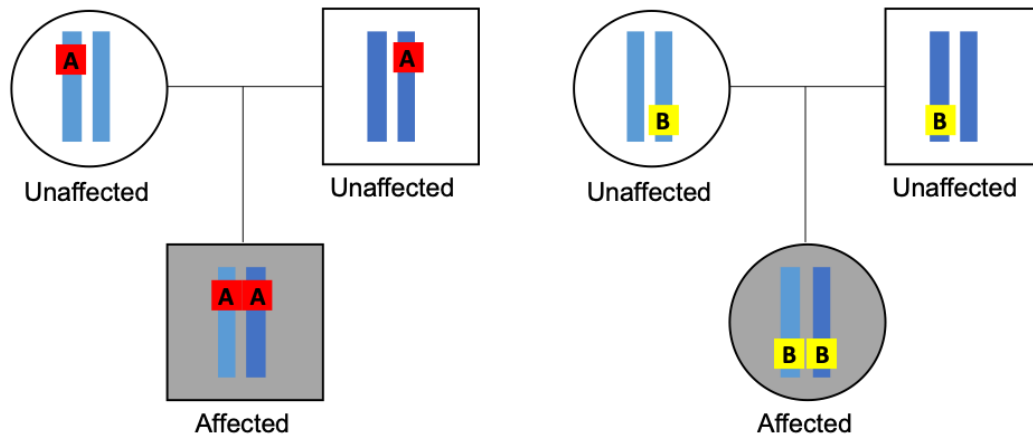
With the larger exome-sequenced dataset, more extensive recessive association analyses can be performed. These include both single-variant tests and gene-based burden tests, where compound heterozygote genotypes can be collapsed together with homozygous genotypes for testing.

An approach to including compound heterozygotes to improve power for recessive testing has been proposed by Lassen et al. 2023, where they phased rare variants in the UK Biobank exome sequencing cohort (n = 175,587), identifying compound heterozygous genotypes that could be combined with homozygous genotypes for gene burden testing. From testing across 311 common diseases, they identified 6 gene-trait associations, of which 5 would have been missed if considering homozygotes alone. Due to the increase homozygosity in G&H, we do not anticipate the gain in power in G&H to be as drastic as UK Biobank, but nevertheless, it is a promising approach to consider. I have included a plot, Figure-5.1, to illustrate how compound heterozygous genotypes can improve power to detect recessive effects. (Lassen et al., 2023)

In addition, there are potentially plans to perform more recall-by-genotype studies, with scRNAseq being performed at higher resolutions (more cells per donor), depending on funding. The extended list of LoFs curated from the exome expansion will contribute to this expanded analysis.

Assume variant A and variant B both have a recessive association with disease X

Couples heterozygous for the same variant may have a chance of having an affected proband:



Couples heterozygous for either variant may also have a chance of having an affected proband, due to compound heterozygosity:

Knowing the phase of the variants is important, as inheriting both variants in *cis*, may not be associated with disease:

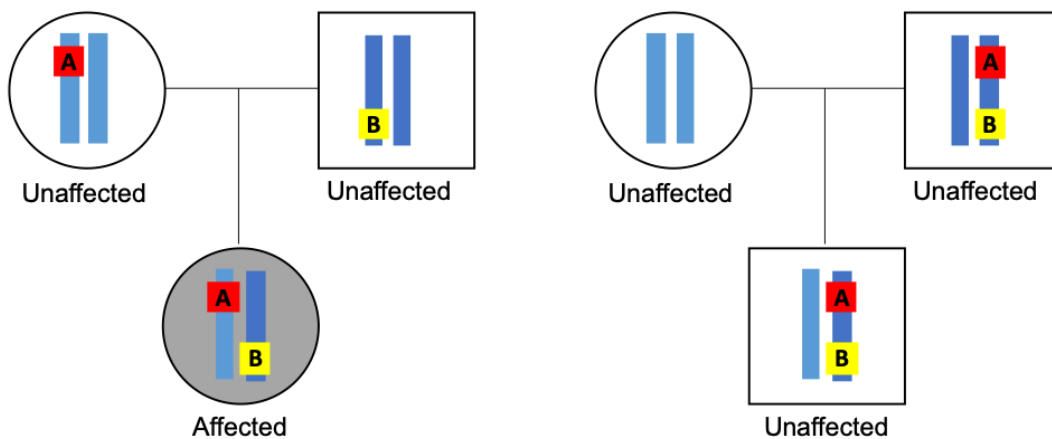


Figure-5.1

Compound heterozygous can contribute to recessive disease inheritance.

More types of variant effects should be considered

As mentioned in the introduction to this thesis, many types of variant effects contribute to the heritability of a trait. Here we have characterised recessive effects and rare, LoF effects in detail. However, many other variant effects remain to be studied, and the constant development of genomic technologies will aid in this.

As previously alluded to, compound heterozygous variants may behave similarly to homozygous variants. With more accurate phase information acquired from sequenced data, these variant effects can be studied in greater detail in G&H, and also add power to recessive gene-based recessive association testing.

Next, the addition of proteomic and other multi-omics assays can allow for more types of variants, such as deleterious missense variants, to be functionally evaluated.

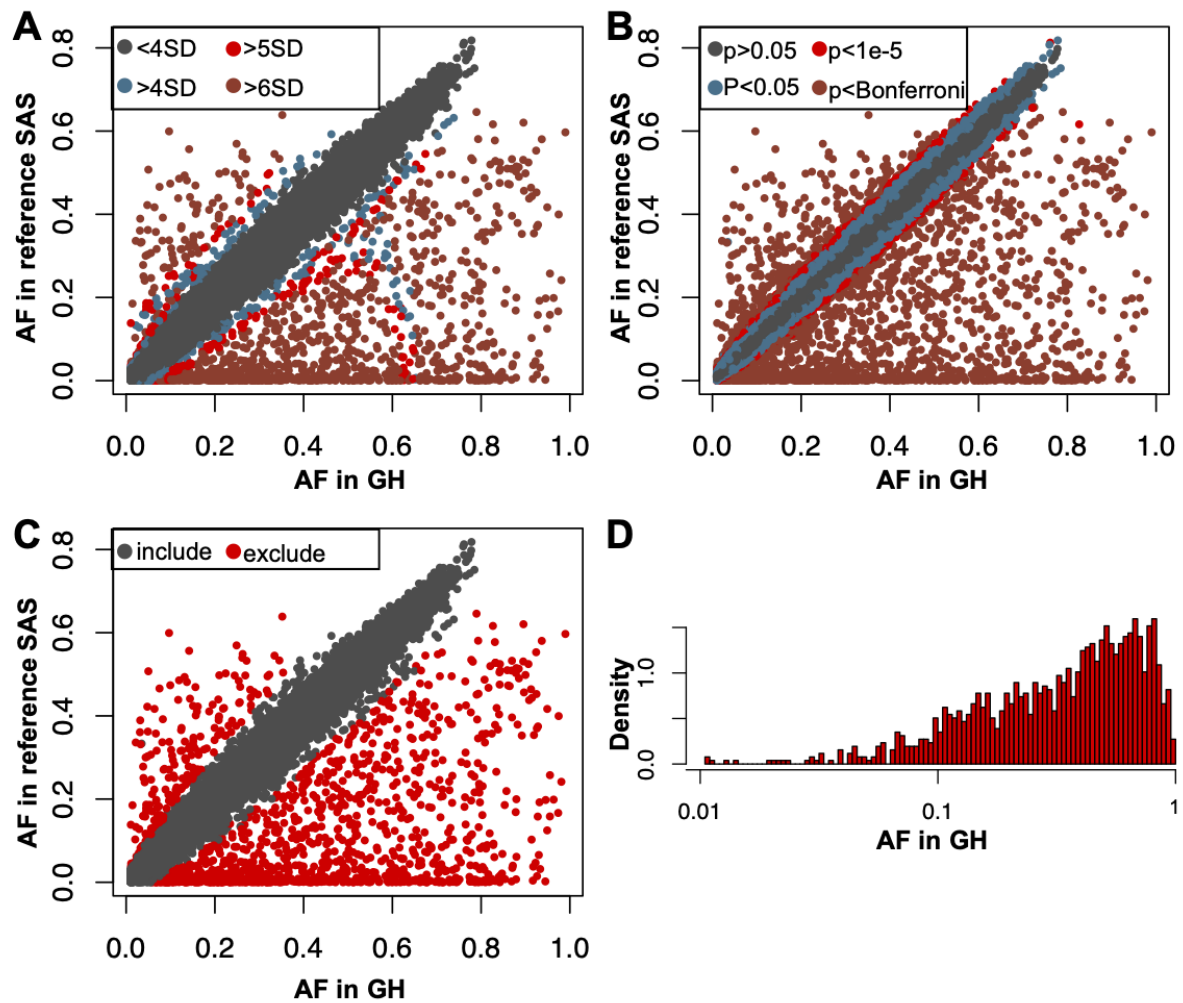
Lastly, with the development of long-read sequencing, large structural variants or copy number variants can be evaluated as well.

Some of these projects are ongoing within the G&H research community. I hope that the findings from this thesis will have contributed as a foundation for these further studies.

Appendices

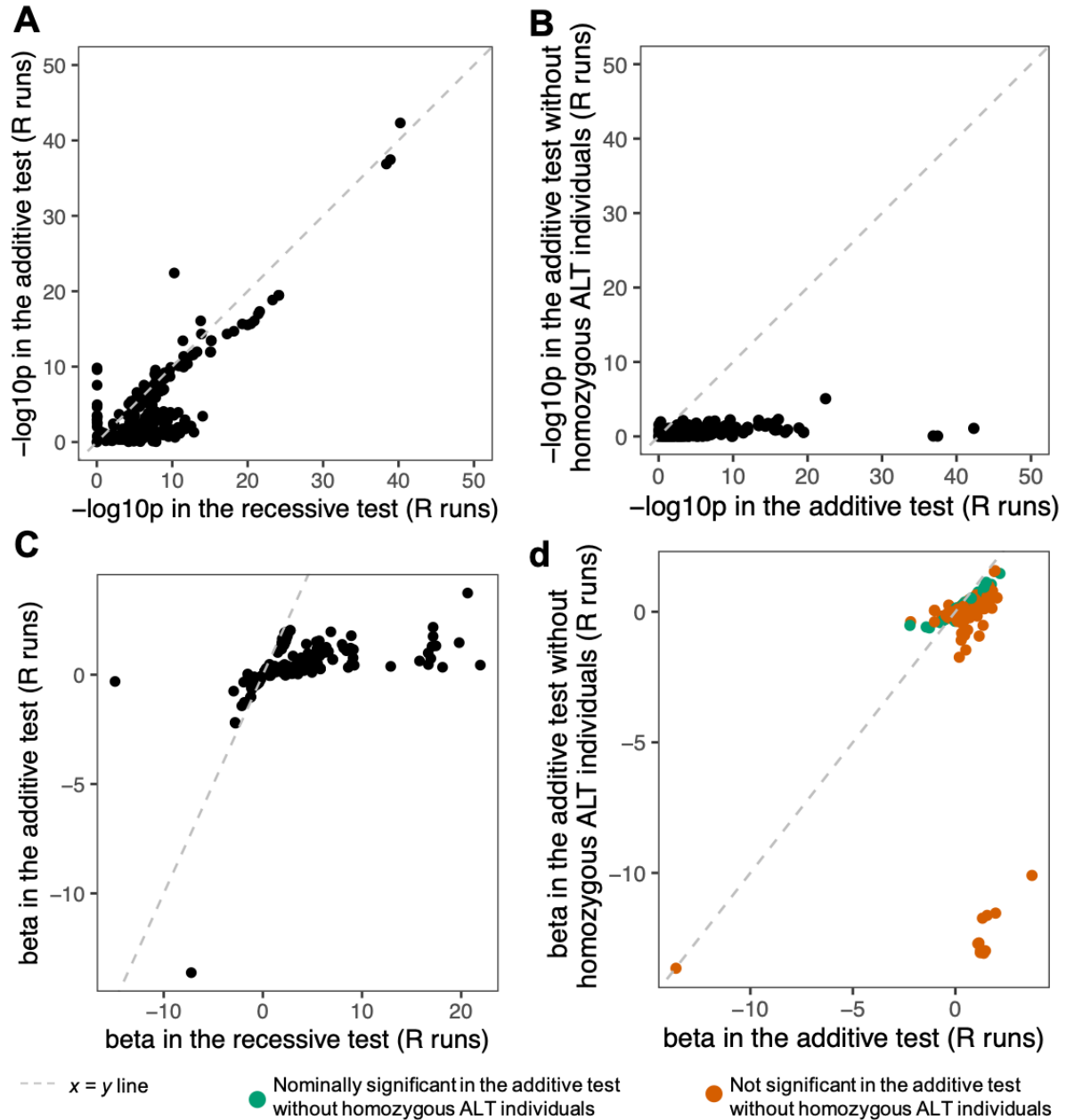
Appendix A - Supplementary data for Chapter 2

Supplementary figures



Supplementary-figure-2.1

Excluding variants with significantly different AF in Genes & Health (GH) and 676 South Asian samples from 1000 Genomes and HGDP (reference SAS). Plotting the AFs in both cohorts for each variant, colour-coding represents in **A**: the different standard deviation (SD) thresholds applied to the mean of the residuals in 0.01 frequency bins, in **B**: the different p-value thresholds for Fisher's exact tests on the genotypes, and in **C**: the variants that were included and excluded from the merged panel of variants. **D**: The AF distribution of the variants excluded.



Supplementary-figure-2.2

Exploring the recessive lead variants with different models of testing in R. A,C: P-values ($-\log_{10}p$, **A**) and betas (**C**) in the recessive tests compared to the additive tests in R. **B,D:** P-values ($-\log_{10}p$, **B**) and betas (**D**) in the additive tests compared to the additive tests without homozygotes in R.

Supplementary tables

Supplementary-table-2.1

(Below) Table of all 207 GWS recessive lead variants:

Lead variant	AF	CSQ	Gene	Phenotype	Beta	p-value	delta log10 p
2:171680912:G:A	3.3E-01	intergenic	.	A09[Other gastroenteritis and colitis of infectious and unspecified origin]	-0.5	2.7E-08	1.5
4:81731628:T:C	4.1E-02	intron	<i>RASGEF1B</i>	B35[Dermatophytosis]	1.1	3.9E-08	5.7
19:32864607:A:C	2.6E-01	intron	<i>SLC7A9</i>	B35[Dermatophytosis]	0.3	3.4E-08	3.2
18:54986413:C:A	4.7E-03	intergenic	.	B85[Pediculosis and phthiriasis]	5.5	3.1E-08	6.6
12:84932173:A:G	4.3E-01	intergenic	.	B85[Pediculosis and phthiriasis]	0.6	1.2E-08	4.5
18:54170645:T:C	1.5E-02	intron	<i>MBD2</i>	B96[Other specified bacterial agents as the cause of diseases classified to other chapters]	3.1	8.5E-09	5.8
6:52012065:G:GT	3.9E-03	intron	<i>PKHD1</i>	B97[Viral agents as the cause of diseases classified to other chapters]	5.2	2.7E-09	6.4
6:52012065:G:GT	3.9E-03	intron	<i>PKHD1</i>	J12[Viral pneumonia, not elsewhere classified]	5.6	3.1E-09	6.7
6:51529735:A:G	2.0E-03	intergenic	.	B97[Viral agents as the cause of diseases classified to other chapters]	6.1	2.4E-08	5.2
2:121338442:C:A	1.9E-01	3 prime utr	<i>CLASP1</i>	B98[Other specified infectious agents as the cause of diseases classified to other chapters]	0.8	3.8E-08	4.3
5:165757061:T:TA	1.9E-01	intron	.	D12[Benign neoplasm of colon, rectum, anus and anal canal]	1.0	1.9E-08	5.6
5:165755237:G:A	1.9E-01	intron	.	D12[Benign neoplasm of colon, rectum, anus and anal canal]	1.0	8.4E-09	6.1
8:134748783:T:C	2.1E-02	intron	.	D22[Melanocytic naevi]	2.5	1.5E-08	5.8
11:5478732:G:C	8.1E-01	intron	<i>HBE1</i>	D56[Thalassaemia]	-0.6	1.9E-16	1.7
11:5477659:C:T	6.7E-01	intron	<i>HBE1</i>	D56[Thalassaemia]	-0.4	2.1E-08	0.2
11:5289803:G:A	9.3E-01	intron	<i>HBE1</i>	D56[Thalassaemia]	-0.4	4.5E-08	0.5
11:5330439:C:T	7.1E-01	intron	<i>HBE1</i>	Thalassaemia	-0.9	1.7E-12	0.6

11:5367606:A:C	9.8E-01	intron	<i>HBE1</i>	D58[Other hereditary haemolytic anaemias]	-1.1	6.2E-36	0.5
11:5401996:G:T	9.3E-01	intron	<i>HBE1</i>	D58[Other hereditary haemolytic anaemias]	-0.6	2.5E-12	0.6
11:5381740:G:A	9.1E-01	intron	<i>HBE1</i>	D58[Other hereditary haemolytic anaemias]	-0.6	9.1E-11	0.7
11:5385153:T:A	8.7E-01	intron	<i>HBE1</i>	D58[Other hereditary haemolytic anaemias]	-0.5	3.4E-08	1.0
11:5328974:T:C	6.7E-01	intron	<i>HBE1</i>	Other_haemolytic_anaemias	-1.0	4.7E-09	0.1
11:5478732:G:C	8.1E-01	intron	<i>HBE1</i>	Other_anaemias	-0.2	4.3E-09	2.5
11:5247392:T:G	6.1E-01	downstream gene	<i>HBG1</i>	D58[Other hereditary haemolytic anaemias]	-1.0	2.3E-19	5.1
11:4931228:C:A	7.9E-01	intron	<i>MMP26</i>	D58[Other hereditary haemolytic anaemias]	-0.9	6.2E-18	0.7
11:4930728:C:T	7.3E-01	intron	<i>MMP26</i>	D58[Other hereditary haemolytic anaemias]	-0.8	4.4E-15	0.2
11:4876745:T:TTA	8.7E-01	intron	<i>MMP26</i>	Other_haemolytic_anaemias	-0.9	1.0E-11	0.7
11:4813157:A:G	8.1E-01	intron	<i>MMP26</i>	Other_haemolytic_anaemias	-0.8	1.6E-08	0.2
11:4908482:G:A	9.5E-01	3 prime utr	<i>OR51A7</i>	D58[Other hereditary haemolytic anaemias]	-1.4	9.6E-67	0.5
11:5351884:G:A	8.8E-01	missense	<i>OR51B6</i>	D58[Other hereditary haemolytic anaemias]	-0.5	2.5E-08	0.8
11:5005537:T:C	9.7E-01	downstream gene	<i>OR51L1</i>	D58[Other hereditary haemolytic anaemias]	-1.5	1.6E-80	0.3
11:5389801:C:T	9.0E-01	missense	<i>OR51M1</i>	D58[Other hereditary haemolytic anaemias]	-0.6	9.6E-10	1.3
11:5197578:AAT:A	5.5E-01	downstream gene	<i>OR51V1</i>	D58[Other hereditary haemolytic anaemias]	-1.7	1.0E-30	1.6
11:5154247:G:A	7.1E-01	intron	<i>OR52A1</i>	D56[Thalassaemia]	0.4	2.8E-09	0.2
11:5139284:G:A	7.6E-01	upstream gene	<i>OR52A5</i>	D56[Thalassaemia]	0.4	3.1E-08	1.4
11:5489111:C:T	8.1E-01	synonymous	<i>OR52D1</i>	D56[Thalassaemia]	-0.5	4.0E-13	1.5
11:5067264:A:G	1.0E+00	upstream gene	<i>OR52E1</i>	D58[Other hereditary haemolytic anaemias]	-2.0	8.7E-140	5.4
11:5058716:C:T	9.9E-01	synonymous	<i>OR52E2</i>	D58[Other hereditary haemolytic anaemias]	-2.0	1.7E-147	7.2
11:5095128:G:A	7.8E-01	downstream gene	<i>OR52E3 P</i>	D56[Thalassaemia]	0.4	2.5E-09	0.4

11:5544800:A:G	9.9E-01	missense	<i>OR52H1</i>	D58[Other hereditary haemolytic anaemias]	-1.1	3.6E-33	2.9
11:5551930:G:C	9.0E-01	non coding transcript exon	<i>OR52H2 P</i>	D58[Other hereditary haemolytic anaemias]	-0.5	2.0E-09	0.1
11:4433683:C:A	6.7E-01	intron	<i>OR52K3 P</i>	D58[Other hereditary haemolytic anaemias]	-0.7	4.9E-10	0.7
11:4410946:C:G	6.6E-01	intron	<i>OR52K3 P</i>	D58[Other hereditary haemolytic anaemias]	-0.6	6.4E-10	0.4
11:4388156:C:T	7.7E-01	intron	<i>TRIM21</i>	D58[Other hereditary haemolytic anaemias]	-0.6	4.1E-09	1.5
11:5609405:T:C	9.1E-01	intron	<i>TRIM6</i>	D56[Thalassaemia]	-0.4	3.2E-10	2.0
11:4598006:G:A	8.1E-01	downstream gene	<i>TRIM68</i>	D58[Other hereditary haemolytic anaemias]	-0.5	9.6E-09	2.2
21:39112030:G:T	5.1E-02	regulatory region	.	D64[Other anaemias]	1.2	5.5E-09	7.6
11:5573872:C:G	9.8E-01	intron	.	D56[Thalassaemia]	-0.4	4.7E-08	0.2
11:5216054:A:G	7.6E-01	intergenic	.	D56[Thalassaemia]	-0.6	5.4E-16	1.4
11:5213312:G:C	7.6E-01	intergenic	.	Thalassaemia	-0.8	2.2E-10	0.1
1:11796321:G:A	1.5E-01	missense	<i>MTHFR</i>	Folate_deficiency_with_and_without_anaemia	0.7	4.9E-09	1.1
8:103355120:G:A	8.1E-01	intergenic	.	Folate_deficiency_with_and_without_anaemia	-0.2	3.8E-08	1.0
2:82023510:T:G	2.7E-01	intergenic	.	D70[Agranulocytosis]	1.2	4.3E-08	3.5
6:1973850:C:T	5.1E-02	intron	<i>GMDS</i>	Hodgkin_Lymphoma	1.6	2.7E-08	4.3
6:1973850:C:T	5.1E-02	intron	<i>GMDS</i>	Non.Hodgkin_Lymphoma	1.6	1.1E-08	4.9
11:74720522:A:C	8.3E-02	intron	<i>CHRD12</i>	Leukaemia	1.7	1.2E-08	4.6
6:41493001:A:G	2.0E-01	intron	<i>FOXP4-AS1</i>	Primary_Malignancy_Other	1.0	6.3E-09	6.1
7:32513757:C:T	5.7E-01	intron	<i>AVL9</i>	E05[Thyrotoxicosis [hyperthyroidism]]	0.6	4.0E-08	3.6
6:140417452:G:A	3.9E-03	intron	.	E05[Thyrotoxicosis [hyperthyroidism]]	6.8	4.8E-10	6.0
6:104459545:G:A	6.6E-03	regulatory region	.	E11[Type 2 diabetes mellitus]	3.6	3.6E-08	7.1
10:92706738:G:C	7.4E-01	downstream gene	<i>Y_RNA</i>	GNH0242_Type_2_Diabetes_narrow	-0.2	8.8E-09	4.E-02

10:92706682:A:G	7.6E-01	downstream gene	Y_RNA	Type_2_Diabetes	-0.2	4.3E-09	0.5
5:5261658:G:A	5.2E-03	intron	ADAMT S16	E14[Unspecified diabetes mellitus]	5.7	5.1E-09	7.3
3:23530638:C:A	8.4E-04	intron	UBE2E2	E14[Unspecified diabetes mellitus]	7.9	4.6E-08	5.3
20:17949555:G:T	1.0E-02	intron	SNX5	GNH0244_Unspecified_or_Rare_Diabetes_narrow	4.1	1.4E-08	4.4
3:176679876:G:C	4.3E-03	intron	LINC01208	E22[Hyperfunction of pituitary gland]	7.1	1.7E-08	5.7
11:116796367:A:G	9.4E-01	upstream gene	APOA5	E78[Disorders of lipoprotein metabolism and other lipidaemias]	-0.2	3.3E-11	1.1
2:233780897:C:G	4.4E-01	intron	MROH2 A	E80[Disorders of porphyrin and bilirubin metabolism]	1.7	1.6E-16	1.4
2:233790144:A:G	4.6E-01	intron	MROH2 A	E80[Disorders of porphyrin and bilirubin metabolism]	1.6	2.7E-14	0.8
2:233670722:G:A	4.0E-01	intron	UGT1A10	E80[Disorders of porphyrin and bilirubin metabolism]	1.6	1.9E-11	1.0
2:233671494:G:A	4.0E-01	intron	UGT1A10	E80[Disorders of porphyrin and bilirubin metabolism]	1.6	2.0E-11	0.9
2:233605627:C:T	4.9E-01	downstream gene	UGT1A11P	E80[Disorders of porphyrin and bilirubin metabolism]	1.3	2.7E-08	0.3
2:233713783:T:C	7.8E-01	synonymous	UGT1A5	E80[Disorders of porphyrin and bilirubin metabolism]	1.4	2.0E-12	0.4
2:233763993:C:A	5.3E-01	intron	UGT1A6	E80[Disorders of porphyrin and bilirubin metabolism]	1.9	5.3E-23	2.9
2:233763993:G:T	5.3E-01	intron	UGT1A6	E80[Disorders of porphyrin and bilirubin metabolism]	1.9	5.3E-23	2.9
2:233738671:G:A	5.0E-01	intron	UGT1A6	E80[Disorders of porphyrin and bilirubin metabolism]	1.8	1.8E-20	2.0
2:233715640:T:C	5.3E-01	intron	UGT1A6	E80[Disorders of porphyrin and bilirubin metabolism]	1.8	1.5E-19	2.6
2:233716651:G:A	5.1E-01	intron	UGT1A6	E80[Disorders of porphyrin and bilirubin metabolism]	1.8	3.2E-19	2.2
2:233703893:C:A	4.5E-01	intron	UGT1A6	E80[Disorders of porphyrin and bilirubin metabolism]	1.7	4.7E-16	1.9
2:233702448:A:C	4.6E-01	intron	UGT1A6	E80[Disorders of porphyrin and bilirubin metabolism]	1.7	5.6E-16	1.8
2:233757013:T:G	7.5E-01	intron	UGT1A6	E80[Disorders of porphyrin and bilirubin metabolism]	1.5	1.1E-14	0.3
2:233750066:G:A	3.7E-01	intron	UGT1A6	E80[Disorders of porphyrin and bilirubin metabolism]	1.8	2.7E-14	0.8
2:233743701:C:G	3.7E-01	intron	UGT1A6	E80[Disorders of porphyrin and bilirubin metabolism]	1.7	5.1E-14	0.6

2:233693319:A:G	5.8E-01	synonymous	UGT1A6	E80[Disorders of porphyrin and bilirubin metabolism]	1.5	1.3E-13	0.2
2:233693556:A:C	5.8E-01	missense	UGT1A6	E80[Disorders of porphyrin and bilirubin metabolism]	1.5	1.4E-13	0.2
2:233681970:C:A	5.1E-01	synonymous	UGT1A7	E80[Disorders of porphyrin and bilirubin metabolism]	1.4	1.1E-11	0.6
2:233755003:C:G	7.6E-01	splice donor region	.	E80[Disorders of porphyrin and bilirubin metabolism]	1.5	3.9E-14	0.3
1:76418868:T:TA	1.3E-02	intron	ST6GALNAC3	E87[Other disorders of fluid, electrolyte and acid-base balance]	3.9	3.3E-09	7.8
21:38092921:C:T	1.1E-01	intron	DSCR4	Hyperparathyroidism	2.2	4.7E-08	3.6
2:621954:T:C	9.5E-01	intergenic	.	Obesity	0.2	8.3E-10	0.3
5:28354818:C:T	1.1E-02	intergenic	.	Osteoporosis	6.3	3.0E-08	7.0
1:194023706:A:G	2.6E-03	intron	.	F11[Mental and behavioural disorders due to use of opioids]	7.8	3.6E-08	5.6
2:146088185:A:T	1.7E-02	intergenic	.	F12[Mental and behavioural disorders due to use of cannabinoids]	4.3	3.0E-08	3.5
2:236834949:C:T	7.1E-03	regulatory region	.	F40[Phobic anxiety disorders]	4.9	2.7E-08	6.1
7:131782270:G:A	7.8E-02	intergenic	.	Anxiety_and_phobia	0.6	3.5E-08	2.7
7:131780486:G:A	8.0E-02	intergenic	.	F41[Other anxiety disorders]	0.6	7.7E-10	4.6
7:131773247:A:G	1.3E-01	regulatory region	.	F41[Other anxiety disorders]	0.6	1.8E-08	3.6
17:9297878:G:A	2.2E-01	intron	STX8	Eating_disorders	0.7	4.7E-09	5.4
15:63724428:A:G	3.0E-01	intron	HERC1	Obsessive.compulsive_disorder	0.7	3.9E-08	3.3
18:48625702:C:T	4.1E-01	intron	CTIF	F53[Mental and behavioural disorders associated with the puerperium, not elsewhere classified]	0.8	1.1E-08	5.6
16:88431904:C:T	1.1E-02	synonymous	ZNF469	Learning_disability	5.2	1.9E-08	6.6
6:168047780:T:C	4.6E-03	downstream gene	KIF25	G62[Other polyneuropathies]	5.0	4.9E-08	6.1
8:143548606:G:A	4.8E-03	intron	.	H02[Other disorders of eyelid]	5.6	9.9E-09	5.3
13:96226624:A:G	3.7E-03	intron	HS6ST3	H25[Senile cataract]	8.9	1.2E-08	7.3

8:131037747:G:A	2.1E-03	intron	<i>ADCY8</i>	H26[Other cataract]	8.6	3.8E-08	6.4
8:131156394:T:C	2.1E-03	intergenic	.	H26[Other cataract]	11.7	2.3E-09	7.8
5:73372589:G:A	2.5E-03	regulatory region	.	H26[Other cataract]	9.9	3.8E-08	6.4
2:40115460:G:T	3.2E-03	synonymous	<i>SLC8A1</i>	H34[Retinal vascular occlusions]	8.7	1.6E-08	6.2
11:117677017:T:T G	3.1E-02	intron	<i>DSCAM L1</i>	H36[Retinal disorders in diseases classified elsewhere]	2.8	3.5E-08	6.6
16:78802406:C:T	3.8E-03	intron	<i>WWOX</i>	H53[Visual disturbances]	5.6	3.5E-08	6.5
1:103412173:G:G A	2.5E-03	upstream gene	.	H54[Visual impairment including blindness (binocular or monocular)]	5.8	1.4E-08	3.0
1:216911690:AG: A	5.3E-03	intron	<i>ESRRG</i>	H60[Otitis externa]	4.0	2.5E-09	6.6
14:47900039:C:T	9.9E-01	intergenic	.	Hearing_loss	-0.3	4.6E-08	0.9
7:117592588:A:G	5.2E-03	missense	<i>CFTR</i>	I07[Rheumatic tricuspid valve diseases]	6.4	2.7E-08	5.5
1:115290827:A:T	2.6E-03	intron	<i>NGF</i>	I42[Cardiomyopathy]	7.7	2.0E-08	5.5
19:10858458:G:A	8.5E-02	intron	<i>C19orf38</i>	I45[Other conduction disorders]	1.8	4.9E-08	4.8
19:10968052:C:T	8.7E-02	intron	<i>SMARCA4</i>	I45[Other conduction disorders]	1.8	1.3E-08	6.0
7:138844895:A:T	7.8E-01	intron	<i>KIAA1549</i>	I63[Cerebral infarction]	0.6	4.5E-08	0.1
9:15897638:CAT: C	2.2E-03	intron	<i>CCDC171</i>	I71[Aortic aneurysm and dissection]	8.9	8.2E-09	5.9
9:15524259:G:A	5.3E-03	downstream gene	<i>FTH1P12</i>	I71[Aortic aneurysm and dissection]	7.8	2.4E-08	7.0
11:1865076:C:T	7.8E-01	intron	<i>LSP1</i>	Hypertension	-0.2	3.3E-08	0.5
1:48228922:G:A	5.2E-02	missense	<i>SLC5A9</i>	Hypertension	-1.4	3.2E-08	5.2
1:48256392:G:GT	5.2E-02	intergenic	.	Hypertension	-1.5	1.7E-08	5.0
1:38245969:G:A	1.8E-02	intergenic	.	Peripheral_arterial_disease	4.2	4.9E-09	7.1
9:36302636:C:T	1.1E-02	intron	<i>CLTA</i>	J02[Acute pharyngitis]	3.2	4.7E-08	6.2
3:194350072:G:A	5.5E-01	intron	<i>CPN2</i>	J03[Acute tonsillitis]	-0.2	1.3E-08	1.4

15:67957738:G:T	6.1E-01	intergenic	.	J03[Acute tonsillitis]	-0.2	9.6E-09	3.2
14:46051834:T:A	3.4E-03	intron	<i>LINC00871</i>	J18[Pneumonia, organism unspecified]	5.6	2.6E-08	6.3
6:42069610:T:C	9.9E-02	intron	<i>TAF8</i>	J22[Unspecified acute lower respiratory infection]	0.7	3.1E-09	7.1
6:42057580:G:A	1.2E-01	3 prime UTR	<i>TAF8</i>	J22[Unspecified acute lower respiratory infection]	0.7	5.4E-09	7.1
20:49755626:A:C	2.7E-01	intergenic	.	J31[Chronic rhinitis, nasopharyngitis and pharyngitis]	-2.4	5.3E-09	4.9
3:64802433:A:G	2.7E-02	intron	<i>ADAMT S9-AS2</i>	J34[Other disorders of nose and nasal sinuses]	2.1	1.5E-08	6.6
7:48916122:T:C	1.6E-02	intergenic	.	J34[Other disorders of nose and nasal sinuses]	2.6	5.0E-09	6.4
5:91760313:C:G	1.1E-02	intron	.	J34[Other disorders of nose and nasal sinuses]	2.4	2.4E-08	6.5
17:82597298:A:C	1.4E-02	intron	<i>FOXK2</i>	J47[Bronchiectasis]	4.6	5.0E-08	4.8
14:104265563:T:A	1.2E-03	intron	<i>LINC02691</i>	Bronchiectasis	6.3	1.5E-08	4.3
14:104293742:C:T	1.2E-03	intron	.	Bronchiectasis	6.3	1.5E-08	4.9
16:1968049:T:C	1.9E-03	intron	<i>RNF151</i>	COPD	7.9	2.2E-08	5.0
5:129239038:C:T	5.5E-03	intergenic	.	GNH0011_COPD	7.5	4.9E-08	7.1
16:85985481:T:A	3.1E-01	downstream gene	.	GNH0017_COPD	0.7	4.9E-09	4.6
16:85972232:C:T	2.7E-01	intron	.	GNH0017_COPD	0.8	8.5E-09	4.1
2:206949775:G:T	3.6E-04	intron	<i>CPO</i>	J84[Other interstitial pulmonary diseases]	8.1	7.0E-09	2.1
12:164855:C:T	4.7E-02	intron	<i>IQSEC3</i>	K04[Diseases of pulp and periapical tissues]	1.5	3.0E-08	5.5
6:119122806:T:C	1.2E-02	intron	<i>FAM184A</i>	K05[Gingivitis and periodontal diseases]	3.6	2.6E-08	6.1
22:24384892:T:C	3.6E-03	intron	<i>SPECC1L</i>	K11[Diseases of salivary glands]	4.9	3.2E-08	6.9
2:170538482:C:T	1.3E-02	intron	<i>MYO3B</i>	K13[Other diseases of lip and oral mucosa]	2.7	3.9E-08	5.3
3:141734432:C:T	1.0E-02	upstream gene	<i>RNF7</i>	K14[Diseases of tongue]	4.4	1.3E-08	6.7
5:129155848:G:A	4.2E-02	intron	.	K21[Gastro-oesophageal reflux disease]	1.0	1.6E-08	4.3

14:83132884:T:C	6.8E-01	intergenic	.	Oesophagitis_and_oesophageal_ulcer	0.3	4.6E-08	0.9
3:60917770:T:C	9.5E-03	intron	<i>FHIT</i>	Peptic_ulcer_disease	3.8	1.9E-08	5.4
7:77206901:G:A	2.8E-03	intron	<i>CCDC146</i>	K44[Diaphragmatic hernia]	5.9	4.7E-08	4.8
9:119995336:G:C	3.2E-01	intergenic	.	K52[Other noninfective gastroenteritis and colitis]	0.5	7.8E-09	2.6
5:64110691:C:T	1.5E-03	intergenic	.	K52[Other noninfective gastroenteritis and colitis]	4.7	1.8E-08	0.9
21:31637098:G:G T	4.6E-02	intergenic	.	K57[Diverticular disease of intestine]	2.2	4.0E-08	7.3
2:69340329:G:A	9.9E-01	intron	<i>GFPT1</i>	K63[Other diseases of intestine]	-0.4	2.8E-09	1.8
11:127012361:A:G	1.8E-01	intergenic	.	K63[Other diseases of intestine]	0.8	2.1E-08	4.7
2:36076367:C:T	1.2E-02	intergenic	.	Chronic_viral_hepatitis	4.8	3.3E-08	4.7
22:43939790:TGG :T	6.1E-01	intron	<i>PNPLA3</i>	Non.alcoholic_fatty_liver_disease_and_steatohepatitis	0.3	2.4E-12	0.9
22:43939790:TGG :T	6.1E-01	intron	<i>PNPLA3</i>	K76[Other diseases of liver]	0.2	3.3E-10	0.7
18:51537133:G:C	2.7E-01	intron	<i>LINC01630</i>	L08[Other local infections of skin and subcutaneous tissue]	-0.7	4.3E-08	5.6
7:157970755:T:G	5.3E-03	intron	<i>PTPRN2</i>	L20[Atopic dermatitis]	4.2	1.1E-08	6.1
1:31215264:T:C	8.8E-01	intron	<i>NKAIN1</i>	L21[Seborrheic dermatitis]	0.2	6.2E-09	0.3
1:93847011:TCGG GCGCGCGGG:*	3.5E-01	intron	<i>BCAR3</i>	Dermatitis_atopc_contact_other_unspecified	0.2	1.6E-08	5.8
8:14479423:T:C	5.2E-02	intron	<i>SGCZ</i>	L50[Urticaria]	1.3	3.9E-08	5.7
13:69293892:G:A	2.6E-03	intron	<i>LINC00383</i>	L57[Skin changes due to chronic exposure to nonionizing radiation]	5.7	4.4E-08	4.0
13:69293892:C:T	2.7E-03	intron	<i>LINC00383</i>	L57[Skin changes due to chronic exposure to nonionizing radiation]	5.7	4.4E-08	4.1
2:66017686:C:T	5.6E-03	intron	<i>LINC02934</i>	L57[Skin changes due to chronic exposure to nonionizing radiation]	6.5	4.5E-08	6.1
2:66017686:G:A	5.7E-03	intron	<i>LINC02934</i>	L57[Skin changes due to chronic exposure to nonionizing radiation]	6.5	4.5E-08	6.2
2:79965309:A:G	1.9E-01	intron	<i>CTNNA2</i>	L73[Other follicular disorders]	1.0	2.9E-08	5.7
2:6740485:C:T	1.8E-03	intron	<i>LINC00487</i>	L73[Other follicular disorders]	5.8	2.2E-08	3.9

6:160102283:C:T	4.6E-03	intron	<i>IGF2R</i>	L82[Seborrheic keratosis]	6.0	2.9E-08	2.6
1:192575098:A:G	6.1E-03	intron	.	M23[Internal derangement of knee]	6.1	3.1E-08	6.8
7:158171977:G:A	5.2E-03	intron	<i>PTPRN2</i>	M41[Scoliosis]	6.2	3.6E-08	5.1
5:130167529:A:G	1.4E-02	intron	<i>CHSY3</i>	M50[Cervical disc disorders]	4.8	3.0E-08	5.5
5:130165261:G:A	1.6E-02	intron	<i>CHSY3</i>	M50[Cervical disc disorders]	4.8	3.0E-08	5.7
9:126052792:A:C	1.7E-02	upstream gene	.	Spinal_stenosis	3.8	3.8E-08	2.2
1:12265958:C:T	2.8E-03	intron	<i>VPS13D</i>	M53[Other dorsopathies, not elsewhere classified]	5.3	1.3E-08	3.1
2:96604271:A:C	3.5E-03	missense	<i>KANSL3</i>	M65[Synovitis and tenosynovitis]	5.3	4.0E-08	7.0
3:54384060:A:G	1.3E-02	intron	<i>CACNA2D3</i>	M70[Soft tissue disorders related to use, overuse and pressure]	4.2	4.5E-08	5.6
1:239913815:A:T	4.2E-03	3 prime utr	<i>CHRM3</i>	M70[Soft tissue disorders related to use, overuse and pressure]	4.9	3.4E-08	6.1
1:2620059:C:A	7.0E-04	intron	<i>MMEL1</i>	M70[Soft tissue disorders related to use, overuse and pressure]	8.1	2.7E-08	0.8
1:2287873:G:A	1.1E-03	intron	<i>SKI</i>	M70[Soft tissue disorders related to use, overuse and pressure]	8.1	2.7E-08	1.9
6:170033874:C:T	1.6E-02	intergenic	.	M72[Fibroblastic disorders]	1.9	4.2E-08	6.0
13:109179501:G:C	1.9E-01	intron	<i>MYO16</i>	M85[Other disorders of bone density and structure]	1.2	2.2E-08	5.6
20:36801415:C:A	2.1E-02	intron	<i>SOGA1</i>	Chronic_fatigue	2.7	3.8E-09	7.3
20:41794342:C:T	5.4E-01	intergenic	.	Fibromyalgia	0.5	3.2E-08	1.4
7:123155315:CT:C	2.1E-03	intron	<i>SLC13A1</i>	N10[Acute tubulo-interstitial nephritis]	6.8	9.1E-09	5.7
8:115346553:C:A	2.3E-02	intergenic	.	N10[Acute tubulo-interstitial nephritis]	2.6	2.1E-08	4.5
6:162067560:G:T	1.5E-02	intron	<i>PRKN</i>	N12[Tubulo-interstitial nephritis, not specified as acute or chronic]	2.6	4.1E-08	5.3
17:66212111:T:A	3.6E-03	3 prime UTR	<i>APOH</i>	Glomerulonephritis_and_other_nephritides	6.2	1.2E-08	5.3
4:81204967:C:G	9.4E-03	synonymous	<i>PRKG2</i>	Glomerulonephritis_and_other_nephritides	4.8	3.2E-08	6.5
10:61112281:ATTCT:A	7.9E-01	intron	<i>LINC00845</i>	Chronic_Kidney_Disease	-0.3	4.8E-08	0.5

8:3932209:A:G	4.3E-01	intron	<i>CSMD1</i>	N28[Other disorders of kidney and ureter, not elsewhere classified]	-1.0	4.9E-08	5.2
2:51166148:A:G	8.6E-02	intron	.	Erectile_dysfunction	-1.4	4.4E-08	7.1
13:22252190:C:T	8.4E-01	intron	<i>LINC00540</i>	N48[Other disorders of penis]	0.5	3.1E-08	0.7
7:27293317:C:T	7.4E-01	intergenic	.	N92[Excessive, frequent and irregular menstruation]	-0.2	2.6E-08	2.2
10:54716681:C:T	9.4E-01	intron	<i>PCDH15</i>	N97[Female infertility]	0.4	4.5E-08	1.2
22:31869259:G:A	3.7E-03	intron	<i>DEPDC5</i>	O20[Haemorrhage in early pregnancy]	7.3	4.4E-08	5.0
1:223842227:G:A	1.5E-03	intron	<i>TP53BP2</i>	O32[Maternal care for known or suspected malpresentation of fetus]	6.9	1.7E-08	3.5
4:179770094:G:A	1.4E-01	intergenic	.	O48[Prolonged pregnancy]	1.0	3.1E-08	5.0
6:144087557:T:C	5.4E-03	regulatory region	.	O80[Single spontaneous delivery]	8.2	1.3E-08	4.4
11:63684628:A:T	7.9E-01	intron	<i>RTN3</i>	O90[Complications of the puerperium, not elsewhere classified]	0.4	2.2E-08	2.0
10:4221063:C:G	9.6E-01	intron	<i>LINC00702</i>	O99[Other maternal diseases classifiable elsewhere but complicating pregnancy, childbirth and the puerperium]	-0.2	3.6E-08	0.8
2:4431718:G:C	1.6E-01	intergenic	.	GNH0247_GestationalHypertension	0.8	3.3E-08	4.8
1:45616613:C:T	1.8E-03	splice polypyrimidine tract	<i>NASP</i>	GNH0249_NormalUncomplicatedTermDelivery	9.2	2.4E-08	7.3
6:24532379:A:G	1.9E-02	intron	<i>ALDH5A1</i>	GNH0250_Preeclampsia_Eclampsia_broad	4.2	1.4E-08	7.6
19:13153584:C:T	4.3E-02	missense	<i>IER2</i>	Q61[Cystic kidney disease]	3.0	4.5E-08	3.4

Supplementary-table-2.2

(Below) Table of phenotype-matching between G&H and FinnGen:

Phenotype_GH	Phenotypes_FG
A09[Other gastroenteritis and colitis of infectious and unspecified origin]	AB1_GASTROENTERITIS_NOS
Anxiety and phobia	KRA_PSY_ANXIETY_EXMORE, F5_ANXIETY,F5_PHOBANX
Aortic aneurysm	I9_AORTANEUR
B35[Dermatophytosis]	AB1_DERMATOPHYTOSIS
B85[Pediculosis and phthiriasis]	AB1_PEDICULOSIS_AND_PHT HIRIAASIS
B96[Other specified bacterial agents as the cause of diseases classified to other chapters]	AB1_OTHER_BACT_INOTHER
B97[Viral agents as the cause of diseases classified to other chapters]	AB1_VIRAL_INOTHER
B98[Other specified infectious agents as the cause of diseases classified to other chapters]	not found
Bronchiectasis	J10_BRONCHIECTASIS
Chronic fatigue	not found
Chronic Kidney Disease	N14_CHRONKIDNEYDIS
Chronic viral hepatitis	not found
COPD	J10_COPD
D12[Benign neoplasm of colon, rectum, anus and anal canal]	CD2_BENIGN_COLORECANI
D22[Melanocytic naevi]	CD2_BENIGN_MELANOCYTIC
D56[Thalassaemia]	D3_ANAEMIANAS
D58[Other hereditary haemolytic anaemias]	D3_ANAEMIANAS,D3_HAEMOL YTICANAEMIA,E4_PORBILIR
D64[Other anaemias]	D3_ANAEMIANAS
D70[Agranulocytosis]	D3_AGRANULOCYTOSIS
Dermatitis atopc contact other unspecified	L12_ATOPIC
E05[Thyrotoxicosis [hyperthyroidism]]	THYROTOXICOSIS
E11[Type 2 diabetes mellitus]	T2D
E14[Unspecified diabetes mellitus]	T1D,T2D
E22[Hyperfunction of pituitary gland]	E4_HYPERPIT
E78[Disorders of lipoprotein metabolism and other lipidaemias]	E4_LIPOPROT
E80[Disorders of porphyrin and bilirubin metabolism]	E4_PORBILIR,D3_HAEMOLYTI CANAEMIA
E87[Other disorders of fluid, electrolyte and acid-base balance]	E4_FLUIDELECTRO
Eating disorders	KRA_PSY_EATING_EXMORE
Erectile dysfunction	ERECTILE_DYSFUNCTION
F11[Mental and behavioural disorders due to use of opioids]	F5_OPIOIDS
F12[Mental and behavioural disorders due to use of cannabinoids]	F5_CANNABIS
F40[Phobic anxiety disorders]	F5_PHOBANX,KRA_PSY_ANXI ETY_EXMORE,F5_ANXIETY
F41[Other anxiety disorders]	F5_ANXIETY,KRA_PSY_ANXIE TY_EXMORE

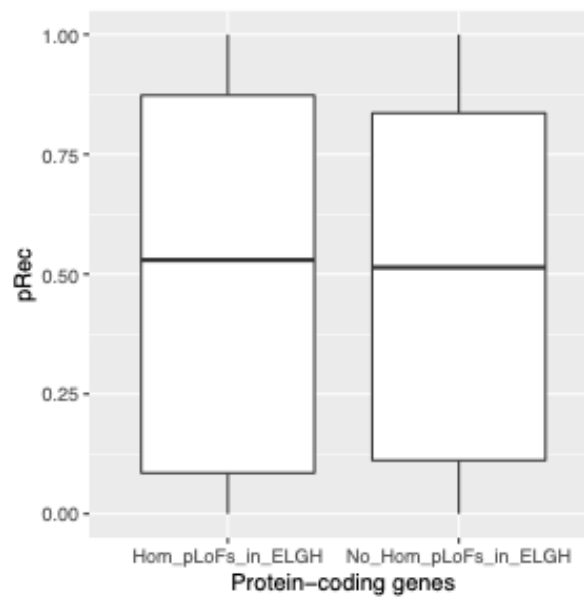
F53[Mental and behavioural disorders associated with the puerperium, not elsewhere classified]	KRA PSY PUERPER EXMORE
Fibromyalgia	M13_FIBROMYALGIA
Folate deficiency with and without anaemia	D3_ANAEMIA_FOLATEDEF
G62[Other polyneuropathies]	G6_POLYOTHUNS
Gastro.oesophageal reflux disease	K11_REFLUX
Glomerulonephritis and other nephritides	N14_GLOMER_NEPHRITIS
GNH0007_Type2Diabetes	T2D
GNH0011_COPD	J10_COPD
GNH0017_COPD	J10_COPD
GNH0018_EssentialHypertension	I9_HYPTENSESS
GNH0242_Type 2 Diabetes narrow	T2D
GNH0244_Unspecified or Rare Diabetes narrow	T1D,T2D
GNH0247_GestationalHypertension	O15_HYPTENSPREG
GNH0249_NormalUncomplicatedTermDelivery	O15_DELIV_SPONT
GNH0250_Preeclampsia Eclampsia broad	O15_PRE_OR_ECLAMPSIA
H02[Other disorders of eyelid]	H7_EYELIDDIS
H25[Senile cataract]	H7_CATARACTSENILE
H26[Other cataract]	H7_CATARACTOTHER
H34[Retinal vascular occlusions]	H7_RETIVASCOCCCLUSION
H36[Retinal disorders in diseases classified elsewhere]	H7_RETINODISINOTH
H53[Visual disturbances]	H7_VISUDISTURB
H54[Visual impairment including blindness (binocular or monocular)]	H7_BLINDANDVISIMPAIRMENT
H60[Otitis externa]	H8_EXTOTITIS
Hearing loss	H8_HL_SEN_NAS
Hodgkin Lymphoma	CD2_HODGKIN_LYMPHOMA_E_XALLC
Hyperparathyroidism	E4_HYPERPARA
Hypertension	I9_HYPTENS,I9_HYPTENSESS
I07[Rheumatic tricuspid valve diseases]	I9_RHEUVALV
I42[Cardiomyopathy]	I9_CARDMYO
I45[Other conduction disorders]	I9_CONDUCTIO
I63[Cerebral infarction]	C_STROKE
I71[Aortic aneurysm and dissection]	I9_AORTANEUR,I9_AORTDIS
J02[Acute pharyngitis]	J10_PHARYNGITIS
J03[Acute tonsillitis]	J10_TONSILLITIS
J12[Viral pneumonia, not elsewhere classified]	J10_VIRALPNEUMO
J18[Pneumonia, organism unspecified]	J10_PNEUMONIA
J22[Unspecified acute lower respiratory infection]	J10_LOWERINF
J31[Chronic rhinitis, nasopharyngitis and pharyngitis]	J10_CHRONRHINITIS
J34[Other disorders of nose and nasal sinuses]	not_found
J47[Bronchiectasis]	J10_BRONCHIECTASIS

J84[Other interstitial pulmonary diseases]	ILD_ENDPOINTS
K04[Diseases of pulp and periapical tissues]	K11_PULP_PERIAPICAL
K05[Gingivitis and periodontal diseases]	K11_GINGIVITIS_PERIODONTAL
K11[Diseases of salivary glands]	K11_SALIVARY
K13[Other diseases of lip and oral mucosa]	K11_LIP_ORAL_MUCOSA
K14[Diseases of tongue]	TONGUEABNORMALITY
K21[Gastro-oesophageal reflux disease]	K11_REFLUX
K44[Diaphragmatic hernia]	K11_DIAHER
K52[Other noninfective gastroenteritis and colitis]	K11_OTHENTERCOL
K57[Diverticular disease of intestine]	K11_DIVERTIC
K63[Other diseases of intestine]	K11_INTESTOTH
K76[Other diseases of liver]	K11_DISLIVOTH
L08[Other local infections of skin and subcutaneous tissue]	L12_INFECT_SKIN_OTHER
L20[Atopic dermatitis]	L12_ATOPIC
L21[Seborrhoeic dermatitis]	L12_SEBORRHOEIC
L50[Urticaria]	L12_URTICARIA
L57[Skin changes due to chronic exposure to nonionizing radiation]	L12_ACTINKERA
L73[Other follicular disorders]	L12_FOLICULAROTH
L82[Seborrhoeic keratosis]	L12_SEBORRKERAT
Learning disability	not found
Leukaemia	CD2_PRIMARY_LYMPHOID_HEMATOPOIETIC_EXALLC
M23[Internal derangement of knee]	M13_KNEEDERANGEMENTS
M41[Scoliosis]	M13_SCOLIOSIS
M50[Cervical disc disorders]	M13_CERVICDISC
M53[Other dorsopathies, not elsewhere classified]	M13_DORSOPATHYOTH
M65[Synovitis and tenosynovitis]	M13_TENDOSYNOVITISNAS
M70[Soft tissue disorders related to use, overuse and pressure]	M13_SOFTOVERUSE
M72[Fibroblastic disorders]	M13_FIBROBLASTIC
M85[Other disorders of bone density and structure]	M13_BONEDENSOTH
N10[Acute tubulo-interstitial nephritis]	N14_PYELONEPHR
N12[Tubulo-interstitial nephritis, not specified as acute or chronic]	N14_TUBULOINTNEPHRNAS
N28[Other disorders of kidney and ureter, not elsewhere classified]	N14_KIDUROTH
N48[Other disorders of penis]	N14_OTHPENIS
N92[Excessive, frequent and irregular menstruation]	N14_MESNRUIRREG
N97[Female infertility]	N14_FEMALEINFERT
Non.alcoholic fatty liver disease and steatohepatitis	NAFLD
Non.Hodgkin_Lymphoma	C3_NONHODGKIN_EXALLC
O20[Haemorrhage in early pregnancy]	O15_HAEMORRH_EARLY_PREG
O32[Maternal care for known or suspected malpresentation of fetus]	O15_MAT_CARE_MALPRESENT

O48[Prolonged pregnancy]	O15_PREG_PROLONGED
O80[Single spontaneous delivery]	O15_DELIV_SPONT
O90[Complications of the puerperium, not elsewhere classified]	O15_PUERP_COMPL_NAS
O99[Other maternal diseases classifiable elsewhere but complicating pregnancy, childbirth and the puerperium]	O15_OTHER_MATERN_DIS_ELSEWHERE
Obesity	E4_OBESITY
Obsessive compulsive disorder	F5_OCD
Oesophagitis and oesophageal ulcer	K11_OESITIS,K11_OESULC
Osteoporosis	M13_OSTEOPOROSIS
Other anaemias	D3_ANAEMIANAS,D3_HAEMOLYTICANAEMIA
Other haemolytic anaemias	D3_HAEMOLYTICANAEMIA,D3_ANAEMIANAS,E4_PORBILIR
Peptic ulcer disease	K11_PULC
Peripheral arterial disease	I9_ATHSCLE
Primary Malignancy Other	not found
Q61[Cystic kidney disease]	Q17_CYSTIC_KIDNEY_DISEA,N14_CHRONKIDNEYDIS
Retinal vascular occlusions	H7_RETIVASCOCCLUSION
Scoliosis	M13_SCOLIOSIS
Seborrheic dermatitis	L12_SEBORRHOEIC
Spinal stenosis	M13_SPINSTENOSIS
Thalassaemia	D3_HAEMOLYTICANAEMIA,D3_ANAEMIANAS
Type 2 Diabetes	T2D

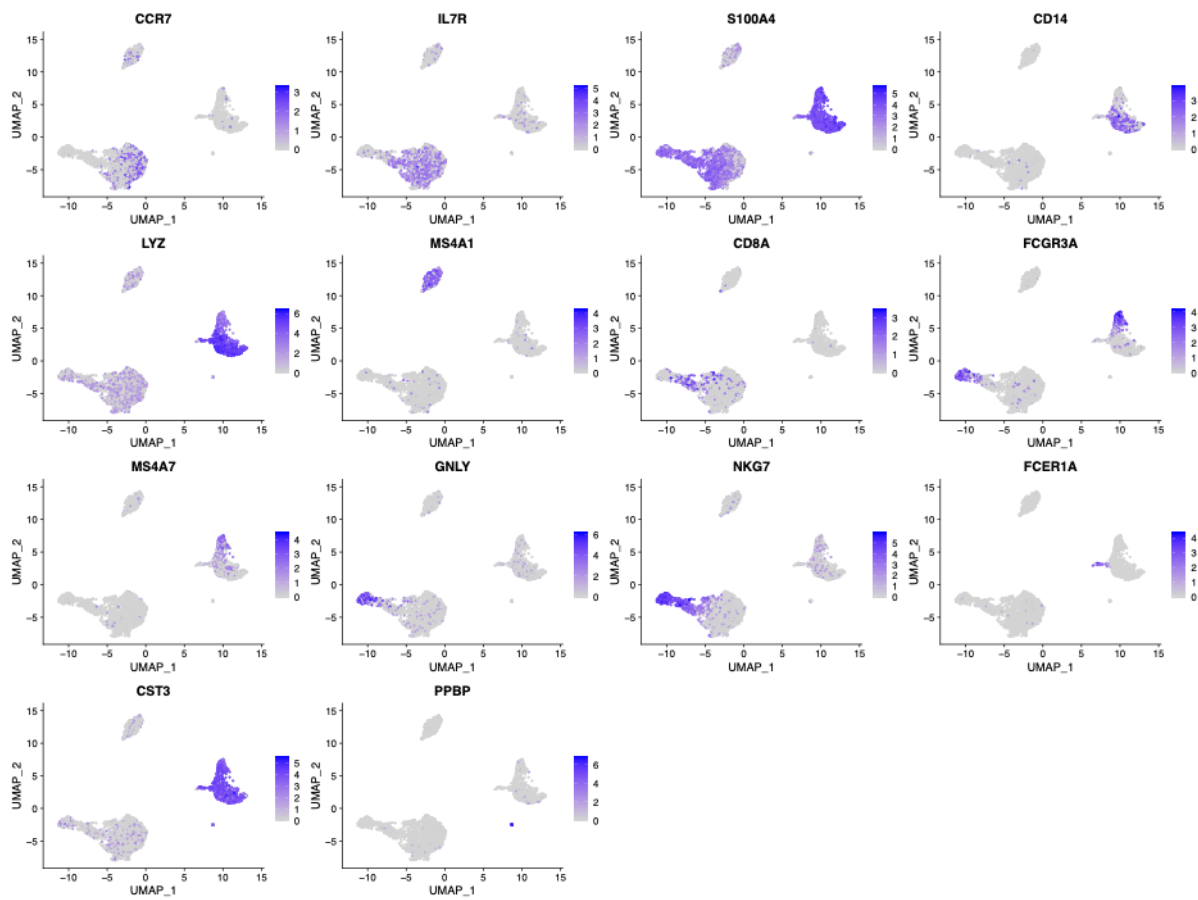
Appendix B - Supplementary data for Chapter 2

Supplementary figures



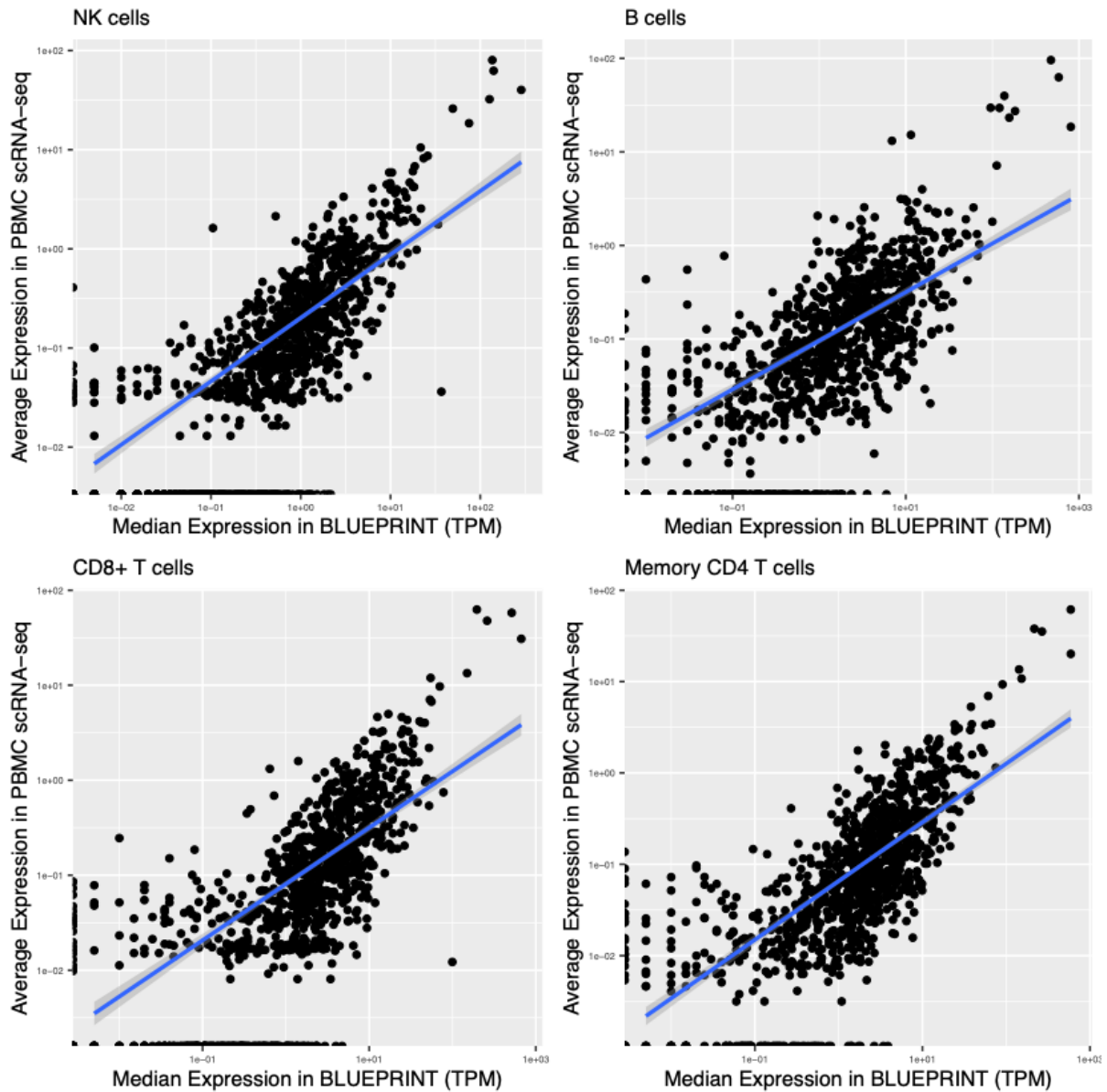
Supplementary-figure-3.1

Distribution of pRec between genes affected by hom-pLoFs and the rest of the protein coding genes. There is no significant difference in the distribution of pRec in genes with hom-pLoFs in G&H compared to the remaining protein-coding genes (Wilcoxon one-sided p-value > 0.05).



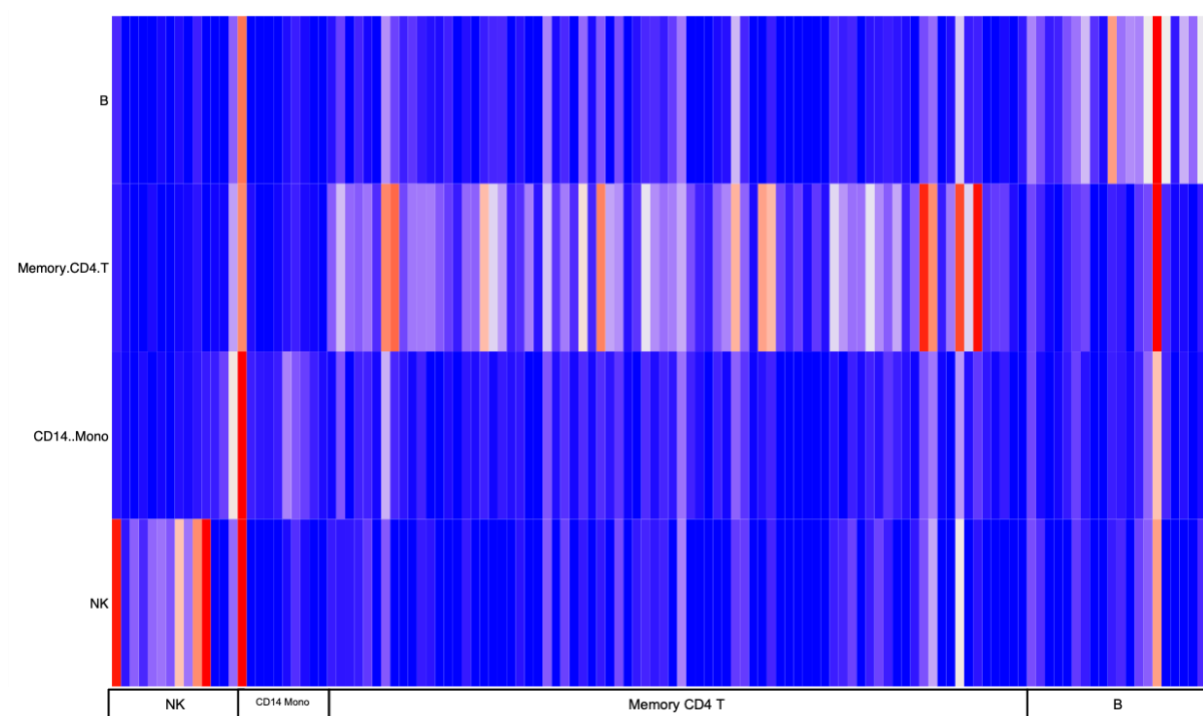
Supplementary-figure-3.3

Feature plots of marker gene expressions in the PBMC scRNAseq UMAP.



Supplementary-figure-3.4.

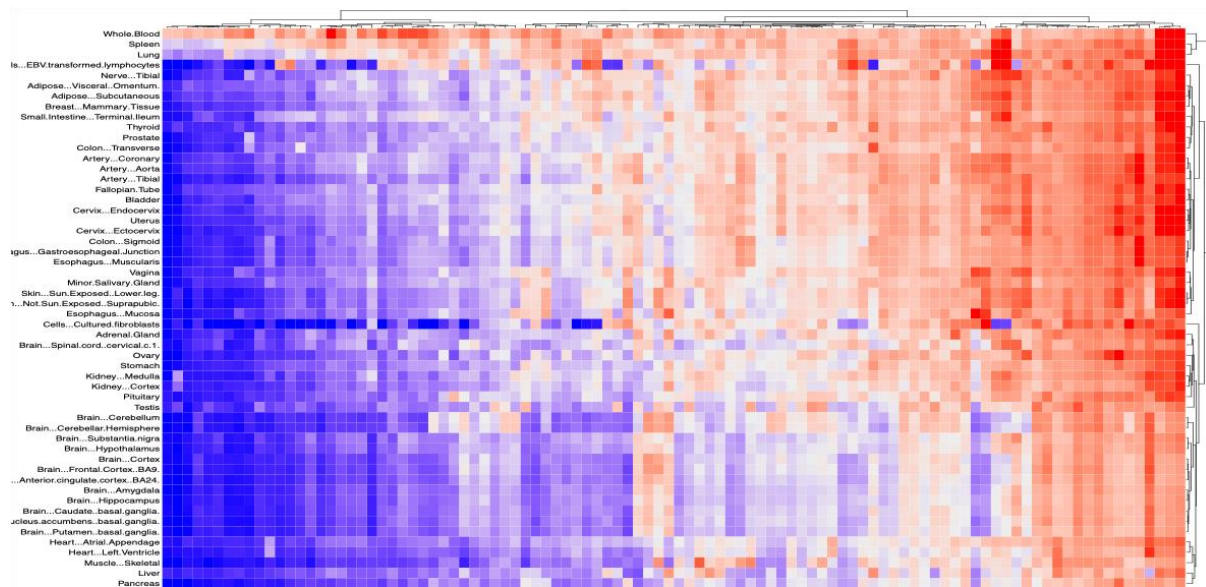
Correlation in cell-type clusters between PBMC scRNAseq and BLUEPRINT. Along with Figure-3.5B, the correlation analyses of the Average Expression of genes in PMBC scRNAseq against the median TPM of genes across samples in BLUEPRINT.



Supplementary-figure-3.5

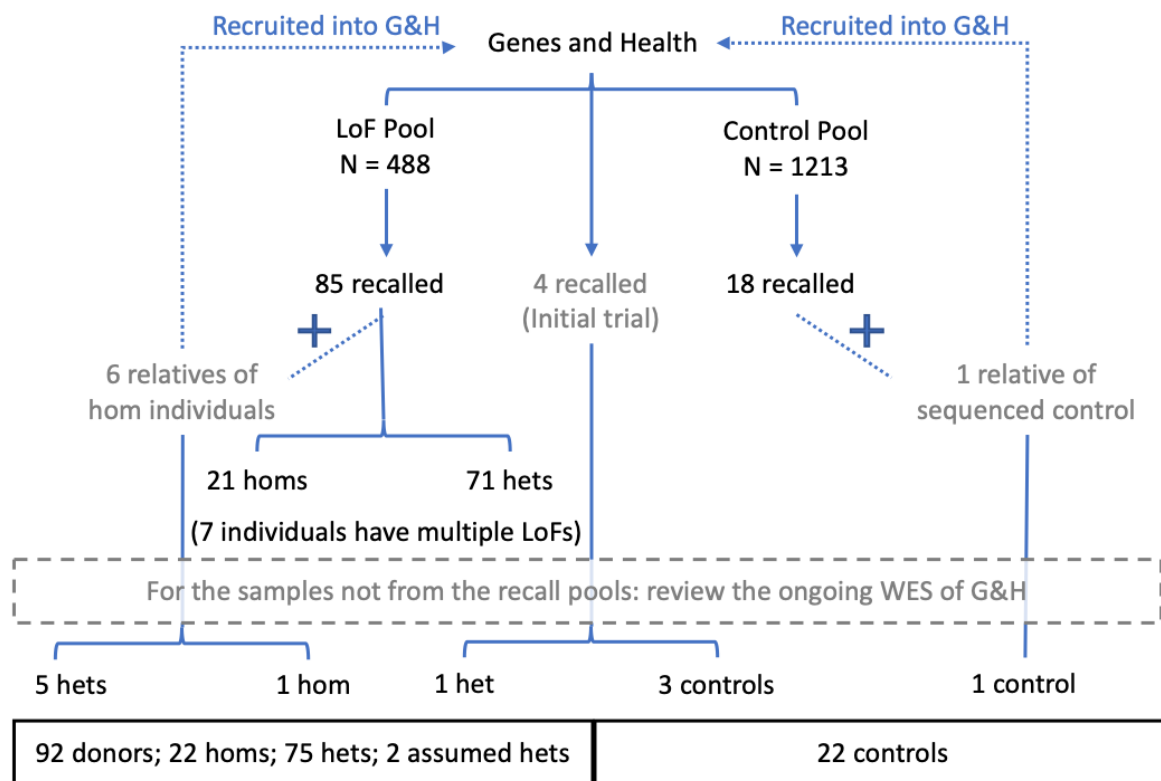
Heatmap of gene affected by hom-pLoFs and their expression in blood.

$\log_2(\text{AverageExpression} + 1)$ of genes with cell-type-specificity in both PBMC scRNAseq and BLUEPRINT expression data.



Supplementary-figure-3.6.

Heatmap of the $\log_2(\text{MedianTPM} + 1)$ of the top 100 genes affected by hom-pLoFs expressed in whole blood.



Supplementary-figure-3.7

Evaluating the pLoF status of the 114 recalled individuals.

Supplementary tables

Criteria								Results (n)		
GQ	Overlaps ROH	AF<0.01	All transcripts affected	pLI≤0.9	PBMC Detectable	pCRISPR	L2G	Sam- ples	Hom- pLoFs	Genes
≥20	Y	Y		Y	Y			361	301	267
≥20	Y	Y	Y	Y	Y			112	100	93
≥30	Y	Y	Y	Y	Y			98	90	85
≥20	Y	Y	Y	Y	Y	Y		2	2	2
≥20	Y	Y	Y	Y	Y		≥0.5	99	88	81
≥20	Y	Y	Y	Y	Y		0.5- 0.8	86	76	70
≥20	Y	Y	Y	Y	Y		≥0.8	28	25	23

Supplementary-table-3.1

Multiple trials of prioritisation with different criteria. The number of samples (n_samples), hom-pLoFs (n_variants), and genes affected (n_genes) when different criteria were applied in several trials of prioritisation. Changes to the criteria at each trial are highlighted in bold.

Supplementary-table-3.2

(Below) Table of pLoF effects on corrected gene expression in all cell-types, tested at different levels (by gene - if no variant is listed - or by pLoF variant), and with different models of association (additive and dominant):

Gene	LoF	Cell-type	Corrected Expression	Group	Additive Beta	Additive p-value	Dominant Beta	Dominant p-value
CAT	chr11 34456818 G C	CD4T	6.92	pLoF	-2.0E-01	3.0E-02	-2.0E-01	3.0E-02
CAT	chr11 34456818 G C	B	4.78	pLoF	-3.3E-01	1.8E-02	-3.3E-01	1.8E-02
CAT	chr11 34456818 G C	DC	5.23	pLoF	-5.0E-01	1.7E-02	-5.0E-01	1.7E-02
CAT	chr11 34456818 G C	Mono	8.40	pLoF	-4.5E-01	1.9E-05	-4.5E-01	1.9E-05
CAT	chr11 34456818 G C	NK	6.34	pLoF	-2.4E-01	1.5E-04	-2.4E-01	1.5E-04
CAT	chr11 34456818 G C	CD8T	6.24	pLoF	-3.2E-01	1.9E-03	-3.2E-01	1.9E-03
CAT	chr11 34456818 G C	Totter	3.32	pLoF	-1.4E-01	3.1E-01	-1.4E-01	3.1E-01
CD109		CD8T	4.82	Gene	-3.1E-02	5.2E-01	-7.1E-02	3.3E-01
CD109		B	2.92	Gene	-3.2E-02	3.2E-01	-4.0E-02	4.2E-01
CD109		CD4T	5.19	Gene	2.7E-02	5.6E-01	1.1E-02	8.7E-01
CD109		Totter	2.42	Gene	-2.4E-03	9.3E-01	-1.7E-02	6.6E-01
CD109		DC	3.01	Gene	-4.2E-02	6.2E-01	-7.4E-02	5.5E-01
CD109		Mono	5.42	Gene	-7.5E-02	2.1E-01	-6.4E-02	4.7E-01
CD109		NK	5.31	Gene	3.2E-02	4.6E-01	5.6E-02	3.8E-01
CD109	chr6 73736509 G A	CD8T	4.82	pLoF	3.3E-02	5.8E-01	-4.0E-03	9.6E-01
CD109	chr6 73736509 G A	B	2.92	pLoF	-2.6E-02	5.4E-01	-2.5E-02	6.7E-01
CD109	chr6 73736509 G A	CD4T	5.19	pLoF	6.9E-02	1.7E-01	7.2E-02	3.1E-01
CD109	chr6 73736509 G A	Totter	2.42	pLoF	9.3E-03	7.7E-01	-8.4E-03	8.5E-01
CD109	chr6 73736509 G A	DC	3.01	pLoF	6.0E-02	5.1E-01	4.3E-02	7.3E-01
CD109	chr6 73736509 G A	Mono	5.42	pLoF	-1.2E-01	8.2E-02	-1.2E-01	1.9E-01
CD109	chr6 73736509 G A	NK	5.31	pLoF	3.4E-02	5.1E-01	6.3E-02	3.8E-01
CD109	chr6 73782709 C T	CD8T	4.82	pLoF	-9.9E-02	3.3E-01	-2.0E-01	3.3E-01
CD109	chr6 73782709 C T	B	2.92	pLoF	-4.9E-02	5.2E-01	-9.9E-02	5.2E-01
CD109	chr6 73782709 C T	CD4T	5.19	pLoF	-3.2E-02	7.0E-01	-6.4E-02	7.0E-01
CD109	chr6 73782709 C T	Totter	2.42	pLoF	-3.5E-02	5.4E-01	-6.9E-02	5.4E-01
CD109	chr6 73782709 C T	DC	3.01	pLoF	-2.6E-01	1.1E-01	-5.2E-01	1.1E-01
CD109	chr6 73782709 C T	Mono	5.42	pLoF	-4.7E-02	4.1E-01	-9.5E-02	4.1E-01
CD109	chr6 73782709 C T	NK	5.31	pLoF	2.3E-02	7.7E-01	4.7E-02	7.7E-01
CDC25B	chr20 3787140 G GT	NK	6.60	pLoF	6.8E-02	1.3E-01	7.7E-02	1.7E-01
CDC25B	chr20 3787140 G GT	Totter	3.84	pLoF	1.5E-02	8.4E-01	8.0E-02	3.9E-01
CDC25B	chr20 3787140 G GT	Mono	6.09	pLoF	-1.2E-01	1.4E-01	-5.7E-02	5.7E-01
CDC25B	chr20 3787140 G GT	DC	3.12	pLoF	6.4E-03	9.5E-01	3.7E-02	7.6E-01
CDC25B	chr20 3787140 G GT	CD8T	6.94	pLoF	5.6E-02	2.3E-01	2.2E-02	7.1E-01
CDC25B	chr20 3787140 G GT	B	4.27	pLoF	-1.9E-02	8.7E-01	1.3E-02	9.3E-01
CDC25B	chr20 3787140 G GT	CD4T	7.88	pLoF	7.9E-02	2.1E-01	8.8E-02	2.6E-01
CST7	chr20 24957358 G T	Mono	6.65	pLoF	1.3E-01	3.1E-01	1.3E-01	3.1E-01
CST7	chr20 24957358 G T	DC	3.92	pLoF	-2.6E-01	3.4E-01	-2.6E-01	3.4E-01
CST7	chr20 24957358 G T	CD8T	9.93	pLoF	-7.6E-01	2.8E-03	-7.6E-01	2.8E-03
CST7	chr20 24957358 G T	NK	9.98	pLoF	-9.4E-01	9.1E-08	-9.4E-01	9.1E-08
CST7	chr20 24957358 G T	Totter	6.75	pLoF	-1.3E+00	5.8E-06	-1.3E+00	5.8E-06
CST7	chr20 24957358 G T	CD4T	8.21	pLoF	-3.0E-01	3.7E-01	-3.0E-01	3.7E-01
CST7	chr20 24957358 G T	B	4.49	pLoF	-1.0E-01	6.6E-01	-1.0E-01	6.6E-01
DDX58		Mono	7.08	Gene	-3.2E-01	1.3E-03	-3.2E-01	9.8E-03
DDX58		B	4.25	Gene	-2.4E-01	9.5E-03	-3.0E-01	8.4E-03
DDX58		CD4T	7.02	Gene	-3.3E-01	2.6E-06	-3.4E-01	1.4E-04
DDX58		DC	3.42	Gene	-2.5E-01	4.1E-02	-3.1E-01	3.3E-02
DDX58		CD8T	6.45	Gene	-3.0E-01	3.2E-06	-3.4E-01	1.4E-05
DDX58		Totter	3.41	Gene	-1.7E-01	7.4E-02	-2.2E-01	5.3E-02
DDX58		NK	6.41	Gene	-1.5E-01	6.0E-03	-2.0E-01	3.0E-03
DDX58	chr9 32466355 C A	Mono	7.08	pLoF	-1.5E-01	2.5E-01	-1.5E-01	2.5E-01
DDX58	chr9 32466355 C A	B	4.25	pLoF	-2.5E-01	6.6E-02	-2.5E-01	6.6E-02
DDX58	chr9 32466355 C A	CD4T	7.02	pLoF	-2.7E-01	3.3E-03	-2.7E-01	3.3E-03
DDX58	chr9 32466355 C A	DC	3.42	pLoF	-2.1E-01	2.3E-01	-2.1E-01	2.3E-01
DDX58	chr9 32466355 C A	CD8T	6.45	pLoF	-3.7E-01	4.8E-05	-3.7E-01	4.8E-05
DDX58	chr9 32466355 C A	Totter	3.41	pLoF	-2.9E-01	4.1E-02	-2.9E-01	4.1E-02
DDX58	chr9 32466355 C A	NK	6.41	pLoF	-1.7E-01	3.2E-02	-1.7E-01	3.2E-02

DDX58	chr9_32488757_GA_G	Mono	7.08	pLoF	-4.3E-01	8.8E-04	-6.1E-01	1.0E-03
DDX58	chr9_32488757_GA_G	B	4.25	pLoF	-3.0E-01	4.8E-05	-4.4E-01	2.1E-05
DDX58	chr9_32488757_GA_G	CD4T	7.02	pLoF	-3.7E-01	7.1E-05	-4.5E-01	9.3E-04
DDX58	chr9_32488757_GA_G	DC	3.42	pLoF	-3.6E-01	1.8E-02	-5.7E-01	6.6E-03
DDX58	chr9_32488757_GA_G	CD8T	6.45	pLoF	-2.3E-01	7.6E-04	-2.7E-01	6.3E-03
DDX58	chr9_32488757_GA_G	Tother	3.41	pLoF	-1.2E-01	1.6E-01	-1.2E-01	3.4E-01
DDX58	chr9_32488757_GA_G	NK	6.41	pLoF	-1.5E-01	4.3E-03	-2.5E-01	6.2E-04
HPSE	chr4_83302158_G_GT	NK	5.51	pLoF	-1.5E-01	1.3E-02	-1.8E-01	2.9E-02
HPSE	chr4_83302158_G_GT	B	3.18	pLoF	-2.0E-01	1.4E-02	-2.6E-01	1.5E-02
HPSE	chr4_83302158_G_GT	DC	2.87	pLoF	-1.0E-01	3.1E-01	-9.1E-02	5.1E-01
HPSE	chr4_83302158_G_GT	CD8T	5.21	pLoF	-1.6E-01	9.4E-03	-2.5E-01	1.2E-03
HPSE	chr4_83302158_G_GT	Tother	2.62	pLoF	-1.0E-01	2.2E-01	-1.1E-01	3.4E-01
HPSE	chr4_83302158_G_GT	CD4T	5.86	pLoF	-2.3E-01	2.2E-03	-2.8E-01	6.3E-03
HPSE	chr4_83302158_G_GT	Mono	7.06	pLoF	-1.4E-01	7.8E-02	-1.2E-01	2.4E-01
IFIH1	chr2_162277515_AC_A	Tother	3.17	pLoF	-2.3E-02	7.8E-01	-1.1E-01	3.0E-01
IFIH1	chr2_162277515_AC_A	Mono	7.29	pLoF	-2.7E-01	3.2E-02	-3.0E-01	6.4E-02
IFIH1	chr2_162277515_AC_A	CD8T	5.98	pLoF	-1.7E-01	2.0E-02	-2.4E-01	1.1E-02
IFIH1	chr2_162277515_AC_A	DC	3.85	pLoF	-2.0E-01	2.2E-01	-3.1E-01	1.3E-01
IFIH1	chr2_162277515_AC_A	NK	6.03	pLoF	-3.5E-01	2.7E-05	-3.9E-01	4.7E-04
IFIH1	chr2_162277515_AC_A	B	4.49	pLoF	-1.9E-01	3.4E-02	-3.1E-01	4.1E-03
IFIH1	chr2_162277515_AC_A	CD4T	6.78	pLoF	-1.8E-01	6.1E-02	-3.0E-01	1.2E-02
MNDA	chr1_158845998_C_T	B	5.06	pLoF	-1.5E-01	2.7E-01	-9.1E-02	5.6E-01
MNDA	chr1_158845998_C_T	DC	6.70	pLoF	-1.0E+00	4.1E-09	-1.1E+00	3.8E-07
MNDA	chr1_158845998_C_T	CD4T	6.79	pLoF	-2.4E-01	1.2E-01	-3.1E-01	8.2E-02
MNDA	chr1_158845998_C_T	CD8T	6.29	pLoF	-2.9E-01	7.1E-03	-3.6E-01	3.9E-03
MNDA	chr1_158845998_C_T	Mono	11.52	pLoF	-7.8E-01	8.4E-11	-7.4E-01	4.0E-06
MNDA	chr1_158845998_C_T	NK	6.21	pLoF	-8.1E-02	3.2E-01	-9.1E-02	3.3E-01
MNDA	chr1_158845998_C_T	Tother	3.29	pLoF	-1.6E-01	1.2E-01	-2.3E-01	5.4E-02
MYO1G	chr7_44964942_TAC_T	DC	4.09	pLoF	1.9E-01	6.1E-01	1.9E-01	6.1E-01
MYO1G	chr7_44964942_TAC_T	Mono	8.44	pLoF	-1.2E-01	3.0E-01	-1.2E-01	3.0E-01
MYO1G	chr7_44964942_TAC_T	Tother	4.09	pLoF	2.0E-02	9.4E-01	2.0E-02	9.4E-01
MYO1G	chr7_44964942_TAC_T	CD4T	7.78	pLoF	-3.8E-02	7.8E-01	-3.8E-02	7.8E-01
MYO1G	chr7_44964942_TAC_T	CD8T	7.43	pLoF	-2.0E-01	3.0E-01	-2.0E-01	3.0E-01
MYO1G	chr7_44964942_TAC_T	NK	7.16	pLoF	-1.4E-01	2.2E-01	-1.4E-01	2.2E-01
MYO1G	chr7_44964942_TAC_T	B	5.02	pLoF	3.2E-01	2.8E-01	3.2E-01	2.8E-01
NOD2	chr16_50716612_C_T	DC	3.34	pLoF	2.3E-02	8.9E-01	2.3E-02	8.9E-01
NOD2	chr16_50716612_C_T	Mono	6.74	pLoF	-1.7E-01	5.1E-02	-1.7E-01	5.1E-02
NOD2	chr16_50716612_C_T	NK	5.47	pLoF	-6.9E-02	5.5E-01	-6.9E-02	5.5E-01
NOD2	chr16_50716612_C_T	Tother	2.50	pLoF	-8.9E-02	3.7E-01	-8.9E-02	3.7E-01
NOD2	chr16_50716612_C_T	CD8T	5.06	pLoF	-1.9E-01	8.5E-02	-1.9E-01	8.5E-02
NOD2	chr16_50716612_C_T	B	3.22	pLoF	8.5E-02	5.4E-01	8.5E-02	5.4E-01
NOD2	chr16_50716612_C_T	CD4T	5.45	pLoF	-2.4E-01	1.2E-01	-2.4E-01	1.2E-01
SIGLEC14		Tother	2.49	Gene	-3.6E-02	3.0E-01	-7.3E-02	3.0E-01
SIGLEC14		DC	3.41	Gene	-3.1E-02	5.4E-01	-6.3E-02	5.4E-01
SIGLEC14		B	3.46	Gene	-6.3E-02	3.2E-01	-1.3E-01	3.2E-01
SIGLEC14		CD8T	5.00	Gene	9.2E-03	8.5E-01	1.8E-02	8.5E-01
SIGLEC14		CD4T	5.20	Gene	-6.4E-02	1.9E-01	-1.3E-01	1.9E-01
SIGLEC14		Mono	6.98	Gene	-5.4E-02	4.2E-01	-1.1E-01	4.2E-01
SIGLEC14		NK	5.35	Gene	-1.5E-02	6.5E-01	-2.9E-02	6.5E-01
SIGLEC14	chr19_51645475_A_G	Tother	2.49	pLoF	-5.5E-02	1.4E-01	-1.1E-01	1.4E-01
SIGLEC14	chr19_51645475_A_G	DC	3.41	pLoF	-1.3E-02	8.1E-01	-2.7E-02	8.1E-01
SIGLEC14	chr19_51645475_A_G	B	3.46	pLoF	-7.8E-02	2.6E-01	-1.6E-01	2.6E-01
SIGLEC14	chr19_51645475_A_G	CD8T	5.00	pLoF	-4.2E-02	3.5E-01	-8.3E-02	3.5E-01
SIGLEC14	chr19_51645475_A_G	CD4T	5.20	pLoF	-1.1E-01	2.0E-02	-2.2E-01	2.0E-02
SIGLEC14	chr19_51645475_A_G	Mono	6.98	pLoF	-2.4E-02	7.4E-01	-4.8E-02	7.4E-01
SIGLEC14	chr19_51645475_A_G	NK	5.35	pLoF	-3.8E-02	2.4E-01	-7.6E-02	2.4E-01
SIGLEC14	chr19_51643777_ACT_A	Tother	2.49	pLoF	5.4E-02	4.8E-01	1.1E-01	4.8E-01
SIGLEC14	chr19_51643777_ACT_A	DC	3.41	pLoF	-1.1E-01	3.4E-01	-2.1E-01	3.4E-01
SIGLEC14	chr19_51643777_ACT_A	B	3.46	pLoF	2.6E-03	9.8E-01	5.3E-03	9.8E-01
SIGLEC14	chr19_51643777_ACT_A	CD8T	5.00	pLoF	2.4E-01	8.0E-03	4.8E-01	8.0E-03
SIGLEC14	chr19_51643777_ACT_A	CD4T	5.20	pLoF	1.6E-01	6.7E-02	3.3E-01	6.7E-02
SIGLEC14	chr19_51643777_ACT_A	Mono	6.98	pLoF	-2.0E-01	1.7E-01	-3.9E-01	1.7E-01
SIGLEC14	chr19_51643777_ACT_A	NK	5.35	pLoF	9.3E-02	1.7E-01	1.9E-01	1.7E-01
SIGLEC5	chr19_51628003_A_AG	NK	5.28	pLoF	-3.3E-02	4.1E-01	-5.2E-02	3.0E-01
SIGLEC5	chr19_51628003_A_AG	Tother	2.45	pLoF	8.8E-03	8.0E-01	2.6E-02	5.5E-01
SIGLEC5	chr19_51628003_A_AG	CD4T	4.98	pLoF	-7.8E-02	2.1E-01	-9.1E-02	2.6E-01

SIGLEC5	chr19_51628003_A_AG	Mono	6.27	pLoF	-2.8E-01	2.6E-03	-3.1E-01	1.3E-02
SIGLEC5	chr19_51628003_A_AG	B	3.05	pLoF	1.2E-01	4.9E-02	1.2E-01	1.4E-01
SIGLEC5	chr19_51628003_A_AG	CD8T	4.74	pLoF	-3.8E-02	4.9E-01	-3.5E-02	6.2E-01
SIGLEC5	chr19_51628003_A_AG	DC	3.08	pLoF	-2.8E-01	2.5E-02	-3.3E-01	3.9E-02
SLC9A3R1	chr17_74749171_G_T	Totter	4.68	pLoF	-2.4E-01	1.4E-02	-2.5E-01	4.8E-02
SLC9A3R1	chr17_74749171_G_T	CD8T	8.05	pLoF	-1.7E-01	2.2E-02	-1.8E-01	7.1E-02
SLC9A3R1	chr17_74749171_G_T	NK	7.80	pLoF	3.8E-02	5.5E-01	6.9E-02	3.9E-01
SLC9A3R1	chr17_74749171_G_T	B	4.86	pLoF	-1.3E-01	3.1E-01	-1.2E-01	4.7E-01
SLC9A3R1	chr17_74749171_G_T	CD4T	8.39	pLoF	-9.3E-02	1.7E-01	-8.0E-02	3.5E-01
SLC9A3R1	chr17_74749171_G_T	DC	4.00	pLoF	1.6E-01	1.3E-01	3.3E-01	1.1E-02
SLC9A3R1	chr17_74749171_G_T	Mono	7.53	pLoF	5.8E-02	3.7E-01	9.1E-02	2.6E-01
TMEM71		B	3.59	Gene	-8.9E-02	1.3E-01	-7.0E-02	4.4E-01
TMEM71		CD4T	7.21	Gene	1.3E-02	7.5E-01	-4.1E-02	5.1E-01
TMEM71		Totter	3.54	Gene	1.4E-02	8.2E-01	2.1E-02	8.2E-01
TMEM71		DC	3.43	Gene	8.7E-02	3.9E-01	1.2E-01	4.1E-01
TMEM71		Mono	7.13	Gene	3.0E-03	9.6E-01	-1.1E-02	9.0E-01
TMEM71		CD8T	6.37	Gene	7.1E-02	9.9E-02	8.6E-02	1.8E-01
TMEM71		NK	6.57	Gene	1.0E-01	1.1E-01	6.8E-02	4.8E-01
TMEM71	chr8_132722068_C_CAGCA	B	3.59	pLoF	4.0E-02	7.2E-01	4.0E-02	7.2E-01
TMEM71	chr8_132722068_C_CAGCA	CD4T	7.21	pLoF	-1.9E-01	1.5E-02	-1.9E-01	1.5E-02
TMEM71	chr8_132722068_C_CAGCA	Totter	3.54	pLoF	5.0E-02	7.1E-01	5.0E-02	7.1E-01
TMEM71	chr8_132722068_C_CAGCA	DC	3.43	pLoF	6.7E-02	7.5E-01	6.7E-02	7.5E-01
TMEM71	chr8_132722068_C_CAGCA	Mono	7.13	pLoF	-1.5E-01	1.8E-01	-1.5E-01	1.8E-01
TMEM71	chr8_132722068_C_CAGCA	CD8T	6.37	pLoF	-3.3E-02	6.8E-01	-3.3E-02	6.8E-01
TMEM71	chr8_132722068_C_CAGCA	NK	6.57	pLoF	-1.4E-01	2.4E-01	-1.4E-01	2.4E-01
TMEM71	chr8_132758873_G_A	B	3.59	pLoF	-1.2E-01	6.5E-02	-1.5E-01	1.9E-01
TMEM71	chr8_132758873_G_A	CD4T	7.21	pLoF	5.8E-02	1.4E-01	7.9E-02	2.6E-01
TMEM71	chr8_132758873_G_A	Totter	3.54	pLoF	9.1E-03	9.0E-01	2.4E-04	1.0E+00
TMEM71	chr8_132758873_G_A	DC	3.43	pLoF	1.0E-01	3.8E-01	1.9E-01	3.5E-01
TMEM71	chr8_132758873_G_A	Mono	7.13	pLoF	4.8E-02	4.4E-01	1.3E-01	2.5E-01
TMEM71	chr8_132758873_G_A	CD8T	6.37	pLoF	9.8E-02	3.9E-02	2.0E-01	1.8E-02
TMEM71	chr8_132758873_G_A	NK	6.57	pLoF	1.4E-01	1.9E-02	2.1E-01	5.2E-02
TRAT1	chr3_108849184_GCTATGAA_G	CD8T	6.81	pLoF	-1.2E-01	2.5E-01	-1.9E-01	1.3E-01
TRAT1	chr3_108849184_GCTATGAA_G	CD4T	9.01	pLoF	-2.6E-01	7.6E-05	-3.3E-01	7.2E-05
TRAT1	chr3_108849184_GCTATGAA_G	Mono	5.87	pLoF	-2.5E-03	9.6E-01	-3.4E-02	6.1E-01
TRAT1	chr3_108849184_GCTATGAA_G	DC	3.14	pLoF	1.8E-01	1.5E-01	8.6E-02	5.9E-01
TRAT1	chr3_108849184_GCTATGAA_G	Totter	3.83	pLoF	-2.0E-01	2.0E-02	-2.6E-01	1.4E-02
TRAT1	chr3_108849184_GCTATGAA_G	NK	5.84	pLoF	-1.8E-01	1.7E-02	-2.8E-01	2.2E-03
TRAT1	chr3_108849184_GCTATGAA_G	B	3.51	pLoF	-3.1E-02	7.8E-01	-3.8E-02	7.9E-01
TRIM58	chr1_247857547_G_T	CD8T	4.81	pLoF	-8.1E-02	4.7E-01	-8.1E-02	4.7E-01
TRIM58	chr1_247857547_G_T	Totter	2.41	pLoF	2.0E-02	8.1E-01	2.0E-02	8.1E-01
TRIM58	chr1_247857547_G_T	Mono	5.45	pLoF	-9.6E-02	5.9E-01	-9.6E-02	5.9E-01
TRIM58	chr1_247857547_G_T	NK	5.29	pLoF	-5.2E-02	4.9E-01	-5.2E-02	4.9E-01
TRIM58	chr1_247857547_G_T	B	3.26	pLoF	-3.2E-01	1.1E-01	-3.2E-01	1.1E-01
TRIM58	chr1_247857547_G_T	DC	2.65	pLoF	8.5E-02	4.2E-01	8.5E-02	4.2E-01
TRIM58	chr1_247857547_G_T	CD4T	5.07	pLoF	-2.1E-01	5.0E-02	-2.1E-01	5.0E-02
ZBP1	chr20_57615536_CA_C	CD4T	6.56	pLoF	-2.5E-01	1.3E-01	-2.5E-01	1.3E-01
ZBP1	chr20_57615536_CA_C	B	4.53	pLoF	-2.6E-01	5.2E-01	-2.6E-01	5.2E-01
ZBP1	chr20_57615536_CA_C	NK	6.33	pLoF	-1.2E-02	9.4E-01	-1.2E-02	9.4E-01
ZBP1	chr20_57615536_CA_C	CD8T	6.20	pLoF	-1.1E-01	4.5E-01	-1.1E-01	4.5E-01
ZBP1	chr20_57615536_CA_C	Mono	5.84	pLoF	-1.1E-01	5.8E-01	-1.1E-01	5.8E-01
ZBP1	chr20_57615536_CA_C	Totter	3.24	pLoF	1.2E-02	9.5E-01	1.2E-02	9.5E-01
ZBP1	chr20_57615536_CA_C	DC	2.88	pLoF	-2.5E-01	3.6E-01	-2.5E-01	3.6E-01
ZC3H8	chr2_112230930_CTTCTT_C	Totter	3.58	pLoF	-1.5E-03	9.8E-01	-1.6E-02	8.9E-01
ZC3H8	chr2_112230930_CTTCTT_C	B	4.47	pLoF	5.9E-02	5.6E-01	8.4E-02	6.2E-01
ZC3H8	chr2_112230930_CTTCTT_C	CD4T	7.54	pLoF	-2.3E-02	5.9E-01	-3.5E-02	6.2E-01
ZC3H8	chr2_112230930_CTTCTT_C	CD8T	6.54	pLoF	1.1E-02	8.7E-01	4.5E-02	6.8E-01
ZC3H8	chr2_112230930_CTTCTT_C	DC	3.58	pLoF	3.1E-02	8.3E-01	-7.9E-02	7.5E-01
ZC3H8	chr2_112230930_CTTCTT_C	Mono	6.42	pLoF	-7.7E-02	1.0E-01	-1.4E-01	8.3E-02
ZC3H8	chr2_112230930_CTTCTT_C	NK	6.44	pLoF	2.0E-02	8.0E-01	2.4E-02	8.5E-01

Supplementary-table-3.3

(Below) Direct pLoF effects on SCTransformed counts. “add_twodf” and “dd_twodf” are the coefficients for the additive component and the dominance deviation component of a genotypic test respectively.

Gene	pLoF	Cell-type	Beta	p-value	Test
CAT	chr11_34456818_G_C	B	-5.6E-02	3.9E-03	additive
CAT	chr11_34456818_G_C	B	-5.6E-02	3.9E-03	dominant
CAT	chr11_34456818_G_C	CD4T	-2.1E-02	1.4E-01	additive
CAT	chr11_34456818_G_C	CD4T	-2.1E-02	1.4E-01	dominant
CAT	chr11_34456818_G_C	CD8T	-3.1E-02	3.8E-02	additive
CAT	chr11_34456818_G_C	CD8T	-3.1E-02	3.8E-02	dominant
CAT	chr11_34456818_G_C	T_other	-2.6E-02	1.8E-01	additive
CAT	chr11_34456818_G_C	T_other	-2.6E-02	1.8E-01	dominant
CAT	chr11_34456818_G_C	NK	-4.1E-02	1.7E-02	additive
CAT	chr11_34456818_G_C	NK	-4.1E-02	1.7E-02	dominant
CAT	chr11_34456818_G_C	Mono	-1.5E-01	4.7E-02	additive
CAT	chr11_34456818_G_C	Mono	-1.5E-01	4.7E-02	dominant
CAT	chr11_34456818_G_C	DC	-1.7E-01	2.9E-02	additive
CAT	chr11_34456818_G_C	DC	-1.7E-01	2.9E-02	dominant
CD109	chr6_73736509_G_A	B	-7.0E-04	5.0E-01	additive
CD109	chr6_73736509_G_A	B	-7.4E-04	5.6E-01	dominant
CD109	chr6_73736509_G_A	B	-1.6E-03	5.7E-01	recessive
CD109	chr6_73736509_G_A	B	-8.5E-04	5.6E-01	add_twodf
CD109	chr6_73736509_G_A	B	3.0E-04	8.8E-01	dd_twodf
CD109	chr6_73736509_G_A	CD4T	6.2E-04	4.3E-01	additive
CD109	chr6_73736509_G_A	CD4T	4.4E-04	6.7E-01	dominant
CD109	chr6_73736509_G_A	CD4T	2.3E-03	2.4E-01	recessive
CD109	chr6_73736509_G_A	CD4T	1.2E-03	2.6E-01	add_twodf
CD109	chr6_73736509_G_A	CD4T	-1.2E-03	4.0E-01	dd_twodf
CD109	chr6_73736509_G_A	CD8T	5.9E-04	3.6E-01	additive
CD109	chr6_73736509_G_A	CD8T	1.4E-04	8.7E-01	dominant
CD109	chr6_73736509_G_A	CD8T	3.5E-03	3.5E-02	recessive
CD109	chr6_73736509_G_A	CD8T	1.7E-03	4.4E-02	add_twodf
CD109	chr6_73736509_G_A	CD8T	-2.3E-03	4.7E-02	dd_twodf
CD109	chr6_73736509_G_A	T_other	2.9E-04	8.0E-01	additive
CD109	chr6_73736509_G_A	T_other	-6.7E-04	6.5E-01	dominant
CD109	chr6_73736509_G_A	T_other	4.4E-03	1.3E-01	recessive
CD109	chr6_73736509_G_A	T_other	2.0E-03	1.6E-01	add_twodf
CD109	chr6_73736509_G_A	T_other	-3.9E-03	5.2E-02	dd_twodf
CD109	chr6_73736509_G_A	NK	8.6E-04	6.5E-01	additive
CD109	chr6_73736509_G_A	NK	9.4E-04	7.2E-01	dominant

CD109	chr6_73736509_G_A	NK	2.1E-03	6.4E-01	recessive
CD109	chr6_73736509_G_A	NK	1.1E-03	6.4E-01	add_twodf
CD109	chr6_73736509_G_A	NK	-6.5E-04	8.6E-01	dd_twodf
CD109	chr6_73736509_G_A	Mono	4.0E-04	8.9E-01	additive
CD109	chr6_73736509_G_A	Mono	3.4E-04	9.3E-01	dominant
CD109	chr6_73736509_G_A	Mono	1.3E-03	8.5E-01	recessive
CD109	chr6_73736509_G_A	Mono	6.3E-04	8.6E-01	add_twodf
CD109	chr6_73736509_G_A	Mono	-5.9E-04	9.1E-01	dd_twodf
CD109	chr6_73736509_G_A	DC	8.1E-03	5.2E-01	additive
CD109	chr6_73736509_G_A	DC	1.0E-02	5.4E-01	dominant
CD109	chr6_73736509_G_A	DC	1.5E-02	6.3E-01	recessive
CD109	chr6_73736509_G_A	DC	7.9E-03	6.1E-01	add_twodf
CD109	chr6_73736509_G_A	DC	1.1E-04	1.0E+00	dd_twodf
CD109	chr6_73758966_T_TTG	B	-1.5E-03	7.5E-01	additive
CD109	chr6_73758966_T_TTG	B	-1.7E-03	7.4E-01	dominant
CD109	chr6_73758966_T_TTG	B	-1.7E-03	9.4E-01	recessive
CD109	chr6_73758966_T_TTG	B	-8.5E-04	9.4E-01	add_twodf
CD109	chr6_73758966_T_TTG	B	-8.5E-04	9.4E-01	dd_twodf
CD109	chr6_73758966_T_TTG	CD4T	-2.6E-03	5.7E-01	additive
CD109	chr6_73758966_T_TTG	CD4T	-2.7E-03	5.7E-01	dominant
CD109	chr6_73758966_T_TTG	CD4T	-2.6E-03	9.5E-01	recessive
CD109	chr6_73758966_T_TTG	CD4T	-1.3E-03	9.5E-01	add_twodf
CD109	chr6_73758966_T_TTG	CD4T	-1.3E-03	9.5E-01	dd_twodf
CD109	chr6_73758966_T_TTG	CD8T	-1.6E-03	7.0E-01	additive
CD109	chr6_73758966_T_TTG	CD8T	-1.7E-03	7.0E-01	dominant
CD109	chr6_73758966_T_TTG	CD8T	-1.7E-03	9.4E-01	recessive
CD109	chr6_73758966_T_TTG	CD8T	-8.7E-04	9.4E-01	add_twodf
CD109	chr6_73758966_T_TTG	CD8T	-8.7E-04	9.4E-01	dd_twodf
CD109	chr6_73758966_T_TTG	T_other	-1.9E-03	8.0E-01	additive
CD109	chr6_73758966_T_TTG	T_other	-1.9E-03	8.0E-01	dominant
CD109	chr6_73758966_T_TTG	NK	-2.7E-03	7.5E-01	additive
CD109	chr6_73758966_T_TTG	NK	-3.0E-03	7.4E-01	dominant
CD109	chr6_73758966_T_TTG	NK	-2.9E-03	9.5E-01	recessive
CD109	chr6_73758966_T_TTG	NK	-1.5E-03	9.4E-01	add_twodf
CD109	chr6_73758966_T_TTG	NK	-1.5E-03	9.5E-01	dd_twodf
CD109	chr6_73758966_T_TTG	Mono	3.4E-02	1.7E-02	additive
CD109	chr6_73758966_T_TTG	Mono	3.4E-02	1.7E-02	dominant
CD109	chr6_73758966_T_TTG	DC	-3.9E-02	5.7E-01	additive
CD109	chr6_73758966_T_TTG	DC	-3.9E-02	5.7E-01	dominant
CD109	chr6_73782709_C_T	B	-8.5E-04	6.0E-01	additive
CD109	chr6_73782709_C_T	B	-1.7E-03	6.0E-01	dominant
CD109	chr6_73782709_C_T	B	-1.7E-03	6.0E-01	recessive

CD109	chr6_73782709_C_T	CD4T	-3.7E-04	7.7E-01	additive
CD109	chr6_73782709_C_T	CD4T	-7.5E-04	7.7E-01	dominant
CD109	chr6_73782709_C_T	CD4T	-7.5E-04	7.7E-01	recessive
CD109	chr6_73782709_C_T	CD8T	-8.7E-04	4.5E-01	additive
CD109	chr6_73782709_C_T	CD8T	-1.7E-03	4.5E-01	dominant
CD109	chr6_73782709_C_T	CD8T	-1.7E-03	4.5E-01	recessive
CD109	chr6_73782709_C_T	T_other	-9.4E-04	8.2E-01	additive
CD109	chr6_73782709_C_T	T_other	-1.9E-03	8.2E-01	dominant
CD109	chr6_73782709_C_T	T_other	-1.9E-03	8.2E-01	recessive
CD109	chr6_73782709_C_T	NK	-6.2E-04	8.5E-01	additive
CD109	chr6_73782709_C_T	NK	-1.2E-03	8.5E-01	dominant
CD109	chr6_73782709_C_T	NK	-1.2E-03	8.5E-01	recessive
CD109	chr6_73782709_C_T	Mono	-2.5E-03	5.6E-01	additive
CD109	chr6_73782709_C_T	Mono	-5.0E-03	5.6E-01	dominant
CD109	chr6_73782709_C_T	Mono	-5.0E-03	5.6E-01	recessive
CD109	chr6_73782709_C_T	DC	-1.9E-02	2.7E-01	additive
CD109	chr6_73782709_C_T	DC	-3.9E-02	2.7E-01	dominant
CD109	chr6_73782709_C_T	DC	-3.9E-02	2.7E-01	recessive
CDC25B	chr20_3787140_G_GT	B	2.9E-03	7.5E-01	additive
CDC25B	chr20_3787140_G_GT	B	2.9E-03	8.0E-01	dominant
CDC25B	chr20_3787140_G_GT	B	7.7E-03	7.6E-01	recessive
CDC25B	chr20_3787140_G_GT	B	4.0E-03	7.6E-01	add_twodf
CDC25B	chr20_3787140_G_GT	B	-2.2E-03	9.0E-01	dd_twodf
CDC25B	chr20_3787140_G_GT	CD4T	2.0E-02	1.6E-01	additive
CDC25B	chr20_3787140_G_GT	CD4T	2.0E-02	2.5E-01	dominant
CDC25B	chr20_3787140_G_GT	CD4T	5.2E-02	1.9E-01	recessive
CDC25B	chr20_3787140_G_GT	CD4T	2.7E-02	1.8E-01	add_twodf
CDC25B	chr20_3787140_G_GT	CD4T	-1.4E-02	6.0E-01	dd_twodf
CDC25B	chr20_3787140_G_GT	CD8T	3.7E-02	1.4E-02	additive
CDC25B	chr20_3787140_G_GT	CD8T	3.8E-02	4.7E-02	dominant
CDC25B	chr20_3787140_G_GT	CD8T	9.1E-02	2.9E-02	recessive
CDC25B	chr20_3787140_G_GT	CD8T	4.8E-02	2.0E-02	add_twodf
CDC25B	chr20_3787140_G_GT	CD8T	-2.2E-02	4.0E-01	dd_twodf
CDC25B	chr20_3787140_G_GT	T_other	9.8E-03	6.3E-01	additive
CDC25B	chr20_3787140_G_GT	T_other	2.2E-02	3.8E-01	dominant
CDC25B	chr20_3787140_G_GT	T_other	-3.3E-02	5.5E-01	recessive
CDC25B	chr20_3787140_G_GT	T_other	-1.4E-02	6.2E-01	add_twodf
CDC25B	chr20_3787140_G_GT	T_other	4.7E-02	2.1E-01	dd_twodf
CDC25B	chr20_3787140_G_GT	NK	2.2E-02	1.9E-01	additive
CDC25B	chr20_3787140_G_GT	NK	2.3E-02	2.6E-01	dominant
CDC25B	chr20_3787140_G_GT	NK	5.1E-02	2.6E-01	recessive
CDC25B	chr20_3787140_G_GT	NK	2.7E-02	2.4E-01	add_twodf

CDC25B	chr20_3787140_G_GT	NK	-9.7E-03	7.5E-01	dd_twodf
CDC25B	chr20_3787140_G_GT	Mono	-4.1E-03	3.7E-01	additive
CDC25B	chr20_3787140_G_GT	Mono	-2.2E-03	6.9E-01	dominant
CDC25B	chr20_3787140_G_GT	Mono	-2.2E-02	8.6E-02	recessive
CDC25B	chr20_3787140_G_GT	Mono	-1.1E-02	9.6E-02	add_twodf
CDC25B	chr20_3787140_G_GT	Mono	1.2E-02	1.5E-01	dd_twodf
CST7	chr20_24957358_G_T	B	-2.9E-02	3.0E-01	additive
CST7	chr20_24957358_G_T	B	-2.9E-02	3.0E-01	dominant
CST7	chr20_24957358_G_T	CD4T	-6.8E-02	2.2E-01	additive
CST7	chr20_24957358_G_T	CD4T	-6.8E-02	2.2E-01	dominant
CST7	chr20_24957358_G_T	CD8T	-2.9E-01	1.2E-01	additive
CST7	chr20_24957358_G_T	CD8T	-2.9E-01	1.2E-01	dominant
CST7	chr20_24957358_G_T	T_other	-2.0E-01	4.1E-01	additive
CST7	chr20_24957358_G_T	T_other	-2.0E-01	4.1E-01	dominant
CST7	chr20_24957358_G_T	NK	-4.4E-01	4.0E-04	additive
CST7	chr20_24957358_G_T	NK	-4.4E-01	4.0E-04	dominant
CST7	chr20_24957358_G_T	Mono	-2.0E-02	4.3E-01	additive
CST7	chr20_24957358_G_T	Mono	-2.0E-02	4.3E-01	dominant
CST7	chr20_24957358_G_T	DC	-5.6E-02	2.5E-01	additive
CST7	chr20_24957358_G_T	DC	-5.6E-02	2.5E-01	dominant
DDX58	chr9_32466355_C_A	B	6.3E-03	5.3E-01	additive
DDX58	chr9_32466355_C_A	B	6.3E-03	5.3E-01	dominant
DDX58	chr9_32466355_C_A	CD4T	-1.7E-02	1.1E-01	additive
DDX58	chr9_32466355_C_A	CD4T	-1.7E-02	1.1E-01	dominant
DDX58	chr9_32466355_C_A	CD8T	-1.8E-02	1.1E-01	additive
DDX58	chr9_32466355_C_A	CD8T	-1.8E-02	1.1E-01	dominant
DDX58	chr9_32466355_C_A	T_other	-1.9E-02	3.5E-01	additive
DDX58	chr9_32466355_C_A	T_other	-1.9E-02	3.5E-01	dominant
DDX58	chr9_32466355_C_A	NK	-5.5E-03	6.3E-01	additive
DDX58	chr9_32466355_C_A	NK	-5.5E-03	6.3E-01	dominant
DDX58	chr9_32466355_C_A	Mono	-3.4E-02	1.7E-01	additive
DDX58	chr9_32466355_C_A	Mono	-3.4E-02	1.7E-01	dominant
DDX58	chr9_32488757_GA_G	B	-1.3E-02	2.7E-02	additive
DDX58	chr9_32488757_GA_G	B	-1.9E-02	5.5E-02	dominant
DDX58	chr9_32488757_GA_G	B	-2.8E-02	2.3E-02	recessive
DDX58	chr9_32488757_GA_G	B	-1.4E-02	2.5E-02	add_twodf
DDX58	chr9_32488757_GA_G	B	1.0E-02	4.9E-01	dd_twodf
DDX58	chr9_32488757_GA_G	CD4T	-1.6E-02	4.7E-02	additive
DDX58	chr9_32488757_GA_G	CD4T	-1.9E-02	1.3E-01	dominant
DDX58	chr9_32488757_GA_G	CD4T	-3.8E-02	3.0E-02	recessive
DDX58	chr9_32488757_GA_G	CD4T	-1.9E-02	3.2E-02	add_twodf
DDX58	chr9_32488757_GA_G	CD4T	1.5E-02	3.6E-01	dd_twodf

DDX58	chr9_32488757_GA_G	CD8T	-1.2E-02	9.7E-02	additive
DDX58	chr9_32488757_GA_G	CD8T	-1.3E-02	2.7E-01	dominant
DDX58	chr9_32488757_GA_G	CD8T	-3.3E-02	3.9E-02	recessive
DDX58	chr9_32488757_GA_G	CD8T	-1.6E-02	4.6E-02	add_twodf
DDX58	chr9_32488757_GA_G	CD8T	2.0E-02	2.1E-01	dd_twodf
DDX58	chr9_32488757_GA_G	T_other	-1.3E-02	3.0E-01	additive
DDX58	chr9_32488757_GA_G	T_other	-1.3E-02	4.8E-01	dominant
DDX58	chr9_32488757_GA_G	T_other	-3.7E-02	2.0E-01	recessive
DDX58	chr9_32488757_GA_G	T_other	-1.8E-02	2.1E-01	add_twodf
DDX58	chr9_32488757_GA_G	T_other	1.9E-02	4.7E-01	dd_twodf
DDX58	chr9_32488757_GA_G	NK	-2.0E-02	3.3E-02	additive
DDX58	chr9_32488757_GA_G	NK	-2.6E-02	6.7E-02	dominant
DDX58	chr9_32488757_GA_G	NK	-4.1E-02	4.1E-02	recessive
DDX58	chr9_32488757_GA_G	NK	-2.1E-02	4.0E-02	add_twodf
DDX58	chr9_32488757_GA_G	NK	7.4E-03	6.9E-01	dd_twodf
DDX58	chr9_32488757_GA_G	Mono	-4.5E-02	1.8E-02	additive
DDX58	chr9_32488757_GA_G	Mono	-7.1E-02	1.5E-02	dominant
DDX58	chr9_32488757_GA_G	Mono	-7.3E-02	9.1E-02	recessive
DDX58	chr9_32488757_GA_G	Mono	-4.0E-02	5.9E-02	add_twodf
DDX58	chr9_32488757_GA_G	Mono	-2.6E-02	5.2E-01	dd_twodf
DDX58	chr9_32488757_GA_G	DC	-2.3E-02	4.5E-01	recessive
HPSE	chr4_83302158_G_GT	B	-3.5E-03	1.1E-01	additive
HPSE	chr4_83302158_G_GT	B	-4.6E-03	1.4E-01	dominant
HPSE	chr4_83302158_G_GT	B	-6.5E-03	2.2E-01	recessive
HPSE	chr4_83302158_G_GT	B	-3.4E-03	2.0E-01	add_twodf
HPSE	chr4_83302158_G_GT	B	-1.3E-04	9.8E-01	dd_twodf
HPSE	chr4_83302158_G_GT	CD4T	-5.0E-03	3.0E-02	additive
HPSE	chr4_83302158_G_GT	CD4T	-6.4E-03	4.5E-02	dominant
HPSE	chr4_83302158_G_GT	CD4T	-9.7E-03	8.7E-02	recessive
HPSE	chr4_83302158_G_GT	CD4T	-5.1E-03	7.2E-02	add_twodf
HPSE	chr4_83302158_G_GT	CD4T	2.2E-04	9.6E-01	dd_twodf
HPSE	chr4_83302158_G_GT	CD8T	-2.9E-03	7.8E-02	additive
HPSE	chr4_83302158_G_GT	CD8T	-4.4E-03	4.4E-02	dominant
HPSE	chr4_83302158_G_GT	CD8T	-2.3E-03	6.0E-01	recessive
HPSE	chr4_83302158_G_GT	CD8T	-1.4E-03	5.1E-01	add_twodf
HPSE	chr4_83302158_G_GT	CD8T	-3.5E-03	2.6E-01	dd_twodf
HPSE	chr4_83302158_G_GT	T_other	-1.6E-03	7.9E-01	additive
HPSE	chr4_83302158_G_GT	T_other	-8.6E-04	9.0E-01	dominant
HPSE	chr4_83302158_G_GT	T_other	-9.9E-03	5.9E-01	recessive
HPSE	chr4_83302158_G_GT	T_other	-4.9E-03	6.0E-01	add_twodf
HPSE	chr4_83302158_G_GT	T_other	5.5E-03	6.4E-01	dd_twodf
HPSE	chr4_83302158_G_GT	NK	-2.5E-03	4.1E-01	additive

HPSE	chr4_83302158_G_GT	NK	-2.3E-03	5.3E-01	dominant
HPSE	chr4_83302158_G_GT	NK	-7.8E-03	3.7E-01	recessive
HPSE	chr4_83302158_G_GT	NK	-4.0E-03	3.6E-01	add_twodf
HPSE	chr4_83302158_G_GT	NK	2.7E-03	6.4E-01	dd_twodf
HPSE	chr4_83302158_G_GT	Mono	-2.3E-02	3.1E-01	additive
HPSE	chr4_83302158_G_GT	Mono	-1.8E-02	5.6E-01	dominant
HPSE	chr4_83302158_G_GT	Mono	-8.4E-02	1.4E-01	recessive
HPSE	chr4_83302158_G_GT	Mono	-4.2E-02	1.5E-01	add_twodf
HPSE	chr4_83302158_G_GT	Mono	4.6E-02	3.0E-01	dd_twodf
IFIH1	chr2_162277515_AC_A	B	-3.8E-02	1.1E-02	additive
IFIH1	chr2_162277515_AC_A	B	-4.4E-02	4.9E-03	dominant
IFIH1	chr2_162277515_AC_A	B	8.7E-02	3.1E-01	recessive
IFIH1	chr2_162277515_AC_A	B	4.1E-02	3.3E-01	add_twodf
IFIH1	chr2_162277515_AC_A	B	-8.8E-02	4.6E-02	dd_twodf
IFIH1	chr2_162277515_AC_A	CD4T	-2.4E-02	1.4E-02	additive
IFIH1	chr2_162277515_AC_A	CD4T	-2.7E-02	1.0E-02	dominant
IFIH1	chr2_162277515_AC_A	CD4T	1.4E-02	7.9E-01	recessive
IFIH1	chr2_162277515_AC_A	CD4T	5.1E-03	8.5E-01	add_twodf
IFIH1	chr2_162277515_AC_A	CD4T	-3.3E-02	2.4E-01	dd_twodf
IFIH1	chr2_162277515_AC_A	CD8T	-1.0E-02	2.0E-01	additive
IFIH1	chr2_162277515_AC_A	CD8T	-1.6E-02	7.8E-02	dominant
IFIH1	chr2_162277515_AC_A	CD8T	2.8E-02	3.3E-01	recessive
IFIH1	chr2_162277515_AC_A	CD8T	1.2E-02	3.7E-01	add_twodf
IFIH1	chr2_162277515_AC_A	CD8T	-3.3E-02	4.8E-02	dd_twodf
IFIH1	chr2_162277515_AC_A	T_other	-2.3E-02	1.0E-01	additive
IFIH1	chr2_162277515_AC_A	T_other	-2.6E-02	6.8E-02	dominant
IFIH1	chr2_162277515_AC_A	T_other	1.2E-01	2.2E-01	recessive
IFIH1	chr2_162277515_AC_A	T_other	5.9E-02	2.3E-01	add_twodf
IFIH1	chr2_162277515_AC_A	T_other	-8.8E-02	8.5E-02	dd_twodf
IFIH1	chr2_162277515_AC_A	NK	-2.4E-02	2.9E-02	additive
IFIH1	chr2_162277515_AC_A	NK	-2.6E-02	3.4E-02	dominant
IFIH1	chr2_162277515_AC_A	NK	-4.5E-02	3.5E-01	recessive
IFIH1	chr2_162277515_AC_A	NK	-2.4E-02	3.2E-01	add_twodf
IFIH1	chr2_162277515_AC_A	NK	-3.2E-04	9.9E-01	dd_twodf
IFIH1	chr2_162277515_AC_A	Mono	-6.0E-02	1.4E-02	additive
IFIH1	chr2_162277515_AC_A	Mono	-7.5E-02	1.3E-02	dominant
IFIH1	chr2_162277515_AC_A	Mono	-8.4E-02	2.3E-01	recessive
IFIH1	chr2_162277515_AC_A	Mono	-4.8E-02	1.5E-01	add_twodf
IFIH1	chr2_162277515_AC_A	Mono	-2.4E-02	5.8E-01	dd_twodf
IFIH1	chr2_162277515_AC_A	DC	-5.5E-02	1.3E-01	additive
IFIH1	chr2_162277515_AC_A	DC	-7.0E-02	8.9E-02	dominant
IFIH1	chr2_162277515_AC_A	DC	-1.2E-02	9.3E-01	recessive

<i>IFIH1</i>	chr2_162277515_AC_A	DC	-9.8E-03	8.7E-01	add_twodf
<i>IFIH1</i>	chr2_162277515_AC_A	DC	-6.6E-02	3.7E-01	dd_twodf
<i>MNDA</i>	chr1_158845998_C_T	B	-1.5E-02	4.2E-01	additive
<i>MNDA</i>	chr1_158845998_C_T	B	-2.0E-02	3.6E-01	dominant
<i>MNDA</i>	chr1_158845998_C_T	B	2.5E-03	9.7E-01	recessive
<i>MNDA</i>	chr1_158845998_C_T	B	-1.9E-03	9.5E-01	add_twodf
<i>MNDA</i>	chr1_158845998_C_T	B	-1.9E-02	6.1E-01	dd_twodf
<i>MNDA</i>	chr1_158845998_C_T	CD4T	-1.2E-02	2.2E-01	additive
<i>MNDA</i>	chr1_158845998_C_T	CD4T	-1.4E-02	1.9E-01	dominant
<i>MNDA</i>	chr1_158845998_C_T	CD4T	-4.2E-03	9.0E-01	recessive
<i>MNDA</i>	chr1_158845998_C_T	CD4T	-4.4E-03	8.0E-01	add_twodf
<i>MNDA</i>	chr1_158845998_C_T	CD4T	-9.8E-03	6.2E-01	dd_twodf
<i>MNDA</i>	chr1_158845998_C_T	CD8T	-1.8E-02	6.8E-02	additive
<i>MNDA</i>	chr1_158845998_C_T	CD8T	-2.2E-02	4.8E-02	dominant
<i>MNDA</i>	chr1_158845998_C_T	CD8T	-6.0E-03	8.7E-01	recessive
<i>MNDA</i>	chr1_158845998_C_T	CD8T	-6.6E-03	7.0E-01	add_twodf
<i>MNDA</i>	chr1_158845998_C_T	CD8T	-1.7E-02	4.0E-01	dd_twodf
<i>MNDA</i>	chr1_158845998_C_T	T_other	-1.2E-02	3.0E-01	additive
<i>MNDA</i>	chr1_158845998_C_T	T_other	-2.1E-02	1.3E-01	dominant
<i>MNDA</i>	chr1_158845998_C_T	T_other	3.0E-02	4.1E-01	recessive
<i>MNDA</i>	chr1_158845998_C_T	T_other	1.1E-02	5.2E-01	add_twodf
<i>MNDA</i>	chr1_158845998_C_T	T_other	-3.7E-02	8.7E-02	dd_twodf
<i>MNDA</i>	chr1_158845998_C_T	NK	-1.4E-02	1.9E-01	additive
<i>MNDA</i>	chr1_158845998_C_T	NK	-1.7E-02	1.4E-01	dominant
<i>MNDA</i>	chr1_158845998_C_T	NK	8.7E-03	8.4E-01	recessive
<i>MNDA</i>	chr1_158845998_C_T	NK	1.7E-03	9.4E-01	add_twodf
<i>MNDA</i>	chr1_158845998_C_T	NK	-2.0E-02	4.2E-01	dd_twodf
<i>MNDA</i>	chr1_158845998_C_T	Mono	-4.8E-01	2.6E-05	additive
<i>MNDA</i>	chr1_158845998_C_T	Mono	-5.2E-01	1.1E-04	dominant
<i>MNDA</i>	chr1_158845998_C_T	Mono	-8.8E-01	3.2E-02	recessive
<i>MNDA</i>	chr1_158845998_C_T	Mono	-5.2E-01	3.2E-03	add_twodf
<i>MNDA</i>	chr1_158845998_C_T	Mono	4.7E-02	8.1E-01	dd_twodf
<i>MNDA</i>	chr1_158845998_C_T	DC	-3.4E-01	1.0E-04	additive
<i>MNDA</i>	chr1_158845998_C_T	DC	-3.5E-01	6.2E-04	dominant
<i>MNDA</i>	chr1_158845998_C_T	DC	-7.6E-01	9.7E-03	recessive
<i>MNDA</i>	chr1_158845998_C_T	DC	-4.2E-01	2.0E-03	add_twodf
<i>MNDA</i>	chr1_158845998_C_T	DC	1.3E-01	4.1E-01	dd_twodf
<i>MYO1F</i>	chr19_8522497_G_A	B	6.1E-01	1.2E-02	additive
<i>MYO1F</i>	chr19_8522497_G_A	B	6.1E-01	1.2E-02	dominant
<i>MYO1F</i>	chr19_8522497_G_A	CD4T	3.6E-01	5.7E-03	additive
<i>MYO1F</i>	chr19_8522497_G_A	CD4T	3.6E-01	5.7E-03	dominant
<i>MYO1F</i>	chr19_8522497_G_A	CD8T	4.5E-01	8.9E-04	additive

MYO1F	chr19_8522497_G_A	CD8T	4.5E-01	8.9E-04	dominant
MYO1F	chr19_8522497_G_A	NK	1.1E+00	1.1E-04	additive
MYO1F	chr19_8522497_G_A	NK	1.1E+00	1.1E-04	dominant
MYO1F	chr19_8522497_G_A	Mono	2.0E-01	3.4E-01	additive
MYO1F	chr19_8522497_G_A	Mono	2.0E-01	3.4E-01	dominant
MYO1G	chr7_44964942_TAC_T	B	-1.2E-03	9.8E-01	additive
MYO1G	chr7_44964942_TAC_T	B	-1.2E-03	9.8E-01	dominant
MYO1G	chr7_44964942_TAC_T	CD4T	-3.5E-03	9.3E-01	additive
MYO1G	chr7_44964942_TAC_T	CD4T	-3.5E-03	9.3E-01	dominant
MYO1G	chr7_44964942_TAC_T	CD8T	-3.0E-02	6.5E-01	additive
MYO1G	chr7_44964942_TAC_T	CD8T	-3.0E-02	6.5E-01	dominant
MYO1G	chr7_44964942_TAC_T	T_other	1.6E-02	8.9E-01	additive
MYO1G	chr7_44964942_TAC_T	T_other	1.6E-02	8.9E-01	dominant
MYO1G	chr7_44964942_TAC_T	NK	-4.4E-02	5.8E-01	additive
MYO1G	chr7_44964942_TAC_T	NK	-4.4E-02	5.8E-01	dominant
MYO1G	chr7_44964942_TAC_T	Mono	4.2E-02	6.6E-01	additive
MYO1G	chr7_44964942_TAC_T	Mono	4.2E-02	6.6E-01	dominant
NOD2	chr16_50716612_C_T	B	2.7E-03	4.2E-01	additive
NOD2	chr16_50716612_C_T	B	2.7E-03	4.2E-01	dominant
NOD2	chr16_50716612_C_T	CD4T	-1.4E-03	6.2E-01	additive
NOD2	chr16_50716612_C_T	CD4T	-1.4E-03	6.2E-01	dominant
NOD2	chr16_50716612_C_T	CD8T	-1.2E-03	5.9E-01	additive
NOD2	chr16_50716612_C_T	CD8T	-1.2E-03	5.9E-01	dominant
NOD2	chr16_50716612_C_T	NK	8.9E-04	8.0E-01	additive
NOD2	chr16_50716612_C_T	NK	8.9E-04	8.0E-01	dominant
NOD2	chr16_50716612_C_T	Mono	-3.1E-02	3.5E-01	additive
NOD2	chr16_50716612_C_T	Mono	-3.1E-02	3.5E-01	dominant
SIGLEC14	chr19_51645475_A_G	B	-2.8E-03	4.0E-01	additive
SIGLEC14	chr19_51645475_A_G	B	-5.7E-03	4.0E-01	dominant
SIGLEC14	chr19_51645475_A_G	B	-5.7E-03	4.0E-01	recessive
SIGLEC14	chr19_51645475_A_G	CD4T	-7.2E-04	1.8E-01	additive
SIGLEC14	chr19_51645475_A_G	CD4T	-1.4E-03	1.8E-01	dominant
SIGLEC14	chr19_51645475_A_G	CD4T	-1.4E-03	1.8E-01	recessive
SIGLEC14	chr19_51645475_A_G	T_other	-1.8E-03	3.5E-01	additive
SIGLEC14	chr19_51645475_A_G	T_other	-3.7E-03	3.5E-01	dominant
SIGLEC14	chr19_51645475_A_G	T_other	-3.7E-03	3.5E-01	recessive
SIGLEC14	chr19_51645475_A_G	NK	-9.8E-04	3.7E-01	additive
SIGLEC14	chr19_51645475_A_G	NK	-2.0E-03	3.7E-01	dominant
SIGLEC14	chr19_51645475_A_G	NK	-2.0E-03	3.7E-01	recessive
SIGLEC14	chr19_51645475_A_G	Mono	-2.3E-02	9.9E-02	additive
SIGLEC14	chr19_51645475_A_G	Mono	-4.5E-02	9.9E-02	dominant
SIGLEC14	chr19_51645475_A_G	Mono	-4.5E-02	9.9E-02	recessive

SIGLEC14	chr19_51645475_A_G	DC	-1.2E-02	3.6E-01	additive
SIGLEC14	chr19_51645475_A_G	DC	-2.3E-02	3.6E-01	dominant
SIGLEC14	chr19_51645475_A_G	DC	-2.3E-02	3.6E-01	recessive
SIGLEC14	chr19_51643777_ACT_A	B	1.7E-03	8.1E-01	additive
SIGLEC14	chr19_51643777_ACT_A	B	3.5E-03	8.1E-01	dominant
SIGLEC14	chr19_51643777_ACT_A	B	3.5E-03	8.1E-01	recessive
SIGLEC14	chr19_51643777_ACT_A	CD4T	2.5E-03	1.2E-02	additive
SIGLEC14	chr19_51643777_ACT_A	CD4T	5.1E-03	1.2E-02	dominant
SIGLEC14	chr19_51643777_ACT_A	CD4T	5.1E-03	1.2E-02	recessive
SIGLEC14	chr19_51643777_ACT_A	T_other	1.0E-03	7.1E-01	additive
SIGLEC14	chr19_51643777_ACT_A	T_other	2.0E-03	7.1E-01	dominant
SIGLEC14	chr19_51643777_ACT_A	T_other	2.0E-03	7.1E-01	recessive
SIGLEC14	chr19_51643777_ACT_A	NK	4.8E-03	5.4E-02	additive
SIGLEC14	chr19_51643777_ACT_A	NK	9.7E-03	5.4E-02	dominant
SIGLEC14	chr19_51643777_ACT_A	NK	9.7E-03	5.4E-02	recessive
SIGLEC14	chr19_51643777_ACT_A	Mono	-2.1E-02	4.4E-01	additive
SIGLEC14	chr19_51643777_ACT_A	Mono	-4.2E-02	4.4E-01	dominant
SIGLEC14	chr19_51643777_ACT_A	Mono	-4.2E-02	4.4E-01	recessive
SIGLEC14	chr19_51643777_ACT_A	DC	1.8E-03	9.5E-01	additive
SIGLEC14	chr19_51643777_ACT_A	DC	3.6E-03	9.5E-01	dominant
SIGLEC14	chr19_51643777_ACT_A	DC	3.6E-03	9.5E-01	recessive
SIGLEC5	chr19_51628003_A_AG	B	4.2E-03	6.3E-02	additive
SIGLEC5	chr19_51628003_A_AG	B	3.2E-03	2.4E-01	dominant
SIGLEC5	chr19_51628003_A_AG	B	9.2E-03	5.2E-03	add_twodf
SIGLEC5	chr19_51628003_A_AG	B	-8.4E-03	4.1E-02	dd_twodf
SIGLEC5	chr19_51628003_A_AG	CD4T	-4.9E-04	1.5E-01	additive
SIGLEC5	chr19_51628003_A_AG	CD4T	-6.4E-04	1.5E-01	dominant
SIGLEC5	chr19_51628003_A_AG	CD4T	-7.0E-04	4.5E-01	recessive
SIGLEC5	chr19_51628003_A_AG	CD4T	-4.0E-04	3.8E-01	add_twodf
SIGLEC5	chr19_51628003_A_AG	CD4T	-2.0E-04	7.5E-01	dd_twodf
SIGLEC5	chr19_51628003_A_AG	CD8T	-1.0E-04	8.1E-01	additive
SIGLEC5	chr19_51628003_A_AG	CD8T	6.3E-06	9.9E-01	dominant
SIGLEC5	chr19_51628003_A_AG	CD8T	-6.8E-04	5.3E-01	recessive
SIGLEC5	chr19_51628003_A_AG	CD8T	-3.3E-04	5.5E-01	add_twodf
SIGLEC5	chr19_51628003_A_AG	CD8T	5.6E-04	5.0E-01	dd_twodf
SIGLEC5	chr19_51628003_A_AG	T_other	-3.8E-05	9.8E-01	additive
SIGLEC5	chr19_51628003_A_AG	T_other	5.8E-04	8.2E-01	dominant
SIGLEC5	chr19_51628003_A_AG	T_other	-2.5E-03	6.2E-01	recessive
SIGLEC5	chr19_51628003_A_AG	T_other	-1.2E-03	6.4E-01	add_twodf
SIGLEC5	chr19_51628003_A_AG	T_other	2.7E-03	4.7E-01	dd_twodf
SIGLEC5	chr19_51628003_A_AG	Mono	-1.6E-02	1.5E-01	additive
SIGLEC5	chr19_51628003_A_AG	Mono	-1.5E-02	3.2E-01	dominant

<i>SIGLEC5</i>	chr19_51628003_A_AG	Mono	-5.0E-02	8.4E-02	recessive
<i>SIGLEC5</i>	chr19_51628003_A_AG	Mono	-2.5E-02	8.7E-02	add_twodf
<i>SIGLEC5</i>	chr19_51628003_A_AG	Mono	2.1E-02	3.2E-01	dd_twodf
<i>SIGLEC5</i>	chr19_51628003_A_AG	DC	-1.7E-02	2.7E-01	additive
<i>SIGLEC5</i>	chr19_51628003_A_AG	DC	-1.9E-02	3.5E-01	dominant
<i>SIGLEC5</i>	chr19_51628003_A_AG	DC	-4.1E-02	3.1E-01	recessive
<i>SIGLEC5</i>	chr19_51628003_A_AG	DC	-2.1E-02	3.0E-01	add_twodf
<i>SIGLEC5</i>	chr19_51628003_A_AG	DC	8.9E-03	7.6E-01	dd_twodf
<i>SLC9A3R1</i>	chr17_74749171_G_T	B	1.8E-02	4.4E-01	additive
<i>SLC9A3R1</i>	chr17_74749171_G_T	B	4.2E-02	1.6E-01	dominant
<i>SLC9A3R1</i>	chr17_74749171_G_T	B	-5.5E-02	3.9E-01	recessive
<i>SLC9A3R1</i>	chr17_74749171_G_T	B	-2.3E-02	4.5E-01	add_twodf
<i>SLC9A3R1</i>	chr17_74749171_G_T	B	8.5E-02	4.3E-02	dd_twodf
<i>SLC9A3R1</i>	chr17_74749171_G_T	CD4T	-6.1E-03	7.9E-01	additive
<i>SLC9A3R1</i>	chr17_74749171_G_T	CD4T	1.4E-02	6.5E-01	dominant
<i>SLC9A3R1</i>	chr17_74749171_G_T	CD4T	-1.0E-01	9.1E-02	recessive
<i>SLC9A3R1</i>	chr17_74749171_G_T	CD4T	-4.8E-02	1.1E-01	add_twodf
<i>SLC9A3R1</i>	chr17_74749171_G_T	CD4T	8.8E-02	3.6E-02	dd_twodf
<i>SLC9A3R1</i>	chr17_74749171_G_T	CD8T	-4.9E-02	1.9E-01	additive
<i>SLC9A3R1</i>	chr17_74749171_G_T	CD8T	-3.8E-02	4.3E-01	dominant
<i>SLC9A3R1</i>	chr17_74749171_G_T	CD8T	-1.8E-01	6.6E-02	recessive
<i>SLC9A3R1</i>	chr17_74749171_G_T	CD8T	-9.0E-02	7.1E-02	add_twodf
<i>SLC9A3R1</i>	chr17_74749171_G_T	CD8T	8.7E-02	2.0E-01	dd_twodf
<i>SLC9A3R1</i>	chr17_74749171_G_T	T_other	-1.8E-02	7.1E-01	additive
<i>SLC9A3R1</i>	chr17_74749171_G_T	T_other	1.4E-02	8.2E-01	dominant
<i>SLC9A3R1</i>	chr17_74749171_G_T	T_other	-2.0E-01	1.3E-01	recessive
<i>SLC9A3R1</i>	chr17_74749171_G_T	T_other	-9.3E-02	1.5E-01	add_twodf
<i>SLC9A3R1</i>	chr17_74749171_G_T	T_other	1.5E-01	8.6E-02	dd_twodf
<i>SLC9A3R1</i>	chr17_74749171_G_T	NK	3.4E-02	4.1E-01	additive
<i>SLC9A3R1</i>	chr17_74749171_G_T	NK	6.9E-02	1.7E-01	dominant
<i>SLC9A3R1</i>	chr17_74749171_G_T	NK	-7.3E-02	5.1E-01	recessive
<i>SLC9A3R1</i>	chr17_74749171_G_T	NK	-2.9E-02	5.9E-01	add_twodf
<i>SLC9A3R1</i>	chr17_74749171_G_T	NK	1.3E-01	8.8E-02	dd_twodf
<i>SLC9A3R1</i>	chr17_74749171_G_T	Mono	1.0E-02	6.6E-01	additive
<i>SLC9A3R1</i>	chr17_74749171_G_T	Mono	2.9E-02	3.5E-01	dominant
<i>SLC9A3R1</i>	chr17_74749171_G_T	Mono	-4.3E-02	4.7E-01	recessive
<i>SLC9A3R1</i>	chr17_74749171_G_T	Mono	-1.8E-02	5.4E-01	add_twodf
<i>SLC9A3R1</i>	chr17_74749171_G_T	Mono	6.6E-02	1.3E-01	dd_twodf
<i>SLC9A3R1</i>	chr17_74749171_G_T	DC	3.9E-03	8.5E-01	additive
<i>SLC9A3R1</i>	chr17_74749171_G_T	DC	3.0E-02	2.9E-01	dominant
<i>SLC9A3R1</i>	chr17_74749171_G_T	DC	-7.6E-02	1.3E-01	recessive
<i>TMEM71</i>	chr8_132722068_C_CAGCA	CD4T	4.7E-03	7.5E-01	additive

TMEM71	chr8_132722068_C_CAGCA	CD4T	4.7E-03	7.5E-01	dominant
TMEM71	chr8_132722068_C_CAGCA	CD8T	-1.5E-03	9.1E-01	additive
TMEM71	chr8_132722068_C_CAGCA	CD8T	-1.5E-03	9.1E-01	dominant
TMEM71	chr8_132722068_C_CAGCA	T_other	2.6E-02	2.3E-01	additive
TMEM71	chr8_132722068_C_CAGCA	T_other	2.6E-02	2.3E-01	dominant
TMEM71	chr8_132722068_C_CAGCA	NK	5.9E-03	8.0E-01	additive
TMEM71	chr8_132722068_C_CAGCA	NK	5.9E-03	8.0E-01	dominant
TMEM71	chr8_132722068_C_CAGCA	Mono	1.5E-03	9.6E-01	additive
TMEM71	chr8_132722068_C_CAGCA	Mono	1.5E-03	9.6E-01	dominant
TMEM71	chr8_132722068_C_CAGCA	DC	-1.0E-02	6.9E-01	additive
TMEM71	chr8_132722068_C_CAGCA	DC	-1.0E-02	6.9E-01	dominant
TMEM71	chr8_132758873_G_A	CD4T	8.6E-03	3.2E-01	additive
TMEM71	chr8_132758873_G_A	CD4T	1.4E-02	3.4E-01	dominant
TMEM71	chr8_132758873_G_A	CD4T	1.7E-02	3.6E-01	recessive
TMEM71	chr8_132758873_G_A	CD4T	8.5E-03	3.6E-01	add_twodf
TMEM71	chr8_132758873_G_A	CD4T	6.2E-04	9.8E-01	dd_twodf
TMEM71	chr8_132758873_G_A	CD8T	1.7E-02	4.4E-02	additive
TMEM71	chr8_132758873_G_A	CD8T	3.1E-02	3.2E-02	dominant
TMEM71	chr8_132758873_G_A	CD8T	2.8E-02	1.2E-01	recessive
TMEM71	chr8_132758873_G_A	CD8T	1.5E-02	9.7E-02	add_twodf
TMEM71	chr8_132758873_G_A	CD8T	1.7E-02	4.6E-01	dd_twodf
TMEM71	chr8_132758873_G_A	T_other	1.1E-02	4.6E-01	additive
TMEM71	chr8_132758873_G_A	T_other	1.9E-02	4.8E-01	dominant
TMEM71	chr8_132758873_G_A	T_other	2.2E-02	4.8E-01	recessive
TMEM71	chr8_132758873_G_A	T_other	1.1E-02	4.8E-01	add_twodf
TMEM71	chr8_132758873_G_A	T_other	-2.5E-04	1.0E+00	dd_twodf
TMEM71	chr8_132758873_G_A	NK	3.7E-02	5.8E-03	additive
TMEM71	chr8_132758873_G_A	NK	6.3E-02	1.0E-02	dominant
TMEM71	chr8_132758873_G_A	NK	7.7E-02	6.1E-03	recessive
TMEM71	chr8_132758873_G_A	NK	3.9E-02	6.3E-03	add_twodf
TMEM71	chr8_132758873_G_A	NK	-1.9E-02	6.8E-01	dd_twodf
TMEM71	chr8_132758873_G_A	Mono	2.3E-03	9.0E-01	additive
TMEM71	chr8_132758873_G_A	Mono	-7.9E-05	1.0E+00	dominant
TMEM71	chr8_132758873_G_A	Mono	1.0E-02	7.9E-01	recessive
TMEM71	chr8_132758873_G_A	Mono	4.6E-03	8.1E-01	add_twodf
TMEM71	chr8_132758873_G_A	Mono	-2.3E-02	6.8E-01	dd_twodf
TMEM71	chr8_132758873_G_A	DC	1.1E-02	4.6E-01	additive
TMEM71	chr8_132758873_G_A	DC	1.9E-02	4.5E-01	dominant
TMEM71	chr8_132758873_G_A	DC	1.9E-02	5.4E-01	recessive
TMEM71	chr8_132758873_G_A	DC	9.7E-03	5.3E-01	add_twodf
TMEM71	chr8_132758873_G_A	DC	7.8E-03	8.5E-01	dd_twodf
TRAT1	chr3_108849184_GCTATGAA_G	B	-7.1E-04	8.3E-01	additive

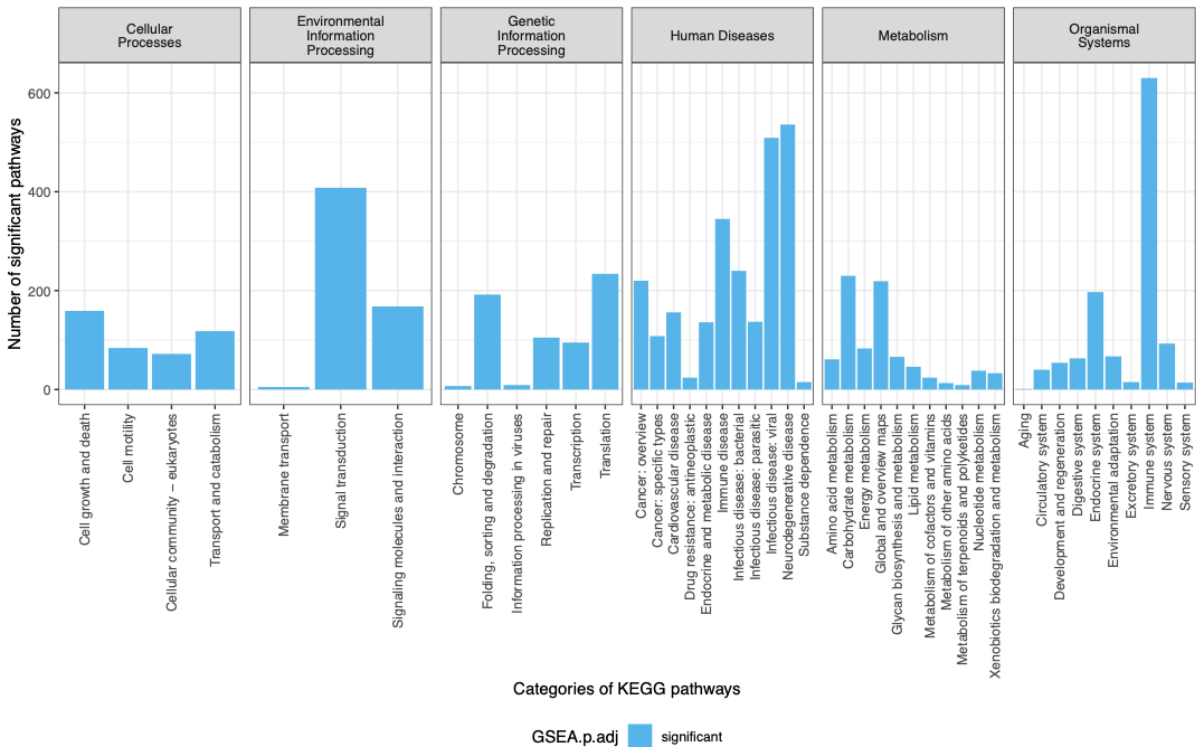
TRAT1	chr3_108849184_GCTATGAA_G	B	-8.5E-04	8.4E-01	dominant
TRAT1	chr3_108849184_GCTATGAA_G	B	-1.2E-03	9.0E-01	recessive
TRAT1	chr3_108849184_GCTATGAA_G	B	-6.8E-04	8.8E-01	add_twodf
TRAT1	chr3_108849184_GCTATGAA_G	B	-1.1E-04	9.9E-01	dd_twodf
TRAT1	chr3_108849184_GCTATGAA_G	CD4T	-7.0E-03	8.4E-01	additive
TRAT1	chr3_108849184_GCTATGAA_G	CD4T	-7.3E-03	8.7E-01	dominant
TRAT1	chr3_108849184_GCTATGAA_G	CD4T	-1.7E-02	8.6E-01	recessive
TRAT1	chr3_108849184_GCTATGAA_G	CD4T	-9.0E-03	8.6E-01	add_twodf
TRAT1	chr3_108849184_GCTATGAA_G	CD4T	3.8E-03	9.5E-01	dd_twodf
TRAT1	chr3_108849184_GCTATGAA_G	CD8T	2.7E-03	9.1E-01	additive
TRAT1	chr3_108849184_GCTATGAA_G	CD8T	-3.0E-03	9.2E-01	dominant
TRAT1	chr3_108849184_GCTATGAA_G	CD8T	3.7E-02	5.8E-01	recessive
TRAT1	chr3_108849184_GCTATGAA_G	CD8T	1.8E-02	6.1E-01	add_twodf
TRAT1	chr3_108849184_GCTATGAA_G	CD8T	-2.8E-02	5.4E-01	dd_twodf
TRAT1	chr3_108849184_GCTATGAA_G	T_other	-4.7E-02	9.8E-02	additive
TRAT1	chr3_108849184_GCTATGAA_G	T_other	-5.0E-02	1.5E-01	dominant
TRAT1	chr3_108849184_GCTATGAA_G	T_other	-1.0E-01	2.0E-01	recessive
TRAT1	chr3_108849184_GCTATGAA_G	T_other	-5.5E-02	1.7E-01	add_twodf
TRAT1	chr3_108849184_GCTATGAA_G	T_other	1.6E-02	7.6E-01	dd_twodf
TRAT1	chr3_108849184_GCTATGAA_G	NK	-9.9E-03	2.0E-02	additive
TRAT1	chr3_108849184_GCTATGAA_G	NK	-1.4E-02	5.9E-03	dominant
TRAT1	chr3_108849184_GCTATGAA_G	NK	-4.2E-03	7.3E-01	recessive
TRAT1	chr3_108849184_GCTATGAA_G	NK	-3.3E-03	5.5E-01	add_twodf
TRAT1	chr3_108849184_GCTATGAA_G	NK	-1.3E-02	9.1E-02	dd_twodf
TRAT1	chr3_108849184_GCTATGAA_G	Mono	-3.9E-04	9.2E-01	additive
TRAT1	chr3_108849184_GCTATGAA_G	Mono	-8.9E-04	8.5E-01	dominant
TRAT1	chr3_108849184_GCTATGAA_G	Mono	2.0E-03	8.6E-01	recessive
TRAT1	chr3_108849184_GCTATGAA_G	Mono	9.0E-04	8.8E-01	add_twodf
TRAT1	chr3_108849184_GCTATGAA_G	Mono	-2.2E-03	7.7E-01	dd_twodf
TRAT1	chr3_108849184_GCTATGAA_G	DC	-7.7E-03	4.6E-01	additive
TRAT1	chr3_108849184_GCTATGAA_G	DC	-1.4E-02	2.4E-01	dominant
TRAT1	chr3_108849184_GCTATGAA_G	DC	3.3E-02	3.5E-01	recessive
TRAT1	chr3_108849184_GCTATGAA_G	DC	1.5E-02	3.9E-01	add_twodf
TRAT1	chr3_108849184_GCTATGAA_G	DC	-3.3E-02	1.1E-01	dd_twodf
TRIM58	chr1_247857547_G_T	B	-5.7E-03	3.1E-01	additive
TRIM58	chr1_247857547_G_T	B	-5.7E-03	3.1E-01	dominant
TRIM58	chr1_247857547_G_T	CD4T	-3.4E-04	7.1E-01	additive
TRIM58	chr1_247857547_G_T	CD4T	-3.4E-04	7.1E-01	dominant
TRIM58	chr1_247857547_G_T	CD8T	-7.9E-04	5.3E-01	additive
TRIM58	chr1_247857547_G_T	CD8T	-7.9E-04	5.3E-01	dominant
TRIM58	chr1_247857547_G_T	T_other	1.7E-03	6.0E-01	additive
TRIM58	chr1_247857547_G_T	T_other	1.7E-03	6.0E-01	dominant

TRIM58	chr1_247857547_G_T	NK	-1.5E-03	4.0E-01	additive
TRIM58	chr1_247857547_G_T	NK	-1.5E-03	4.0E-01	dominant
TRIM58	chr1_247857547_G_T	Mono	-6.5E-04	8.7E-01	additive
TRIM58	chr1_247857547_G_T	Mono	-6.5E-04	8.7E-01	dominant
TRIM58	chr1_247857547_G_T	DC	3.1E-03	6.4E-01	additive
TRIM58	chr1_247857547_G_T	DC	3.1E-03	6.4E-01	dominant
ZBP1	chr20_57615536_CA_C	B	-1.7E-02	7.1E-01	additive
ZBP1	chr20_57615536_CA_C	B	-1.7E-02	7.1E-01	dominant
ZBP1	chr20_57615536_CA_C	CD4T	-1.2E-02	4.3E-01	additive
ZBP1	chr20_57615536_CA_C	CD4T	-1.2E-02	4.3E-01	dominant
ZBP1	chr20_57615536_CA_C	CD8T	5.0E-03	8.5E-01	additive
ZBP1	chr20_57615536_CA_C	CD8T	5.0E-03	8.5E-01	dominant
ZBP1	chr20_57615536_CA_C	T_other	1.2E-02	7.3E-01	additive
ZBP1	chr20_57615536_CA_C	T_other	1.2E-02	7.3E-01	dominant
ZBP1	chr20_57615536_CA_C	NK	4.2E-03	9.2E-01	additive
ZBP1	chr20_57615536_CA_C	NK	4.2E-03	9.2E-01	dominant
ZBP1	chr20_57615536_CA_C	Mono	-1.5E-02	3.5E-01	additive
ZBP1	chr20_57615536_CA_C	Mono	-1.5E-02	3.5E-01	dominant
ZBP1	chr20_57615536_CA_C	DC	-1.5E-02	5.1E-01	additive
ZBP1	chr20_57615536_CA_C	DC	-1.5E-02	5.1E-01	dominant
ZC3H8	chr2_112230930_CTTCTT_C	B	3.2E-03	8.3E-01	additive
ZC3H8	chr2_112230930_CTTCTT_C	B	-3.8E-03	8.7E-01	dominant
ZC3H8	chr2_112230930_CTTCTT_C	B	2.2E-02	5.0E-01	recessive
ZC3H8	chr2_112230930_CTTCTT_C	B	1.0E-02	5.2E-01	add_twodf
ZC3H8	chr2_112230930_CTTCTT_C	B	-4.2E-02	2.6E-01	dd_twodf
ZC3H8	chr2_112230930_CTTCTT_C	CD4T	9.6E-06	1.0E+00	additive
ZC3H8	chr2_112230930_CTTCTT_C	CD4T	-3.7E-03	8.4E-01	dominant
ZC3H8	chr2_112230930_CTTCTT_C	CD4T	6.6E-03	7.8E-01	recessive
ZC3H8	chr2_112230930_CTTCTT_C	CD4T	3.0E-03	8.1E-01	add_twodf
ZC3H8	chr2_112230930_CTTCTT_C	CD4T	-1.8E-02	5.3E-01	dd_twodf
ZC3H8	chr2_112230930_CTTCTT_C	CD8T	2.7E-05	1.0E+00	additive
ZC3H8	chr2_112230930_CTTCTT_C	CD8T	-1.7E-03	9.3E-01	dominant
ZC3H8	chr2_112230930_CTTCTT_C	CD8T	3.1E-03	9.0E-01	recessive
ZC3H8	chr2_112230930_CTTCTT_C	CD8T	1.4E-03	9.2E-01	add_twodf
ZC3H8	chr2_112230930_CTTCTT_C	CD8T	-8.6E-03	7.8E-01	dd_twodf
ZC3H8	chr2_112230930_CTTCTT_C	T_other	3.2E-03	6.8E-01	additive
ZC3H8	chr2_112230930_CTTCTT_C	T_other	4.5E-03	7.4E-01	dominant
ZC3H8	chr2_112230930_CTTCTT_C	T_other	7.6E-03	6.4E-01	recessive
ZC3H8	chr2_112230930_CTTCTT_C	T_other	3.7E-03	6.6E-01	add_twodf
ZC3H8	chr2_112230930_CTTCTT_C	T_other	-7.4E-03	7.7E-01	dd_twodf
ZC3H8	chr2_112230930_CTTCTT_C	NK	1.1E-02	3.8E-01	additive
ZC3H8	chr2_112230930_CTTCTT_C	NK	1.0E-02	6.2E-01	dominant

ZC3H8	chr2_112230930_CTTCTT_C	NK	3.4E-02	2.2E-01	recessive
ZC3H8	chr2_112230930_CTTCTT_C	NK	1.7E-02	2.4E-01	add_twodf
ZC3H8	chr2_112230930_CTTCTT_C	NK	-3.0E-02	3.3E-01	dd_twodf
ZC3H8	chr2_112230930_CTTCTT_C	Mono	-2.5E-04	9.7E-01	additive
ZC3H8	chr2_112230930_CTTCTT_C	Mono	-3.4E-03	7.8E-01	dominant
ZC3H8	chr2_112230930_CTTCTT_C	Mono	4.6E-03	7.7E-01	recessive
ZC3H8	chr2_112230930_CTTCTT_C	Mono	2.0E-03	8.0E-01	add_twodf
ZC3H8	chr2_112230930_CTTCTT_C	Mono	-1.5E-02	4.5E-01	dd_twodf
ZC3H8	chr2_112230930_CTTCTT_C	DC	4.8E-03	7.7E-01	additive
ZC3H8	chr2_112230930_CTTCTT_C	DC	-8.3E-03	7.6E-01	dominant
ZC3H8	chr2_112230930_CTTCTT_C	DC	3.6E-02	2.9E-01	recessive

Appendix C - Supplementary data for Chapter 4

Supplementary figures



Supplementary-figure-4.1

Number of significantly dysregulated pathways in pLoF carriers, sorted over KEGG categories.

Appendix D - General supplementary data

The Genes & Health Research Team

[Accessed September 2024] Genes & Health Research Team (in alphabetical order by surname): Shaheen Akhtar, Mohammad Anwar, Omar Asgar, Samina Ashraf, Saeed Bidi, Gerome Breen, James Broster, Raymond Chung, David Collier, Charles J Curtis, Shabana Chaudhary, Grainne Colligan, Panos Deloukas, Ceri Durham, Faiza Durrani, Fabiola Eto, Sarah Finer, Joseph Gafton, Ana Angel, Chris Griffiths, Joanne Henry, Teng Heng, Sam Hodgson, Qin Qin Huang, Matt Hurles, Karen A Hunt, Shapna Hussain, Kamrul Islam, Vivek Iyer, Benjamin M Jacobs, Georgios Kalantzis, Ahsan Khan, Claudia Langenberg, Cath Lavery, Sang Hyuck Lee, Daniel MacArthur, Eamonn Maher, Sidra Malik, Daniel Malawsky, Hilary Martin, Dan Mason, Rohini Mathur, Mohammed Bodrul Mazid, John McDermott, Caroline Morton, Bill Newman, Vladimir Ovchinnikov, Elizabeth Owor, Iaroslav Popov, Asma Qureshi, Mehru Raza, Jessry Russell, Nishat Safa, Miriam Samuel, Moneeza Siddiqui, Michael Simpson, John Solly, Marie Spreckley, Daniel Stow, Michael Taylor, Richard C Trembath, Karen Tricker, David A van Heel, Klaudia Walter, Caroline Winckley, Suzanne Wood, John Wright, Sabina Yasmin, Ishevanhu Zengeya, Julia Zöllner.

Bibliography

- 10xGenomics. (n.d.). *Datasets*. 10x Genomics. Retrieved March 29, 2022, from <https://www.10xgenomics.com/resources/datasets?menu%5Bproducts.name%5D=Single%20Cell%20Gene%20Expression&query=&page=1&configure%5Bfacets%5D%5B0%5D=chemistryVersionAndThroughput&configure%5Bfacets%5D%5B1%5D=pipeline.version&configure%5BhitsPerPage%5D=500>
- 1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., & McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), 1061–1073. <https://doi.org/10.1038/nature09534>
- 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. <https://doi.org/10.1038/nature15393>
- Adli, M. (2018). The CRISPR tool kit for genome editing and beyond. *Nature Communications*, 9(1), 1911. <https://doi.org/10.1038/s41467-018-04252-2>
- Agelidis, A., Turturice, B. A., Suryawanshi, R. K., Yadavalli, T., Jaishankar, D., Ames, J., Hopkins, J., Koujah, L., Patil, C. D., Hadigal, S. R., Kyzar, E. J., Campeau, A., Wozniak, J. M., Gonzalez, D. J., Vlodavsky, I., Li, J.-P., Perkins, D. L., Finn, P. W., & Shukla, D. (2021). Disruption of innate defense responses by endoglycosidase HPSE promotes cell survival. *JCI Insight*, 6(7). <https://doi.org/10.1172/jci.insight.144255>
- Alberts, B. (2002). *Molecular biology of the cell*. American Society for Cell Biology. <https://play.google.com/store/books/details?id=n-pFAQAIAAJ>
- Aleksander, S. A., Balhoff, J., Carbon, S., Michael, J., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., Hill, D. P., Lee, R., Mi, H., Moxon, S., Mungall, C. J., Muruganugan, A., Mushayahama, T., Sternberg, P. W., Thomas, P. D., Auken, V., ... Westerfield, M. (2023). The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1), iyad031. <https://doi.org/10.1093/genetics/iyad031>
- Al-Mufti, R., Hambley, H., Farzaneh, F., & Nicolaidis, K. H. (2000). Fetal and embryonic hemoglobins in erythroblasts of chromosomally normal and abnormal fetuses at 10-40 weeks of gestation. *Haematologica*, 85(7), 690–693. <https://www.ncbi.nlm.nih.gov/pubmed/10897119>
- Alsalem, A. B., Halees, A. S., Anazi, S., Alshamekh, S., & Alkuraya, F. S. (2013). Autozygome sequencing expands the horizon of human knockout research and provides novel insights into human phenotypic variation. *PLoS Genetics*, 9(12), e1004030. <https://doi.org/10.1371/journal.pgen.1004030>

- Antonarakis, S. E., & Beckmann, J. S. (2006). Mendelian disorders deserve more attention. *Nature Reviews. Genetics*, 7(4), 277–282. <https://doi.org/10.1038/nrg1826>
- Aparici-Herraiz, I., Sánchez-Sánchez, G., Batlle, C., Rehues, P., López-Serrat, M., Valverde-Estrella, L., Lloberas, J., & Celada, A. (2023). IRF1 Is Required for MDA5 (IFIH1) Induction by IFN- α , LPS, and poly(I:C) in Murine Macrophages. *Journal of Innate Immunity*, 15(1), 297–316. <https://doi.org/10.1159/000527008>
- Arciero, E., Dogra, S. A., Malawsky, D. S., Mezzavilla, M., Tsimentzoglou, T., Huang, Q. Q., Hunt, K. A., Mason, D., Sharif, S. M., van Heel, D. A., Sheridan, E., Wright, J., Small, N., Carmi, S., Iles, M. M., & Martin, H. C. (2021). Fine-scale population structure and demographic history of British Pakistanis. *Nature Communications*, 12(1), 7189. <https://doi.org/10.1038/s41467-021-27394-2>
- Aristizábal, B., & González, Á. (2013). *Innate immune system*. El Rosario University Press. <https://www.ncbi.nlm.nih.gov/books/NBK459455/>
- Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M. A., Lambourne, J. J., Sivapalaratnam, S., Downes, K., Kundu, K., Bomba, L., Berentsen, K., Bradley, J. R., Daugherty, L. C., Delaneau, O., ... Soranzo, N. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*, 167(5), 1415–1429.e19. <https://doi.org/10.1016/j.cell.2016.10.042>
- Auer, P. L., & Lettre, G. (2015). Rare variant association studies: considerations, challenges and opportunities. *Genome Medicine*, 7(1), 16. <https://doi.org/10.1186/s13073-015-0138-2>
- Ayadi, A., Birling, M.-C., Bottomley, J., Bussell, J., Fuchs, H., Fray, M., Gailus-Durner, V., Greenaway, S., Houghton, R., Karp, N., Leblanc, S., Lengger, C., Maier, H., Mallon, A.-M., Marschall, S., Melvin, D., Morgan, H., Pavlovic, G., Ryder, E., ... Herault, Y. (2012). Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, 23(9-10), 600–610. <https://doi.org/10.1007/s00335-012-9418-y>
- Backman, J. D., Li, A. H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M. D., Benner, C., Liu, D., Locke, A. E., Balasubramanian, S., Yadav, A., Banerjee, N., Gillies, C. E., Damask, A., Liu, S., Bai, X., Hawes, A., Maxwell, E., Gurski, L., ... Ferreira, M. A. R. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*, 599(7886), 628–634. <https://doi.org/10.1038/s41586-021-04103-z>
- Bagos, P. G. (2013). Genetic model selection in genome-wide association studies: robust methods and the use of meta-analysis. *Statistical Applications in Genetics and Molecular Biology*, 12(3), 285–308. <https://doi.org/10.1515/sagmb-2012-0016>

- Bailon-Zambrano, R., Sucharov, J., Mumme-Monheit, A., Murry, M., Stenzel, A., Pulvino, A. T., Mitchell, J. M., Colborn, K. L., & Nichols, J. T. (2022). *Variable paralog expression underlies phenotype variation*. <https://doi.org/10.7554/eLife.79247>
- Baird, P. A., Anderson, T. W., Newcombe, H. B., & Lowry, R. B. (1988). Genetic disorders in children and young adults: a population study. *American Journal of Human Genetics*, 42(5), 677–693. <https://www.ncbi.nlm.nih.gov/pubmed/3358420>
- Bais, A. S., & Kostka, D. (2019). scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics*, 36(4), 1150–1158. <https://doi.org/10.1093/bioinformatics/btz698>
- Baptista, M. A. S., Dave, K. D., Frasier, M. A., Sherer, T. B., Greeley, M., Beck, M. J., Varsho, J. S., Parker, G. A., Moore, C., Churchill, M. J., Meshul, C. K., & Fiske, B. K. (2013). Loss of leucine-rich repeat kinase 2 (LRRK2) in rats leads to progressive abnormal phenotypes in peripheral organs. *PloS One*, 8(11), e80705. <https://doi.org/10.1371/journal.pone.0080705>
- Barnett, A. H., Dixon, A. N., Bellary, S., Hanif, M. W., O'Hare, J. P., Raymond, N. T., & Kumar, S. (2006). Type 2 diabetes and cardiovascular risk in the UK south Asian community. *Diabetologia*, 49(10), 2234–2246. <https://doi.org/10.1007/s00125-006-0325-1>
- Barrett, J. C., Dunham, I., & Birney, E. (2015). Using human genetics to make new medicines. *Nature Reviews. Genetics*, 16(10), 561–562. <https://doi.org/10.1038/nrg3998>
- Barton, A. R., Hujoel, M. L. A., Mukamel, R. E., Sherman, M. A., & Loh, P.-R. (2022). A spectrum of recessiveness among Mendelian disease variants in UK Biobank. *American Journal of Human Genetics*, 109(7), 1298–1307. <https://doi.org/10.1016/j.ajhg.2022.05.008>
- Barton, A. R., Sherman, M. A., Mukamel, R. E., & Loh, P.-R. (2021). Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nature Genetics*, 53(8), 1260–1269. <https://doi.org/10.1038/s41588-021-00892-1>
- Bastarache, L., Delozier, S., Pandit, A., He, J., Lewis, A., Annis, A. C., LeFaive, J., Denny, J. C., Carroll, R. J., Altman, R. B., Hughey, J. J., Zawistowski, M., & Peterson, J. F. (2023). The phenotype-genotype reference map: Improving biobank data science through replication. *American Journal of Human Genetics*, 110(9), 1522. <https://doi.org/10.1016/j.ajhg.2023.07.012>
- Basu Ray, S. (2019). PNPLA3-I148M: a problem of plenty in non-alcoholic fatty liver disease. *Adipocyte*, 8(1), 201–208. <https://doi.org/10.1080/21623945.2019.1607423>
- Baulande, S., Lasnier, F., Lucas, M., & Pairault, J. (2001). Adiponutrin, a transmembrane protein corresponding to a novel dietary- and obesity-linked mRNA specifically

- expressed in the adipose lineage. *The Journal of Biological Chemistry*, 276(36), 33336–33344. <https://doi.org/10.1074/jbc.M105193200>
- Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J.-F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., ... Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367(6484). <https://doi.org/10.1126/science.aay5012>
- Blighe, K., Brown, A. L., Carey, V., Hooiveld, G., & Lun, A. (2022). *PCAtools: PCAtools: Everything Principal Components Analysis*. 10.18129/B9.bioc.PCAtools.
- Bodmer, W., & Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics*, 40(6), 695–701. <https://doi.org/10.1038/ng.f.136>
- Boesgaard, T. W., Gjesing, A. P., Grarup, N., Rutanen, J., Jansson, P.-A., Hribal, M. L., Sesti, G., Fritsche, A., Stefan, N., Staiger, H., Häring, H., Smith, U., Laakso, M., Pedersen, O., Hansen, T., & EUGENE2 Consortium. (2009). Variant near ADAMTS9 known to associate with type 2 diabetes is related to insulin resistance in offspring of type 2 diabetes patients--EUGENE2 study. *PloS One*, 4(9), e7236. <https://doi.org/10.1371/journal.pone.0007236>
- Bouland, G. A., Mahfouz, A., & Reinders, M. J. T. (2023). Consequences and opportunities arising due to sparser single-cell RNA-seq datasets. *Genome Biology*, 24(1), 1–10. <https://doi.org/10.1186/s13059-023-02933-w>
- Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., & Sherlock, G. (2004). GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics (Oxford, England)*, 20(18), 3710–3715.
- Briggs, J. A., Burrus, G. R., Stickney, B. D., & Briggs, R. C. (1992). Cloning and expression of the human myeloid cell nuclear differentiation antigen: regulation by interferon alpha. *Journal of Cellular Biochemistry*, 49(1), 82–92. <https://doi.org/10.1002/jcb.240490114>
- Briggs, R. C., Shults, K. E., Flye, L. A., McClintock-Treep, S. A., Jagasia, M. H., Goodman, S. A., Boulos, F. I., Jacobberger, J. W., Stelzer, G. T., & Head, D. R. (2006). Dysregulated human myeloid nuclear differentiation antigen expression in myelodysplastic syndromes: evidence for a role in apoptosis. *Cancer Research*, 66(9), 4645–4651. <https://doi.org/10.1158/0008-5472.CAN-06-0229>
- Bruyns, E., Marie-Cardine, A., Kirchgessner, H., Sagolla, K., Shevchenko, A., Mann, M., Autschbach, F., Bensussan, A., Meuer, S., & Schraven, B. (1998). T cell receptor (TCR) interacting molecule (TRIM), a novel disulfide-linked dimer associated with the TCR-CD3-zeta complex, recruits intracellular signaling proteins to the plasma membrane. *The Journal of Experimental Medicine*, 188(3), 561–575.

<https://doi.org/10.1084/jem.188.3.561>

- Bulger, M., Bender, M. A., van Doorninck, J. H., Wertman, B., Farrell, C. M., Felsenfeld, G., Groudine, M., & Hardison, R. (2000). Comparative structural and functional analysis of the olfactory receptor genes flanking the human and mouse beta-globin gene clusters. *Proceedings of the National Academy of Sciences of the United States of America*, 97(26), 14560–14565. <https://doi.org/10.1073/pnas.97.26.14560>
- Burgess, D. J. (2016). Loss-of-function variants — not always what they seem. *Nature Reviews. Genetics*, 17(5), 251–251. <https://doi.org/10.1038/nrg.2016.33>
- 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. <https://doi.org/10.1038/s41586-018-0579-z>
- Calafell, F., Roubinet, F., Ramírez-Soriano, A., Saitou, N., Bertranpetit, J., & Blancher, A. (2008). Evolutionary dynamics of the human ABO gene. *Human Genetics*, 124(2), 123–135. <https://doi.org/10.1007/s00439-008-0530-8>
- Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics*, 11, 505357. <https://doi.org/10.3389/fgene.2020.00424>
- Cao, A., & Galanello, R. (2010). Beta-thalassemia. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 12(2), 61–76. <https://doi.org/10.1097/GIM.0b013e3181cd68ed>
- Cassa, C. A., Weghorn, D., Balick, D. J., Jordan, D. M., Nusinow, D., Samocha, K. E., O'Donnell-Luria, A., MacArthur, D. G., Daly, M. J., Beier, D. R., & Sunyaev, S. R. (2017). Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nature Genetics*, 49(5), 806–810. <https://doi.org/10.1038/ng.3831>
- Castel, S. E., Cervera, A., Mohammadi, P., Aguet, F., Reverter, F., Wolman, A., Guigo, R., Iossifov, I., Vasileva, A., & Lappalainen, T. (2018). Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nature Genetics*, 50(9), 1327–1334. <https://doi.org/10.1038/s41588-018-0192-y>
- Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4, 7. <https://doi.org/10.1186/s13742-015-0047-8>
- Chang, Y.-F., Imam, J. S., & Wilkinson, M. F. (2007). The nonsense-mediated decay RNA surveillance pathway. *Annual Review of Biochemistry*, 76, 51–74. <https://doi.org/10.1146/annurev.biochem.76.050106.093909>
- Chaplin, D. D. (2010). Overview of the immune response. *The Journal of Allergy and Clinical*

- Immunology*, 125(2 Suppl 2), S3–S23. <https://doi.org/10.1016/j.jaci.2009.12.980>
- Charlesworth, D., & Willis, J. H. (2009). The genetics of inbreeding depression. *Nature Reviews. Genetics*, 10(11), 783–796. <https://doi.org/10.1038/nrg2664>
- Chelikani, P., Fita, I., & Loewen, P. C. (2004). Diversity of structures and properties among catalases. *Cellular and Molecular Life Sciences: CMLS*, 61(2), 192–208. <https://doi.org/10.1007/s00018-003-3206-5>
- Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., & Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14, 128. <https://doi.org/10.1186/1471-2105-14-128>
- Chen, S., Francioli, L. C., Goodrich, J. K., Collins, R. L., Kanai, M., Wang, Q., Alföldi, J., Watts, N. A., Vittal, C., Gauthier, L. D., Poterba, T., Wilson, M. W., Tarasova, Y., Phu, W., Yohannes, M. T., Koenig, Z., Farjoun, Y., Banks, E., Donnelly, S., ... Karczewski, K. J. (2022). A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. In *bioRxiv* (p. 2022.03.20.485034). <https://doi.org/10.1101/2022.03.20.485034>
- Chen, X., Kuja-Halkola, R., Rahman, I., Arpegård, J., Viktorin, A., Karlsson, R., Hägg, S., Svensson, P., Pedersen, N. L., & Magnusson, P. K. E. (2015). Dominant Genetic Variation and Missing Heritability for Human Complex Traits: Insights from Twin versus Genome-wide Common SNP Models. *American Journal of Human Genetics*, 97(5), 708. <https://doi.org/10.1016/j.ajhg.2015.10.004>
- Chen, Y., Lun, A. T. L., & Smyth, G. K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research*, 5(1438), 1438. <https://doi.org/10.12688/f1000research.8987.2>
- Chen, Y., McCarthy, D., Baldoni, P., Ritchie, M., Robinson, M., & Smyth, G. (2024, April 29). *edgeR: differential analysis of sequence read count data User's Guide*. Bioconductor. <http://bioconductor.org/packages/edgeR/>
- Chi, H., Pepper, M., & Thomas, P. G. (2024). Principles and therapeutic applications of adaptive immunity. *Cell*, 187(9), 2052–2078. <https://doi.org/10.1016/j.cell.2024.03.037>
- Chong, J. X., Buckingham, K. J., Jhangiani, S. N., Boehm, C., Sobreira, N., Smith, J. D., Harrell, T. M., McMillin, M. J., Wiszniewski, W., Gambin, T., Coban Akdemir, Z. H., Doheny, K., Scott, A. F., Avramopoulos, D., Chakravarti, A., Hoover-Fong, J., Mathews, D., Witmer, P. D., Ling, H., ... Bamshad, M. J. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *American Journal of Human Genetics*, 97(2), 199–215. <https://doi.org/10.1016/j.ajhg.2015.06.009>
- Chopp, L., Redmond, C., O'Shea, J. J., & Schwartz, D. M. (2023). From thymus to tissues and tumors: A review of T-cell biology. *The Journal of Allergy and Clinical Immunology*,

- 151(1), 81–97. <https://doi.org/10.1016/j.jaci.2022.10.011>
- Ciechanover, A., & Schwartz, A. L. (1998). The ubiquitin-proteasome pathway: The complexity and myriad functions of proteins death. *Proceedings of the National Academy of Sciences of the United States of America*, 95(6), 2727. <https://doi.org/10.1073/pnas.95.6.2727>
- Cirulli, E. T., & Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews. Genetics*, 11(6), 415–425. <https://doi.org/10.1038/nrg2779>
- Cohen, J. C., Boerwinkle, E., Mosley, T. H., Jr, & Hobbs, H. H. (2006). Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *The New England Journal of Medicine*, 354(12), 1264–1272. <https://doi.org/10.1056/NEJMoa054013>
- Connolly, D. J., & Bowie, A. G. (2014). The emerging role of human PYHIN proteins in innate immunity: implications for health and disease. *Biochemical Pharmacology*, 92(3), 405–414. <https://doi.org/10.1016/j.bcp.2014.08.031>
- Cooper, M. D. (2015). The early history of B cells. *Nature Reviews. Immunology*, 15(3), 191–197. <https://doi.org/10.1038/nri3801>
- Crigler, J. F., Jr, & Najjar, V. A. (1952). Congenital familial nonhemolytic jaundice with kernicterus. *Pediatrics*, 10(2), 169–180. <https://www.ncbi.nlm.nih.gov/pubmed/12983120>
- Cummings, B. B., Karczewski, K. J., Kosmicki, J. A., Seaby, E. G., Watts, N. A., Singer-Berk, M., Mudge, J. M., Karjalainen, J., Satterstrom, F. K., O'Donnell-Luria, A. H., Potterba, T., Seed, C., Solomonson, M., Alföldi, J., Genome Aggregation Database Production Team, Genome Aggregation Database Consortium, Daly, M. J., & MacArthur, D. G. (2020). Transcript expression-aware annotation improves rare variant interpretation. *Nature*, 581(7809), 452–458. <https://doi.org/10.1038/s41586-020-2329-2>
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2). <https://doi.org/10.1093/gigascience/giab008>
- Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D., & Marioni, J. C. (2021). Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology*, 40(2), 245–253. <https://doi.org/10.1038/s41587-021-01033-z>
- Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., ... Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, 48(10), 1284–1287. <https://doi.org/10.1038/ng.3656>
- DePasquale, E. A. K., Schnell, D. J., Van Camp, P.-J., Valiente-Alandí, Í., Blaxall, B. C., Leighton Grimes, H., Singh, H., & Salomonis, N. (2019). DoubletDecon: Deconvoluting

- Doublets from Single-Cell RNA-Sequencing Data. *Cell Reports*, 29(6), 1718–1727.e8. <https://doi.org/10.1016/j.celrep.2019.09.082>
- DepMap. (2021). *DepMap Data Downloads*. <https://depmap.org/portal/download/>
- De Sanctis, V., Kattamis, C., Canatan, D., Soliman, A. T., Elsedfy, H., Karimi, M., Daar, S., Wali, Y., Yassin, M., Soliman, N., Sobti, P., Al Jaouni, S., El Kholy, M., Fiscina, B., & Angastiniotis, M. (2017). β -Thalassemia Distribution in the Old World: an Ancient Disease Seen from a Historical Standpoint. *Mediterranean Journal of Hematology and Infectious Diseases*, 9(1), e2017018. <https://doi.org/10.4084/MJHID.2017.018>
- de Weerd, N. A., & Nguyen, T. (2012). The interferons and their receptors--distribution and regulation. *Immunology and Cell Biology*, 90(5), 483–491. <https://doi.org/10.1038/icb.2012.9>
- Dharia, A., Khan, A., Sridhar, V. S., & Cherney, D. Z. I. (2023). SGLT2 Inhibitors: The Sweet Success for Kidneys. *Annual Review of Medicine*, 74, 369–384. <https://doi.org/10.1146/annurev-med-042921-102135>
- Ding, R., Wang, Q., Gong, L., Zhang, T., Zou, X., Xiong, K., Liao, Q., Plass, M., & Li, L. (2023). scQTLbase: an integrated human single-cell eQTL database. *Nucleic Acids Research*, 52(D1), D1010–D1017. <https://doi.org/10.1093/nar/gkad781>
- Donnard, E., Shu, H., & Garber, M. (2022). Single cell transcriptomics reveals dysregulated cellular and molecular networks in a fragile X syndrome model. *PLoS Genetics*, 18(6), e1010221. <https://doi.org/10.1371/journal.pgen.1010221>
- Duan, D., Goemans, N., Takeda, S. 'ichi, Mercuri, E., & Aartsma-Rus, A. (2021). Duchenne muscular dystrophy. *Nature Reviews. Disease Primers*, 7(1), 13. <https://doi.org/10.1038/s41572-021-00248-3>
- Dubail, J., & Apte, S. S. (2015). Insights on ADAMTS proteases and ADAMTS-like proteins from mammalian genetics. *Matrix Biology: Journal of the International Society for Matrix Biology*, 44-46, 24–37. <https://doi.org/10.1016/j.matbio.2015.03.001>
- Eiz-Vesper, B., & Schmetzer, H. M. (2020). Antigen-Presenting Cells: Potential of Proven und New Players in Immune Therapies. *Transfusion Medicine and Hemotherapy: Offizielles Organ Der Deutschen Gesellschaft Fur Transfusionsmedizin Und Immunhamatologie*, 47(6), 429–431. <https://doi.org/10.1159/000512729>
- Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., & Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, 411(6836), 494–498. <https://doi.org/10.1038/35078107>
- Eraslan, G., Drokhlyansky, E., Anand, S., Fiskin, E., Subramanian, A., Slyper, M., Wang, J., Van Wittenberghe, N., Rouhana, J. M., Waldman, J., Ashenberg, O., Lek, M., Dionne, D., Win, T. S., Cuoco, M. S., Kuksenko, O., Tsankov, A. M., Branton, P. A., Marshall, J. L., ... Regev, A. (2022). Single-nucleus cross-tissue molecular reference maps toward

- understanding disease gene function. *Science*. <https://doi.org/10.1126/science.abl4290>
- Eyking, A., Ferber, F., Köhler, S., Reis, H., & Cario, E. (2019). TRIM58 Restrains Intestinal Mucosal Inflammation by Negatively Regulating TLR2 in Myeloid Cells. *Journal of Immunology*, 203(6), 1636–1649. <https://doi.org/10.4049/jimmunol.1900413>
- 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*. <https://doi.org/10.1126/science.1246949>
- Fatumo, S., Chikowore, T., Choudhury, A., Ayub, M., Martin, A. R., & Kuchenbäcker, K. (2022). Diversity in Genomic Studies: A Roadmap to Address the Imbalance. *Nature Medicine*, 28(2), 243. <https://doi.org/10.1038/s41591-021-01672-4>
- Fearnhead, N. S., Winney, B., & Bodmer, W. F. (2005). Rare variant hypothesis for multifactorial inheritance: susceptibility to colorectal adenomas as a model. *Cell Cycle*, 4(4), 521–525. <https://doi.org/10.4161/cc.4.4.1591>
- Fernández, J. M., de la Torre, V., Richardson, D., Royo, R., Puiggròs, M., Moncunill, V., Fragkogianni, S., Clarke, L., BLUEPRINT Consortium, Flicek, P., Rico, D., Torrents, D., Carrillo de Santa Pau, E., & Valencia, A. (2016). The BLUEPRINT Data Analysis Portal. *Cell Systems*, 3(5), 491–495.e5. <https://doi.org/10.1016/j.cels.2016.10.021>
- Finer, S., Martin, H. C., Khan, A., Hunt, K. A., MacLaughlin, B., Ahmed, Z., Ashcroft, R., Durham, C., MacArthur, D. G., McCarthy, M. I., Robson, J., Trivedi, B., Griffiths, C., Wright, J., Trembath, R. C., & van Heel, D. A. (2020). Cohort Profile: East London Genes & Health (ELGH), a community-based population genomics and health study in British Bangladeshi and British Pakistani people. *International Journal of Epidemiology*, 49(1), 20. <https://doi.org/10.1093/ije/dyz174>
- Finkel, R. S., Mercuri, E., Darras, B. T., Connolly, A. M., Kuntz, N. L., Kirschner, J., Chiriboga, C. A., Saito, K., Servais, L., Tizzano, E., Topaloglu, H., Tulinius, M., Montes, J., Glanzman, A. M., Bishop, K., Zhong, Z. J., Gheuens, S., Bennett, C. F., Schneider, E., ... De Vivo, D. C. (2017). Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy. *The New England Journal of Medicine*, 377(18). <https://doi.org/10.1056/NEJMoa1702752>
- FinnGen. (2022). *FinnGen Results Freeze 6*. FINNGEN_R6. <https://r6.finnngen.fi/>
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27–38. <https://doi.org/10.1093/biomet/80.1.27>
- Fisher, R. A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2), 399–433. <https://doi.org/10.1017/S0080456800012163>
- Fleming, S. J., Chaffin, M. D., Arduini, A., Akkad, A.-D., Banks, E., Marioni, J. C., Philippakis,

- A. A., Ellinor, P. T., & Babadi, M. (2023). Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nature Methods*, 20(9), 1323–1335. <https://doi.org/10.1038/s41592-023-01943-7>
- Folkersen, L., Gustafsson, S., Wang, Q., Hansen, D. H., Hedman, Å. K., Schork, A., Page, K., Zhernakova, D. V., Wu, Y., Peters, J., Eriksson, N., Bergen, S. E., Boutin, T. S., Bretherick, A. D., Enroth, S., Kalnapenkis, A., Gådin, J. R., Suur, B. E., Chen, Y., ... Mälarstig, A. (2020). Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nature Metabolism*, 2(10), 1135–1148. <https://doi.org/10.1038/s42255-020-00287-2>
- Forrest, I. S., Chaudhary, K., Vy, H. M. T., Petrazzini, B. O., Bafna, S., Jordan, D. M., Rocheleau, G., Loos, R. J. F., Nadkarni, G. N., Cho, J. H., & Do, R. (2022). Population-Based Penetrance of Deleterious Clinical Variants. *JAMA: The Journal of the American Medical Association*, 327(4), 350–359. <https://doi.org/10.1001/jama.2021.23686>
- Fotouhi-Ardakani, N., Kebir, D. E., Pierre-Charles, N., Wang, L., Ahern, S. P., Filep, J. G., & Milot, E. (2010). Role for myeloid nuclear differentiation antigen in the regulation of neutrophil apoptosis during sepsis. *American Journal of Respiratory and Critical Care Medicine*, 182(3), 341–350. <https://doi.org/10.1164/rccm.201001-0075OC>
- Frazer, K. A., Murray, S. S., Schork, N. J., & Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews. Genetics*, 10(4), 241–251. <https://doi.org/10.1038/nrg2554>
- Frosst, P., Blom, H. J., Milos, R., Goyette, P., Sheppard, C. A., Matthews, R. G., Boers, G. J., den Heijer, M., Kluijtmans, L. A., & van den Heuvel, L. P. (1995). A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. *Nature Genetics*, 10(1), 111–113. <https://doi.org/10.1038/ng0595-111>
- Fuchsberger, C., Abecasis, G. R., & Hinds, D. A. (2015). minimac2: faster genotype imputation. *Bioinformatics*, 31(5), 782–784. <https://doi.org/10.1093/bioinformatics/btu704>
- Futuyma, D. J. (1986). *Evolutionary Biology*. Sinauer Associates. <https://play.google.com/store/books/details?id=g6XuAAAAMAAJ>
- Gan, Y., Cao, C., Li, A., Song, H., Kuang, G., Ma, B., Zhang, Q., & Zhang, Q. (2021). Silencing of the TRIM58 Gene by Aberrant Promoter Methylation is Associated with a Poor Patient Outcome and Promotes Cell Proliferation and Migration in Clear Cell Renal Cell Carcinoma. *Frontiers in Molecular Biosciences*, 8, 655126. <https://doi.org/10.3389/fmolb.2021.655126>
- Garapati, P. V., Rudd, C., & de Bono, B. (2008). *PI3K bound to TRAT1 phosphorylates PIP2 to PIP3* [Dataset]. <https://reactome.org/content/detail/R-HSA-202365>
- Gayashani Sandamalika, W. M., Kwon, H., Lim, C., Yang, H., & Lee, J. (2021). The possible

- role of catalase in innate immunity and diminution of cellular oxidative stress: Insights into its molecular characteristics, antioxidant activity, DNA protection, and transcriptional regulation in response to immune stimuli in yellowtail clownfish (*Amphiprion clarkii*). *Fish & Shellfish Immunology*, 113, 106–117. <https://doi.org/10.1016/j.fsi.2021.03.022>
- Gebreyesus, G., Sahana, G., Christian Sørensen, A., Lund, M. S., & Su, G. (2020). Novel approach to incorporate information about recessive lethal genes increases the accuracy of genomic prediction for mortality traits. *Heredity*, 125(3), 155–166. <https://doi.org/10.1038/s41437-020-0329-5>
- Genes & Health. (2019, September). *Sept 2019: SUMMARY FILES, Exome sequencing loss of function variant data from 8921 individuals (East London, Birmingham, Born In Bradford)*. <https://www.genesandhealth.org/research/scientific-data-downloads/sept-2019-summary-files-exome-sequencing-loss-function-variant>
- Genes & Health. (2024). *genesandhealth.org*. <https://www.genesandhealth.org/>
- Gerasimavicius, L., Livesey, B. J., & Marsh, J. A. (2022). Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nature Communications*, 13(1), 1–15. <https://doi.org/10.1038/s41467-022-31686-6>
- Germain, P.-L., Lun, A., Meixide, C. G., Macnair, W., & Robinson, M. D. (2022). Doublet identification in single-cell sequencing data using *scDbIFinder*. *F1000Research*, 10(979), 979. <https://doi.org/10.12688/f1000research.73600.2>
- Ghoussaini, M., Mountjoy, E., Carmona, M., Peat, G., Schmidt, E. M., Hercules, A., Fumis, L., Miranda, A., Carvalho-Silva, D., Buniello, A., Burdett, T., Hayhurst, J., Baker, J., Ferrer, J., Gonzalez-Uriarte, A., Jupp, S., Karim, M. A., Koscielny, G., Machlitt-Northen, S., ... Dunham, I. (2021). Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Research*, 49(D1), D1311–D1320. <https://doi.org/10.1093/nar/gkaa840>
- Ghoussaini, M., Nelson, M. R., & Dunham, I. (2023). Future prospects for human genetics and genomics in drug discovery. *Current Opinion in Structural Biology*, 80. <https://doi.org/10.1016/j.sbi.2023.102568>
- Giani, A. M., Gallo, G. R., Gianfranceschi, L., & Formenti, G. (2020). Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal*, 18, 9–19. <https://doi.org/10.1016/j.csbj.2019.11.002>
- Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews. Genetics*, 13(2), 135–145. <https://doi.org/10.1038/nrg3118>
- Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., Deng, C., Varusai, T., Ragueneau, E., Haider, Y., May, B., Shamovsky, V., Weiser, J., Brunson, T., Sanati, N., ... D'Eustachio, P. (2022).

- The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1), D687–D692. <https://doi.org/10.1093/nar/gkab1028>
- Goldberger, A., Hnilica, L. S., Casey, S. B., & Briggs, R. C. (1986). Properties of a nuclear protein marker of human myeloid cell differentiation. *The Journal of Biological Chemistry*, 261(10), 4726–4731. [https://doi.org/10.1016/S0021-9258\(17\)38562-9](https://doi.org/10.1016/S0021-9258(17)38562-9)
- Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews. Genetics*, 17(6), 333–351. <https://doi.org/10.1038/nrg.2016.49>
- Gorzynski, J. E., Goenka, S. D., Shafin, K., Jensen, T. D., Fisk, D. G., Grove, M. E., Spiteri, E., Pesout, T., Monlong, J., Baid, G., Bernstein, J. A., Ceresnak, S., Chang, P.-C., Christle, J. W., Chubb, H., Dalton, K. P., Dunn, K., Garalde, D. R., Guillory, J., ... Ashley, E. A. (2022). Ultrarapid Nanopore Genome Sequencing in a Critical Care Setting. *The New England Journal of Medicine*. <https://doi.org/10.1056/NEJMc2112090>
- Goubau, D., Deddouch, S., & Reis e Sousa, C. (2013). Cytosolic sensing of viruses. *Immunity*, 38(5), 855–869. <https://doi.org/10.1016/j.immuni.2013.05.007>
- Goyette, P., Pai, A., Milos, R., Frosst, P., Tran, P., Chen, Z., Chan, M., & Rozen, R. (1998). Gene structure of human and mouse methylenetetrahydrofolate reductase (MTHFR). *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, 9(8), 652–656. <https://doi.org/10.1007/s003359900838>
- GTEX Consortium, Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration & Visualization—EBI, Genome Browser Data Integration & Visualization—UCSC Genomics Institute, University of California Santa Cruz, Lead analysts:, Laboratory, Data Analysis & Coordinating Center (LDACC):, ... Montgomery. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675), 204–213. <https://doi.org/10.1038/nature24277>
- Guindo-Martínez, M., Amela, R., Bonàs-Guarch, S., Puiggròs, M., Salvoró, C., Miguel-Escalada, I., Carey, C. E., Cole, J. B., Rüeger, S., Atkinson, E., Leong, A., Sanchez, F., Ramon-Cortes, C., Ejarque, J., Palmer, D. S., Kurki, M., FinnGen Consortium, Aragam, K., Florez, J. C., ... Torrents, D. (2021). The impact of non-additive genetic associations on age-related complex diseases. *Nature Communications*, 12(1), 2436. <https://doi.org/10.1038/s41467-021-21952-4>
- Gujral, U. P., Pradeepa, R., Weber, M. B., Narayan, K. M. V., & Mohan, V. (2013). Type 2

- diabetes in South Asians: similarities and differences with white Caucasian and other populations. *Annals of the New York Academy of Sciences*, 1281(1), 51–63.
<https://doi.org/10.1111/j.1749-6632.2012.06838.x>
- Gu, L., Casserly, D., Brady, G., Carpenter, S., Bracken, A. P., Fitzgerald, K. A., Unterholzner, L., & Bowie, A. G. (2022). Myeloid cell nuclear differentiation antigen controls the pathogen-stimulated type I interferon cascade in human monocytes by transcriptional regulation of IRF7. *Nature Communications*, 13(1), 14.
<https://doi.org/10.1038/s41467-021-27701-x>
- Gupta, N., Roychoudry, S., Sticco, K. L., Hsu, P., Zhang, X., & Sheikh-Fayyaz, S. (2023). Study of the Utility of Myeloid Cell Nuclear Differentiation Antigen (MNDA) in the Diagnosis of Marginal Zone Lymphoma. *Applied Immunohistochemistry & Molecular Morphology: AIMM / Official Publication of the Society for Applied Immunohistochemistry*, 31(4), 217–223. <https://doi.org/10.1097/PAI.0000000000001115>
- Guttormsen, A. B., Ueland, P. M., Nesthus, I., Nygård, O., Schneede, J., Vollset, S. E., & Refsum, H. (11 1996). Determinants and vitamin responsiveness of intermediate hyperhomocysteinemia (> or = 40 micromol/liter). The Hordaland Homocysteine Study. *The Journal of Clinical Investigation*, 98(9), 2174–2183.
<https://doi.org/10.1172/JCI119024>
- Gyimesi, G., Pujol-Giménez, J., Kanai, Y., & Hediger, M. A. (2020). Sodium-coupled glucose transport, the SLC5 family, and therapeutically relevant inhibitors: from molecular discovery to clinical application. *Pflügers Archiv: European Journal of Physiology*, 472(9), 1177–1206. <https://doi.org/10.1007/s00424-020-02433-x>
- Habel, J. C., Zachos, F. E., Dapporto, L., Rödder, D., Radespiel, U., Tellier, A., & Schmitt, T. (2015). Population genetics revisited – towards a multidisciplinary research field. *Biological Journal of the Linnean Society. Linnean Society of London*, 115(1), 1–12.
<https://doi.org/10.1111/bij.12481>
- Hamilton company. (1972). *Analytical Chemistry*, 44(7), 29A.
<https://doi.org/10.1021/ac60315a721>
- Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(suppl_1), D514–D517.
<https://doi.org/10.1093/nar/gki033>
- Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., 3rd, Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., ... Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13), 3573–3587.e29.
<https://doi.org/10.1016/j.cell.2021.04.048>

- Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., & Satija, R. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 42(2), 293–304. <https://doi.org/10.1038/s41587-023-01767-y>
- Hasin, Y., Seldin, M., & Lusi, A. (2017). Multi-omics approaches to disease. *Genome Biology*, 18(1), 1–15. <https://doi.org/10.1186/s13059-017-1215-1>
- Heid, I. M., Jackson, A. U., Randall, J. C., Winkler, T. W., Qi, L., Steinthorsdottir, V., Thorleifsson, G., Zillikens, M. C., Speliotes, E. K., Mägi, R., Workalemahu, T., White, C. C., Bouatia-Naji, N., Harris, T. B., Berndt, S. I., Ingelsson, E., Willer, C. J., Weedon, M. N., Luan, J., 'an, ... Lindgren, C. M. (2010). Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature Genetics*, 42(11), 949–960. <https://doi.org/10.1038/ng.685>
- Heilbron, K., Mozaffari, S. V., Vacic, V., Yue, P., Wang, W., Shi, J., Jubb, A. M., Pitts, S. J., & Wang, X. (2021). Advancing drug discovery using the power of the human genome. *The Journal of Pathology*, 254(4), 418. <https://doi.org/10.1002/path.5664>
- Hemmings, B. A., & Restuccia, D. F. (2012). PI3K-PKB/Akt pathway. *Cold Spring Harbor Perspectives in Biology*, 4(9), a011189. <https://doi.org/10.1101/cshperspect.a011189>
- Heyne, H. O., Karjalainen, J., Karczewski, K. J., Lemmelä, S. M., Zhou, W., FinnGen, Havulinna, A. S., Kurki, M., Rehm, H. L., Palotie, A., & Daly, M. J. (2023). Mono- and biallelic variant effects on disease at biobank scale. *Nature*, 613(7944), 519–525. <https://doi.org/10.1038/s41586-022-05420-7>
- He, Z., & Russell, J. E. (2002). A human embryonic hemoglobin inhibits Hb S polymerization in vitro and restores a normal phenotype to mouse models of sickle cell disease. *Proceedings of the National Academy of Sciences of the United States of America*, 99(16), 10635–10640. <https://doi.org/10.1073/pnas.162269099>
- Higgs, D. R., Vickers, M. A., Wilkie, A. O., Pretorius, I. M., Jarman, A. P., & Weatherall, D. J. (1989). A review of the molecular genetics of the human alpha-globin gene cluster. *Blood*, 73(5), 1081–1104. <https://doi.org/10.1182/blood.V73.5.1081.1081>
- Hill, W. G., Goddard, M. E., & Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, 4(2), e1000008. <https://doi.org/10.1371/journal.pgen.1000008>
- Hinney, A., Vogel, C. I. G., & Hebebrand, J. (2010). From monogenic to polygenic obesity: recent advances. *European Child & Adolescent Psychiatry*, 19(3), 297–310. <https://doi.org/10.1007/s00787-010-0096-6>
- Hivert, V., Sidorenko, J., Rohart, F., Goddard, M. E., Yang, J., Wray, N. R., Yengo, L., & Visscher, P. M. (2021). Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *American Journal of Human*

- Genetics*, 108(5), 786. <https://doi.org/10.1016/j.ajhg.2021.02.014>
- Hoffman, H. M., Mueller, J. L., Broide, D. H., Wanderer, A. A., & Kolodner, R. D. (2001). Mutation of a new gene encoding a putative pyrin-like protein causes familial cold autoinflammatory syndrome and Muckle–Wells syndrome. *Nature Genetics*, 29(3), 301–305. <https://doi.org/10.1038/ng756>
- Horton, J. D., Cohen, J. C., & Hobbs, H. H. (2007). Molecular biology of PCSK9: its role in LDL metabolism. *Trends in Biochemical Sciences*, 32(2), 71–77. <https://doi.org/10.1016/j.tibs.2006.12.008>
- Huang, Q. Q., Sallah, N., Dunca, D., Trivedi, B., Hunt, K. A., Hodgson, S., Lambert, S. A., Arciero, E., Wright, J., Griffiths, C., Trembath, R. C., Hemingway, H., Inouye, M., Finer, S., van Heel, D. A., Lumbers, R. T., Martin, H. C., & Kuchenbaecker, K. (2022). Transferability of genetic loci and polygenic scores for cardiometabolic traits in British Pakistani and Bangladeshi individuals. *Nature Communications*, 13(1), 4664. <https://doi.org/10.1038/s41467-022-32095-5>
- Huang, X., & Huang, Y. (2021). Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics*, 37(23), 4569–4571. <https://doi.org/10.1093/bioinformatics/btab358>
- 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), 1–12. <https://doi.org/10.1186/s13059-019-1865-2>
- Hübener, C., Mincheva, A., Lichter, P., Schraven, B., & Bruyns, E. (2000). Genomic organization and chromosomal localization of the human gene encoding the T-cell receptor-interacting molecule (TRIM). *Immunogenetics*, 51(2), 154–158. <https://doi.org/10.1007/s002510050024>
- Hunter, D. J. (2005). Gene–environment interactions in human diseases. *Nature Reviews. Genetics*, 6(4), 287–298. <https://doi.org/10.1038/nrg1578>
- Hustad, S. (2007). The methylenetetrahydrofolate reductase 677C-T polymorphism as a modulator of a B vitamin network with major effects on homocysteine metabolism. *Am. J. Hum. Genet.*, 80, 846–855. [https://www.cell.com/ajhg/pdf/S0002-9297\(07\)60940-9.pdf](https://www.cell.com/ajhg/pdf/S0002-9297(07)60940-9.pdf)
- Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11), 801–811. <https://doi.org/10.1016/j.humimm.2021.02.012>
- Hu, X., Li, J., Fu, M., Zhao, X., & Wang, W. (2021). The JAK/STAT signaling pathway: from bench to clinic. *Signal Transduction and Targeted Therapy*, 6(1), 402. <https://doi.org/10.1038/s41392-021-00791-1>
- Hu, Y., Stilp, A. M., McHugh, C. P., Rao, S., Jain, D., Zheng, X., Lane, J., Méric de Bellefon, S., Raffield, L. M., Chen, M.-H., Yanek, L. R., Wheeler, M., Yao, Y., Ren, C., Broome,

- J., Moon, J.-Y., de Vries, P. S., Hobbs, B. D., Sun, Q., ... NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium. (2021). Whole-genome sequencing association analysis of quantitative red blood cell phenotypes: The NHLBI TOPMed program. *American Journal of Human Genetics*, 108(5), 874–893.
<https://doi.org/10.1016/j.ajhg.2021.04.003>
- Hwang, E. S., & Ho, I.-C. (2002). Regulation of thymocyte homeostasis by Fliz1. *Immunology*, 106(4), 464–469. <https://doi.org/10.1046/j.1365-2567.2002.01455.x>
- Icking, A., Matt, S., Opitz, N., Wiesenthal, A., Müller-Esterl, W., & Schilling, K. (2005). NOSTRIN functions as a homotrimeric adaptor protein facilitating internalization of eNOS. *Journal of Cell Science*, 118(Pt 21), 5059–5069.
<https://doi.org/10.1242/jcs.02620>
- International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., ... Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164), 851–861. <https://doi.org/10.1038/nature06258>
- Iowa Institute of Human Genetics. (2024). *Microarrays and Fees*. Iowa Institute of Human Genetics - Carver College of Medicine | The University of Iowa.
<https://humangenetics.medicine.uiowa.edu/genomics-division/microarray/microarrays-and-fees>
- Isaacs, A., & Lindenmann, J. (1957). Virus interference. I. The interferon. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character*. Royal Society, 147(927), 258–267. <https://doi.org/10.1098/rspb.1957.0048>
- Jackson, M., Marks, L., May, G. H. W., & Wilson, J. B. (2018). The genetic basis of disease. *Essays in Biochemistry*, 62(5), 643–723. <https://doi.org/10.1042/EBC20170053>
- Jacques, P. F., Bostom, A. G., Williams, R. R., Ellison, R. C., Eckfeldt, J. H., Rosenberg, I. H., Selhub, J., & Rozen, R. (1996). Relation between folate status, a common mutation in methylenetetrahydrofolate reductase, and plasma homocysteine concentrations. *Circulation*, 93(1), 7–9. <https://doi.org/10.1161/01.cir.93.1.7>
- Janeway, C. A., Jr. (1989). Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harbor Symposia on Quantitative Biology*, 54 Pt 1, 1–13.
<https://doi.org/10.1101/sqb.1989.054.01.003>
- Janeway, C. A., Jr, Travers, P., Walport, M., & Shlomchik, M. J. (2001). *The components of the immune system*. Garland Science. <https://www.ncbi.nlm.nih.gov/books/NBK27092/>
- Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., & Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, 51(12), 1749–1755. <https://doi.org/10.1038/s41588-019-0530-8>

- Jiang, R., Sun, T., Song, D., & Li, J. J. (2022). Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biology*, 23(1), 1–24.
<https://doi.org/10.1186/s13059-022-02601-5>
- Johnstone, R. W., & Trapani, J. A. (1999). Transcription and growth regulatory functions of the HIN-200 family of proteins. *Molecular and Cellular Biology*, 19(9), 5833–5838.
<https://doi.org/10.1128/MCB.19.9.5833>
- Jonkers, I. H., & Wijmenga, C. (2017). Context-specific effects of genetic variants associated with autoimmune disease. *Human Molecular Genetics*, 26(R2), R185–R192.
<https://doi.org/10.1093/hmg/ddx254>
- Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F., & Luo, Y. (2022). Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, 12(3). <https://doi.org/10.1002/ctm2.694>
- Kaddoura, R., Orabi, B., & Salam, A. M. (2020). PCSK9 Monoclonal Antibodies: An Overview. *Heart Views: The Official Journal of the Gulf Heart Association*, 21(2), 97–103. https://doi.org/10.4103/HEARTVIEWS.HEARTVIEWS_20_20
- Kaiser, V. B., Svinti, V., Prendergast, J. G., Chau, Y.-Y., Campbell, A., Patarcic, I., Barroso, I., Joshi, P. K., Hastie, N. D., Miljkovic, A., Taylor, M. S., Generation Scotland, UK10K, Enroth, S., Memari, Y., Kolb-Kokocinski, A., Wright, A. F., Gyllensten, U., Durbin, R., ... Wilson, J. F. (2015). Homozygous loss-of-function variants in European cosmopolitan and isolate populations. *Human Molecular Genetics*, 24(19), 5464–5474.
<https://doi.org/10.1093/hmg/ddv272>
- Kals, M., Nikopensius, T., Läll, K., Pärn, K., Sikka, T. T., Suvisaari, J., Salomaa, V., Ripatti, S., Palotie, A., Metspalu, A., Esko, T., Palta, P., & Mägi, R. (2019). Advantages of genotype imputation with ethnically matched reference panel for rare variant association analyses. In *bioRxiv* (p. 579201). <https://doi.org/10.1101/579201>
- Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1). <https://doi.org/10.1093/nar/28.1.27>
- Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4), 348–354.
<https://doi.org/10.1038/ng.548>
- Kang, S. S., Zhou, J., Wong, P. W., Kowalisyn, J., & Strokosch, G. (1988). Intermediate homocysteinemia: a thermolabile variant of methylenetetrahydrofolate reductase. *American Journal of Human Genetics*, 43(4), 414–421.
<https://www.ncbi.nlm.nih.gov/pubmed/3177384>
- Karczewski, K. (2020). *loftree*. Github. <https://github.com/konradjk/loftree>
- Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins,

- R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., ... MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809), 434–443. <https://doi.org/10.1038/s41586-020-2308-7>
- Karczewski, K. J., Solomonson, M., Chao, K. R., Goodrich, J. K., Tiao, G., Lu, W., Riley-Gillis, B. M., Tsai, E. A., Kim, H. I., Zheng, X., Rahimov, F., Esmaeeli, S., Grundstad, A. J., Reppell, M., Waring, J., Jacob, H., Sexton, D., Bronson, P. G., Chen, X., ... Neale, B. M. (2022). Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genomics*, 2(9), 100168. <https://doi.org/10.1016/j.xgen.2022.100168>
- Kartik Chundru, V., Zhang, Z., Walter, K., Lindsay, S., Danecek, P., Eberhardt, R. Y., Gardner, E. J., Malawsky, D. S., Wigdor, E. M., Torene, R., Retterer, K., Wright, C. F., McWalter, K., Sheridan, E., Firth, H. V., Hurles, M. E., Samocha, K. E., Ustach, V. D., & Martin, H. C. (2023). Federated analysis of the contribution of recessive coding variants to 29,745 developmental disorder patients from diverse populations. In *medRxiv* (p. 2023.07.24.23293070). <https://doi.org/10.1101/2023.07.24.23293070>
- Kato, H., Takeuchi, O., Mikamo-Satoh, E., Hirai, R., Kawai, T., Matsushita, K., Hiiragi, A., Dermody, T. S., Fujita, T., & Akira, S. (2008). Length-dependent recognition of double-stranded ribonucleic acids by retinoic acid–inducible gene-I and melanoma differentiation–associated gene 5. *The Journal of Experimental Medicine*, 205(7), 1601–1610. <https://doi.org/10.1084/jem.20080091>
- Kattamis, A., Forni, G. L., Aydinok, Y., & Viprakasit, V. (2020). Changing patterns in the epidemiology of β -thalassemia. *European Journal of Haematology*, 105(6), 692–703. <https://doi.org/10.1111/ejh.13512>
- Kell, A. M., & Gale, M., Jr. (2015). RIG-I in RNA virus recognition. *Virology*, 479-480, 110–121. <https://doi.org/10.1016/j.virol.2015.02.017>
- Kerimov, N., Hayhurst, J. D., Peikova, K., Manning, J. R., Walter, P., Kolberg, L., Samoviča, M., Sakthivel, M. P., Kuzmin, I., Trevanion, S. J., Burdett, T., Jupp, S., Parkinson, H., Papatheodorou, I., Yates, A. D., Zerbino, D. R., & Alasoo, K. (2021). A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nature Genetics*, 53(9), 1290–1299. <https://doi.org/10.1038/s41588-021-00924-w>
- 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>

- King, C. D., Rios, G. R., Green, M. D., & Tephly, T. R. (2000). UDP-glucuronosyltransferases. *Current Drug Metabolism*, 1(2), 143–161.
<https://doi.org/10.2174/1389200003339171>
- Kisak, P. F. (2016). *The Human Genome: The Complete Nucleic Acid Sequence for Human DNA*. Createspace Independent Publishing Platform.
https://books.google.com/books/about/The_Human_Genome.html?hl=&id=vJz9jwEACA
AJ
- Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., & Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720), 385–389.
<https://doi.org/10.1126/science.1109557>
- Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N., & Sergushichev, A. (2016). Fast gene set enrichment analysis. In *Bioinformatics* (No. biorxiv;060012v3). bioRxiv. <https://www.biorxiv.org/content/10.1101/060012v3>
- Kos, J., Nanut, M. P., Prunk, M., Sabotič, J., Dautović, E., & Jewett, A. (2018). Cystatin F as a regulator of immune cell cytotoxicity. *Cancer Immunology, Immunotherapy: CII*, 67(12), 1931–1938. <https://doi.org/10.1007/s00262-018-2165-5>
- Kotowski, I. K., Pertsemliadis, A., Luke, A., Cooper, R. S., Vega, G. L., Cohen, J. C., & Hobbs, H. H. (2006). A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. *American Journal of Human Genetics*, 78(3), 410–422.
<https://doi.org/10.1086/500615>
- Koyasu, S. (2003). The role of PI3K in immune cells. *Nature Immunology*, 4(4), 313–319.
<https://doi.org/10.1038/ni0403-313>
- Krishna, S., Andersson, A. M. C., Semsey, S., & Sneppen, K. (2006). Structure and function of negative feedback loops at the interface of genetic and metabolic networks. *Nucleic Acids Research*, 34(8), 2455. <https://doi.org/10.1093/nar/gkl140>
- Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., Donner, K. M., Reeve, M. P., Laivuori, H., Aavikko, M., Kaunisto, M. A., Loukola, A., Lahtela, E., Mattsson, H., Laiho, P., Della Briotta Parolo, P., Lehisto, A. A., Kanai, M., Mars, N., Rämö, J., ... Palotie, A. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*, 613(7944), 508–518. <https://doi.org/10.1038/s41586-022-05473-8>
- Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics. *Korean Journal of Anesthesiology*, 70(2), 144–156.
<https://doi.org/10.4097/kjae.2017.70.2.144>
- Lambert, S. A., Abraham, G., & Inouye, M. (2019). Towards clinical utility of polygenic risk scores. *Human Molecular Genetics*, 28(R2), R133–R142.

<https://doi.org/10.1093/hmg/ddz187>

- Lam, B. Y. H., Williamson, A., Finer, S., Day, F. R., Tadross, J. A., Gonçalves Soares, A., Wade, K., Sweeney, P., Bedenbaugh, M. N., Porter, D. T., Melvin, A., Ellacott, K. L. J., Lippert, R. N., Buller, S., Rosmaninho-Salgado, J., Dowsett, G. K. C., Ridley, K. E., Xu, Z., Cimino, I., ... O'Rahilly, S. (2021). MC3R links nutritional state to childhood growth and the timing of puberty. *Nature*, 599(7885), 436–441. <https://doi.org/10.1038/s41586-021-04088-9>
- Lamy, P., Grove, J., & Wiuf, C. (2011). A review of software for microarray genotyping. *Human Genomics*, 5(4), 304–309. <https://doi.org/10.1186/1479-7364-5-4-304>
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., ... International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921. <https://doi.org/10.1038/35057062>
- Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(Database issue), D980–D985. <https://doi.org/10.1093/nar/gkt1113>
- Lassen, F. H., Venkatesh, S. S., Baya, N., Zhou, W., Bloemendal, A., Neale, B. M., Kessler, B. M., Whiffin, N., Lindgren, C. M., & Palmer, D. S. (2023). Exome-wide evidence of compound heterozygous effects across common phenotypes in the UK Biobank. *medRxiv : The Preprint Server for Health Sciences*. <https://doi.org/10.1101/2023.06.29.23291992>
- Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2), 1–17. <https://doi.org/10.1186/gb-2014-15-2-r29>
- Lee, J. Y., & Hong, S.-H. (2020). Hematopoietic Stem Cells and Their Roles in Tissue Regeneration. *International Journal of Stem Cells*, 13(1), 1–12. <https://doi.org/10.15283/ijsc19127>
- Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-Variant Association Analysis: Study Designs and Statistical Tests. *American Journal of Human Genetics*, 95(1), 5. <https://doi.org/10.1016/j.ajhg.2014.06.009>
- 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.

<https://doi.org/10.1038/nature19057>

- Lettre, G., Lange, C., & Hirschhorn, J. N. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genetic Epidemiology*, 31(4), 358–362. <https://doi.org/10.1002/gepi.20217>
- Lewis, C. M. (2002). Genetic association studies: design, analysis and interpretation. *Briefings in Bioinformatics*, 3(2), 146–153. <https://doi.org/10.1093/bib/3.2.146>
- Lichou, F., & Trynka, G. (2020). Functional studies of GWAS variants are gaining momentum. *Nature Communications*, 11(1), 1–4. <https://doi.org/10.1038/s41467-020-20188-y>
- Liew, S.-C., & Gupta, E. D. (2015). Methylenetetrahydrofolate reductase (MTHFR) C677T polymorphism: epidemiology, metabolism and the associated diseases. *European Journal of Medical Genetics*, 58(1), 1–10. <https://doi.org/10.1016/j.ejmg.2014.10.004>
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In *arXiv [q-bio.GN]*. <http://arxiv.org/abs/1303.3997>
- Li, L., Huang, P., Sun, X., Wang, S., Xu, M., Liu, S., Feng, Z., Zhang, Q., Wang, X., Zheng, X., Dai, M., Bi, Y., Ning, G., Cao, Y., & Wang, W. (2021). The ChinaMAP reference panel for the accurate genotype imputation in Chinese populations. *Cell Research*, 31(12), 1308–1310. <https://doi.org/10.1038/s41422-021-00564-z>
- Lim, W. F., Muniandi, L., George, E., Sathar, J., Teh, L. K., & Lai, M. I. (2015). HbF in HbE/ β -thalassemia: A clinical and laboratory correlation. *Hematology*, 20(6), 349–353. <https://doi.org/10.1179/1607845414Y.0000000203>
- Lindeboom, R. G. H., Supek, F., & Lehner, B. (2016). The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nature Genetics*, 48(10), 1112. <https://doi.org/10.1038/ng.3664>
- Lin, M., Sutherland, D. R., Horsfall, W., Totty, N., Yeo, E., Nayar, R., Wu, X.-F., & Schuh, A. C. (2002). Cell surface antigen CD109 is a novel member of the alpha(2) macroglobulin/C3, C4, C5 family of thioester-containing proteins. *Blood*, 99(5), 1683–1691. <https://doi.org/10.1182/blood.v99.5.1683>
- Li, T., Zhang, Y., Patil, P., & Johnson, W. E. (2023). Overcoming the impacts of two-step batch effect correction on gene expression estimation and inference. *Biostatistics*, 24(3), 635–652. <https://doi.org/10.1093/biostatistics/kxab039>
- Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype imputation. *Annual Review of Genomics and Human Genetics*, 10, 387–406. <https://doi.org/10.1146/annurev.genom.9.081307.164242>
- Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R., Hernandez, R. D., Hubisz, M. J., Sninsky, J. J., White, T. J., Sunyaev, S. R., Nielsen, R., Clark, A. G., & Bustamante, C. D. (2008). Proportionally more deleterious genetic variation in European than in African

- populations. *Nature*, 451(7181), 994–997. <https://doi.org/10.1038/nature06611>
- Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., & L Price, A. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, 48(11), 1443–1448. <https://doi.org/10.1038/ng.3679>
- Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N., & Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3), 284–290. <https://doi.org/10.1038/ng.3190>
- Londe, A. C., Fernandez-Ruiz, R., Julio, P. R., Appenzeller, S., & Niewold, T. B. (2023). Type I Interferons in Autoimmunity: Implications in Clinical Phenotypes and Treatment Response. *The Journal of Rheumatology*, 50(9), 1103–1113. <https://doi.org/10.3899/jrheum.2022-0827>
- 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. <https://doi.org/10.1186/s13059-014-0550-8>
- Luo, W., Pant, G., Bhavnasi, Y. K., Blanchard, S. G., Jr, & Brouwer, C. (2017). Pathview Web: user friendly pathway visualization and data integration. *Nucleic Acids Research*, 45(W1), W501–W508. <https://doi.org/10.1093/nar/gkx372>
- MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J. K., Montgomery, S. B., Albers, C. A., Zhang, Z. D., Conrad, D. F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M. A., Banks, E., Hu, M., ... Tyler-Smith, C. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070), 823–828. <https://doi.org/10.1126/science.1215040>
- MacArthur, D. G., & Tyler-Smith, C. (2010). Loss-of-function variants in the genomes of healthy humans. *Human Molecular Genetics*, 19(R2), R125. <https://doi.org/10.1093/hmg/ddq365>
- Maddirevula, S., Kuwahara, H., Ewida, N., Shamseldin, H. E., Patel, N., Alzahrani, F., AlSheddi, T., AlObeid, E., Alenazi, M., Alsaif, H. S., Alqahtani, M., AlAli, M., Al Ali, H., Helaby, R., Ibrahim, N., Abdulwahab, F., Hashem, M., Hanna, N., Monies, D., ... Alkuraya, F. S. (2020). Analysis of transcript-deleterious variants in Mendelian disorders: implications for RNA-based diagnostics. *Genome Biology*, 21(1), 145. <https://doi.org/10.1186/s13059-020-02053-9>
- Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., Payne, A. J., Steinthorsdottir, V., Scott, R. A., Grarup, N., Cook, J. P., Schmidt, E. M., Wuttke, M., Sarnowski, C., Mägi, R., Nano, J., Gieger, C., Trompet, S., Lecoeur, C., ...

- McCarthy, M. I. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics*, 50(11), 1505–1513. <https://doi.org/10.1038/s41588-018-0241-6>
- Maher, B. (2008, November 5). *Personal genomes: The case of the missing heritability*. Nature Publishing Group UK. <https://doi.org/10.1038/456018a>
- Malawsky, D. S. (2021). *Population structure, consanguinity patterns, and clinical impact of homozygosity in British Bangladeshis and Pakistanis* (H. C. Martin (Ed.)) [Master of Philosophy]. University of Cambridge.
- Malawsky, D. S., van Walree, E., Jacobs, B. M., Heng, T. H., Huang, Q. Q., Sabir, A. H., Rahman, S., Sharif, S. M., Khan, A., Mirkov, M. U., Kuwahara, H., Gao, X., Alkuraya, F. S., Newman, W. G., Griffiths, C. J., Mathur, R., van Heel, D. A., Finer, S., O'Connell, J., & Martin, H. C. (2023). Influence of autozygosity on common disease risk across the phenotypic spectrum. *Cell*, 186(21), 4514. <https://doi.org/10.1016/j.cell.2023.08.028>
- Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22), 2867–2873. <https://doi.org/10.1093/bioinformatics/btq559>
- Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews. Genetics*, 11(7), 499–511. <https://doi.org/10.1038/nrg2796>
- Marshall, J. S., Warrington, R., Watson, W., & Kim, H. L. (2018). An introduction to immunology and immunopathology. *Allergy, Asthma, and Clinical Immunology: Official Journal of the Canadian Society of Allergy and Clinical Immunology*, 14(2), 1–10. <https://doi.org/10.1186/s13223-018-0278-1>
- Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4), 584–591. <https://doi.org/10.1038/s41588-019-0379-x>
- Martin, H. C., Jones, W. D., McIntyre, R., Sanchez-Andrade, G., Sanderson, M., Stephenson, J. D., Jones, C. P., Handsaker, J., Gallone, G., Bruntraeger, M., McRae, J. F., Prigmore, E., Short, P., Niemi, M., Kaplanis, J., Radford, E. J., Akawi, N., Balasubramanian, M., Dean, J., ... on behalf of the Deciphering Developmental Disorders Study. (2018). Quantifying the contribution of recessive coding variation to developmental disorders. *Science*. <https://doi.org/10.1126/science.aar6731>
- Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305–315. <https://doi.org/10.1016/j.jml.2017.01.001>
- Matzinger, P. (2002). The danger model: a renewed sense of self. *Science*, 296(5566), 301–305. <https://doi.org/10.1126/science.1071059>
- Mazidi, M., Rezaie, P., Gao, H.-K., & Kengne, A. P. (2017). Effect of Sodium-Glucose

- Cotransport-2 Inhibitors on Blood Pressure in People With Type 2 Diabetes Mellitus: A Systematic Review and Meta-Analysis of 43 Randomized Control Trials With 22 528 Patients. *Journal of the American Heart Association*, 6(6).
<https://doi.org/10.1161/JAHA.116.004007>
- Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., Habegger, L., Ferreira, M., Baras, A., Reid, J., Abecasis, G., Maxwell, E., & Marchini, J. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, 53(7), 1097–1103. <https://doi.org/10.1038/s41588-021-00870-7>
- McGinnis, C. S., Murrow, L. M., & Gartner, Z. J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems*, 8(4), 329–337.e4. <https://doi.org/10.1016/j.cels.2019.03.003>
- McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. In *arXiv [stat.ML]*. <http://arxiv.org/abs/1802.03426>
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), 1297–1303. <https://doi.org/10.1101/gr.107524.110>
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1), 122. <https://doi.org/10.1186/s13059-016-0974-4>
- Medzhitov, R., & Iwasaki, A. (2024). Exploring new perspectives in immunology. *Cell*, 187(9), 2079–2094. <https://doi.org/10.1016/j.cell.2024.03.038>
- Medzhitov, R., Preston-Hurlburt, P., & Janeway, C. A., Jr. (1997). A human homologue of the Drosophila Toll protein signals activation of adaptive immunity. *Nature*, 388(6640), 394–397. <https://doi.org/10.1038/41131>
- Menon, A. P., Moreno, B., Meraviglia-Crivelli, D., Nonatelli, F., Villanueva, H., Barainka, M., Zheleva, A., van Santen, H. M., & Pastor, F. (2023). Modulating T Cell Responses by Targeting CD3. *Cancers*, 15(4). <https://doi.org/10.3390/cancers15041189>
- Menzel, S., Garner, C., Rooks, H., Spector, T. D., & Thein, S. L. (2013). HbA2 levels in normal adults are influenced by two distinct genetic mechanisms. *British Journal of Haematology*, 160(1), 101–105. <https://doi.org/10.1111/bjh.12084>
- Métais, J.-Y., Doerfler, P. A., Mayuranathan, T., Bauer, D. E., Fowler, S. C., Hsieh, M. M., Katta, V., Keriwala, S., Lazzarotto, C. R., Luk, K., Neel, M. D., Perry, S. S., Peters, S. T., Porter, S. N., Ryu, B. Y., Sharma, A., Shea, D., Tisdale, J. F., Uchida, N., ... Weiss, M. J. (2019). Genome editing of HBG1 and HBG2 to induce fetal hemoglobin. *Blood Advances*, 3(21), 3379–3392. <https://doi.org/10.1182/bloodadvances.2019000820>

- Meyer, H. V. (2020). *plinkQC: R package for quality control of plink genetic datasets*. Github.
<https://github.com/meyer-lab-cshl/plinkQC>
- Michael Conn, P. (2011). *Animal Models of Molecular Pathology*. Academic Press.
<https://play.google.com/store/books/details?id=qmeyAM5x3HYC>
- Minikel, E. V., Karczewski, K. J., Martin, H. C., Cummings, B. B., Whiffin, N., Rhodes, D., Alföldi, J., Trembath, R. C., van Heel, D. A., Daly, M. J., Schreiber, S. L., & MacArthur, D. G. (2020). Evaluating drug targets through human loss-of-function genetic variation. *Nature*, 581(7809), 459–464. <https://doi.org/10.1038/s41586-020-2267-z>
- Miranda, R. N., Briggs, R. C., Shults, K., Kinney, M. C., Jensen, R. A., & Cousar, J. B. (1999). Immunocytochemical analysis of MNDA in tissue sections and sorted normal bone marrow cells documents expression only in maturing normal and neoplastic myelomonocytic cells and a subset of normal and neoplastic B lymphocytes. *Human Pathology*, 30(9), 1040–1049. [https://doi.org/10.1016/s0046-8177\(99\)90221-6](https://doi.org/10.1016/s0046-8177(99)90221-6)
- Mitt, M., Kals, M., Pärn, K., Gabriel, S. B., Lander, E. S., Palotie, A., Ripatti, S., Morris, A. P., Metspalu, A., Esko, T., Mägi, R., & Palta, P. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *European Journal of Human Genetics: EJHG*, 25(7), 869–876. <https://doi.org/10.1038/ejhg.2017.51>
- Mizrahi-Man, O., Woehrmann, M. H., Webster, T. A., Gollub, J., Bivol, A., Keeble, S. M., Aull, K. H., Mittal, A., Roter, A. H., Wong, B. A., & Schmidt, J. P. (2022). Novel genotyping algorithms for rare variants significantly improve the accuracy of Applied Biosystems™ Axiom™ array genotyping calls: Retrospective evaluation of UK Biobank array data. *PloS One*, 17(11). <https://doi.org/10.1371/journal.pone.0277680>
- Mohr, S., & Liew, C.-C. (2007). The peripheral-blood transcriptome: new insights into disease and risk assessment. *Trends in Molecular Medicine*, 13(10), 422–432.
<https://doi.org/10.1016/j.molmed.2007.08.003>
- Moore, C. M., Jacobson, S. A., & Fingerlin, T. E. (2019). Power and Sample Size Calculations for Genetic Association Studies in the Presence of Genetic Model Misspecification. *Human Heredity*, 84(6), 256–271. <https://doi.org/10.1159/000508558>
- Mulder, N., Lombard, Z., Owolabi, M. O., & Ofori-Acquah, S. F. (2020). *The Genetic and Environmental Basis for Diseases in Understudied Populations*. Frontiers Media SA.
<https://play.google.com/store/books/details?id=r88OEAAAQBAJ>
- Nandi, A., Yan, L.-J., Jana, C. K., & Das, N. (2019). Role of Catalase in Oxidative Stress- and Age-Associated Degenerative Diseases. *Oxidative Medicine and Cellular Longevity*, 2019, 9613090. <https://doi.org/10.1155/2019/9613090>
- Naparstek, Y., Cohen, I. R., Fuks, Z., & Vlodavsky, I. (1984). Activated T lymphocytes produce a matrix-degrading heparan sulphate endoglycosidase. *Nature*, 310(5974),

- 241–244. <https://doi.org/10.1038/310241a0>
- Narayan, K. M. V., & Kanaya, A. M. (2020). Why are South Asians prone to type 2 diabetes? A hypothesis based on underexplored pathways. *Diabetologia*, 63(6), 1103–1109. <https://doi.org/10.1007/s00125-020-05132-5>
- Nassir, N., Bankapur, A., Samara, B., Ali, A., Ahmed, A., Inuwa, I. M., Zarrei, M., Safizadeh Shabestari, S. A., AlBanna, A., Howe, J. L., Berdiev, B. K., Scherer, S. W., Woodbury-Smith, M., & Uddin, M. (2021). Single-cell transcriptome identifies molecular subtype of autism spectrum disorder impacted by de novo loss-of-function variants regulating glial cells. *Human Genomics*, 15(1), 68. <https://doi.org/10.1186/s40246-021-00368-7>
- Neale, B. (2018, August 1). *UK biobank GWAS round 2 results [released 1st August 2018]*. Neale Lab. <http://www.nealelab.is/uk-biobank/>
- Need, A. C., & Goldstein, D. B. (2009). Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics: TIG*, 25(11). <https://doi.org/10.1016/j.tig.2009.09.012>
- Negishi, H., Taniguchi, T., & Yanai, H. (2018). The Interferon (IFN) Class of Cytokines and the IFN Regulatory Factor (IRF) Transcription Factor Family. *Cold Spring Harbor Perspectives in Biology*, 10(11). <https://doi.org/10.1101/cshperspect.a028423>
- Ng, P. C., Levy, S., Huang, J., Stockwell, T. B., Walenz, B. P., Li, K., Axelrod, N., Busam, D. A., Strausberg, R. L., & Venter, J. C. (2008). Genetic variation in an individual human exome. *PLoS Genetics*, 4(8), e1000160. <https://doi.org/10.1371/journal.pgen.1000160>
- O'Connell, J., Yun, T., Moreno, M., Li, H., Litterman, N., Kolesnikov, A., Noblin, E., Chang, P.-C., Shastri, A., Dorfman, E. H., Shringarpure, S., 23andMe Research Team, Auton, A., Carroll, A., & McLean, C. Y. (2021). A population-specific reference panel for improved genotype imputation in African Americans. *Communications Biology*, 4(1), 1269. <https://doi.org/10.1038/s42003-021-02777-9>
- O'Connor, M. J., Schroeder, P., Huerta-Chagoya, A., Cortés-Sánchez, P., Bonàs-Guarch, S., Guindo-Martínez, M., Cole, J. B., Kaur, V., Torrents, D., Veerapen, K., Grarup, N., Kurki, M., Rundsten, C. F., Pedersen, O., Brandslund, I., Linneberg, A., Hansen, T., Leong, A., Florez, J. C., & Mercader, J. M. (2022). Recessive Genome-Wide Meta-analysis Illuminates Genetic Architecture of Type 2 Diabetes. *Diabetes*, 71(3), 554–565. <https://doi.org/10.2337/db21-0545>
- Oelen, R., de Vries, D. H., Brugge, H., Gordon, G., Vochteloo, M., BIOS Consortium, Ye, C. J., Westra, H.-J., Franke, L., & van der Wijst, M. G. P. (2021). Single-cell RNA-sequencing reveals widespread personalized, context-specific gene expression regulation in immune cells. In *bioRxiv* (p. 2021.06.04.447088). <https://doi.org/10.1101/2021.06.04.447088>
- Ogata, M. (1991). Acatalasemia. *Human Genetics*, 86(4), 331–340.

- <https://doi.org/10.1007/BF00201829>
- Okada, Y., Momozawa, Y., Ashikawa, K., Kanai, M., Matsuda, K., Kamatani, Y., Takahashi, A., & Kubo, M. (2015). Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese. *Nature Genetics*, 47(7), 798–802. <https://doi.org/10.1038/ng.3310>
- Ong, T., & Ramsey, B. W. (2023). Cystic Fibrosis: A Review. *JAMA: The Journal of the American Medical Association*, 329(21), 1859–1871. <https://doi.org/10.1001/jama.2023.8120>
- Ozols, M. (2023). *yascp: scRNA analysis pipeline*. Github. <https://github.com/wtsi-hgi/yascp>
- Palmer, D. S., Zhou, W., Abbott, L., Wigdor, E. M., Baya, N., Churchhouse, C., Seed, C., Potterba, T., King, D., Kanai, M., Bloemendal, A., & Neale, B. M. (2023). Analysis of genetic dominance in the UK Biobank. *Science*, 379(6639), 1341–1348. <https://doi.org/10.1126/science.abn8455>
- Pålsson-McDermott, E. M., & O'Neill, L. A. J. (2004). Signal transduction by the lipopolysaccharide receptor, Toll-like receptor-4. *Immunology*, 113(2), 153–162. <https://doi.org/10.1111/j.1365-2567.2004.01976.x>
- Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., & Schacht, A. L. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews. Drug Discovery*, 9(3), 203–214. <https://doi.org/10.1038/nrd3078>
- Pazoki, R., Vujkovic, M., Elliott, J., Evangelou, E., Gill, D., Ghanbari, M., van der Most, P. J., Pinto, R. C., Wielscher, M., Farlik, M., Zuber, V., de Knegt, R. J., Snieder, H., Uitterlinden, A. G., Lifelines Cohort Study, Lynch, J. A., Jiang, X., Said, S., Kaplan, D. E., ... VA Million Veteran Program. (2021). Genetic analysis in European ancestry individuals identifies 517 loci associated with liver enzymes. *Nature Communications*, 12(1), 2579. <https://doi.org/10.1038/s41467-021-22338-2>
- Pazokitoroudi, A., Chiu, A. M., Burch, K. S., Pasaniuc, B., & Sankararaman, S. (2021). Quantifying the contribution of dominance deviation effects to complex trait variation in biobank-scale data. *American Journal of Human Genetics*, 108(5), 799. <https://doi.org/10.1016/j.ajhg.2021.03.018>
- Peisley, A., Lin, C., Wu, B., Orme-Johnson, M., Liu, M., Walz, T., & Hur, S. (2011). Cooperative assembly and dynamic disassembly of MDA5 filaments for viral dsRNA recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 108(52), 21010–21015. <https://doi.org/10.1073/pnas.1113651108>
- Pelak, K., Shianna, K. V., Ge, D., Maia, J. M., Zhu, M., Smith, J. P., Cirulli, E. T., Fellay, J., Dickson, S. P., Gumbs, C. E., Heinzen, E. L., Need, A. C., Ruzzo, E. K., Singh, A., Campbell, C. R., Hong, L. K., Lornsen, K. A., McKenzie, A. M., Sobreira, N. L. M., ...

- Goldstein, D. B. (2010). The characterization of twenty sequenced human genomes. *PLoS Genetics*, 6(9), e1001111. <https://doi.org/10.1371/journal.pgen.1001111>
- Perišić Nanut, M., Sabotič, J., Jewett, A., & Kos, J. (2014). Cysteine cathepsins as regulators of the cytotoxicity of NK and T cells. *Frontiers in Immunology*, 5, 616. <https://doi.org/10.3389/fimmu.2014.00616>
- Perišić Nanut, M., Sabotič, J., Švajger, U., Jewett, A., & Kos, J. (2017). Cystatin F Affects Natural Killer Cell Cytotoxicity. *Frontiers in Immunology*, 8, 1459. <https://doi.org/10.3389/fimmu.2017.01459>
- Pestka, S., Krause, C. D., & Walter, M. R. (2004). Interferons, interferon-like cytokines, and their receptors. *Immunological Reviews*, 202, 8–32. <https://doi.org/10.1111/j.0105-2896.2004.00204.x>
- Plenge, R. M., Scolnick, E. M., & Altshuler, D. (2013). Validating therapeutic targets through human genetics. *Nature Reviews. Drug Discovery*, 12(8), 581–594. <https://doi.org/10.1038/nrd4051>
- Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, 538(7624), 161. <https://doi.org/10.1038/538161a>
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8), 904–909. <https://doi.org/10.1038/ng1847>
- Pritchard, J. K., & Cox, N. J. (2002). The allelic architecture of human disease genes: common disease–common variant... or not? *Human Molecular Genetics*, 11(20), 2417–2423. <https://doi.org/10.1093/hmg/11.20.2417>
- 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. <https://doi.org/10.1086/519795>
- Purwar, N., McGarry, J. M., Kostera, J., Pacheco, A. A., & Schmidt, M. (2011). Interaction of nitric oxide with catalase: structural and kinetic analysis. *Biochemistry*, 50(21), 4491–4503. <https://doi.org/10.1021/bi200130r>
- Rehwinkel, J., & Gack, M. U. (2020). RIG-I-like receptors: their regulation and roles in RNA sensing. *Nature Reviews. Immunology*, 20(9), 537–551. <https://doi.org/10.1038/s41577-020-0288-3>
- Reilly, M. P., Li, M., He, J., Ferguson, J. F., Stylianou, I. M., Mehta, N. N., Burnett, M. S., Devaney, J. M., Knouff, C. W., Thompson, J. R., Horne, B. D., Stewart, A. F. R., Assimes, T. L., Wild, P. S., Allayee, H., Nitschke, P. L., Patel, R. S., Myocardial Infarction Genetics Consortium, Wellcome Trust Case Control Consortium, ... Rader, D. J. (2011). Identification of ADAMTS7 as a novel locus for coronary atherosclerosis and

- association of ABO with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies. *The Lancet*, 377(9763), 383–392. [https://doi.org/10.1016/S0140-6736\(10\)61996-4](https://doi.org/10.1016/S0140-6736(10)61996-4)
- Richard Boland, C. (2017). Non-coding RNA: It's Not Junk. *Digestive Diseases and Sciences*, 62(5), 1107. <https://doi.org/10.1007/s10620-017-4506-1>
- Ritari, J., Hyvärinen, K., Clancy, J., FinnGen, Partanen, J., & Koskela, S. (2020). Increasing accuracy of HLA imputation by a population-specific reference panel in a FinnGen biobank cohort. *NAR Genomics and Bioinformatics*, 2(2), lqaa030. <https://doi.org/10.1093/nargab/lqaa030>
- 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–e47. <https://doi.org/10.1093/nar/gkv007>
- Rivas, M. A., Pirinen, M., Conrad, D. F., Lek, M., Tsang, E. K., Karczewski, K. J., Maller, J. B., Kukurba, K. R., DeLuca, D. S., Fromer, M., Ferreira, P. G., Smith, K. S., Zhang, R., Zhao, F., Banks, E., Poplin, R., Ruderfer, D. M., Purcell, S. M., Tukiainen, T., ... MacArthur, D. G. (2015). Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science*, 348(6235). <https://doi.org/10.1126/science.1261877>
- Robertson, A. (1952). The Effect of Inbreeding on the Variation Due to Recessive Genes. *Genetics*, 37(2). <https://doi.org/10.1093/genetics/37.2.189>
- Romeo, S., Kozlitina, J., Xing, C., Pertsemliadis, A., Cox, D., Pennacchio, L. A., Boerwinkle, E., Cohen, J. C., & Hobbs, H. H. (2008). Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nature Genetics*, 40(12), 1461–1465. <https://doi.org/10.1038/ng.257>
- Ruan, Y., Lin, Y.-F., Feng, Y.-C. A., Chen, C.-Y., Lam, M., Guo, Z., Initiatives, S. G. A., He, L., Sawa, A., Martin, A. R., Qin, S., Huang, H., & Ge, T. (2022). Improving Polygenic Prediction in Ancestrally Diverse Populations. *Nature Genetics*, 54(5), 573. <https://doi.org/10.1038/s41588-022-01054-7>
- Russell, J. E., & Liehaber, S. A. (1998). Reversal of lethal alpha- and beta-thalassemias in mice by expression of human embryonic globins. *Blood*, 92(9), 3057–3063. <https://www.ncbi.nlm.nih.gov/pubmed/9787139>
- Ryu, Y., Han, G. H., Jung, E., & Hwang, D. (2023). Integration of Single-Cell RNA-Seq Datasets: A Review of Computational Methods. *Molecules and Cells*, 46(2), 106. <https://doi.org/10.14348/molcells.2023.0009>
- Sahu, M., & Prasuna, J. G. (2016). Twin Studies: A Unique Epidemiological Tool. *Indian Journal of Community Medicine: Official Publication of Indian Association of Preventive & Social Medicine*, 41(3), 177. <https://doi.org/10.4103/0970-0218.183593>

- Sanger, F., Nicklen, S., & Coulson, A. R. (1992). DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnology*, 24, 104–108.
<https://www.ncbi.nlm.nih.gov/pubmed/1422003>
- satijalab.org. (2020, October 2). *Seurat - Guided Clustering Tutorial*.
https://satijalab.org/seurat/archive/v3.2/pbmc3k_tutorial.html
- Sattar, N., & Gill, J. M. R. (2015). Type 2 diabetes in migrant south Asians: mechanisms, mitigation, and management. *The Lancet Diabetes & Endocrinology*, 3(12), 1004–1016.
[https://doi.org/10.1016/S2213-8587\(15\)00326-5](https://doi.org/10.1016/S2213-8587(15)00326-5)
- Schoenborn, J. R., & Wilson, C. B. (2007). Regulation of Interferon- γ During Innate and Adaptive Immune Responses. In *Advances in Immunology* (Vol. 96, pp. 41–101). Academic Press. [https://doi.org/10.1016/S0065-2776\(07\)96002-2](https://doi.org/10.1016/S0065-2776(07)96002-2)
- Schork, A., Saynisch, J., Vosseler, A., Jaghutriz, B. A., Heyne, N., Peter, A., Häring, H.-U., Stefan, N., Fritsche, A., & Artunc, F. (2019). Effect of SGLT2 inhibitors on body composition, fluid status and renin-angiotensin-aldosterone system in type 2 diabetes: a prospective study using bioimpedance spectroscopy. *Cardiovascular Diabetology*, 18(1), 46. <https://doi.org/10.1186/s12933-019-0852-y>
- Schwarze, K., Buchanan, J., Taylor, J. C., & Wordsworth, S. (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 20(10), 1122–1130. <https://doi.org/10.1038/gim.2017.247>
- Scriber, C. R. (2001). *The Metabolic & Molecular Bases of Inherited Disease*. McGraw-Hill.
<https://play.google.com/store/books/details?id=1XLynQEACAAJ>
- Seaby, E. G., Rehm, H. L., & O'Donnell-Luria, A. (2021). Strategies to Uplift Novel Mendelian Gene Discovery for Improved Clinical Outcomes. *Frontiers in Genetics*, 12, 674295. <https://doi.org/10.3389/fgene.2021.674295>
- Seal, R. L., Braschi, B., Gray, K., Jones, T. E. M., Tweedie, S., Haim-Vilmovsky, L., & Bruford, E. A. (2023). Genenames.org: the HGNC resources in 2023. *Nucleic Acids Research*, 51(D1), D1003–D1009. <https://doi.org/10.1093/nar/gkac888>
- Seo, J. Y., Lee, J.-E., Chung, G. E., Shin, E., Kwak, M.-S., Yang, J. I., & Yim, J. Y. (2020). A genome-wide association study on liver enzymes in Korean population. *PloS One*, 15(2), e0229374. <https://doi.org/10.1371/journal.pone.0229374>
- Shi, H., Doench, J. G., & Chi, H. (2023). CRISPR screens for functional interrogation of immunity. *Nature Reviews. Immunology*, 23(6), 363–380.
<https://doi.org/10.1038/s41577-022-00802-4>
- Shubbrook, J. H., Bokaie, B. B., & Adkins, S. E. (2015). Empagliflozin in the treatment of type 2 diabetes: evidence to date. *Drug Design, Development and Therapy*, 9, 5793–5803.
<https://doi.org/10.2147/DDDT.S69926>

- Singer-Berk, M., Gudmundsson, S., Baxter, S., Seaby, E. G., England, E., Wood, J. C., Son, R. G., Watts, N. A., Karczewski, K. J., Harrison, S. M., MacArthur, D. G., Rehm, H. L., & O'Donnell-Luria, A. (2023). Advanced variant classification framework reduces the false positive rate of predicted loss-of-function variants in population sequencing data. *American Journal of Human Genetics*, 110(9).
<https://doi.org/10.1016/j.ajhg.2023.08.005>
- Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., Venkataraman, G. R., Wainberg, M., Ollila, H. M., Kiiskinen, T., Havulinna, A. S., Pirruccello, J. P., Qian, J., Shcherbina, A., FinnGen, Rodriguez, F., Assimes, T. L., Agarwala, V., Tibshirani, R., ... Rivas, M. A. (2021). Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nature Genetics*, 53(2), 185–194.
<https://doi.org/10.1038/s41588-020-00757-z>
- Slight-Webb, S., Thomas, K., Smith, M., Wagner, C. A., Macwana, S., Bylinska, A., Donato, M., Dvorak, M., Chang, S. E., Kuo, A., Cheung, P., Kalesinskas, L., Ganesan, A., Dermadi, D., Guthridge, C. J., DeJager, W., Wright, C., Foecke, M. H., Merrill, J. T., ... Guthridge, J. M. (2023). Ancestry-based differences in the immune phenotype are associated with lupus activity. *JCI Insight*, 8(16).
<https://doi.org/10.1172/jci.insight.169584>
- Smallwood, M. J., Nissim, A., Knight, A. R., Whiteman, M., Haigh, R., & Winyard, P. G. (2018). Oxidative stress in autoimmune rheumatic diseases. *Free Radical Biology & Medicine*, 125, 3–14. <https://doi.org/10.1016/j.freeradbiomed.2018.05.086>
- Smith, C. (2015). *DNA Microarrays: A Trusted Tool Keeps Evolving*.
<https://www.biocompare.com/Editorial-Articles/172195-DNA-Microarrays-A-Trusted-Tool-Keeps-Evolving/>
- Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, J. A. L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., ... Harris, L. W. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1), D977–D985. <https://doi.org/10.1093/nar/gkac1010>
- Soskic, B., Cano-Gamez, E., Smyth, D. J., Ambridge, K., Ke, Z., Matte, J. C., Bossini-Castillo, L., Kaplanis, J., Ramirez-Navarro, L., Lorenc, A., Nakic, N., Esparza-Gordillo, J., Rowan, W., Wille, D., Tough, D. F., Bronson, P. G., & Trynka, G. (2022). Immune disease risk variants regulate gene expression dynamics during CD4+ T cell activation. *Nature Genetics*, 54(6), 817–826. <https://doi.org/10.1038/s41588-022-01066-3>
- Soskic, B., Cano-Gamez, E., Smyth, D. J., Rowan, W. C., Nakic, N., Esparza-Gordillo, J., Bossini-Castillo, L., Tough, D. F., Larminie, C. G. C., Bronson, P. G., Willé, D., & Trynka, G. (2019). Chromatin activity at GWAS loci identifies T cell states driving

- complex immune diseases. *Nature Genetics*, 51(10), 1486–1493.
<https://doi.org/10.1038/s41588-019-0493-9>
- STEMCELL Technologies. (n.d.). *EasySep™ Cell Separation Technology*. Retrieved May 9, 2024, from <https://www.stemcell.com/products/brands/easysep-cell-separation.html>
- Suciu-Foca, N., Reed, E., Rubinstein, P., MacKenzie, W., Ng, A. K., & King, D. W. (1985). A late-differentiation antigen associated with the helper inducer function of human T cells. *Nature*, 318(6045), 465–467. <https://doi.org/10.1038/318465a0>
- Sun, B. B., Kurki, M. I., Foley, C. N., Mechakra, A., Chen, C.-Y., Marshall, E., Wilk, J. B., Biogen Biobank Team, Chahine, M., Chevalier, P., Christé, G., FinnGen, Palotie, A., Daly, M. J., & Runz, H. (2022). Genetic associations of protein-coding variants in human disease. *Nature*, 603(7899), 95–102. <https://doi.org/10.1038/s41586-022-04394-w>
- Sun, L., Wu, J., Du, F., Chen, X., & Chen, Z. J. (2013). Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. *Science*, 339(6121), 786–791. <https://doi.org/10.1126/science.1232458>
- Szeinberg, A., de Vries, A., Pinkhas, J., Djaldetti, M., & Ezra, R. (1963). A DUAL HEREDITARY RED BLOOD CELL DEFECT IN ONE FAMILY: HYPOCATALASEMIA AND GLUCOSE-6-PHOSPHATE DEHYDROGENASE DEFICIENCY. *Acta Geneticae Medicae et Gemellologiae*, 12, 247–255. <https://doi.org/10.1017/s1120962300016231>
- Takeuchi, A., Miyamoto, T., Yamaji, K., Masuho, Y., Hayashi, M., Hayashi, H., & Onozaki, K. (1995). A human erythrocyte-derived growth-promoting factor with a wide target cell spectrum: identification as catalase. *Cancer Research*, 55(7), 1586–1589.
<https://www.ncbi.nlm.nih.gov/pubmed/7882369>
- Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S.-B., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., ... Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, 590(7845), 290–299. <https://doi.org/10.1038/s41586-021-03205-y>
- Tamayo, P., Steinhardt, G., Liberzon, A., & Mesirov, J. P. (2016). The limitations of simple gene set enrichment analysis assuming gene independence. *Statistical Methods in Medical Research*, 25(1), 472–487.
- Tazawa, S., Yamato, T., Fujikura, H., Hiratochi, M., Itoh, F., Tomae, M., Takemura, Y., Maruyama, H., Sugiyama, T., Wakamatsu, A., Isogai, T., & Isaji, M. (2005). SLC5A9/SGLT4, a new Na⁺-dependent glucose transporter, is an essential transporter for mannose, 1,5-anhydro-D-glucitol, and fructose. *Life Sciences*, 76(9), 1039–1050.
<https://doi.org/10.1016/j.lfs.2004.10.016>
- Telek, E., Kengyel, A., & Bugyi, B. (2020). Myosin XVI in the Nervous System. *Cells*, 9(8).

<https://doi.org/10.3390/cells9081903>

- Thomas, M. C., & Cherney, D. Z. I. (2018). The actions of SGLT2 inhibitors on metabolism, renal function and blood pressure. *Diabetologia*, 61(10), 2098–2107.
<https://doi.org/10.1007/s00125-018-4669-0>
- Thoresen, D. T., Galls, D., Götte, B., Wang, W., & Pyle, A. M. (2023). A rapid RIG-I signaling relay mediates efficient antiviral response. *Molecular Cell*, 83(1), 90–104.e4.
<https://doi.org/10.1016/j.molcel.2022.11.018>
- Trajanoska, K., Bhérrer, C., Taliun, D., Zhou, S., Richards, J. B., & Mooser, V. (2023). From target discovery to clinical drug development with human genetics. *Nature*, 620(7975), 737–745. <https://doi.org/10.1038/s41586-023-06388-8>
- Truong, B., Hull, L. E., Ruan, Y., Huang, Q. Q., Hornsby, W., Martin, H., van Heel, D. A., Wang, Y., Martin, A. R., Hong Lee, S., & Natarajan, P. (2024). Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases. *Cell Genomics*, 4(4). <https://doi.org/10.1016/j.xgen.2024.100523>
- Tweedie, S., Braschi, B., Gray, K., Jones, T. E. M., Seal, R. L., Yates, B., & Bruford, E. A. (2021). Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Research*, 49(D1), D939–D946. <https://doi.org/10.1093/nar/gkaa980>
- Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., & Lappalainen, T. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1), 1–21. <https://doi.org/10.1038/s43586-021-00056-9>
- Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigartyo, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., ... Pontén, F. (2015). Proteomics. Tissue-based map of the human proteome. *Science*, 347(6220), 1260419. <https://doi.org/10.1126/science.1260419>
- UK Biobank: Resource 593. (n.d.). Retrieved August 12, 2024, from <https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=593>
- VanRaden, P. M., Null, D. J., & Al-Khudhair, A. S. (2022). 349. Recent advances and future needs in genotype imputation. In *Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP)* (pp. 1463–1466). Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-940-4_349
- van Rooij, F. J. A., Qayyum, R., Smith, A. V., Zhou, Y., Trompet, S., Tanaka, T., Keller, M. F., Chang, L.-C., Schmidt, H., Yang, M.-L., Chen, M.-H., Hayes, J., Johnson, A. D., Yanek, L. R., Mueller, C., Lange, L., Floyd, J. S., Ghanbari, M., Zonderman, A. B., ... Ganesh, S. K. (2017). Genome-wide Trans-ethnic Meta-analysis Identifies Seven Genetic Loci Influencing Erythrocyte Traits and a Role for RBPMS in Erythropoiesis. *American Journal of Human Genetics*, 100(1), 51–63.

- <https://doi.org/10.1016/j.ajhg.2016.11.016>
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., ... Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507), 1304–1351. <https://doi.org/10.1126/science.1058040>
- Vestal, B. E., Wynn, E., & Moore, C. M. (2022). ImerSeq: an R package for analyzing transformed RNA-Seq data with linear mixed effects models. *BMC Bioinformatics*, 23(1), 489. <https://doi.org/10.1186/s12859-022-05019-9>
- Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, 90(1), 7–24. <https://doi.org/10.1016/j.ajhg.2011.11.029>
- Visscher, P. M., Hill, W. G., & Wray, N. R. (2008). Heritability in the genomics era — concepts and misconceptions. *Nature Reviews. Genetics*, 9(4), 255–266. <https://doi.org/10.1038/nrg2322>
- 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. [https://www.cell.com/ajhg/pdf/S0002-9297\(17\)30240-9.pdf](https://www.cell.com/ajhg/pdf/S0002-9297(17)30240-9.pdf)
- Vissers, L. E. L. M., Gilissen, C., & Veltman, J. A. (2015). Genetic studies in intellectual disability and related disorders. *Nature Reviews. Genetics*, 17(1), 9–18. <https://doi.org/10.1038/nrg3999>
- Vivier, E., Tomasello, E., Baratin, M., Walzer, T., & Ugolini, S. (2008). Functions of natural killer cells. *Nature Immunology*, 9(5), 503–510. <https://doi.org/10.1038/ni1582>
- Vlodavsky, I., Friedmann, Y., Elkin, M., Aingorn, H., Atzmon, R., Ishai-Michaeli, R., Bitan, M., Pappo, O., Peretz, T., Michal, I., Spector, L., & Pecker, I. (1999). Mammalian heparanase: gene cloning, expression and function in tumor progression and metastasis. *Nature Medicine*, 5(7), 793–802. <https://doi.org/10.1038/10518>
- Vlodavsky, I., Ilan, N., & Sanderson, R. D. (2020). Forty Years of Basic and Translational Heparanase Research. *Advances in Experimental Medicine and Biology*, 1221, 3–59. https://doi.org/10.1007/978-3-030-34521-1_1
- Vujkovic, M., Keaton, J. M., Lynch, J. A., Miller, D. R., Zhou, J., Tcheandjieu, C., Huffman, J. E., Assimes, T. L., Lorenz, K., Zhu, X., Hilliard, A. T., Judy, R. L., Huang, J., Lee, K. M., Klarin, D., Pyarajan, S., Danesh, J., Melander, O., Rasheed, A., ... Saleheen, D. (2020). Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nature Genetics*, 52(7), 680–691. <https://doi.org/10.1038/s41588-020-0637-y>

- Wainschtein, P., Jain, D., Zheng, Z., TOPMed Anthropometry Working Group, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Adrienne Cupples, L., Shadyab, A. H., McKnight, B., Shoemaker, B. M., Mitchell, B. D., Psaty, B. M., Kooperberg, C., Liu, C.-T., Albert, C. M., Roden, D., Chasman, D. I., Darbar, D., Lloyd-Jones, D. M., Arnett, D. K., ... Visscher, P. M. (2022). Assessing the contribution of rare variants to complex trait heritability from whole genome sequence data. *Nature Genetics*, 54(3), 263. <https://doi.org/10.1038/s41588-021-00997-7>
- Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., ... Lander, E. S. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366). <https://doi.org/10.1126/science.280.5366.1077>
- Wang, J., Yang, F., Zhuang, J., Huo, Q., Li, J., & Xie, N. (2022). TRIM58 inactivates p53/p21 to promote chemoresistance via ubiquitination of DDX3 in breast cancer. *The International Journal of Biochemistry & Cell Biology*, 143, 106140. <https://doi.org/10.1016/j.biocel.2021.106140>
- Wang, K.-Y., Huang, R.-Y., Tong, X.-Z., Zhang, K.-N., Liu, Y.-W., Zeng, F., Hu, H.-M., & Jiang, T. (2019). Molecular and clinical characterization of TMEM71 expression at the transcriptional level in glioma. *CNS Neuroscience & Therapeutics*, 25(9), 965–975. <https://doi.org/10.1111/cns.13137>
- Wang, X. (2014). Firth logistic regression for rare variant association tests. *Frontiers in Genetics*, 5, 187. <https://doi.org/10.3389/fgene.2014.00187>
- Ward, S. G., & Cantrell, D. A. (2001). Phosphoinositide 3-kinases in T lymphocyte activation. *Current Opinion in Immunology*, 13(3), 332–338. [https://doi.org/10.1016/s0952-7915\(00\)00223-5](https://doi.org/10.1016/s0952-7915(00)00223-5)
- Weedon, M. N., Jackson, L., Harrison, J. W., Ruth, K. S., Tyrrell, J., Hattersley, A. T., & Wright, C. F. (2021). Use of SNP chips to detect rare pathogenic variants: retrospective, population based diagnostic evaluation. *BMJ*, 372. <https://doi.org/10.1136/bmj.n214>
- 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. <https://doi.org/10.1038/nature05911>
- West, A. B., Moore, D. J., Biskup, S., Bugayenko, A., Smith, W. W., Ross, C. A., Dawson, V. L., & Dawson, T. M. (2005). Parkinson's disease-associated mutations in leucine-rich repeat kinase 2 augment kinase activity. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16842–16847. <https://doi.org/10.1073/pnas.0507360102>

- Whiffin, N., Armean, I. M., Kleinman, A., Marshall, J. L., Minikel, E. V., Goodrich, J. K., Quaife, N. M., Cole, J. B., Wang, Q., Karczewski, K. J., Cummings, B. B., Francioli, L., Laricchia, K., Guan, A., Alipanahi, B., Morrison, P., Baptista, M. A. S., Merchant, K. M., Genome Aggregation Database Production Team, ... MacArthur. (2020). The effect of LRRK2 loss-of-function variants in humans. *Nature Medicine*, 26(6), 869–877. <https://doi.org/10.1038/s41591-020-0893-5>
- Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., Belbin, G. M., Bien, S. A., Cheng, I., Cullina, S., Hodonsky, C. J., Hu, Y., Huckins, L. M., Jeff, J., Justice, A. E., ... Carlson, C. S. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, 570(7762), 514–518. <https://doi.org/10.1038/s41586-019-1310-4>
- Wolock, S. L., Lopez, R., & Klein, A. M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems*, 8(4). <https://doi.org/10.1016/j.cels.2018.11.005>
- Wright, C. F., FitzPatrick, D. R., & Firth, H. V. (2018). Paediatric genomics: diagnosing rare disease in children. *Nature Reviews. Genetics*, 19(5), 253–268. <https://doi.org/10.1038/nrg.2017.116>
- Wright, C. F., West, B., Tuke, M., Jones, S. E., Patel, K., Laver, T. W., Beaumont, R. N., Tyrrell, J., Wood, A. R., Frayling, T. M., Hattersley, A. T., & Weedon, M. N. (2019). Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting. *American Journal of Human Genetics*, 104(2), 275–286. <https://doi.org/10.1016/j.ajhg.2018.12.015>
- Wright, S. (1934). An Analysis of Variability in Number of Digits in an Inbred Strain of Guinea Pigs. *Genetics*, 19(6), 506–536. <https://doi.org/10.1093/genetics/19.6.506>
- Xiang, H., Wu, Z., Wang, J., & Wu, T. (2021). Research progress, challenges and perspectives on PNPLA3 and its variants in Liver Diseases. *Journal of Cancer*, 12(19), 5929–5937. <https://doi.org/10.7150/jca.57951>
- Xiao, Q., & Lauschke, V. M. (2021). The prevalence, genetic complexity and population-specific founder effects of human autosomal recessive disorders. *Npj Genomic Medicine*, 6(1), 1–7. <https://doi.org/10.1038/s41525-021-00203-x>
- Xie, J., Briggs, J. A., & Briggs, R. C. (1998). Human hematopoietic cell specific nuclear protein MNDA interacts with the multifunctional transcription factor YY1 and stimulates YY1 DNA binding. *Journal of Cellular Biochemistry*, 70(4), 489–506. <https://www.ncbi.nlm.nih.gov/pubmed/9712147>
- Xie, J., Briggs, J. A., Morris, S. W., Olson, M. O., Kinney, M. C., & Briggs, R. C. (1997). MNDA binds NPM/B23 and the NPM-MLF1 chimera generated by the t(3;5) associated with myelodysplastic syndrome and acute myeloid leukemia. *Experimental Hematology*,

- 25(11), 1111–1117. <https://www.ncbi.nlm.nih.gov/pubmed/9328447>
- Xie, J., Briggs, J. A., Olson, M. O., Sipos, K., & Briggs, R. C. (1995). Human myeloid cell nuclear differentiation antigen binds specifically to nucleolin. *Journal of Cellular Biochemistry*, 59(4), 529–536. <https://doi.org/10.1002/jcb.240590412>
- Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., & Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2), 100–106. <https://doi.org/10.1038/ng.2876>
- Yngvadottir, B., Xue, Y., Searle, S., Hunt, S., Delgado, M., Morrison, J., Whittaker, P., Deloukas, P., & Tyler-Smith, C. (2009). A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense SNPs. *American Journal of Human Genetics*, 84(2), 224–234. <https://doi.org/10.1016/j.ajhg.2009.01.008>
- Yoneyama, M., Kikuchi, M., Matsumoto, K., Imaizumi, T., Miyagishi, M., Taira, K., Foy, E., Loo, Y.-M., Gale, M., Jr, Akira, S., Yonehara, S., Kato, A., & Fujita, T. (2005). Shared and unique functions of the DExD/H-box helicases RIG-I, MDA5, and LGP2 in antiviral innate immunity. *Journal of Immunology*, 175(5), 2851–2858. <https://doi.org/10.4049/jimmunol.175.5.2851>
- Young, A. I. (2019). Solving the missing heritability problem. *PLoS Genetics*, 15(6), e1008222. <https://doi.org/10.1371/journal.pgen.1008222>
- Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omics: A Journal of Integrative Biology*, 16(5), 284. <https://doi.org/10.1089/omi.2011.0118>
- Zaccardi, F., Webb, D. R., Htike, Z. Z., Youssef, D., Khunti, K., & Davies, M. J. (2016). Efficacy and safety of sodium-glucose co-transporter-2 inhibitors in type 2 diabetes mellitus: systematic review and network meta-analysis. *Diabetes, Obesity & Metabolism*, 18(8), 783–794. <https://doi.org/10.1111/dom.12670>
- Zaitlen, N., & Kraft, P. (2012). Heritability in the genome-wide association era. *Human Genetics*, 131(10), 1655. <https://doi.org/10.1007/s00439-012-1199-6>
- Zande, P. V., Siddiq, M. A., Hodgins-Davis, A., Kim, L., & Wittkopp, P. J. (2023). Active compensation for changes in TDH3 expression mediated by direct regulators of TDH3 in *Saccharomyces cerevisiae*. *PLoS Genetics*, 19(12). <https://doi.org/10.1371/journal.pgen.1011078>
- Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I. W., Abecasis, G. R., Almgren, P., Andersen, G., Ardlie, K., Boström, K. B., Bergman, R. N., Bonnycastle, L. L., Borch-Johnsen, K., Burt, N. P., Chen, H., Chines, P. S., Daly, M. J., ... Altshuler, D. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics*, 40(5), 638–645. <https://doi.org/10.1038/ng.120>

- Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., Bastarache, L. A., Wei, W.-Q., Denny, J. C., Lin, M., Hveem, K., Kang, H. M., Abecasis, G. R., Willer, C. J., & Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, 50(9), 1335–1341.
<https://doi.org/10.1038/s41588-018-0184-y>
- Zhu, Z., Bakshi, A., Vinkhuyzen, A. A. E., Hemani, G., Lee, S. H., Nolte, I. M., van Vliet-Ostaptchouk, J. V., Snieder, H., The LifeLines Cohort Study, Esko, T., Milani, L., Mägi, R., Metspalu, A., Hill, W. G., Weir, B. S., Goddard, M. E., Visscher, P. M., & Yang, J. (2015). Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits. *American Journal of Human Genetics*, 96(3), 377.
<https://doi.org/10.1016/j.ajhg.2015.01.001>
- Zimmermann, K., Opitz, N., Dedio, J., Renne, C., Muller-Esterl, W., & Oess, S. (2002). NOSTRIN: a protein modulating nitric oxide release and subcellular distribution of endothelial nitric oxide synthase. *Proceedings of the National Academy of Sciences of the United States of America*, 99(26), 17167–17172.
<https://doi.org/10.1073/pnas.252345399>
- Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4), 1193–1198. <https://doi.org/10.1073/pnas.1119675109>
- Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., & Lander, E. S. (2014). Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America*, 111(4), E455.
<https://doi.org/10.1073/pnas.1322563111>